

**UNIVERSIDADE FEDERAL DE PERNAMBUCO  
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA**

**GABRIEL DE FRANÇA PEREIRA E SILVA**

**ALGORITMOS PARA CLASSIFICAÇÃO,  
FILTRAGEM E TRANSCRIÇÃO DE IMAGENS DE  
DOCUMENTOS**

Recife, 31 de julho de 2014.

**UNIVERSIDADE FEDERAL DE PERNAMBUCO  
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA**

**ALGORITMOS PARA CLASSIFICAÇÃO,  
FILTRAGEM E TRANSCRIÇÃO DE IMAGENS DE  
DOCUMENTOS**

por

**GABRIEL DE FRANÇA PEREIRA E SILVA**

Tese submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco como parte dos requisitos para a obtenção do grau de Doutor em Engenharia Elétrica.

**ORIENTADOR: PROF RAFAEL DUEIRE LINS, PhD**

Recife, julho de 2014.

Gabriel de França Pereira e Silva, 2014

Catálogo na fonte  
Bibliotecária Margareth Malta, CRB-4 / 1198

S586a Silva, Gabriel de França Pereira e.  
Algoritmos para classificação, filtragem e transcrição de imagens de documentos / Gabriel de França Pereira e Silva. - Recife: O Autor, 2014.  
104 folhas, il., gráfs., tabs.

Orientador: Prof. Dr. Rafael Dueire Lins.  
Tese (Doutorado) – Universidade Federal de Pernambuco. CTG.  
Programa de Pós-Graduação em Engenharia Elétrica, 2014.  
Inclui Referências e Apêndices.

1. Engenharia Elétrica. 2. Processamento de imagens. 3. Classificação de imagens. 4. Filtragem de documentos. 5. Transcrição automática. I. Lins, Rafael Dueire. (Orientador). II. Título.

UFPE

621.3 CDD (22. ed.)

BCTG/2014-229



Universidade Federal de Pernambuco  
*Pós-Graduação em Engenharia Elétrica*

PARECER DA COMISSÃO EXAMINADORA DE DEFESA DE  
TESE DE DOUTORADO


**GABRIEL DE FRANÇA PEREIRA E SILVA**


TÍTULO


**“ALGORITMOS PARA CLASSIFICAÇÃO, FILTRAGEM E  
TRANSCRIÇÃO DE IMAGENS DE DOCUMENTOS”**

A comissão examinadora composta pelos professores: RAFAEL DUEIRE LINS, CIN/UFPE; VALDEMAR CARDOSO DA ROCHA JÚNIOR, DES/UFPE; RICARDO MENEZES CAMPELLO DE SOUZA, DES/UFPE; BYRON LEITE DANTAS BEZERRA, POLI/UPE e DANIEL MARQUES OLIVEIRA, CHEMTECH/BRASIL sob a presidência do primeiro, consideram o candidato **GABRIEL DE FRANÇA PEREIRA E SILVA APROVADO.**


Recife, 31 de julho de 2014.

  
**CECILIO JOSÉ LINS PIMENTEL**  
Coordenador do PPGE

  
**RAFAEL DUEIRE LINS**  
Orientador e Membro Titular Interno

  
**BYRON LEITE DANTAS BEZERRA**  
Membro Titular Externo

  
**VALDEMAR CARDOSO DA ROCHA JÚNIOR**  
Membro Titular Interno

  
**DANIEL MARQUES OLIVEIRA**  
Membro Titular Externo

  
**RICARDO MENEZES CAMPELLO DE SOUZA**  
Membro Titular Interno

*Dedico este trabalho  
aos meus pais,  
à minha esposa,  
à minha família.*

# AGRADECIMENTOS

Primeiramente agradeço aos meus pais, José e Libânia. Vocês me deram todo o apoio e amor que um filho poderia pedir; serei eternamente grato nunca esquecendo os exemplos de caráter e perseverança.

A Renata, minha esposa, pelo grande amor dedicado a mim e pela paciência, ajuda, conselhos, apoio e carinho neste percurso.

Ao professor Rafael Dueire Lins, meus sinceros agradecimentos, por esses 10 anos de lições e ensinamentos não só no meio acadêmico, mas de vida.

À banca examinadora composta pelos pesquisadores Rafael Dueire Lins, Byron Leite Dantas Bezerra, Daniel Marques Oliveira, Ricardo Campello Menezes de Souza, Valdemar Cardoso da Rocha Junior pelas contribuições no desenvolvimento desta tese através de suas sugestões.

Aos Professores Hélio Magalhães de Oliveira e Cecílio José Lins Pimentel, pelo exemplo de profissionais que são.

A Andréa Tenório Pinto secretária da PPGEE pela paciência e ajuda.

Aos amigos da UFRPE Rodrigo Rocha, Ryan Ribeiro e Sérgio Mendonça companheiros de pesquisa e ensino, pela ajuda e sugestões.

Agradeço a minha avó, tios, primos e outros familiares, por sempre desejarem o melhor para mim. Em especial a minha Tia Leônia pelo carinho e apoio.

Agradeço aos amigos da pós-graduação que compartilharam as dificuldades nas disciplinas e no desenvolvimento desta tese.

Resumo da Tese apresentada à UFPE como parte dos requisitos necessários para a obtenção do grau de Doutor em Engenharia Elétrica.

# **ALGORITMOS PARA CLASSIFICAÇÃO, FILTRAGEM E TRANSCRIÇÃO DE IMAGENS DE DOCUMENTOS**

**GABRIEL DE FRANÇA PEREIRA E SILVA**

julho/2014

Orientador: Prof. Dr. Rafael Dueire Lins.

Área de Concentração: Telecomunicações.

Palavras-chave: processamento de imagens, classificação de imagens, filtragem de documentos, transcrição automática.

Número de Páginas: 113

O aumento dos esforços para digitalizar grandes coleções de documentos e a facilidade do uso de ferramentas de transcrição automática gerou uma maior heterogeneidade dos documentos digitalizados em termos de conteúdo, condições de preservação e qualidade de digitalização. O aumento da escala e complexidade exige a automatização do *workflow* relacionado ao processamento de imagens de documentos. Esta tese apresenta contribuições para os problemas de classificação automática, filtragem de ruídos e transcrição de documentos, com o objetivo de automatizar bases heterogêneas de imagens.

Abstract of Thesis presented to UFPE as a partial fulfillment of the requirements for the degree of Doctor in Electrical Engineering.

# **ALGORITMOS PARA CLASSIFICAÇÃO, FILTRAGEM E TRANSCRIÇÃO DE IMAGENS DE DOCUMENTOS**

**GABRIEL DE FRANÇA PEREIRA E SILVA**

July/2014

Supervisor: Prof. Dr. Rafael Dueire Lins.

Area of Concentration: Telecommunications.

Keywords: image processing, image classification, document filtering,

OCR. Number of pages:113

Increased efforts to digitize large collections of documents and the ease of using OCR tools generated a large heterogeneity of digital documents in terms of content, preservation conditions and digitization quality. The increasing scale and complexity requires automation of workflow related to document image processing. This thesis presents contributions to the problems of automatic classification, noise filtering and transcription of documents, with the goal of automating heterogeneous databases of images.



# SUMÁRIO

<b>CAPÍTULO 1</b>	<b>INTRODUÇÃO</b>	<b>16</b>
1.1	Objetivos	17
1.2	Contribuições	18
1.3	Imagens e Equipamentos utilizados	20
1.4	Organização da Tese	21
<b>CAPÍTULO 2</b>	<b>CLASSIFICAÇÃO</b>	<b>23</b>
2.1	Classificação de Imagens	23
2.1.1	Trabalhos Relacionados	24
2.1.2	Contribuições	26
2.1.3	Experimentos e Resultados	27
2.2	Classificação de Dispositivos de Captura	34
2.2.1	Trabalhos Relacionados	34
2.2.2	Contribuições	35
2.2.3	Experimentos e Resultados	36
2.3	Classificação e Caracterização de Ruídos	42
2.3.1	Trabalhos Relacionados	44
2.3.2	Contribuições	45
2.3.3	Experimentos e Resultados	46
<b>CAPÍTULO 3</b>	<b>FILTRAGEM DE RUÍDOS EM IMAGENS DE DOCUMENTOS</b>	<b>53</b>
3.1	Remoção de Borda	53
3.1.1	Trabalhos Relacionados	54
3.1.2	Algoritmo para Remoção de Borda	55
3.1.3	Experimentos e Resultados	56
3.2	Interferência Frente e Verso	60
3.2.1	Trabalhos relacionados	60

3.2.2	Algoritmo para filtragem de Interferência Frente e Verso.....	61
3.2.3	Experimentos e Resultados .....	62
<b>3.3</b>	<b>Remoção de Ruído Especular .....</b>	<b>65</b>
3.3.1	Trabalhos Relacionados.....	65
3.3.2	Algoritmo para Remoção do Ruído Especular .....	66
<b>3.4</b>	<b>Processamento Inteligente de Imagens de Documentos.....</b>	<b>70</b>
3.4.1	Trabalhos Relacionados.....	70
3.4.2	Método para Processamento Inteligente de Documentos .....	70
3.4.3	Experimentos e Resultados .....	73
 <b>CAPÍTULO 4 TRANSCRIÇÃO AUTOMÁTICA DE DOCUMENTOS HISTÓRICOS... 75</b>		
<b>4.1</b>	<b>Trabalhos Relacionados.....</b>	<b>75</b>
<b>4.2</b>	<b>Contribuições.....</b>	<b>77</b>
4.2.1	Método 1 (Reconstrução da Informação Textual) .....	77
4.2.2	Experimentos e Resultados .....	80
4.2.3	Método 2 (Geração automática de conjuntos de treinamento) .....	80
4.2.4	Experimentos e Resultados .....	82
 <b>CAPÍTULO 5 CONCLUSÕES E TRABALHOS FUTUROS ..... 85</b>		
<b>5.1</b>	<b>Trabalhos futuros.....</b>	<b>87</b>
 <b>REFERÊNCIAS..... 89</b>		
 <b>APÊNDICE A..... 105</b>		
<b>A.1</b>	<b>Publicações sobre Classificação em Imagens de documentos .....</b>	<b>105</b>
<b>A.2</b>	<b>Publicações sobre Classificação de Dispositivos de Captura.....</b>	<b>106</b>
<b>A.3</b>	<b>Publicações sobre Classificação de Ruído .....</b>	<b>107</b>
<b>A.4</b>	<b>Publicações sobre Ferramentas para Processamento de Imagens de Documentos.....</b>	<b>108</b>
<b>A.5</b>	<b>Publicações sobre remoção de Interferência Frente e Verso .....</b>	<b>109</b>

<b>A.6 Publicações sobre remoção de ruído Especular .....</b>	<b>110</b>
<b>A.7 Publicação sobre remoção de Embaçamento .....</b>	<b>111</b>
<b>A.8 Publicação sobre remoção de Bordas .....</b>	<b>112</b>
<b>A.9 Publicações sobre Transcrição automática de imagens de Documentos Históricos.....</b>	<b>113</b>

## LISTA DE FIGURAS

Figura 1.1 Aumento da heterogeneidade da digitalização de documentos.....	16
Figura 2.1 Classificação de imagens de documentos.....	24
Figura 2.2 Etapa de demosaico na formação de imagens digitais.....	26
Figura 2.3 Arquitetura em cascata de classificação.....	27
Figura 2.4 Exemplos de erros de classificação. ....	30
Figura 2.5 Precisão em relação aos classificadores e características usadas no treinamento. ....	31
Figura 2.6 Cobertura em relação aos classificadores e características usadas no treinamento. ....	31
Figura 2.7 <i>F-Measure</i> em relação aos classificadores e características usadas no treinamento.....	32
Figura 2.8 Classificação de imagens por diferentes tipos de dispositivos .....	34
Figura 2.9 Arquitetura em cascata de classificação.....	36
Figura 2.10 Medida de Precisão do classificador SVM em relação aos conjuntos de características.....	40
Figura 2.11 Medida de Cobertura do classificador SVM em relação aos conjuntos de características. ....	40
Figura 2.12 Medida de <i>F-Measure</i> do classificador SVM em relação aos conjuntos de características. ....	41
Figura 2.13 Classes de ruídos estudados .....	43
Figura 2.14 Arquitetura de classificação de ruído em paralelo.....	45
Figura 2.15 Geração de imagem com ruído sintético .....	46
Figura 3.1 Documento fotografado .....	54
Figura 3.2 Documento fotografado com as cinco janelas iniciais.....	55
Figura 3.3 Tipos de ruído de bordas .....	56
Figura 3.4 Trecho de documento com Interferência Frente e Verso .....	60
Figura 3.5 Exemplo de imagem sintética com diferentes níveis de <i>fade</i> .....	62
Figura 3.6 Exemplo de histograma de imagem de documento com interferência.....	62
Figura 3.7 Medida de Erro de Texto dos algoritmos de.....	63
Figura 3.8 Medida de Erro de Papel dos algoritmos de.....	63
Figura 3.9 Medida de Erro de Interferência dos algoritmos de.....	64
Figura 3.10 Resultado da filtragem de interferência .....	64
Figura 3.11 Exemplos de imagens com Ruído Especular.....	65
Figura 3.12 <i>Scanner</i> 3D HP TopShot Laser Printer.....	66
Figura 3.13 Sequência de imagens capturadas pelo.....	67
Figura 3.14 Geração da imagem filtrada a partir da sequência de imagens do .....	67
Figura 3.15 Resultado da remoção do ruído Especular em documentos.....	68
Figura 3.16 Exemplos de objetos capturados por meio de <i>scanner</i> 3D .....	68
Figura 3.17 Resultado da remoção do ruído Especular em objetos. ....	69
Figura 3.18 Geração de ruído sintético.....	71
Figura 3.19 Geração de casos.....	72

Figura 3.20 Trecho da tabela de casos do SRBC proposto. ....	73
Figura 4.1 Exemplo de reconstrução de informação textual .....	78
Figura 4.2 Arquitetura de reconstrução de áreas degradadas .....	79
Figura 4.3 Etapas de geração de <i>fontset</i> .....	81
Figura 4.4 Exemplos de características holísticas.....	81
Figura 4.5 Carta manuscrita de Joaquim Nabuco.....	82
Figura 4.6 Palavras originais x número de palavras corretamente transcritas .....	83
Figura 4.7 Exemplo de certidão de óbito da base de dados (TJPE/Family Search).....	83
Figura 5.1 Carta de George Washington.....	88

## LISTA DE TABELAS

Tabela 2.1 Imagens divididas por classes. ....	27
Tabela 2.2 Matriz de confusão do classificador Random Forest. ....	29
Tabela 2.3 Matriz de confusão do classificador KNN. ....	29
Tabela 2.4 Acurácia dos classificadores em relação ao conjunto de características. ....	32
Tabela 2.5 Acurácia dos classificadores em relação ao conjunto de características. ....	32
Tabela 2.6 Relação de Documentos Escaneados.....	36
Tabela 2.7 Relação de Documentos Fotografados no formato JPG <i>truecolor</i> .....	37
Tabela 2.8 Resultados do Classificador SVM para tarefa de classificação .....	38
Tabela 2.9 Matriz de confusão <i>scanner</i> x características de captura da câmera. ....	39
Tabela 2.10 Distribuição do rótulo dos blocos por classe.....	47
Tabela 2.11 Apresentação do quadro geral dos documentos utilizados nos experimentos de classificação de ruídos.....	48
Tabela 2.12 Resultado do classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para o ruído de Orientação.....	49
Tabela 2.13 Resultado do classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para o ruído de Inclinação. ....	49
Tabela 2.14 Resultado do classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para o ruído Especular.....	49
Tabela 2.15 Resultado do classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para o ruído de Borda.....	49
Tabela 2.16 Resultado do classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para o ruído de Sal e Pimenta.....	50
Tabela 2.17 Resultado do classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para o ruído de Interferência.....	50
Tabela 2.18 Resultado classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para o ruído de Embaçamento.....	50
Tabela 2.19 Resultado do classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para o ruído de Furo.....	50
Tabela 2.20 Distribuição do rotulo dos blocos por subclasse. ....	51
Tabela 2.21 Resultado do Classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para a intensidade do .....	51
Tabela 2.22 Resultado do Classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para a intensidade do .....	51
Tabela 2.23 Resultado do Classificador Random Forest sob as características de (LINS <i>et al.</i> , 2010d; SILVA e LINS, 2011) para a intensidade do .....	52

Tabela 3.1 Especificação e distribuição da base de dados.....	57
Tabela 3.2 Fator de aceitação ( $0 \leq f_e < 0,01$ ) para as técnicas de remoção de bordas.....	58
Tabela 3.3 Tempo de processamento dos métodos.....	59
Tabela 3.4 Imagens utilizadas para o experimento.....	73
Tabela 4.1 Resultado da transcrição dos campos dos certificados. ....	84

## **LISTA DE SÍMBOLOS**

3D - tridimensional

BW - Black and White (termo usado para imagens binárias 0 ou 1)

CF - Com Flash

GIF - Graphics Interchange Format (formato de arquivo de imagem)

HP - Hewlett Packard (Empresa Multinacional)

JPEG - Joint Photographic Experts Group (formato de arquivo de imagem)

KNN - K-Nearest Neighbors (algoritmo de classificação)

LED - Light Emitting Diode (fonte de emissão de luz)

MQI - Medidas de Qualidade de Imagem

OCR - Optical Character Recognition (transcrição automática de imagens de caracteres)

PLN - Processamento de Linguagem Natural

PNG - Portable Network Graphics (formato de arquivo de imagem)

RBC - Raciocínio Baseado em Casos (paradigma de aprendizagem de máquina)

RF - Random Forests (algoritmo de classificação)

RNRMD - Redes Neurais Recorrentes Multidimensionais

SF - Sem Flash

SRBC - Sistema de Raciocínio Baseado em Casos

SVM - Support Vector Machine (algoritmo de classificação)

TIFF - Tagged Image File Format (formato de arquivo de imagem)



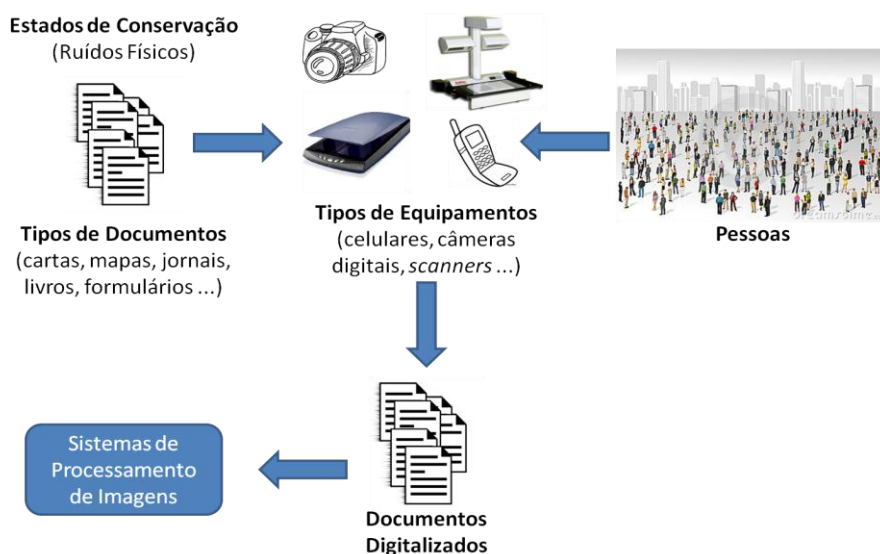
# Capítulo 1

## Introdução

Há séculos a humanidade utiliza o papel como meio de armazenamento e transmissão de informação. No entanto, o papel sofre desgaste natural pelo envelhecimento e pode ser danificado pelo manuseio incorreto. Além disso, o transporte físico de documentos pode demandar custo e tempo elevados. Essas desvantagens podem ser contornadas através da digitalização, que apresenta inúmeras vantagens, como a cópia e o armazenamento de documentos de maneira fácil.

O documento no formato digital, aliado à infraestrutura de redes de computadores, tal como a Internet, possibilita uma maior divulgação, com custo e tempo menores, comparado ao transporte físico (Lins *et al.*, 2006). Essa digitalização permite o uso de ferramentas de transcrição automática de texto (OCR), o que permite a extração de informações desses documentos de forma automatizada.

A redução de preço e o aumento da qualidade dos dispositivos de captura fomentou um novo cenário, em que a tarefa de digitalização passou a ser corriqueira nos mais diversos ambientes e seguimentos da sociedade (ver Figura 1.1). Esse cenário, aliado a popularização das ferramentas de OCRs, proporcionou um aumento na digitalização de documentos. Serviços como o Google Drive, que armazena diariamente dezenas de milhares de imagens fotografadas, escaneadas ou *print screen*, acoplaram em seu portfólio de produtos ferramentas de OCR (GOOGLE OCR, 2014).



**Figura 1.1** Aumento da heterogeneidade da digitalização de documentos.

Essas questões aumentaram o interesse na digitalização de grandes repositórios de documentos. Há uma série de projetos que continuam a exigir avanços na área de Engenharia de Documentos, como Impact (IMPACT, 2014), Kb Digitization (KB DIGITIZATION, 2014), Google Book (GOOGLE BOOK, 2014) e Million Book (MILLION BOOK, 2014). O cenário descrito representa um grande desafio, uma vez que implica na elaboração de algoritmos capazes de lidar com problemas relacionados à classificação, filtragem e transcrição de documentos digitalizados. Esta tese apresenta avanços na área de Engenharia de Documentos para automatização do processamento e análise de acervos heterogêneos de imagens. Desenvolver um método automático, que direcione e ajuste os algoritmos já existentes, é uma solução viável, como foi demonstrado em (SILVA *et al.*, 2010a; SILVA *et al.*, 2010b; SILVA *et al.*, 2010c; SILVA e LINS, 2011).

A primeira área de contribuição desta tese é no problema de classificação em imagens de documentos, que é dividido em três etapas: Classificação de Imagens; Classificação de Dispositivos de Captura; Classificação e Caracterização de Ruídos.

A segunda área de contribuição é o melhoramento da qualidade das imagens digitalizadas. São propostos algoritmos para remoção dos ruídos de borda, interferência frente e verso e especular. Um sistema de filtragem inteligente para imagens de documentos também é apresentado.

Em seguida, dois sistemas de transcrição automática para documentos históricos são apresentados. O primeiro trata o problema de imagens com texto impresso, enquanto o segundo trata texto manuscrito cursivo.

## 1.1 Objetivos

O objetivo desta tese é melhorar o processamento de acervos heterogêneos de documentos. Para atingir o objetivo, três etapas foram propostas:

- Classificação (imagens, dispositivos e ruídos), o que permite extrair informações relevantes para o ajuste e escolha de técnicas de processamento;
- Desenvolvimento de novos algoritmos para remoção de ruídos e de um sistema inteligente de filtragem;
- Transcrição automática das imagens de documentos históricos.

## 1.2 Contribuições

As contribuições de pesquisa desta tese são reportadas abaixo:

- Classificação em Imagens de documentos
  - (LINS *et al.*, 2009) – R. D. Lins; G. F. P. Silva, J. S. Simske, J. Fan, M. Shaw, P. Sá, M Thielo. Image Classification to Improve Printing Quality of Mixed-Type Documents. In: International Conference on Document Analysis and Recognition, pp: 1106-1110.
  - (GIANETTI *et al.*, 2010) – F. Gianetti; G. Dispoto ; R. D. Lins; G. F. P. Silva; G. Torreato; A. Cabeda. PDF profiling for B&W versus color pages cost estimation for efficient on-demand book printing. In: ACM-Symposium on Document Engineering, vol.1, pp: 177-188.
  - (SILVA e LINS, 2012a) G. F. P. Silva e R. D. Lins. Automatic Content Recognition of Teaching Boards in the Tableau Platform. In: International Conference on Pattern Recognition, vol.1, pp: 1-5.
  
- Classificação de Dispositivos de Captura
  - (SILVA *et al.*, 2009a) – Silva, G. F. P; Lins, R. D.; Miro B.; Simske, S.; Thielo, M. Automatically Deciding if a Document was Scanned or Photographed. In: Journal of Universal Computer Science, 2009, v.15, pp: 3364-3366.
  - (LINS *et al.*, 2011a) – R. D. Lins; G. F. P. Silva; J. S. Simske. Automatically Discriminating between Digital and Scanned Photographs. In: International Conference on Document Analysis and Recognition, vol.1, pp: 1280-1284.
  
- Classificação de Ruído
  - (SILVA *et al.*, 2010b) – G. F. P. Silva, R. D. Lins, S. Banergee, A. Kuchibhotla, M Thielo. Automatically Detecting and Classifying Noises in Document Images. In: ACM-Symposium on Applied Computing, vol.1, pp: 33-39.
  - (SILVA *et al.*, 2010c) – G. F. P. Silva e R. D. Lins; S. Banergee; A. Kuchibhotla; M Thielo. Enhancing the Filtering-out of the Back-to-Front

Interference in Color Documents with a Neural Classifier. In: International Conference on Pattern Recognition, vol.1, pp: 2415-2419.

- Ferramentas para Processamento de Imagens de Documentos
  - (SILVA *et al.*, 2010b) – G. F. P. Silva; R. D. Lins; J. M. M. Silva. HistDoc - A Toolbox for Processing Images of Historical Documents. Lecture Notes in Computer Science, vol.6112, pp: 409-419.
  - (LINS *et al.*, 2011b) – R. D. Lins; G. F. P. Silva; A. Formiga. HistDoc v. 2.0: enhancing a platform to process historical documents. In: Workshop on Historical Document Imaging and Processing, vol.1. pp: 169-176.
  
- Algoritmos para remoção de Interferência Frente e Verso
  - (SILVA *et al.*, 2009) J. M. Silva, R. D. Lins and G. F. P. Silva. Enhancing the Quality of Color Documents with Back-to-Front Interference. Image Analysis and Recognition, 1rd, Ed. Springer, pp: 875-885.
  - (SILVA *et al.*, 2010a) – G. F. P. Silva e R. D. Lins; S. Banerjee; A. Kuchibhotla; M Thielo. Enhancing the Filtering-out of the Back-to-Front Interference in Color Documents with a Neural Classifier. In: International Conference on Pattern Recognition, vol.1, pp: 2415-2419.
  
- Algoritmos para remoção de ruído Especular
  - (MARIANO *et al.*, 2011) E. Mariano, R. D. lins, G. F. P. Silva and J. Fan. Correcting Specular Noise in Multiple Images of Photographed Documents. In: International Conference on Document Analysis and Recognition, pp: 915-919.
  - (LINS *et al.*, 2013) R. D. Lins; G. F. P. Silva; E. Mariano, F. Fan, P. Majewicz and M. Thielo. Removing Shade and Specular Noise in Images of Objects and Documents Acquired with a 3D-Scanner. In: Lecture Notes in Computer Science, vol.1, pp: 299-307.
  
- Algoritmo para remoção de Embaçamento

- (OLIVEIRA *et al.*, 2011) – D. Oliveira; G. F. P. Silva e R. D. Lins. Deblurring Textual Document Images. In: The Ninth International Workshop on Graphics Recognition, vol.1, pp: 154-157.
- Algoritmo para remoção de Borda
  - (SILVA *et al.*, 2013) G. F. P. Silva ; R. D. Lins ; A. R. Silva. A New Algorithm for Background Removal of Document Images Acquired Using Portable Digital Cameras. In: Lecture Notes in Computer Science, Ed. Springer, vol.1, pp: 290-298.
- Transcrição automática de imagens de Documentos Históricos
  - (SILVA e LINS, 2011) – G. F. P. Silva; R. D. Lins. An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. In: International Conference on Document Analysis and Recognition, vol.1, pp: 553-557.
  - (ALMEIDA *et al.*, 2011) – A. B. S. Almeida; R.D. Lins; G. F. P. Silva. Thanatos: automatically retrieving information from death certificates in Brazil. In: Workshop on Historical Document Imaging and Processing, vol.1, pp: 146-153.
  - (SILVA e LINS, 2012b) – G. F. P. Silva; R. D. Lins. Generating Training Sets for the Automatic Recognition of Handwritten Documents. In: Advances in Character Recognition, 1ed, New York: InTech, 2012, pp: 155-174.
  - (SILVA e LINS, 2014) G. F. P. Silva; R. D. Lins. Automatic Training Set Generation for Better Historic Document Transcription and Compression. In: International Workshop on Document Analysis Systems, vol.1. pp: 20-31.

Os artigos listados acima se encontram no Apêndice A desta tese.

### **1.3 Imagens e Equipamentos utilizados**

Os experimentos dessa tese foram executados em um servidor Dell Power Edge 2900 (Intel Xeon Quad-Core 3.3 GHz com 16GB de memória). Para realizar os estudos e testes apresentados nesta tese, foram utilizadas imagens das seguintes bases:

- Projeto Nabuco (LINS, 2010) – imagens digitalizadas por *scanner* das cartas de Joaquim Nabuco. (Não disponível para distribuição)
- LiveMemory (LINS *et al.*, 2009) – imagens digitalizadas por *scanner* dos volumes dos anais da SBrT desde 1982. (DVD da tese)
- CBDAR 2007 Dewarping dataset (SHAFAIT e BREUEL, 2007) – documentos fotografados. (Direitos de distribuição controlada em: <http://www.dfki.uni-kl.de/~shafait/downloads.html>)
- HP Better-Printing (DES-UFPE e HP-USA) – Fotos, Documentos, Logotipos, Gráficos, Tabelas e Sintetizadas por computador. (Não disponível para distribuição)
- HP Multi-PiC (DES-UFPE e HP-India) – imagens de documentos contendo ruídos de borda, orienta, inclinação, sal e pimenta e interferência frente e verso. (Não disponível para distribuição)
- HP Better-Printing II (DES-UFPE e HP-USA) - imagens contendo ruídos de *warp*, especular e embaçamento. (Não disponível para distribuição)
- HP PROPRIO (DES-UFPE e HP-USA) - livros, manuais, revistas e relatórios.
- Cartórios do Estado de Pernambuco - Imagens digitalizadas de certidões de obituários e casamentos do Estado de Pernambuco (TJPE/ Family Search). (Não disponível para distribuição)
- BigBatch - Documentos escaneados (LINS *et al.*, 2006). (Não disponível para distribuição)

#### 1.4 Organização da Tese

Esta tese contém cinco capítulos. Em cada capítulo expõem-se soluções para os problemas mencionados na seção de objetivos. Os capítulos estão estruturados conforme descrito a seguir.

- Introdução sobre o problema a ser solucionado.
- Revisão bibliográfica dos trabalhos relacionados.
- Descrição das contribuições.
- Experimentos e Resultados.

Anexo a esta tese pode ser encontrado um DVD, que contém a base de imagens com os resultados de processamento de cada proposta, bem como os códigos fonte e executável na

plataforma Matlab, Java e C++ dos algoritmos desenvolvidos que não possuem restrições de confidencialidade.

## Capítulo 2

# Classificação

A tarefa de classificação de imagens atribui um ou mais rótulos de categoria para uma imagem. É um dos problemas fundamentais em Visão Computacional e Reconhecimento de Padrões, e possui variadas aplicações, por exemplo, vigilância de vídeo (COLLINS *et al.*, 2000), recuperação de imagem e vídeo (VAILAYA *et al.*, 2001), interação humano-computador (KOSALA e BLOCKEEL, 2000), biometria (JAIN *et al.*, 2004), impressão (LINS *et al.*, 2009) e tratamento de ruídos (SILVA e LINS, 2011).

O cenário apresentado na Introdução dessa tese destacou o aumento da heterogeneidade dos documentos (diversos tipos de documentos e estados de preservação) digitalizados por diferentes tipos de dispositivos (*scanners* e câmeras digitais). Essas coleções desafiam os pressupostos básicos do estado da arte em relação à qualidade, conteúdo e *layout*. Normalmente, as aplicações para processamento de documentos são limitadas a domínios restritos ou formatos regulares e assumem uma digitalização livre de ruídos (AGRAWAL e DOERMANN, 2011).

Neste capítulo é apresentado um método automático de classificação para o processamento de bases de imagens heterogêneas. O sistema consiste em três etapas de classificação (classificação de imagens, classificação de dispositivos e classificação de ruídos), que fornecem dados que permitem escolher e ajustar os algoritmos de filtragem (SILVA *et al.*, 2010a).

### 2.1 Classificação de Imagens

A recuperação de informações através de análise de imagens de documentos pode determinar quais metodologias de investigação são mais apropriadas ao processamento dessas imagens. Algumas aplicações necessitam processar diferencialmente imagens pertencentes a grupos distintos (LINS *et al.*, 2009; SILVA e LINS, 2012a). Em particular, documentos, fotografias, logotipos e gráficos exigem diferentes tratamentos para otimizar a sua aparência quando copiados ou impressos e este tratamento pode ser aplicado ao problema de filtragem de ruídos (CHOWDHURY *et al.*, 2003; PHAM, 2003; STROUTHOPOULOS *et al.*, 2002; ZHU *et al.*, 2006).



O problema reportado nesta tese foi proposto pela Hewlett Packard/Palo Alto-USA e seu objetivo foi informar às impressoras o que será impresso, a partir do estudo de seis classes (Figura 2.1).



(a) Foto



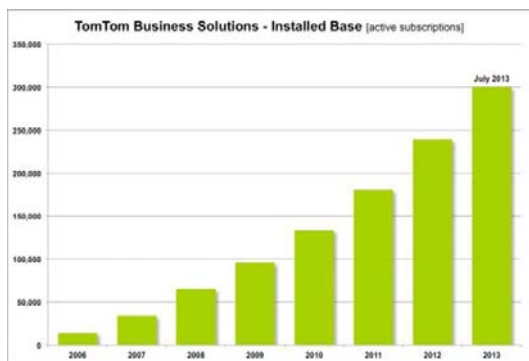
(b) Logotipo



(c) Sintetizada



(d) Documento



(e) Gráfico

Table 6. Effect of Roundup on the 1997 yield of lemon trees when Roundup was sprayed on the bottom 20 to 24 inches of the tree canopies. Data are means  $\pm$  standard deviations.

Roundup Rate	Lemon Tree Yield			
	11/4/96	12/2/96	Total Yield	First Harvest
(lb. a.i./acre)	(lb./tree)	(lb./tree)	(lb./tree)	(% of total)
0	85.6 $\pm$ 23.1	50.8 $\pm$ 17.4	136.4 $\pm$ 26.8	62.6 $\pm$ 10.6
0.5	81.9 $\pm$ 22.9	43.5 $\pm$ 16.9	125.4 $\pm$ 30.7	65.3 $\pm$ 9.2
0.75	82.7 $\pm$ 19.0	42.9 $\pm$ 16.1	125.7 $\pm$ 20.0	65.8 $\pm$ 11.9
1.0	79.9 $\pm$ 21.5	57.1 $\pm$ 20.6	137.0 $\pm$ 20.4	58.5 $\pm$ 12.9
1.25	81.0 $\pm$ 31.7	47.7 $\pm$ 18.0	128.5 $\pm$ 45.0	62.4 $\pm$ 8.7
1.5	81.6 $\pm$ 27.9	42.4 $\pm$ 13.7	124.0 $\pm$ 36.1	64.8 $\pm$ 9.7

(f) Tabela

Figura 2.1 Classificação de imagens de documentos

### 2.1.1 Trabalhos Relacionados

Agrupamento de imagens é tema de pesquisa na comunidade Banco de Dados desde a década de 90 e tem como objetivo a recuperação eficiente de informação (FRIGUI e KRISHNAPURAM, 2001; HEARST e PEDERSEN, 1996). A ideia básica é tentar

organizar as imagens em bases de dados usando algumas características semelhantes (HEARST e PEDERSEN, 1996; KRISHNAMACHARI e MOTTALEB, 1999).

Um dos métodos que tem apresentado maior sucesso na recuperação de imagem é a análise e agrupamento pelo histograma de cor (SCHEUNDERS, 1997; PARK *et al.*, 2002). A semântica de palavras e adaptações do modelo *bag-of-words* tem sido usadas para recuperação de imagens (BARNARD e FORSYTH, 2001; ÁVILA *et al.*, 2011). Esse método utiliza informações textuais para ajudar na classificação.

Recentemente, modelos baseados em *bag-of-features* receberam destaque devido à simplicidade, robustez e bom desempenho (HUANG *et al.*, 2014). A quantização vetorial de pequenas janelas (*keyblocks*) extraídas das imagens é utilizada como padrão de classificação (ZHU *et al.*, 2002). No entanto, os *keyblocks* não possuem propriedades de invariância. A ideia de *clusters* de descritores invariáveis pode ser utilizada para extrair características das imagens (CSURKA *et al.*, 2004).

Simske (2005) propõe um conjunto de características estatísticas e assume uma distribuição gaussiana para cada uma (SIMSKE, 2005). Nos trabalhos recentes (TIAN, 2013; HUANG *et al.*, 2014) são apresentadas revisões sobre as características utilizadas na classificação de imagens.

Por fim, a extração das características de histogramas usando wavelets é utilizada para resolver o problema de classificação entre imagens sintéticas e reais (WANG, 2006). A ideia é então estendida (DIRIK, 2007), incluindo novas funcionalidades para detectar o uso de matriz de cor do filtro de Bayer, durante a etapa de demosaico (BAYRAM *et al.*, 2005; BAYRAM *et al.*, 2006) (ver Figura 2.2). Essas características são utilizadas para capturar as irregularidades estatísticas de imagens reais (LYU e FARID, 2005).

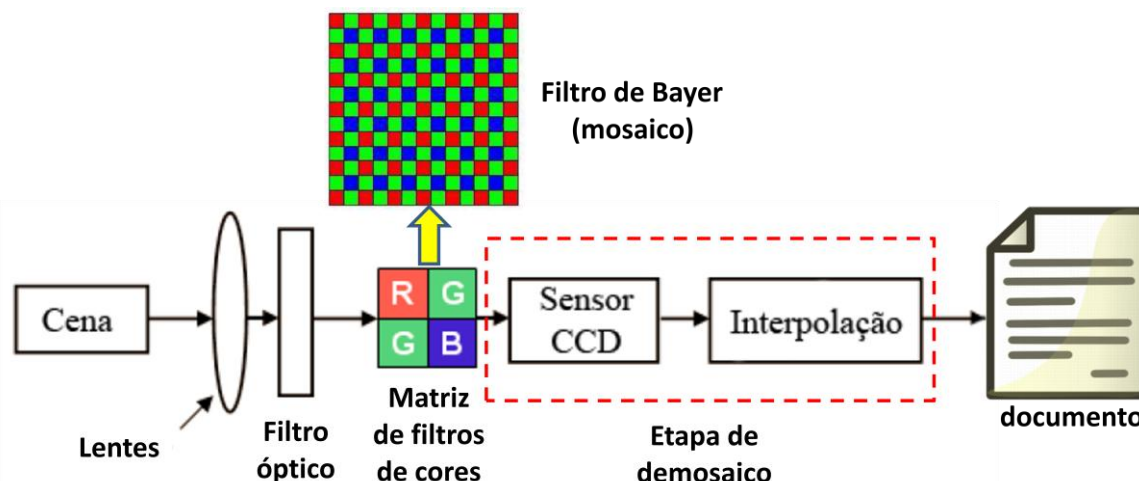


Figura 2.2 Etapa de demosaico na formação de imagens digitais.

### 2.1.2 Contribuições

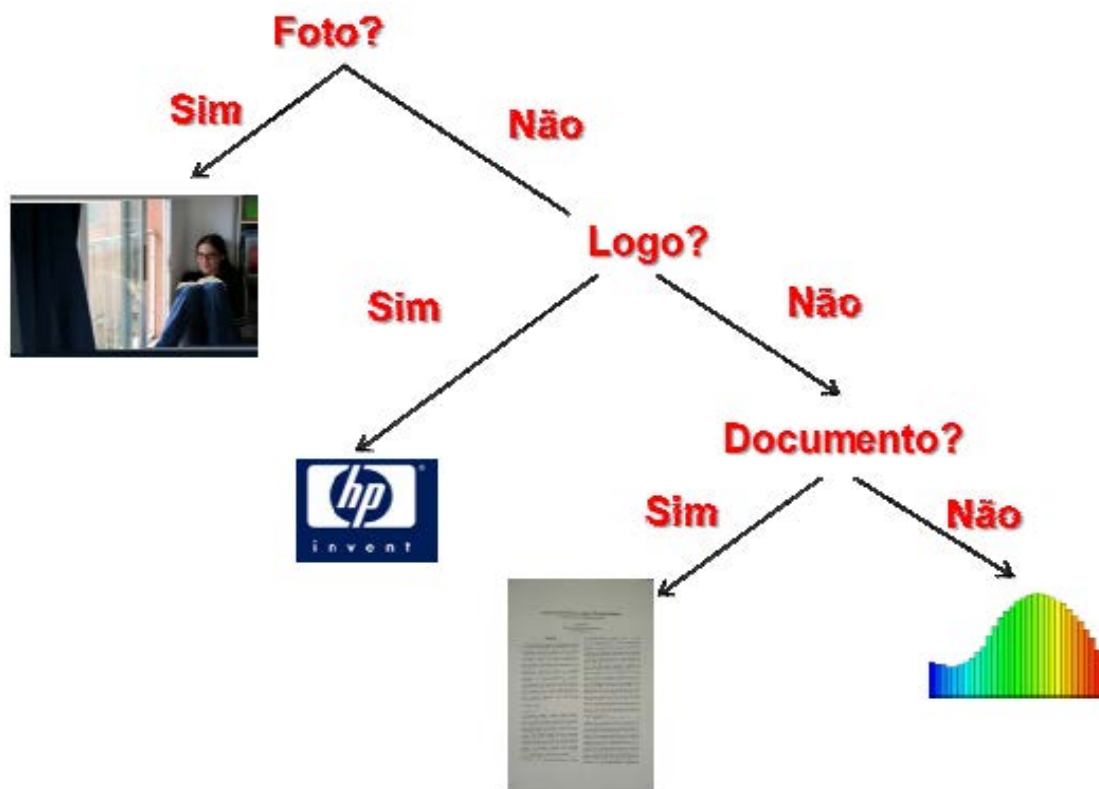
O trabalho de classificação de imagens desenvolvido nesta tese foi fruto de um projeto de P&D entre a Hewlett-Packard (HP) e o DES-UFPE e está em uso desde Janeiro de 2010 nas impressoras da HP. O reconhecimento automático foi proposto para seis classes (Foto, Logo, Documento, Sintética, Gráfico e Tabela) e possibilitou impressões de melhor qualidade.

O trabalho focou na extração de características de imagens e no desenvolvimento de um modelo de classificação. A descrição da solução foi publicada na *10th International Conference on Document Analysis and Recognition* (LINS *et al.*, 2009) e encontra-se disponível no Apêndice A desta tese.

O conjunto de características extraído é baseado na paleta da imagem:

- Palette (true-color/grayscale).
- Gamut (true-color/grayscale).
- Número de black pixels (OTSU, 1979).
- $(\#Black\_pixels/Total\_#\_pixels)*100\%$ .
- $(Gamut/Palette)*100\%$  (true-color/grayscale).

A arquitetura em cascata foi adotada, em que cada nó é um classificador binário (Figura 2.3).



**Figura 2.3 Arquitetura em cascata de classificação.**

### 2.1.3 Experimentos e Resultados

A base de dados utilizada nos experimentos é composta por 38.804 documentos digitalizados por diferentes tipos de dispositivos e em diferentes resoluções. Parte dessas imagens foram extraídas de documentos PDF, por uma ferramenta desenvolvida juntamente com a HP para o problema de distribuição de carga de impressão (GIANNETTI *et al.*, 2010). As distribuições das bases de dado são apresentadas na Tabela 2.1.

**Tabela 2.1 Imagens divididas por classes.**

Classes	Total
Fotos	10.000
Logotipo	5.000
Documento	17.804
Gráficos	2.000
Tabelas	1.000
Sintéticas	3.000
<b>Total</b>	<b>38.804</b>

A comparação entre os trabalhos de Lins (LINS *et al.*, 2009) e Simske (SIMSKE, 2005) é apresentada na forma da matriz de confusão dos classificadores e por meio das medidas de Precisão, Cobertura e *F-Measure*.

A matriz de confusão (STEHMAN, 1997) é um *layout* de tabela específico que permite visualizar o desempenho de um classificador. Cada coluna da matriz representa as instâncias de uma classe prevista, enquanto cada linha representa as instâncias de uma classe real. O nome deriva do fato que ela torna mais fácil visualizar se o sistema está confundindo duas ou mais classes.

Já as medidas de Precisão, Cobertura e *F-Measure* foram inicialmente propostas para avaliar sistemas de Recuperação de Informação (SALTON e BUCKLEY, 1988). Na tarefa de classificação, a Precisão de uma classe é o número de verdadeiros positivos (isto é, o número de itens corretamente marcados como pertencendo à classe positiva) divididos pelo número total de elementos marcados como pertencendo à classe positiva (a soma de verdadeiros positivos e falsos positivos, que são instância incorretamente identificados como pertencendo à classe). A Cobertura, neste contexto, é definida como o número de verdadeiros positivos dividido pelo número total de elementos que, na verdade, pertencem à classe positiva (soma dos verdadeiros positivos e falsos negativos, que são elementos que não foram marcados como pertencendo à classe positiva, mas deveriam ter sido). Já a *F-Measure* é dada por:  $2 \times \frac{\text{Precisão} \times \text{Cobertura}}{\text{Precisão} + \text{Cobertura}}$ .

Os classificadores KNN (AHA e KIBLER, 1991) e Random Forests (BREIMAN, 2001) foram utilizados para a avaliação da classificação. Para validar os classificadores, foi utilizada a técnica de Validação Cruzada. A Validação Cruzada é uma técnica usada para avaliar a capacidade de generalização de um modelo de classificação, a partir de um conjunto de dados (KOHAVI, 1995). Esta técnica é empregada em problemas onde o objetivo da modelagem é a predição. Ela busca estimar o quão acurado é o modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados. O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutuamente excludentes (*k-fold*), e posteriormente, utiliza-se alguns destes subconjuntos para estimar os parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo.

Para os experimentos, cada classificador foi gerado dez vezes, com diferentes configurações. O *Paired Corrected T-Tester* (significância  $p=0.05$ ) foi usado para verificar a diferença estatística entre os classificadores, em torno da medida F-Measure.

**Tabela 2.2 Matriz de confusão do classificador Random Forest.**

Modelo treinado com as características (LINS <i>et al.</i> , 2009)						
	Foto	Logotipo	Documento	Tabela	Gráfico	Sintética
Foto	9.881	98	16	0	0	5
Logotipo	100	4.854	21	0	0	25
Documento	27	10	17.763	3	1	0
Tabela	0	0	70	914	4	12
Gráfico	2	23	4	1	1.802	168
Sintética	82	13	1	0	98	2.806
Modelo treinado com as características (SIMSKE, 2005)						
Foto	8.544	932	216	107	98	103
Logotipo	1.203	3.459	189	30	10	109
Documento	1.922	2.488	11.358	1.576	259	201
Tabela	63	20	271	604	22	20
Gráfico	10	297	81	42	1.108	462
Sintética	202	420	32	21	337	1.988

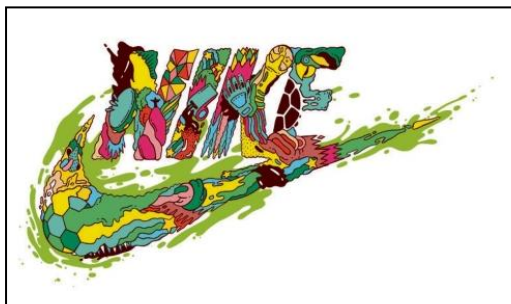
**Tabela 2.3 Matriz de confusão do classificador KNN.**

Modelo treinado com as características (LINS <i>et al.</i> , 2009)						
	Foto	Logotipo	Documento	Tabela	Gráfico	Sintética
Foto	9.866	112	16	1	0	5
Logotipo	100	4.854	21	0	0	25
Documento	27	10	17.761	5	1	0
Tabela	0	0	71	904	5	20
Gráfico	2	23	4	1	1.802	168
Sintética	82	13	1	0	98	2.806
Modelo treinado com as características (SIMSKE, 2005)						
	Foto	Logotipo	Documento	Tabela	Gráfico	Sintética
Foto	8.219	931	216	107	99	428
Logotipo	1.200	3459	192	30	10	109
Documento	1.922	2502	11.313	1.594	272	201
Tabela	63	20	271	604	5	20
Gráfico	10	297	81	42	1.108	462
Sintética	202	489	32	21	355	1.901

A maior confusão observada nas matrizes foi entre as classes Logotipo, Gráfico e Sintética. Esta maior taxa é decorrente da natureza semelhante dessas imagens, que foram geradas por ferramentas computacionais. As distribuições de confusão para os classificadores treinados com o mesmo conjunto de características apresentam um menor valor para (LINS *et al.*, 2009). A confusão registrada na classe foto foi decorrente de

imagens com ruído especular (80% dos casos) que diminuiram os níveis de informações disponíveis nas imagens. Os resultados apresentados pelas matrizes de confusão demonstram um melhor desempenho do conjunto de características proposto nesta tese.

A Figura 2.4 apresenta algumas imagens que o sistema apresentado nesta tese classificou incorretamente.



(a) Classe Logo "confudida"  
com a Classe Foto



(b) Classe Foto "confudida"  
com a Classe Sintética

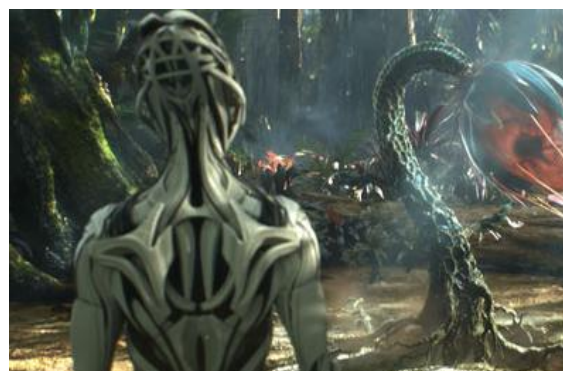
**Periodic Table of Elements**

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
1 H																	2 He			
3 Li	4 Be											10 Ne	11 Na	12 Mg	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr			
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe			
55 Cs	56 Ba	57-71	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn			
87 Fr	88 Ra	89-103	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Nh	114 Fl	115 Mc	116 Lv	117 Ts	118 Og			

For elements with no stable isotopes, the mass number of the isotope with the longest half-life is in parentheses.

Design and Interface Copyright © 1997 Michael Dayak (michael@dayak.com) <http://www.ptable.com>

(c) Classe Tabela "confudida"  
com a Classe Foto



(d) Classe Sintética "confudida"  
com a Classe Foto



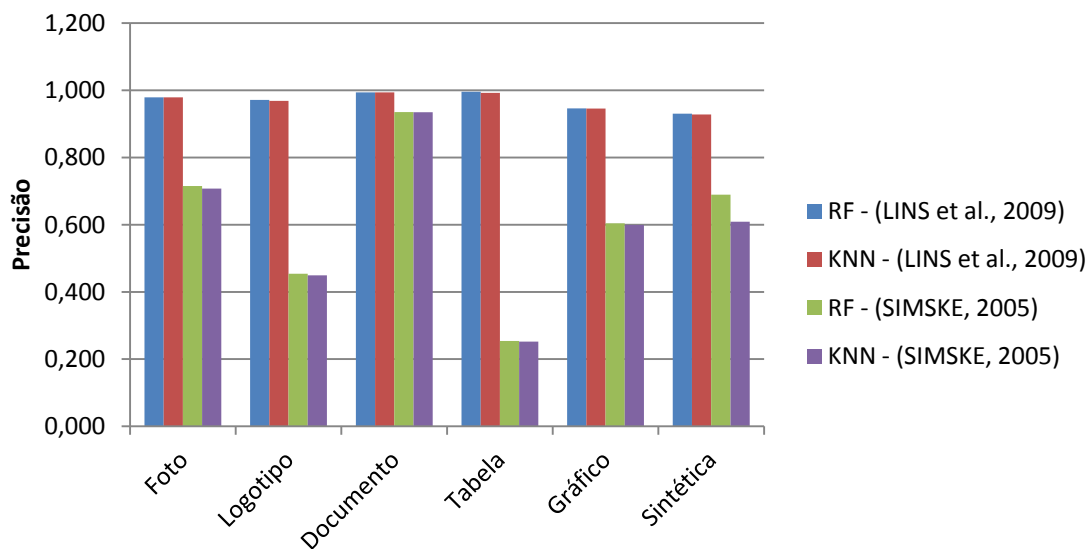
(e) Classe Gráfico "confudida"  
com a Classe Sintética



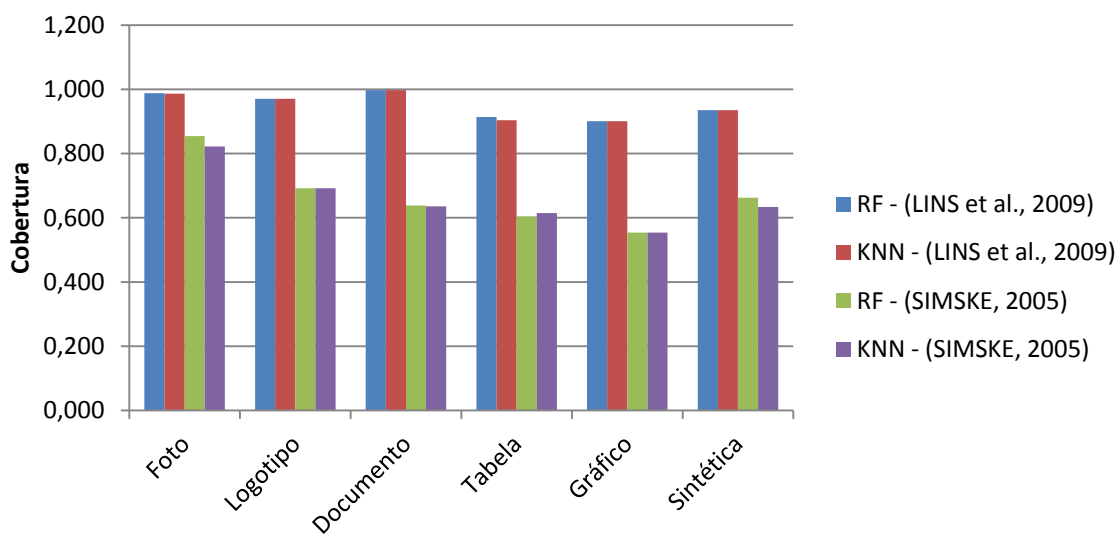
(f) Classe Documento "confudida"  
com a Classe Sintética

**Figura 2.4 Exemplos de erros de classificação.**

As comparações das medidas de Precisão, Cobertura e F-Measure dos classificadores Random Forests (RF) e KNN, treinados com as características propostas por (LINS *et al.*, 2009) e (SIMSKE, 2005), são apresentada a seguir.

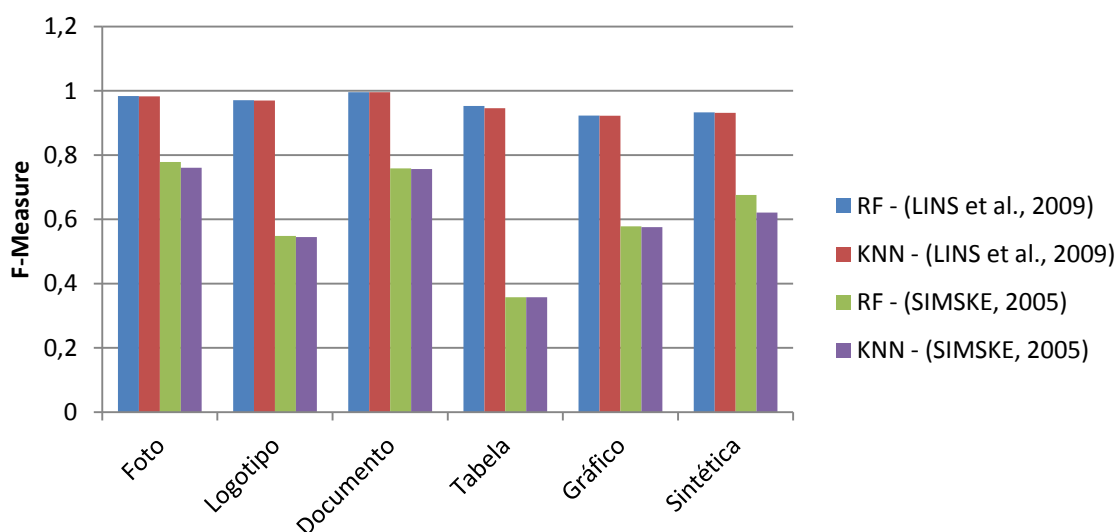


**Figura 2.5 Precisão em relação aos classificadores e características usadas no treinamento.**



**Figura 2.6 Cobertura em relação aos classificadores e características usadas no treinamento.**





**Figura 2.7 F-Measure em relação aos classificadores e características usadas no treinamento.**

O conjunto de características proposto nesta tese (LINS *et al.*, 2009) foi superior nas três medidas em relação ao conjunto de comparação (SIMSKE, 2005). A Tabela 2.4 apresenta a acurácia de cada classificador em relação ao conjunto de características utilizado.

**Tabela 2.4 Acurácia dos classificadores em relação ao conjunto de características.**

	(LINS <i>et al.</i> , 2009)	(SIMSKE, 2005)
<b>Random Forests (RF)</b>	0.96	0.62
<b>KNN</b>	0.96	0.60

A Tabela 2.5 apresenta os resultados do classificador Random Forest em relação aos conjuntos de características usados para o treinamento.

**Tabela 2.5 Acurácia dos classificadores em relação ao conjunto de características.**

Classe	Precisão		Cobertura		F-Measure	
	(SIMSKE, 2005)	(LINS <i>et al.</i> , 2009)	(SIMSKE, 2005)	(LINS <i>et al.</i> , 2009)	(SIMSKE, 2005)	(LINS <i>et al.</i> , 2009)
Foto	0.708	0.979	0.822	0.987	0.76	0.983
Logotipo	0.449	0.968	0.692	0.971	0.545	0.97
Documento	0.935	0.994	0.635	0.998	0.756	0.996
Tabela	0.252	0.992	0.614	0.904	0.357	0.946
Gráfico	0.599	0.945	0.554	0.901	0.576	0.923
Sintética	0.609	0.928	0.634	0.935	0.621	0.932

A análise da diferença estatística entre os classificadores treinados com as mesmas características revelou que não há diferença em torno das medidas de Precisão, Cobertura e F-Measure entre o Random Forests e o KNN para o conjunto (LINS *et al.*, 2009). Os resultados comprovam a eficiência do método proposto nesta tese em termos de classificação e tempo de extração de características 11,02 vezes mais rápido (ambos os algoritmos na linguagem Java).

## 2.2 Classificação de Dispositivos de Captura

Os *scanners* e câmeras digitais são os dispositivos atualmente mais utilizados para a digitalização de documentos. O reconhecimento automático do tipo *scanner* ou câmera, permite uma melhor filtragem das imagens de documento. As características dos sensores de digitalização e a prevalência de certos ruídos na imagem permitem um processamento inteligente dos documentos.

O problema de classificação de dispositivos de captura é visto nesta tese como um subproblema da seção anterior. Nessa área, dois aspectos são fundamentais: entender que tipo de dispositivo capturou a imagem (por exemplo, um *scanner*, uma câmera digital ou se elas são geradas por computador) e reconhecer o modelo e a marca (ver Figura 2.8).



**Figura 2.8 Classificação de imagens por diferentes tipos de dispositivos**

Os métodos disponíveis na literatura são complexos e de custo computacional elevado (CELIKUTAN, 2007; McKAY, 2008). O trabalho desenvolvido nesta tese busca identificar se o dispositivo usado para digitalização de documentos foi uma câmera ou um *scanner* (SILVA *et al.*, 2009a; LINS *et al.*, 2011a). Um conjunto de características simples e de baixo custo computacional é proposto e validado neste trabalho.

### 2.2.1 Trabalhos Relacionados

Técnicas de identificação de dispositivo estão focadas em avaliar a origem dos dados na forma de imagens ou vídeos. Três linhas de pesquisa recebem destaque: Análise de Metadados, Medidas de Qualidade de Imagem (MQI) e Extração de Características (estatísticas, geométricas, cor e propriedades físicas dos sensores de captura).

Imagens digitais podem ser armazenadas em uma variedade de formatos, tais como JPEG, GIF, PNG, TIFF e esses podem ser tão informativos quanto à imagem (LINS e MACHADO, 2004). Por exemplo, arquivos JPEG contêm um conjunto de características

bem definido, que inclui metadados, tabelas de quantização para compressão de imagens e dados compactados com perdas. Os metadados descrevem a origem da imagem, podendo incluir o nome do dispositivo, resolução e outras características (COHEN, 2007). O problema dessa abordagem é que nem sempre a informação é disponibilizada e informações falsas podem ser inseridas no arquivo de imagem.

Estudos propõem o uso de Medidas de Qualidade de Imagem (MQI) combinadas com características de cor (desvio de correlação de níveis de cinza, correlação entre banda, fator gama) e com estatísticas de coeficiente de wavelet, para classificação de dispositivos de captura (MEHDI *et al.*, 2004; CELIKTUTAN, 2007). A extração e cálculo das medidas de qualidade possuem alto custo computacional em relação às características extraídas da paleta da imagem propostas em (SILVA *et al.*, 2009a; LINS *et al.*, 2011a).

Características baseadas nas propriedades físicas dos sensores de captura buscam extrair informações dos padrões de ruídos gerados pelos sensores (GOU *et al.*, 2007a; KHANNA, 2007b; CHEN *et al.*, 2008). A mesma ideia, com a adição de coeficientes de interpolação de cor, é usada para identificar imagens produzidas por câmeras, *scanners* e sintetizadas por computador (McKAY, 2008). Um ponto negativo dessas abordagens é a sensibilidade a ruídos e a intensidade luminosa a que a imagem foi submetida (CELIKTUTAN *et al.*, 2008). A busca por padrões de assinaturas em algoritmos de interpolação proprietários é outro importante indicador do tipo e marca dos dispositivos (GALLAGHER, 2005; BABAK, 2008). Em geral, essas técnicas são afetadas pelas etapas de conversão para o JPEG que distorcem os padrões de interpolação.

### 2.2.2 Contribuições

A contribuição dessa tese é um sistema de classificação que permite a distinção das imagens em cinco classes:

- Escaneada.
- Fotografada mão livre (SF - Sem *Flash*).
- Fotografada c/ suporte (SF - Sem *Flash*).
- Fotografada mão livre (CF - Com *Flash*).
- Fotografada c/ suporte (CF - Com *Flash*).

As cinco classes representam as formas de digitalizações de documentos atualmente mais utilizadas (LINS *et al.*, 2011a). Essas classes apresentam peculiaridades que

permitem direcionar os algoritmos das etapas de classificação e filtragem de ruídos e obter melhores resultados (SILVA *et al.*, 2010a). Um novo arranjo de classificadores em cascata foi proposto para esse problema de classificação (Figura 2.9).



**Figura 2.9 Arquitetura em cascata de classificação.**

A publicação sobre classificação de imagens quanto ao dispositivo encontra-se disponível no Apêndice A desta tese, publicado no *Journal of Universal Computer Science* (SILVA *et al.*, 2009).

### 2.2.3 Experimentos e Resultados

A base de dados utilizada nos experimentos é composta por 17.804 documentos digitalizados por diferentes tipos de dispositivos em diferentes resoluções. As distribuições das bases de dados são apresentadas nas Tabelas 2.6 e 2.7.

**Tabela 2.6 Relação de Documentos Escaneados**

Dispositivo	Resolução (DPIs)	JPEG	PNG	TIFF	BMP	Quantidade
Ricoh Affício 1075	100	537	537	537	537	2.148
Ricoh Affício 1075	200	537	537	537	537	2.148
Ricoh Affício 1075	300	537	537	537	537	2.148
HP 5300c	300	0	0	300	0	300
HP 5300c	200	0	0	300	0	300
EPSON L355	200	100	100	100	100	400
EPSON L355	200	100	100	100	100	400
Total						7.844

Tabela 2.7 Relação de Documentos Fotografados no formato JPG *truecolor*.

Dispositivo	Resolução (MPixel)	Suporte	Flash	Quantidade
<b>LG Shine ME970</b>	2	Não	Não	60
<b>Samsung Galaxy Mini</b>	3	Não	Não	200
<b>Samsung Galaxy Mini</b>	3	Sim	Não	200
<b>Motorola MotoG</b>	5	Não	Não	100
<b>Motorola MotoG</b>	5	Não	Sim	100
<b>Motorola MotoG</b>	5	Sim	Não	100
<b>Motorola MotoG</b>	5	Sim	Sim	100
<b>Sony DSC-S40</b>	4	Não	Não	200
<b>Sony DSC-S40</b>	4	Não	Sim	200
<b>Sony DSC-W55</b>	5	Não	Não	2.000
<b>Sony DSC-W55</b>	5	Não	Sim	2.000
<b>Sony DSC-W55</b>	7	Sim	Não	2.000
<b>Sony DSC-W55</b>	7	Sim	Sim	2.000
<b>Nikon S2700</b>	16	Não	Não	150
<b>Nikon S2700</b>	16	Não	Sim	150
<b>Canon 60D</b>	18	Não	Não	200
<b>Canon 60D</b>	18	Não	Sim	200
<b>Total</b>				9.960

O conjunto de características proposto nesta tese (SILVA *et al.*, 2009) foi comparado com o trabalho apresentado em (McKAY *et al.*, 2008). O classificador adotado no trabalho de Macky foi uma variação do Support Vector Machine (SVM) (HSU e LIN, 2002). O mesmo classificador foi utilizado para conduzir os experimentos nesta tese.

A comparação é apresentada na forma de matriz de confusão e nas medidas de Precisão, Cobertura e *F-Measure* (Salton e Buckley 1988). A técnica de Validação Cruzada (KOHAVI, 1995) com *k-fold*, onde  $k=10$ , foi usada para validação. Para os experimentos, dez SVM com diferentes configurações foram gerados para o conjunto de características proposto por (McKAY *et al.*, 2008) e outros dez para o conjunto proposto

por (SILVA *et al.*, 2009). O *Paired Corrected T-Tester* com significância  $p=0.05$  foi usado para verificar a diferença estatística entre os SVMs com o mesmo conjunto de características, em torno da medida F-Measure. O conjunto de características proposto neste trabalho não apresentou diferença estatística entre os dez classificadores SVM:

- Palette (true-color/grayscale).
- Gamut (true-color/grayscale).
- Número de black pixels (Binarização de Otsu).
- $(\#Black\_pixels/Total\_\#\_pixels)*100\%$ .
- $(Gamut/Palette)*100\%$  (true-color/grayscale).

O conjunto de características proposto por (McKAY *et al.*, 2008) apresentou cinco casos com nível de significância  $p>0.05$ . Esses dados demonstram uma maior estabilidade das características propostas nesta tese.

O primeiro experimento desta seção apresenta o resultado da distinção entre documentos escaneados e fotografados; o melhor classificador treinado com as características apresentadas em (McKAY *et al.*, 2008) foi usado na comparação (Tabela 2.8).

**Tabela 2.8 Resultados do Classificador SVM para tarefa de classificação (scanner x câmera)**

Classificador SVM treinado com as características (SILVA <i>et al.</i> , 2009)					
	Matriz de confusão		Medidas de Avaliação do Classificador		
	Escaneado	Fotografado	Precisão	Cobertura	F-Measure
Escaneada	7.170	674	0,99	0,91	0,95
Câmera	28	9.932	0,93	0,99	0,96
Classificador SVM treinado com as características (McKAY <i>et al.</i> , 2008)					
	Escaneado	Fotografado	Precisão	Cobertura	F-Measure
Escaneada	7.066	778	0,88	0,90	0,89
Câmera	898	9.062	0,92	0,90	0,91

A análise dos dados mostrou que o conjunto de características de (McKAY *et al.*, 2008) apresentou uma confusão 36 vezes maior entre imagens fotografadas e escaneadas. Esta maior confusão é refletida nas medidas de Precisão para documentos escaneados (11% inferior) e na Cobertura dos documentos fotografados (9% inferior). O conjunto de características proposto em (SILVA *et al.*, 2009) apresentou resultados superiores para

Precisão e Cobertura, para o problema de classificação proposto por (McKAY *et al.*, 2008).

O segundo experimento foi aplicado para o problema de classificação em cinco classes: Escaneada; Fotografada mão livre (SF - Sem Flash); Fotografada c/ suporte (SF - Sem Flash); Fotografada mão livre (CF - Com Flash); Fotografada c/ suporte (CF - Com Flash). Esse problema é um refinamento do primeiro (*scanner* x câmera) e busca identificar os parâmetros de captura das câmeras digitais. A partir da distribuição obtida nesse, será possível entender melhor os erros de classificação da proposta de (McKAY *et al.*, 2008). A metodologia para avaliação dos resultados foi a mesma aplicada ao primeiro experimento. Os resultados da matriz de confusão são apresentados na Tabela 2.9.

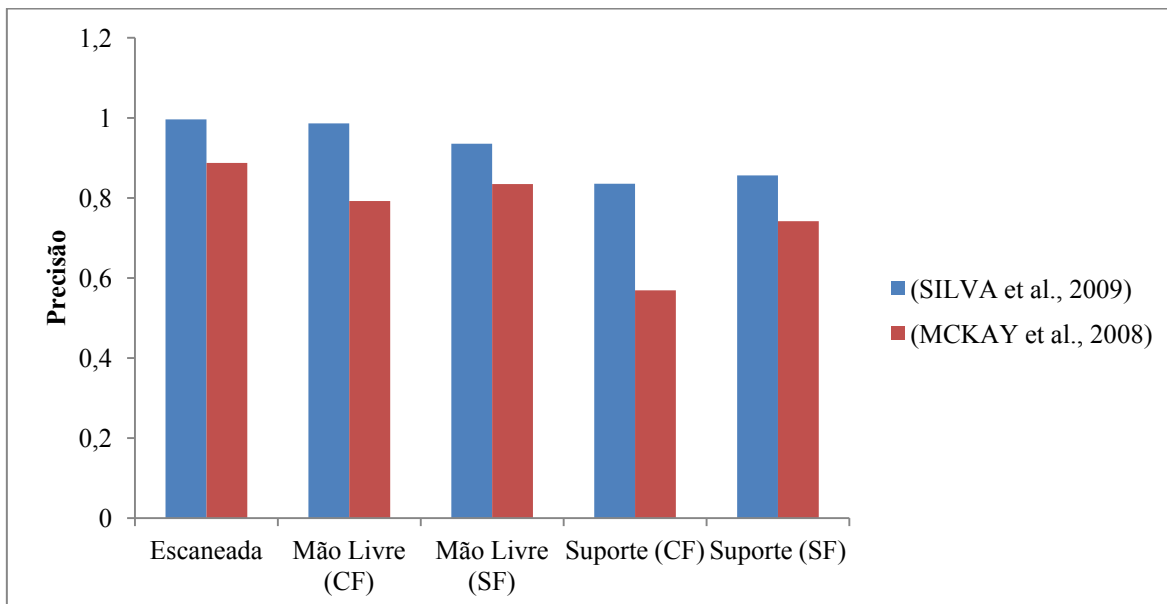
**Tabela 2.9 Matriz de confusão *scanner* x características de captura da câmera.**

Classificador SVM treinado com as características (SILVA <i>et al.</i> , 2009)					
	Escaneado	Livre (CF)	Livre (SF)	Suporte (CF)	Suporte (SF)
Escaneada	7.828	0	1	10	5
Livre (CF)	2	2.639	4	3	2
Livre (SF)	3	24	2.881	0	2
Suporte (CF)	13	1	0	2.068	18
Suporte (SF)	10	1	0	14	2.275
Classificador SVM treinado com as características (McKAY <i>et al.</i> , 2008)					
	Escaneado	Livre (CF)	Livre (SF)	Suporte (CF)	Suporte (SF)
Escaneada	6.866	0	1	631	346
Livre (CF)	42	2.174	382	37	15
Livre (SF)	26	513	2.191	102	78
Suporte (CF)	628	27	34	1.299	112
Suporte (SF)	202	30	26	458	1.584

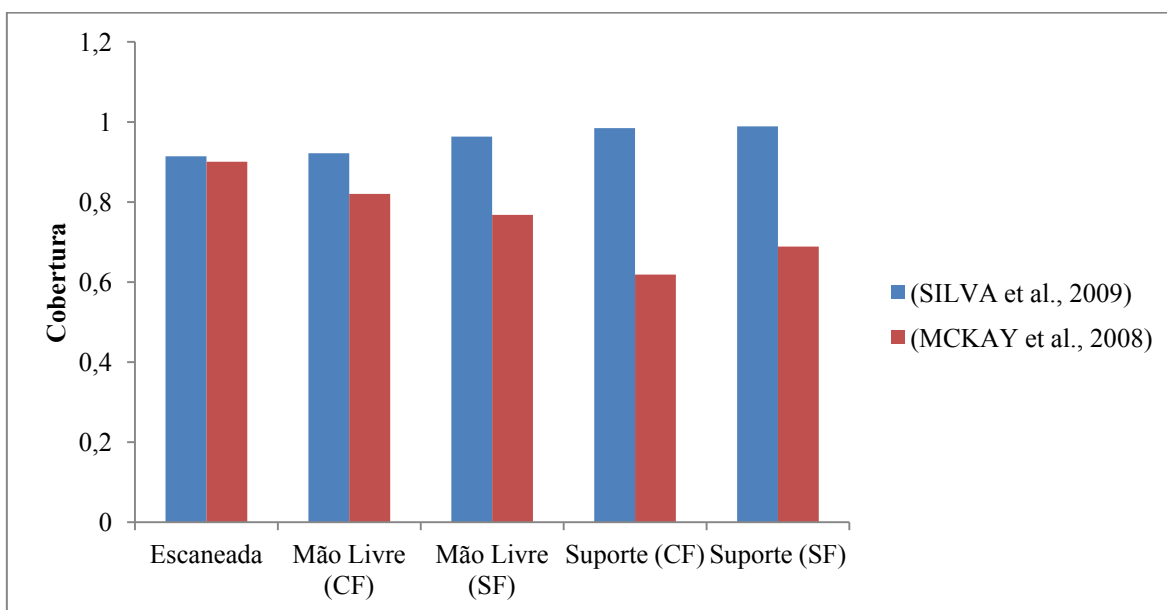
Os dados da matriz de confusão comprovaram a eficiência do método proposto nesta tese para o problema de classificação de dispositivo. A análise e comparação dos dados das Tabelas 2.8 e 2.9, em relação a *scanners* x câmeras, apontou uma distribuição da confusão semelhante. Os dados evidenciam para o método de (McKAY *et al.*, 2008) problemas de classificação em relação a documentos fotografados com o uso de suporte. Isso pode ser explicado pelas propriedades das características usadas, como propriedades de interpolação, que são afetadas pela compressão do formato JPEG.



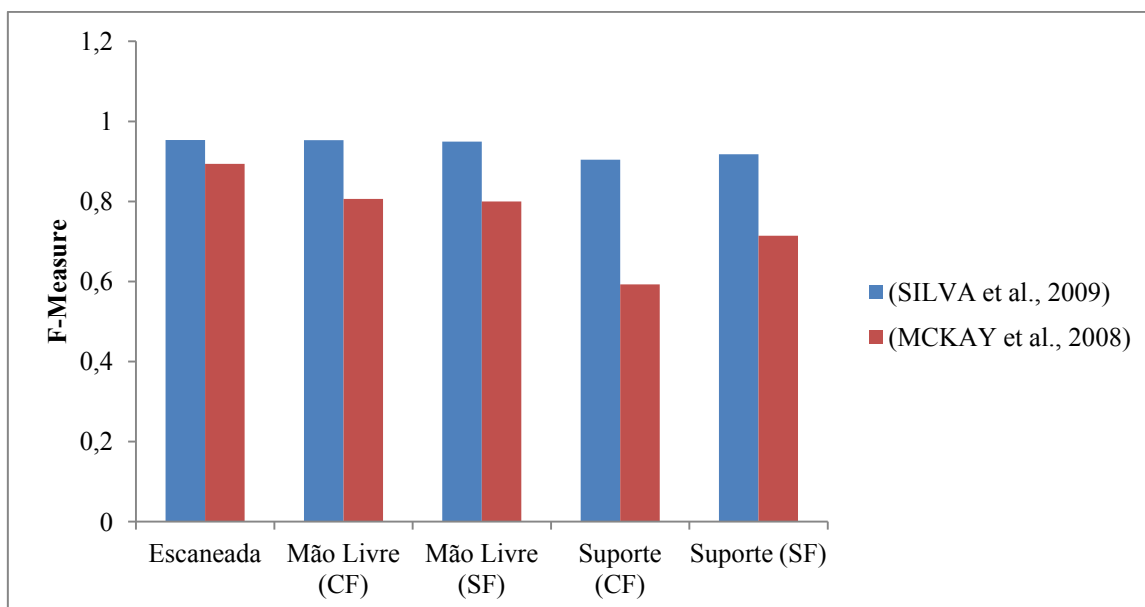
Os resultados das medidas Precisão, Cobertura e F-Measure do classificador SVM para os conjuntos de características de McKAY (McKAY *et al.*, 2008) e SILVA (SILVA *et al.*, 2009) são apresentados na forma de gráficos.



**Figura 2.10 Medida de Precisão do classificador SVM em relação aos conjuntos de características.**



**Figura 2.11 Medida de Cobertura do classificador SVM em relação aos conjuntos de características.**



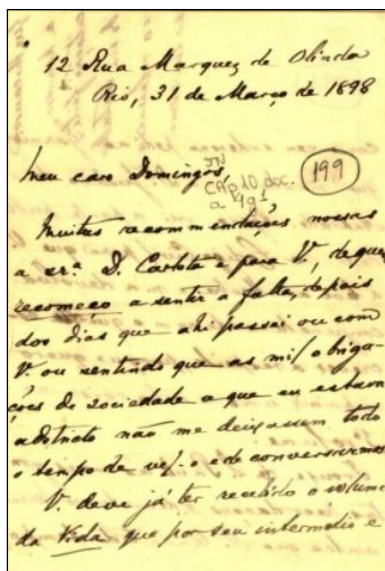
**Figura 2.12 Medida de *F-Measure* do classificador SVM em relação aos conjuntos de características.**

A análise dos gráficos aponta um melhor desempenho do método proposto, 15% superior para Precisão, 19% para Cobertura e 17% para a F-Measure. Os resultados experimentais atestam o melhor desempenho do método de classificação proposto nesta tese. Outra vantagem da proposta desta tese é o tempo de extração de características, 6 vezes menor.

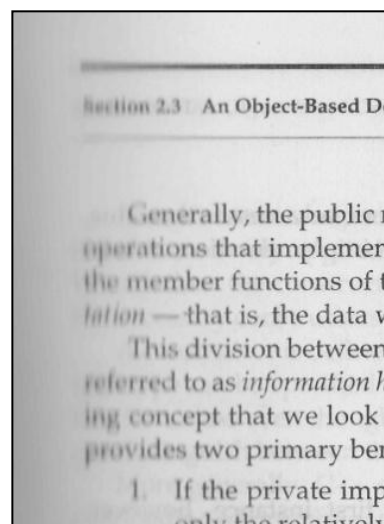
## 2.3 Classificação e Caracterização de Ruídos

A maioria dos estudos em processamento de imagens de documentos foca na remoção do ruído, e não na sua detecção e caracterização, o que leva a uma filtragem "cega". Determinar o tipo de ruído permite uma filtragem inteligente, o que pode melhorar a eficiência dos algoritmos e uma melhor qualidade das imagens. Reconhecer o ruído, classificar e compreender sua natureza e intensidade é fundamental para a sua remoção adequada. O problema de classificação e caracterização de ruídos é de grande importância para tratar bases que contenham documentos heterogêneos. Diferentes tipos de ruídos estão presentes nessas bases, com intensidade e distribuição variáveis (Figura 2.13). Nesta tese, as características indesejadas encontradas frequentemente em documentos digitais, como borda, orientação e inclinação, são assumidas como ruído.

O trabalho desenvolvido nesta tese é focado na identificação e caracterização de oito tipos de ruído: Inclinação (*skew*), Orientação (*orientation*), existência de Borda (*border*), Interferência Frente e Verso (*back to front interference*, *bleeding* ou *show-through*), Embaçamento (*blur*), Iluminação (*uneven illumination*), Furos e Rasgos (*punching* e *torn off regions*) e Sal e Pimenta (*salt and pepper*).

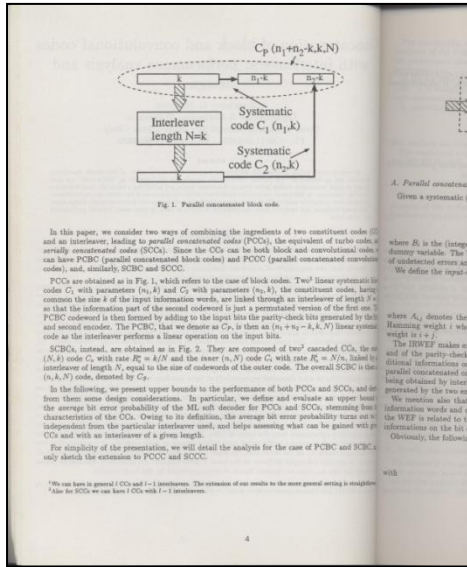


(a) Interferência frente e verso.



(b) Embaçamento.

(Continua)



(c) Borda.

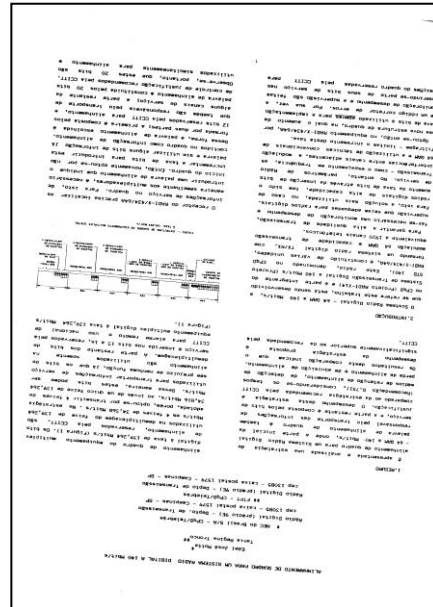


(d) Especular.

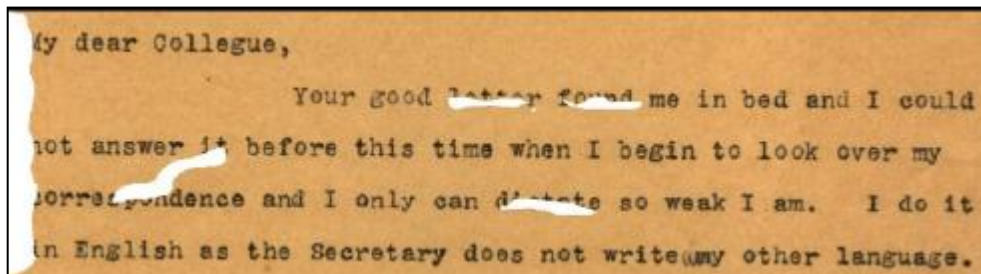
pois melhoram as características de transmissão uma vez que separam os enlaces de subida e descida. Além disso, permitem processamento e eventual armazenamento a bordo.

Para isso, técnicas de estimação e detecção espectral foram estudadas [2 a 6] para se avaliar o método de detecção mais eficiente à aplicação, considerando-se os limites do sistema. Foram considerados não apenas o desempenho das técnicas de estimação, mas também a complexidade computacional e a capacidade de se processar em tempo real sinais com um número de PCDs desconhecido a priori, obtendo uma estimativa linear direta do seu espectro de potências e decidiu-se pela utilização do estimador direto constituído pelo periodograma com janela temporal de dados prolate de ordem zero. Com os resultados da análise foi desenvolvido e validado um algoritmo de detecção das PCDs utilizando a técnica de processamento digital de sinais na busca em frequência dos sinais de PCDs. Como na prática não se dispõe de dados reais da aplicação, foram utilizados para a validação do algoritmo os sinais reais recebidos no *front end* dos demoduladores e processadores de dados (PROCODs)1e2 – processadores codificadores de dados de PCDs não embarcados da estação receptora terrena de Cuiabá, que recebem os sinais na mesma faixa de frequências do detector proposto. Desta forma, como contribuição adicional, com pequenas alterações o algoritmo de detecção desenvolvido, foi também validada para o detector não embarcado PROCOD3, em desenvolvimento, que utilizará processamento digital de sinais.

(e) Sal e Pimenta



(f) Orientação e Inclinação.



(g) Furos e Rasgos.

Figura 2.13 Classes de ruídos estudados

### 2.3.1 Trabalhos Relacionados

A classificação e caracterização de ruído é uma área de pesquisa relativamente nova (LINS, 2009). Falhas podem manifestar-se sobre o documento físico ou durante a digitalização, transmissão, armazenamento e conversão de um formato para outro, e são popularmente conhecidas como ruído (LINS, 2009). Um conjunto de técnicas de análise de documentos, como segmentação e reconhecimento de caracteres, normalmente trabalha melhor em documentos "limpos". Esses algoritmos por vezes dependem de componentes conectados como unidades básicas, que são sensíveis a vários tipos de ruído.

A detecção e remoção de ruído é geralmente baseada em suas propriedades, como sua forma, posição, frequência, níveis de cinza, densidade ou periodicidade de ocorrência no documento (FAN *et al.*, 2001; ZHENG *et al.*, 2001; ZHENG *et al.*, 2003; DONG *et al.*, 2007). O ruído de borda mostra regularidade em suas posições (FAN *et al.*, 2001), enquanto linhas de texto mostram periodicidade em suas posições e consistência na direção (ZHENG *et al.*, 2001; ZHENG *et al.*, 2003). Por outro lado, interferências como manchas de tinta são mais densas do que o texto, enquanto que o Sal e Pimenta é um ruído impulsivo e apresenta maior espaçamento entre os *pixels* de ruído em relação aos de conteúdo (CHAN *et al.*, 2005). Se o ruído apresenta um comportamento consistente em termos dessas propriedades, é mais fácil detectá-lo e separá-lo do conteúdo (ALI, 1996; CHINNASARN *et al.*, 1998).

A extração de características relativas a momentos geométricos e invariantes (GONZALEZ e WOODS, 2008), descritores de textura e evolução de altas frequências é utilizada para detecção do ruído de embaçamento e borda (RUGNA e KONIK, 2003). A extração dessas características implica em um alto custo computacional e apresenta sensibilidade à presença de outros ruídos na imagem.

Outro conjunto de características é obtido usando a resposta de filtros (mediana, Wiener e Gaussiano) para detectar ruídos de Sal e Pimenta e de alta frequência (GOU *et al.*, 2007b). Esses métodos dependem do correto ajuste dos parâmetros dos filtros, por exemplo, o tamanho da janela do filtro de mediana. Técnicas para extrair estimativas de parâmetros foram propostas, mas elas tendem a ser heurísticas (GRAY e COK, 1997; SNYDER *et al.*, 1999).

Na verdade, poucos trabalhos tratam diretamente o problema de classificação e caracterização de ruído. Os avanços na área de classificação de ruídos propostos nessa tese serão apresentados na próxima seção.

### 2.3.2 Contribuições

As contribuições desta seção estão focadas na detecção e caracterização da intensidade de ruídos. Um conjunto de características é proposto para solução do problema de classificação de ruído:

- Palette (true-color/grayscale).
- Gamut (true-color/grayscale).
- Número de black pixels (Binarização de Otsu).
- $(\#Black\_pixels/Total\_\#\_pixels)*100\%$ .
- $(Gamut/Palette)*100\%$  (true-color/grayscale).
- Máxima Saturação.
- Variação Local do Espectro de Potência.
- Entropia.

O modelo de classificação proposto é formado por oito classificadores em paralelo (Figura 2.14).

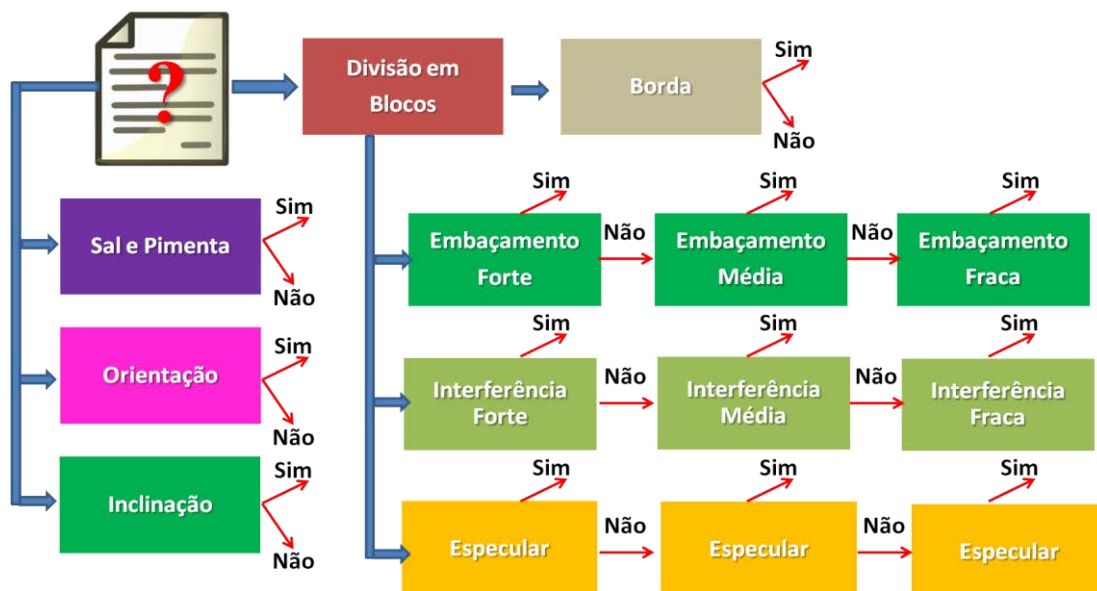


Figura 2.14 Arquitetura de classificação de ruído em paralelo

Essa arquitetura foi adotada de forma a tratar imagens que possuam uma ou mais classes de ruído. Os artigos publicados (LINS *et al.*, 2010d; SILVA e LINS, 2011) referentes ao problema de classificação de ruído, encontram-se disponíveis no Apêndice A desta tese.

### 2.3.3 Experimentos e Resultados

Os experimentos realizados para validação desta contribuição contam com a maior base de dados presente na literatura. A base de dados utilizada nos experimentos é composta por 29.583 imagens reais e 25.800 sintéticas, totalizando 55.383 documentos. As imagens sintéticas foram geradas a partir de 500 documentos no formato JPEG *truecolor* e não apresentaram ruído de Borda, Furos e Rasgos, Inclinação, Orientação, Sal e Pimenta, Interferência Frente e Verso, Especular e Embaçamento.

As imagens originais tiveram o ruído mapeado manualmente, enquanto nas sintéticas as informações de localidade foram previamente definidas. A geração do ruído artificial respeitou as características de localização e/ou aleatoriedade de cada classe (Figura 2.15). O mapeamento consistiu em dividir as imagens em 100 blocos e identifica-lós em uma ou cinco classes.

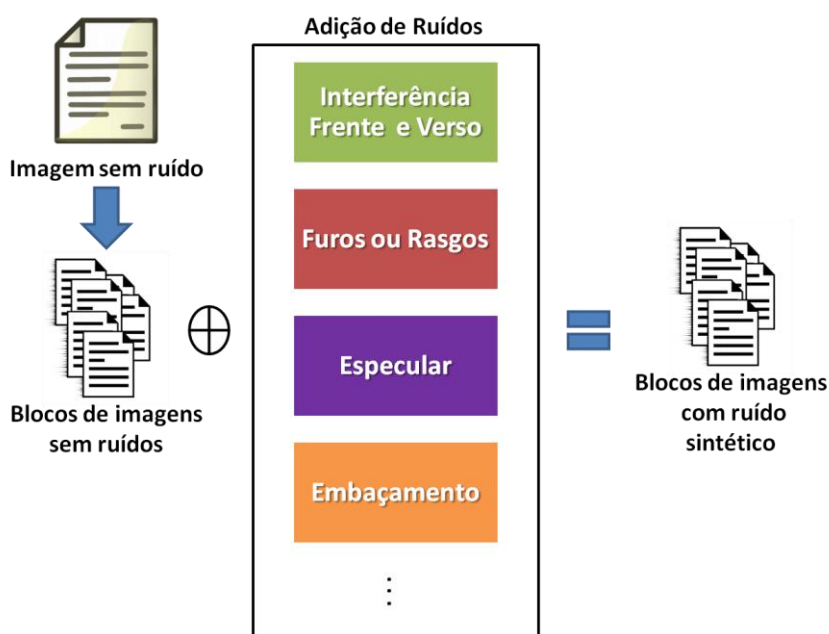


Figura 2.15 Geração de imagem com ruído sintético

Três classes de ruído não foram subdivididas em blocos: Orientação, Inclinação e Sal e Pimenta, uma vez que foi assumida uma distribuição global do ruído. A distribuição dos blocos com e sem ruído é apresentada na Tabela 2.10.

**Tabela 2.10 Distribuição do rótulo dos blocos por classe.**

<b>Nome</b>	<b>Bloco c/ Ruído</b>	<b>Bloco s/Ruído</b>
<b>Orientação</b>	---	---
<b>Inclinação</b>	---	---
<b>Especular</b>	1.500	28.500
<b>Borda</b>	89.968	1.034.632
<b>Interferência</b>	16.108	386.592
<b>Sal e Pimenta</b>	---	---
<b>Embaçado</b>	28.800	331.200
<b>Furos</b>	5.850	111.150



<b>Tabela 2.11 Apresentação do quadro geral dos documentos utilizados nos experimentos de classificação de ruídos.</b>									
	<b>Orientação</b>	<b>Inclinação</b>	<b>Especular</b>	<b>Borda</b>	<b>Interferência</b>	<b>Sal e Pimenta</b>	<b>Embaçado</b>	<b>Furos</b>	
<b>Sintética</b>	6.200	6.200	---	5.200	2.000	2.000	3.200	1.000	
<b>TIFF(BW)</b>	3.000	3.000	---	500	---	640	---	---	
<b>TIFF(cinza)</b>	3.000	3.000	---	2.600	---	---	---	120	
<b>PNG (color)</b>	3.000	3.000	---	2.000	---	---	---	---	
<b>JPEG (color)</b>	1.000	1.000	300	946	2.027	---	400	50	
<b>Quantidade</b>	<b>16.200</b>	<b>16.200</b>	<b>300</b>	<b>11.246</b>	<b>4.027</b>	<b>2.640</b>	<b>3.600</b>	<b>1.170</b>	

Os resultados são apresentados na forma de matriz de confusão e nas medidas de Precisão, Cobertura e F-Measure (SALTON e BUCKLEY 1988). A técnica de Validação Cruzada (KOHAVI, 1995) com *k-folder*, para  $k=10$ , foi usada. Para os experimentos, o classificador Random Forests (BREIMAN, 2001) foi adotado. Os resultados são apresentados a seguir.

**Tabela 2.12 Resultado do classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para o ruído de Orientação.**

Orientação	Matriz de Confusão		Medidas de Avaliação do Classificador		
	Ruído	Sem Ruído	Precisão	Cobertura	F-Measure
<b>Ruído</b>	15.832	368	0,93	0,97	0,95
<b>Sem Ruído</b>	1.097	15.103	0,97	0,93	0,95

**Tabela 2.13 Resultado do classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para o ruído de Inclinação.**

Inclinação	Matriz de Confusão		Medidas de Avaliação do Classificador		
	Ruído	Sem Ruído	Precisão	Cobertura	F-Measure
<b>Ruído</b>	15.391	809	0,94	0,95	0,94
<b>Sem Ruído</b>	895	15.305	0,94	0,94	0,94

**Tabela 2.14 Resultado do classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para o ruído Especular.**

Especular	Matriz de Confusão		Medidas de Avaliação do Classificador		
	Ruído	Sem Ruído	Precisão	Cobertura	F-Measure
<b>Ruído</b>	1.408	92	0,80	0,93	0,86
<b>Sem Ruído</b>	345	28.155	0,99	0,98	0,99

**Tabela 2.15 Resultado do classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para o ruído de Borda.**

Borda	Matriz de Confusão		Medidas de Avaliação do Classificador		
	Ruído	Sem Ruído	Precisão	Cobertura	F-Measure
<b>Ruído</b>	85.422	4.546	0,90	0,94	0,92
<b>Sem Ruído</b>	9.191	1.025.441	0,99	0,99	0,99

**Tabela 2.16 Resultado do classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para o ruído de Sal e Pimenta.**

Sal e Pimenta	Matriz de Confusão		Medidas de Avaliação do Classificador		
	Ruído	Sem Ruído	Precisão	Cobertura	F-Measure
<b>Ruído</b>	2.428	212	0,92	0,91	0,92
<b>Sem Ruído</b>	184	2.316	0,91	0,92	0,92

**Tabela 2.17 Resultado do classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para o ruído de Interferência.**

Interferência	Matriz de Confusão		Medidas de Avaliação do Classificador		
	Ruído	Sem Ruído	Precisão	Cobertura	F-Measure
<b>Ruído</b>	15.032	1.076	0,77	0,93	0,84
<b>Sem Ruído</b>	4.415	382.177	0,99	0,98	0,99

**Tabela 2.18 Resultado classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para o ruído de Embaçamento.**

Embaçado	Matriz de Confusão		Medidas de Avaliação do Classificador		
	Ruído	Sem Ruído	Precisão	Cobertura	F-Measure
<b>Ruído</b>	26.954	1.846	0,67	0,93	0,78
<b>Sem Ruído</b>	12.786	318.414	0,99	0,96	0,97

**Tabela 2.19 Resultado do classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para o ruído de Furo.**

Furo	Matriz de Confusão		Medidas de Avaliação do Classificador		
	Ruído	Sem Ruído	Precisão	Cobertura	F-Measure
<b>Ruído</b>	5.772	78	0,94	0,98	0,96
<b>Sem Ruído</b>	324	110.826	0,99	0,99	0,99

A análise dos resultados acima mostra valores médios de precisão e de cobertura de 91%. Alguns classificadores apresentaram valores de precisão abaixo de 80% (Tabelas 2.14, 2.17 e 2.18). Na classificação do ruído especular, 1,45% dos blocos apresentaram erro. Desse total, 1,15% foram referentes a blocos "limpos" classificados como ruidosos. Para o ruído de interferência frente e verso, o erro foi de 1,4%, com 1,09% referente a blocos sem ruídos. O ruído de Embaçamento apresentou o maior erro (4,06%), com 3,55% de erros de classificação de blocos "limpos". Isso implica no aumento do custo

computacional e na possível degradação de blocos sem ruídos, causados por uma filtragem desnecessária. Os blocos de ruídos que foram considerados "limpos" representam menos de 0,5% dos casos, o que significa que parte do ruído não será filtrado.

Para melhor entender a distribuição desse erro, os blocos de ruídos das três classes que apresentaram baixo rendimento foram subdivididos em três níveis (Forte, Médio e Fraco). A distribuição dessas subclasses é apresentada na Tabela 2.20.

**Tabela 2.20 Distribuição do rotulo dos blocos por subclasse.**

Nome	Forte	Médio	Fraco
<b>Especular</b>	381	625	494
<b>Interferência</b>	4.862	5.140	6.106
<b>Embaçado</b>	9.521	9.471	9.808

Os problemas de precisão relatados anteriormente foram mapeados em suas respectivas subclasses e apresentaram uma distribuição superior a 96% na forma de ruído fraco. Já os blocos com ruído que não foram classificados como "limpos" são em sua maioria de intensidade fraca. As tabelas a seguir apresentam o resultado da classificação das subclasses.

**Tabela 2.21 Resultado do Classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para a intensidade do ruído Especular.**

Especular	Matriz de Confusão			Medidas de Avaliação do Classificador		
	Forte	Médio	Fraco	Precisão	Cobertura	F-Measure
<b>Forte</b>	369	10	2	0,99	0,96	0,97
<b>Média</b>	4	612	9	0,97	0,98	0,97
<b>Fraca</b>	0	6	488	0,98	0,98	0,98

**Tabela 2.22 Resultado do Classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para a intensidade do ruído de Interferência Frente e Verso.**

Interferência	Matriz de Confusão			Medidas de Avaliação do Classificador		
	Forte	Médio	Fraco	Precisão	Cobertura	F-Measure
<b>Forte</b>	4.647	186	29	0,97	0,95	0,96
<b>Média</b>	83	5.028	29	0,95	0,97	0,96
<b>Fraca</b>	13	52	6.041	0,99	0,98	0,98

**Tabela 2.23 Resultado do Classificador Random Forest sob as características de (LINS *et al.*, 2010d; SILVA e LINS, 2011) para a intensidade do ruído de Embaçamento.**

<b>Embaçamento</b>	<b>Matriz de Confusão</b>			<b>Medidas de Avaliação do Classificador</b>		
	<b>Forte</b>	<b>Médio</b>	<b>Fraço</b>	<b>Precisão</b>	<b>Cobertura</b>	<b>F-Measure</b>
<b>Forte</b>	9.383	97	41	0,97	0,98	0,98
<b>Média</b>	89	9.304	78	0,97	0,98	0,97
<b>Fraca</b>	128	174	9.506	0,98	0,97	0,97

Os resultados apresentados mostram a viabilidade da classificação de ruído em imagens de documentos. A caracterização do ruído em níveis distintos é fundamental para guiar os algoritmos de filtragem. Isso diminui a degradação dos blocos sem ruídos incorretamente classificados, já que algoritmos e parâmetros menos agressivos serão usados para filtrar blocos com baixa intensidade de ruído, que são em sua maioria mapeados nessa intensidade.

## Capítulo 3

# Filtragem de Ruídos em Imagens de Documentos

Ruídos em imagens podem ser vistos como dependentes ou independentes do conteúdo do documento. Manchas de tinta e Sal e Pimenta, são, em geral, independentes da localização, tamanho ou outras propriedades do conteúdo de texto (CHINNASARN *et al.*, 1998). Por outro lado, quando o ruído está incluído no domínio de frequência espacial da imagem e não pode ser suprimido sem o conhecimento prévio do conteúdo, é referido como ruído dependente (WINDYGA, 2001). Exemplos de ruídos dependentes são o embaçamento e a interferência frente e verso (WANG e TAN, 2001b), que são muitas vezes multiplicativos.

A filtragem de ruído em imagens de documentos pode ser realizada de duas maneiras fundamentais. A primeira consiste em extrair o conteúdo de "interesse" da imagem, enquanto a outra consiste em detectar e remover o ruído (STROUTHOPOULOS *et al.*, 2002; LI e FAN, 2009). A primeira abordagem é frequentemente utilizada em casos em que há um número limitado de tipos de conteúdo.

Em particular, a extração de texto a partir de imagens é a abordagem que apresentou mais trabalhos nos últimos anos (WU *et al.*, 1999; WANG e TAN, 2001a; STROUTHOPOULOS *et al.*, 2002; BUKHARI *et al.*, 2012). No entanto, para conteúdo variado, como logotipos, figuras, selos, diagramas, equações e desenhos, processos de extração individuais podem ser necessárias (STROUTHOPOULOS *et al.*, 2002; PHAM, 2003), o que torna essa abordagem menos prática. Estes processos de extração individuais são normalmente dependentes de análise do layout, que por sua vez são dependentes de um documento "limpo" para bons resultados.

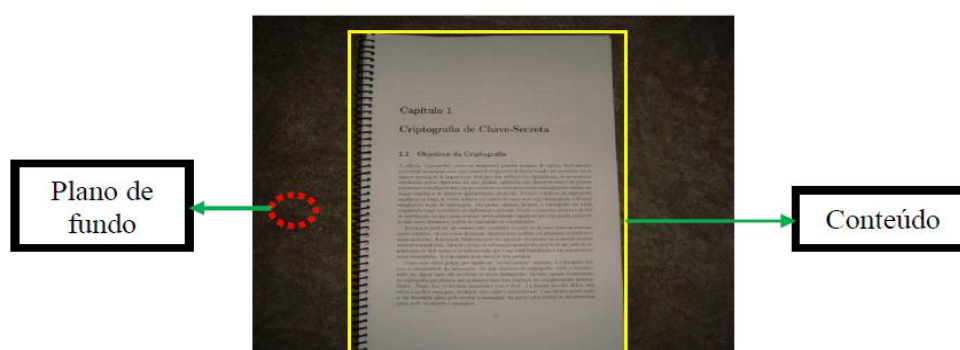
Essa seção irá apresentar quatro novos algoritmos para filtragem de ruído em imagens e um método inteligente de automação da etapa de filtragem de documentos.

### 3.1 Remoção de Borda

As bordas de documentos são definidas como os contornos que delimitam o conteúdo e a região externa dos documentos. O ruído de borda são áreas que circundam a imagem do documento e que não representam informação relevante. Esse ruído geralmente é adquirido

no processo de digitalização. No caso de *scanners*, as bordas correspondem a páginas adjacentes ou à área da bandeja não ocupada pelo documento. Nas câmeras digitais, o fenômeno é causado por um mau enquadramento. Esse último caso apresenta um desafio maior, pois as bordas geradas apresentam uma diversidade maior (textura e tamanho) o que inviabiliza grande parte dos algoritmos (LINS *et al.*, 2007b).

De modo geral, a borda ruidosa acarreta em vários problemas: (1) degrada a imagem visualmente; (2) aumenta o espaço necessário para armazenamento e transmissão via rede; (3) aumenta o gasto de tinta para impressão do documento e; (4) degrada a eficiência dos algoritmos de processamento de imagem (Figura 3.1).



**Figura 3.1 Documento fotografado**

### 3.1.1 Trabalhos Relacionados

Os métodos para detecção e remoção de ruído de borda podem ser divididos em duas categorias. A primeira identifica e remove componentes ruidosos; a segunda foca na identificação da área de conteúdo ou página (PEERAWIT e KAWTRAKUL, 2004). A maioria das técnicas presentes na literatura trata o problema de bordas em documentos adquiridos por *scanners* (LINS *et al.*, 2007b).

O método de projeção de perfil vertical foi um dos primeiros a ser empregado para recuperar imagens de documentos que contêm ruído de borda (ZHANG e TAN, 2001). Apesar de ser uma técnica simples, apresenta restrições quanto à complexidade do plano de fundo. Geralmente, isso inviabiliza o seu uso em documentos digitalizados por câmeras (LINS *et al.*, 2007b).

Uma classe de algoritmos para remoção de borda usa a segmentação de linhas de texto para fazer a separação entre o conteúdo e a borda (BUKHARI *et al.*, 2012). Essas técnicas não tratam o problema da aderência da borda ao texto e dependem da binarização e da segmentação do conteúdo textual para remover o ruído.

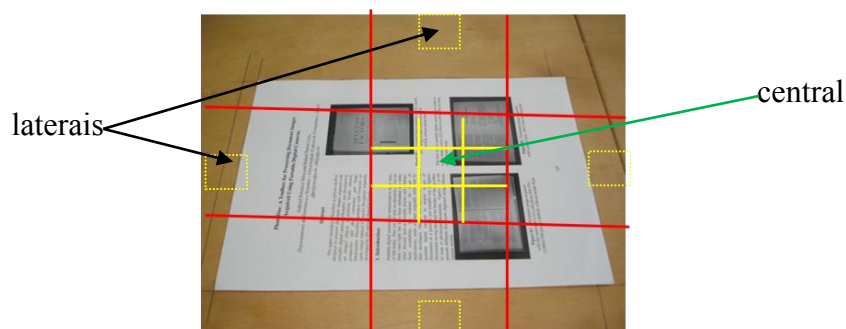
A ideia de um *flood-fill* recursivo é utilizada para lidar com bordas que aderem ao texto do documento (ÁVILA e LINS, 2004), onde o algoritmo é referência para remoção de bordas em documentos escaneados (AGRAWAL e DOERMANN, 2011; BUKHARI *et al.*, 2012; SILVA *et al.*, 2013). No entanto, problemas de bordas texturizadas e duplas (página adjacente) não são tratados. Outras técnicas usam operadores de Sobel (GONZALEZ e WOODS, 2008) e Canny (CANNY, 1986) para buscar os contornos do documento (CUMANI, 1991; PEERAWIT e KAWTRAKUL, 2004). Algumas adaptações foram propostas para tratar imagens de câmeras digitais (STAMATOPOULOS *et al.*, 2007; BUKHARI *et al.*, 2009). Esses métodos assumem resolução de captura adequada, bordas não complexas (sem textura) e *layouts* simples.

A entropia de Shannon (SHANNON, 1948) é utilizada para separar o conteúdo do documento da borda (SILVA *et al.*, 2013). Esse algoritmo é capaz de remover bordas complexas (texturizadas, duplas e aderentes). Na próxima seção serão apresentadas as contribuições que buscam resolver os problemas mencionados na discussão acima.

### 3.1.2 Algoritmo para Remoção de Borda

O algoritmo desenvolvido foi publicado na *Lecture Notes in Computer Science* (SILVA *et al.*, 2013) e encontra-se anexo ao Apêndice A desta tese. O algoritmo proposto é dividido em quatro etapas:

- extração de cinco janelas (Figura 3.2);
- para cada bloco calcular a entropia de Shannon (SHANNON, 1948) por componente de cor;
- calcular a diferença entre os valores de entropia dos blocos laterais com o central;
- mapeamento dos contornos do documento com o auxílio do operador de Sobel (GONZALEZ e WOODS, 2008).



**Figura 3.2 Documento fotografado com as cinco janelas iniciais**



O algoritmo traz uma série de vantagens em relação ao estado da arte:

- 10,2 ms de processamento na linguagem Java para uma imagem de documento fotografado com borda texturizada a 16 Mpixel com flash (Canon 60D);
- bom funcionamento em imagens de baixa resolução e *layout* complexo;
- tratamento do problema de aderência da borda ao texto.

### 3.1.3 Experimentos e Resultados

O conjunto utilizado para os testes conta com 3.800 imagens de diversos tipos de documentos (burocráticos, revistas, livros, cartas e formulários) e com diferentes bordas (Figura 3.3).

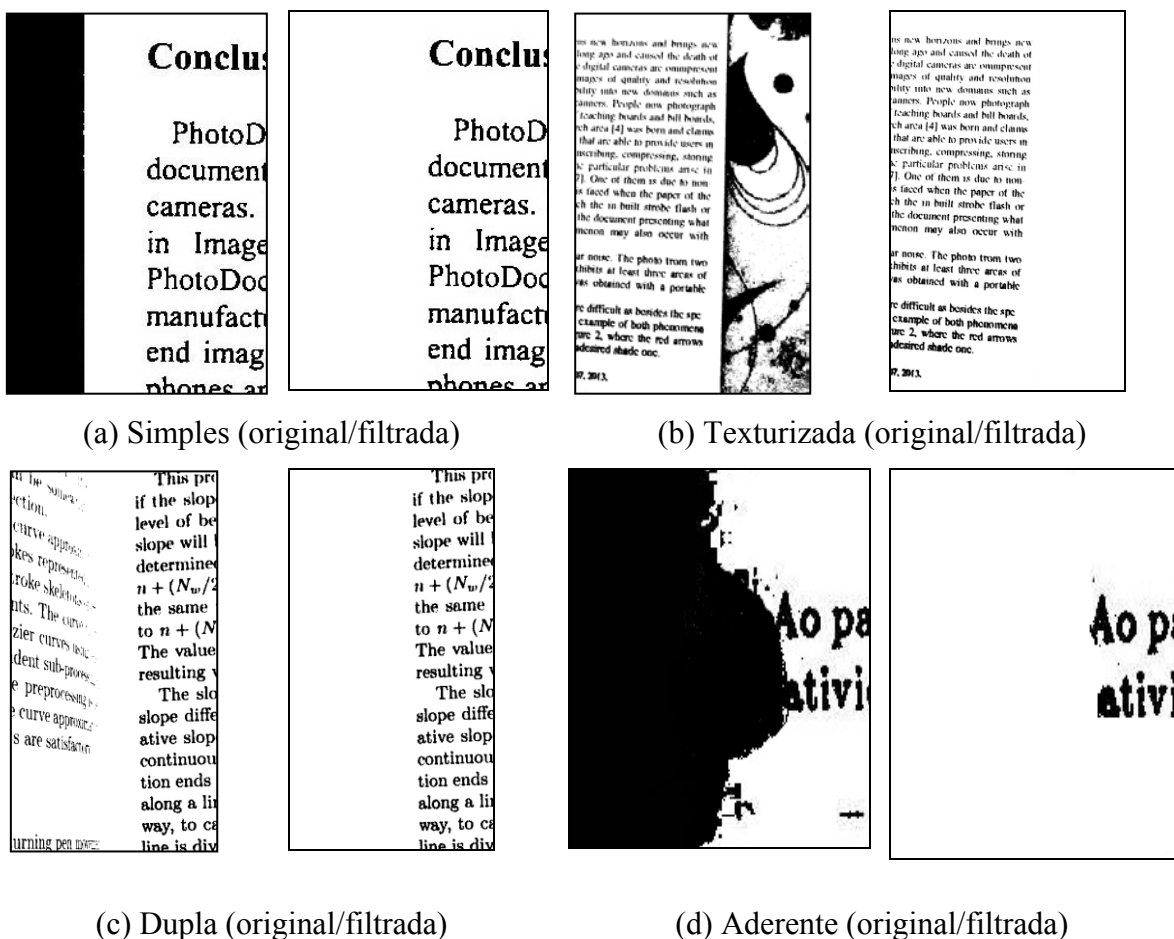


Figura 3.3 Tipos de ruído de bordas

A distribuição dos tipos de bordas é apresentada abaixo (Tabela 3.1).

**Tabela 3.1 Especificação e distribuição da base de dados**

	<b>Paleta</b>	<b>Tipo de borda</b>	<b>Quantidade</b>
<b>Scanner</b>	<i>Truecolor</i>	Simple	600
	<i>Truecolor</i>	Dupla	400
	Escala de cinza	Simple	400
	BW	Simple	600
	BW	Aderente	220
<b>Total</b>			2.220
<b>Câmera</b>	<i>Truecolor</i>	Simple	840
	<i>Truecolor</i>	Texturizada	340
	<i>Truecolor</i>	Dupla	400
<b>Total</b>			1.580

A metodologia de avaliação utilizou a comparação entre as imagens:

- original binarizada ( $I_O$ );
- original binarizada e remoção manual da borda ( $I_M$ );
- filtrada binarizada ( $I_F$ ).

Neste experimento, os pixels de borda foram substituídos por pixels brancos, de forma a preservar as dimensões das imagens. O fator de aceitação ( $f_a$ ) foi desenvolvido nesta tese e busca compensar erros provenientes da filtragem manual das imagens. O fator é definido pela equação abaixo:

$$f_a = 1 - \begin{cases} \left( \frac{SSIM(I_O|I_F)}{SSIM(I_O|I_M)} \right) & \text{para imagem original BW} \\ SSIM(I_F|I_M) & \text{para imagem original não BW} \end{cases}$$

A medida SSIM (WANG *et al.*, 2004) indica o grau de similaridade entre duas imagens compreendidas entre 0 e 1, em que quanto mais perceptível for a disparidade entre as imagens, mais próximo de 0 será o seu valor. Por outro lado, o valor 1 indica que as imagens são idênticas. Para os experimentos desta tese foi adotado  $0 \leq f_a < 0,01$ .

Os algoritmos de Ávila (ÁVILA e LINS, 2004) e Bukhari (BUKHARI *et al.*, 2012) foram comparados ao método proposto (Tabela 3.2).

**Tabela 3.2 Fator de aceitação ( $0 \leq f_a < 0,01$ ) para as técnicas de remoção de bordas**

<b>Imagens</b>	<b>Tipo de Borda</b>	<b>ÁVILA e LINS, 2004</b>	<b>BUKHARI et al., 2012</b>	<b>SILVA et al., 2013</b>
<i>Imagens em truecolor</i>				
<b>Escaneada</b>	Simple	0.93	0.90	0.93
<b>Escaneada</b>	Dupla	0.22	0.84	0.84
<b>Fotografada</b>	Simple	0.87	0.92	0.92
<b>Fotografada</b>	Texturizada	0.36	0.90	0.93
<b>Fotografada</b>	Dupla	0.31	0.96	0.92
<i>Imagens em escala de cinza</i>				
<b>Escaneada</b>	Simple	0.93	0.90	0.93
<i>Imagens BW</i>				
<b>Imagens</b>	Tipo de Borda	ÁVILA e LINS, 2004.	BUKHARI Et al., 2012	SILVA Et al., 2013
<b>Escaneada</b>	Simple	0.93	0.90	0.90
<b>Escaneada</b>	Aderente	0.89	0.12	0.85

O desempenho médio dos algoritmos foi de 68% (ÁVILA e LINS, 2004), 80% (BUKHARI *et al.*, 2012) e 90% (SILVA *et al.*, 2013) para a base de dados adotada. A análise dos dados aponta vantagens e desvantagens das três técnicas (Tabela 3.2).

Para documentos fotografados e que possuem bordas do tipo duplas, o algoritmo de (BUKHARI *et al.*, 2012) apresentou o melhor desempenho. O problema das bordas aderentes foi melhor tratado por (ÁVILA e LINS, 2004). Já para o caso das bordas texturizadas, o melhor algoritmo foi o de (SILVA *et al.*, 2013).

Outro parâmetro analisado é o tempo de processamento dessas técnicas (Tabela 3.3). A análise dos dados forneceu os tempos de 2,5 ms/imagem para o algoritmo (ÁVILA e LINS, 2004), 3,9 ms/imagem (BUKHARI *et al.*, 2012) e 3,4 ms/imagem (SILVA *et al.*, 2013). Os dados comprovam a eficiência da contribuição desta tese no problema de remoção de bordas em imagens de documentos. O mesmo apresenta um resultado médio para filtragem 10% superior ao segundo melhor e é apenas 0,9 ms em média mais lento que o algoritmo mais rápido.

**Tabela 3.3 Tempo de processamento dos métodos de remoção do ruído de borda**

<b>Técnica</b>	<b>Tipo de Borda</b>	<b>Tempo médio por imagem</b>
ÁVILA e LINS, 2004.	Simples	2,6 ms
BUKHARI <i>et al.</i> , 2012		3,8 ms
SILVA <i>et al.</i> , 2013		2,8 ms
ÁVILA e LINS, 2004.	Dupla	2,1 ms
BUKHARI <i>et al.</i> , 2012		4,2 ms
SILVA <i>et al.</i> , 2013		3,5 ms
ÁVILA e LINS, 2004.	Texturizada	2,1 ms
BUKHARI <i>et al.</i> , 2012		3,9 ms
SILVA <i>et al.</i> , 2013		3,5 ms
ÁVILA e LINS, 2004.	Aderente	3,2 ms
BUKHARI <i>et al.</i> , 2012		3,8 ms
SILVA <i>et al.</i> , 2013		3,8 ms

## 3.2 Interferência Frente e Verso

O ruído de interferência frente e verso pode ocorrer quando as duas faces de uma folha de papel são utilizadas, em decorrência do uso de papel translúcido e pela infiltração da tinta de uma face à outra. Seu efeito é similar aos discutidos no ruído de borda: (1) degrada a imagem visualmente; (2) aumenta o espaço necessário para armazenamento e transmissão via rede; (3) aumenta o gasto de tinta para impressão do documento e; (4) degrada a eficiência dos algoritmos de processamento de imagem (Figura 3.4).

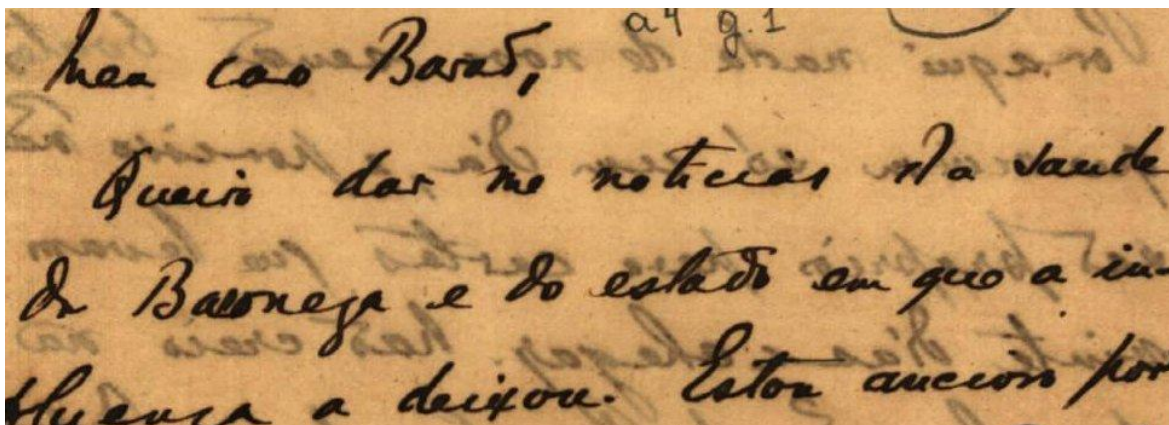


Figura 3.4 Trecho de documento com Interferência Frente e Verso

### 3.2.1 Trabalhos relacionados

A descrição deste fenômeno foi abordada pela primeira vez na literatura em 1995 por Lins (LINS *et al.*, 1995), que o chamou de "*back-to-front interference*". Outros trabalhos atribuíram novos nomes para o ruído como "*show-through*" (SHARMA, 2001) ou "*bleeding*" (KASTURI, 2002).

Para lidar com o problema, Lins e seus colaboradores propuseram uma filtragem em espelho, onde as duas faces do papel são escaneadas e alinhadas. A análise simultânea dos *pixels* de maior intensidade determina a sua origem em relação à face (frente ou verso) do documento. O grande problema dessa técnica, conforme documentado por Lins (LINS *et al.*, 1995) está na dificuldade no alinhamento preciso das imagens das faces em casos onde o documento foi dobrado ou rasgado, por exemplo, situação frequentemente encontrada nas cartas de Joaquim Nabuco.

A técnica de espelhamento foi patenteada pela Xerox em 2001 (SHARMA, 2001). O trabalho de (SHARMA, 2001) usa o alinhamento das faces para remover a interferência e

assume documentos sem problemas de dobramento e rasgos (NISHIDA e SUZUKI, 2003). Recentemente foram propostos novos trabalhos com o uso da técnica de espelhamento (SU e DJAFARI, 2007; TONAZZINI *et al.*, 2007; MOGHADDAM e CHERIET, 2009), porém os trabalhos não superaram os problemas relatados em (LINS, 1995).

A literatura apresenta diversos outros esquemas para a remoção de interferência frente e verso: modelos de *waterflow* (OHA *et al.*, 2005) e filtragem por wavelet (CAO *et al.*, 2001). As mais bem sucedidas parecem ser técnicas baseadas em limiarização (LINS e SILVA, 2007).

Alguns algoritmos baseados em entropia foram propostos para lidar diretamente com o problema de interferência frente e verso (MELLO e LINS, 2002) e (SILVA *et al.*, 2005). O algoritmo de binarização de Gatos (GATOS *et al.*, 2004; GATOS *et al.*, 2006) é adaptado, acrescentando limiares secundários em (BURGOYNE *et al.*, 2008). Já o algoritmo adaptativo de Sauvola e Pietikäinen (SAUVOLA e PIETIKAINEN, 2000), é melhorado por um classificador *fuzzy* (CASTRO *et al.*, 2007). Diversos outros métodos de limiarização foram propostos para o tratamento desta classe de ruído (KAPUR *et al.*, 1985; LEEDHAM *et al.*, 2002; KAVALLIERATOU e ANTONOPOULOU, 2005).

### 3.2.2 Algoritmo para filtragem de Interferência Frente e Verso

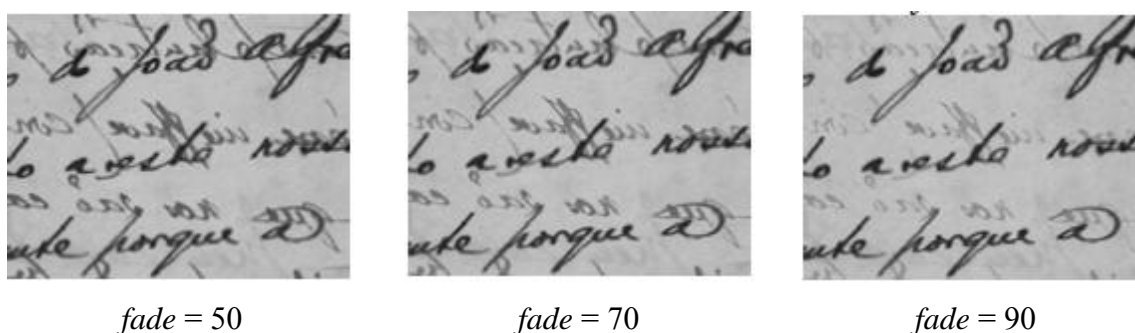
Os resultados descritos nesta tese apresentam um algoritmo que:

- é relativamente rápido, 140 ms de processamento em média por imagem RGB (1066x1612);
- funciona em imagens com textos manuscritos e impressos;
- funciona com intensidades variáveis de ruído.

A ideia do algoritmo é otimizar o fator de perda  $\alpha$ . O ajuste desse parâmetro permite uma melhor afinação estatística entre as distribuições dos histogramas originais e binários (SILVA *et al.*, 2007). A classificação e caracterização de ruído de interferência, apresentada na seção 2.3 desta tese, permite um ajuste local do fator de perda, melhorando a filtragem e a reconstrução da imagem original. Os trabalhos para remoção da interferência frente e verso propostos nesta tese podem ser visualizados nas referências (SILVA *et al.*, 2009; SILVA *et al.*, 2010c), anexos ao Apêndice A.

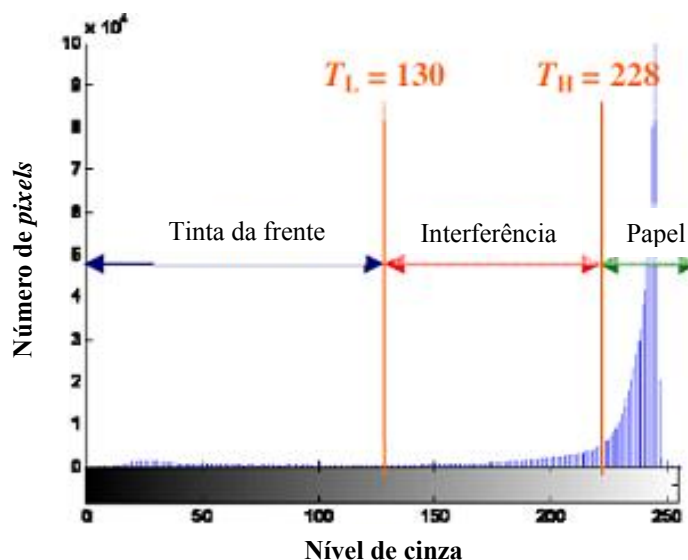
### 3.2.3 Experimentos e Resultados

A base de dados é formada por 1.000 imagens em escala de cinza que serviram de base para gerar 3.000 imagens com interferência frente e verso sintética. A inserção de ruído é controlada pelo parâmetro de infiltração (fator *fade*) que variou aleatoriamente entre três intervalos: Fraco (80, 100], Médio (60,80] e Forte [30, 60] nos blocos de imagens (Figura 3.5). Os intervalos de interferência foram criados considerando o mapeamento dos blocos de ruído apresentados na seção 2.3 desta tese.



**Figura 3.5** Exemplo de imagem sintética com diferentes níveis de *fade*

O modelo matemático de geração das imagens sintéticas é apresentado em (LINS *et al.*, 2007a). A ideia do método é inserir a interferência frente e verso na imagem considerando o intervalo do histograma da imagem em nível de cinza (Figura 3.6).

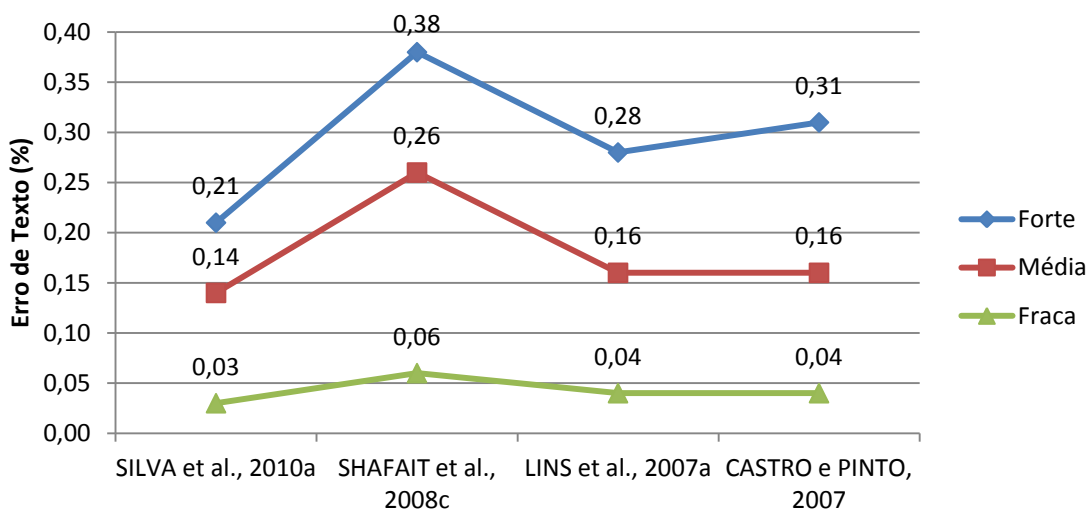


**Figura 3.6** Exemplo de histograma de imagem de documento com interferência.

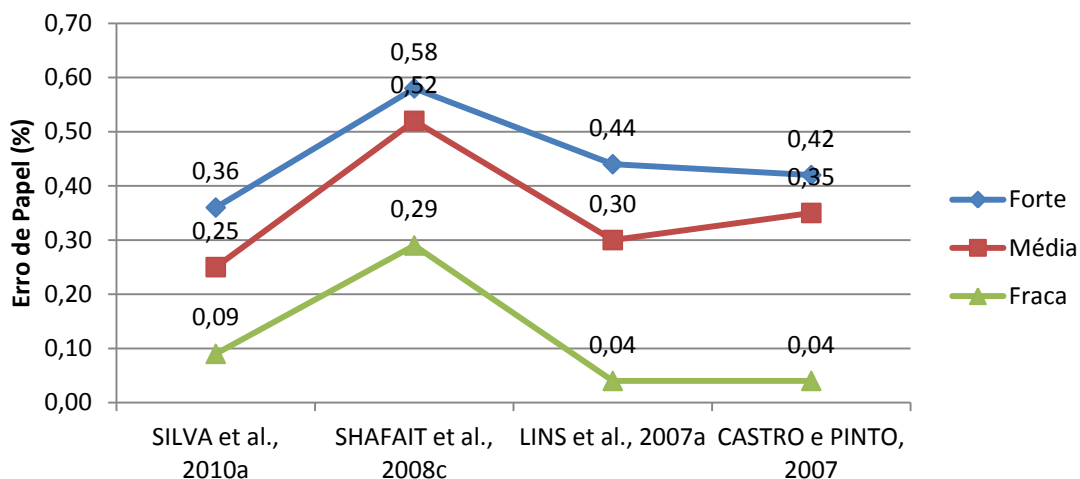
O método de avaliação dos resultados consiste em comparar duas imagens binárias, que são a imagem original e sua imagem sintética filtrada. Três medidas de qualidade são extraídas dessa comparação:

- Erro de texto: número de *pixels* de texto da imagem original que foram apagados na imagem filtrada (Figura 3.7);
- Erro de Papel: número de *pixels* de fundo da imagem original que não são interferência, mas que foram incluídos (Figura 3.8);
- Erro de Interferência: número de *pixels* de interferência não filtrados (Figura 3.9).

O resultado do método proposto foi comparado aos algoritmos de (SILVA *et al.*, 2007; CASTRO e PINTO, 2007; SHAFAIT *et al.*, 2008c). Os resultados são apresentados na forma de gráfico.

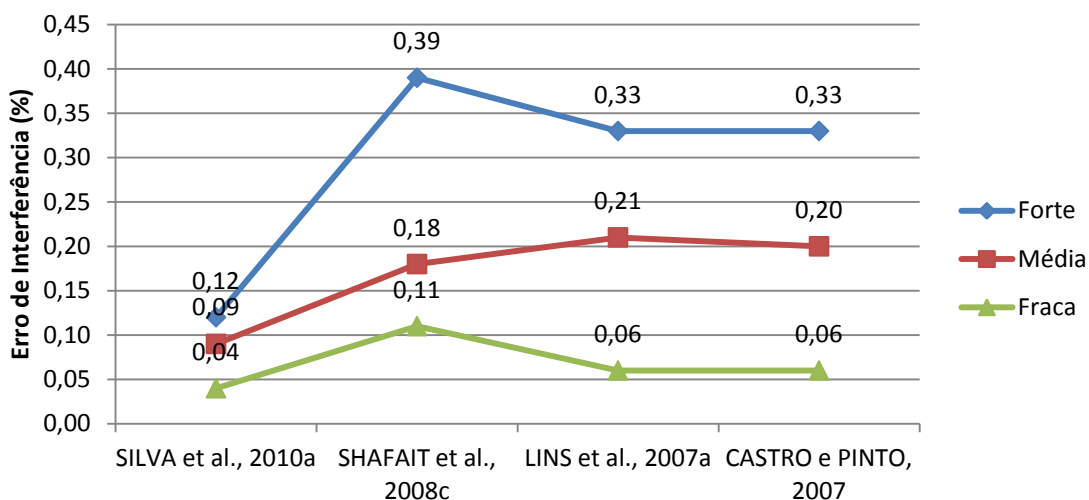


**Figura 3.7 Medida de Erro de Texto dos algoritmos de remoção de interferência frente e verso.**

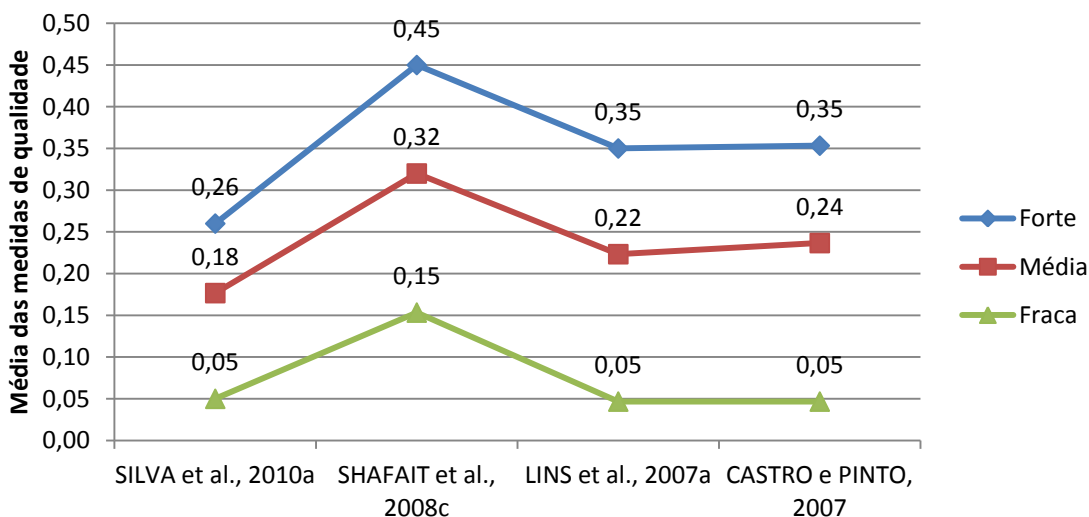


**Figura 3.8 Medida de Erro de Papel dos algoritmos de remoção de interferência frente e verso.**





**Figura 3.9 Medida de Erro de Interferência dos algoritmos de remoção de interferência frente e verso..**



**Figura 3.10 Resultado da filtragem de interferência em relação as medidas de qualidade.**

A análise dos dados apontou resultados próximos para o nível de ruído fraco. Para os demais níveis, o algoritmo proposto apresentou melhores resultados. O algoritmo proposto por esta tese utiliza informações da classificação de ruído (Seção 2.3) o que representa um acréscimo de tempo de processamento. Em relação ao tempo, esse algoritmo é três vezes mais lento que o (LINS *et al.*, 2007a), mas apresenta resultados 20% melhores para as três medidas adotadas.

### 3.3 Remoção de Ruído Especular

Alguns problemas particulares surgem na digitalização de documentos usando câmeras digitais portáteis (LINS *et al.*, 2010d). Um deles é devido a não uniformidade da iluminação. Uma situação complexa é enfrentada quando o papel do documento é brilhante. O flash ou a iluminação intensa do ambiente "apaga" partes do documento, o que é chamado o "ruído especular" (GONZALEZ e WOODS, 2008). O mesmo fenômeno pode ocorrer com objetos 3D se os mesmos refletirem parte da iluminação incidente (Figura 3.11).



(a) documento



(b) objeto

**Figura 3.11 Exemplos de imagens com Ruído Especular.**

#### 3.3.1 Trabalhos Relacionados

O componente de reflectância difusa pode ser descrito pelo modelo Lambertiano e aproximações podem ser usadas com sucesso no mundo real (cenas não Lambertianas) (MALLICK *et al.*, 2006). Um conjunto de algoritmos de visão computacional usa este componente para realizar suas tarefas (BLAKE E BRELSTAFF, 1988; BROCKETT E MARAGOS, 1994; OSADCHY *et al.*, 2003). O modelo de reflectância dicromático (SHAFER, 1985) assume uma distribuição espectral de componente especular semelhante a de um aparelho de iluminação, enquanto que a distribuição do componente difuso depende fortemente das propriedades do material da superfície. O modelo dicromático sugere a possibilidade de decompor uma imagem em seus componentes especular e difuso, com base na informação de cor.

Além das abordagens globais mencionados acima, há um considerável interesse em separar os componentes de reflexão através de interações puramente locais. A maior parte

dos métodos locais assume que a cor da luz é conhecida a priori o que não é uma restrição severa, pois pode ser estimada usando métodos globais (LEE, 1986). Um dos primeiros métodos locais utiliza a cromaticidade como uma polarização adicional para permitir a recuperação de uma cor fonte espacialmente variável (NAYAR *et al.*, 1997).

O método proposto na referência (TAN *et al.*, 2003) permite que o usuário especifique uma curva fechada em torno da região especular e, em seguida, minimize a função objetivo (PING *et al.*, 2003) com base em variações locais de cromaticidade difusa e intensidade especular. Já o método desenvolvido no domínio contínuo da imagem efetua interações locais regidas por equações diferenciais parciais (EDPs) (MALLICK *et al.*, 2006).

### 3.3.2 Algoritmo para Remoção do Ruído Especular

O interesse da pesquisa apresentada nesta tese foi o problema da remoção do ruído especular nas imagens de objetos e documento capturadas pelo *scanner* 3D da Hewlett-Packard (Figura 3.12).



**Figura 3.12** *Scanner* 3D HP TopShot Laser Printer.

O *scanner* conta com uma câmera de 8 *Megapixels* e três LEDs (esquerdo, central e direito). O processo de digitalização deste *scanner* consiste na captura de três imagens sequenciais alternando a fonte de iluminação dos LEDs (Figura 3.13). A ideia do trabalho é utilizar a combinação das três imagens para gerar uma quarta de melhor qualidade. Dois algoritmos foram propostos (MARIANO *et al.*, 2011; LINS *et al.*, 2013). O primeiro busca eliminar o ruído especular e o segundo remove as áreas de sombreamentos geradas por objetos 3D.



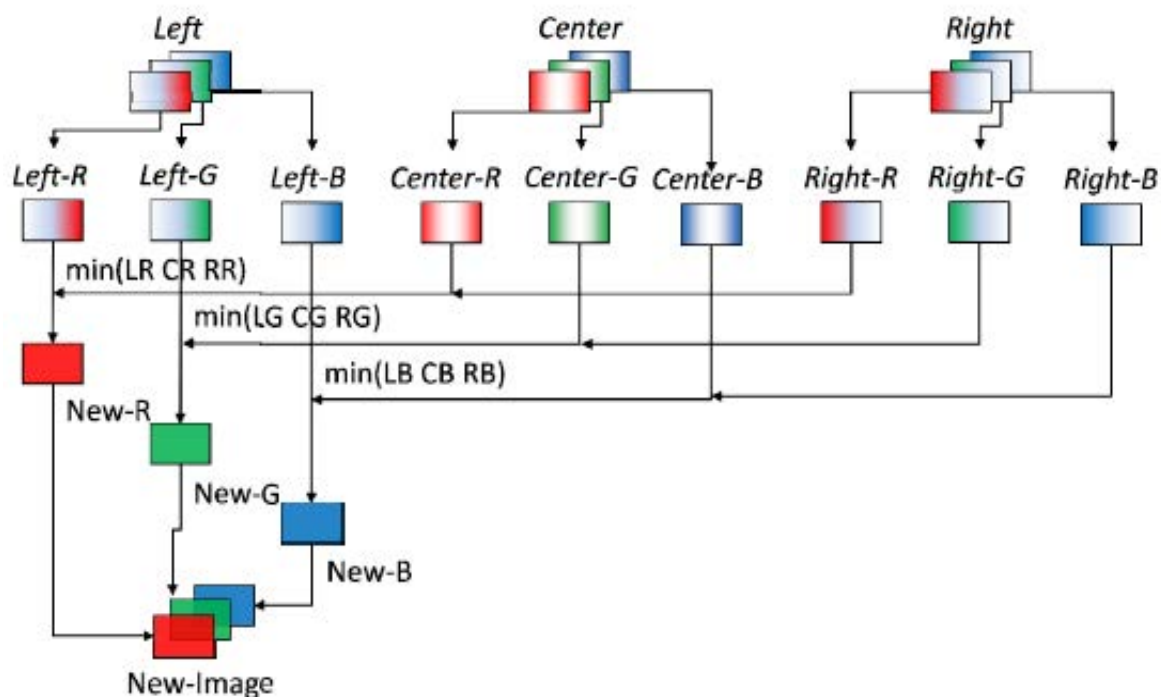
LED esquerdo acionado

LED central acionado

LED direito acionado

**Figura 3.13 Sequência de imagens capturadas pelo  
Scanner 3D HP TopShot Laser Printer.**

As áreas de incidência do ruído especular são detectadas e classificadas com o método descrito na seção 2.3 desta tese. Em seguida, são escolhidos os blocos que, entre as três imagens, serão combinadas para formar a nova imagem. Por fim, são comparadas as intensidade dos *pixels* entre as componentes das imagens. O *pixel* de menor intensidade em cada componente de cor é escolhido (MARIANO *et al.*, 2011). Esses *pixels* são combinados para formar a nova imagem colorida (Figura 3.14).



**Figura 3.14 Geração da imagem filtrada a partir da sequência de imagens do  
Scanner 3D HP TopShot Laser Printer**

O resultado do algoritmo de remoção do ruído especular é apresentado na Figura 3.15.



(a) LED esquerdo acionado

(b) LED central acionado

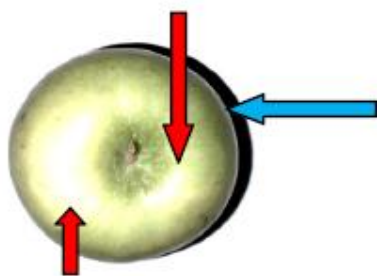


(c) LED direito acionado

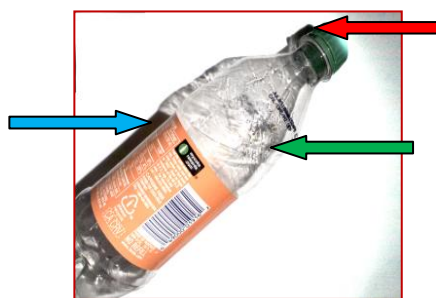
(d) Combinação das outras três Imagens

### Figura 3.15 Resultado da remoção do ruído Especular em documentos.

Algumas características são próprias dos objetos e, além do ruído especular, podem ocorrer áreas de sombreamento e material transparente (Figura 3.16). Desta forma, uma nova etapa foi incluída para eliminar áreas com sombras nas imagens (LINS et al., 2013).



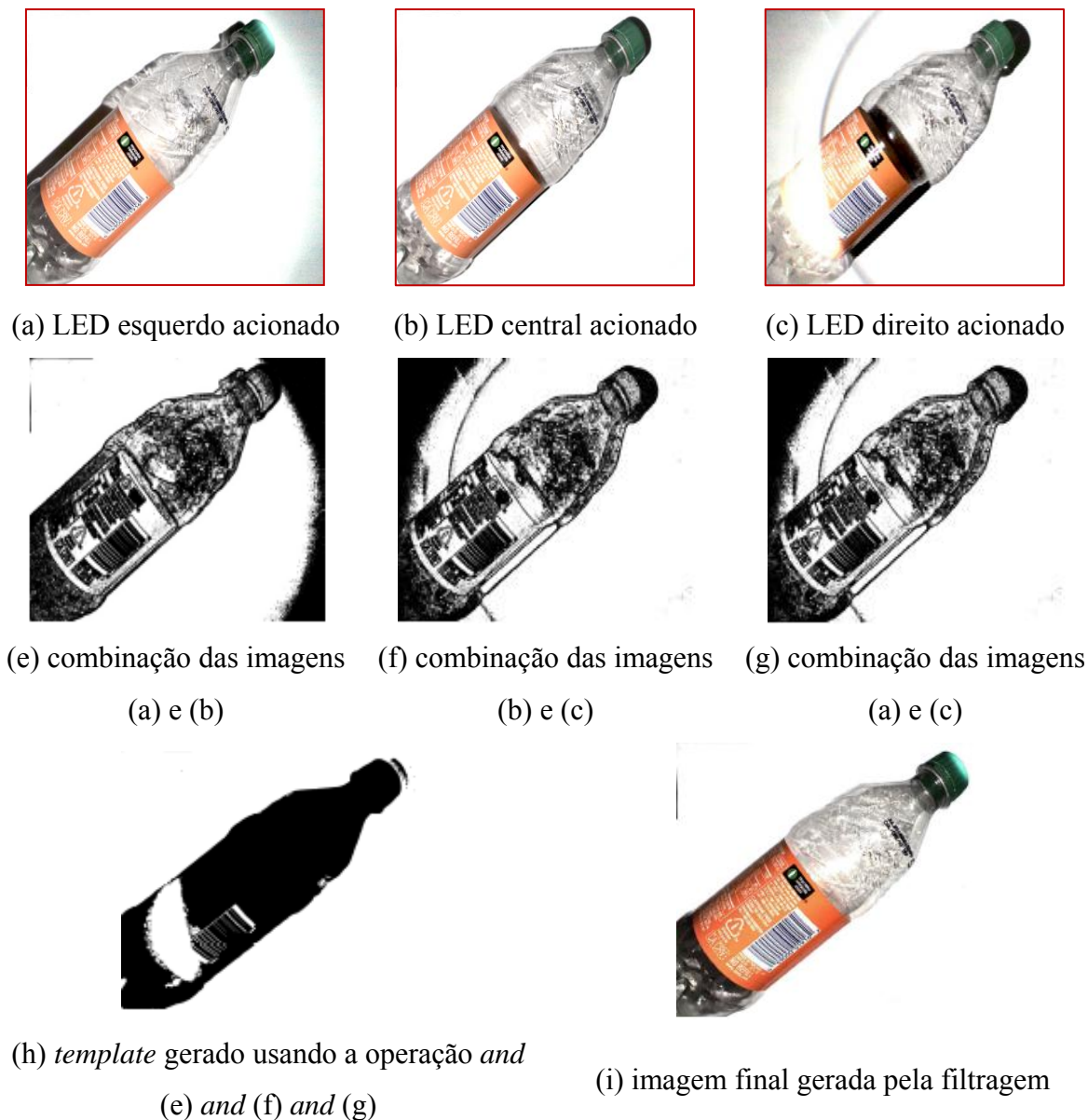
(a) ruído especular e sombreamento



(b) ruído especular, sombreamento e transparência

### Figura 3.16 Exemplos de objetos capturados por meio de scanner 3D

O tratamento das regiões sombreadas na imagem é realizado por meio de uma máscara binária. A máscara é utilizada como *template* para remover as regiões de sombreado da imagem (LINS *et al.*, 2013). O procedimento para remover o ruído especular é o mesmo adotado no primeiro algoritmo. A filtragem de um objeto 3D com ruído especular, sombreado e transparência é apresentada na Figura 3.17.



**Figura 3.17** Resultado da remoção do ruído Especular em objetos.

### 3.4 Processamento Inteligente de Imagens de Documentos

O aumento da escala e da complexidade relacionada ao processamento de imagens de documentos exige a automatização do fluxo de trabalho. Tradicionalmente, as ferramentas de processamento tratam as bases de documentos de forma uniforme com o mesmo conjunto de algoritmos em que os ajustes são realizados manualmente. Isso torna o tratamento de bases heterogêneas ineficiente e custoso.

#### 3.4.1 Trabalhos Relacionados

Diversas ferramentas tratam o processamento de imagens de documentos em *batch* (LINS *et al.*, 2006; SILVA *et al.*, 2010a; LAZZARA *et al.*, 2011; GATOS *et al.*, 2014). Essas ferramentas assumem para cada tarefa de filtragem um único algoritmo "ótimo" e que as imagens possuem sempre um conjunto de problemas bem definido. O problema dessas abordagens é que não existe um algoritmo capaz de obter sempre o melhor resultado e a filtragem "cega" do ruído pode causar perda de informação. Dessa forma, as ferramentas tradicionais não atendem a atual demanda causada pelo aumento da heterogeneidade dos documentos. O levantamento da literatura científica e das ferramentas comerciais não apresentou uma solução próxima da proposta por esta tese.

#### 3.4.2 Método para Processamento Inteligente de Documentos

A ideia apresentada nesta tese para o problema de processamento inteligente de documentos é composta por um Sistema de Raciocínio Baseado em Casos (SRBC). Trata-se de uma abordagem de apoio à decisão semelhante ao modelo usado por humanos na resolução de problemas. O desenvolvimento de um SRBC permitiu a escolha e o ajuste inteligente de algoritmos para o processamento de imagens de documentos.

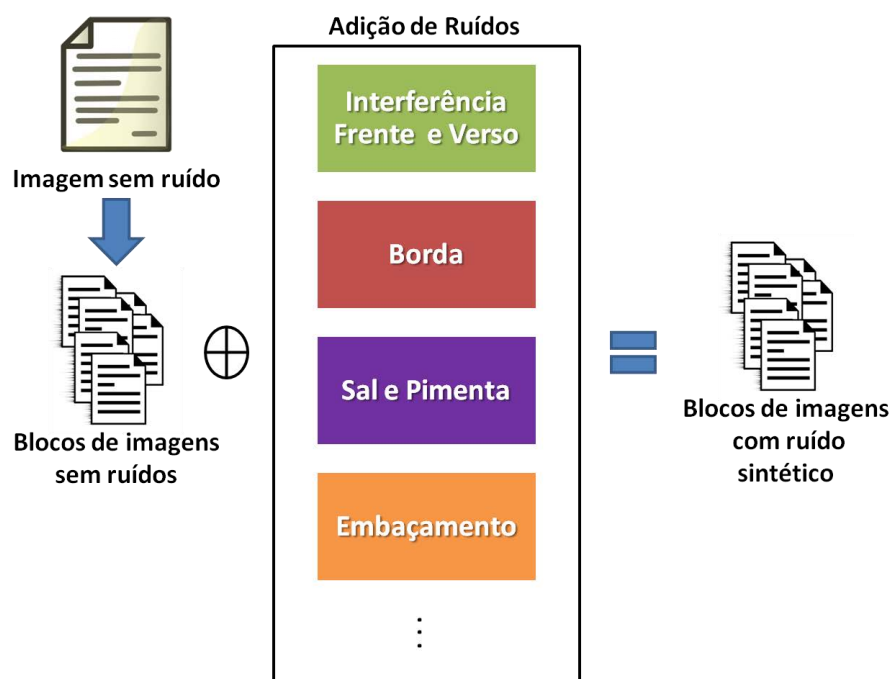
O Raciocínio Baseado em Casos (RBC) utiliza a experiência de casos anteriores para resolver novos problemas (LÓPES *et al.*, 2005). O ciclo de um RBC pode ser definido como uma série de quatro processos consecutivos (AAMODT e PLAZA, 1994), conhecidos como os quatro **REs**:

- **RE**cuperar: dado um problema atual, um ou mais casos bem-sucedidos são recuperados do repositório de casos. A recuperação de casos é feita de acordo

com a similaridade entre os mesmos;

- **REusar**: é sugerido o uso ou adaptação de um caso recuperado na fase anterior para resolver o problema atual;
- **REvisar**: os resultados são avaliados, revisados e ajustados;
- **REter**: a solução revisada pode ser retida como um novo caso, expandindo, portanto, o repositório de casos.

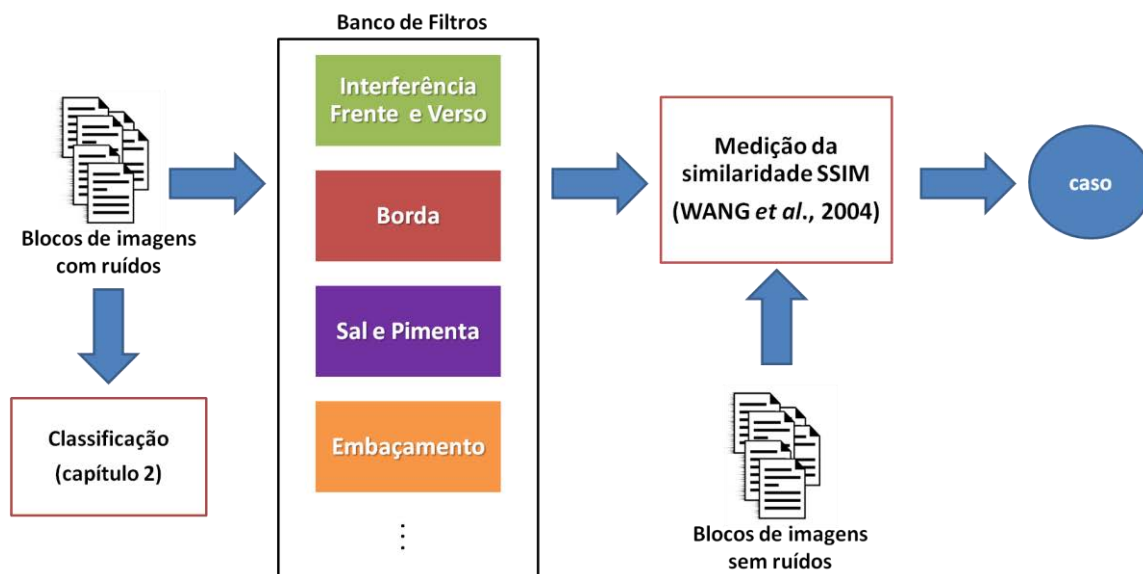
A arquitetura adotada no método proposto utiliza um método de geração automática de casos, que é dividido em duas etapas. A primeira etapa consiste em gerar blocos de imagens com ruído sintético a partir de uma imagem sem ruído (Figura 3.18). Nessa etapa, são armazenadas informações sobre os parâmetros de ruído sintético, o tipo de imagem e o dispositivo de captura.



**Figura 3.18 Geração de ruído sintético.**

A segunda etapa é responsável por gerar os casos. Nessa etapa, os blocos são filtrados para o tipo de ruído indicado na etapa anterior. Para cada ruído são testados diferentes algoritmos de filtragem variando-se seus parâmetros. Por fim, as saídas do banco de filtro são comparadas aos blocos da imagem original (Figura 3.19).





**Figura 3.19** Geração de casos.

Os casos aceitos são aqueles que apresentam medida de similaridade entre os blocos  $\geq 0,995$ , ou seja, foi adotado um valor de significância  $p < 0.05$ . A medida SSIM foi adotada para medir o grau de similaridade entre os blocos, em que quanto mais perceptível for a disparidade entre as imagens, mais próxima de 0 será o seu valor. Por outro lado, o valor 1 indica que as imagens são idênticas.

No sistema proposto foi adotada uma representação vetorial. Os vetores de caso são formados pelos atributos:

- gerados pelos sistemas de classificação e pelas características extraídas apresentadas no capítulo 2 desta tese;
- média das características de todos os blocos de ruído que formam a imagem;
- rótulo do ruído armazenado na etapa de geração das imagens sintéticas;
- algoritmo e parâmetros utilizados na filtragem.

O banco de filtro conta os algoritmos presentes nos *softwares* (ImageJ, LINS *et al.*, 2011b; LAZZARA *et al.*, 2011; GATOS *et al.*, 2014). Foram gerados em torno de 3 milhões de casos, o que tornou a etapa de recuperação lenta. Para reduzir a granularidade dos casos, o algoritmo de clusterização k-means (ARTHUR e VASSILVITSKII, 2007) foi adotado para reduzir o espaço de casos para 150 mil. Os resultados das classificações

apresentadas no capítulo 2 desta tese, são usados como índices da tabela de casos (Figura 3.20).

<b>Classe de Imagem</b>	<b>Dispositivos de Captura</b>	<b>Ruído</b>	<b>Algoritmo</b>	<b>Parâmetros utilizados</b>	<b>...</b>
<b>documento</b>	<b>escaneado</b>	<b>Interferência Frente e Verso (Forte)</b>	<b>Silva et al ICPR 2010</b>	<b><math>\alpha = 0.64</math></b>	<b>...</b>
<b>logo</b>	<b>escaneado</b>	<b>Sal e Pimenta</b>	<b>Mediana</b>	<b>Janela 5x5</b>	<b>...</b>

**Figura 3.20 Trecho da tabela de casos do SRBC proposto.**

### 3.4.3 Experimentos e Resultados

Em sistemas RBC, um ou mais casos são recuperados de uma base de casos de acordo com as similaridades com o problema atual. Esse procedimento é realizado na fase "REcuperar" do ciclo RBC. A distância de Hamming (HAMMING, 1950) foi adotada para a recuperação dos casos. O tempo médio para recuperação de casos foi de 3,21 ms.

Para a validação do método proposto, foram utilizadas imagens com ruídos natural e sintético (Tabela 3.4).

**Tabela 3.4 Imagens utilizadas para o experimento.**

<b>Tipo de Ruído</b>	<b>Ruído Natural</b>	<b>Ruído Sintético</b>
<b>Bordas</b>	3.800	3.000
<b>Interferência frente e verso</b>	400	3.000
<b>Sal e Pimenta</b>	0	3.000
<b>Orientação e Inclinação</b>	1.000	3.000
<b>Embaçamento</b>	0	3.000

O uso do sistema de RBC proposto nesta tese apresentou melhorias na filtragem em todas as classes de ruídos. Para as imagens com ruído natural de borda, o resultado foi de 98% das imagens corretamente filtradas, o que representou uma melhora de 8% em relação ao melhor algoritmo testado e para as imagens com borda sintética o ganho foi de 13%.

O ruído de interferência frente e verso apresentou uma diminuição do erro médio das medidas de qualidade em 20% , 23% e 2% para os níveis de interferência sintética forte, média e fraca, respectivamente. No caso das imagens com ruído natural, o ganho médio foi

de 27%. A remoção do ruído de Sal e Pimenta obteve um ganho de 38% em relação ao melhor resultado obtido com uma filtragem "cega". O ganho na correção de Orientação e Inclinação foi de 12%. Finalmente o ganho apresentado pelo algoritmo de correção de embaçamento proposto por (OLIVEIRA *et al.*, 2013) foi de 33%. Apesar do tempo médio por imagem ter aumentado 232 ms, o ganho de qualidade das imagens torna o uso do método desenvolvido nesta tese viável para o processamento de grandes massas de imagens de documentos heterogêneos.

## Capítulo 4

# Transcrição Automática de Documentos Históricos

Nos últimos anos, as aplicações de transcrição automática têm se popularizado, devido ao seu fácil manuseio e bons resultados obtidos. Porém, este sucesso é limitado ao reconhecimento de caracteres impressos e textos cursivos capturados *on-line* (movimentos da caneta são registrados em tempo real) (PLAMONDON e SRIHARI, 2000). Já a transcrição a partir de uma imagem com texto cursivo manuscrito mostra bons resultados em domínios restritos (vocabulários pequenos e redundantes), como processamento de cheques e serviços postais (KORNAI *et al.*, 1996).

A qualidade dos documentos é também um fator que influencia a transcrição, uma vez que as ferramentas assumem que os documentos estão livres de ruídos e com os caracteres bem conservados. No caso de documentos históricos, essa premissa quase nunca é verdadeira. Os recentes esforços em digitalizar grandes coleções de documentos históricos reacenderam o interesse na tarefa de transcrição de documentos degradados e com vocabulários extensos (Impact; Kb Digitization; Google Book; Million Book).

Este capítulo apresenta dois novos métodos para melhorar a transcrição de documentos históricos.

### 4.1 Trabalhos Relacionados

A maioria das abordagens para transcrição de texto em imagens segmenta as palavras em fragmentos menores (geralmente caracteres) que são reconhecidos separadamente. O resultado da transcrição de uma palavra é a composição dos fragmentos reconhecidos individualmente. No entanto, esses métodos são dependentes da segmentação, tornando-se um problema complexo. A alternativa é reconhecer a palavra como um todo e essa abordagem é viável em bases com vocabulário pequeno (KORNAI *et al.*, 1996). Porém, quando o vocabulário é extenso, essa abordagem se torna custosa, pois seria necessária a extração de pelo menos um exemplo para cada palavra (RATH e MANMATHA, 2003).

Documentos impressos em bom estado de conservação apresentam uma taxa elevada de transcrição (maior que 90%) pelas melhores ferramentas comerciais (LINS e SILVA,

2007). Entretanto, o desempenho dessas ferramentas é reduzido em imagens de documentos históricos que apresentam degradação da informação textual.

A adição de uma etapa de pré-processamento para melhorar a informação textual (palavras/caracteres) é proposta por diversos trabalhos (LINS e SILVA, 2007b; BEUSEKOM *et al.*, 2007; DRIRA *et al.*, 2007; LIKFORMAN *et al.*, 2011; GANGAMMA *et al.*, 2012). Essa etapa pode ser realizada através da eliminação dos ruídos da imagem e/ou da reconstrução da informação textual. A primeira abordagem apresenta bons resultados tanto para texto cursivo manuscrito quanto para impresso (SILVA *et al.*, 2010a). A segunda é utilizada em caracteres impressos, devido a seu menor grau de variabilidade de estilos em relação aos manuscritos (MOGHADDAM e CHERIET, 2011; SHIRAI *et al.*, 2013).

Diferentes esquemas de binarização são usados para recuperar caracteres e gerar possíveis sequências utilizando uma estrutura de árvore para reconstruir palavras (RAY *et al.*, 2013). No entanto, essa abordagem é limitada ao nível de degradação da unidade (palavra/caractere) que será recuperada. A reconstrução usando duas imagens do mesmo documento histórico de fontes diferentes foi proposta para tratar rasgos no documento e manchas de tintas (BEUSEKOM *et al.*, 2007). Essa abordagem depende da existência de outro exemplar do documento e do correto alinhamento entre as páginas.

Outra linha de estudo é inspirada na psicologia cognitiva, que propõe uma abordagem holística para o reconhecimento de palavras. O levantamento do paradigma holístico e suas aplicações em reconhecimento de palavras manuscritas são apresentados em (MADHVANATH e GOVINDARAJU, 2001). Duas abordagens holísticas são apresentadas. A primeira trata uma palavra como uma coleção de subunidades mais simples (caracteres). A outra abordagem foi inspirada em estudos psicológicos da leitura humana, em que foi evidenciado que os seres humanos usam características, como comprimento, ascendentes e descendentes na leitura, para o reconhecimento das palavras.

O trabalho apresentado em (LAVRENKO *et al.*, 2004) utiliza o conceito de reconhecimento usando características holísticas para reconhecer palavras manuscritas por um mesmo autor. O Hidden Markov Models (HMM) é um exemplo comum e bem sucedido para reconhecimento de amostras não segmentadas (RABINER, 1989; LU *et al.*, 1999; MARTI e BUNKE, 2002; INDERMUHLE *et al.*, 2008). O conhecimento do léxico da linguagem pode ser incorporado ao processo de reconhecimento do texto e essa modelagem pode ser estendida ao HMM (MARTI e BUNKE, 2002). Os manuscritos de

George Washington foram utilizados nos trabalhos de (LAVRENKO *et al.*, 2004; FENG *et al.*, 2006), em que um HMM e características holísticas foram usados para o reconhecimento de palavras.

Recentemente, o uso de classificadores multidimensionais/multiníveis (BERNARD *et al.*, 2007; GRAVES *et al.*, 2007), sistemas híbridos (BEZERRA *et al.*, 2012; ZANCHETTIN *et al.*, 2012) e Redes Neurais Recorrentes (GRAVES e SCHMIDHUBER, 2009; BEZERRA *et al.*, 2012), apresentaram resultados promissores para a tarefa de reconhecimento *off-line* de caracteres manuscritos. A competição para o reconhecimento de escrita manual (EL ABED *et al.*, 2009) apresentou como ganhador um modelo de reconhecimento híbrido. O modelo vencedor é composto por uma hierarquia de Redes Neurais Recorrentes Multidimensionais (RNRMD) (GRAVES e SCHMIDHUBER, 2009) que utiliza a arquitetura de memória de curto prazo para termos longos (GRAVES *et al.*, 2007) e uma classificação temporal (GRAVES *et al.*, 2006). Esse modelo apresenta propriedades desejáveis para o problema de reconhecimento de escrita: alfabeto independente, globalmente treinável e não precisa de um processo de extração de características explícito. Esta última propriedade pode causar erros de classificação em caracteres/palavras semelhantes (BEZERRA *et al.*, 2012). O problema de erros de classificação em caracteres semelhantes é tratado no trabalho (BEZERRA *et al.*, 2012) em que um SVM é utilizado para corrigir os erros de classificação.

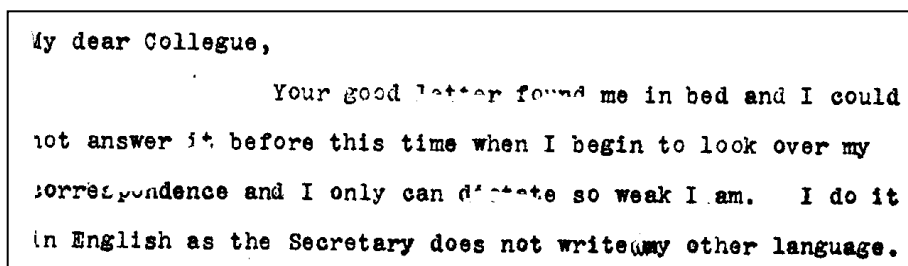
## 4.2 Contribuições

Esta seção apresenta dois métodos para transcrição de imagens de documentos históricos. O primeiro método busca reconstruir a informação textual das imagens de documentos históricos impressos. O outro apresenta uma metodologia para geração automática de *fontsets* para documentos manuscritos cursivos.

### 4.2.1 Método 1 (Reconstrução da Informação Textual)

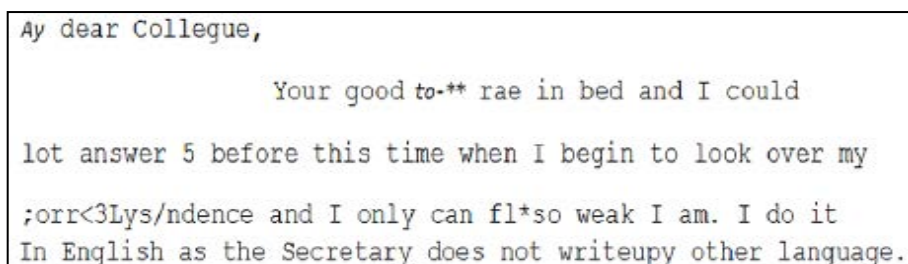
O método apresenta uma nova abordagem para a melhoria do reconhecimento de caracteres em documentos históricos degradados. Algumas das limitações de outros métodos, como existência de mais de um exemplar do documento e alinhamento entre as páginas, são superadas no método aqui apresentado. O sistema proposto consiste na identificação de regiões em que há perda de informação devido a ruídos físicos ou à

filtragem dos documentos. A ideia é reconstruir o documento com seu próprio *fontset* gerando possíveis candidatos para a transcrição do texto correto (Figura 4.1).



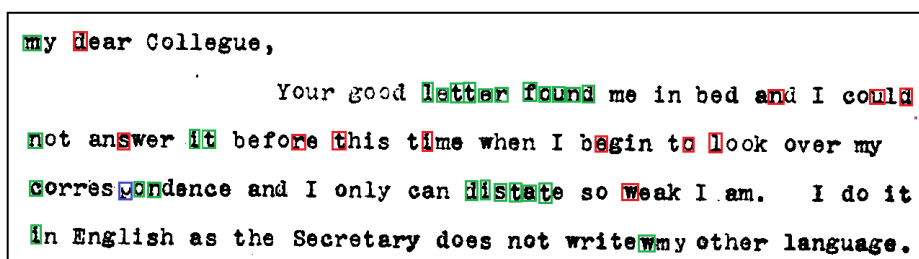
My dear Colleague,  
Your good letter found me in bed and I could  
not answer it before this time when I begin to look over my  
correspondence and I only can distate so weak I am. I do it  
in English as the Secretary does not write my other language.

(a) Trecho de carta de Nabuco binarizado.



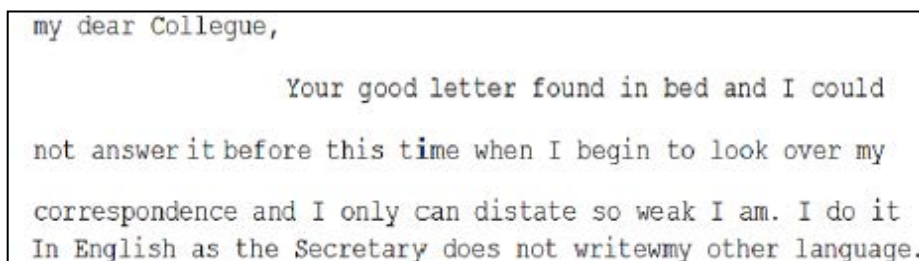
My dear Colleague,  
Your good to-\*\* rae in bed and I could  
lot answer 5 before this time when I begin to look over my  
;orr<3Lys/ndence and I only can fl\*so weak I am. I do it  
In English as the Secretary does not writeupy other language.

(b) Transcrição da imagem (a) pela ferramenta ABBYY FineReader Pro 12



my dear Colleague,  
Your good letter found me in bed and I could  
not answer it before this time when I begin to look over my  
correspondence and I only can distate so weak I am. I do it  
In English as the Secretary does not write my other language.

(c) Trecho de carta de Nabuco reconstruído pelo método proposto.



my dear Colleague,  
Your good letter found in bed and I could  
not answer it before this time when I begin to look over my  
correspondence and I only can distate so weak I am. I do it  
In English as the Secretary does not writewmy other language.

(d) Transcrição da imagem (c) pela ferramenta ABBYY FineReader Pro 12

### Figura 4.1 Exemplo de reconstrução de informação textual

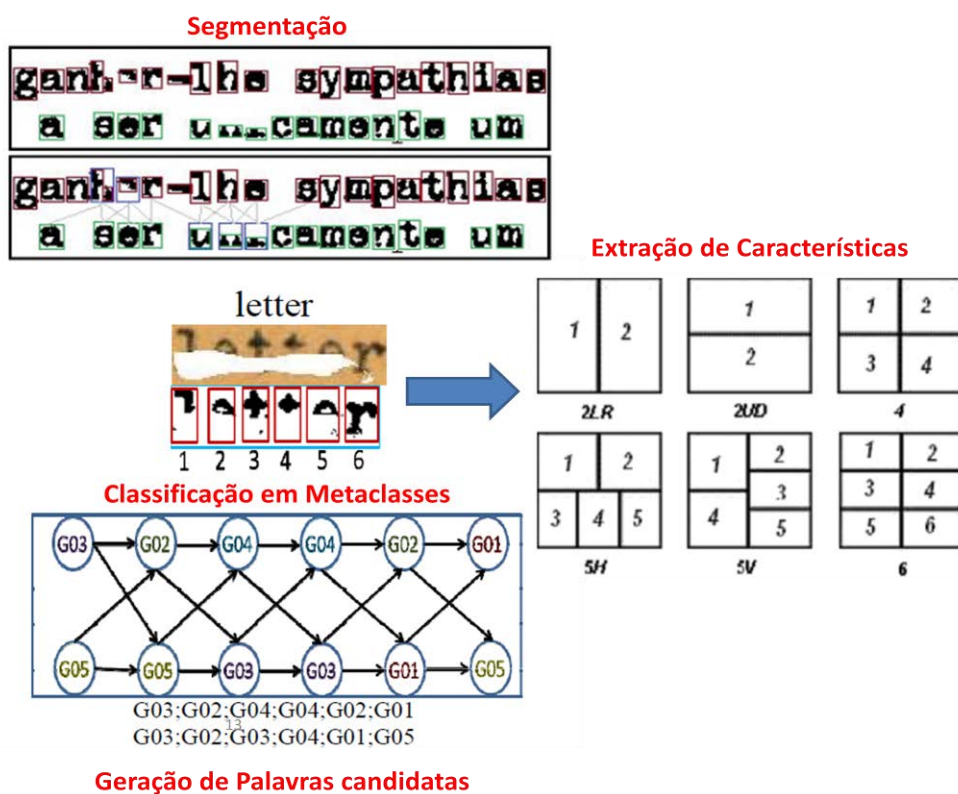
A reconstrução da imagem permite uma melhor transcrição automática pelas ferramentas comerciais de OCR.

A detecção e a caracterização de ruídos apresentadas na seção 2.3 desta tese são utilizadas para identificar áreas com perda de informação (SILVA e LINS, 2011). O processo de segmentação dos componentes textuais é realizado por uma adaptação do

algoritmo proposto por (OLIVEIRA *et al.*, 2011a). Os segmentos de informações remanescentes são divididos em zonas (RADTKE *et al.*, 2003) e submetidos ao processo de extração de características:

- Momentos Geométricos.
- Representação de Forma.
- Massa.
- Inclinação.
- Quantidade de concavidades e *loops*.

Esses caracteres são classificados em metaclasses e é construído um grafo em que são geradas palavras candidatas.



**Figura 4.2** Arquitetura de reconstrução de áreas degradadas

O artigo referente ao método descrito foi publicado na *11th International Conference on Document Analysis and Recognition (ICDAR 2011)* e está disponível no Apêndice A desta tese (SILVA e LINS, 2011).



## 4.2.2 Experimentos e Resultados

O método proposto foi testado em 100 documentos históricos mantidos pela Fundação Joaquim Nabuco (NABUCO). Essas cartas foram transcritas e corrigidas e renderam 13.833 palavras; dessas, 3.814 apresentam problemas de perda de informação. As ferramentas comerciais (ABBYY FineReader Pro 12; NUANCE OmniPage 18; OCRopus 0.3.1) com dicionários de suporte foram utilizadas para transcrição. A contribuição desta tese aumentou o número de palavras corretamente transcritas pelas ferramentas de OCR em 23,7%, 28,2% e 46,8%, respectivamente.

## 4.2.3 Método 2 (Geração automática de conjuntos de treinamento)

O segundo método foi proposto para o reconhecimento de palavras manuscritas cursivas de um mesmo autor. Cada pessoa tem um estilo próprio de escrita, que pode variar de acordo com o estado psicológico, o tipo de documento escrito e até elementos físicos, como a textura do papel e tipo de lápis ou caneta usados. Alguns desses elementos tendem a permanecer inalterados para um mesmo indivíduo. Os artigos (SILVA e LINS, 2012b; SILVA e LINS, 2014) que abordam o problema de transcrição automática estão disponíveis no Apêndice A desta tese.

A ideia é gerar automaticamente palavras sintéticas, a partir de um conjunto de palavras transcritas manualmente, que contêm as características das palavras originais. Essas palavras são então utilizadas como conjunto de treinamento para o modelo de classificação. A sintetização ocorre por meio da aproximação entre as palavras geradas por um *fontset* cursivo e a palavra original. As palavras sintéticas guardam informações sobre as ligações entre os caracteres das palavras. Dessa forma, durante a aproximação, as palavras sintéticas são usadas para extrair características das ligações entre os caracteres. A informação sobre essas ligações é utilizada para gerar novas palavras (Figura 4.3).

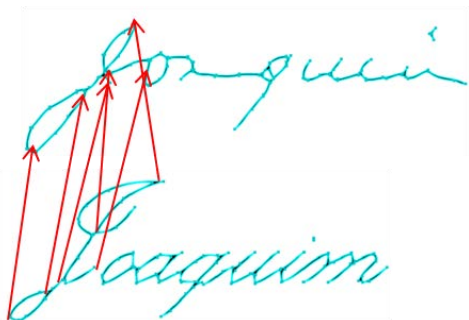
A sintetização das palavras ocorre em quatro etapas:

- "esqueletizar" e vetorizar as imagens: para cada 15 *pixels* consecutivos é gerado um vetor;
- mapear *pixels* de estrutura da palavra *fontset* na palavra original: os *pixels* estruturais são as transições entre as estruturas (*loops*, retas, máximos e mínimos) que compõem os caracteres;
- aproximar os vetores da palavra *fontset* na palavra original;

- extrair informações da ligação entre os *pixels*.



(a) palavra original


(b) palavra *fontset*(c) palavras esqueletizada com *pixels* estruturais


(d) aproximação do primeiro caractere

#### Figura 4.3 Etapas de geração de *fontset*

Um conjunto de características holísticas (LAVRENKO et al., 2004) é extraído das imagens das palavras (Figura 4.4):

- Momentos Geométricos.
- Representação de Forma.
- Massa.
- Inclinação.
- Quantidade de concavidades e *loops*.

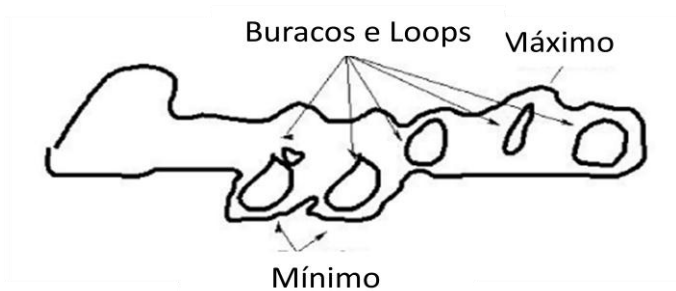


Figura 4.4 Exemplos de características holísticas

#### 4.2.4 Experimentos e Resultados

O método foi testado para a tarefa de transcrição de vocabulário extenso e vocabulário restrito. Para o reconhecimento de vocabulário extenso, foram utilizadas 80 cartas de Joaquim Nabuco, totalizando 4.293 palavras (Figura 4.5). Desse total, foram selecionadas 20 cartas, com obtenção de 1.469 palavras, que foram usadas para geração das palavras sintéticas. O resultado da transcrição das 2.824 palavras pertencentes às 60 cartas não utilizadas na etapa de geração da base sintética apresentou 1.868 palavras corretamente transcritas. Esses resultados mostram que a transcrição de 34% das palavras da base representam 72% de acerto para transcrição automática.

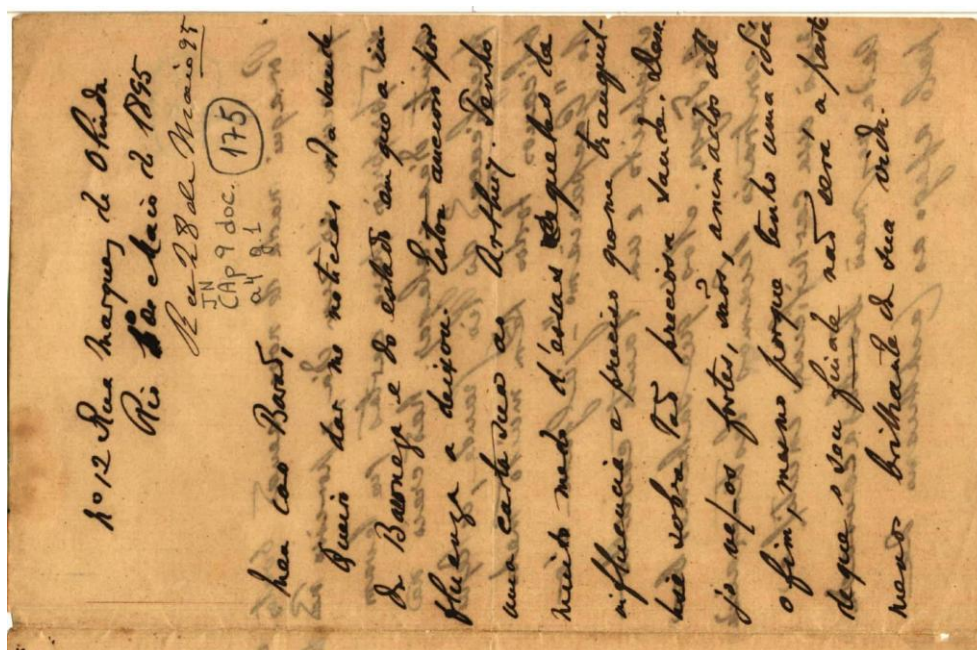
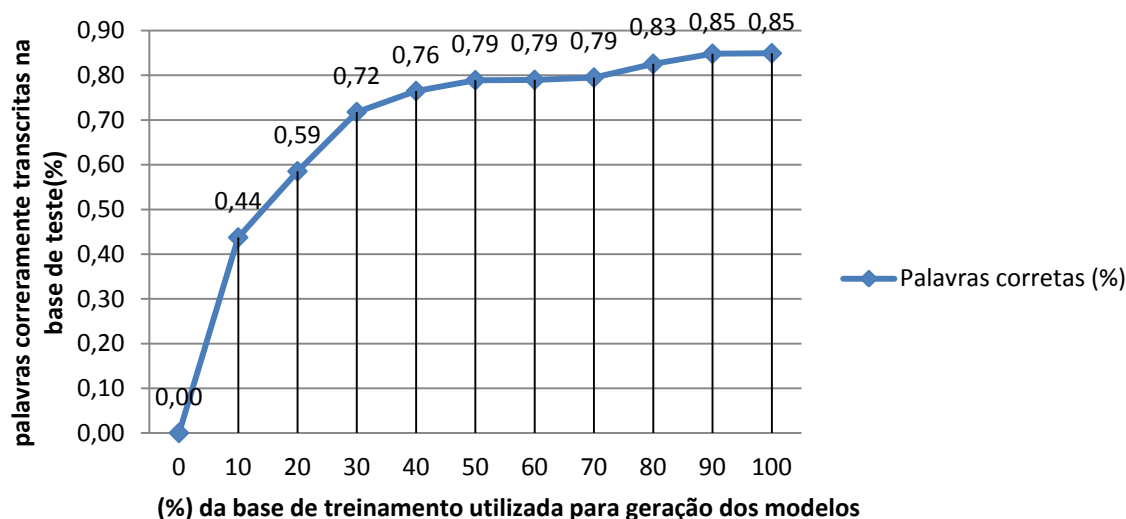


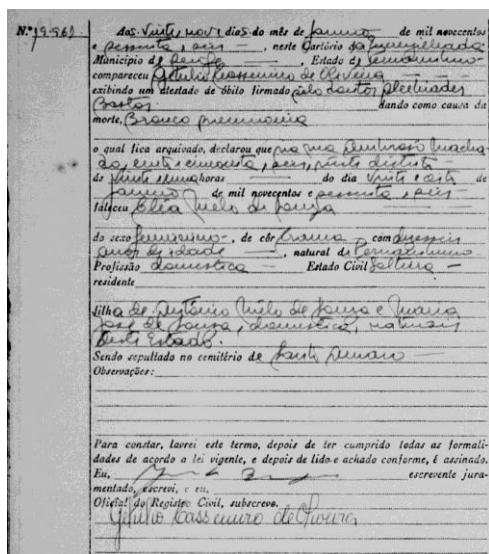
Figura 4.5 Carta manuscrita de Joaquim Nabuco

Outro experimento foi realizado para aferir o comportamento da transcrição em relação ao número de palavras originais usadas para geração do conjunto de treinamento (Figura 4.6). Para esse experimento, as 4.293 palavras foram divididas em dois conjuntos. O primeiro conjunto foi composto com 2.146 palavras e foi utilizado para gerar o *fontset* de treinamento para o modelo de classificação Random Forest (BREIMAN, 2001; BERNARD *et al.*, 2007). O número de palavras usadas no treinamento variou em 10% para cada experimento até atingir as 2.146 palavras que representam o total do conjunto de treinamento. O conjunto de validação usado é composto por 2.147 e é disjunto ao conjunto de treinamento.

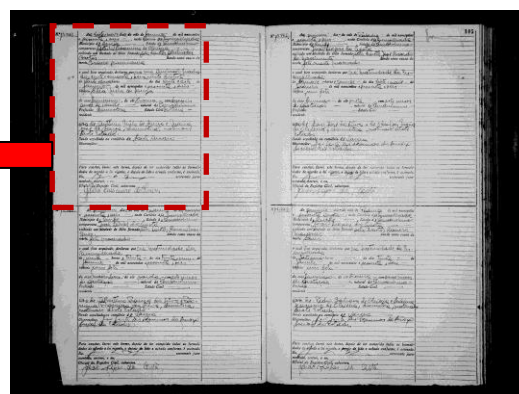


**Figura 4.6 Palavras originais x número de palavras corretamente transcritas**

Para o experimento de transcrição automática de vocabulário restrito, foram usadas as imagens das certidões de óbitários e casamentos do Estado de Pernambuco (TJPE/Family Search).



(a) certidão de óbito



(b) livro do cartório

**Figura 4.7 Exemplo de certidão de óbito da base de dados (TJPE/Family Search).**

Os campos das certidões foram segmentados automaticamente (ALMEIDA, 2011) e 5% das imagens de cada campo foram utilizadas para gerar o *fontset* de treinamento. A partir do subconjunto de palavras segmentadas, foram escolhidos os *fontset* Windows mais próximos das imagens originais. O modelo de classificação foi gerado individualmente para cada campo. O classificador adotado foi o Random Forests (BREIMAN, 2001). Os campos reconhecidos e o resultado da transcrição são apresentados a seguir (Tabela 4.1).

**Tabela 4.1 Resultado da transcrição dos campos dos certificados.**

<b>Campo</b>	<b>Total de campos transcritos corretamente (%)</b>
<b>Nome do Cartório</b>	98,5%
<b>Cidade do Cartório</b>	95,2%
<b>Estado do Cartório</b>	100%
<b>Local da Morte</b>	82,7%
<b>Data de Nascimento</b>	94%
<b>Data da Morte</b>	94%
<b>Estado civil</b>	100%
<b>Cor da pele</b>	100%

Os dados experimentais comprovaram a eficiência do método proposto para vocabulários específicos e abrangentes.

## Capítulo 5

### Conclusões e trabalhos futuros

O processamento de imagens de documentos em larga escala continua a demandar soluções para os problemas de classificação, filtragem e transcrição. O aumento dos esforços para digitalizar grandes coleções de documentos e a facilidade do uso de ferramentas de transcrição automática gerou uma maior heterogeneidade dos documentos em termos de conteúdo, condições de preservação e qualidade de digitalização. Esta tese apresenta contribuições para melhorar e automatizar o processamento de imagens de bases heterogêneas de documentos e para transcrição automática de documentos históricos.

No Capítulo 2, foram apresentados três modelos de classificação (classificação de imagens, classificação de dispositivos e classificação de ruídos), fornecendo dados que permitem melhorar as tarefas de filtragem e impressão dos documentos (SILVA *et al.*, 2010a). O foco da pesquisa foi na extração de características das imagens. O conjunto de características extraído é baseado na paleta da imagem e apresentou um desempenho superior a 92% na tarefa de classificação nos três modelos abordados.

A primeira contribuição apresentada no Capítulo 2 é um modelo de classificação que trata seis classes de imagens (Foto, Logo, Documento, Sintética, Gráfico e Tabelas) e foi incorporado pela Hewlett Packard em suas impressoras, proporcionando uma melhora na qualidade das impressões. Esse modelo possibilita novas pesquisas, como o desenvolvimento de novos algoritmos de filtragem de ruído (CHOWDHURY *et al.*, 2003; PHAM, 2003; STROUTHOPOULOS *et al.*, 2002; ZHU *et al.*, 2006). Já a segunda contribuição é um modelo de classificação do tipo de dispositivo usado na digitalização de documentos. As imagens capturadas por *scanners* e câmeras necessitam de um processamento diferenciado (LINS e SILVA, 2007b), o que permite uma filtragem mais eficiente.

A terceira contribuição do Capítulo 2 é um modelo para classificação em oito classes de ruído (Orientação, Inclinação, Interferência Frente e Verso, Embaçamento, Especular, Sal e Pimenta, Borda e Furos). A classificação permite uma melhor filtragem das imagens (SILVA *et al.*, 2010c). Essa última contribuição é um tema pouco explorado na literatura, mas de grande relevância para o processamento de imagens de documentos. Alguns

problemas relativos à classificação de ruído estão em aberto, como a extensão do classificador para outras classes e a melhor ordem de remoção dos ruídos; são temas fundamentais de pesquisa para a área de Engenharia de Documentos.

No Capítulo 3, novos algoritmos foram propostos para remoção de ruídos em imagens de documentos. O primeiro trata o problema do ruído de borda, o algoritmo faz uso da entropia de Shannon (SHANNON, 1948) e apresenta avanços em relação ao estado da arte ao tratar diferentes tipos de bordas (simples, aderente, texturizada e dupla), diferentes níveis de representação (RGB, *grayscale* e binárias) e imagens capturadas em baixa resolução. O segundo algoritmo remove o ruído de interferência frente e verso em imagens de documento. Esse algoritmo usa a classificação do ruído apresentada no Capítulo 2 desta tese para removê-lo. O terceiro algoritmo foi desenvolvido em parceria com a Hewlett Packard para remoção do ruído especular em imagens. O algoritmo foi incorporado ao *Scanner 3D HP TopShot Laser Printer* e trouxe ganhos em relação à qualidade das imagens capturadas. Diversos outros desafios, como a curvatura da superfície do documento em relação ao plano do *scanner*, sombras e distorções de perspectiva são problemas que requerem novos algoritmos para melhorar a qualidade das imagens.

Ao final do Capítulo 3, foi apresentado um sistema de Raciocínio Baseado em Casos para filtragem automática de grande fluxo de documentos. O SRBC permite o uso eficiente dos algoritmos desenvolvidos nesta tese e o uso de algoritmos desenvolvidos por terceiros. O resultado obtido é superior ao método tradicional de filtragem "cega" em *batch*. A melhoria do SRBC em termos de granularidade e tempo de respostas, juntamente com desenvolvimento de uma medida específica para o problema recuperação de casos, ainda são desafios de pesquisa. O SRBC apresenta indícios da melhor ordem de aplicação de algoritmos em imagens afetadas por diferentes tipos de ruído. Novos trabalhos na área da Engenharia de Documentos são necessários para validar e definir a melhor ordem para remoção de diferentes tipos de ruído em uma mesma imagem (LINS *et al.*, 2010d).

No Capítulo 4, foram apresentados dois métodos para transcrição de imagens de documentos históricos. Esses métodos apresentam ideias inovadoras para o problema de transcrição de documentos impressos e com escrita manual e cursiva. O primeiro método busca reconstruir a informação textual com o uso do *fontset* do documento. Esse método usa informações parciais para gerar possíveis palavras e reconstruir a informação do texto. O resultado da aplicação da reconstrução nas imagens de documento apresenta um ganho de 23,7%, 28,2% e 46,8%%, para as ferramentas de OCR, FineReader Pro 12, NUANCE

OmniPage 18 e OCRopus 0.3.1, respectivamente. O segundo método utiliza *fontset* cursivos do Microsoft Windows como "guia" para aprender o estilo de escrita de um autor. A ideia é gerar um conjunto de treinamento com palavras sintéticas e construir um modelo de classificação baseado no reconhecimento das palavras.

Os métodos de transcrição necessitam de uma melhor validação dos resultados. A validação dos métodos propostos deve ser melhorada com o uso de diferentes acervos históricos. A literatura apresenta diversos estudos com as cartas de George Washington e com os livros históricos da Biblioteca do Congresso Americano, que possuem transcrição manual disponível. Outra fonte de pesquisa é o uso dos métodos propostos em modelos em que não é necessário um processo de extração de características explícita (GRAVES e SCHMIDHUBER, 2009). Nos últimos anos, esses modelos apresentaram bons resultados para a tarefa de transcrição automática.

Por fim, o objetivo de melhorar o processamento de acervos heterogêneos de documentos foi alcançado nesta tese. A melhora do processamento foi comprovada experimentalmente com o uso de um sistema inteligente. O sistema descrito na seção 3.4 proporcionou melhorias significativas (8% para Borda, 27% para Interferência Frente e Verso, 38% para Sal e Pimenta, 12% para Orientação/Inclinação, 33% para Embaçamento) em relação ao processamento tradicional em *batch*. A melhora da transcrição automática de imagens de documentos históricos foi obtida por meio de duas estratégias: reconstrução da informação textual e a geração de conjuntos de treinamento sintéticos.

## 5.1 Trabalhos futuros

Diversas melhorias podem ser aplicadas na pesquisa apresentada nesta tese. Com relação à classificação, sugere-se como trabalho futuro o refinamento da classificação e testes com outras classes de imagens e ruídos.

A melhora do SRBC em termos de granularidade e tempo de respostas, juntamente com desenvolvimento de uma medida específica para a recuperação de casos também são temas interessantes de pesquisa. O SRBC apresentou indícios da melhor ordem de aplicação de algoritmos em imagens com múltiplos ruídos. A análise desses indícios e validação são importantes para a melhora da filtragem de ruídos.



Os métodos de transcrição serão testados em outras bases de dados para uma melhor validação dos resultados, como no acervo de cartas de George Washington e os livros históricos da Biblioteca do Congresso Americano (Figura 5.1).

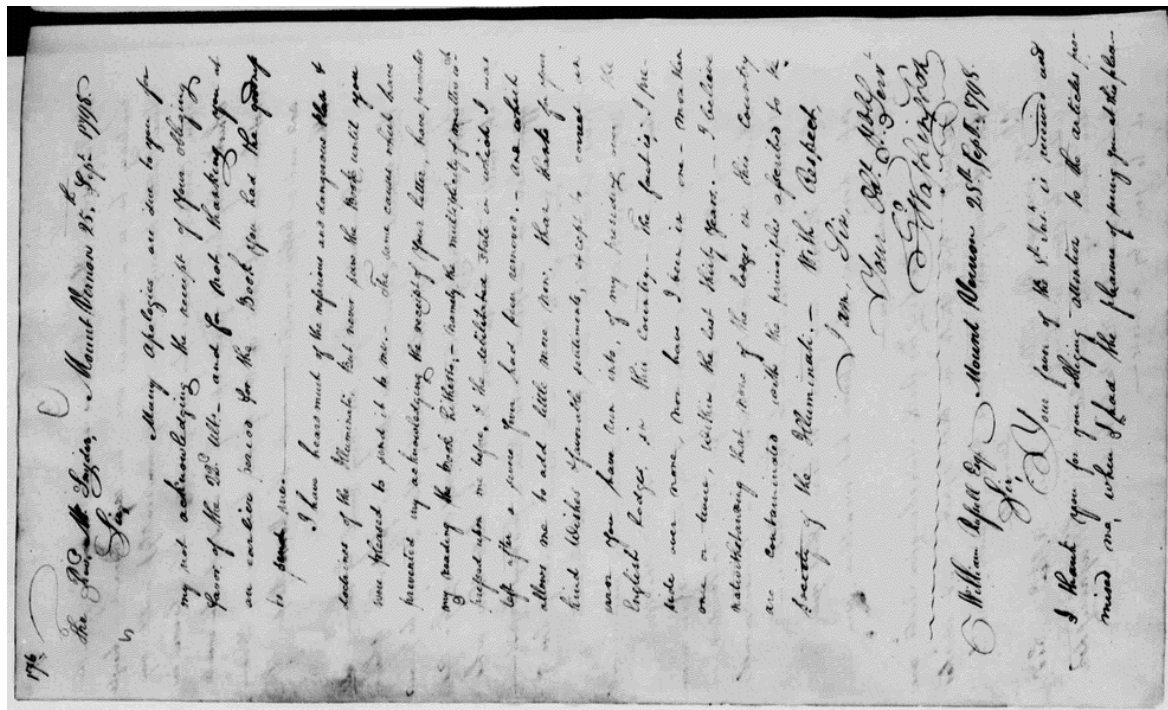


Figura 5.1 Carta de George Washington.

Por fim, pesquisas nas áreas de Processamento de Linguagem Natural (PLN) serão temas de trabalhos futuros. O uso das técnicas de PLN podem melhorar a reconstrução de documentos degradados e as ferramentas de OCR. Nesse mesmo contexto, a compressão (sumarização de texto) e recuperação de informação serão abordados para organizar o acesso aos documentos após a transcrição automática.

## REFERÊNCIAS

(AAMODT e PLAZA, 1994) A. Aamodt, E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. In: AI Communications, vol.7, pp: 39-59.

(AGRAWAL e DOERMANN, 2011) M. Agrawal and D. Doermann. Stroke-like Pattern Noise Removal in Binary Document Images. In: International Conference on Document Analysis and Recognition, pp: 17-21.

(ALI, 1996) M.B.H. Ali. Background noise detection and cleaning in document images. In: International Conference on Pattern and Recognition, vol. 3, pp: 39-59.

(ALMEIDA, 2011) A. B. S. Almeida. Pré-Processamento de Imagens na Plataforma Thanatos. In: Dissertação de Mestrado do Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) da Universidade Federal de Pernambuco. Data da Defesa: 31 de outubro de 2011.

(AHA e KIBLER, 1991) D. Aha and D. Kibler (1991). Instance-based learning algorithms. In: Machine Learning, vol.6, pp: 37-66.

(ÁVILA e LINS, 2004) B. T. Ávila and R. D. Lins. A New Algorithm for Removing Noisy Borders from Monochromatic Documents. In: ACM Symposium on Applied Computing, vol.2, pp:1219-1225.

(ÁVILA *et al.*, 2011) S. Ávila, N. Thome, M. Cord, E. Valle, and A. A. Araújo. Extended Bag-of-Words Formalism for Image Classification. In: 18th IEEE International Conference on Image Processing, vol.2, pp:2909-2912.

(ARTHUR e VASSILVITSKII, 2007) D. Arthur, S. Vassilvitskii. k-means++: the advantages of carefull seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp:1027-1035.

(BAYRAM *et al.*, 2005) S. Bayram, H. Sencar, N. Memon and I. Avcibas. Source camera identification based on CFA interpolation. In: IEEE International Conference on Image Processing, vol.3, pp: 69-72.

- (BAYRAM *et al.*, 2006) S. Bayram, H. Sencar and N. Memon. Improvements on Source Camera-Model Identification Based on CFA Interpolation. In: International Conference on Digital Forensics, pp: 24-27.
- (BARNARD e FORSYTH, 2001) K. Barnard and D. Forsyth. Learning the Semantics of Words and Pictures. In: International Conf. Computer Vision, pp: 1-8.
- (BERNARD *et al.*, 2007) S. Bernard, L. Heutte, S. Adam. Using Random Forests for Handwritten Digit Recognition. In: Ninth International Conference on Document Analysis and Recognition, pp:1520-5363.
- (BEUSEKOM *et al.*, 2007) J. van Beusekom, F. Shafait and T. M. Breuel. Image-Matching for Revision Detection in Printed Historical Documents. In: Pattern Recognition Lecture Notes in Computer Science, vol.4713, pp: 507-516.
- (BEZERRA *et al.*, 2012) B. L. D. Bezerra, C. Zanchettin, and V. B Andrade. A MDRNN-SVM Hybrid Model for Cursive Offline Handwriting Recognition. In: International Conference on Artificial Neural Networks, vol.2, pp: 246-254.
- (BLAKE E BRELSTAFF, 1988) A. Blake and G. Brelstaff. Geometry from specularities. In: International Conference on Computer Vision, pp: 394-403.
- (BREIMAN, 2001) L. Breiman. Random Forests. In: Machine Learning, vol.45, pp: 5-32.
- (BROCKETT E MARAGOS, 1994) R. Brockett and P. Maragos. Evolution equations for continuous scale morphology. In: IEEE Trans. on Signal. Processing, vol.42, pp: 3377-3386.
- (BUKHARI *et al.*, 2009) S. S. Bukhari, F. Shafait e T. M. Breuel. Textline information extraction from grayscale camera-captured document images. In: 13th International Conference on Image Processing, pp: 1522-4880.
- (BUKHARI *et al.*, 2012) S. S. Bukhari, F. Shafaity e T. M. Breuel. Border Noise Removal of Camera-Captured Document Images using Page Frame Detection. In: Lecture Notes in Computer Science, vol.7139, pp: 126-137.
- (BURGOYNE *et al.*, 2008) J. A. Burgoyne, J. Devaney, L. Pugin, e I. Fujinaga. Enhanced bleedthrough correction for early music documents with recto-verso registration. In: International Conference Music Information Retrieval, pp: 407-412.

- (CANNY, 1986) J. Canny. A Computational Approach to Edge Detection. In: IEEE Trans. Pattern Analysis and Machine Intelligence, vol.8, pp: 679-698.
- (CAO *et al.*, 2001) R. Cao, C. L. Tan e P. Shen, A wavelet approach to double-sided document image pair processing. In: Proc. International. Conference in Image Processing, pp: 174-177.
- (CASTRO e PINTO, 2007) P. Castro, J. R. C. Pinto. Methods for Written Ancient Music Restoration. In: Image Analysis and Recognition, pp: 1194-1205.
- (CASTRO *et al.*, 2007) P. Castro, R. J. Almeida e J. R. C. Pinto. Restoration of double-sided ancient music documents with bleed-through. In: Progress in Pattern Recognition, Image Analysis and Applications, pp: 940-949.
- (CELIKTUTAN *et al.*, 2008) O. Celiktutan, I.Avcibas e B. Sankur. Blind identification of source cell-phone model. In: IEEE Transactions on Information Forensics and Security, vol.3, pp: 553-566.
- (CHAN *et al.*, 2005) R. H. Chan, C-W Ho, M. Nikolova, Salt-and-pepper Noise Removal by Median-type Noise Detectors and Detail-preserving Regularization. In: IEEE Transactions Image Process, vol.14, pp: 1479-1485.
- (CHEN *et al.*, 2008) M. Chen, J. Fridrich, M. Goljan, J. Lukas(2008). Determining Image Origin and Integrity Using Sensor Noise. In: IEEE Transactions on Information Forensics and Security. vol.3, pp: 74-90.
- (CHINNASARN *et al.*, 1998) K. Chinnasarn, Y. Rangsanseri, and P. Thitimajshima. Removing salt-and-pepper noise in text/graphics images. In: Asia-Pacific Conference on Circuits and Systems, pp: 459-462.
- (CHOWDHURY *et al.*, 2003) S. P. Chowdhury, S. Mandal, A. K. Das, and Bhabatosh Chanda. Automated segmentation of math-zones from document images. In: International Conf. on Document Analysis and Recognition, pp: 755-759.
- (COHEN, 2007) K. Cohen. Digital Still Camera Forensics. In: Small scale digital device forensics journal, vol.1, pp: 1-8.
- (COLLINS *et al.*, 2000) R. Collins, A. Lipton, T. Kanade, H. Fujuyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa. A System for Video Surveillance and

Monitoring. In: Technical Report CMU-RITR-00-12, Carnegie Mellon University, Pittsburgh.

(CSURKA *et al.*, 2004) G. Csurka, C. R. Dance, L. Fan , J. Willamowski , C. Bray (2004). Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, vol.1, pp: 1-16.

(CUMANI, 1991) A. Cumani. Edge detection in multispectral images. G. Models and Image Processing, vol.53, pp: 40-51.

(DRIRA *et al.*, 2007) F. Drira, F. Lebourgeois, H. Emptoz. OCR accuracy improvement through a PDE-based approach. In: International Conference on Document Analysis and Recognition, pp: 1068-1072.

(DIRIK *et al.*, 2007) A.E. Dirik, S. Bayram, H.T. Sencar, N. Memon (2007). New Features to Identify Computer Generated Images. In: IEEE International Conference on Image Processing, vol.4, pp: 433-436.

(DONG *et al.*, 2007) Y. Dong, R. H. Chan, and S. Xu. A Detection Statistic for random-valued impulse noise. In: IEEE Trans. Image Processing, vol.16 , pp: 1112-1120.

(EL ABED *et al.*, 2009) H. El Abed, V. Margner, M. Kherallah, A.M. Alimi. Handwriting Recognition Competition. In: International Conference on Document Analysis and Recognition, pp: 1388-1392.

(FAN *et al.*, 2001) K. C. Fan, Y. KaiWang, and T. R. Lay. Marginal noise removal of document images. In: International Conference on Document Analysis and Recognition pp: 317-321.

(FENG *et al.*, 2006) S. Feng, R. Manmatha, and A. McCallum. Exploring the use of conditional random field models and HMMs for historical handwritten document recognition. In: International Conference on Document Image Analysis for Libraries, pp: 30-37.

(FRIGUI e KRISHNAPURAM, 2001) H.Frigui and R.Krishnapuram. Clustering by competitive agglomeration. In: Pattern Recognition, vol.30, pp: 1109-1119.

(GANGAMMA *et al.*, 2012) B. Gangamma, M. K. Srikanta, A. V. Singh. Restoration of Degraded Historical Document Image. In: Journal of Emerging Trends in Computing and Information Sciences, vol.3, pp: 792-798.

- (GATOS *et al.*, 2004) B. Gatos, I. Pratikakis e S.J. Perantonis. An Adaptive Binarization Technique for Low Quality Historical Documents. In: International Workshop on Document Analysis Systems, pp: 102-113.
- (GATOS *et al.*, 2006) B. Gatos, I. Pratikakis e S.J. Perantonis. Adaptive degraded document image binarization. Pattern Recognition, vol.39, pp: 317-327.
- (GATOS *et al.*, 2014) B. Gatos, N. Stamatopoulos, G. Louloudis, S. Perantonis. H-DocPro: A Document Image Processing Platform for Historical Documents. In: Digital Access to Textual Cultural Heritage, vol.1, pp: 131-136.
- (GIANNETTI *et al.*, 2010) F. Giannetti, G. Dispoto, R. D. Lins, G. F. P. Silva and A. Cabeda. PDF Profiling for B&W versus Color Pages Cost Estimation for Efficient On-Demand Book Printing. In: ACM symposium on Document engineering, pp: 177-180.
- (GOOGLE BOOK, 2014) Google book search. <http://books.google.com/>, acessado em 05/07/2014.
- (GOOGLE OCR, 2014) <https://support.google.com/drive/answer/176692?hl=pt-BR>, acessado em 05/07/2014.
- (GONZALEZ e WOODS, 2008) R. C. Gonzalez e R. E. Woods. Digital Image Processing. 3rd ed. Prentice Hall, 2008.
- (GOU *et al.*, 2007a) Gou, H., Swaminathan, A., & Wu, M. Robust Scanner Identification Based on Noise Features. In: Conference on Security, Steganography, and Watermarking of Multimedia Contents, vol.6505, pp: 1-8.
- (GOU *et al.*, 2007b) H. Gou, S. Swaminathan and M. Wu. Noise Features for Image Tampering Detection and Steganalysis. In: IEEE International Conference on Image Processing, pp: 97-100.
- (GRAY e COK, 1997) R.T. Gray and D.R. Cok, "Adjusting Film Grain Properties in Digital Images," US patent number: 5641596, 1997.
- (GRAVES *et al.*, 2006) A. Graves, S. Fernández, F. Gomez, J. Schmidhuber. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In: International Conference on Machine Learning, pp: 369-376.

- (GRAVES *et al.*, 2007) A. Graves, S. Fernandez, J. Schmidhuber. Multidimensional Recurrent Neural Networks. In: International Conference on Artificial Neural Networks, pp: 549-558.
- (GRAVES e SCHMIDHUBER, 2009) A. Graves e J. Schmidhuber: Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. In: Advancer in Neural Information Processing System, pp: 545-552.
- (HAMMING, 1950) R. W. Hamming. Error detecting and error correcting codes. In: Bell System Technical Journal, vol.29, pp: 147–160.
- (HAR-PELED, 2011) S. Har-Peled. Data structures for geometric approximation. In: American Mathematical Society.
- (HAYKIN, 2002) S. Haykin. Redes Neurais - Princípios e Praticas. Editora: BOOKMAN 2ed, 2002.
- (HEARST e PEDERSEN, 1996) M.A.Hearst and J.O.Pedersen. Reexamining the Cluster Hypotesis: Scatter Gathet on Retrieval Results. In: Special Interest Group On Information Retrieval, pp: 1-10.
- (HSU e LIN, 2002) C. W. Hsu, C. J. Lin, A comparison of methods for multiclass Support Vector Machines. In: IEEE Trans on Neural Networks, v.13, pp: 415-425.
- (HUANG *et al.*, 2014) Y. Huang, Z. Wu, L. Wang e T. Tan. Feature Coding in Image Classification: A Comprehensive Study. In: IEEE Transactions on Pattern Analysis and Machine Intelligence vol.36, pp: 493-506.
- (IMPACT, 2014) Impact. <http://www.digitisation.eu/>, acessado em 05/07/2014.
- (INDERMUHLE *et al.*, 2008) E. Indermuhle, M. Liwicki and H. Bunke1. Recognition of Handwritten Historical Documents: HMM-Adaptation vs. Writer Specific Training. In: Conference on Frontiers in Handwriting Recognition, pp: 186-191.
- (JAIN *et al.*, 2004) A.K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. In: IEEE Trans. Circuits and Systems for Video Technology, vol.14, pp: 4-20.
- (KAPUR *et al.*, 1985) J. N. Kapur, P. K. Sahoo and A. K. C. Wong. A New Method for Gray-Level Picture Thresholding using the Entropy of the Histogram. In: Vision Graphics and Image Processing, vol.29, pp: 1-10.

- (KAVALLIERATOU e ANTONOPOULOU, 2005) E. Kavallieratou and H. Antonopoulou. Cleaning and Enhancing Historical Document Images. In: Intelligent Vision Systems, vol.3708, pp: 681-688.
- (KB DIGITIZATION, 2014) Kb digitization. <http://www.kb.nl/hrd/digitalisering/index-en.html>, acessado em 05/07/2014.
- (KORNAI *et al.*, 1996) A. Kornai, K. M. Mohiuddin and S. D. Connell. Recognition of Cursive Writing on Personal Checks. In: International Workshop on Frontiers in Handwriting Recognition, pp: 1-10.
- (KOSALA e BLOCKEEL, 2000) R. Kosala and H. Blockeel. Web Mining Research: A Survey. In: ACM SIGKDD Explorations Newsletter, vol.2, pp. 1-15.
- (KRISHNAMACHARI e MOTTALEB, 1999) S. Krishnamachari and M. Abdel-Mottaleb. Image Browsing using Hierarchical Clustering. In: IEEE Symposium on Computers and Communications, pp:12-20.
- (KASTURI, 2002) R. Kasturi, L. O’Gorman and V. Govindaraju. Document image analysis: A primer. In: Sadhana, vol.27, pp: 3-22.
- (KOHAVI, 1995) R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence, v.14, pp: 1137-1145.
- (LAVRENKO *et al.*, 2004) V. Lavrenko, T. Rath, and R. Manmatha. Holistic word recognition for handwritten historical documents. In: International Workshop on Document Image Analysis for Libraries, pp: 278-287.
- (LAZZARA *et al.*, 2011) G. Lazzara, R. Levillain, T. Géraud, Y. Jacquélet, J. Marquegnies, and A. Crepin-Leblond. The SCRIBO Module of the Olena Platform: A Free Software Framework for Document Image Analysis. In: International Conference Document Image Analysis and Recognition, pp: 252-258.
- (LEE, 1986) H. S. Lee. Method for computing the scene-illuminant chromaticity from specular highlights. In: Journal of the Optical Society of America A, vol.3 pp: 1694-1699.
- (LEEDHAM *et al.*, 2002) G. Leedham, S. Varma, A. Patankar, V. Govindaraju. Separating text and background in degraded document images-a comparison of global



thresholding techniques for multi-stage thresholding. In: International Workshop on Frontiers in Handwritten Recognition, pp: 244-249.

(LI e FAN, 2009) F. Li e J. Fan (2009). Salt and Pepper Noise Removal by Adaptive Median Filter and Minimal Surface Inpainting. In: International Congress on Image and Signal Processing, vol.1, pp: 1-5.

(LIKFORMAN-SULEMA *et al.*, 2011) L. Likforman-Sulema, J. Darbonb, E. H. B. Smithd. Enhancement of historical printed document images by combining Total Variation regularization and Non-local Means filtering. *Image and Vision Computing*, vol.29, pp: 351-363.

(LINS *et al.*, 1995) R. D. Lins, M. G. Neto, L. F. Neto e L. G. Rosa. An Environment for Processing Images of Historical Documents. In: *Microproc. & Microprogramming*, pp: 111-121.

(LINS *et al.*, 2006) R. D. Lins, Ávila B. T., A. F. Araújo. BigBatch An Environment for Processing Monochromatic Documents. In: *Lecture Notes in Computer Science*, v. 4142, pp: 886-896.

(LINS e MACHADO, 2004) R. D. Lins e D.S.A. Machado. A Comparative Study of File Formats for Image Storage and Transmission. In: *Journal of Electronic Imaging*, vol.13, pp: 175-183.

(LINS e SILVA, 2007) R. D. Lins and J. M. M. da Silva, A Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents, In: *ACM Symposium on Applied Computing*, pp: 610-616.

(LINS e SILVA, 2007b) LINS, R. D. ; MIRO, B. ; SILVA, Gabriel de França Pereira e . An OCR Assessment of the Quality of Document Images Acquired with Portable Digital Cameras. In: *International Workshop on Camera-Based Document Analysis and Recognition*, vol.1, pp: 106-111.

(LINS *et al.*, 2007a) R. D. Lins, J. M. Silva, G. F. P. Silva. A quantitative method for assessing algorithms to remove back-to-front interference in documents. In: *ACM Symposium On Applied Computing*, pp: 610-616.

(LINS *et al.*, 2007b) R. D. Lins, A. R. G. Silva e G. F. P. Silva. Enhancing Document Images Acquired Using Portable Digital Cameras. In: Lecture Notes in Computer Science, pp: 1229-1241.

(LINS, 2009) R. D. Lins. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. In: Lecture Notes in Computer Science, vol.5627, pp: 844-854.

(LINS *et al.*, 2009) R. D. Lins, G. F. P. Silva, S. Simske, J. Fan, M. Shaw, P. Sá, M. Thielo. Image Classification to Improve Printing Quality of Mixed-Type Documents. In: International Conference on Image Analysis and Recognition, pp: 1106-1110.

(LINS *et al.*, 2010a) R. D. Lins, D. M. Oliveira, G. Torreão, J. Fan e M. Thielo. A Dewarping Algorithm to Compensate Volume Binding Distortion in Scanned Documents. In: ACM Symposium On Applied Computing, pp: 61-62.

(LINS *et al.*, 2010b) R. D. Lins, D. M. Oliveira, G. Torreão, J. Fan e M. Thielo. Correcting Book Binding Distortion in Scanned Documents. In: Lecture Notes in Computer Science, vol.6112, pp: 355-365.

(LINS *et al.*, 2010d) R. D Lins, G. F. P. Silva, S. Banergee, A. Kuchibhotla e M. Thielo. Automatically Detecting and Classifying Noises in Document Images. In: ACM Symposium On Applied Computing, vol.1, pp: 33-39.

(LINS *et al.*, 2011a) R. D. Lins; G. F. P. Silva; J. S. Simske. Automatically Discriminating between Digital and Scanned Photographs. In: International Conference on Image Analysis and Recognition, vol.1, pp: 1280-1284.

(LINS *et al.*, 2011b) - Lins, R. D.; Silva, G. F. P; Formiga, A. A. HistDoc v. 2.0: enhancing a platform to process historical documents. In: Workshop on Historical Document Imaging and Processing, 2011, Beijing. HIP '11 Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. New York: ACM, 2011. v. 1. p. 169-176.

(LINS *et al.*, 2013) R.D. Lins, G. F. P. Silva, E. Mariano, F. Fan, P. Majewicz and M. Thielo . Removing Shade and Specular Noise in Images of Objects and Documents Acquired with a 3D-Scanner. In: Lecture Notes in Computer Science, vol.1, pp: 299-307.

- (LU *et al.*, 1999) Z. Lu, R. M. Schwartz, P. Natarajan, I. Bazzi, and J. Makhoul. Advances in the BBN BYBLOS OCR System. In: International Conference on Image Analysis and Recognition, pp: 337-340.
- (LYU e FARID, 2005) S. Lyu, & H Farid. How realistic is photorealistic. In: IEEE Transactions on Signal Processing, vol.53, pp: 845-850.
- (MADHVANATH e GOVINDARAJU, 2001) S. Madhvanath and V. Govindaraju: The Role of Holistic Paradigms in Handwritten Word Recognition. In: Transactions on Pattern Analysis and Machine Intelligence, vol.23, pp: 149-164.
- (MAKHOUL *et al.*, 2010) J. Makhoul, F. Kubala, R. Schwartz e R. Weischedel. Performance measures for information extraction. In: Workshop on Broadcast News Understanding, pp: 249-252.
- (MARIANO *et al.*, 2011) E. Mariano, R. D. lins, G. F. P. Silva and J. Fan. Correcting Specular Noise in Multiple Images of Photographed Documents. In: International Conference on Image Analysis and Recognition, pp: 915-919.
- (MARTI e BUNKE, 2002) U.-V. Marti and H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition systems. Hidden Markov models. In: World Scientific Publishing Co, pp: 65-90.
- (MCKAY *et al.*, 2008) C. McKay, A. Swaminathan, G. Hongmei, Min Wu. Image acquisition forensics: Forensic analysis to identify imaging source. In: IEEE International Conference on Acoustics, Speech and Signal Processing: vol.48, pp: 1657-1660.
- (MALLICK *et al.*, 2006) S. P. Mallick, T. Zickler, P. N. Belhumeur e D. J. Kriegman (2006). Specularity Removal in Images and Videos: A PDE Approach. In: Lecture Notes in Computer Science, vol.3951, pp: 550-563.
- (MEHDI *et al*, 2004) Mehdi, K.L., Sencar, H.T., & Memon, N. Blind source camera identification. In: International Conference on Image Processing, vol.1, pp: 709-712.
- (MILLION BOOK, 2014) Million book project. <http://www.archive.org/details/millionbooks>, acessado em 05/07/2014.
- (MOGHADDAM e CHERIET, 2009) R. F. Moghaddam and M. Cheriet. Low quality document image modeling and enhancement. In: International Journal on Document Analysis and Recognition, vol.11, pp:183-201.

- (MOGHADDAM e CHERIET, 2011) R. F. Moghaddam, M. Cheriet. Beyond *pixels* and regions: A non-local patch means (NLPM) method for content-level restoration, enhancement, and reconstruction of degraded document images. In: Pattern Recognition vol.44, pp: 363-374.
- (NAYAR *et al.*, 1997) S. Nayar, X. Fang, and T. Boult. Separation of reflection components using color and polarization. In: International Journal of Computer Vision, vol.21, pp: 163-186.
- (NEELAMANI *et al.*, 1989) R. Neelamani, H. Choi e R. G. Baraniuk. Wavelet-based deconvolution for ill-conditioned systems. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol.6, pp: 3241-3244.
- (NISHIDA E SUZUKI, 2003) H. Nishida and T. Suzuki. A Multiscale Approach to Restoring Scanned Color Document Images with Show-trough Effects. In: International Conference on Document Analysis and Recognition, pp: 584-588.
- (OHA *et al.*, 2005) Hyun-Hwa Oha, Kil-Taek Limb and Sung-Il Chienc. An improved binarization algorithm based on a waterflow model for document image with inhomogeneous backgrounds. In: Pattern Recognition vol.38, pp: 2612-2625.
- (OLIVEIRA *et al.*, 2013) D. M. Oliveira, R. D. Lins, G. F. P. Silva, J. Fan, M. Thielo. De-blurring Textual Document Images. Lecture Notes in Computer Science. 1ed.: vol. 7423, pp: 238-250.
- (OSADCHY *et al.*, 2003) M. Osadchy, D. Jacobs, and R. Ramamoorthi. Using specularities for recognition. In: International Conference on Computer Vision, pp: 1512-1519.
- (OTSU, 1979) N. Otsu. A threshold selection method from gray level histograms. In: IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, pp: 62-66.
- (PARK *et al.*, 2002) G.Park, Y.Baek and L.Heung-Kyu. A Ranking Algorithm Using Dynamic Clustering for Content-Based Image Retrieval. In: Lecture Notes in Computer Science, vol.2383, pp: 328-337.
- (PEERAWIT e KAWTRAKUL, 2004) W. Peerawit and A. Kawtrakul. Marginal Noise Removal from Document Images Using Edge Density. In: Proceedings of Fourth Information and Computer Eng. Postgraduate Workshop, pp: 1-23.

- (PING *et al.*, 2003) T. Ping, S. Lin, L. Quan, and H.-Y. Shum. Highlight removal by illumination-constrained inpainting. In: International Conference on Computer Vision, pp: 164-169.
- (PHAM, 2003) T. D. Pham. Unconstrained logo detection in document images. In: Pattern Recognition, vol.36, pp: 3023-3025.
- (PLAMONDON e SRIHARI , 2000) R. Plamondon and S. N. Srihari: On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. In: IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.22, pp: 63-84.
- (RABINER, 1989) L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, vol.77, pp: 257–286.
- (RADTKE *et al.*, 2003)P. V. W. Radtke, L.S. Oliveira, R. Sabourin e T. Wong. Intelligent Zoning Design Using Multi-Objective Evolutionary Algorithms. In: International Conference on Document Analysis and Recognition, pp: 824-828.
- (RAMSAK, 2002) F. Ramsak. Towards a general-purpose, multidimensional index: Integration, Optimization e Enhancement of UB-Trees. Munich University of Technology. PhD Thesis. Munique, 2002.
- (RAY *et al.*, 2013) A. Ray, A. Chandawala and S. Chaudhary. Character Recognition using Conditional Random Field based Matching Engine. In: International Conference on Document Analysis and Recognition, pp: 18-22.
- (RATH e MANMATHA, 2003) T. M. Rath and R. Manmatha: Word Image Matching Using Dynamic Time Warping. In: Conference on Computer Vision and Pattern Recognition, vol.2, pp: 521-527.
- (RUGNA e KONIK, 2003) J. Da Rugna and H. Konik. Automatic blur detection for metadata extraction in content-based retrieval context. In: SPIE Proceedings, vol.5304, pp: 285.294.
- (SAUVOLA e PIETIKAINEN, 2000) J. Sauvola e M. Pietikainen. Adaptive document image binarization. In: Pattern Recognition, vol.33, pp: 225-236.
- (SCHEUNDERS, 1997) P. Scheunders. Comparison of Clustering Algorithms Applied to Color Image Quantization. In: Pattern Recognition Letters, vol.18, pp: 1379-1384.

(SHAFAIT e BREUEL, 2007) F. Shafait and T. M. Breuel. Document Image Dewarping Contest. International Workshop on Camera-Based Document Analysis and Recognition, pp: 181-188.

(SHAFAIT *et al.*, 2008c) F. Shafait, D. Keysers, and T. M. Breuel. Efficient implementation of local adaptive thresholding techniques using integral images. In: Document Recognition and Retrieval, pp. 101–106.

(SHAFER, 1985) S. Shafer. Using color to separate reflection components. In: COLOR research and applications, vol.10, pp: 210-218.

(SHANNON, 1948) C. Shannon, A mathematical theory of communication. In: Bell System Technology Journal, vol.27, pp: 370-423 and 623-656.

(SHARMA, 2001) G. Sharma. Show-trough cancellation in scans of duplex printed documents. In: IEEE Transaction Image Processing, vol.10, pp: 736-754.

(SHIRAI *et al.*, 2013) K. Shirai, Y. Endo, A. Kitadai, S. Inoue, N. Kurushima, H. Baba, A. Watanabe, M. Nakagawa. Character Shape Restoration of Binarized Historical Documents by Smoothing via Geodesic Morphology. In: International Conference on Document Analysis and Recognition, pp: 1285-1289.

(SILVA e LINS, 2005) A. R. G. Silva e R. D. Lins. Background Removal of Document Images Acquired Using Portable Digital Cameras. In: Lecture Notes in Computer Science, vol.3656, pp: 278-285.

(SILVA, 2006) A. R. G. Silva. Análise e Melhoria da Qualidade de Documentos Fotografados. Dissertação de Mestrado. Universidade Federal de Pernambuco, Departamento de Eletrônica e Sistemas, Pernambuco, Recife, 2006.

(SILVA e LINS, 2007) G. F. P. Silva e R. D. Lins. PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. In: Camera-Based Document Analysis and Recognition, pp: 107-115.

(SILVA *et al.*, 2007) J. M. Silva, R. D. Lins and G. F. P. Silva. A Fast Algorithm to Binarize and Filter Documents with Back-to-Front Interference. In: ACM Symposium On Applied Computing, pp: 639-640.

(SILVA *et al.*, 2009a) Silva, G. F. P.; Lins, R. D.; Miro B.; Simske, S.; Thielo, M. Automatically Deciding if a Document was Scanned or Photographed. In: Journal of Universal Computer Science, 2009, v.15, pp: 3364-3366.

(SILVA *et al.*, 2009b) J. M. Silva, R. D. Lins and G. F. P. Silva. Enhancing the Quality of Color Documents with Back-to-Front Interference. Image Analysis and Recognition, pp: 875-885.

(SILVA *et al.*, 2010a) G. F. P. Silva, R. D. Lins e J. M. M. da Silva. HistDoc - A Toolbox for Processing Images of Historical Documents. In International Conference on Image Analysis and Recognition, pp. 409-419.

(SILVA *et al.*, 2010b) G. F. P. Silva, R. D. Lins, S. Banerjee, A. Kuchibhotla, M. Thielo. Automatically Detecting and Classifying Noises in Document Images. In: Symposium On Applied Computing, vol.1, pp: 33-39.

(SILVA *et al.*, 2010c) G. F. P. Silva, R. D. Lins, S. Banerjee, A. Kuchibhotla, M. Thielo. Enhancing the Filtering-out of the Back-to-Front Interference in Color Documents with a Neural Classifier. In: International Conference on Pattern Recognition, pp: 2415-2419.

(SILVA e LINS, 2011) G. F. P. Silva e R. D. Lins. An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. In: International Conference on Document Analysis and Recognition, vol.1, pp: 553-557.

(SILVA e LINS, 2012a) G. F. P. Silva e R. D. Lins. Automatic Content Recognition of Teaching Boards in the Tableau Platform. In: International Conference on Pattern Recognition, vol.1, pp: 1-5.

(SILVA e LINS, 2012b) G. F. P. Silva; R. D. Lins. Generating Training Sets for the Automatic Recognition of Handwritten Documents. In: Advances in Character Recognition. 1ed. New York: InTech, 2012, vol.1, pp: 155-174.

(SILVA e LINS, 2014) G. F. P. Silva e R. D. Lins. Automatic Training Set Generation for Better Historic Document Transcription and Compression. In: International Workshop on Document Analysis Systems, vol.1. pp: 20-31.

(SILVA *et al.*, 2013) G. F. P. Silva ; R. D. Lins ; A. R. Silva. A New Algorithm for Background Removal of Document Images Acquired Using Portable Digital Cameras. In: Lecture Notes in Computer Science, vol.1, pp: 290-298.

(SIMSKE, 2005) S.J. Simske. Low-resolution photo/drawing classification: metrics, method and archiving optimization. In: IEEE International Conference on Image Processing, pp: 534-537.

(SNYDER *et al.*, 1999) P.D. Snyder, T.F. Pawlicki, and R.S. Gaborski. Apparatus and Method for Signal Dependent Noise Estimation and Reduction in Digital Images. In: US patent 5923775, 1999.

(STAMATOPOULOS *et al.*, 2007) N. Stamatopoulos, B. Gatos, and A. Kesidis, Automatic borders detection of camera document images. In: International Workshop on Camera-Based Document Analysis and Recognition, pp: 71-78.

(STEHMAN, 1997) S. V. Stehman. Selecting and interpreting measures of thematic classification accuracy. In: Remote Sensing of Environment , vol.62, pp: 77-89.

(STROUTHOPOULOS *et al.*, 2002) C. Strouthopoulos, N. Papamarkos, and A. E. Atsalakis. Text extraction in complex color documents. In: Pattern Recognition, vol.35, pp: 1743-1758.

(SU e DJAFARI, 2007) F. Su and A. Mohammad-Djafari. Bayesian Separation of Document Images with Hidden Markov Model. In: International Conference on Computer Vision Theory and Applications, pp:1-8.

(TAN *et al.*, 2003) T. Tan, R. K. Nishino, and K. Ikeuchi. Highlight removal by illumination-constrained inpainting. In: International Conference on Computer Vision, pp: 164-169.

(TIAN, 2013) D. P. Tian. A Review on Image Feature Extraction and Representation Techniques. In: International Journal of Multimedia and Ubiquitous Engineering, vol.8, pp: 385-396.

(TONAZZINI *et al.*, 2007) A. Tonazzini, E. Salerno, e L. Bedini. Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. In: International Journal on Document Analysis and Recognition, vol.10, pp: 17-25.

(VAILAYA *et al.*, 2001) A. Vailaya, M.A.T. Figueiredo, A.K. Jain, and H.J. Zhang. Image Classification for Content-Based Indexing. In: IEEE Transactions Image Processing, vol.10, pp: 117-130.



(WANG *et al.*, 2004) Z. Wang, A. C. Bovik, H. R. Sheikh e E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. In: IEEE Transactions on Image Processing, vol.13, pp: 600-612. Disponível em <https://ece.uwaterloo.ca/~z70wang/research/ssim/> (acessado em 11/07/2014 ).

(WANG, 2006) Y. Wang and P. Moulin. On Discrimination between Photorealistic and Photographic Images. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol.2, pp: 1-8.

(ZANCHETTIN *et al.*, 2012) C. Zanchettin, B. L. D. Bezerra, W. W. Azevedo. A KNN-SVM Hybrid Model for Cursive Handwriting Recognition. In: International Joint Conference on Neural Networks, pp: 1-8.

(ZHANG e TAN, 2001) Zheng Zhang and Chew Lim Tan. Recovery of Distorted Document Images from Bound Volumes. In: International Conference on Document Analysis and Recognition, pp: 429-433.

(ZHENG *et al.*, 2001)Y. Zheng, C. Liu, X. Ding, and Shiyan Pan. Form frame line detection with directional single-connected chain. In: International Conference on Document Analysis and Recognition, pp: 699-703.

(ZHENG *et al.*, 2003) Y. Zheng, H. Li, and D. Doermann. A model-based line detection algorithm in documents. In: International Conference on Document Analysis and Recognition, pp: 44-48.

(ZHU *et al.*, 2002) L. Zhu, A. Rao and A. Zhang. Theory of Keyblock-based image retrieval. In: ACM Transactions on Information Systems, vol.20, pp: 224-257.

(ZHU *et al.*, 2006) G. Zhu, S. Jaeger, and D. Doermann. A Robust Stamp Detection Framework on Degraded Documents. In: International Conference on Document Recognition and Retrieval, pp: 1-9.

## APÊNDICE A

### A.1 Publicações sobre Classificação em Imagens de documentos

(LINS et al., 2009) - R. D. Lins; G. F. P. Silva, J. S. Simske, J. Fan, M. Shaw, P. Sá, M Thielo. Image Classification to Improve Printing Quality of Mixed-Type Documents. In: International Conference on Document Analysis and Recognition, pp: 1106-1110.

(GIANETTI et al., 2010) - F. Gianetti; G. Dispoto ; R. D. Lins; G. F. P. Silva; G. Torreao; A. Cabeda. PDF profiling for B&W versus color pages cost estimation for efficient on-demand book printing. In: ACM-Symposium on Document Engineering, vol.1, pp: 177-188.

(SILVA e LINS, 2012a) G. F. P. Silva e R. D. Lins. Automatic Content Recognition of Teaching Boards in the Tableau Platform. In: International Conference on Pattern Recognition, vol.1, pp: 1-5.

## Image Classification to Improve Printing Quality of Mixed-Type Documents

Rafael Dueire Lins,  
Gabriel Pereira e Silva  
UFPE, Recife, Brazil  
rdl@ufpe.br,  
gfps@cin.ufpe.br

Steven J. Simske, Jian Fan,  
Mark Shaw,  
HP Labs., Palo Alto, USA  
{steven.simske, jian.fan,  
mark.q.shaw}@hp.com

Paulo Sá,  
Marcelo Thielo  
HP Labs., Porto Alegre, Brazil  
paulo.sa@hp.com  
marcelo.resende.thielo@hp.com

### Abstract

*Functional image classification is the assignment of different image types to separate classes to optimize their rendering for reading or other specific end task, and is an important area of research in the publishing and multi-Average industries. This paper presents recent research on optimizing the simultaneous classification of documents, photos and logos. Each of these is handled during printing with a class-specific pipeline of image transformation algorithms, and misclassification results in pejorative imaging effects. This paper reports on replacing an existing classifier with a Weka-based classifier that simultaneously improves accuracy (from 85.3% to 90.8%) and performance (from 1458 msec to 418 msec/image). Generic subsampling of the images further improved the performance (to 199 msec/image) with only a modest impact on accuracy (to 90.4%). A staggered subsampling approach, finally, improved both accuracy (to 96.4%) and performance (to 147 msec/image) for the Weka-base classifier. This approach did not appreciable benefit the HP classifier (85.4% accuracy, 497 msec/image). These data indicate staggered subsampling using the optimized Weka classifier substantially improves the classification accuracy and performance without resulting in additional “egregious” misclassifications (assigning photos or logos to the “document” class).*

### 1. Introduction

Image clustering has been researched by the database community since the early 1980's aiming to make efficient information retrieval in image databases [1][2]. In that kind of application one image is used to search the database looking for either the same or similar images. The basic idea is to try to organise the images in the database using some “common” features [3][2]. The same “features” are used to analyse the

image that will serve as the “search-key”. Instead of stepping through the whole database image-by-image, the retrieval process tries to match the properties of the search-key image, also known as *query image*, with the different image clusters in the database. This largely reduces the search-space making the retrieval process far more efficient. One of the features that has presented greater success in image retrieval was the analysis and clustering by the colour histogram [4][5]. The semantics of images have also been used as a clustering method [6] in database retrieval. Images that have similar “motifs” are most likely to have properties that are common to each other forming clusters. On the other hand, images whose theme completely uncorrelated should exhibit very different properties. Image classification is used in all-in-one and multi-functional devices to differentially render images belonging to different clusters. In particular, document, photo and logo images require widely different imaging pipelines to optimize their appearance when copied or printed. Documents (text, tables), for example, require sharpening that would damage the appearance of photos and logos. Logos use a palette that would “posterize” photos. Photos, in turn, can be rendered with a lower resolution (but greater bit depth) than either documents or logos. Figure 01 shows examples of typical representatives of the three classes of interest herein.

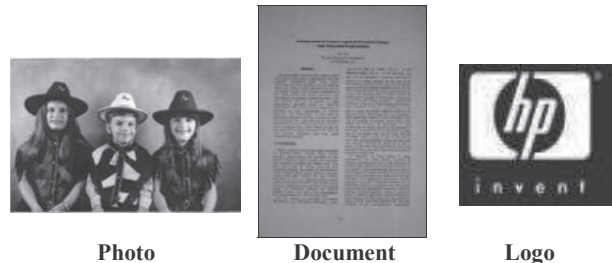


Photo                      Document                      Logo

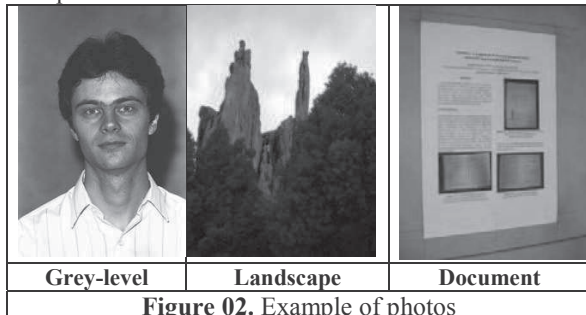
**Figure 01.** Example images of the clusters of interest  
This paper improves the results of the classifier described in reference [7] and presents a new, Weka-

based classifier [8] used to distinguish between these three types of images. Part 2 describes the experiments performed. Part 3 summarizes the results. We conclude with a discussion of the results.

## 2. Experiments Performed

The starting point for this work was collecting images that are representative of the different clusters of interest. Images were classified one-by-one by one person and the data-set was checked by three other people to avoid misclassifications and repetitions of the same image. Sometimes the “same” image appears in the test set in different file formats, for instance an image may appear in jpg, tiff, and bmp, as their features (palette, gamut, size) change from a format to another. Figure 01 shows an example of each of the classes of image of interest for this work. Images that do not belong to any of those classes are classified as “Don’t know”.

The “Photo” cluster encompassed many different sorts of photos, which ranged from people, landscapes, objects and even documents. Most photos were true-color although there were grey scale ones. The resolution also varied widely from VGA (480x640 pixels) to 7.2 Mpixels. The photos were collected from family albums of the people linked to the authors to ones obtained from the Internet. As professionals of many different areas start to use portable digital cameras to acquire images of documents, such images were included in this study, bringing an extra level of difficulty: a document acquired with a camera is classified as a photo or a document? The answer to that question is not straightforward and may puzzle even humans. The criterion adopted here was that if the image encompasses only the document it is classified as “document” if parts of the surroundings are included it is classified as “photo”. That means that if the document image in the rightmost part of Figure 02 is classified as “photo” and the same image after being processed in a tool such as PhotoDoc [9]. The “photo” test set has 7,968 photos of people and landscape and 500 photos of documents.



The 3,051 logos in the test set were collected from the Internet and from many different sources. Logos

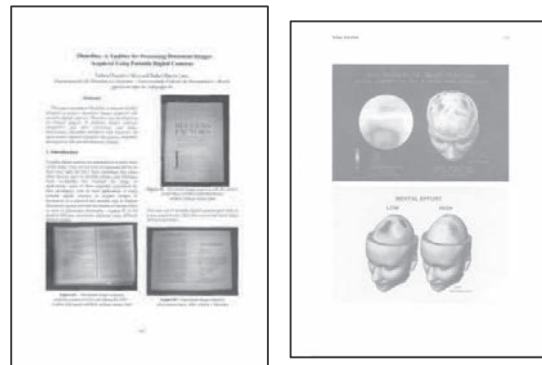
tend to exhibit a palette with a small number of colors, although very often they are saved in jpeg file format, introducing hues not originally intended. Figure 03 presents some examples of images classified as logos.



Figure 03. Examples of logos

The “Document” cluster was formed by 3,856 images of documents acquired from several different ways. Five hundred documents were photographed with a Sony Cybershot digital camera DSC-W55 in 5 and 7.2 Mpixels, with and without mechanical support, in-built strobe flash on and off, and then processed with PhotoDoc [9] that crops the framing border and corrects perspective and skew, should be classified as “document”. About half of the remaining documents were scanned documents with different resolutions (from 100 to 300 dpi) and saved in bmp, tiff, and jpeg, which although not suitable for such kind of image is often used by people in general [10]. The remaining photos were obtained by saving Adobe pdf documents into tiff and jpeg.

Figure 04 presents some examples of document images used in this work.



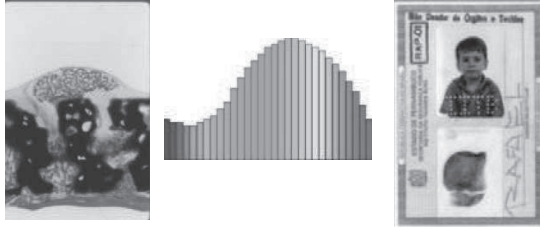
PhotoDoc processed

Scanned document

Figure 04. Examples of documents

The last cluster of images in the test set is the “Don’t Know” images. These images were included as to increase the possibility of misclassifications. They are images that appear in the “real world” and range widely in nature from biological images, to vector graphics (obtained by softwares such as Excell®,

Powerpoint®, etc.). Figure 05 shows examples of images labeled as “Don’t know”.



**Biomedical      Vector graphics      Document**

**Figure 05.** Examples of “Don’t Know” images

Table 01 shows the numbers of images per file format in the test set.

	JPG	TIFF	BMP	Total
<b>Photo</b>	7,476	35	457	<b>7,968</b>
<b>Logo</b>	2,984	0	67	<b>3,051</b>
<b>Doc</b>	3,048	808	0	<b>3,856</b>
<b>Don’t know</b>	202	0	327	<b>529</b>
<b>Total</b>	<b>13,710</b>	<b>843</b>	<b>851</b>	<b>15,404</b>

**Table 01 – Images per file formats**

## 2.1 Features Tested

The choice of the features to be extracted and tested is the key to the success and performance of the classification. Image entropy is often used as the key for classification [7]. It has a large computational cost, however. Entropy calculation demands a scan in the image to calculate the relative frequency of a given color, for instance, which is then multiplied for its logarithm and added up. The existing classifier is based on the binary classification approach originally described in [7]. It assumes a Gaussian distribution for each of the features, and its performance degrades in proportion to the non-Gaussian nature of the data.

This work assumed that decreasing the gamut of an image, analyzed together with its grey scale and monochromatic equivalents would provide enough elements for a fast and efficient image classification. The features tested are:

- Palette (true-color/grayscale)
- Gamut
- Conversion into Grayscale (if RGB)
- Gamut in Grayscale (if RGB)
- Conversion into Binary (Otsu)
- Number of black pixels in binary image.
- $(\#Black\_pixels/Total\ \#\_pixels)*100\%$
- $(Gamut/Palette)*100\%$  (true-color/grayscale)

Image binarization is performed by using Otsu [11] algorithm. The data above are extracted for each image and placed in a vector of features.

## 2.2 Training and test sets

Table 02 summarizes some of the features of the images in the test set. The height and width stand for the number of pixels in the image. RGB size stands for the true color size of the image (if a color image). 8-bits size is either the size of the original image if in gray scale or the size of the grey-scale converted from true-color. #B\_pixels stands for the number of black pixels in the monochromatic converted image.

Photo	Average	Median	Variance	Deviation
height	1138	1104	387968	622
width	1274	1132	548014	740
RGB size	1.98MB	1.49MB	460646	214627
8-bits size	667KB	578KB	675668	7548028
gamut RGB	13641	9956	171547287	5884
gamut gray	231	247	1288	0,707
#B_pixels	1373240	755150	23511	153332
Logo	Average	Median	Variance	Deviation
height	232	180	19324	139
width	253	221	13066	114
RGB size	15KB	7KB	45800	214010
8-bits size	9KB	5KB	50658	609718
gamut RGB	5324	3848	37149265	6095
gamut gray	230	241	1102	33
#B_pixels	38785	6579	59580	244095
Doc	Average	Median	Variance	Deviation
height	1896	1734	325997	570
width	1437	1328	201513	448
RGB size	1.10MB	879KB	111962	334581
8-bits size	674KB	568KB	207357	143999
gamut RGB	2700	2097	16317414	4039
gamut gray	181	211	5502	74
#B_pixels	1310386	749770	157416	3967564
Don’t Know	Average	Median	Variance	Deviation
height	477	363	180173	424
width	574	490	208282	456
RGB size	1.55MB	1.33MB	129678	3387135
8-bits size	520KB	386KB	323578	172717
gamut RGB	3954	2934	207165	5514
gamut gray	217	198	959	28
#B_pixels	101896	54475	82345	3169822

**Table 02 – Main features on the images in the test set**

The training set was carefully selected to guarantee the diversity of the images in the test set, having in mind that quality matters more than size. Table 03 presents the relative size of the training and test sets.

	Test	Training	%
Photo	7,968	668	8.34
Logo	3,051	412	10.22
Doc	3,856	276	4.70
Don't know	529	0	0
<b>Total</b>	<b>15,404</b>	<b>1,356</b>	<b>8.80</b>

**Table 03** – Sizes of Training x Test sets

The Weka [8] classification strategy used was the Random Forests (number of trees equal to 10) [12].

### 2.3 Sub-sampling

The factors for the feature extractor to improve were:

- 1- The larger the file - the richer in data redundancy - thus if the redundant data are thrown away the efficiency both in time and classification.
- 2- The selection of points should not be random. It should somehow provide a "reduced" version of the original image (although in some cases it may be distorted by unequal scaling!).

Twenty different sub sampling strategies were evaluated on the images. Two are presented here. The first, designated the "simple" subsampling technique, consisted of splitting the image in blocks of 4x4 pixels and averaging their values. This sub sampler provided the best overall improvement in performance for a subsampling approach that did not involve a decision tree. The second, the cascaded subsampling strategy, consisted of removing more points from the larger image files and provided the best overall accuracy of any classification schema, while simultaneously significantly improving performance, as shown in the next section. The cascaded subsampler performs the following operations:

size = height\*width

- If size ≤ 300,000 break;
- If 300,000 < size ≤ 500,000:  
remove 1 line or 1 column (whatever the larger);
- If 500,000 < size ≤ 700,000:  
remove 1 line and 1 column;
- If 700,000 < size ≤ 900,000:  
remove 2 lines and 1 column, (if height>width)  
remove 1 line and 2 columns, otherwise;
- If 900,000 < size remove 2 lines and 2 columns;

**Code for the "cascaded" sub-sampler**

## 3. Results

This section presents the results of classification of the images in the test set comparing the current classifier [7] and the one proposed herein. Both classifiers had the same training and test sets. The results are divided into three groups: original, simple (averaging), and cascaded subsampling.

### 3.1 Original Data

The results of classification are presented by the

confusion matrices obtained. Table 04 presents the results for the current classifier, while Table 05 shows the results obtained with the new classifier.

Current	Photo	Logo	Document	DK	A
Photo	7280	620	12	56	0.914
Logo	429	2104	96	422	0.690
Document	206	351	3299	0	0.856
Don't Know	70	225	0	234	0.442

**Table 04** – Confusion matrix of the current classifier with original images

The mean accuracy (A) for the Photo, Logo and Document images was  $12638/14875 = 85.3\%$ .

New	Photo	Logo	Document	DK	A
Photo	7554	363	14	37	0.948
Logo	282	2730	23	16	0.894
Document	277	266	3314	0	0.859
Don't Know	151	309	17	52	0.098

**Table 05** – Confusion matrix of the new classifier with original images

The mean accuracy (A) for the Photo, Logo and Document images was  $12893/14875 = 86.3\%$ , which shows that the proposed classifier is slightly better than the current one.

### 3.2 Simple Sub-sampler Performance

The results for the 4x4-pixel averaging subsampler are shown in Tables 06 and 07.

Current	Photo	Logo	Document	DK	A
Photo	5009	2209	586	164	0.628
Logo	1280	1005	115	651	0.329
Document	1245	376	2183	52	0.566
Don't Know	101	154	15	259	0.489

**Table 06** – Confusion matrix of the current classifier with subsampled images (simple)

The mean accuracy (A) for Photo, Logo and Document for the simple subsampler with the current classifier was  $8197/14875 = 55.1\%$ .

New	Photo	Logo	Document	DK	A
Photo	6674	1043	138	113	0.837
Logo	263	2751	20	17	0.901
Document	381	61	3414	0	0.885
Don't Know	124	330	26	49	0.092

**Table 07** – Confusion matrix of the new classifier with subsampled images (simple)

The mean accuracy (A) for the Photo, Logo and Document images was  $12839/14875 = 86.3\%$ .

The simple subsampler performed as well as the original version of the new classifier, but drastically degraded the accuracy of the current one. The simple subsampler halved the time for feature extraction of the new classifier as is shown in section 4.

### 3.3 Cascaded Subsampler Performance

The results for the cascaded subsampler are found in tables 08 and 09.

Current	Photo	Logo	Document	DK	A
Photo	7603	275	18	72	0.954
Logo	385	1929	135	602	0.632
Document	311	373	3167	5	0.821
DK	77	174	128	150	0.283

**Table 08** – Confusion matrix of the **current** classifier with **cascaded** subsampled images

The mean accuracy (A) for the Photo, Logo and Document images for the cascaded subsampler using the current classifier is  $12699/14875 = 85.3\%$

New	Photo	Logo	Document	DK	A
Photo	7740	164	34	30	0.971
Logo	258	2761	11	21	0.904
Document	93	41	3722	0	0.965
DK	110	300	30	89	0.343

**Table 09** – Confusion matrix of the **new** classifier with **cascaded** subsampled images

The mean accuracy (A) for the Photo, Logo and Document images was  $14223/14875 = 95.6\%$ . As one may observe the cascaded subsampler largely improved the performance of the new classifier and has a positive effect in performance, as shown below.

### 4. Time Performance

Table 10 presents the feature extraction and classification times together with information about the language those procedures were implemented into. Besides classification accuracy per cluster, the average feature extraction and classification times are presented. The entries with an “S” superscript denote the “simple” subsampler, while de “C” superscript stand for the “cascaded” subsampler. One should also remark that there is a difference in time scale between feature extraction and classification.

	Feature extraction		Classification	
	Time (s)	Language	Time (ms)	Language
<b>Current</b>	<b>1.4576</b>	<b>C#</b>	<b>6.13</b>	<b>C#</b>
<b>New</b>	<b>0.4174</b>	<b>C++</b>	<b>0.12</b>	<b>C#</b>
<b>Current<sup>S</sup></b>	<b>0.719</b>	<b>C#</b>	<b>6.13</b>	<b>Java</b>
<b>New<sup>S</sup></b>	<b>0.199</b>	<b>C++</b>	<b>0.12</b>	<b>C#</b>
<b>Current<sup>C</sup></b>	<b>0.497</b>	<b>C#</b>	<b>6.13</b>	<b>Java</b>
<b>New<sup>C</sup></b>	<b>0.1470</b>	<b>C++</b>	<b>0.12</b>	<b>C#</b>

**Table 10** – Feature extraction and classification times

### 5. Discussion and Conclusions

Weka has shown to be an excellent test bed for statistical analysis. The choice for a Random tree classifier was made after performing several experiments with the large number of alternatives offered by Weka, although results did not vary widely. Amongst them a preliminary comparison between the

new statistical classifier proposed here and a MLP neural classifier provided worse results (Photos 91.37%, Logos 85.48%, and Documents 94.54%).

The choice of the images in the training set is of paramount importance to the performance of the classifier. Quality has proved more important than size. For some reason not fully understood, the current classifier seems to be more sensitive to the quality of the training set than the one proposed herein. Enlarging the training set with incorrectly recognized data has proved efficient, but should be used with parsimony. Very small images seem to pose a higher degree of difficulty for classification, as they were more often misclassified.

The test set used here attempted to be representative of the universe of images of interest and incorporated two other test sets developed by Steven Simske and Mark Shaw that proved effective in the tuning of the current classifier. Every effort was made in the correct labeling of images and to avoid image duplication.

The new classification scheme provided here decreased the error rate by a factor of 3.2 (from 14.6% to 4.4%) while simultaneously improving performance by a factor of ten (from 1458 to 147 msec/image processing time) compared to the current scheme based on [7]. The increased accuracy improves the appearance of the printed output while the greatly improved performance frees up computing resources for additional printing tasks.

### 5. References

- [1] H.Frigui and R.Krishnapuram. Clustering by competitive agglomeration. *P. Recognition*, 30(7), 2001.
- [2] M.A.Hearst and J.O.Pedersen. Reexamining the Cluster Hypothesis: Scatter Gathet on Retrieval Results, SIGIR, 1996.
- [3] S.Krishnamachari and M.Abdel-Mottaleb. Image Browsing using Hierarchical Clustering. IEEE Symposium on Computers and Communications, ISCC'99, July 99.
- [4] P.Scheunders. Comparison of Clustering Algorithms Applied to Color Image Quantization, *Patt. Recog. Letters*, v18(11-13):1379-1384, 1997.
- [5] G.Park, Y.Baek and L.Heung-Kyu. A Ranking Algorithm Using Dynamic Clustering for Content-Based Image Retrieval. CIVR'2002, pp.328—337, LNCS 2383, Springer Verlag, 2002.
- [6] K.Barnard and D.Forsyth. Learning the Semantics of Words and Pictures, *Inter. Conf. C. Vision*, 2001.
- [7] S.J. Simske, “Low-resolution photo/drawing classification: metrics, method and archiving optimization,” *Proceedings IEEE ICIP*, IEEE, Genoa, Italy, pp. 534-537, 2005.
- [8] Weka 3: Data Mining Software in Java, website <http://www.cs.waikato.ac.nz/ml/weka/>.
- [9] G.Pereira e Silva and R.D.Lins. PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. CBDAR'2007, pp.107-114, 2007.
- [10] Lins, R.D. and D.S.A. Machado, A Comparative Study of File Formats for Image Storage and Trans., v13(1):175-183, *Journal of Electronic Imaging*, 2004.
- [11] N. Otsu. "A threshold selection method from gray level histograms". *IEEE Trans.Syst.Man Cybern.* v(9):62-66, 1979.
- [12] L. Breiman, “Random Forests”, *Machine Learning*, 45(1), pp. 5-32, 2001.

# PDF Profiling for B&W versus Color Pages Cost Estimation for Efficient On-Demand Book Printing

Fabio Giannetti  
Gary Dispoto  
HP Laboratories  
1501 Page Mill Rd. M/S 1160  
Palo Alto, CA 94304  
+1 650 8575085  
{fabio.giannetti,  
gary.dispoto}@hp.com

Rafael D. Lins  
Gabriel F. P e Silva  
Universidade Federal de Pernambuco  
Recife, PE, BRAZIL  
+55 81 2126 8210  
{rdl, gabriel.psilva}@ufpe.br

Alexis Cabeda  
Hewlett-Packard Brasil  
Av. Ipiranga, 6.681 - Prédio 95  
Porto Alegre, RS, Brazil  
+55 51 2121 3559  
alexis.cabeda@hp.com

## ABSTRACT

Today, the way books, magazines and newspapers are published is undergoing a democratic revolution. Digital Presses have enabled the on-demand model, which provides individuals with the opportunity to produce and publish their own books with very low upfront cost. With these new markets, opportunities, and challenges have arisen. In a traditional environment, black-and-white and color pages were printed using different presses. Later on, the book was assembled combining the pages accordingly. In a digital workflow all the pages are printed with the same press, although the page cost varies significantly between color and b/w pages. Having an accurate printing cost profiler for pdf-files is fundamental for the print-on-demand business, as jobs often have a mix of color and b/w pages. To meet the expectations of some of HP customers in the large Print Service Providers (PSPs) business, a profiler was developed which yielded a reasonable cost estimate. The industrial use of such a tool showed some discrepancies between estimated and printer log, however. The new profiler presented herein provides a more accurate account of pdf jobs to be printed. Tested on 79 "real world" pdf jobs, totaling 7,088 pages, the new profiler made only one page misclassification, while the previous one yielded 54 classification errors.

## Categories and Subject Descriptors

H.4.0 [Information Systems]: Applications, printing, profiling.

## General Terms

Management, Measurement, Performance, Economics, Reliability, Experimentation.

## Keywords

Pdf profiling, printing costs, digital presses.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*DocEng'10*, September 21–24, 2010, Manchester, UK.  
Copyright 2010 ACM 978-1-4503-0231-9/10/09...\$10.00.

## 1. INTRODUCTION

Book printing has been re-invented with the adoption of Digital Presses. The entire production and re-ordering process has been simplified and the time has been reduced significantly. In offset printing, the cost of making press templates for each page yields unviable to have customized documents, besides making prohibitive the cost of printing only a few copies. Digital Presses have allowed the on-demand model, which opens the opportunity to print customized documents and books with very low upfront cost. The cost per page in Digital Presses for printing one copy is now basically the same as printing 10,000 copies, although it is still higher than for an Offset Press. The price gap is even higher for color pages since printer manufacturers in the commercial printing market, charge Print Service Providers (PSPs) a certain amount per impression. A CMYK print job requires 4 impressions versus a "Black ink only print job", which only requires 1 impression. The so-called "click charge" for a color page is thus 4 times as high as the one for a b/w page. This difference is due to the fact that not only more ink is used, but the Press has to image the sheet up to four times, using more of the consumables in the transfer mechanism, which needs to be replaced after a certain amount of impressions. Books submitted directly through self-publishing web sites [1], old books and out-of-print books restoration services [2] have a very unpredictable mix of color and b/w pages. Sometimes, as result of the scanning and processing, the book content is stored in an RGB encoding even if the original content is Grayscale or monochromatic preventing the publication to be printed at a lower price. A further issue is the "click" charge is based on sheets. This means that the overall cost can only be computed after the book is imposed. Many PSPs, though, have several presses (roll and sheet fed) and they may employ different imposition strategies depending on e.g. press availability, and substrate requirements.

The Job Profiler tool analyzes the print job, provides an accurate count of color vs. b/w pages and stores page-by-page meta-information in a database. It also identifies grayscale content that has been stored in a sub-optimal way resulting in unnecessary color impressions. HP PSP customers believe that applying a cost optimization process to the print job in question will significantly reduce the printing cost and make a better estimate of these costs. It is important to stress that the profiler is targeted at book printing, in which the majority of pages is monochromatic and the few color pages within have a large impact on printing costs. In the case of other printed materials, such as customized magazines, which tend to be colorful, the cost parameters are different.



## 2. THE PROFILER

The proposed solution is to profile the print job and extract per page color information. This step is supposed to be performed during the publication upload phase for computational efficiency. The Job Profiler parses the individual pages of the print job, identifies potential color objects and marks the corresponding pages accordingly either as color or b/w pages. The final result is a Job Profile containing the color information (color or b/w) for each page. The Job Profile can then be used with a second application, Cost Planner, that virtually applies several imposition templates to the job producing a set of prognosticated costs based on the number of color or b/w sheets produced. This helps the PSP to provide a realistic pricing to the customer as well as planning to which presses jobs should be sent to in order to save money. Figure 1 illustrates the proposed workflow with the various steps. Once one imposition strategy is chosen, the job is loaded, sent to a compatible press, printed and shipped to the customer. The Job Profile metadata is stored with the job in the PSP database such that job re-orders (even with imposition changes) will not require an additional profiling operation.

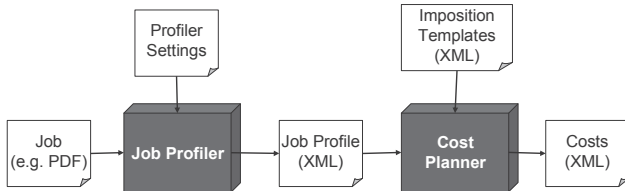


Figure 1: Proposed workflow with Job Profiler and Cost Planner.

The Job Profiler acts as a color related preflight tool. In fact, it is also possible to perform a more thorough analysis parsing the object streams to identify if the object color encoding corresponds to the original intended color. Often due to scanner settings and mixed color and b/w content the objects are stored as RGB even if the original content is a grayscale image or text. Documents containing objects with R=G=B will commonly be transformed during the ripping process into CMYK data, where C, M and Y are unequal to zero. This has been tested and confirmed using the default settings for most of the industrial digital presses. Even CMYK objects, with C=0,M=0,Y=0 can potentially be transformed into device dependent CMYK values where C, M and Y are unequal to zero if the RIP is set to color manage the object using a CMYK source and destination ICC profile. Moreover, due to the scanning limitations or poor PDF conversion, a print job could contain spurious CMY information. A threshold based algorithm could optimise these values avoiding an unnecessary color print. The Job Profiler is capable of identifying color mismatch cases and issuing a warning. The coordinates about the mismatched object are detailed in order to simplify its retrieval. The optimization task can be performed in a subsequent step as illustrated in Figure 2. Once the job has been optimized an updated version of the profile is generated to reflect the changes. The Cost Planner module can now perform the evaluation based on the optimized job.

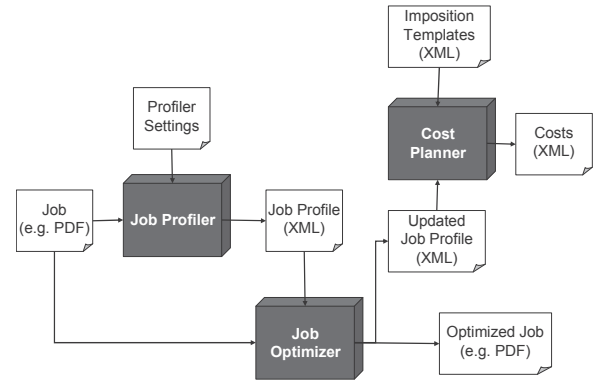


Figure 2: Expanded workflow with Job Optimizer step.

## 3. PROFILER ARCHITECTURE

The pdf Profiler was developed to analyze a document in pdf and to extract the information on the pages of the document. The tool was implemented using the PDF-Box\_0.7.3 [5] library in Java to search and classify the pdf structural elements. The main objective of the pdf Profiler is to extract the needed information for printing each page. The PDF format was studied in detail in order to provide a better understanding of its objects and how they can be represented. The PDF format is a document description language that uses vector graphics to represent the document content, providing several advantages to the PSPs, such as document resolution independence. The PDF specification [7] describes commands to define or use primitive types, such as, integer and floating point numbers, strings, boolean values, etc. Furthermore, through the use of the commands, it is also possible to define the elements which will be painted in the pages of a document.

### 3.1 Profiler Modules

On top of the library Java PDF-Box\_0.7.3 [5] it was possible to define the profiler in four modules: *GSave*, *GRestore*, *TextShow*, *PaintXObject*. Their functionality and challenges in the code development are discussed.

The modules *GRestore* and *GSave* are directly related to the state of the graphical scope. In a pdf document the graphical state may be restored and saved in those two modules, which control the graphical scope of objects using a stack. The top of the stack holds the graphical state when an object scope is saved and pops it out if there is a re-store command. Thus, the top of the stack holds the current state that is valid to the graphical objects in a pdf page. Thus, this kind of control allows verifying for each graphical object that is under analysis to check if there is transparency applied to it. This kind of control allows checking for each graphical object under analysis if there is transparency applied on it. For the Profiler to work with these two modules there was no need to modify the basic structure of the PDF-Box\_0.7.3 library.

The module *TextShow* checks if the current page has text or "transparent text". This module is instantiated whenever there is text to be plot in the pdf document. Once created, it verifies the actual state of the graphics to see if there is transparency and marks the current page as either having text or transparent text; at last it checks the color of the text font. This last phase demanded the PDF-Box\_0.7.3 library to be re-coded, as it only informs the

existing text together with its color space. That means that the colors of the text were not decoded.

Finally, the module *PaintXObject* handles *PDXObjectImage*, for: (i) Checking if the XObject is being re-used or not; (ii) Testing whether it is transparent. (iii) Retrieving the area of objects. To perform the first of the three tasks, such module holds a map with the encoded images. Thus, whenever a new image is inserted, it first checks that map to see if the image is re-used. The graphical state is read from the stack and evaluated. At last, the PDF Profiler retrieves all the properties of *PDXObjectImage*, through the *XObjects* dictionary, that means that it starts the image decoding process. The encoding performed by Adobe for PNG images in the CMYK color space is not always compatible with the decoding standards adopted by most image processing libraries, however.

Here, rested the greatest difficulty of the development of the PDF Profiler, as the PDF-Box 0.7.3 library did not decode the *PDXObjectImage* adequately. For instance, often monochromatic images were decoded as color images and color images of type (C=1, M=1, Y=1 e K=0) were decoded as binary. To overcome this problem, it was necessary to re-write the code of the PDF-Box\_0.7.3 library to decode the primitive components (stream parsing). This task was performed through adapting the code of the JPedal [6] library, which offered a more accurate response.

#### 4. TESTS AND PERFORMANCE

The Job Profiler was tested with a library of jobs collected from different HP PSP industrial customers. Figure 3 illustrates the profiling times with the deeper analysis enabled for real book examples. It can be evinced that the longest time is not directly correlated to the file size, since re-usability and correctly color specified embedded data streams speed-up the process.

FileName	Size (KB)	Parsing (s)	Profiling (s)	Out.Writing (s)
Book_01.pdf	11,265	0.03	1.54	0.03
Book_02.pdf	1,783	0.03	2.19	0.03
Book_03.pdf	15,571	0.03	0.34	0.02
Book_04.pdf	51,823	0.05	0.90	0.05
Book_05.pdf	8,057	0.05	2.95	0.05
Book_06.pdf	25,197	0.03	0.48	0.03
Book_07.pdf	61,229	0.03	0.39	0.03
Book_08.pdf	121,017	0.03	1.51	0.03

Figure 3: Job Profiler Processing Times for a PSP customer and Output example.

Figure 4 illustrates the output of the Job Profiler for a simple 32 page print job, which contains grayscale images encoded in form of sRGB. The Cost Planner module prognosticates the results if this job is printing using two different impositioning strategies: 2-up perfect bound and 4-up perfect bound. In the first case the following pages will be combined, (32,1),(31,2),(30,3),(29,4)...(17,16) and result in 13 color and 3 b/w sheet impressions. In the second case, combining pages (32,1,18,15), (31,2,17,16) ... (26,7,24,9), (25,8,23,10) will lead to 8 color sheet impressions. Under the current pricing module of most industrial printers the cost for the PSP would be lower for the 2-up imposition than for the 4-up imposition.

```
<?xml version="1.0" encoding="UTF-8" ?>
<profile>
  <overallInformation bw="14" colorful="18" pages="32" />
  <pagesInformation>
    <page color="NO" number="1" />
    <page color="YES" number="2" />
    <page color="NO" number="3" />
    <page color="NO" number="4" />
    <page color="YES" number="5" />
    <page color="YES" number="6" />
    <page color="YES" number="7" />
    <page color="YES" number="8" />
    ...
    <page color="YES" number="27" />
    <page color="NO" number="28" />
    <page color="NO" number="29" />
    <page color="NO" number="30" />
    <page color="NO" number="31" />
    <page color="NO" number="32" />
  </pagesInformation>
</profile>
```

```
<?xml version="1.0" encoding="UTF-8" ?>
<profile>
  <overallInformation bw="2" colorful="206" pages="208" />
  <pagesInformation>
    <page color="NO" number="1" />
    <page color="NO" number="2" />
    <page color="YES" number="3">
      <warning identified-colorspace="Grayscale" object="10" stated-colorspace="DeviceRGB" />
    </page>
    <page color="YES" number="4">
      <warning identified-colorspace="Grayscale" object="14" stated-colorspace="DeviceRGB" />
    </page>
    <page color="YES" number="5">
      <warning identified-colorspace="Grayscale" object="18" stated-colorspace="DeviceRGB" />
    </page>
  </pagesInformation>
</profile>
```

Figure 4: Output of the Job Profiler (top) and with Cost Optimization Identification (bottom).

The new profiler was tested on a set of 79 books provided by one of the largest HP PSP customers. That test set encompassed 7,088 pdf pages generated by different professional and non-professional tools. Often, the book pages had color, grayscale, and binary images. Transparency was used in some of them. Similarly, the images were processed using several environments. Pdf files were generated using different tools. The diversity of sources and tools used in the generation of the test set was considered as representative of the universe of jobs to be printed by a real PSP commercial enterprise.

The new pdf Profiler showed an accuracy of 99.98%, while the previous version of the profiler yielded an accuracy of 99.238%. Unfortunately, errors were non-uniformly distributed in the test universe, as they tended to be concentrated in some jobs, which could have their printing cost estimation unviable. That may be an indication that some of the tools used in the generation of the books provided non-standard outputs, which made less accurate their profiling and cost estimation.

Profiler	Time(s)
HP (PDFBOX)	42.99
UFPE-HP_v1 (PDFBOX)	41.31
UFPE-HP_v2 (PDFBOX+JPEDAL)	38.23

Table 1 : Average time performance for 79 pdf files in 10 runs.

Another advantage of the new profiler is that the accuracy gains attained did not bring performance penalties as shown in Table 1, the final version of the pdf Profiler that made use of PDF-Box\_0.7.3 and JPedal, developed jointly between HP and UFPE (Brazil) was almost 10% faster than the first version of the pdf Profiler, developed by HP.

## 5. CONCLUSIONS AND FURTHER WORK

Competitive solutions could be implemented by leveraging either pre-flight tools or RIPs. Nevertheless, none of them provide the requested flexibility. In fact, RIP solutions would require a pre-imposition and ripping of the job just for cost evaluation purposes. If a PSP would like to compare the costs for a variety of imposition strategies they would need to RIP it several times. Pre-flight tools may be more flexible but are focused on checking image dpi, fonts and every aspect that can prevent the document from being incorrectly printed [3][4]. Color or b/w pages are not relevant from a pre-flight tool perspective, but are of paramount importance in printing cost estimation. This was the central motivation for this work which largely improved and corrected a previous pdf Profiler developed by HP labs (Palo Alto), already in beta-testing in some of the largest HP PSP customers.

The new profiler, developed in collaboration between HP and UFPE (Brazil), claimed for a thorough analysis of the previous profiler for minimizing misclassifications. Tested on a set of 79 books totalling 7,088 pages, the new profiler made only one page misclassification, while the previous one yielded 54 classification errors. Such profiling standard is considered acceptable and may be considered for a more sophisticated cost optimisation process. This can be achieved for the proposed Cost Optimization module. The profiler used an integration of PDF Box [5] and JPedal [6], two free licensed tools, as neither of them alone provided the full functionality needed. The use of PDFNET [8] or the Adobe PDF Library may make the implementation of the profiler simpler as they provide more powerful PDF analysis environments. On the other hand, the profiler as developed here allows the direct correction of the false B/W into “pure” B/W.

The Cost Planner and the Job Optimizer are under investigation and will be added to the solution as soon as the integration with the customer’s workflow becomes clearer. The Cost Planner can be implemented using an XSL-T transformation applying a PPML

imposition template. The Job Optimizer, on the other hand, requires the specification and testing of the parameters for the color optimization and the development of a stream conversion tool.

Another possibility of use of the pdf Profiler is to spot the outputs of non-conventional environments and correcting them into standards used.

## ACKNOWLEDGEMENTS

The research reported herein was partly sponsored by an MCT-Brazilian Government R&D Grant between Universidade Federal de Pernambuco and Hewlett-Packard do Brasil Ltda. Rafael Lins and Gabriel Silva also are grateful for CNPq and CAPES funding.

## REFERENCES

- [1] I Universe, Self-Publishing Website, Visited on July, 7th 2010. <http://www.iuniverse.com/>
- [2] Bookprep, <http://www.hp.com/idealab/us/en/bookprep.html>
- [3] Pitstop, Enfocuse, <http://www.enfocuse.com/>
- [4] Flightcheck, Markzware, <http://www.markzware.com/>
- [5] PDF-Box Home Page. Extracted from <http://www.pdfbox.org>, December 20th 2009.
- [6] JPedal Home Page. Extracted from <http://www.jpedal.org/>, December 20th 2009.
- [7] Adobe Systems. PDF Reference. Adobe Systems Incorporated, San Jose, USA, 4th edition, 2003.
- [8] PDFNET. <http://www.pdftron.com/pdfnet/downloads.html>. Visited on July, 7th 2010.
- [9] Adobe PDF Library. Visited on July, 7th 2010. <http://www.datalogics.com/products/pdf/pdflibrary.asp>.

# Automatic Content Recognition of Teaching Boards in the Tableau Platform

Gabriel de França Pereira e Silva and Rafael Dueire Lins  
*Universidade Federal de Pernambuco, Recife, Brazil*  
{gfps, rdl}@cin.ufpe.br

## Abstract

Teaching boards are omnipresent in classrooms throughout the world. Tableau is a software environment for processing images from teaching-boards acquired using portable digital cameras and cell-phones, being the first software environment able to process non-white boards. The aim of the Tableau environment goes far beyond processing and compressing teaching-board images, it also targets at content analysis and recognition. This paper presents the content analysis tool in Tableau that is able to automatically segment and classify textual and pictorial areas.

## 1. Introduction

Anyone in the teaching area today often sees students in classroom taking photos of the slides being presented or from the teaching board. This phenomenon also happens in conferences and work meetings. It is due not only to the fact that portable digital cameras have become omnipresent, either in their own or embedded in cell phones and tablets, but also because the quality of images acquired has become increasingly better.

Teaching boards are possibly the oldest and most universal didactic tool used throughout the world. Originally, they were made with slices of dark stone to be written on with a piece of chalk. Over the centuries it evolved to whiteboards used with erasable felt pens. Figure 1(a) presents two images of “real world” teaching boards, in which one may observe some of the problems met in such images: perspective distortion, uneven illumination with specular noise, inclusion of background areas and non board elements, etc. Figure 1(b) exemplifies the complexity of a board image in which textual elements are mixed together with drawings, tables, and text in a total absence of pattern and organization.

WhiteboardIt [4, 9, 10, 11] pioneered the research in white board image processing and became a commercial product. Independent work lead to the

development of Tableau [2], a simple software tool to process images of teaching boards acquired using portable digital cameras, implemented as an ImageJ plugin, freely available and distributed. Tableau aims to provide a way to generate digital content for courses, respecting particular aspects of the group such as syllabus, class learning speed, teacher experience, regional content, local culture, etc. Although whiteboards have the advantage over chalkboards of generating no chalk dust, that causes allergy to teachers and students, chalkboards are still of widespread use, overall in developing countries. Thus, to reach its original aim to help students and teachers to generate didactic content, Tableau was generalized to process images of any color of board [2]. Until today, this is a unique feature of Tableau, as no other similar tool is able to process non-white boards.

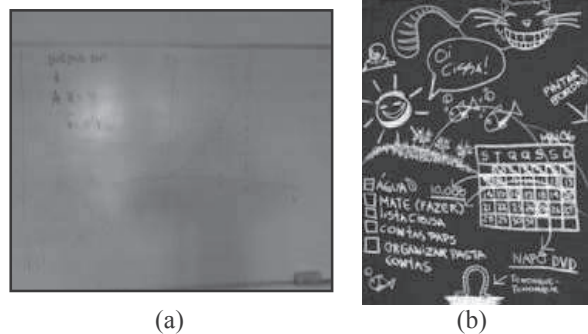


Figure 1 – Images of a “real world” teaching board.

The Tableau platform consists of three parts. The first is database formation. As soon as images are transferred from the camera to the PC, information is collected to generate a simple database that will organize images for later content formation. Information such as teacher name, course name, discipline, subject, class number, group number, etc. are requested. The second module is for image processing. This module improves the image acquired in a number of ways: background (non-board) removal, image segmentation, skew and perspective correction,

image enhancement, etc. The third part of the processing environment deals with outputting the content. Three different ways are under development: printed handouts, webpage generation and Powerpoint™ slide production. Each of these media receives the information of the processed image part of the environment and makes it suitable to its best use. Image “understanding” is needed in order to be able to correctly lay-out the content of the acquired image. This paper targets at such a difficult task and presents the first version of the content analysis tool in Tableau, which is able to discriminate textual and pictorial areas. The text classification is further refined into cursive and block writing, while the pictorial elements are classified into tables, pizza diagram, histogram, and line plotting.

## 2. Board Graphical Elements

A board image may convey a great variety of graphical elements, such as text, tables, drawings, etc. In general there is no high-level descriptor one could associate with parts of the board image. Trying to understand how such elements are organized in the board image is fundamental to the correct interpretation of its content. The layout analysis has as objective to detect and form the different areas in the image to identify graphical elements as a whole and relation between them. The human brain uses several clues such as contextual information and past experiences about layout recognition together with a sophisticated and complex reasoning mechanism to segment an image classify its elements and “understand” them. The machine, on the other hand, has to infer semantics only from syntax, or in this case from image layout. The idea here is to try to cluster similar pixels forming pictorial elements or text together and “interpret” them by pattern-matching them with “syntactically” similar elements for which semantic value have been attributed to in “artificially intelligent” systems. This is the reason for which automatic layout and structural analysis of arbitrary documents and images is such a challenging research problem. While in some printed documents one may assume some a priori knowledge about layout [3][10][11] such as artifact and graphical settings of textual elements (position of document heading, type of fonts, size, etc.) [3], this is not the case in board images in general.

The system proposed here prioritizes text finding in the board image. Once a cluster of pixels is recognized as text the classification is refined into cursive and block writing. This strategy has shown valuable in making the problem less complex and increasing the correct identification rate. In general, people writing style follow standards, written blocks

have more complex patterns (loops and hollows) in a small area in relation to patterns typically found in graphics, which are larger and less dense.

## 3. Tableau Board Recognition System

The board element identifier in Tableau encompasses four phases: pre-processing, segmentation, layout analysis and recognition, as sketched in Figure 2.

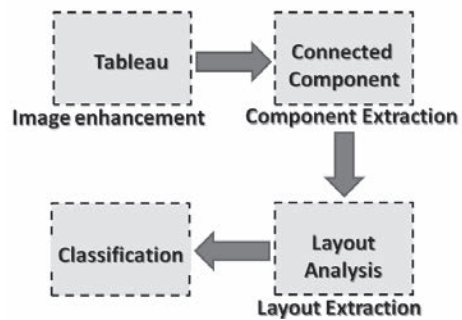


Figure 2 -Block diagram of the element recognition system in Tableau

Image preprocessing is fundamental to the success of the segmentation of the components in the board image. This phase makes use of specific algorithms for noise removal [4] and for the processing of images of teaching boards [2] developed for the Tableau platform. At the end of the preprocessing phase, a binary image is generated by applying the Sauvola-Pietikainen algorithm [5].

## 4. Layout Analysis and Segmentation

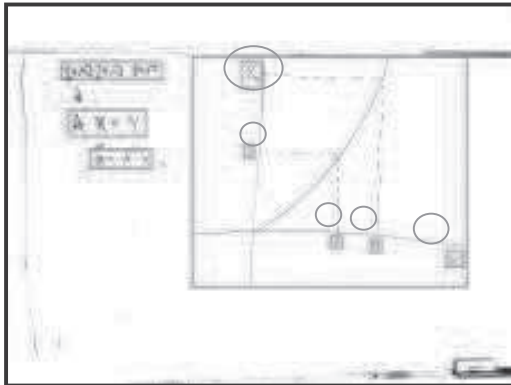
At this stage we use the Tableau [2] to filter out the noise acquired in the digitalization process (specular noise from uneven lighting and perspective) and then to perform background segmentation.

The second segmentation step attempts to identify the regions in the board image that encompass graphical or textual elements. The segmentation phase is performed in two steps. First, non-background areas are clustered by using the connected component algorithm. Then, the image blocks formed by clusters are analyzed to be possibly merged by taking into account the average size and distance between graphical and textual elements, defining an envelope-box that should contain the whole element.

The detection of a graphical element should be as coarse as possible. For instance, table recognition must encompass all its elements including the textual information on coordinate names, etc. In general a teaching board has a high density of elements and information, thus the conventional content extraction techniques are not good enough and inefficient. To circumvent such problem a hierarchical information

arrangement is adopted here, allowing to link attributes to the extracted content, as shown in Figure 3.

The test set used encompasses 348 “real-world” teaching board images acquired from different courses and classrooms using a portable digital camera manufactured by Sony DSC-T10 of 7.2 MPixels and by the embedded camera in a Samsung Galaxy Mini cell-phone of 5 MPixels. The test images were “hand” labeled and segmented to be used as ground truth in the automatic tests performed. Table 1 presents the number of occurrences of such elements in the dataset.



**Figure 3** – Hierarchical component connection of the elements of Fig 1(a). In red: textual areas, in blue graphical elements, in green hierarquical association of textual areas to the graphical element.

Type	Total
Tables	136
Cursive Text (words)	2,983
Text in Block format (words)	1,254
Picture	211
Pizza diagrams	321
Histograms	254
Line Plottings	437
Don't know	62

**Table 1**-Total of classes manually labeled in the test images.

The test images were submitted to the segmentation algorithm described and tested against the “hand-made” segmentation. The accuracy of the segmentation algorithm in finding the “bounding-box” that frames each of the graphical elements in the test images is shown in Table 2

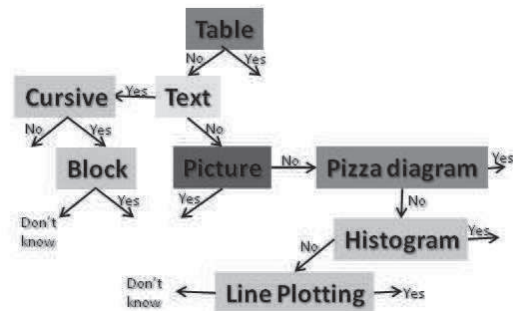
Type	Total	Accuracy
Tables	97	71.32%
Cursive Text (words)	2,132	71.47%
Text in Block format (words)	968	77.19%
Picture	183	86.72%
Pizza diagrams	298	92.83%
Histograms	213	83.85%
Line Plottings	297	67.96%
Don't know	25	40.32%

**Table 2** – Accuracy of the automatic segmentation process.

Considering that the test images are “real-world” ones and the teacher did not know *a priori* that the board images would be used for processing, the segmentation rate is good enough, showing that the segmentation tool in Tableau environment is useful.

## 5. The Element Classifier

In the last processing phase of the Tableau recognition system performs the automatic classification of the graphic and textual elements of the segmented blocks. A cascaded binary classifier was used, in which each of them is a Random Forest [6] classifier. Making the classifier binary allows to reduce the complexity of the classification problem as well as it allows that new branches to be introduced into the classifier in a simple way [12]. The current version of the classifier, as shown in Figure 4, starts by attempting to differentiate between tables and other classes. If a block area is recognized as textual, it goes into another classifier that decides whether the writing is in cursive or in block format. Further classification as a pizza (or pie) diagram, a histogram and a line plotting is made. If the classifier does not “guess” the “nature” of the image block it outputs “Don't\_know”.



**Figure 4** – Tree of binary RandomForest classifiers.

### 5.1 Feature Extraction

The recent work in image classification [13] points at the model called “bag of visual word” (BoW) [13] based on the quantization of local descriptors [14][15] and SVM (support vector machines). Such technique is refined with the use of multi-scale spatial pyramids (SPM) [15] together with BoW or with the features of orientation of the histogram of the gradient (HGO) [16]. The current “state-of-the-art” apply multiple descriptors and kernels are combined using learning kernel approaches [17][18]. Such methods did not show suitable to address the classification problem presented here, however. This may be due to the dimension of the data set or to the diversity in the class description universe. Here a Gaussian data distribution is assumed and its performance degrades as their nature distance away from such model. The efficiency

in assuming such distributions is reported in references [7][12]. The following features were extracted from each segmented area from the enhanced board image: Palette (true-color/grayscale); Gamut; Gamut in Grayscale; Number of black pixels in binary image;  $(\#Black\_pixels/Total\ \#\_pixels)*100\%$ ;  $(Gamut/Palette)*100\%$ ; (true-color/grayscale); Number of open and closed loops; Size of the segmented blocks.

The average time for feature extraction was 1.02 s in a Intel® Core 2 Quad Q8300 with 4 GB of RAM.

## 5.2 Classification Accuracy

The size of the training set used with the different classes is shown in Table 3. The classifier used was Random Forest [6], with cross-validation implemented in Weka [8]. The images in the training set were automatically segmented by the algorithm presented.

Type	Total
Tables	28
Cursive Text (words)	300
Text in Block format (words)	150
Picture	30
Pizza diagrams	60
Histograms	40
Line Plottings	60

**Table 3** – Size of the training set.

The performance of the classifier was tested on manually segmented blocks. No block used for training is part of the test set. The classification accuracy is shown in Table 4. Classification time is 7 ms / image.

Type	Total	Accuracy
Tables	87	80.55%
Cursive Text (words)	2,683	91.91 %
Text in Block format (words)	1,104	83.42 %
Picture	131	72.37 %
Pizza diagrams	261	77.01 %
Histograms	214	85.51 %
Line Plottings	377	90.71 %
Don't know	41	66.12%

**Table 4** – Classification accuracy for the segmented board elements.

## Conclusions

The image segmentation and classifier presented here is an important tool in the Tableau environment, because it will allow it to meet its original purpose and functionality of generating digital content from teaching board images. The automatic segmentation part reached over 70% accuracy, while the classifier reached accuracy over 80% in “real world” teaching board images. It is reasonable to assume that if teachers knew *a priori* that their boards would be processed a better layout would be obtained, increasing

the performance of the tool. Another possibility is to customize the tool to a certain teacher, by making his/her “private” training set. Although the Tableau environment aims to ease users’ life by inferring the content of teaching board areas, it is recommendable that the user checks the results obtained. This is especially important if the image is to be automatically transcribed to generate a vectorized image.

## References

- [1] G.F.P e Silva and R.D.Lins. PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras CBDAR 2007. p.107 - 115.
- [2] D.M.Oliveira, and R.D. Lins. Generalizing Tableau to Any Color of Teaching Boards. ICPR 2010. pp: 2411-2414. IEEE Press.
- [3] R.D. Lins, P.H.Espirito-Santo, G.F.P. e Silva. Academus - Generating Digital Libraries of M.Sc. and Ph.D. Thesis. Submitted to ICPR 2012.
- [4] G.F.P e Silva, R.D. Lins, J.M.M.Silva, S. Banergee, A.Kuchibhotla, M.Thielo. Enhancing the Filtering-out of the Back-to-Front Interference in Color Documents with a Neural Classifier. ICPR 2010.
- [5] J. Sauvola, M. Pietikainen, Adaptive document image binarization, Pattern Recognition 33 (2) (2000) 225–236.
- [6] L. Breiman, "Random Forests", Machine Learning, 45(1), pp. 5-32, 2001.
- [7] S.J.Simske, Low-resolution photo/drawing classification: metrics, method and archiving optimization, Proceedings IEEE ICIP, pp. 534-537, 2005.
- [8] Weka 3: Data Mining Software in Java, website <http://www.cs.waikato.ac.nz/ml/weka/>
- [9] R.D.Lins. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. ICIAR, Springer Verlag, 2009. v.5627. p.844 - 854.
- [10] A. Antonacopoulos, S. Pletschacher, C. Clausner and C. Papadopoulos. Historical Document Layout Analysis Competition, ICDAR 2011, Beijing.
- [11] C. Clausner, S. Pletschacher and A. Antonacopoulos. Aletheia - an advanced document layout and text ground-truthing system for production environments, ICDAR 2011.
- [12] R.D.Lins, G.F.P e Silva, S.J.Simske, J. Fan, M. Shaw, Mark, M.Thielo. Image Classification to Improve Printing Quality of Mixed-Type. ICDAR 2009. IEEE Press, 2009. p.1106 – 1110.
- [13] C.Dance, J. Willamowski, L.Fan, C.Bray, G. Csurka. Visual categorization with bags of keypoints. I. Workshop on Statistical Learning in Computer Vision. 2004.
- [14] D.Lowe. Distinctive image features from scale-invariant keypoints. IJCV (2004) 91-110.
- [15] S.Lazebnik, C.Schmid, J.Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. CVPR06. (2006).
- [16] N.Dalal, B.Triggs. Histograms of oriented gradients for human detection. In: CVPR05. (2005) p. 886-893.
- [17] A.Vedaldi, V.Gulshan, M.Varma, A.Zisserman. Multiple kernels for object detection. In: ICCV. (2009)
- [18] P.V.Gehler, S.Nowozin. On feature combination for multiclass object classification. In: ICCV. (2009).

## **A.2 Publicações sobre Classificação de Dispositivos de Captura**

(SILVA et al., 2009a) - Silva, G. F. P.; Lins, R. D.; Miro B.; Simske, S.; Thielo, M. Automatically Deciding if a Document was Scanned or Photographed. In: Journal of Universal Computer Science, 2009, v.15, pp: 3364-3366.

(LINS et al., 2011a) - R. D. Lins; G. F. P. Silva; J. S. Simske. Automatically Discriminating between Digital and Scanned Photographs. In: International Conference on Document Analysis and Recognition, vol.1, pp: 1280-1284.



## **Automatically Deciding if a Document was Scanned or Photographed**

**Gabriel Pereira e Silva, Rafael Dueire Lins, Brenno Miro**

(Federal University of Pernambuco, Recife, Brazil  
gfps@cin.ufpe.br, rdl@ufpe.br)

**Steven J. Simske**

(HP Labs, Fort Collins, USA  
steven.simske@hp.com)

**Marcelo Thielo**

(HP Labs, Porto Alegre, Brazil  
marcelo.resende.thielo@hp.com)

**Abstract:** Portable digital cameras are being used widely by students and professionals in different fields as a practical way to digitize documents. Tools such as PhotoDoc enable the batch processing of such documents, performing automatic border removal and perspective correction. A PhotoDoc processed document and a scanned one look very similar to the human eye if both are in true color. However, if one tries to automatically binarize a batch of documents digitized from portable cameras compared to scanners, they have different features. The knowledge of their source is fundamental for successful processing. This paper presents a classification strategy to distinguish between scanned and photographed documents. Over 16,000 documents were tested with a correct classification rate of over 99.96%.

**Keywords:** MPEG-7, content-based Multimedia Retrieval, Hypermedia systems, Web-based services, XML, Semantic Web, Multimedia

**Categories:** H.3.1, H.3.2, H.3.3, H.3.7, H.5.1

### **1 Introduction**

Portable digital cameras are ubiquitous. Either in standalone versions, or incorporated in cell phones, the quality of the images has risen at a fast pace while their price has dropped drastically. Such pervasiveness has given rise to unforeseen applications such as using portable digital cameras for digitalizing documents by users of many different professional areas. For instance, students and professionals are taking photos of writing boards instead of taking notes; lawyers are taking photos of legal processes instead of going through a difficult bureaucratic path to take documents out of court to photocopy them, etc. This new research area [Doermann, 03] [Liang, 05] is evolving fast in many directions. People in general are non-specialized in image processing and claim for new algorithms, tools and processing environments to be able to provide simple and user-friendly ways of visualizing, printing, transcribing, compressing, storing and transmitting document images. Figure 1 presents an example of a document acquired with a portable digital camera. Reference [Lins, 07] points out some particular problems that arise in this document digitalization process:

the first is background removal. Very often the document photograph goes beyond the document size and incorporates parts of the area that served as mechanical support for taking the photo of the document. The second problem is due to the skew often found in the image in relation to the photograph axes. As portable cameras have no fixed mechanical support, often there is some inclination in the document image. The third problem is non-frontal perspective, due to the same reasons that give rise to skew. A fourth problem is caused by the distortion of the lens of the camera. This means that the perspective distortion is not a straight line, but a convex arc, depending on the quality of the lens and the relative position of the camera and the document. The fifth difficulty in processing document images acquired with portable cameras is non-uniform illumination.

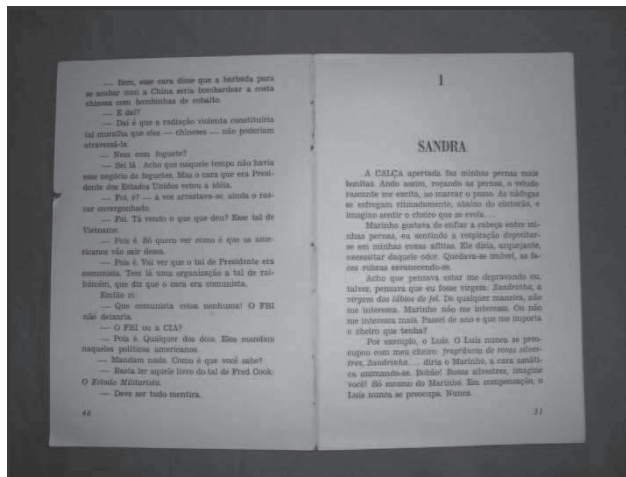


Figure 1: Example of a photo document

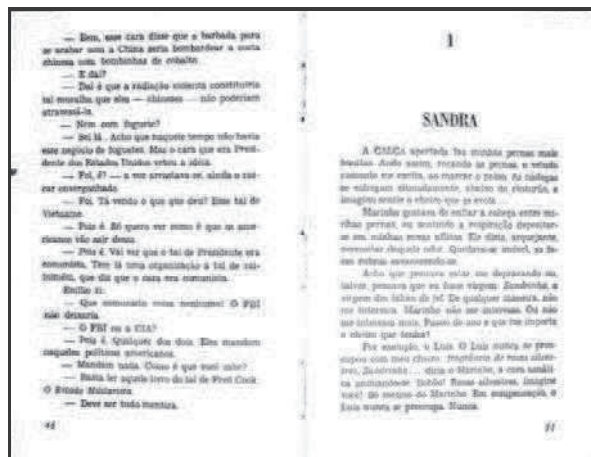


Figure 2: PhotoDoc processed photo document

Reference [Silva, 07] presents PhotoDoc, a freely available toolbox for processing document images acquired with portable digital cameras, which is implemented as a plugin in ImageJ [ImageJ, 09]. Figure 2 presents an example of a photo document processed with PhotoDoc, which is implemented as a Plugin in ImageJ [ImageJ, 09].

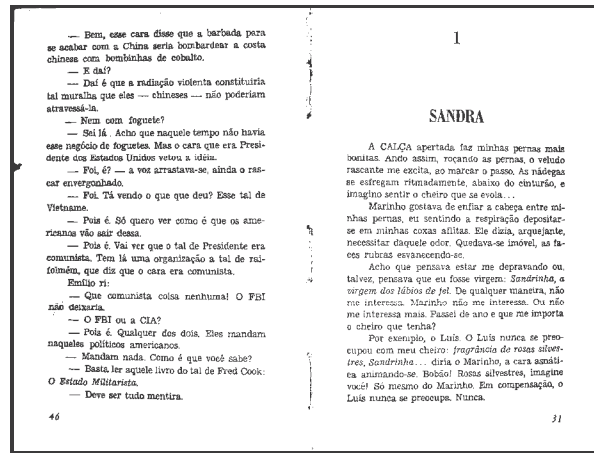


Figure 3: Binarization of a photo document using a global algorithm [Otsu, 79]

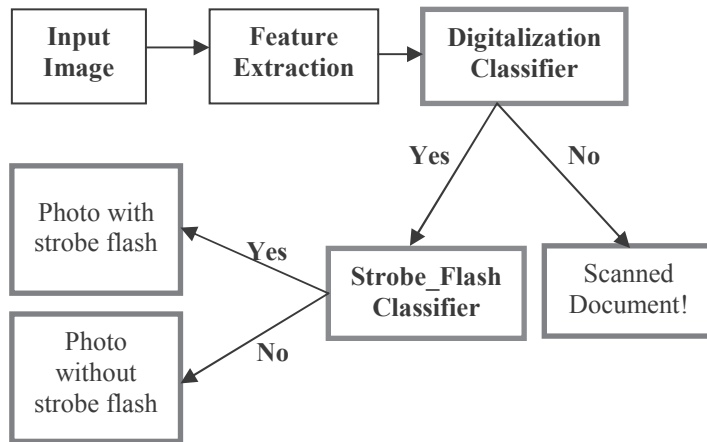
Illumination is less uniform for documents captured with digital cameras in comparison to scanned images. It may not be easy for a person to differentiate between a document processed using PhotoDoc and the same document captured with a scanner. Distinguishing between them is important in the case, for example, of image binarization. The irregular illumination in general tends to provide shaded black areas in the direct binarization of a photo document as shown in Figure 3. Color images such as the one in Figure 4, both scanned and photographed, are also present in the test set used here.

Once the digitalization device for a given image is known, one has valuable information about the nature of the possible noises present in the captured image. According to the taxonomy proposed in [Lins, 09a] the kinds of noises present in scanned documents are *physical* noises (considering an adequate scanner manipulation and its perfect functioning). Physical noises, such as stains, folding marks, annotations, etc. may be difficult to be removed, and some may even consider them as part of the document information. On the other hand, photographed documents, besides being passive of having the same noises as the scanned ones, may include the *digitalization* noises that if known their existence is known a priori may be suitably removed. One of the few digitalization noises one finds in scanned documents is found in the digitalization of bound volumes in flatbed scanners, as the distance from the object to the flatbed caused a document warping. Also, in this case, it is of paramount importance the information of how the document image was obtained as different de-warping algorithms are used depending on the digitalization source.



Figure 4: PhotoDoc processed color photo document

This paper focuses on a classification strategy to distinguish, in a batch of documents, the scanned documents from documents acquired with portable digital cameras. Camera documents are further classified based on whether a strobe flash was used, as shown below:



The classification strategy depends on the following:

- The choice of the set of features to be extracted. The features selected must provide enough elements to distinguish between the clusters of interest. Feature extraction has also impact in classification time.

- The choice of the classifier. Some classifiers are able to perform better than some others depending on the nature and class of the problem, the representativeness of the features selected, etc.
- The quality and size of the training set used for the classifier. The training set must be carefully chosen to encompass the whole diversity of the universe of objects to be classified, with as less redundancy as possible.

This paper shows that the classifier presented in reference [Lins, 09] presents excellent performance for distinguishing between documents obtained from scanners and portable digital cameras with or without the strobe flash on. The results obtained are compared with the classification strategy in reference [Lins, 09a]. The new classifier not only reached a higher correct classification rate, but besides that, elapsed much less time for feature extraction and classification. The classifier presented herein was implemented using Weka [Witten, 05] [Weka, 09], an excellent, user friendly and open-source platform developed at the University of Waikato. The test set encompassed 17,781 documents of which only 3 documents were misclassified, yielding a correct classification rate of 99.98%.

## 2 Experiments Performed

The starting point for this work was collecting images that are representative of the two different clusters of interest: scanned and photographed documents. The photographed documents were split in two sub-clusters: images acquired with and without the strobe flash on.

The test set for the photo document cluster is formed by 9,573 documents acquired with a Sony Cybershot digital camera DSC-W55 in 5 and 7.2 Mpixels, with and without mechanical support, in-built strobe flash on and off. In the camera set there are also 404 photos taken with a portable camera Sony DSC-S40 and 60 photos from a cell phone LG Shine ME970, both without any mechanical support. All photo documents were processed with PhotoDoc [Silva, 07] a photo document processing tool that crops the framing border and corrects perspective and skew, should be classified as "document".

The 6,444 of the scanned documents were digitized with a Ricoh Afficio 1075 flatbed scanner in 100, 200 and 300 dpi saved into four different file formats: bmp (uncompressed), jpg (1% losses), png (lossless), and tiff (uncompressed), using the software provided by the scanner manufacturer. Although the jpeg file format may be seen as unsuitable for such kind of image it is often used by people in general [Lins, 04]. In addition, 300 images were acquired with a scanner HP 5300c in 300 dpi, true color, stored in tiff (uncompressed) and 1000 jpeg images in different resolutions were collected from the Internet.

Table 1 shows the numbers of images per file format in the test set.

	JPG	PNG	TIFF	BMP	Total
Photo	10,037	***	***	***	10,037
Scanned	2,611	1,611	2,522	1,000	7,744
Total	12,648	1,611	2,522	1,000	17,821

Table 1: Images per file formats

## 2.1 Features Tested

The choice of the features to be extracted and tested is the key to the success and performance of the classification. Image entropy is often used as the key for classification [Simske, 05]. It has a large computational cost, however. Entropy calculation demands a scan in the image to calculate the relative frequency of a given color, for instance, which is then multiplied for its logarithm and added up. The classifier described in reference [Simske, 05] is based on the binary classification approach, and assumes a Gaussian distribution for each of the features. Its performance degrades in proportion to the non-Gaussian nature of the data. We designate this the entropy-based classifier, as the set of features chosen herein has entropy calculation as its key.

The work presented in reference [Lins, 09] proposes a new classification strategy that assumes that decreasing the gamut of an image, analyzed together with its grey scale and monochromatic equivalents would provide enough elements for a fast and efficient image classification. The features tested are:

- Palette (true-color/grayscale)
- Gamut
- Conversion into Grayscale
- Gamut in Grayscale (if RGB)
- Conversion into Binary (Otsu)
- Number of black pixels in binary image.
- $(\#Black\_pixels/Total\_#\_pixels)*100\%$
- $(Gamut/Palette)*100\%$  (true-color/grayscale)

Image binarization is performed by using Otsu [Otsu, 79] algorithm. The data above are extracted for each image and placed in a vector of features. The classification strategy adopted herein follows the feature set proposed in reference [Lins, 09]. The training set used had size of about 8% of the test set and was selected from within the images of Sony Cybershot digital camera DSC-W55 in 5 and 7.2 Mpixel and the Ricoh Afficio 1075 flatbed scanner in 100, 200 and 300 dpi. The images in the training set were not part of the test set. The entropy-based classifier [Simske, 05] was used to compare the results obtained. Both classifiers used the same training and test sets.



Classifier		Photo +sf	Photo -sf	Scanned	Accuracy %
Random Forest 5-trees	Photo +sf	4029	0	0	100
	Photo -sf	4	5534	6	99,81962
	Scanned	0	0	6444	100
Random Forest 10-trees	Photo +sf	4,029	0	0	100
	Photo -sf	4	5,537	3	99.8737
	Scanned	0	0	6,444	100
Random Forest 15-trees	Photo +sf	4029	0	0	100
	Photo -sf	7	5535	2	99,83766
	Scanned	0	0	6444	100
Random Forest 20-trees	Photo +sf	4029	0	0	100
	Photo -sf	8	5534	2	99,81962
	Scanned	0	0	6444	100
Random Forest 100-trees	Photo +sf	4029	0	0	100
	Photo -sf	7	5535	2	99,83766
	Scanned	0	0	6444	100
MLP	Photo +sf	4029	0	0	100
	Photo -sf	13	5531	0	99,76551
	Scanned	0	1	6443	99,98448
RBF	Photo +sf	3975	54	0	98,65972
	Photo -sf	47	5497	0	99,15224
	Scanned	0	5	6439	99,92241

Table 2: Confusion matrix of the proposed classifier with 16,017 original images

Table 3 shows the results obtained for the same set of classifiers trained and tested with sub-sampled images.



Classifier		Photo +sf	Photo -sf	Scanned	Accuracy %
Random Forest 5-trees	Photo +sf	4029	0	0	100
	Photo -sf	4	5534	6	99,81962
	Scanned	4029	0	0	100
Random Forest 10-trees	Photo +sf	2	5525	17	99,65729
	Photo -sf	0	0	6444	100
	Scanned	4,029	0	0	100
Random Forest 15-trees	Photo +sf	0	5,540	4	99,9278
	Photo -sf	0	0	6,444	100
	Scanned	4029	0	0	100
Random Forest 20-trees	Photo +sf	2	5539	3	99,90981
	Photo -sf	0	0	6444	100
	Scanned	4029	0	0	100
Random Forest 100-trees	Photo +sf	2	5539	3	99,90981
	Photo -sf	0	0	6444	100
	Scanned	4029	0	0	100
MLP	Photo +sf	3	5539	2	99,90981
	Photo -sf	0	0	6444	100
	Scanned	4029	0	0	100
RBF	Photo +sf	3971	58	0	98,56043
	Photo -sf	48	5496	0	99,13419
	Scanned	0	5	6439	99,92240

Table 3: Confusion matrix of the proposed classifiers with 16,017 subsampled images

Using sub-sampling, the relative performance of the classifiers was stable. Again, Random-forests with 10 trees provided the best results. Curiously, sub-sampling, besides speeding-up the feature extraction time, increased correct classification rate. One important point worth noting is that the misclassified documents, when binarized using a global algorithm, performed satisfactorily. Having the strobe flash off may resemble a scanned document, provided there is enough uniform illumination from the environment. Then, the misclassification errors in this case do not cause serious problems to the overall process.

Now, the entropy-based set of features for classification proposed by reference [Simske, 05] was tested on the original data and the results obtained are presented on Table 4.

<b>Proposed Classifier</b>	Photo +sf	Photo -sf	Scanned	Accuracy
Photo +sf	3402	272	355	84.4378 %
Photo -sf	71	4466	1007	80.5555 %
Scanned	32	152	6260	97.1446 %

*Table 4: Confusion matrix of the entropy-based classifier with original images*

The results obtained for entropy based classifier with subsampled images are shown on Table 5.

<b>Proposed Classifier</b>	Photo +sf	Photo -sf	Scanned	Accuracy
Photo +sf	3402	270	357	84.4378 %
Photo -sf	69	4562	913	82.2871 %
Scanned	24	158	6262	97.1756 %

*Table 5: Confusion matrix of the entropy-based classifier with subsampled images*

The comparison between the entropy-based and the new one proposed here shows that the new one is about 10% better than the previous one.

The classification of the 404 photos taken with a portable camera Sony DSC-S40 and 60 photos from a cell phone LG Shine ME970, both without any mechanical support, and the images obtained with scanner HP 5300c and the images collected from the Internet did not bring any misclassification at all.

#### 4 Time Performance

Table 6 presents the feature extraction and classification times along with the programming language used for implementation. Besides classification accuracy per cluster, the average feature extraction and classification times are presented. One

should also remark that there is a difference in time scale between feature extraction and classification.

	Feature extraction		Classification	
	Time (s)	Language	Time (ms)	Language
<b>Original</b>	<b>0.4174</b>	<b>C++</b>	<b>0.12</b>	<b>C#</b>
<b>Subsampled</b>	<b>0.1470</b>	<b>C++</b>	<b>0.12</b>	<b>C#</b>
<b>Original</b>	<b>0.4174</b>	<b>C++</b>	<b>0.10</b>	<b>C++</b>
<b>Subsampled</b>	<b>0.1470</b>	<b>C++</b>	<b>0.10</b>	<b>C++</b>
<b>Entropy Or.</b>	<b>1.4576</b>	<b>C#</b>	<b>6.13</b>	<b>C#</b>
<b>Entropy Ss.</b>	<b>0.497</b>	<b>C#</b>	<b>6.13</b>	<b>C#</b>

Table 6: Feature extraction and classification times

Table 6 shows that the set of features used for image classification based on image palette conversion outperforms the entropy-based classifier by a factor of four for feature extraction and by a factor of fifty for image classification. ("Entropy Or." stands for the Entropy-based classifier [Simske, 05] with the original images, while "Entropy Ss." corresponds to the Entropy-based classifier with subsampled images).

The figures of the relative performance of the classifiers for the proposed set of features varying the number of trees and the MLP implemented in Weka (Java) are shown on Table 7.

Proposed Classifier	Java -Time (ms)	C++ - Time (ms)
<b>Random Forest 5-trees</b>	5.4	3.7
<b>Random Forest 10-trees</b>	6.1	5.0
<b>Random Forest 15-trees</b>	6.7	5.1
<b>Random Forest 20-trees</b>	7.9	6.4
<b>Random Forest 100-trees</b>	9.5	6.9
<b>MLP</b>	6.8	***

Table 07: Classification times in Random Forest [Breiman, 01]

One may observe that the Random-forests classifier reaches the best trade-off classification and time efficiency.

## 5 Conclusions

Weka [Witten, 05] [Weka, 09] has shown to be an excellent test bed for statistical analysis. The choice for a Random tree classifier was made after performing several experiments with the large number of alternatives offered by Weka, although results did not vary widely. Amongst them a preliminary comparison between the new statistical classifier proposed here and a MLP neural classifier provided worse results (around 94% of accuracy).

The choice of the images in the training set is of paramount importance to the performance of the classifier. They must be representative of the whole universe of images in a cluster.

The classification scheme presented in this paper increased the correct classification rate by more than 10%. This automatic classification allows distinguishing scanned from photographed document images yielding better ways to suitably process document images.

### Acknowledgements

Research presented herein was partly sponsored by CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, and HP - UFPE Project TechDoc sponsored by MCT, both of the Brazilian Government.

### References

- [Breiman, 01] Breiman, L.: Random Forests, *Machine Learning*, 45(1), pp. 5-32, 2001.
- [Doermann, 03] Doermann D. and Liang J., Li H.: Progress in Camera-Based Document Image Analysis, *ICDAR'03*, Volume (1): 606, 2003.
- [Silva, 07] Silva, G.P and Lins, R.D.: PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. *CBDAR 2007*, pp.107-114, 2007.
- [Liang, 05] Liang, J., Doermann, D., and Li, H.: Camera-Based Analysis of Text and Documents: A Survey. *International Journal on Document Analysis and Recognition*, 2005.
- [Lins, 04] Lins, R.D. and Machado, D.S.A.: A Comparative Study of File Formats for Image Storage and Transmission, vol. 13(1):175-183, *Journal of Electronic Imaging*, 2004.
- [Lins, 07] Lins, R.D., Gomes, A.R. e Silva and Silva, G.P.: Enhancing Document Images Acquired Using Portable Digital Cameras, *ICIAR'07*, LNCS 4633, pp. 1229-1241, Springer-Verlag, 2007.
- [Lins, 09] Lins, R.D., Silva, G.P, Simske, S.J., Fan,J., Shaw, M.S., Sá, P., Thiello,M.R.: Image Classification to Improve Printing Quality of Mixed-Type Documents. *ICDAR 2009*, IAPR Press, Barcelona, 2009.
- [Lins, 09a] Lins, R.D.: A taxonomy for noises in paper documents – the physical noises, LNCS 5624, pp. 844-854, Springer Verlag, 2009.
- [Otsu, 79] Otsu, N.: A threshold selection method from gray level histograms. *IEEETrans.Syst.Man Cybern.* Vol. (9):62-66, 1979.
- [Simske, 05] Simske, S.J.: Low-resolution photo/drawing classification: metrics, method and archiving optimization, *Proceedings IEEE ICIP*, IEEE, Genoa, Italy, pp. 534-537, 2005.
- [Witten, 05] Witten, I.H. , Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition)* -- Morgan Kaufmann, June 2005. ISBN 0-12-088407-0.
- [ImageJ, 09] ImageJ <http://rsb.info.nih.gov/ij/>; last visited 09.02.2010.
- [Weka, 09] Weka 3: Data Mining Software in Java, website <http://www.cs.waikato.ac.nz/ml/weka/>; last visited 09.02.2010.

## Automatically Discriminating between Digital and Scanned Photographs

Rafael Dueire Lins  
Gabriel de França Pereira e Silva  
Universidade Federal de Pernambuco  
Recife, Brazil  
{gabriel.psilva, rdl}@ufpe.br

Steven J. Simske  
Hewlett-Packard Labs  
Fort Collins, USA  
steven.simske@hp.com

**Abstract** — True digital photos and the digital images of scanned photographs have very different properties. The illumination pattern and palette of the two kinds of images are different. Being able to distinguish between them is important, as each of these should be handled during printing with a class-specific pipeline of image transformation algorithms, and misclassification results in detrimental imaging effects. This paper presents an automatic classifier to discriminate between the two sources. The classifier proposed is fast enough to be embedded in the driver of any printing device today.

**Keywords**- digital photo, analogical photos, printing.

### I. INTRODUCTION

Color perception goes beyond the psycho-physical phenomenon usually described in the literature. Cultural elements also influence the way people see printed images. One typical example of that is photo printing – the use of a color palette in place of a richer set of hues looks unpleasantly flat and pale. People expect photos to have sharp bright colors, rich in different hues, while keeping an overall rich color balance. Figure 1 presents an image which was obtained by scanning a printed analogical photograph (saved in JPEG) with a resolution of 600 dots per inch, while Figure 02 exhibits an image of a digital photograph. The difference between the two images is easily observable. The scanned photograph looks much “paler” than the digital photograph, but whoever scans and reprints a photograph expects it to look as “sharp and bright” as the digital one.



Figure 1. Scanned photo 600 dpi – 3.57 Mbytes – JPEG

Functional image classification is the assignment of different image types to separate classes to optimize their rendering for printing or another specific end task, and is an important area of research in the publishing and industries. To meet customer expectations, the printer needs to print each image with the correct color palette, balance and other image processing operations applied. To perform this task automatically in the absence of image metadata, the printer must perform accurate image classification based solely on the image raster information. This classification must be both accurate and fast due to the constraints of the printer embedded processor.

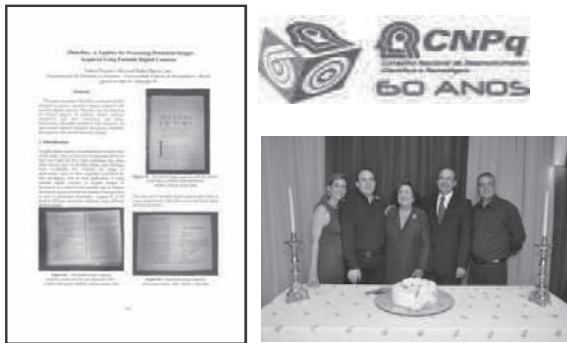
Image classification is used in all-in-one and multi-functional devices to differentially render images belonging to different clusters. In particular, document, photo and logo images require widely different imaging pipelines to optimize their appearance when copied or printed. Documents (text, tables), for example, require sharpening that would damage the appearance of photos and logos. Logos use a palette that would “posterize” photos. Photos, in turn, can be rendered with a lower resolution (but greater bit depth) than either documents or logos. Reference [6] presents an image classifier that replaced the previous one embedded in HP printers. The new classifier [6] largely outperformed the previous one [11] both in accuracy and time performance, an important feature for an algorithm to be embedded in low-cost, low-power consuming, fast printing devices. The present paper introduces a new classifier for photos: whether they are digital or from a printed (hardcopy) source which was later digitized.



Figure 2. Digital photo taken with a Sony Cyber-shot 7.2 MPixels portable camera. 3.01 MBytes - JPEG

## II. MOTIVATION

Image clustering [2][3] has a long tradition in the database community for efficient information retrieval from image databases [4][8]. The classifier described in reference [6] is able to discriminate with over ninety percent accuracy between three clusters: documents, logo and photos. The document clusters include scanned, digitally generated (such as image files from pdf files), and photographed ones (processed through PhotoDoc, a software platform that removes borders, corrects perspective and skew, etc). The logo cluster includes color and monochromatic images. The photo cluster covers a wide range of images varying in palette (color, sepia, and monochromatic) and theme (landscapes, people, objects, and PhotoDoc unprocessed documents). Figure 3 presents examples of images in the three clusters.



**Figure 03.** Examples images of the different clusters discriminated by the classifier in [6]. Left-document, Right\_T-logo, Right\_B-photo

It is extremely difficult to an observer to distinguish between a scanned and PhotoDoc processed document image, until one tries to binarize them. The illumination pattern of the photographed document, although imperceptible to the naked eye, is non-uniform and the direct binarization using a global algorithm leads to some black areas as may be observed in Figure 04.



**Figure 4.** Binarization of a photo document using Otsu global algorithm [7].

Thus, for batch processing such images an image classifier to discriminate between the two sources is most desirable and this was the motivation for the work reported in [10], which also served as inspiration to the current one.

## III. METHODOLOGY

The “Photo” cluster in [6] encompassed many different sorts of photos, which ranged from people (approximately 4,000), landscapes (about 3,700), objects (just under 400) and even documents (500) in different file formats (7,476 JPEG, 35 TIFF, and 457 BMP) and varied from true-color to grey scale ones. The resolution also varied widely from VGA (480x640 pixels) to 7.2 Mpixels. The photos were collected from family albums of the people linked to the authors to ones obtained from the Internet.

The current study is far more restrictive and limited the test set of to only one theme – people. The starting point for this work was scanning a set of photos from a family photo album. All photos were printed in 10 x 15 cm on glossy paper without texture at a professional printing house. They were scanned with a 600 dots per inch resolution with an HP flatbed scanner model ScanJet 5300C. The photos were stored in JPEG file format with 1% loss, the standard used by portable digital cameras [5]. The choice of the resolution adopted was such as the size of the scanned photo was similar to the size of the photographed ones. The photos were taken with a Sony Cyber-shot digital camera DSC-W55 in 5 and 7.2 Mpixels and a Sony Cyber-shot DSC-T10 7.2 MPixels portable cameras. Table 1 presents some of the features of the test images.

JPG	Average Size (MB)	Variance (MB)	Total Number
Photo	3.18	0.10	241
Scanned	3.48	0.08	95

**Table 1.** Data of dimension of the test set and the size and variance of images (in JPEG)

The last cluster of images in the classifier in [6] is “Don’t Know” images (unassigned). These 529 images were included as to increase the possibility of misclassifications. They are images that appear in the “real world” and range widely in nature from biological images, to vector graphics (obtained by softwares such as Excell®, Powerpoint®, etc.), of which 202 are JPEG and 327 in BMP.

### A. The Classifier

The choice of the features to be extracted and tested is the key to the success and performance of the classification. Image entropy is often used as the key for classification [11]. It has a large computational cost, however. Entropy calculation demands a scan in the image to calculate the relative frequency of a given

color, for instance, which is then multiplied for its logarithm and added up.

The classifier in reference [6] assumed that decreasing the gamut of an image, analyzed together with its grey scale and monochromatic equivalents would provide enough elements for a fast and efficient image classification. The features tested are:

- Palette (true-color/grayscale)
- Gamut
- Conversion into Grayscale (if RGB)
- Gamut in Grayscale (if RGB)
- Conversion into Binary (Otsu)
- Number of black pixels in binary image.
- (#Black\_pixels/Total\_#\_pixels)\*100%
- (Gamut/Palette)\*100% (true-color/grayscale)

Image binarization is performed by using Otsu [7] algorithm. The data above are extracted for each image and placed in a vector of features.

The classifier “architecture” is made by cascading binary classifiers. The order they appear has an effect on the final classification accuracy. Figure 5 shows the way they are cascaded.

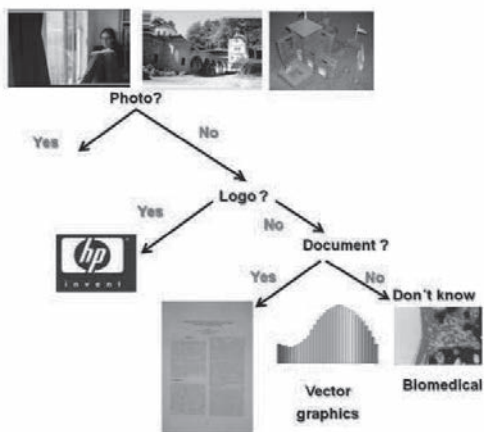


Figure 5. Cascaded binary classifier “architecture”.

### B. Recognizing test images as Photos

We wish to use the classifier presented here to refine the discriminator in [6], thus the first test performed was to submit the test images directly to that classifier and analyze to see if it was able to correctly recognize the images as photos. The test set was fed directly to the general (photo-document-logo) classifier without any training or tuning. The result obtained may be found in the confusion matrix presented in Table 2.

General	Photo	Logo	Doc	DK	Accuracy
Scanned Photos	91	0	4	0	95.7%
Digital Photos	268	3	5	5	95.3%
Total	359	3	9	5	95.3%

Table 2 – Confusion matrix of the general classifier with Scanned/Digital Photo test set

As one may observe from Table 2, the general classifier was able to correctly recognize 95.3 % of the photos in the Scanned/Digital test set as belonging to the cluster “Photo”. Five of the Digital photos were misclassified as “Don’t Know” (DK).

### C. Sub-sampling

Time performance is of paramount importance for embedded software such as an image classifier to run on printing devices. Image sub-sampling may be used as a way to reduce the time elapsed in feature extraction of images to be classified. The key points in image sub-sampling are:

- 1- The larger the image files, the richer in data redundancy; thus, if the redundant data are thrown away the efficiency both in feature-collection time and classification may rise.
- 2- The selection of points to be analyzed for feature collection should not be random. It should somehow provide a "reduced" version of the original image (although in some cases it may be distorted by unequal scaling!).

Twenty different sub sampling strategies were evaluated in [6]. The cascaded sub-sampling strategy consisted of removing more points from the larger image files and provided the best overall accuracy of any classification schema. The pseudo-code for the cascaded sub-sampler is shown below:

```

size = height*width
• If size ≤ 300,000 break;
• If 300,000 < size ≤ 500,000:
  remove even lines or columns
  (whatever the larger);
• If 500,000 < size ≤ 700,000:
  remove even lines and columns;
• If 700,000 < size ≤ 900,000:
  remove 2 lines in every 3 lines and even columns,
  (if height>width)
  remove even lines and
  2 columns in every 3 columns, otherwise;
• If 900,000 < size remove 2 lines and 2 columns
  in every 3 lines and columns;
  
```

**Code of the “cascaded” sub-sampler**

Table 3 presents the results for the General classifier for the sub-sampled images in the Scanned/Digital Photo test set.

S sampled	Photo	Logo	Document	DK	Accuracy
Scanned Photos	90	0	4	1	94.7%
Digital Photos	268	3	4	6	95.3%
Total	359	3	8	7	95.0%

Table 3 – Confusion matrix of the general classifier with sub-sampled Scanned/Digital Photo test set

The data presented in the confusion matrix shown in Table 3 supports the conclusion that the sub-sampling procedure proposed marginal degrades the performance of the classifier. Later on, it will be shown that the performance gain largely compensates the accuracy degradation.

#### D. Training and test sets

To increase the difficulty of discriminating between scanned and digital photos, the images chosen are as similar in theme (people) and size as possible. All images are in color and stored in JPEG file format. Table 1 summarizes some of the features of the images in the test set. The training set was carefully selected to guarantee the diversity of the images in the test set, having in mind that quality matters more than size. Table 4 presents the relative size of the training and test sets.

	Test	Training	%
Scanned Photos	91	31	34.06
Digital Photos	281	60	21.35
<b>Total</b>	<b>372</b>	<b>91</b>	<b>24.46</b>

**Table 4 – Sizes of Training x Test sets**

The Weka [12] classification strategy used was the Random Forests (number of trees equal to 10) [1].

#### E. Results

This section presents the results of clustering of the images in the Scanned/Digital photo test set after the classifier was specially trained for discriminating between scanned and digital photos.

Scanned/Digital	Scanned Photos	Digital Photos	Accuracy
Scanned Photos	95	0	100.00%
Digital Photos	1	280	99.99%
<b>Total</b>	<b>96</b>	<b>280</b>	<b>99.99%</b>

**Table 4 – Confusion matrix of the Scanned/ Digital classifier with original images**

Table 4 shows that the results obtained for the original images are extremely good with accuracy close to 100%.

Scanned/Digital	Scanned Photos	Digital Photos	Accuracy
Scanned Photos	95	0	100.00%
Digital Photos	1	280	99.99%
<b>Total</b>	<b>96</b>	<b>280</b>	<b>99.99%</b>

**Table 5 – Confusion matrix of the Scanned/ Digital classifier with sub-sampled images**

Image sub-sampling, as demonstrated by the data that are shown in Table 5, does not introduce any detrimental effects on image classification, and brings performance gains to the

feature extraction phase of the classifier. In both cases, exactly one digital photo was misclassified as being a scanned photo either original or in the sub-sampled case. That artistic photo, shown in Figure 6, has a complex illumination pattern that makes difficult its correct classification.



Figure 6. Misclassified digital photo

#### IV. FURTHER IMPROVEMENTS

Analyzing the results presented in Tables 2 and 3 one may observe that the general (photo-logo-document) classifier is less accurate than the scanned-digital photo classifier for the same test set. In particular, the number of images that were not classified and thus left in the Don't know (DK) set is not negligible. It is also important to note that the digital photo in Figure 6 when assigned by the general classifier was inserted in the Don't know set. If one observes the classifier architecture from Figure 5 and includes the new Scanned/Digital photo classifier, one may re-structure it to feed-back the Don't know cluster into the Scanned/Digital photo classifier, yielding the architecture shown in Figure 6.

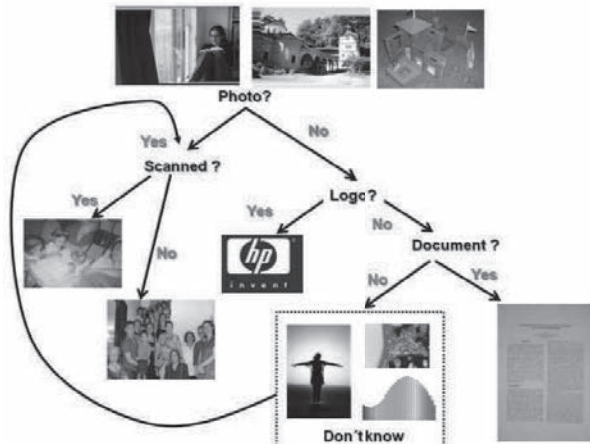


Figure 6. New classifier "architecture"



The new architecture proposed will raise the accuracy of the general classifier from 93.3% to 96.4%, as may be seen in Table 6 for the original images.

General	Photo	Logo	Document	DK	Accuracy
Scanned Photos	91	0	4	0	95.7%
Digital Photos	272	3	5	1	96.7%
Total	363	3	9	1	96.4%

**Table 6** – Confusion matrix of the **new architecture** general classifier with Scanned/Digital Photo test set.

Table 7 shows the results of the new architecture for the sub-sampled images increasing the overall accuracy.

S sampled	Photo	Logo	Document	DK	Accuracy
Scanned Photos	91	0	4	0	95.7%
Digital Photos	273	3	4	1	97.1%
Total	364	3	8	1	96.7%

**Table 7** – Confusion matrix of the **new architecture** general classifier with sub-sampled images.

## V. TIME PERFORMANCE

Table 8 presents the feature extraction and classification times together with information about the language those procedures were implemented into. Besides classification accuracy per cluster, the average feature extraction and classification times are presented. Note that there is a difference in time scale between feature extraction and classification.

	Feature extraction		Classification	
	Time (s)	Language	Time (ms)	Language
<b>Original</b>	<b>0.4382</b>	<b>C++</b>	<b>0.10</b>	<b>C#</b>
<b>Sub-sampled</b>	<b>0.1502</b>	<b>C++</b>	<b>0.10</b>	<b>C#</b>

**Table 8** – Feature extraction and classification times Processor Pentium IV 2.4GHZ 2GB RAM

The performance results presented shows that sub-sampling reduces the feature extraction time to one third of that needed for the original images. In some cases, as in the test data set here, sub-sampling also yielded an increase in accuracy.

## VI. DISCUSSION AND CONCLUSIONS

Scanned and Digital photos have different features. For cultural reasons, people today are more acquainted with the

way digital photos look with sharp bright colors than what one tends to get from scanning printer photos. The correct classification allows the printer to automatically meet the users' expectations.

Weka [8], as in previous research [6], proved an excellent test bed for statistical analysis. The choice of the Random tree classifier [1] was made after performing several experiments with the large number of alternatives offered by Weka, although results did not vary widely. In the current case of the scanned/digital photo classifier, in opposition to the results of [6], the choice of the images in the training set was not of paramount importance to the performance of the classifier.

The new classifier "architecture" proposed here, besides improving the appearance of the printed output in the case of scanned and digital photos, also benefits the overall classification accuracy. It is important to note that the classifier "architecture" with feedback presented in this paper opens a new way of using binary classifiers for multiple classification.

## VII. REFERENCES

- [1] L. Breiman, "Random Forests", *Machine Learning*, 45(1), pp. 5-32, 2001.
- [2] H. Frigui and R. Krishnapuram. Clustering by competitive agglomeration. *P. Recognition*, 30(7), 2001.
- [3] M. A. Hearst and J. O. Pedersen. Reexamining the Cluster Hypothesis: Scatter Gathet on Retrieval Results, SIGIR, 1996.
- [4] S. Krishnamachari and M. Abdel-Mottaleb. Image Browsing using Hierarchical Clustering, IEEE Symposium on Computers and Communications, ISCC'99, July 99.
- [5] R. D. Lins and D. S. A. Machado, A Comparative Study of File Formats for Image Storage and Trans., v13(1):175-183, *Journal of Electronic Imaging*, 2004.
- [6] R. D. Lins, G. F. P. e Silva, B. S.J. Simske, J. Fan, M. Shaw, P. Sá, M. Thielo. Image Classification to Improve Printing Quality of Mixed-Type Documents, ICDAR 2009. IEEE Press, 2009. p.1106 - 1110.
- [7] N. Otsu. "A threshold selection method from gray level histograms". *IEEE Trans. Syst. Man Cybern.* v(9):62-66, 1979.
- [8] P. Scheunders. Comparison of Clustering Algorithms Applied to Color Image Quantization, *Patt. Recog. Letters*, v18(11-13):1379-1384, 1997.
- [9] G. F. P e Silva and R. D. Lins. PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. *CB DAR '2007*, pp.107-114, 2007.
- [10] G. F. P. e Silva, R. D. Lins, B. Miro, S.J. Simske, M. Thielo, Automatically Deciding if a Document was Scanned or Photographed. *Journal of Universal Computer Science.*, v.15, p.3364 - 3366, 2009.
- [11] S.J. Simske, "Low-resolution photo/drawing classification: metrics, method and archiving optimization," *Proceedings IEEE ICIP*, IEEE, Genoa, Italy, pp. 534-537, 2005.
- [12] Weka 3: Data Mining Software in Java, website <http://www.cs.waikato.ac.nz/ml/weka/>.

### **A.3 Publicações sobre Classificação de Ruído**

(SILVA et al., 2010b) - G. F. P. Silva, R. D. Lins, S. Banergee, A. Kuchibhotla, M Thielo. Automatically Detecting and Classifying Noises in Document Images. In: ACM-Symposium on Applied Computing, vol.1, pp: 33-39.

(SILVA et al., 2010c) - G. F. P. Silva e R. D. Lins; S. Banergee; A. Kuchibhotla; M Thielo. Enhancing the Filtering-out of the Back-to-Front Interference in Color Documents with a Neural Classifier. In: International Conference on Pattern Recognition, vol.1, pp: 2415-2419.

# Automatically Detecting and Classifying Noises in Document Images

Rafael Dueire Lins  
Gabriel Pereira e Silva

Universidade Federal de Pernambuco  
Recife - Pernambuco  
BRAZIL  
+55 81 2126-8210 x-241

{rdl, gabriel.psilva}@ufpe.br

Serene Banerjee  
Anjaneyulu Kuchibhotla

HP Labs.  
Bangalore  
INDIA

+91 80 2504-2238

{anji, serene.banerjee}@hp.com

Marcelo Thielo

HP Brazil R&D.  
Porto Alegre  
BRAZIL

+55 51 2121-3583

marcelo.resende.thielo@hp.com

## ABSTRACT

Image filtering to remove noise in document images follows two different approaches. The first one uses human classification of the noise present in an image for identifying a noise filter to use. The second approach is to blindly apply a batch of filters to an image. The former approach, although widely used, may insert noise in the filtering process due to the incorrect classification of the noise or even unsuitable filtering parameters. This paper presents a new paradigm for document image filtering. It aims at doing a more accurate and computationally efficient document cleanup by pre-characterizing the noise that is present in the document based on a set of human labeled training samples. The current focus of the project is on pre-characterization of the following types of noise: back-to-front interference or bleed through, skew and orientation, blur and framing.

## Categories and Subject Descriptors

I.4.9 [Image Processing and Computer Vision]: Applications.

## General Terms

Measurement, Documentation, Performance.

## Keywords

Noise characterization, documents, borders, skew, back-to-front interference, bleeding, show-through, orientation, classification.

## 1. INTRODUCTION

Finding ways to classify images and grouping them in sets of similar features has been researched by the database community for almost three decades aiming to make efficient information retrieval in image databases [1][2]. In such systems one image, known as a *query image*, is used to search the database looking for either the same or similar images. The basic idea is to try to organise the images in the database using some “common” features [3][2]. The same “features” are used to analyse the image

that will serve as the “search-key”. Instead of stepping through the whole database image-by-image, the retrieval process tries to match the properties of the search-key image with the different image clusters in the database. This largely reduces the search-space making the retrieval process far more efficient.

One of the features that achieved greater success in image retrieval was the analysis and clustering by using colour histograms [4], [5]. The semantics of images have also been used as a clustering method [06] in database retrieval. Images that have similar “motifs” are most likely to have properties that are common to each other forming clusters. On the other hand, images whose theme is completely uncorrelated should exhibit very different properties. Recent works [7][8] address the problem of image classification from the perspective of a printer that is fed with a raster file and classifies its content as belonging to one of the four clusters: photo, logo, document and complex. Such classification allows the printer to load color enhancement filters which yield better printing quality. Along the same research line of image classification, reference [9] attempts to identify the digitalization device of a given document deciding whether a document image was scanned or photographed. The photographed documents were processed using PhotoDoc [10], a tool developed for processing document images acquired with portable digital cameras. The latter group of images is further split into images acquired with and without the strobe flash on.

Document cleanup is important while scanning and copying documents. Current approaches for document cleanup typically are either based on human noise detection and filter selection or is performed automatically. They try to focus on certain types of cleanups and perform the filtering “blindly” on documents. The problem is that the latter strategy leads to inefficiency of document cleanup and, more importantly, could yield to degradation of the document image if the type of noise being cleaned up does not exist in the document or does not match the strength of the filter. If one could determine the type of noise and could do more intelligent document cleanup on the document it could enhance both the efficiency and the quality of the cleanup. Noise recognition, classification, understanding of its nature and strength is fundamental for suitable noise removal.

Noise characterization and even classification is a relatively new area of research [11]. The important point is to determine which features of a document should be used for noise characterization in a very efficient manner. Given a database of document images with ground truths that have been manually labeled indicating the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC'10, March 22-26, 2010, Sierre, Switzerland.

Copyright 2010 ACM 978-1-60558-638-0/10/03...\$10.00.

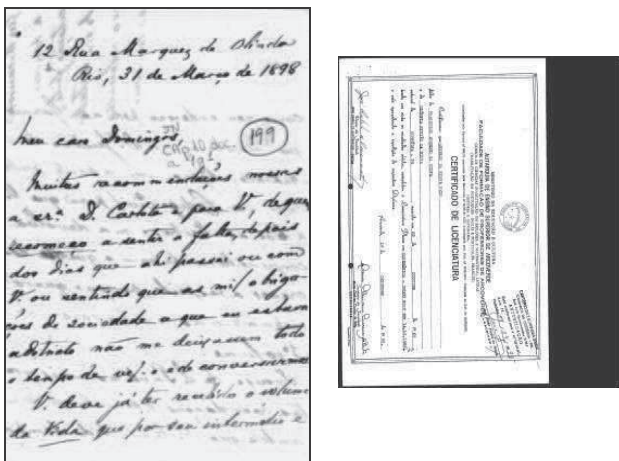
type of noise, can one then determine the classification of the noise for a document the system has not seen before given that:

- The document may have no noise (i.e. no noise should also be one class)
- The document may have more than one type of noise.

The classifier reported in this work is able to classify the existence of noise in a given document into the following six categories:

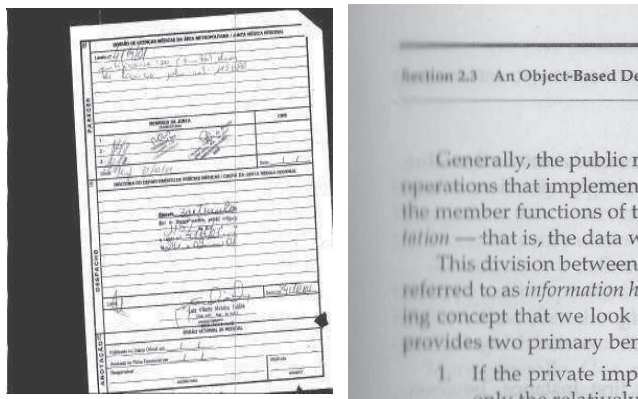
- Back-to-front interference (or bleed or show through)
- Frame or border noise
- Skew
- Orientation (0, 90, 180, 270 degrees)
- Blur
- No noise

Figure 01 presents some sample images with the different kinds of noise classified here. As one may observe, one often finds more than one kind of noise per image.



Back-to-front interference

Frame and orientation noises



Skew and black frame

Blur

Figure 1 – Document images with noises of interest

## 2. CLASSIFICATION STRATEGY

The architecture of the classifier is shown in Figure 4. The classifier used is Random Forest [14] which was implemented in Weka [12], an open source tool for statistical analysis developed at the University of Waikato, New Zealand. A number of features were extracted from each image to allow classification. The details of the training and test sets are provided in Table 01.

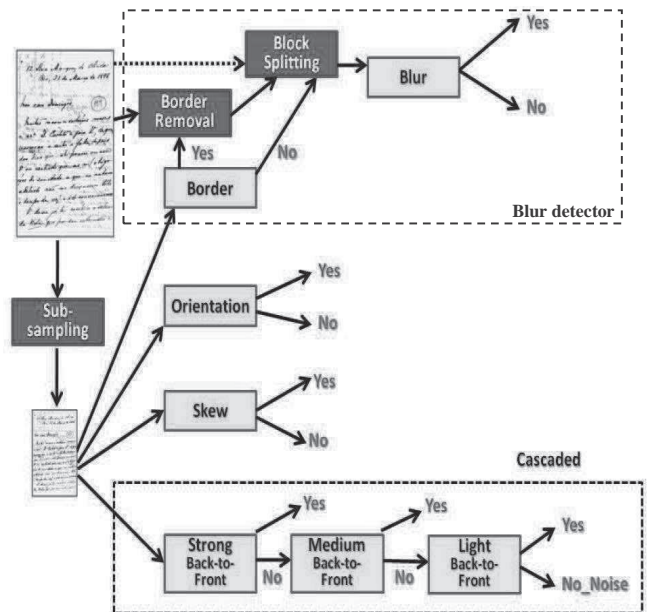


Figure 2 – Classifier architecture

The classifier developed herein works in parallel for the detection of the different clusters of noises. In the case of back-to-front interference the overall classifier is the result of cascading three classifiers that split the noise into strong, medium and light interference. Blur noise is seldom a global one. In general it affects areas of a document. In the case of Blur detection the classifier works with the aid of the Border noise detector. If the image has border noise it has to be removed prior to splitting the image in blocks.

Figure 3 presents a document image which was obtained from scanning a hard bound volume. In such image one may observe in the right hand side the existence of edge which interferes with the classifier responsible for detecting the presence (or absence) of blur. Figure 4 shows the same document of Figure 3 after the removal of its border by PhotoDoc [10]. The adoption of this pre-classification procedure provides a gain of about 11.2% to the classifier responsible for detecting the presence of the blur noise in images, since the images containing only framing noises are properly classified.

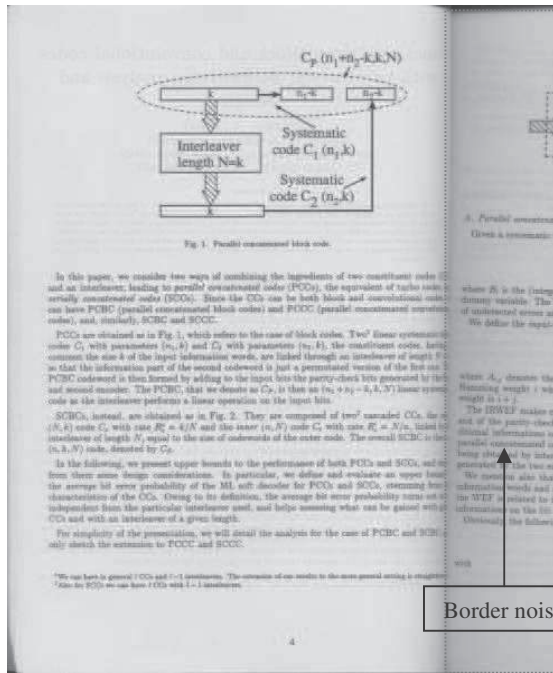


Figure 3 – Document images with Frame noise

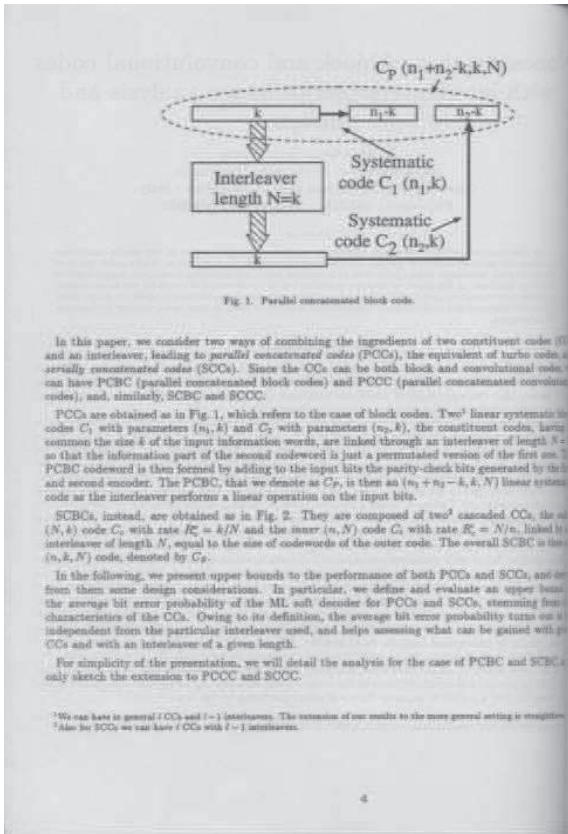


Figure 4 – Document images after removal of Border noise by PhotoDoc.

Skew	Noise	No_Noise
Synthetic	6,200	8,000
Original	3,800	2,000
Tiff (BW)	3,000	2,400
Tiff (gray)	3,000	3,000
png (color)	3,000	3,600
Jpg (color)	1,000	1,000
Orientation	Noise	No_Noise
Synthetic	6,200	8,000
Original	3,800	2,000
Tiff (BW)	3,000	2,400
Tiff (gray)	3,000	3,000
png (color)	3,000	3,600
Jpg (color)	1,000	1,000
Border	Noise	No_Noise
Synthetic	5,200	7,000
Original	5,346	2,011
Tiff (BW)	500	2,011
Tiff (gray)	2,600	5,000
png (color)	2,000	1,000
Jpg (color)	946	1,000
Back-to-Front	Noise	No_Noise
Synthetic	-----	-----
Original	2,027	3,000
Tiff (BW)	-----	-----
Tiff (gray)	-----	-----
png (color)	-----	-----
Jpg (color)	2,027	3,000
Blur	Noise	No_Noise
Synthetic	3,200	3,150
Original	-----	-----
Tiff (BW)	-----	-----
Tiff (gray)	-----	-----
png (color)	-----	-----
Jpg (color)	3,200	3,150

Table 1 – Main features on the images in the test set

The training set was carefully selected to guarantee the diversity of the images in the test set, keeping in mind that quality matters more than size. Table 02 presents the relative size of the training and test sets.

	Test	Training	%
Skew	20,000	1,600	8.00
Orientation	20,000	1,600	8.00
Border	19,557	1,651	8.44
Back-to-Front	5,027	510	10.14
Blur	3,000	350	5.83
<b>Total</b>	<b>67,584</b>	<b>5,711</b>	<b>8.45</b>

Table 2 – Sizes of Training x Test sets

## 2.1 Sub-sampling

Very often classifiers do not use the whole original image for classification, as their feature extraction is a time intensive task. The larger the image file, the richer it is in data redundancy. Thus, if the redundant data is thrown away, the efficiency both in time and classification increase. The selection of points should not be random. It should somehow provide a "reduced" version of the original image (although in some cases it may be distorted by unequal scaling!). The cascaded sub-sampler presented in reference [8] was used here. It performs the following operations:

<pre>size = height*width</pre> <ul style="list-style-type: none"> <li>• If <math>size \leq 300,000</math> break;</li> <li>• If <math>300,000 &lt; size \leq 500,000</math>: remove even lines or columns (whatever the larger);</li> <li>• If <math>500,000 &lt; size \leq 700,000</math>: remove even lines and columns;</li> <li>• If <math>700,000 &lt; size \leq 900,000</math>: remove 2 lines in every 3 lines and even columns, (if height&gt;width) remove even lines and 2 columns in every 3 columns, otherwise;</li> <li>• If <math>900,000 &lt; size</math>: remove 2 lines and 2 columns in every 3 lines/columns;</li> </ul> <p style="text-align: center;"><b>Code for the "cascaded" sub-sampler</b></p>
--

## 2.2 Classification features

The choice of the features to be extracted from each image is of paramount importance to the success of the classifier. The following set of features, based on the classifier described in reference [8], was chosen:

- Palette (true-color/grayscale)
- Gamut
- Conversion into Grayscale (if RGB)
- Gamut in Grayscale (if RGB)
- Conversion into Binary (Otsu)
- Number of black pixels in binary image.
- $(\#Black\_pixels/Total\_#\_pixels)*100\%$
- $(Gamut/Palette)*100\%$  (true-color/grayscale)
- Shannon's entropy in three different regions of the document shown in Figure 5.

Image binarization is performed by using Otsu [11] algorithm. The height and width stand for the number of pixels in the image. RGB size stands for the true color size of the image (if it is a color image). 8-bits size is either the size of the original image if in grey scale or the size of the grey-scale converted from true-color. #B\_pixels stands for the number of black pixels in the monochromatic converted image. The features above are extracted for each sub-sampled image and placed in a vector of features.

As will be explained later on in this paper, not all the features from the list above are used for detecting all the noises of concern in this research.

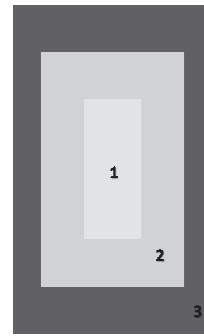


Figure 5 – Areas of interest for entropy calculation

## 3. RESULTS

This section presents the results of noise classification of the images in the test set for each of the noise classifiers. It is worth observing that the classifiers act in parallel as shown in Figure 2. Thus given an image the different noises may be observed simultaneously.

### 3.1 Border Noise Detection

The monochromatic images in the test set presented borders of all kinds:

- White borders,
- Black borders (uniform, irregular, etc.),
- Textured black borders.

Figure 6 presents examples of the different kinds of border noise.

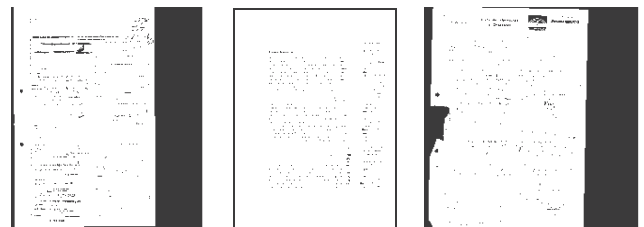


Figure 6 – Different kinds of framing borders in documents (from left-to-right: solid black, white framing border, and textured noisy border in a thorn off document).

The results of classification for the border noise detection classifier are presented by the confusion matrix presented in Table 3. One should stress that all images were sub-sampled for reasons of increasing the time efficiency of the feature extractor and classifier.

Border Noise	With	Without	Ratio
With	9,514	1,032	90.2142
Without	1,425	7,586	84.1859

Table 3 – Confusion matrix of the border noise classifier with sub-sampled images

Although the results obtained for detecting the border noise shown in the confusion matrix in Table 3 are quite reasonable, this classifier may be improved by being broken into parallel sub-classifiers for each of the kinds of borders surrounding the documents above.

### 3.2 Skew Detection

Skew noise is often found in the digitalization of large quantity of documents overall when performed by automatic feed scanners. In order to test the presence of skew noise in document images, the images considered as having original skew (inserted by the digitalization process) have rotation angles of less than 3 degrees in over 96% of cases. The remaining ones have rotation angles of less than 5 degrees. The synthetic images were generated by rotating straight-up images 2, 5 and 10 degrees. Some of the images were of handwritten documents such as the one on the top left hand corner of Figure 1. This although considered as being a non-skewed document, has a visible skew in the handwritten lines and so poses extra difficulty to the classifier. The confusion matrix for classifying images with skew is shown in Table 4.

Skew	With	Without	Accuracy %
With	9,671	329	96.71
Without	198	9,802	98.02

**Table 4** – Confusion matrix of the skew noise classifier with sub-sampled images

As one may observe from the results presented in Table 4 the classifier correctly detected most of the images. From the 329 images with skew that were classified as being without skew, there are 100 synthetic images of historical documents that, the rotation imposed compensated the skew in the handwriting.

### 3.3 Wrong Orientation Detection

The mass digitalization of batches of documents very often includes incorrectly placed ones, either upside down or sideways. The images included in the test set with original orientation noise are the result of such accidental misplacement in a large batch of documents from a real-world digitalization bureau. The synthetic documents were obtained by rotating the images 90, 180, and 270 degrees. The straight-up documents are documents whose orientation is correct.

Orientation	Misplaced	Straight-up	Ratio
Misplaced	9548	452	95.48
Straight-up	200	9,800	98.00

**Table 5** – Confusion matrix of the orientation noise classifier with sub-sampled images

The classification results shown in Table 5 present a high accuracy reaching over 96% of the documents analyzed. The detection of upside-down documents is responsible for most of the incorrect orientations found in the classifier results, as one would expect.

### 3.4 Back-to-Front Noise Detection

Back-to-front noise, also known as bleeding or show-through depending on its strength may make document binarization impossible. As most OCR software takes input as a binary image, this fact has as a consequence that documents with such a noise cannot be automatically transcribed. Researchers [15][16] have pointed out that no algorithm in the literature is good enough to remove bleeding noise in all sorts of documents. Depending on the strength of the noise, some algorithms may perform better than others. Unfortunately, the back-to-front noise appears more often in the digitalization of documents than one may assume to start with. The test set of documents we used with show-through had 2,027 real-world documents (no synthetic ones) which were obtained either from historical files (such as the one shown in the top-left hand corner of Figure 1) or from the scanning of printed proceedings of technical events. Images were hand labeled according to four levels of interference as: strong, medium, light and none. The classifiers for this noise were cascaded, as shown in Figure 2. The strong-classifier was trained with the images tagged as strong in the training set, against all the remaining images (Medium-Light-None) from the training set. Similarly, the medium-classifier was trained with the images labeled as medium, against the others with a lighter or no interference. The classification results obtained are shown in Table 6.

Back-to-front	Strong	Medium	Light	None	Accuracy %
Strong	1,073	65	3	1	93.95
Medium	91	638	15	19	83.61
Light	5	9	96	12	76.22
None	24	53	106	2,817	93.90

**Table 6** – Confusion matrix of the back-to-front noise classifier with sub-sampled images

The analysis of the data obtained shows that the classifier was able to detect the back-to-front noise in 90.97% of the noisy images and also to classify 93.90% of the noise-less images correctly. It is also worth mentioning that the misclassification of the images without noise was in the direction that they had a light back-to front interference. If one takes into account that such images were in JPEG format and that the background of many documents was not solid white, but also encompassed other noises due to aging, stains, etc, the results obtained are quite reasonable.

One should also note that the noisy documents, whenever misclassified, tend to be placed in the group immediately below. For instance 91 of the documents labeled as having strong bleeding noise were classified as having a medium noise, an acceptable result as the tagging followed no quantitative criteria. The adoption of synthetic noisy images could be of some help in solving the aforementioned problems, but their generation is far from being a simple task as it involves not only the overlapping of two images, one of which is faded. The image in the background also presents some degree of blur and this scenario gets complicated further in the case of the simulation of aged documents, a situation very often found whenever dealing with historical documents.

### 3.4 Blur Detection

One of the targets of this study is the analysis of images with blur and subsequently to propose a new method to automatically detect the presence or absence of noise. For such, the image is divided into small blocks might easily extracted information from the analysis of these blocks can be classified into two categories: blocks with and without blur noise. As already mentioned, blur is a noise that tends to appear in some regions of a document image instead of being a global noise. Its presence may be an indicator of low digitalization quality, but may also be associated with some other problems such as digitalizing hard-bound volumes. Although issues related to the analysis of blur in image have attracted much attention of researchers in recent years, the work reported in the literature is concentrated mostly in solving the de-blurring problem. Since the detection of blur is rarely explored and is still far from practical. Da Rugna and Konik [17] introduced a method of learning to classify image regions with or without blur. This method is based on an observation that blurry regions are more invariant to low pass filtering. But using only the technique described in [17] one obtains as result a lower detection rate than using the noise classification/detection architecture proposed here. For the set of images tested the performance of the Da Rugna-Konik classifier is around 16% lower than obtained by the classifier proposed. The confusion matrix for classifying images with blur is shown in Table 7 and Table 8.

Orientation	With	Without	Accuracy %
With	2,483	517	82.76
Without	614	2,386	79.53

**Table 7** – Confusion matrix of the **blur noise** classifier with sub-sampled images in proposed architecture

Orientation	With	Without	Accuracy %
With	1,987	1,013	66.23
Without	1,031	1,969	65.63

**Table 8** –Confusion matrix of the **blur noise** classifier with sub-sampled images

### 4. TIME PERFORMANCE

This section presents the time performance of the feature extractor and classifier, which used as hardware platform a machine running a processor Intel(R) Core(TM)2 Duo CPU E7400 @ 2.80GHz, with 4,00 GB RAM. The feature extractor and the sub-sampler were implemented in C++. Together, they take 93 ms per image, on average in the case of the detection of all other noises but the blur one that uses images without sub-sampling and claims 87 ms of processing time. For the global noise detection one takes 180 ms for feature extraction per image.

Table 9 presents details of the Random Forest [14] classifier time

performance, which was implemented in Weka [12], using Java as implementation language.

Classification	Number of Trees	Time (ms)	Language
Skew	10	3.1	Java
Orientation	10	3.1	Java
Board	10	3.1	Java
Back-to-front interference	30	5.9	Java
Blur	10	367	Java
Total time per image		394	

**Table 9** – Weka Random Forest classifier time performance

As shown in Table 9, the overall classification time per image is 394 ms. Such a time can be made smaller by a carefully re-coding the classifier in a lower level language such as C++. Time performance was not one of the concerns of the authors at present. The aim of the data provided is only to show that the approach of having noise classification prior to filtering is a viable strategy.

### 5. DISCUSSION AND CONCLUSIONS

Any document digitalization process introduces some sort of undesirable noise that makes more difficult and even impossible its readability by humans, degrades its automatic transcription via OCR, and claims for unnecessary storage space or network bandwidth for transmission. The automatic detection of noise in document images allows for better document filtering and enhancement. The classifier proposed herein presented a performance standard that is reliable enough to free humans of the burden of choosing which filters to use to remove the noises studied: border, skew, orientation and back-to-front interference. Besides that, it also helps avoiding document degradation by blindly processing document images through a bank of filters. It is also important to mention that although the classifier takes 487 ms (93 ms for feature extraction and sub-sampling plus 394 ms for classification) to be able to decide about the presence of the four studied noises in an image this time is far less than what would be needed to run the image through the unnecessary filters.

Weka [8] has provided an excellent testbed for statistical analysis. The choice for a Random tree classifier was made after performing several experiments with a large number of alternatives offered by Weka, including a MLP neural classifier although results did not vary widely.

The choice of the images in the training set is of paramount importance to the performance of the classifier. Quality has proved more important than size. The test set used here attempted to be representative of the universe of images of interest. Every effort was made to ensure correct labeling of images and to avoid image duplication.

This paper has proposed a change in the paradigm of document image filtering that may be also valid in many other fields of



image processing that has to process a large quantity of items. Further refinement of the classifier architecture must be pursued. For instance, the framing noises shown here encompass a variety of noises that may be sub-split into different classes of interest, as they are removed by different algorithms. The solid white or black framing noises may be removed with much less effort than whenever one has a textured black frame or even a torn-off region. Similarly, the back-to-front interference may be better classified if instead of having classes as presented here (strong, medium, and light) one has a quantitative measure of the interference. Such refinement in classification allows a better selection and tuning of the filters used to remove such noise.

## 6. ACKNOWLEDGEMENTS

The research reported herein was partly sponsored by CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico and CAPES – Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior. Collaboration between HP Labs. and UFPE was funded by MCT, Ministry of Science and Technology, Informatics Law, Brazilian Government.

## 7. REFERENCES

- [1] H.Frigui and R.Krishnapuram. Clustering by competitive agglomeration. *P. Recognition*, 30(7), 2001.
- [2] M.A.Hearst and J.O.Pedersen. Reexamining the Cluster Hypothesis: Scatter Gathet on Retrieval Results, SIGIR, 1996.
- [3] S.Krishnamachari and M.Abdel-Mottaleb. Image Browsing using Hierarchical Clustering, IEEE Symposium on Computers and Communications, ISCC'99, July 99.
- [4] P.Scheunders. Comparison of Clustering Algorithms Applied to Color Image Quantization, *Patt. Recog. Letters*, v18(11-13):1379-1384, 1997.
- [5] G.Park, Y.Baek and L.Heung-Kyu. A Ranking Algorithm Using Dynamic Clustering for Content-Based Image Retrieval. CIVR'2002, pp.328—337, LNCS 2383, Springer Verlag, 2002.
- [6] K.Barnard and D.Forsyth. Learning the Semantics of Words and Pictures, *Inter. Conf. C. Vision*, 2001.
- [7] S.J. Simske, "Low-resolution photo/drawing classification: metrics, method and archiving optimization," *Proceedings IEEE ICIP*, IEEE, Genoa, Italy, pp. 534-537, 2005.
- [8] R.D. Lins; G.F.P. Silva; S.J. Simske; J. Fan; M. Shaw; P. Sá; M Thielo. Image Classification to Improve Printing Quality of Mixed-Type Documents. In: International Conference on Document Analysis and Recognition, 2009, Barcelona. Proceedings of ICDAR 2009. New York: IEEE Press, 2009. p. 1106-1110.
- [9] G.F.P. Silva; R.D. Lins; B. Miro; S.J. Simske; M. Thielo. Scanned or Photographed? Automatically Deciding How a Document was Digitized. International Workshop on Camera-Based Document Analysis and Recognition, p. 1-10, IAPR Press, 2009.
- [10] G.F.P. Silva and R.D.Lins. PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. *ICDAR'2007*, pp.107-114, 2007.
- [11] R.D. Lins. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. International Conference on Image Analysis and Recognition, LNCS 5627, pp 844-854. Springer Verlag, 2009.
- [12] Weka 3: Data Mining Software in Java, website <http://www.cs.waikato.ac.nz/ml/weka/>.
- [13] N. Otsu. "A threshold selection method from gray level histograms". *IEEETrans.Syst.Man Cybern.* v(9):62-66, 1979.
- [14] L. Breiman, "Random Forests", *Machine Learning*, 45(1), pp. 5-32, 2001.
- [15] R.D Lins; J.M.M. Silva; F.M.J. Martins. Detailing a Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents. *Journal of Universal Computer Science*, v. 14, pp. 299-313, 2008.
- [16] P. Stathis; E. Kavallieratou; N. Papamarkos. An Evaluation Technique for Binarization Algorithms. *Journal of Universal Computer Science*, v. 14, pp. 3011-3030, 2008.
- [17] J. Da Rugna and H. Konik. Automatic blur detection for metadata extraction in content-based retrieval context. In *SPIE*, volume 5304, pages 285.294, 2003.

## Enhancing the Filtering-out of the Back-to-Front Interference in Color Documents with a Neural Classifier

G.F. P e Silva<sup>(1)</sup>, R. D. Lins<sup>(1)</sup>, J.M. Silva<sup>(1)</sup>, S. Banerjee<sup>(2)</sup>, A. Kuchibhotla<sup>(2)</sup> and M. Thielo<sup>(3)</sup>  
<sup>(1)</sup> UFPE-Brazil, <sup>(2)</sup> HP Labs.-India, <sup>(3)</sup> HP Brazil R&D-Brazil

{ gabriel.psilva, rdl}@ufpe.br, {serene.banerjee, anji, marcelo.resende.thielo}@hp.com

### Abstract

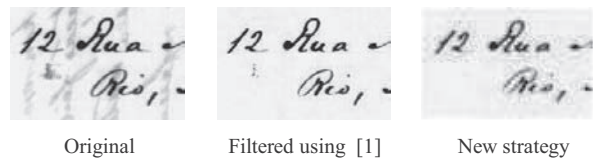
*Back-to-front, show-through, or bleeding are the names given to the interference that appears whenever one writes or prints on both sides of translucent paper. Such interference degrades image binarization and document transcription via OCR. The technical literature presents several algorithms to remove the back-to-front noise, but no algorithm is good enough in all cases. This article presents a new technique to remove such noise in color documents which makes use of neural classifiers to evaluate the degree of intensity of the interference and besides that to indicate the existence of blur. Such classifier allows tuning the parameters of an algorithm for back-to-front interference and document enhancement.*

### 1. Introduction

At the beginning of the 1990s, the important file of over 6,000 private letters of Joaquim Nabuco, a statesman, writer and diplomat, one of the leading figures of the freedom of black slaves in Brazil and the first Brazilian ambassador to the US, were digitized in a joint preservation effort between the Joaquim Nabuco Foundation [2] and the Universidade Federal de Pernambuco. About 10% of the images presented a noise which had not been previously described in the technical literature, which was called “back-to-front” interference [1]. Much later, other people called it “bleeding” [14] or “show-through” [15].

The back-to-front interference appears whenever the content of the verso side of a document is visible on the front side due to paper translucidity (Figure 1). Such artifact degrades the automatic document transcription via OCR and there is often the superposition of both sides whenever the image is binarized, yielding an unreadable document. In the case of historical

documents paper aging is a complicating factor as it darkens the paper and causes an overlapping of the distributions of the RGB components of the ink on both sides of the paper.



**Figure 1.** Zoom in a document from Nabuco bequest with back-to-front interference filtered out using the algorithm described in reference [1] and the new strategy herein.

The technical literature presents several algorithms to remove the back-to-front noise, but no algorithm is good enough in all cases [12]. Depending on the degree of translucidity of the paper, the kind of the ink used in printing or writing, the porosity of the paper, etc. the interference may show itself stronger or weaker. Some algorithms perform better than others in different degrees of interference and even one chosen algorithm may perform better if its parameters are tuned to the intensity of the noise.

This article presents a new strategy to select and tune an algorithm to remove the back-to-front interference in color documents. It makes use of a set of neural classifiers to assess the intensity of the back-to-front interference and to automatically adjust the parameters of the algorithm described in reference [1] to filter the noise out of a given a document. The blanks yielded by removing the artifact are filled in with pixels that correspond to the paper area in the document in such a way to provide the reader with “a natural” look of the document as if it were written on one side only. Figure 1 presents a sample of the results obtained by the algorithm proposed herein, in which one may observe its efficacy.

This paper is organized as follows. Section 2 describes the new filtering strategy. The document

features extracted are presented in Section 3. Section 4 details the noise detection mechanism. The results obtained are presented and analyzed in Section 5. The paper ends presenting its conclusions and draws lines for further work.

## 2. The Filtering System

This section presents a new strategy to remove the back-to-front interference in color documents. First, one needs to remove the framing borders in the document image. Such borders act as noise that interferes with the analysis performed by other algorithms [3]. PhotoDoc [4] was used to pre-process the whole file of documents. Then, there is the use of the classifiers to verify the existence and degree of back-to-front interference. In parallel to that analysis another classifier checks the presence of blur in blocks of the image. Once the former classifier detects the presence and intensity of back-to-front noise, the global threshold algorithm presented in reference [1] is tuned to remove the artifact. At the end the interfering pixels are painted with the colors of pixels that correspond to the sheet of paper, removing the interference in the resulting image.

### 2.1. Classification strategy

The architecture of the classifier is presented in Figure 2.

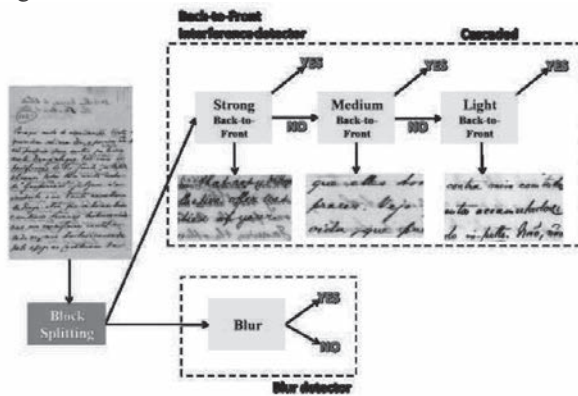


Figure 2. Classifier architecture

The classifier used is Random Forest [6], implemented in Weka [5], an open source tool developed at Waikato University, New Zeland, which offers a wide variety of classifiers implemented. A set of features is extracted from each image to allow its classification. The details about the training and test sets are provided in Section 3.

The classifier developed works in parallel for the detection in image blocks of two different kinds of

noise: the back-to-front interference and blur. In the case of back-to-front interference the classification is performed by three cascaded classifiers that split the bleeding noise into three categories: strong, medium and weak.

### 2.2. Discriminating the interfering pixels

The entropy-based segmentation algorithm by Silva-Lins-Rocha [1] is used twice to find the back-to-front interference area. The first time, to split the text from the rest of the document. The second time to separate the interference from the paper. The algorithm uses the grayscale converted image as an intermediate to split the histogram into three different areas of interest (see Figure 3).

The loss factor  $\alpha$  is a parameter of the algorithm that allows a better statistical tuning between the distributions of the original and binary histograms and it is based on Shannon entropy [8]. For the second image filtering using the algorithm by Silva-Lins-Rocha, a new  $\alpha$  is defined taking into account the intensity of the back-to-front interference and the presence of blur in the image, Table 1 indicates the suggested values  $\alpha$ , such as to allow a better separation between the interference and the paper distribution. The values for  $\alpha$  were experimentally found.

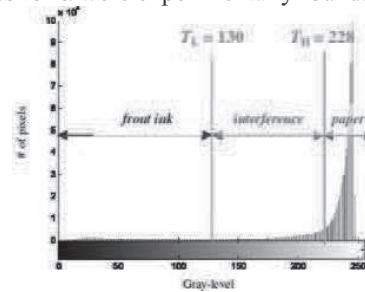


Figure 3. Image histogram of document with back-to-front interference - segmentation details.

Interference	Blur	No Blur
Weak	0.90	1.00
Medium	0.78	0.90
Strong	0.60	0.70

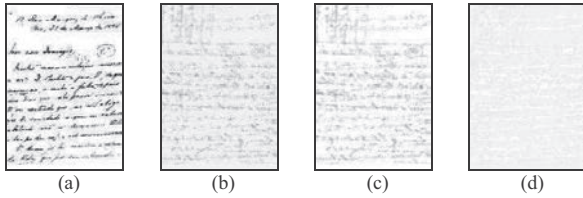
Table 1. Values for alpha.

In summary, to detect the interference area:

1. Apply the segmentation algorithm by Silva-Lins-Rocha to sieve the foreground ink from the rest of the document (see Figures 4a /4b);
2. For each image block classified as having back-to-front interference a new loss factor  $\alpha$  (see Table 1) is chosen. Filter it using the algorithm by Silva-Lins-Rocha to separate the interference ink from the paper (see Figures 4c

and 4d), yielding a blank sheet of paper with white holes where there was ink and ink interference in the original document image.

To illustrate the process, in Figure 4 the first threshold,  $T_L$ , is obtained by the first application of the Silva-Lins-Rocha algorithm and the blocks threshold,  $T_H$ , by the second. The pixels for which their gray-levels are less than  $T_L$  are classified as ink of the front face. The pixels with gray-level greater than  $T_H$  are classified as belonging to the paper. Pixels with gray-levels between  $T_L$  and  $T_H$  are discriminated as interference. The difference here is the application on each image block of a fine tuning between the thresholds  $T_L$  and  $T_H$  taking into account local image information. It is also worth stressing that this new filtering procedure has the advantage of reducing the risk of “damaging” image areas in which there is no interference, once the segmentation algorithm is not applied on them.



**Figure 04.** Views of a document with back-to-front interference: (a) ink of the front face and (b) paper with interference. Image segments of Figure 3b: (c) interference and (d) paper.

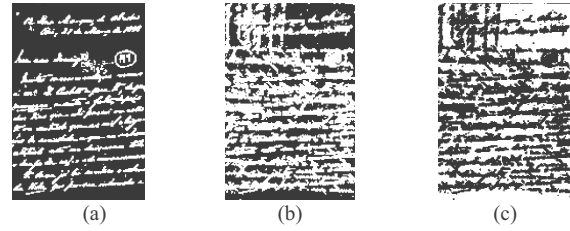
### 2.3. Document Reconstruction

The process proposed here makes use of a “linear” interpolation to fill in the blank pixels that originally corresponded to the interference area. Two binary masks are defined: TEXT and INTERF. The first one identifies the pixels from the ink of the front text (see Figure 5a); the second one highlights the interference area (see Figure 5b). One could assume that only the INTERF mask would be sufficient to the fulfillment process, because the pixels to be replaced “are known already”. Some difficulties appear, however. The key idea is to replace the colors of the noisy pixels with colors as close as possible to the paper in their neighborhood. This is achieved by interpolation, using the colors of the pixels that surround the area to be filled in. There is still the need to remove some of the vestigial shades surrounding the ink pixels in the resulting image; otherwise those pixels will “damage” the interpolation process, bringing in noisy dark colors to the interference area. To solve this problem, one should apply the dilate morphological expansion operation to both masks, with that, the text and

interference contours will be properly classified as “text” and “interference”, respectively (see Figure 6a and Figure 6b). As mentioned earlier on, the pixels that are used in the interpolation process are surrounding the interference area and with the pixels belonging only to the paper. This mask, PAPER, is obtained by the complement of the logical OR operation between the TEXT and INTERF dilated masks (see Figure 6c). Equation 1 calculates a weighed mean, where the intensity of the nearest pixel from the pixel P has the greatest weight. This is reasonable, because in a neighborhood, generally, the closer a pixel is from another, the more alike they should look. Figure 7b shows the result of the application of the proposed filtering strategy applied to the image in Figure 7a.



**Figure 05.** Masks that identify (a) the text and (b) the interference.



**Figure 06.** Dilated masks: (a) text (T) and (b) interference (I) (c) T or I.

Now, the interpolation process is presented. Let the coordinates be as depicted in Figure 7b:

- $(x_0, y_0)$  of a pixel  $P$  from the interval to be interpolated;
- $(x_0, y_1)$  of pixel  $P_N$  – first pixel north  $P$ ;
- $(x_0, y_2)$  of pixel  $P_S$  – first pixel south  $P$ ;
- $(x_1, y_0)$  of pixel  $P_W$  – first pixel west  $P$ ;
- $(x_2, y_0)$  of pixel  $P_E$  – first pixel east  $P$ ,

Where  $i_C(x, y)$  is the value of the component  $C$  (R, G or B) of the pixel  $(x, y)$ . The intensity of the interpolated pixel ( $P$ ) is given by

$$i_c(x_0, y_0) = \frac{d_4 \times i_1 + d_3 \times i_2 + d_2 \times i_3 + d_1 \times i_4}{d_4 + d_3 + d_2 + d_1}, \quad (1)$$

where the  $i_k$  and  $d_k$  ( $k = 1, \dots, 4$ ) represent the intensities and the distances from the pixels –  $P_N, P_S, P_W$  and  $P_E$  – to  $P$ , sorted by increasing distances. For example, the closest pixel to  $P$  has distance  $d_1$  and intensity  $i_1$ , the second closest one has distance  $d_2$  and

intensity  $i_2$ , and so on. The distance between any two pixels A e B with coordinates  $(x_a, y_a)$  and  $(x_b, y_b)$ , is the standard Euclidian distance:

$$d_{A,B} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}. \quad (2)$$

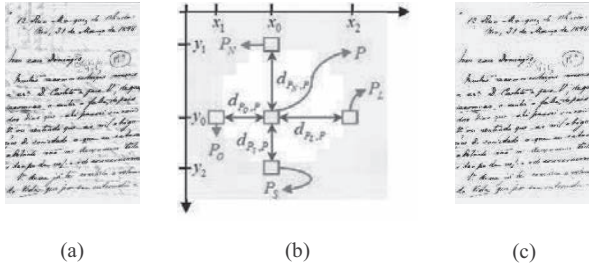


Figure 7. Images: (a) original, (b) Interpolation process and (c) filtered by the new strategy proposed here.

### 3. Classification Features

The choice of the features to be extracted from each image is of paramount importance to the success of the classifier. The following set of features, based a combination on the classifiers described in references [9]. Image binarization is performed by using Otsu [10] algorithm. The height and width stand for the number of pixels in the image. RGB size stands for the true color size of the image (if it is a color image). 8-bits size is either the size of the original image if in grey scale or the size of the grey-scale converted from true-color. #Black\_pixels stands for the number of black pixels in the monochromatic converted image. The combination of the features presented in [9] and the two new features (Local Power Spectrum Slope and Maximum Saturation) in [11] yielded a relative gain to the performance of the classifier. Each of these features was taken on nine blocks of the image (see Figure 5).



Figure 8. Areas of interest for extract features

### 4. Back-To-Front Noise Detection

Back-to-front noise, depending on its strength, may make document binarization unviable. As most OCRs take a binary image as input, thus documents with such noise may not be automatically transcribed. Researchers [12][13] have pointed out that no algorithm in the literature is good enough to remove bleeding noise in all sorts of documents. Depending on the strength of the noise, some algorithms may perform

better than others. Unfortunately, the back-to-front noise appears more often in the digitalization of documents than one may assume to start with. The test set of documents we used with show-though had 260 real-world documents (no synthetic ones) which were obtained either from historical files (as shown in Figure 1). The images were divided into nine blocks totaling 2,340 blocks and were hand labeled according to four levels of interference as: strong (773), medium (856), light (524) and none (187). The classifiers for this noise were cascaded, as shown in Figure 2. The strong-classifier was trained with the blocks tagged as strong in the training set, against all the remaining images (Medium-Light-None) from the training set (strong 150; medium 150; light 100; and 20 none). Similarly, the medium-classifier was trained with the blocks labeled as medium, against the others with a lighter or no interference. The classification results obtained are shown in Table 2.

Back-to-front	Strong	Medium	Light	None	Accuracy %
Strong	703	58	11	1	90.94
Medium	27	816	4	9	95.32
Light	5	9	96	12	92.93
None	1	2	11	187	92.51

Table 2. Confusion matrix of the back-to-front noise classifier with sub-sampled block images

### 5. Blur Detection

The presence of blur may be an indicator of low quality digitalization, but can also be associated with other problems such as the scanning of hard-bound volumes. The blur noise is seldom global. In general, it affects some areas of a document. In the case of the documents studied here, the blur noise is originated from the spreading out of the ink in the verso face of the document. While the issues related to the analysis of image-blur have attracted much attention from researchers in recent years, the work reported in the literature focus mainly on solving the problem of deblurring. Blur detection is a complex task. The work reported in [7], points at blur as one of the greatest difficulties for the filtering out of the back-to-front noise in historical documents. Bluer detection is solved here by using the classifier presented in subsection 2.1 and the characteristics described in section 3. The image blocks were classified manually, 124 blocks with and 2216 without blur. The training set used 20 blurred blocks and 150 unblurred ones. Besides those, 500 blocks with synthetic blur were used to validate the

classifier. The result is shown by the myo classifier confusion matrix (see Table 3).

Orientation	With	Without	Accuracy %
With	615	9	98.55
Without	3	2,213	99.86

**Table 3.** Confusion matrix of the **blur noise** classifier with sub-sampled images in proposed architecture

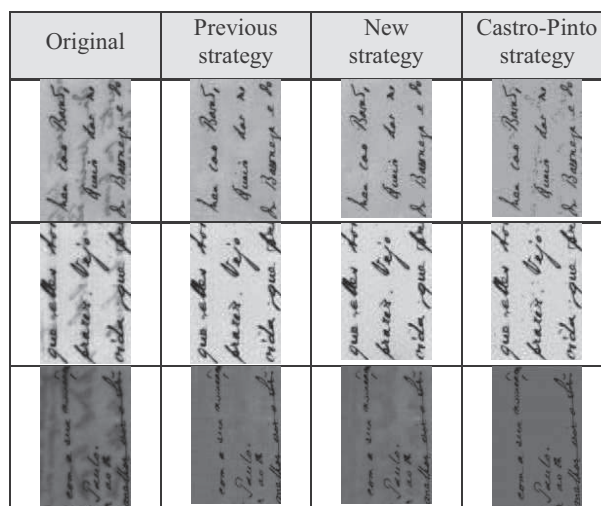
## 6. Results and Analysis

The proposed algorithm was tested in a set of 260 images from the Joaquim Nabuco bequest of digitalized documents [2], yielding good results. Evidences of the efficiency of the new filtering technique are shown in Figure 9, as the back-to-front interference was removed yielding a more readable document with a “natural” look. Figure 9 provides the results of using different strategies, amongst them using as fulfillment for the blanks the result of the interpolation based on Laplace’s equation (the MATLAB function “*roifill*” was used). The third alternative is one of the strategies proposed by Castro and Pinto [16] that uses the algorithm by Salvola and Pietikainen [17] which defines a mask that identifies the pixels of the foreground and background objects. The final image is obtained through keeping the object pixels and replacing the background pixels with the average of the colors of the pixels in that class. The latter strategy yielded the best results. The two strategies proposed herein yielded very similar quality results. However, the one based on Laplace interpolation leaves the filled-in area look undesirably uniform with a “flat” color. On the other hand, the linear interpolation yields a residual pattern of vertical/horizontal stripes. The strategy proposed by Castro and Pinto [16] aims to yield a uniform paper surface with unchanged text, while the ones presented here try to remove only the interference, keeping the pixels from the paper and text unchanged. However, in the very few images in the Nabuco file that the back-to-front interference looks very “blurred” (see last segments of Nabuco file in Figure 9), the proposed algorithm did not perform well. The detection of the whole back-to-front interference area is far from being a trivial task.

## References

[1] J. M. M. da Silva; *et al.* A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference". *J. Universal Computer Science*, v. (14):299-313, 2008.  
 [2] FUNDAJ: [www.fundaj.gov.br](http://www.fundaj.gov.br).

[3] R. D. Lins. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. *ICIAR 2009*. LNCS v. 5627. p. 844-854, Springer Verlag, 2009.  
 [4] G. F. P. e Silva and R. D. Lins. PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. *Proceedings of CBDAR 2007*. IAPR Press, 2007. p. 107-115.  
 [5] Weka 3: Data Mining Software in Java, website <http://www.cs.waikato.ac.nz/ml/weka/>.  
 [6] L. Breiman, "Random Forests", *Machine Learning*, 45(1), pp. 5-32, 2001.  
 [7] J. M. M. da Silva; *et al.* Enhancing the Quality of Color Documents with Back-to-Front Interference. *ICIAR 2009*. LNCS, v. 5627, p. 875-885, Springer Verlag, 2009.  
 [8] N. Abramson, "*Information Theory and Coding*", McGraw-Hill Book Co, 1963.  
 [9] R.D Lins; G.F.P. Silva; S. Banergee; A. Kuchibhotla and M. Thielo. "Automatically Detecting and Classifying Noises in Document Images", *ACM-SAC '2010*, ACM Press, March 2010.  
 [10] N. Otsu. "A threshold selection method from gray level histograms". *IEEE Trans. Syst. Man Cybernetics SMC-9*, 62-66, 1979.  
 [11] L. Renting, L. Zhaorong, and J. Jiaya, Image Partial Blur Detection and Classification, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.  
 [12] R.D Lins; J. M. M. da Silva; F. M. J. Martins. Detailing a Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents. *Journal of Universal Computer Science*, v. 14, pp. 299-313, 2008.  
 [13] P. Stathis; E. Kavallieratou; N. Papamarkos. An Evaluation Technique for Binarization Algorithms. *Journal of Universal Computer Science*, v. 14, pp. 3011-3030, 2008.  
 [14] R. Kasturi, L. O’Gorman and V. Govindaraju, "Document image analysis: A primer", *Sadhana*, (27):3-22, 2002.  
 [15] G. Sharma, "Show-through cancellation in scans of duplex printed documents", *IEEE Trans. Image Processing*, v10(5):736-754, 2001.  
 [16] P. Castro, J. R. C. Pinto: "Methods for Written Ancient Music Restoration". *ICIAR 2007*: 1194-1205.  
 [17] J. Sauvola, M. Pietikainen: "Adaptive document image binarization", *Patt. Recognition* 33(2):225-236, 2000.



**Figure 9** - Parts of documents from the Nabuco file: original and filtered.

#### **A.4 Publicações sobre Ferramentas para Processamento de Imagens de Documentos**

(SILVA et al., 2010b) - G. F. P. Silva; R. D. Lins; J. M. M. Silva. HistDoc - A Toolbox for Processing Images of Historical Documents. Lecture Notes in Computer Science, vol.6112, pp: 409-419.

(LINS et al., 2011b) - R. D. Lins; G. F. P. Silva; A. Formiga. HistDoc v. 2.0: enhancing a platform to process historical documents. In: Workshop on Historical Document Imaging and Processing, vol.1. pp: 169-176.

# HistDoc - A Toolbox for Processing Images of Historical Documents

Gabriel Pereira e Silva, Rafael Dueire Lins, and João Marcelo Silva

Universidade Federal de Pernambuco, Recife, Brazil

gfps.cin@gmail.com, rdl@ufpe.br, joao.mmsilva@ufpe.br

**Abstract.** HistDoc is a software tool designed to process images of historical documents. It has two operation modes: standalone mode - one can process one image a time; and batch mode - one can process thousands of documents automatically. This tool automatically detects noises present in the document image including back-to-front interference (also called bleeding or show-through) and uses the best techniques to filter it out. Besides that it removes noisy borders and salt-and-pepper degradation introduced during the digitalization process. PhotoDoc also allows document binarization and image compression.

**Keywords:** Back-to-front interference, bleeding, show-through, historical documents, border removal, binarization, document enhancement.

## 1 Introduction

Document images - acquired either by scanners or digital cameras - almost always present some kind of noisy artifacts. This statement is particularly true in the case of images of historical documents, in which one often finds back-to-front interference [10] (also known as bleeding [6] or show-through [19]), darkened paper, faded ink, folding marks, stains and damaged or torn off regions. The bequest of the letters of Joaquim Nabuco, a Brazilian statesman, writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil (b.1861-d.1910) is a file of historical documents of paramount importance to understand the formation of the political and social structure of the countries in the Americas and their relationship with other countries. This rich file is kept by the Joaquim Nabuco Foundation [3] (a social science research institute in Recife - Brazil). It encompasses over 6,000 letters of active and passive correspondence. The HistDoc tool presented here was conceived as a way to preserve this important heritage, as the chemical process used in producing paper in the late 19<sup>th</sup> century used too much beach and the papers are in a fast decomposition process. An example of a document of the Nabuco collection is presented in Figure 1, in which one may observe back-to-front interference, paper darkened, document filing annotation, and writing in different directions, a feature often found in such documents.

HistDoc was conceived as a device independent software tool to run on PCs. Whenever the user unloads the images of the historical documents, he will be able to run the tool prior to storing, printing or sending through networks the document images. HistDoc works in two different ways user driven standalone mode and batch mode. In standalone mode the user chooses which filters to use to enhance the document image.



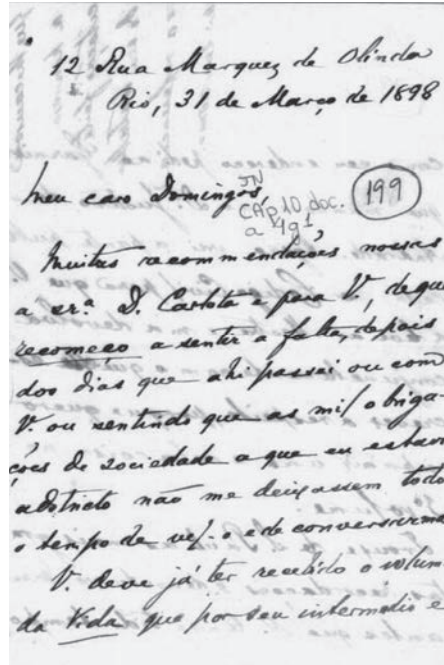


Fig. 1. Letter from Nabuco data base

In batch mode HistDoc uses the noise classifier presented in reference [33] specially tuned for historical documents, which automatically detects which undesirable artifacts are present in each document image and applies the suitable filtering technique. One should observe that such *a priori* noise classification is an important new feature in batch image processing.

HistDoc also encompasses a document compression module which decomposes the document image into paper background and writing. The color distribution and texture of the paper and writing are collected and the monochromatic image of the document is stored. Whenever the document is to be printed the data collected allows to colorize the monochromatic image yielding an image similar to the original one, at the cost of storing (or network transmitting) a compressed monochromatic document.

This paper is organized as follows. Section 2 briefly sketches the automatic noise classifier. The image filters implemented and the user interface for the standalone operation mode is presented in Section 3. The document image compression scheme is presented in Section 4. Conclusions and lines for further work are drawn at the final section.

## 2 The HistDoc Noise Classifier

Each document exhibits different noises and in general batch processing applies filters blindly and this may even cause document image degradation. In this section, the HistDoc Noise Classifier is outlined.

A number of features are extracted from each image to allow classification and training set as specified in [33]. The noise classifier used is Random Forest [31] which was implemented in Weka [32], an open source tool for statistical analysis developed at the University of Waikato, New Zealand. The noise detection architecture is formed by parallel classifiers that detect framing border noises, skew, orientation and back-to-front interference. The first three classifiers detect noises with almost 100 % accuracy, while the last one due to its complex nature claimed for a more sophisticated noise detection and classification strategy as explained below.

### 2.1 The Back-to-Front Interference Classifier

Researchers [29] [30] have pointed out that no algorithm in the literature is good enough to remove bleeding noise in all sorts of documents. Depending on the strength of the noise, some algorithms may perform better than others. Unfortunately, the back-to-front noise appears more often in the digitalization of documents than one may assume to start with. The test set of documents with show-through used was formed by 2,027 real-world documents (no synthetic ones) which were obtained either from historical files (such as the one shown in Figure 1 from Nabuco bequest) or from the scanning of printed proceedings of technical events. Images were hand labeled according to four levels of interference as: strong, medium, light and none. The classifiers for this noise were cascaded. The architecture of the cascaded classifier to handle the spotting of the back-to-front noise is shown in Figure 2.

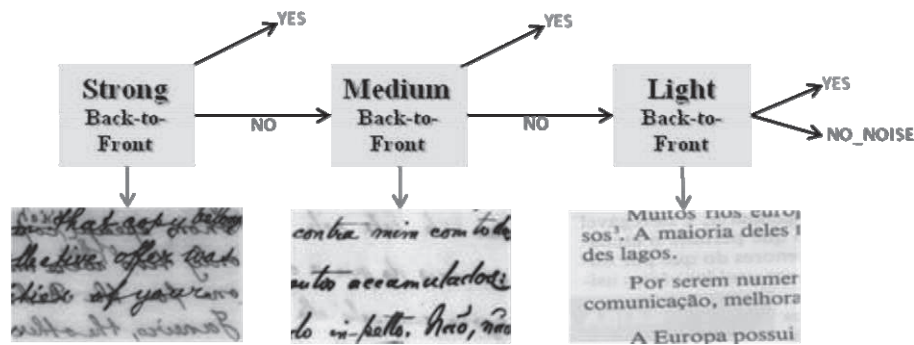


Fig. 2. Cascaded back-to-front noise detection architecture

The strong-classifier was trained with the images human tagged as strong in the training set, against all the remaining images (Medium-Light-None) from the training set. Similarly, the medium-classifier was trained with the images labeled as medium, against the others with a lighter or no interference. The classification results obtained are shown in Table 1.

The analysis of the data obtained shows that the classifier was able to detect the back-to-front noise in 90.97% of the noisy images and also to classify 93.90% of the noise-less images correctly. It is also worth mentioning that the misclassification of the images without noise was in the direction that they had a light back to front interference. If one takes into account that such images were in JPEG format and that the

**Table 1.** Confusion matrix of the back-to-front noise classifier with sub-sampled images

<b>Back-to-front</b>	Strong	Medium	Light	None	Accuracy %
Strong	1,073	65	3	1	93.95
Medium	91	638	15	19	83.61
Light	5	9	96	12	76.22
None	24	53	106	2,817	93.90

background of many documents was not solid white, but also encompassed other noises due to aging, stains, etc, the results obtained are quite reasonable.

One should also note that the noisy documents, whenever misclassified, tend to be placed in the group immediately below. For instance 91 of the documents labeled as having strong bleeding noise were classified as having a medium noise, an acceptable result as the tagging followed no quantitative criteria. The adoption of synthetic noisy images could be of some help in solving the aforementioned problems, but their generation is far from being a simple task as it involves not only the overlapping of two images, one of which is faded. The image in the background also presents some degree of blur and this scenario gets complicated further in the case of the simulation of aged documents, a situation very often found whenever dealing with historical documents.

### 3 HistDoc Filters

This section explains the filters implemented in HistDoc and presents the user interface for operating them in standalone user driven filtering. The filters developed in HistDoc are able to process the kinds of noise often found in historical document. The same filters are used in batch mode processing. The current version of HistDoc is implemented as an ImageJ [4] plug-in. Figure 3 shows a screen shot of HistDoc being activated from ImageJ.

As one may observe in Figure 3, the present version of the HistDoc plug-in offers five different filters, which appear in alphabetical order:

1. Back-to-front interference removal
2. Binarization
3. Border removal
4. Document Enhancement
5. Compression

The fact that HistDoc is now in ImageJ also allows the user to experiment with the different filters and other plug-ins already present in ImageJ.

ImageJ, as an open code library, allows the developer to extract from it only the needed functionality in such a way that the developer may provide to ordinary user a HistDoc interface that looks independent from ImageJ. At present, the authors of this paper consider such possibility premature. Such tool particularization seems to be more adequate if the processing tool becomes embedded into a particular device, which allows also a better tuning of the algorithms implemented in HistDoc developed for such a device. In what follows the HistDoc filter operations are described.



Fig. 3. HistDoc plugin in ImageJ

### 3.1 Border Removal

Very often document binarization either performed with scanners or cameras yield an image framed with some background which served of physical support to the document, an instance of which may be found in Figure 4 left. There are obvious drawbacks in keeping such frame: larger space and network bandwidth are needed for storage and transmission, respectively; The visualization area in a device such as a CRT is wasted in exhibiting pixels that convey no information and ink or toner are used in printing such border noise. Besides that, the digitalization border has a serious

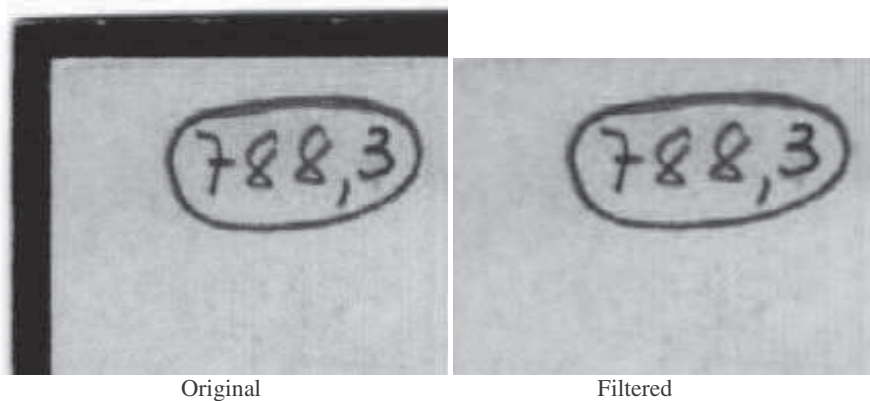


Fig. 4. Images: original and border removed

deleterious impact in the quality of the image subjected to palette reduction. This brings important implication as most automatic transcription tools (OCR and ICR) pre-process their input images into grayscale or binary before character recognition. The very first step to perform in processing a document image in HistDoc is to detect the actual physical limits of the original document [3]. Reference [3] reports on the binarization of documents. HistDoc offers to the user 16 thresholding techniques suitable for this sort of document, as it is detailed in the next section. Global and even local binarization algorithms take into account a statistical analysis of the document image, thus the presence of such border mislays the binarization process.

The algorithm presented in reference [20] was used in the development of the HistDoc (see Figure 4).

### 3.2 Back-to-Front Interference Removal

The HistDoc document processing environment offers three different strategies for filtering out the back-to-front noise [11], [21], [25] (see Figure 5). Whenever HistDoc is used in the user driven mode the user may select the most suitable algorithm for removing the back-to-front noise present in the document. If operated in batch mode the noise classifier will automatically choose the filter to be applied based on the strength of the interfering artifact.

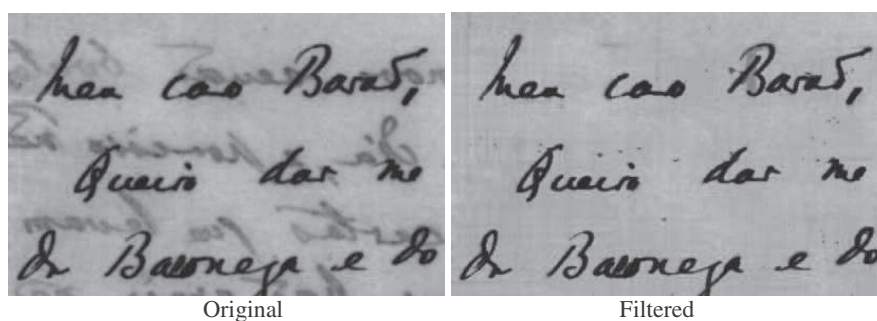


Fig. 5. Zoom into parts with back-to-front interference

The basic idea of the algorithm presented in reference [25], the most sophisticated and general of the algorithms implemented in HistDoc is to segment the three components of the document (background, ink and interference). Figure 6 shows the segmentation of the components of the document. The scheme used applies twice a global entropy-based thresholding algorithm. The first application of the algorithm separates the text from the rest of the document. The second pass separates the back-to-front interference from the rest of the paper background. Different loss factors  $\alpha$ , an empirically found adjustment parameter that allows a better adjustment between the distributions of the original and binarized images, are used in the two applications of the algorithm. In the case of the batch automatic application of this algorithm three pairs of are used to suitably remove the strong, medium and weak back-to-front interference. The result of the application of such scheme to the document shown in Figure 1 appears in Figure 6. This scheme is also of central importance in the parametric image compression strategy presented in Section 4 below, also implemented in HistDoc.

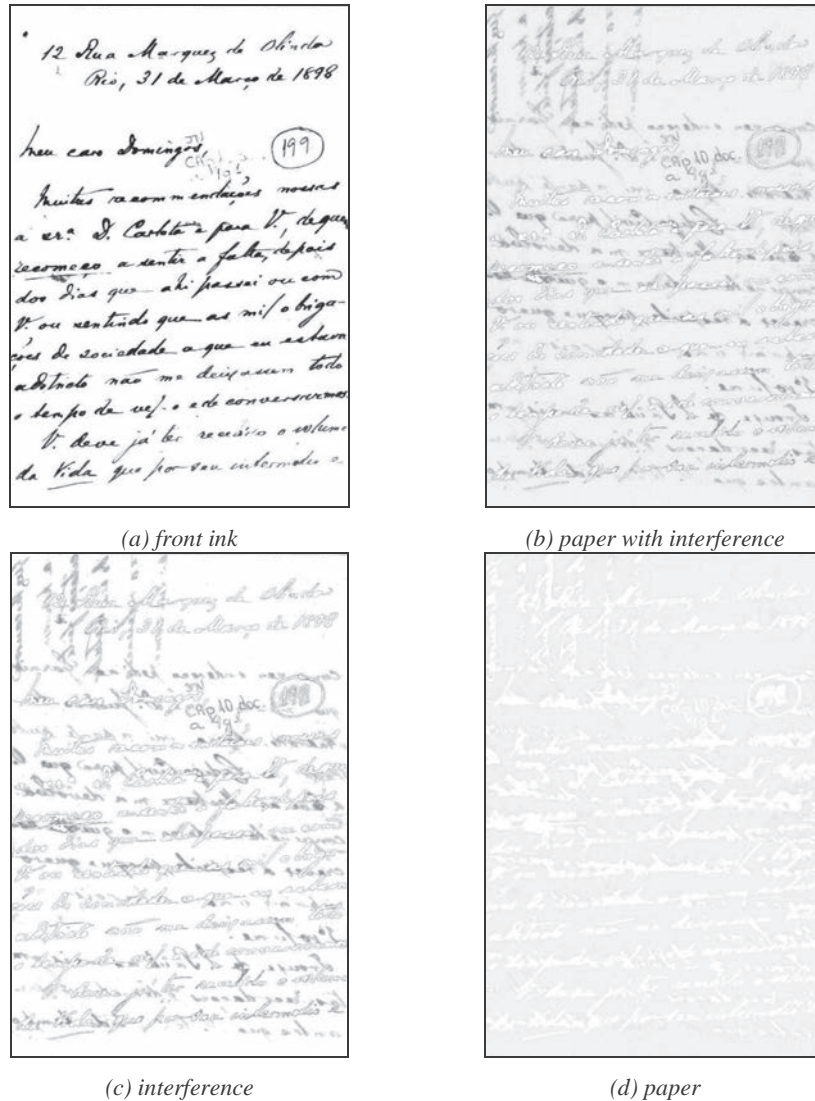


Fig. 6. Image segments of a document with back-to-front interference

### 3.3 Binarization

Document binarization is an important operation not only because a binary image is much smaller than its color counterpart but also due to most automatic transcription tools (OCR and ICR) pre-process their input images into grayscale or binary before character recognition. Reference [34] presents a survey of the most important binarization techniques applied to documents. HistDoc in user driven mode offers to the user 16 thresholding techniques suitable for this sort of document:

- 11 global ([5], [7], [8], [9], [13], [15], [17], [23], [24], [27], [28]) and,
- 5 local ([1], [14], [16], [18], [26]).

Figure 7 shows the result of the binarization of the image in Figure 3 and provides an account that the removal of the back-to-front interference prior to binarization is mandatory; otherwise the show-trough noise irrecoverably degrades document information in the monochromatic version.

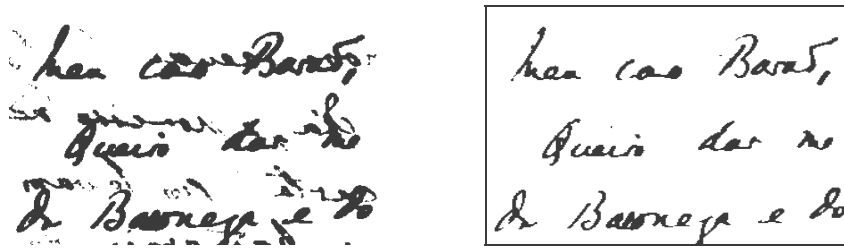


Fig. 7. Direct binarization (left) and after back-to-front interference removed (right)

In the case of using HistDoc in the automatic batch mode the binarization algorithm is called from the back-to-front noise detector.

### 3.4 Document Enhancement

This task creates a mask that identifies the pixels of the foreground and background objects. The final image is obtained through keeping the object pixels and replacing the background pixels with the average of the colors of the pixels in that class. HistDoc brings two strategies to do this. The first is the proposed by Castro and Pinto [2], that uses the Sauvola and Pietikainen's binarization algorithm [18] to determine the mask. The second strategy is based on the algorithm in reference [23]. Figure 8 presents the results of the latter algorithm.

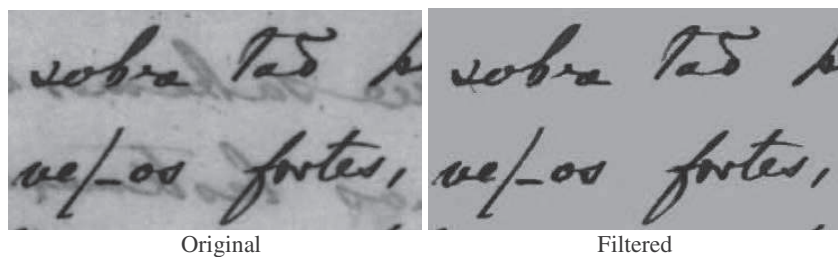


Fig. 8. Images: original and enhanced (filtered)

## 4 HistDoc Compression Module

If the user wants to obtain an image that resembles the color original image, but is very efficiently compressed, HistDoc offers the compression scheme described in

reference [22], in which the image is decomposed and stored as a compressed monochromatic image together with the colors and texture of the different graphical elements in the document (paper, printing, signature, etc.). The basic principle adopted in this compression scheme is shown in Figure 9.

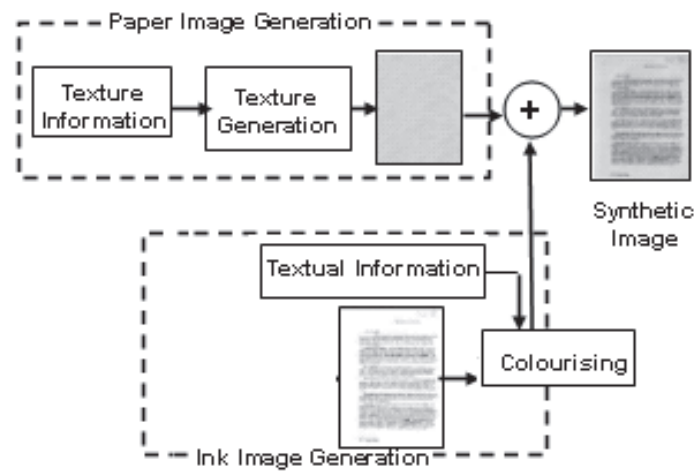


Fig. 9. Parametric generation of synthetic color document images

The user may also save images with the several file formats available in ImageJ (jpg, jpeg2000, png, tiff, etc), with and without losses.

## 5 Conclusions and Lines for Further Work

HistDoc is a user friendly tool for processing images of historical documents. It works in two different modes: user driven and automatic batch filtering mode. The batch mode makes use of a noise detecting tool that automatically detects and removes noisy framing borders, skew, orientation and back-to-front interference. The output may be either a binary image, a color image in the same file format of the input image or a parametrically compressed image which closely resembles the original one but is far more efficiently compressed.

The user driven operating mode of HistDoc provides a wide range of filters to enhance the document image at will. The first version was developed using the MATLAB [12] environment. It can be used as a MATLAB Tool, but a standalone version is also available. Aiming to speed-up the document processing phase, some of the algorithms are implemented in C.

The current version of HistDoc was developed as a plug-in in ImageJ, an open source portable Java library freely available. HistDoc runs on the users' PC and has the advantage of the great portability of Java. The executable code of HistDoc is freely available and may be obtained by requesting to the authors of this paper.

Several lines may be followed to provide further improvements to HistDoc filters and environment. Some of them are: being able to easily erase marks and stains from



the digital version of the document, incorporate screens in which the user may provide annotations, interface with an OCR to automatically transcribe or find keywords in documents. The interfacing of Tesseract [35] with HistDoc is on progress. Incorporating into HistDoc some of the functionalities of Gamera [36] another free platform of similar purpose is also a possibility. Preliminary tests performed with Gamera showed that although its OCR mechanism presents a much lower recognition performance than Tesseract, it allows the user to train the OCR recognizer with new font types, for instance, which may be of interest in some files of historical documents in which all documents were typed using a particular machine. Gamera is implemented in Python and C++ and is slightly faster than the current version HistDoc which is implemented as an ImageJ plugin in Java.

## Acknowledgments

Research reported herein was partly sponsored by CNPq – Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico and CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazilian Government.

The authors also express their gratitude to the Fundação Joaquim Nabuco, for granting the permission to use the images from Nabuco bequest.

## References

1. Lins, R.D., et al.: An Environment for Processing Images of Historical Documents. In: *Microprocessing & Microprogramming*, pp. 111–121. North-Holland, Amsterdam (1994)
2. Kasturi, R., ÓGorman, L., Govindaraju, V.: Document image analysis: A primer. *Sadhana* (27), 3–22 (2002)
3. Sharma, G.: Show-through cancellation in scans of duplex printed documents. *IEEE Trans. Image Processing* 10(5), 736–754 (2001)
4. FUNDAJ, <http://www.fundaj.gov.br> (accessed on 20/03/2010)
5. Lins, R.D., Silva, G.F.P., Banergee, S., Kuchibhotla, A., Thielo, M.: Automatically Detecting and Classifying Noises in Document Images. In: *ACM-SAC 2010*, ACM, New York (2010)
6. Breiman, L.: Random Forests. *Machine Learning* 45(1), 5–32 (2001)
7. Weka 3: Data Mining Software in Java, <http://www.cs.waikato.ac.nz/ml/weka/>
8. Lins, R.D., Silva, J.M.M., Martins, F.M.J.: Detailing a Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents. *Journal of Universal Computer Science* 14, 299–313 (2008)
9. Stathis, P., Kavallieratou, E., Papamarkos, N.: An Evaluation Technique for Binarization Algorithms. *Journal of Universal Computer Science* 14, 3011–3030 (2008)
10. IMAGEJ, <http://rsbweb.nih.gov/ij/> (accessed on 20/03/2010)
11. e Silva, A.R.G., Lins, R.D.: Background removal of document images acquired using portable digital cameras. In: Kamel, M.S., Campilho, A.C. (eds.) *ICIAR 2005*. LNCS, vol. 3656, pp. 278–285. Springer, Heidelberg (2005)
12. Lins, R.D., Netto, I.G.: Uma Nova Estratégia para Filtrar a Interferência Frente-Verso em Documentos Históricos. In: *XXV SBrT*, Recife, Brazil (2007)
13. da Silva, J.M.M., Lins, R.D.: Um Novo Método de Filtragem de Interferência Frente-Verso em Documentos Coloridos. In: *XXV SBrT*, Recife, Brazil (2007)

14. da Silva, J.M.M., Lins, R.D., Silva, G.F.P.: Enhancing the quality of color documents with back-to-front interference. In: Kamel, M., Campilho, A. (eds.) ICIAR 2009. LNCS, vol. 5627, pp. 875–885. Springer, Heidelberg (2009)
15. Sezgin, M., Sankur, B.: Survey over Image Thresholding Techniques and Quantitative Performance Evaluation. *Journal of Electronic Imaging* 13(1), 145–165 (2004)
16. Bernsen, J.: Dynamic thresholding of gray level images. In: ICPR'86: Proc. Intl. Conf. Patt. Recog., pp. 1251–1255 (1986)
17. Castro, P., Pinto, J.R.C.: Methods for Written Ancient Music Restoration. In: Kamel, M.S., Campilho, A. (eds.) ICIAR 2007. LNCS, vol. 4633, pp. 1194–1205. Springer, Heidelberg (2007)
18. Kapur, J.N., Sahoo, P.K., Wong, A.K.C.: A new method for gray-level picture thresholding using the entropy of the histogram. *G. Models I. Process.* 29, 273–285 (1985)
19. Kavalliaratou, E., Antonopoulou, H.: Cleaning and Enhancing Historical Document Images. In: Blanc-Talon, J., Philips, W., Popescu, D.C., Scheunders, P. (eds.) ACIVS 2005. LNCS, vol. 3708, pp. 681–688. Springer, Heidelberg (2005)
20. Khashman, A., Sekeroglu, B.: A Novel Thresholding Method for Text Separation and Document Enhancement. In: 11th Panhellenic Conf. on Informatics, Greece, May 18–20 (2007)
21. Kittler, J., Illingworth, J.: Minimum error thresholding. *Patt. Recog.* 19, 41–47 (1986)
22. Mello, C.A.B., Lins, R.D.: Generation of images of historical documents by composition. In: ACM Document Engineering 2002, McLean, VA, USA (2002)
23. Niblack, W.: An Introduction to Image Processing, pp. 115–116. Prentice-Hall, Englewood Cliffs (1986)
24. Otsu, N.: A threshold selection method from gray level histograms. *IEEE Trans. Syst. Man Cybernetics* 9, 62–66 (1979)
25. Palumbo, P.W., Swaminathan, P., Srihari, S.N.: Document image binarization: Evaluation of algorithms. In: Proc. SPIE, vol. 697, pp. 278–286 (1986)
26. Ridler, T.W., Calvard, S.: Picture thresholding using an iterative selection method. *IEEE Trans. Syst. Man Cybern. SMC-8*, 630–632 (1978)
27. Sauvola, J., Pietaksinen, M.: Adaptive document image binarization. *Pattern Recog.* 33, 225–236 (2000)
28. da Silva, J.M.M., Lins, R.D., Martins, F.M.J., Wachenchauser, R.: A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference. *Journal of Universal Computer Science* 14, 299–313 (2008)
29. da Silva, J.M.M., Lins, R.D., da Rocha Jr., V.C.: Binarizing and Filtering Historical Documents with Back-to-Front Interference. In: ACM-SAC'06, ACM Press, New York (2006)
30. White, J.M., Rohrer, G.D.: Image thresholding for optical character recognition and other applications requiring char. image extraction. *IBM J. Res. Dev.* 27(4), 400–411 (1983)
31. Wu, L.U., Songde, M.A., Hanqing, L.U.: An effective entropic thresholding for ultrasonic imaging. In: ICPR'98: Intl. Conference Pattern Recognition, pp. 1522–1524 (1998)
32. Yen, J.C., Chang, F.J., Chang, S.: A new criterion for automatic multilevel thresholding. *IEEE Trans. Image Process.* IP-4, 370–378 (1995)
33. da Silva, J.M.M., Lins, R.D.: Color Document Synthesis as a Compression Strategy. In: ICDAR 2007, vol. 1, pp. 466–470. IEEE Press, Los Alamitos (2007)
34. MATHWORKS, <http://www.mathworks.com/>
35. Tesseract OCR, <http://code.google.com/p/tesseract-ocr/> (accessed on 20/03/2010)
36. Gamera Project, <http://gamera.informatik.hsnr.de/> (accessed on 20/03/2010)

# HistDoc v. 2.0

## Enhancing a Platform to Process Historical Documents

Rafael Dueire Lins  
U.F.PE.  
Recife - Pernambuco - Brazil  
+55 81 8896-0698  
rdl.ufpe@gmail.com

Gabriel de F. Pereira e Silva  
PPGEE - U.F.PE.  
Recife - Pernambuco - Brazil  
+55 81 8803-8715  
gfps.cin@gmail.com

Andrei de A. Formiga  
U.F.PB.  
J.P. - Paraiba - Brazil  
+55 83 8126-0808  
andrei@dce.ufpb.br

### ABSTRACT

The first version of the HistDoc platform was designed as an ImageJ plugin to process images of historical documents. This paper presents the second version of HistDoc that besides updating the image processing capabilities of HistDoc in a number of ways, including processing images of monochromatic documents and incorporating newer and better algorithms for the old functionality, it allows document images to be batch processed in standalone mode in a single machine and in parallel distributed architectures in cluster and grids.

### Categories and Subject Descriptors

I.4.9 [Image Processing and Computer Vision]: Applications.

### General Terms

Algorithms, Document image analysis.

### Keywords

Document processing, image processing, historical documents.

## 1. INTRODUCTION

The Internet has revolutionized all areas of human activity from personal relations to shopping, but none of them will have as much impact as the access to knowledge. Every day more and more information sources are made available to the whole world without geographical frontiers or barriers of any kind. Historical books and documents are being “published” in the Internet at an astonishing pace by libraries, universities, research centers and even individuals. Virtual libraries and museums are making available new documental sources every day casting a “new light” on them making possible a plural view of history. Document engineering a new area of knowledge that acquire, process, index, and makes available all sorts of documents has had a rapid development in many different branches. Historical document processing is one of its central pillars. The preservation of the bequest the present generation has inherited needs special attention and claims for the development of special tools and algorithms to overcome the difficulties posed the age: darkened

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HIP'11, September 16 - September 17 2011, Beijing, China  
Copyright 2011 ACM 978-1-4503-0916-5/11/09 ...\$10.00.

paper, stains, fungus, folding marks, worm attacks, torn-off parts, etc. Extracting information from such sources in a way that allows for easy access and knowledge correlation is a challenge of paramount importance.

HistDoc was conceived as a device independent software tool to run on PCs. The first version of HistDoc released in 2009 and described in reference [1] works in two different ways: user driven mode and standalone batch mode. In user driven mode the operator chooses which filters to use to enhance the document image. In standalone batch mode HistDoc uses the noise classifier presented in reference [2] specially tuned for historical documents, which automatically detects which undesirable artifacts are present in each document image and applies the suitable filtering technique. One should observe that such *a priori* noise classification is an important new feature in batch image processing.

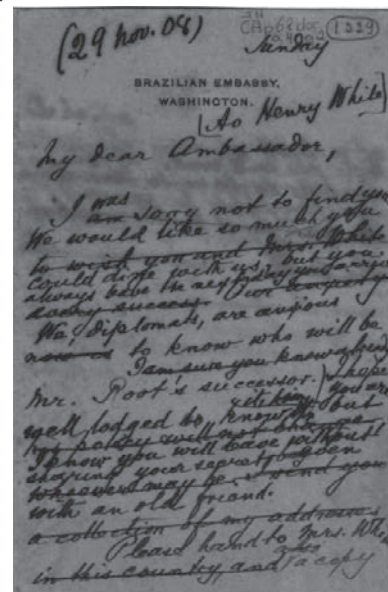


Figure 1 – Draft of a letter of Joaquim Nabuco.

HistDoc is indeed the result of a long evolution and two decades of research in historical document processing [3] within a pioneer initiative in Latin America. The environment described in reference [4] reports on the genesis of the HistDoc platform, an environment for processing images of historical documents developed for digitizing, enhancing and making available the bequest of the letters of Joaquim Nabuco, a Brazilian statesman, writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil (b.1861-d.1910). The Nabuco file of active and passive correspondence is of paramount importance to

understand the formation of the political and social structure of the countries in the Americas and their relationship with other countries. This rich file is kept by the Joaquim Nabuco Foundation [5] (a social science research institute in Recife - Brazil). Figure 1 presents a document from Nabuco bequest, the draft version of a letter of Joaquim Nabuco dated on 29 November 1908, to Ambassador Henry White.

About 10% of the documents in the Nabuco file were written on both sides of translucent paper allowing the verso writing to be visible on the front side, as may be observed in the document of Figure 1. Such noise makes the direct binarization yield an unreadable document and was first addressed in the literature in reference [4], which named the phenomenon “back-to-front interference”. Much later, the back-to-front interference was called *bleeding* [6] and *show-through* [7].

This paper presents the new version of HistDoc which encompasses several new features such as:

- Handling documents digitized with cameras.
- Processing monochromatic documents.
- Working in clusters and grids in batch mode.

This paper is organized as follows. Section 2 briefly sketches the functionality of the user driven mode of HistDoc. Section 3 details the use of the platform in cluster and grid modes for processing batches of documents. Conclusions and lines for further work are drawn at the final section.

## 2. HistDoc in User Driven Mode

The user driven mode of the HistDoc platform was built as an ImageJ [8] plugin, similarly to the first version of the platform. ImageJ, as an open code library, allows the developer to extract from it only the needed functionality in such a way that the developer may provide to ordinary user a HistDoc interface that looks independent from ImageJ. The difference of the new version rests on the increased functionality that allows for working with monochromatic documents. Figure 2 presents a screen shot of HistDoc under ImageJ.

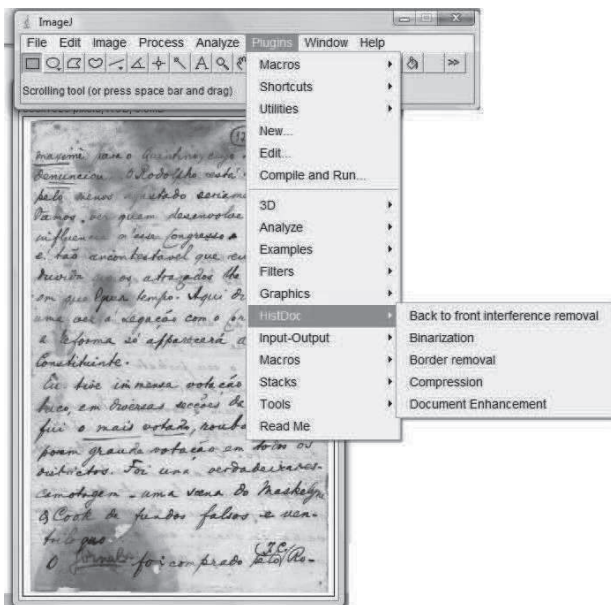


Figure 2 - Screen shot of the HistDoc plug-in in ImageJ

As one may observe in Figure 3, the present version of the HistDoc plug-in offers five different options, which appear in alphabetical order:

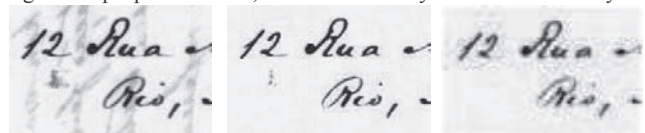
1. Back-to-front interference removal
2. Binarization
3. Border removal
4. Compression
5. Document Enhancement

The fact that HistDoc is now in ImageJ also allows the user to experiment with the different filters and other plug-ins already present in ImageJ. Each of the options in the HistDoc plug-in in ImageJ is now detailed.

### 2.1 Back-to-front Interference Removal

The technical literature presents several algorithms to remove the back-to-front noise, but no algorithm is good enough in all cases [9]. Depending on the degree of translucidity of the paper, the kind of the ink used in printing or writing, the porosity of the paper, etc. the interference may show itself stronger or weaker. Some algorithms perform better than others in different degrees of interference and even one chosen algorithm may perform better if its parameters are tuned to the intensity of the noise.

This version of HistDoc replaced the previous algorithms implemented for a newer and better one [10] which presents a new strategy to select and tune an algorithm to remove the back-to-front interference in color documents. It makes use of a set of neural classifiers to assess the intensity of the back-to-front interference and to adjust the parameters of the algorithm described in reference [11] to filter the noise out of a given a document. The neural classifier was implemented using Weka [12], an open-code classification tool developed at The University of Waikato, New Zeland. The blanks yielded by removing the artifact are filled in with pixels that correspond to the paper area in the document in such a way to provide the reader with “a natural” look of the document as if it were written on one side only. Figure 3 presents a sample of the results obtained by the algorithm proposed herein, in which one may observe its efficacy.



Original                      Filtered using [12]                      Filtered using [11]

Figure 3. Zoom in a document from Nabuco bequest with back-to-front interference filtered out using the algorithm described in reference [12] and the algorithm tuned with a neural classifier [11].

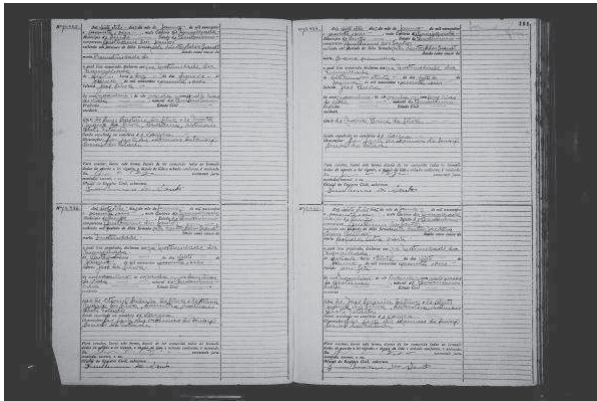
The new algorithm presented is also suitable for working in standalone batch mode as the neural classifier automatically sets the best parameters for filtering out the back-to-front interference.

### 2.2 Binarization

Document binarization is an important operation not only because a binary image is much smaller than its color counterpart but also due to most automatic transcription tools (OCR and ICR) pre-process their input images into grayscale or binary before character recognition. Reference [13] presents a survey of the most important binarization techniques applied to documents. HistDoc in user driven mode offers to the user 16 thresholding algorithms suitable for this sort of document:

- 11 global ([14], [15], [16], [17], [18], [19], [12], [20], [21], [22], [23]) and,
- 5 local ([24], [25], [26], [27], [28]).

Although the binarization algorithms above were already implemented in the first version of HistDoc, the algorithm presented in [10] for removing back-to-front interference may be

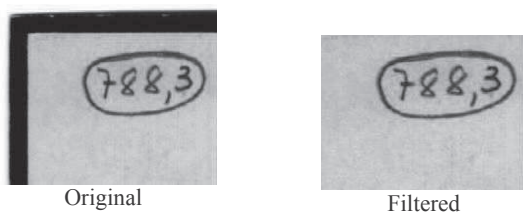


**Figure 5** – Two-page image of a “book” showing four Death Certificates from the Mormon-TJPE image set.

used as an important intermediate step for the binarization of documents with such noise.

## 2.3 Border Removal

Document digitalization either performed with scanners or cameras yield an image framed with some background which served of physical support to the document, an instance of which may be found in Figure 4 left. Such frame claims for larger space for storage and more bandwidth for network transmission. The visualization area in a device such as a CRT is wasted in exhibiting pixels that convey no information and ink or toner are used in printing such border noise. Besides that, the digitalization border has a serious deleterious impact in the quality of the image subjected to palette reduction. This brings important implication as most automatic transcription tools (OCR and ICR) pre-process their input images into grayscale or binary before character recognition.



**Figure 4** – Scanned document images: original and border removed.

Depending on the “nature” of the document HistDoc version 2.0 offers three different approaches to border removal:

- Border removal for monochromatic documents (either scanned or photographed).
- Border removal of color photographed documents.
- Border removal of color scanned documents.

The three options listed above apply different algorithms, which are detailed next.

### 2.3.1 Monochromatic Documents

Documents such as the ones from the Mormon-TDJP file processed by the *Thanatos* project [29] were digitized in grayscale but most processing is performed in their monochromatic version. Figure 5 presents an example of such a document.

The new algorithm [30] was incorporated to HistDoc version 2.0 and performs border removal in an efficient way. This algorithm may be used in the standalone version in batch mode.

### 2.3.2 Color Photographed Documents

Documents acquired with cameras almost always encompass in the image parts of the background framing the area of interest. The use of portable digital cameras for document image acquisition has become more frequent lately. The new version of HistDoc incorporated the border removal algorithm of PhotoDoc [31][32], a tool developed with the purpose of processing document images acquired with portable digital cameras.

Document images such as the one presented in Figure 6 pose an enormous degree of complexity for automatic border removal. The algorithm implemented in HistDoc asks for the user help either to confirm or adjust the frame boundary window automatically found. After the user confirmation, the algorithm makes perspective correction and crops the document image.



**Figure 6** – Example of color document acquired with portable digital camera on uneven textured color background.

Due to the iterative nature of the algorithm it should not be used in the standalone version of the HistDoc platform.

### 2.3.3 Color Scanned Documents

Documents such as the one shown in Figure 4 are often found. The document is framed by part of the lid of the scanner flatbed. Such “almost uniform” either “white” or “black” frame is detected and removed by the algorithm described in reference [33], which is implemented in HistDoc v. 2.0. Special attention is needed whenever the document background is white and the frame is also white. This case seldom occurs in the case of historical documents as paper aging is one of their features.

## 2.4 Compression

HistDoc offers the compression scheme described in reference [34], in which the image is decomposed and stored as a compressed monochromatic image together with the colors and texture of the different graphical elements in the document (paper, printing, signature, etc.). The basic principle adopted in this compression scheme is shown in Figure 7, this HistDoc

compressed file has extension “hdc”. The user may also save images with the several other file formats available in ImageJ (jpg, jpeg2000, png, tiff, etc), with and without losses.

The compression scheme presented here allows the user to have the “impression” of visualizing the original document, but the cost of storing it is much lower, basically the cost of storing a monochromatic image. The addition of a visualization plugin for the hdc format is at the end user allows network transmitting much faster than a “real” color one.

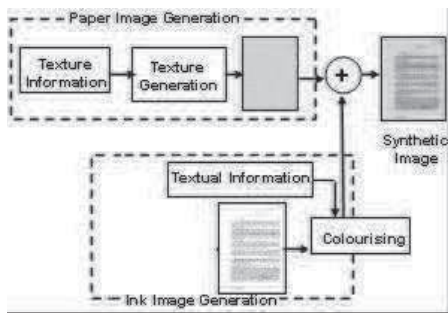


Figure 7 – Parametric generation of synthetic color document images for the hdc (HistDoc) fileformat.

## 2.5 Document Enhancement

One of the advantages of having HistDoc implemented as a plugin in ImageJ [8] is being able to use all the filters, tools, and other plugins already implemented in it for image filtering and enhancement.

Three new features were introduced for enhancing images in HistDoc version 2.0:

- Correcting book-binding distortion in scanned documents.
- Removing highlighting.
- Character recognition enhancement.

The algorithms implemented are briefly explained as follows and, at present may only be used in user driven mode.

### 2.5.1 Book-binding Warp Compensation

The digitalization of a bound document such as a book using a flatbed scanner causes a distortion similar to the one that may be observed in the image of Figure 8 (left) and better observed in the zoomed image shown to the right hand side of the same figure in which one may see that font width becomes narrower closer to the left margin (brochure). The book-binding warp not degrades OCR transcription, but also brings problems for image binarization.

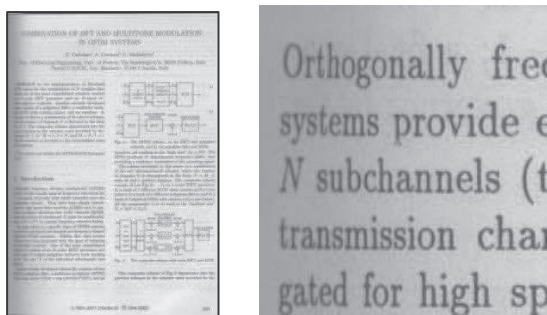


Figure 8 – Page of a bound volume scanned with a flatbed scanner (left). Zoom into part of the image (right).

HistDoc version 2.0 makes use of the recent algorithm presented in reference [35] to compensate book-binding warp. Although a similar phenomenon may be observed in the photographed images from the Mormon-TJPE image set (see Figure 5 left margin close to the brochure) the algorithm implemented [35] makes use of “shape-from-shading” information from the scanned image, thus it is unsuitable to handle such images.

### 2.5.2 Highlighting Removal.

Over the centuries, the interested readers often underlined texts to somehow emphasize parts of a text for further reference. There is a personal outlook in document highlighting. What one reader may stress and emphasize may be considered irrelevant to another. Thus, in general, highlighting may be perceived as “noise” physically damaging the document [36].

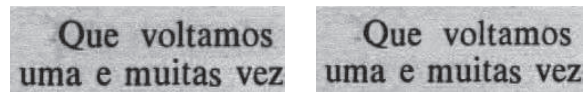


Figure 9 - Example of highlighted images of aged paper background on the left hand side and its removal using the algorithm described in [37]

Highlighting removal is a complex task, because the ink fades, sometimes non-uniformly, and interacts with the paper background. Figure 9 presents an example of a highlighted part of a printed document in aged paper background and the result of being processed with the algorithm presented in reference [37], which is able to suitably process several different colors of felt pens, such as the ones shown in Figure 10.

Highlight	Color
	Yellow
	Blue
	Green
	Orange
	Cyan
	Magenta

Figure 10 – Different color of highlighting that are filtered out in the HistDoc platform.

### 2.5.3 Character recognition enhancement.

Often historical documents, such as the part of the one shown in Figure 11, suffer physical damage such as thorn-off parts from handling or attacks caused by worms that “dig” holes in documents [36].

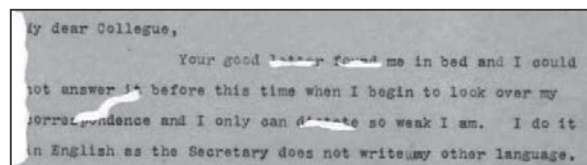


Figure 11 - Image from Nabuco's bequest with thorn-off regions and holes.

The physical noises listed seriously degrade OCR performance. The recent paper [38] presents a system to look for torn-off regions or holes and try to “complete” such areas with possible images in such a way as to maximize the probability of the correct

transcription of the word as a whole. This way, instead of performing character-to-character recognition as used in conventional OCR tools, the system proposed here infers a set of possible words and chooses the one with the highest probability to occur. Incomplete words in holes or torn off areas are completed taking into account the parts left of characters completing them with characters that may possibly “fit” the remaining parts. The OCR drives the choice of the most suitable part to fill in the holes. The choice of the most probable word may be helped by using a dictionary of terms already recognized in the document or in the file as a whole. The system showed satisfactory results in conjunction with the ABBY FineReader 10 Professional Edition [39].

### 3. HistDoc in Standalone Mode

Document digitalization of legated and historical files is happening at a very fast pace. Every day new and better algorithms appear to enhance the quality of images making possible content correlation and information extraction. In the last section some of the most recent algorithms that have already been introduced to the HistDoc platform were presented. The union of larger sets of documents to be processed together with more sophisticated image processing algorithms demand for an enormous amount of processing time that is impossible to be handled by human operators.

HistDoc version 2.0 provides the possibility of users to batch processing document images. A directory is passed as source for the image processing filters and another directory is indicated to store the processed files. There are three different working modes:

- **User selected filtering mode** - The user selects which filters and in what order they should be applied on each document image in the batch of documents in the source directory.
- **Automatic filtering mode** - HistDoc v.2.0 implemented image classifiers that with a high accuracy rate discover the “nature” of images and which noises are present automatically choosing which filters to use to enhance the batch of images. Filters are applied once and in a default sequence (Border detection and removal-Orientation and Skew correction-Back-to-front interference filtering).
- **Selected automatic filtering mode** - In this operating mode the user selects which filters will be active to be used, in which order and the number of times they are applied, subject to the noise presence being detected by the classifier. The noise classifier is also used to tune the filtering algorithms.

The automatic filtering modes in reality make use of two cascaded classifiers. The first classifier, originally described in reference [40] but enhanced with the feedback technique introduces in reference [41], discriminates between scanned and photographed documents. Then, a second classifier “investigates” which are the noises present at each document image. In some cases, the noise classifier goes further and is able to assess the suitable parameters to yield the best results. The block diagram of such classifier is shown in Figure 11 and is detailed in reference [2].

### 4. HistDoc in Clusters and Grids

Batch processing offers a great advance in meeting the enormous and growing demand for quality document image processing, overall in the case of historical ones. The use of

parallel and distributed architectures is an option that became reality with today technology. The new version of HistDoc borrowed from BigBatch [42, 43] the control mechanisms for those parallel architectures in clusters and grids. It is important to stress that the distributed HistDoc platform works in terms of image filtering and enhancement similarly to the standalone mode described in Section 3 of this paper. In what follows there are some details of the distributed mechanisms used in the HistDoc platform.

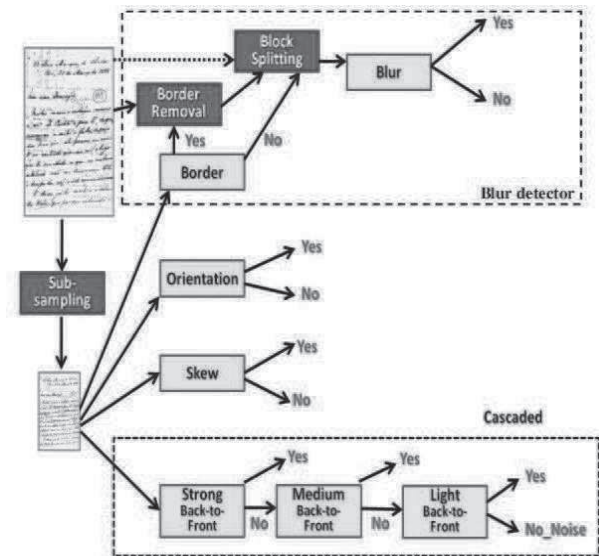


Figure 11. Noise classifier “architecture”

#### 4.1 The HistDoc Cluster Configuration

A cluster is composed by a collection of computers in a Local Area Network, called nodes, which work together in the execution of computationally demanding tasks that would not be feasible to execute on a single computer. Clusters may be formed from standard desktop computers, connected by an Ethernet network; however, high-performance clusters are usually created by the connection of specialized computers in special bus architectures. The level at which an application is programmed to run in the cluster can also vary from the manual separation of the problem into sub problems and allocation of sub problems to nodes, to high-level programming architectures in which the whole cluster can be considered as a single computer. Nodes in a cluster are normally homogeneous and often dedicated to execution of cluster tasks.

There is a wide variety of cluster software libraries and middleware programs that can be used to help managing tasks in a cluster, e.g. openMosix [44] and Microsoft Cluster Server [45]. It is more common that applications must be explicitly written with the cluster in mind, incorporating the division of tasks between nodes and the communication between them. The programming of cluster tasks often uses specialized libraries such as MPI [46] and OpenMP [47].

The support for distributing HistDoc tasks to nodes in a cluster configuration was custom written for this application, using the Scala programming language [48]. Scala is a functional and

object-oriented language that was designed to run in the Java Virtual Machine and interoperate with Java libraries and APIs. It was chosen because it was desirable to work with the Java platform, leveraging its portability and the availability of libraries; further, Scala includes good support for distributed programming using Actors [49]. Another reason for having the HistDoc application running over the Java platform is to ease integration with the grid component, as will be explained in the next subsection.

Nodes in the image processing application are divided into *worker* nodes and a single *master* node which coordinates the distribution of tasks between the workers. The computer where the main BigBatch application is executed is the master node, while worker nodes must execute a smaller component called the BigBatch Client Module. Communication between the nodes is done by message-passing, always from the master to the workers or from one worker to the master, never between workers. The master dynamically balances the load by distributing tasks to the available worker nodes, maintaining a list of tasks that need to be executed and available worker nodes. Whenever there are pending tasks and workers are available, the master assigns tasks to the workers in some arbitrary order (as nodes are homogeneous, it makes no sense to select one over another for a given task). A worker node that receives a task is marked as busy, and it stays in this state until the task is completed and a message is sent to the master to signal that; the master then marks the node as available again, adding it to the list of available nodes. This process continues until there are no more tasks to be executed. The experience with the BigBatch platform [43] showed that load-balancing is very simple, due to the homogeneous nature of the cluster architecture and the data-parallelism nature of the problem.

## 4.2 The HistDoc Grid Configuration

Computational grids are also formed from a collection of general-use computers that are coordinated to the execution of related tasks. The main difference between a grid and a cluster is that the latter tends to be established using dedicated resources that are local to a single organization, while the former may include non-dedicated, non-local computers as nodes. It is common for grid software to take over a workstation computer (that would normally be available to human users) to execute tasks while it is idle. Therefore, grids are a distributed computing environment that features lower coupling than what is expected of clusters.

The low-coupling between nodes and the distributed nature of processing makes the programming of applications over grids more complex and challenging than is the case with clusters. A special case of problems that can be solved with grid platforms are the ones whose sub problems are independent and need no communication between the nodes themselves. This class of applications is commonly called *bag-of-tasks* applications, and their execution is simpler to manage in grids. Taking advantage of that, a number of software systems have been developed to support the execution of this kind of tasks on computational grids, the so-called grid *middleware* systems. One such system is OurGrid [50], which was selected to be used in the support for grids provided by BigBatch and HistDoc. Document processing tasks generated by HistDoc fulfill the bag-of-tasks requirements.

OurGrid is both an open-source grid middleware and a grid infrastructure where sites may make available their computational resources from idle computers; in exchange, a participating site can obtain access to the computational resources from other sites,

whenever necessary. To organize the exchange of computational favors, OurGrid establishes a peer-to-peer network between interested sites, in which the “currency of exchange” is computational time. This is done to assure that participation in the grid and allocation of resources is fair, and the peer-to-peer network formed is called a “network of favors” [51]. Participation in the network of favors is optional and an organization may use the OurGrid middleware only internally, as is the case reported in this paper.

In the OurGrid solution there are three main components: **MyGrid**, the **Peer**, and the **UserAgent**. The MyGrid component is responsible for the management and scheduling of grid tasks – organized in collections called *jobs*. The Peer manages nodes in a site and the exchange of computational resources with other sites. Finally, the userAgent is a small program that must be installed in each node that will be part of the grid. The grid needs a node executing the MyGrid component and a node executing the Peer component (these two may execute in a single node), in addition to the userAgent executing in each worker node.

MyGrid is further subdivided into two modules: the scheduler and the replica executor. The scheduler is responsible for receiving new tasks from users and managing them, allocating nodes for their execution; it creates replicas of the tasks (if necessary) and communicates with the Peer requesting nodes for execution of the replicas. The nodes returned by the Peer may be local, or may be obtained from remote sites through the network of favors. The replica executor manages the execution of replicas of tasks and the sending of task results to the scheduler.

Currently, MyGrid works with two scheduling strategies: Workqueue with Replication [52] and Storage Affinity [53]. The first was designed for CPU-intensive applications, while the latter was created to improve the performance of applications that process large data sets.

A collection of tasks related to the same problem is called a job in OurGrid. A job is composed of independent tasks, each one composed of three phases: init, remote and final. These phases are executed on sequence, with the init and final phases being mostly used to transfer files needed for execution of the task; they are thus executed on the MyGrid node. The remote phase is executed in one or more worker nodes (depending on the replication strategy), and comprises the computation needed by the job. While executing a job, MyGrid requests nodes from the Peer to assign tasks to them.

Execution of the job is managed by the MyGrid component, which schedules tasks between the nodes made available by the Peer following the chosen scheduling method. This proceeds until all tasks have been executed. In the case of HistDoc, the application creates the job, based on the batch of document images that must be processed, and communicates this job to the MyGrid component. As both this component and the main HistDoc application run over the Java platform, this communication is easily performed using the Java Remote Method Invocation (RMI) mechanism.

## 5. Conclusions and lines for further work

The HistDoc platform is the result of a two decade evolution of algorithms and filters to process images of historical documents. HistDoc version 2.0 kept the ease of use and friendliness of the



first version of HistDoc. It encompasses now a much wider range of functionalities as it now:

- Is able to work with monochromatic documents removing framing noisy borders, correcting orientation and skew, and removing salt-and-pepper noise in images.
- Encompasses possibly the best algorithm in the literature for removing back-to-front interference in documents.
- Processes images of documents acquired with portable digital cameras allowing in user driven mode to find framing borders even with textured color background.
- Corrects book binding distortion in scanned monochromatic documents for which the shape-from shading hypothesis is applicable.
- Offers a tool to enhance character and word recognition in damaged text areas.

The HistDoc platform was developed as a plug-in in ImageJ, an open source portable Java library freely available. HistDoc runs on the users' PC and has the advantage of the great portability of Java. Aiming to speed-up the document processing phase, some of the algorithms are implemented in C.

Besides all that, HistDoc offers four different operating modes, being completely innovative in its kind: user driven, standalone automatic batch filtering mode in uniprocessors, cluster parallel document processing, and distributed grid-based image filtering. In the autonomous version of the HistDoc platform, statistical and neural classifiers were incorporated allowing "analyzing" the "nature" and noises present in each document prior to image filtering. The output may be either a binary image, a color image in the same file format of the input image or a parametrically compressed image which closely resembles the original one but is far more efficiently compressed.

Several lines may be followed to provide further improvements to HistDoc filters and environment. Some of them are:

- Being able to easily erase marks and stains from the digital version of the document.
- Interfacing with OCRs to automatically transcribe or find keywords in documents. The interfacing of Tesseract [54] with HistDoc is on progress.
- Incorporating into HistDoc some of the functionalities of Gamera [55] another free platform of similar purpose is also a possibility. Preliminary tests performed with Gamera showed that although its OCR mechanism presents a much lower recognition performance than Tesseract, it allows the user to train the OCR recognizer with new font types, for instance, which may be of interest in some files of historical documents in which all documents were typed using a particular machine. Gamera is implemented in Python and C++ and is slightly faster than the current version HistDoc which is implemented as an ImageJ plugin in Java.

The executable code for the HistDoc v.2.0 platform is available by requesting it via e-mail to any of the authors of this paper.

## 6. Acknowledgements

Research presented here is partly sponsored by CNPq-Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico, Brazilian Government.

## 7. REFERENCES

- [1] G. F. P e Silva, R. D. Lins, J. M. M. da Silva. HistDoc - A Toolbox for Processing Images of Historical Documents, ICIAR 2010, LNCS v.6112, p.1 – 11. Springer Verlag, 2010.
- [2] R. D. Lins, G. F. P. Silva, S. Banergee, A. Kuchibhotla, M. Thielo. Automatically Detecting and Classifying Noises in Document Images. SAC 2010, v. 1. p. 33-39. ACM Press, 2010.
- [3] R.D.Lins. Nabuco - Two Decades of Document Processing in Latin America, Journal of Universal Computer Science, v. 17(1), pp. 151-161, 2011.
- [4] R. D. Lins, L. G. Rosa, L. R. França Neto, M. S. Guimarães Neto. An Environment for Processing Images of Historical Documents. Microprocessing & Microprogramming, pp. 111-121, North-Holland, 1994.
- [5] FUNDAJ: [www.fundaj.gov.br](http://www.fundaj.gov.br), accessed on 20/06/2011.
- [6] R. Kasturi, L. O'Gorman and V. Govindaraju, Document image analysis: A primer, Sadhana, (27):3-22, 2002.
- [7] G.Sharma, Show-through cancellation in scans of duplex printed documents, IEEE Trans. Image Processing, v.10(5):736-754, 2001.
- [8] IMAGEJ: <http://rsbweb.nih.gov/ij/>, accessed on 20/06/2011.
- [9] R.D. Lins; J. M. M. da Silva; F. M. J. Martins. Detailing a Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents. Journal of Universal Computer Science, v. 14, pp. 299-313, 2008.
- [10] G. F. P e Silva, R. D. Lins, J. M. M. da Silva, S. Banergee, A. Kuchibhotla, M. Thielo. Enhancing the Filtering-Out of the Back-to-Front Interference in Color Documents with a Neural Classifier. ICPR 2010. pp: 2415-2419. IEEE Press.
- [11] J. M. M. da Silva; R. D. Lins; F. M. J. Martins; R. Wachenchauser. A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference. Journal Universal Computer Science, v.14, p. 299-313, 2008.
- [12] Weka 3: Data Mining Software in Java, website <http://www.cs.waikato.ac.nz/ml/weka/>.
- [13] M. Sezgin and B. Sankur. Survey over Image Thresholding Techniques and Quantitative Performance Evaluation. Journal of Eletronic Imaging, 13(1), pp 145-165, 2004;
- [14] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram". G. Models I. Process. 29, 273-285, 1985.
- [15] E. Kavallieratou and H. Antonopoulou, "Cleaning and Enhancing Historical Document Images", Intelligent Vision Systems, LNCS 3708:pp. 681-688, Springer-Verlag, 2005.
- [16] A. Khashman and B. Sekeroglu. "A Novel Thresholding Method for Text Separation and Document Enhancement", 11th Panhellenic Conf. on Informatics, 18-20 May 2007.
- [17] J. Kittler and J. Illingworth, "Minimum error thresholding," Pattern Recognition, 19, 41-47, 1986.
- [18] C. A. B. Mello and R. D. Lins, "Generation of images of historical documents by composition". ACM Document Engineering 2002, McLean, VA, USA.
- [19] N. Otsu. "A threshold selection method from gray level histograms". IEEE Trans. Syst. Man Cybernetics SMC-9, 62-66, 1979.

- [20] T. W. Ridler and S. Calvard, "Picture thresholding using an iterative selection method," *IEEE Trans. Syst. Man Cybern.* SMC-8, 630-632, 1978.
- [21] J. M. M. da Silva, R. D. Lins and V. C. da Rocha Jr. "Binarizing and Filtering Historical Documents with Back-to-Front Interference", *ACM-SAC'06*, ACM Press, 2006.
- [22] L. U. Wu, M. A. Songde, and L. U. Hanqing. "An effective entropic thresholding for ultrasonic imaging". *ICPR'98: Intl. Conference Pattern Recognition*, pp. 1522-1524, 1998.
- [23] J. C. Yen, F. J. Chang, and S. Chang. "A new criterion for automatic multilevel thresholding". *IEEE Trans. Image Process.* IP-4, 370-378, 1995.
- [24] J. Bernsen. "Dynamic thresholding of gray level images". *ICPR'86: Proc. Intl. Conf. Patt. Recog.*, pp. 1251-1255, 1986.
- [25] W. Niblack. "An Introduction to Image Processing". pp. 115-116, Prentice-Hall, 1986.
- [26] P. W. Palumbo, P. Swaminathan, and S. N. Srihari. "Document image binarization: Evaluation of algorithms". *Proc. SPIE* 697, 278-286, 1986.
- [27] J. Sauvola and M. Pietaksinen. "Adaptive document image binarization". *Pattern Recogn.* 33, 225-236, 2000.
- [28] J. M. White and G. D. Rohrer. "Image thresholding for optical character recognition and other applications requiring char. image extraction". *IBM J.Res.Dev.* 27(4):400-411, 1983.
- [29] A. B. S. Almeida, R. D. Lins and G. F. P. e Silva. *Thanatos - Automatically Retrieving Information from Death Certificates in Brazil*, HIP 2011, ACM Press, 2011.
- [30] A. de A. Formiga and R. D. Lins. *Efficient Removal of Noisy Borders of Monochromatic Documents*. *International Conference on Image Analysis and Recognition*, 2009, LNCS v.5627. p.158 – 167, Springer Verlag, 2009.
- [31] G. F. P. e Silva and R. D. Lins. *PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras*, *Camera Based Document Analysis and Recognition*, 2007, Curitiba (Brazil). *Proceedings of CBDAR 2007*. IAPR Press, 2007. p.107 – 115
- [32] R. D. Lins, A. R. G. e Silva, G. F. P. e Silva. *Enhancing Document Images Acquired Using Portable Digital Cameras*. *ICIAR 2007*, v.LNCS. p.1229 – 1241Springer Verlag, 2007.
- [33] A. R. G. e Silva and R. D. Lins. *Background Removal of Document Images Acquired Using Portable Digital Cameras*, *ICIAR 2005*, LNCS 3656, pp.278-285, Springer Verlag, 2005.
- [34] J. M. M. da Silva and R. D. Lins. *Color Document Synthesis as a Compression Strategy*. *ICDAR 2007*. v. 1. p. 466-470. IEEE Press, 2007.
- [35] R. D. Lins, D. M. Oliveira, G. Torreão, J. Fan, M. Thielo. *Correcting Book Binding Distortion in Scanned Documents*. *ICIAR 2010*, LNCS 6112. p. 355-365. Springer Verlag, 2010.
- [36] R. D. Lins. *A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises*. *ICIAR 2009*. LNCS v. 5627. p. 844-854, Springer Verlag, 2009.
- [37] R. S. Barbosa, R. D. Lins, V. M. de S. Pereira. *Using Readers' Highlighting on Monochromatic Documents for Automatic Text Transcription and Summarization*, *ICDAR 2011*, Beijing, IEEE Press, 2011.
- [38] G. F. P. e Silva and R. D. Lins. *An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents*. *ICDAR 2011*, Beijing, Sep., IEEE Press, 2011.
- [39] ABBYY FineReader 10 Professional Editor, <http://finereader.abbyy.com/>.
- [40] G. F. P. e Silva, R. D. Lins, B. Miro, S. J. Siemke, M. Thielo. *Automatically Deciding if a Document was Scanned or Photographed*. *Journal of Universal Computer Science*. v.15, p.3364 - 3366, 2009.
- [41] R. D. Lins, G. F. P. e Silva, S. J. Siemske, *Automatically Discriminating between Digital and Scanned Photographs*. *ICDAR 2011*, Beijing, IEEE Press, 2011.
- [42] R. D. Lins, B. T. Ávila, A. de A. Formiga. *BigBatch: An Environment for Processing Monochromatic Documents* *ICIAR 2006*, v.4142. p.886 – 896, Springer Verlag, 2006.
- [43] G. G. de Mattos, A. de A. Formiga, R. D. Lins. *BigBatch: a document processing platform for clusters and grids*. *23rd ACM Symposium on Applied Computing*, ACM Press, ACM-SAC 2008. v.I. p.434 – 441, ACM Press, 2008.
- [44] openMosix, <http://openmosix.sourceforge.net/>. (27.06.2011).
- [45] Microsoft Cluster Server. Visited in 29.06.2011. <http://www.microsoft.com/windowsserver2008/en/us/default.aspx>.
- [46] M. Snir and W. Gropp. 1998. *MPI: The Complete Reference*, 2nd. Ed., MIT Press.
- [47] R. Chandra, R. Menon, *et al.* 2000. *Parallel Programming in OpenMP*, Morgan Kaufmann.
- [48] M. Odersky. *Scalable Component Abstractions*. *OOPSLA 2005*: pp. 41-57. 2005.
- [49] P. Haller and M. Odersky, *Event-Based Programming without Inversion of Control*. LNCS. 4228, pp. 4-22. Springer Verlag, 2006.
- [50] W. Cirne, *et al.* "Labs of the World, Unite!!!" *Journal of Grid Computing*, v. 4, n. 3, pp.225-246. 2006.
- [51] N. Andrade, F. Brasileiro, W. Cirne, and M. Mowbray, *Discouraging Free-riding in a Peer-to-Peer Grid*. *Proceedings of the 13<sup>th</sup> IEEE International Symposium on High-Performance Distributed Computing (HPDC13)*.
- [52] D. Paranhos, W. Cirne, W and F. Brasileiro. *Trading cycles for information: Using replication to schedule bag-of-tasks applicatoins on computational grids*. *Euro-Par 2003*, LNCS v. 2790, pp. 169-180, Springer Verlag, 2003.
- [53] E. Santos-Neto, W. Cirne, F. Brasileiro, and A. Lima, *Exploiting replication and data reuse to efficiently schedule data-intensive applications on grids*. LNCS, v. 3277, pp. 210-232. Springer Verlag, 2005.
- [54] Tesseract OCR - <http://code.google.com/p/tesseract-ocr/> ; accessed on 27.06.2011.
- [55] Gamera Project (<http://gamera.informatik.hsrn.de/>) accessed on 27.06.2011.

## **A.5 Publicações sobre remoção de Interferência Frente e Verso**

(SILVA et al., 2009) J. M. Silva, R. D. Lins and G. F. P. Silva. Enhancing the Quality of Color Documents with Back-to-Front Interference. *Image Analysis and Recognition*, 1rd, Ed. Springer, pp: 875-885.

(SILVA et al., 2010a) - G. F. P. Silva e R. D. Lins; S. Banerjee; A. Kuchibhotla; M Thielo. Enhancing the Filtering-out of the Back-to-Front Interference in Color Documents with a Neural Classifier. In: *International Conference on Pattern Recognition*, vol.1, pp: 2415-2419.

# Enhancing the Quality of Color Documents with Back-to-Front Interference

João Marcelo Silva, Rafael Dueire Lins, Gabriel Pereira e Silva

Universidade Federal de Pernambuco, Brazil  
joaommsilva@gmail.com, rdl@ufpe.br, gfps.cin@gmail.com

**Abstract.** Back-to-front, show-through, or bleeding are the names given to the overlapping interference whenever a document is written (or printed) on both sides of a translucent paper. Such interference makes more difficult, if not impossible, document transcription and binarization. This paper presents a new technique to filter out such interference in color documents, enhancing their readability.

**Keywords:** Back-to-Front interference, Bleeding, Show-through, Document Enhancement.

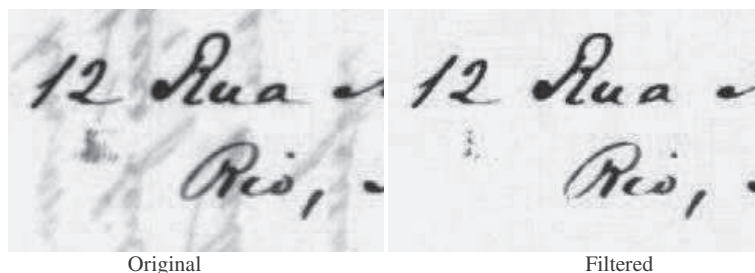
## 1 Introduction

In beginning of the 1990s, the historically relevant file of 6,500 letters by Joaquim Nabuco were digitalized through the partnership between the Joaquim Nabuco Foundation and the Federal University of Pernambuco. About 10% of the scanned document images presented a feature not previously described in the literature, which was called back-to-front interference [5]. Much later, other authors addressed the same phenomenon and called it bleeding [4] and show-through [8].

The back-to-front interference occurs whenever the verso face content of a document becomes visible on its front. Such interference appears in a document, whenever it is written (or printed) on both sides of translucent paper (see Fig. 1 - left). The motivation for removing such artifact is that it degrades document transcription and the binarization process as front and verso images often overlap yielding an unreadable monochromatic document. In the case of historical documents, ageing is a complicating factor as paper darkens overlapping the RGB-distributions of the ink on each side and paper.

This article presents a new filtering strategy to remove back-to-front interference in images of color documents. The idea herein is to discriminate the interference area and replace interference pixels with blank paper ones in such a way as to remove the interference providing a "natural" look under visual inspection. Such fulfillment is done by a linear interpolation of the pixels in surrounding areas. Fig. 1 provides a sample of the results obtained by the algorithm proposed herein, in which one may witness its effectiveness.

Section 2 of this paper details the new filtering strategy. The results and analyses are presented in Section 3. Finally, Section 4 draws our conclusions and guidelines for further works.



**Fig. 1.** Zoom into a document from the Nabuco bequest with back-to-front interference, filtered using the proposed strategy.

## 2 The filtering system

This section presents the new strategy to remove the back-to-front interference from images of color documents. First, one discriminates the area corresponding to such interference; in a second step, the interference pixels are replaced by others that resemble to the paper pixels, removing the back-to-front interference from the resulting image.

### 2.1 Discrimination of the noisy pixels

To find the interference area, the segmentation algorithm by Silva-Lins-Rocha [9] is used twice: first, to separate the text from the rest of document, and second, to highlight the interference from the paper. That algorithm is an entropy-based global algorithm that uses the gray-level document image as an intermediate step to chop-off the gray-level histogram in three different areas of interest (see Fig. 2), as explained later on.

The empirically found loss factor ( $\alpha$ ) is a parameter of the segmentation algorithm that yields a better statistical adjustment between the distributions of the original and binarized images, based on the Shannon entropy [1]. For the second application of Silva-Lins-Rocha algorithm, one adopts a constant ( $\alpha=1$ ) factor, ensuring a better separation between interference and paper distributions.

Summarizing, to detect the interference area:

1. Silva-Lins-Rocha segmentation algorithm is applied to separate the front ink from the rest of document (see Fig. 3a and 3b);
2. The same algorithm, with the new loss factor value, is applied on the (paper+interference) image to separate the interference ink from the paper (see Fig. 3c and 3d), yielding a blank sheet of paper with white holes where there was ink and the verso ink interference in the original document image.

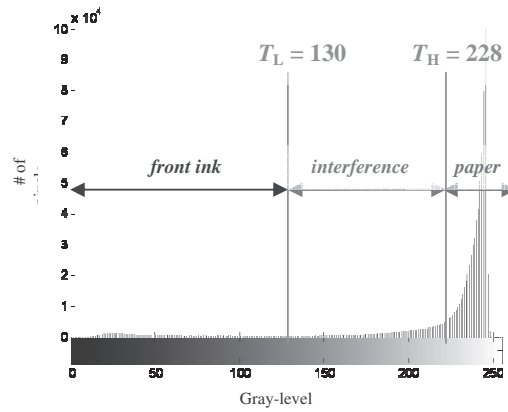


Fig. 2. Image histogram of document with back-to-front interference - segmentation details.

To illustrate the process, in Fig. 2 the first threshold,  $T_L$ , is obtained by the first application of the Silva-Lins-Rocha algorithm and the second threshold,  $T_H$ , by the second. The pixels for which their gray-levels are less than  $T_L$  are classified as ink of the front face. The pixels with gray-level greater than the  $T_H$  are classified as belonging to the paper. Pixels with gray-levels between  $T_L$  and  $T_H$  are discriminated as interference.

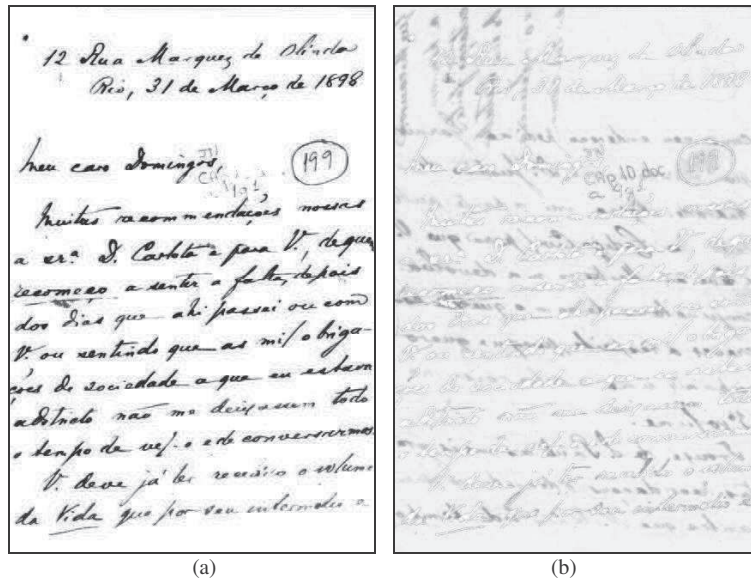
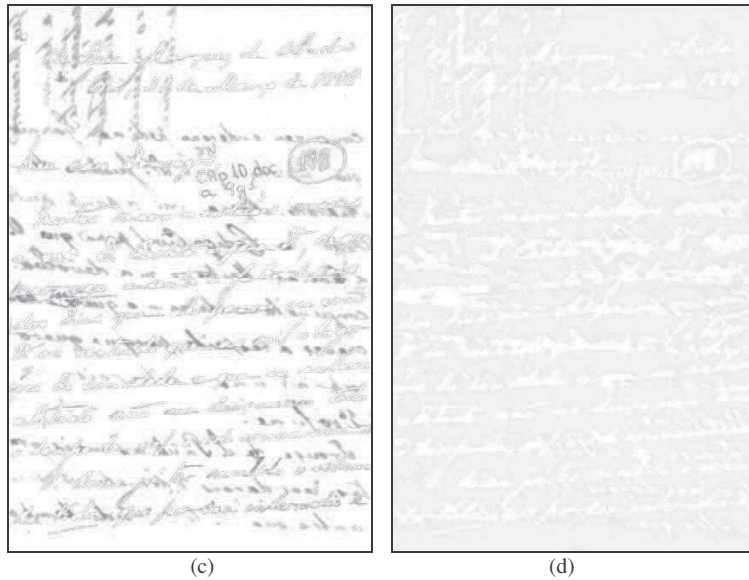


Fig. 3. Image segments of a document with back-to-front interference: (a) ink of the front face and (b) paper with interference. Image segments of Figure 3b: (c) interference and (d) paper.



**Fig. 3.** Image segments of a document with back-to-front interference: (a) ink of the front face and (b) paper with interference. Image segments of Figure 3b: (c) interference and (d) paper.

## 2.2 Fulfillment of the Blank Areas

The process proposed here uses a "linear" interpolation to fill in the blank pixels that originally corresponded to the interference area. Two binary masks are defined: TEXT and INTERF. The first one identifies the pixels from the ink of the front text (see Fig. 4a); the second one highlights the interference area (see Fig. 4b). One could assume that only the INTERF mask would be sufficient to the fulfillment process, because the pixels to be replaced "are known already". Some difficulties appear, however.

The key idea is to replace the colors of the noisy pixels with colors as close as possible to the paper in their neighborhood. This is achieved by interpolation, using the colors of the pixels that surround the area to be filled in. There is still the need to remove some of the vestigial shades surrounding the ink pixels in the resulting image; otherwise those pixels will "damage" the interpolation process, bringing in noisy dark colors to the interference area. To solve this problem, one should apply a "dilate" morphological expansion operation to both masks, with that, the text and interference contours will be properly classified as "text" and "interference", respectively (see Fig. 5a and 5b).

As mentioned earlier on, the pixels that are used in the interpolation process are surrounding the interference area and with the pixels belonging only to the paper. This mask, PAPER, is obtained by the complement of the logical OR operation between the TEXT and INTERF dilated masks (see Fig. 5c).

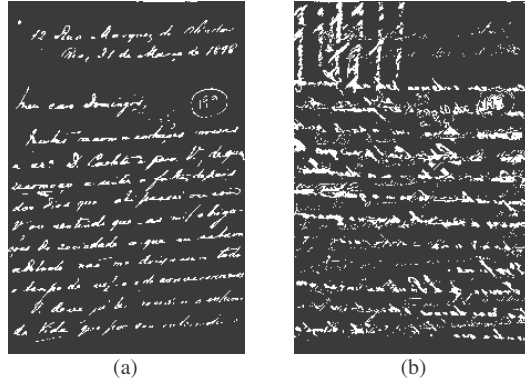


Fig. 4. Masks that identify (a) the text and (b) the interference.

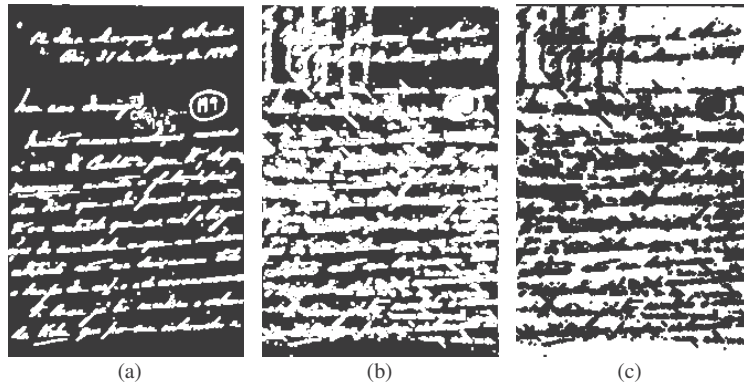


Fig. 5. Dilated masks: (a) text (T) and (b) interference (I) (c) T or I.

Now, the interpolation process is presented. Let the coordinates be as depicted in Fig. 6:

- $(x_0, y_0)$  of a pixel  $P$  from the interval to be interpolated;
- $(x_0, y_1)$  of pixel  $P_N$  – first pixel north  $P$ ;
- $(x_0, y_2)$  of pixel  $P_S$  – first pixel south  $P$ ;
- $(x_1, y_0)$  of pixel  $P_W$  – first pixel west  $P$ ;
- $(x_2, y_0)$  of pixel  $P_E$  – first pixel east  $P$ ,

Where  $i_C(x, y)$  is the value of the component  $C$  (R, G or B) of the pixel  $(x, y)$ . The intensity of the interpolated pixel ( $P$ ) is given by

$$i_C(x_0, y_0) = \frac{d_4 \mathcal{X}_1 + d_3 \mathcal{X}_2 + d_2 \mathcal{X}_3 + d_1 \mathcal{X}_4}{d_4 + d_3 + d_2 + d_1}, \quad (1)$$

where the  $i_k$  and  $d_k$  ( $k = 1, \dots, 4$ ) represent the intensities and the distances from the



pixels –  $P_N$ ,  $P_S$ ,  $P_W$  and  $P_E$  – to  $P$ , sorted by increasing distances. For example, the closest pixel to  $P$  has distance  $d_1$  and intensity  $i_1$ , the second closest one has distance  $d_2$  and intensity  $i_2$ , and so on.

The distance between any two pixels A e B with coordinates  $(x_a, y_a)$  and  $(x_b, y_b)$ , is the standard Euclidean distance:

$$d_{A,B} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}.$$

Equation 1 calculates a weighed mean, where the intensity of the nearest pixel from the pixel  $P$  has the greatest weight. This is reasonable, because in a neighborhood, generally, the closer a pixel is from another, the more alike they should look. Fig. 7b shows the result of the application of the proposed filtering strategy applied to the image in Fig. 7a.

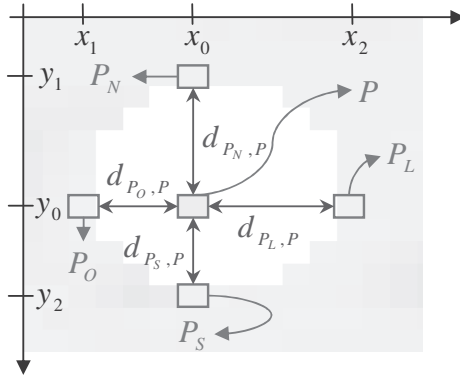
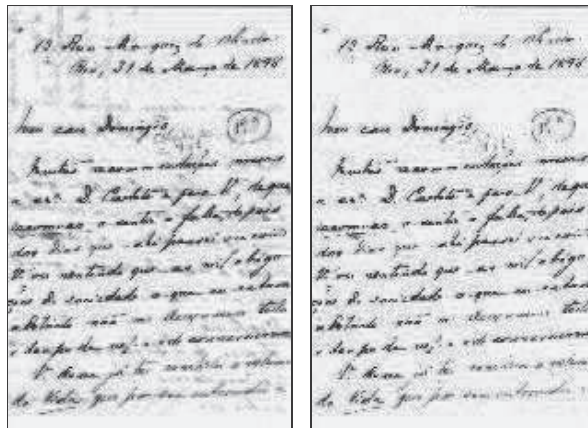


Fig. 6. Interpolation process.



(a)

(b)

Fig. 7. Images: (a) original and (b) filtered by the new strategy proposed here.

### 3 Results and analysis

The proposed algorithm was tested in a set of 260 images from the Joaquim Nabuco bequest of digitalized documents [2], yielding good results. Evidences of the efficiency of the new filtering technique are shown in Fig. 7, 8 and 9, as the back-to-front interference was removed yielding a more readable document with a “natural” look.

Fig. 8, 9, and 10 provide the results of using different strategies, amongst them using as fulfilment for the blanks the result of the interpolation based on Laplace’s equation (the MATLAB function “*roifill*” was used). The third alternative is one of the strategies proposed by Castro and Pinto [2] that uses the algorithm by Sauvola and Pietikainen [7] which define a mask that identifies the pixels of the foreground and background objects. The final image is obtained through keeping the object pixels and replacing the background pixels with the average of the colours of the pixels in that class. The latter strategy yielded the best results in [2].

The two strategies proposed herein yielded very similar quality results. However, the one based on Laplace interpolation leaves the filled-in area look undesirably uniform with a “flat” colour. On the other hand, the linear interpolation yields a residual pattern of vertical/horizontal stripes.

The strategy proposed by Castro and Pinto [2] aims to yield a uniform paper surface with unchanged text, while the ones presented here try to remove only the interference, keeping the pixels from the paper and text unchanged.

However, in the very few images in the Nabuco file that the back-to-front interference looks very “blurred” (see Fig. 10), the proposed algorithm did not perform too well.

The effective detection of whole interference is not a trivial task. Even when “almost all interference” is detected (that was archived making a greater dilatation in INTERF mask) the area to be filled is large (because the interference is scattered). With a larger area to be filled in, the interpolation process proposed here does not yield a “natural” aspect in the final image. This occurs with the Laplace interpolation, also. Fig. 11 illustrates that problem. The first and second image contains the same part observed in Fig. 10; however, it corresponds to the image filtered by the new strategy using the INTERF mask with a greater dilatation. If one observes the Fig. 10 and 11a, one will see that such part was enhanced. On the other hand, if one takes another part (Fig. 11b), one will evidence the problem that appears when one tries to interpolate a “relatively large” area. To reduce such a problem, one may try to interpolate a larger number of pixels in a larger “neighbouring area”.

### 4 Conclusions and lines for further work

This paper proposes a new strategy for filtering the back-to-front interference from images of colour documents. Such system uses the segmentation algorithm proposed in reference [8] twice to discriminate the noisy pixels. After the discrimination phase, the pixels that margin the blank areas are interpolated. The result of interpolation step

by step fills in the blank spaces in the interference area. The proposed algorithm yielded satisfactory results in 260 images from Nabuco bequest.

There are several lines to improve the results obtained here. One of them is to instead of using the same dilatation filter in all images to tune it according to the blur factor in each image. Ways of measuring the degree of interference dispersion (blur) are being analyzed by measuring the gradient between the interference and paper.

Another aspect not mentioned before is the rise of high-frequency components in the resulting image. This occurs because new intensity variations may be introduced in the blank fulfilment process. To avoid such problem, one could verify the maximum frequency that appears in the original document, and with that, use a low-pass filter in the final image to smooth to transitions between interpolated and text areas, bringing a more natural aspect to the final document.

Along the lines for further work, the authors intend to compare the strategies proposed herein and the work of Sharman [8] and Nishida and Suzuki [6]. Sharman makes use of the information on both sides of the document implementing a mirror transformation as suggested in [4]. The first step in Sharman solution is image alignment, which is extremely difficult to be performed adequately overall in the case of documents that were folded, as already pointed out in [4]. Sharman presents no solution to this problem, thus the applicability of his solution is still to be seen. The strategy proposed by Nishida and Suzuki [6] starts by performing a border detection to discriminate the text from its background. Observing the image presented in Fig. 8, one may say that such strategy is no good for that image, as it is most likely that the interference would be classified as object. The implementation of both algorithms is needed to allow further conclusions and a fair comparison with the results obtained here.

**Acknowledgments.** Research reported herein was partly sponsored by CNPq - Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico and CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazilian Government. The authors also express their gratitude to the Fundação Joaquim Nabuco, for granting the permission to use the images from Nabuco bequest..

## References

1. N. Abramson, "Information Theory and Coding", McGraw-Hill Book Co, 1963.
2. P. Castro, J. R. C. Pinto: "Methods for Written Ancient Music Restoration". Proceedings of ICIAR 2007: 1194-1205.
3. FUNDAJ: [www.fundaj.gov.br](http://www.fundaj.gov.br)
4. R. Kasturi, L. O'Gorman and V. Govindaraju, "Document image analysis: A primer", *Sadhana*, (27):3-22, 2002.
5. R. D. Lins, *et al.* "An Environment for Processing Images of Historical Documents. Microprocessing & Microprogramming", pp. 111-121, North-Holland, 1993.
6. H. Nishida, T. Suzuki: "A Multiscale Approach to Restoring Scanned Color Document Images with Show-Through Effects", Proceedings of ICDAR'03: 584-588.
7. J. Sauvola, M. Pietikainen: "Adaptive document image binarization", *Pattern Recognition* 33(2) (February 2000) 225-236.
8. G. Sharma, "Show-through cancellation in scans of duplex printed documents", *IEEE*

Trans. Image Processing, v10(5):736-754, 2001.

9. J. M. M. da Silva; R. D. Lins; F. M. J. Martins; R. Wachenchauser. "A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference". Journal of Universal Computer Science, v. 14, p. 299-313, 2008.

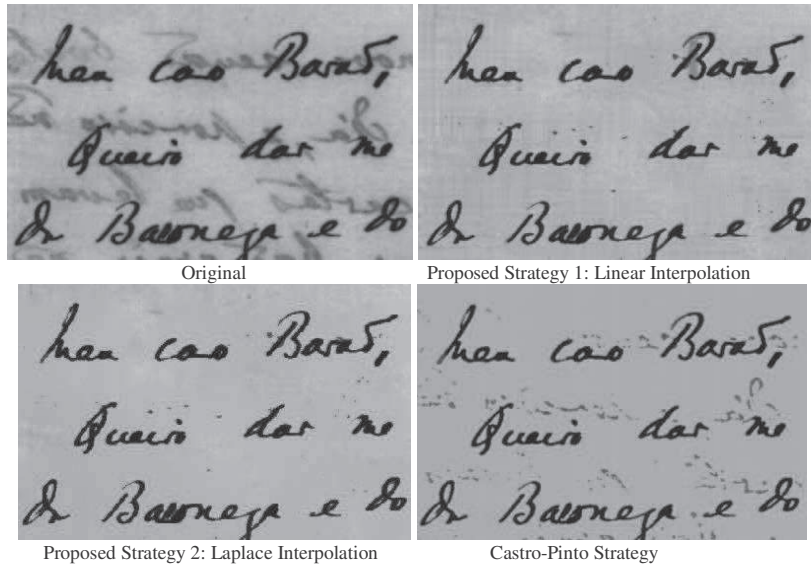


Fig. 8. Parts of documents from the Nabuco file: original and filtered.

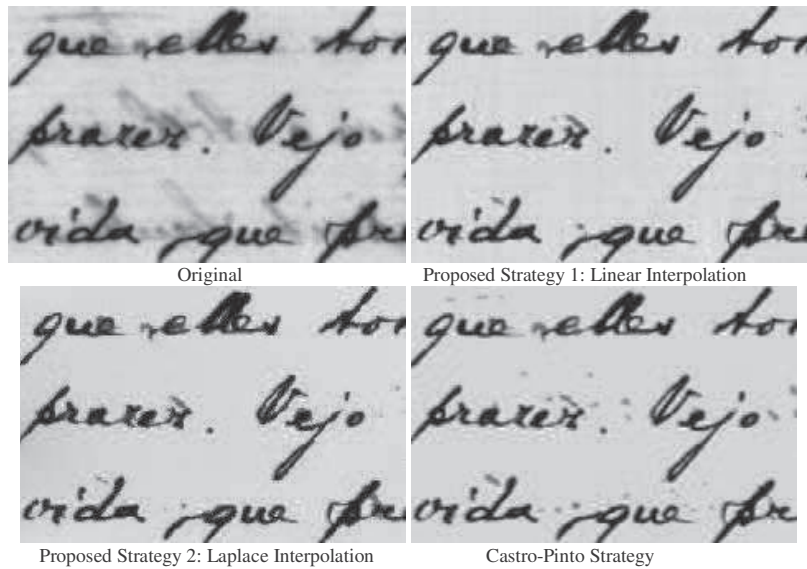


Fig. 9. Parts of documents from the Nabuco file: original and filtered.

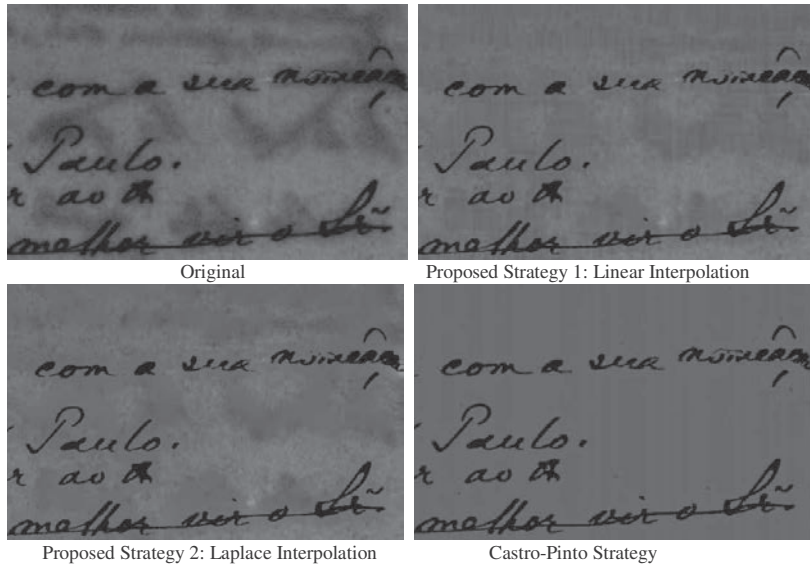
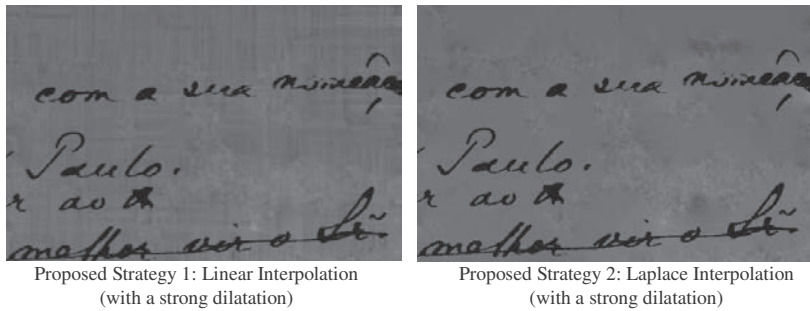
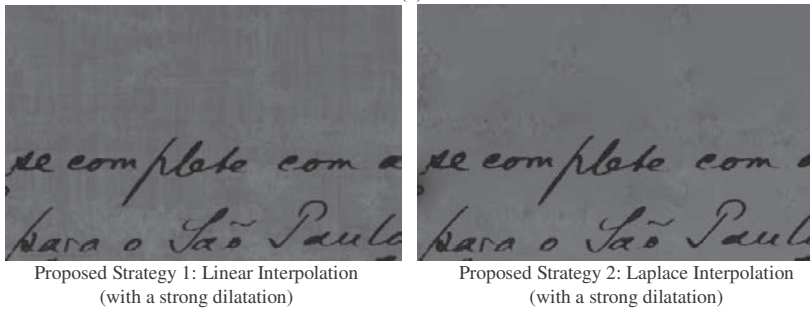


Fig. 10. Parts of documents from the Nabuco file: original and filtered.



(a)



(b)

Fig. 11. (a) Fig. 10, filtered with the new strategy, using a stronger dilatation. (b) Another part of the same document.

## Enhancing the Filtering-out of the Back-to-Front Interference in Color Documents with a Neural Classifier

G.F. P e Silva<sup>(1)</sup>, R. D. Lins<sup>(1)</sup>, J.M. Silva<sup>(1)</sup>, S. Banerjee<sup>(2)</sup>, A. Kuchibhotla<sup>(2)</sup> and M. Thielo<sup>(3)</sup>  
<sup>(1)</sup> UFPE-Brazil, <sup>(2)</sup> HP Labs.-India, <sup>(3)</sup> HP Brazil R&D-Brazil

{ gabriel.psilva, rdl}@ufpe.br, {serene.banerjee, anji, marcelo.resende.thielo}@hp.com

### Abstract

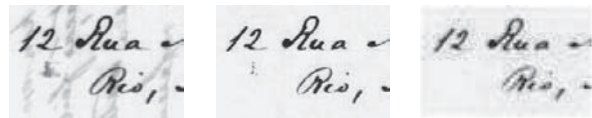
*Back-to-front, show-through, or bleeding are the names given to the interference that appears whenever one writes or prints on both sides of translucent paper. Such interference degrades image binarization and document transcription via OCR. The technical literature presents several algorithms to remove the back-to-front noise, but no algorithm is good enough in all cases. This article presents a new technique to remove such noise in color documents which makes use of neural classifiers to evaluate the degree of intensity of the interference and besides that to indicate the existence of blur. Such classifier allows tuning the parameters of an algorithm for back-to-front interference and document enhancement.*

### 1. Introduction

At the beginning of the 1990s, the important file of over 6,000 private letters of Joaquim Nabuco, a statesman, writer and diplomat, one of the leading figures of the freedom of black slaves in Brazil and the first Brazilian ambassador to the US, were digitized in a joint preservation effort between the Joaquim Nabuco Foundation [2] and the Universidade Federal de Pernambuco. About 10% of the images presented a noise which had not been previously described in the technical literature, which was called “back-to-front” interference [1]. Much later, other people called it “bleeding” [14] or “show-through” [15].

The back-to-front interference appears whenever the content of the verso side of a document is visible on the front side due to paper translucidity (Figure 1). Such artifact degrades the automatic document transcription via OCR and there is often the superposition of both sides whenever the image is binarized, yielding an unreadable document. In the case of historical

documents paper aging is a complicating factor as it darkens the paper and causes an overlapping of the distributions of the RGB components of the ink on both sides of the paper.



Original

Filtered using [1]

New strategy

**Figure 1.** Zoom in a document from Nabuco bequest with back-to-front interference filtered out using the algorithm described in reference [1] and the new strategy herein.

The technical literature presents several algorithms to remove the back-to-front noise, but no algorithm is good enough in all cases [12]. Depending on the degree of translucidity of the paper, the kind of the ink used in printing or writing, the porosity of the paper, etc. the interference may show itself stronger or weaker. Some algorithms perform better than others in different degrees of interference and even one chosen algorithm may perform better if its parameters are tuned to the intensity of the noise.

This article presents a new strategy to select and tune an algorithm to remove the back-to-front interference in color documents. It makes use of a set of neural classifiers to assess the intensity of the back-to-front interference and to automatically adjust the parameters of the algorithm described in reference [1] to filter the noise out of a given a document. The blanks yielded by removing the artifact are filled in with pixels that correspond to the paper area in the document in such a way to provide the reader with “a natural” look of the document as if it were written on one side only. Figure 1 presents a sample of the results obtained by the algorithm proposed herein, in which one may observe its efficacy.

This paper is organized as follows. Section 2 describes the new filtering strategy. The document

features extracted are presented in Section 3. Section 4 details the noise detection mechanism. The results obtained are presented and analyzed in Section 5. The paper ends presenting its conclusions and draws lines for further work.

## 2. The Filtering System

This section presents a new strategy to remove the back-to-front interference in color documents. First, one needs to remove the framing borders in the document image. Such borders act as noise that interferes with the analysis performed by other algorithms [3]. PhotoDoc [4] was used to pre-process the whole file of documents. Then, there is the use of the classifiers to verify the existence and degree of back-to-front interference. In parallel to that analysis another classifier checks the presence of blur in blocks of the image. Once the former classifier detects the presence and intensity of back-to-front noise, the global threshold algorithm presented in reference [1] is tuned to remove the artifact. At the end the interfering pixels are painted with the colors of pixels that correspond to the sheet of paper, removing the interference in the resulting image.

### 2.1. Classification strategy

The architecture of the classifier is presented in Figure 2.

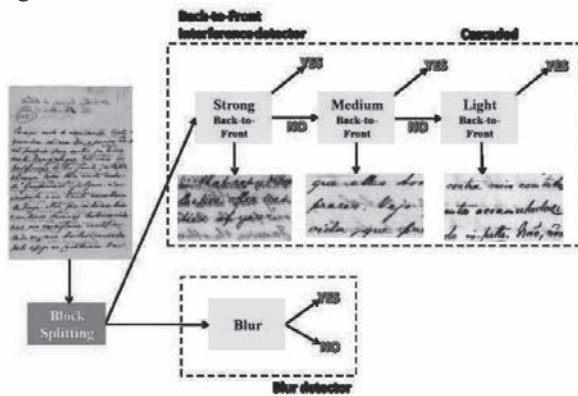


Figure 2. Classifier architecture

The classifier used is Random Forest [6], implemented in Weka [5], an open source tool developed at Waikato University, New Zeland, which offers a wide variety of classifiers implemented. A set of features is extracted from each image to allow its classification. The details about the training and test sets are provided in Section 3.

The classifier developed works in parallel for the detection in image blocks of two different kinds of

noise: the back-to-front interference and blur. In the case of back-to-front interference the classification is performed by three cascaded classifiers that split the bleeding noise into three categories: strong, medium and weak.

### 2.2. Discriminating the interfering pixels

The entropy-based segmentation algorithm by Silva-Lins-Rocha [1] is used twice to find the back-to-front interference area. The first time, to split the text from the rest of the document. The second time to separate the interference from the paper. The algorithm uses the grayscale converted image as an intermediate to split the histogram into three different areas of interest (see Figure 3).

The loss factor  $\alpha$  is a parameter of the algorithm that allows a better statistical tuning between the distributions of the original and binary histograms and it is based on Shannon entropy [8]. For the second image filtering using the algorithm by Silva-Lins-Rocha, a new  $\alpha$  is defined taking into account the intensity of the back-to-front interference and the presence of blur in the image, Table 1 indicates the suggested values  $\alpha$ , such as to allow a better separation between the interference and the paper distribution. The values for  $\alpha$  were experimentally found.

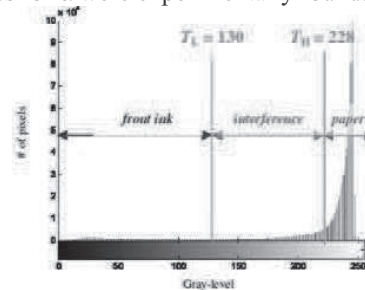


Figure 3. Image histogram of document with back-to-front interference - segmentation details.

Interference	Blur	No Blur
Weak	0.90	1.00
Medium	0.78	0.90
Strong	0.60	0.70

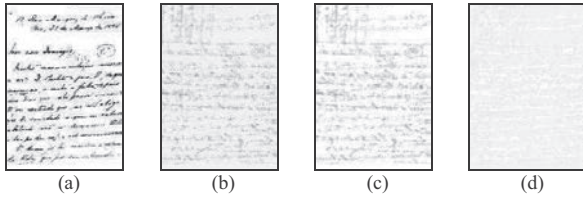
Table 1. Values for alpha.

In summary, to detect the interference area:

1. Apply the segmentation algorithm by Silva-Lins-Rocha to sieve the foreground ink from the rest of the document (see Figures 4a /4b);
2. For each image block classified as having back-to-front interference a new loss factor  $\alpha$  (see Table 1) is chosen. Filter it using the algorithm by Silva-Lins-Rocha to separate the interference ink from the paper (see Figures 4c

and 4d), yielding a blank sheet of paper with white holes where there was ink and ink interference in the original document image.

To illustrate the process, in Figure 4 the first threshold,  $T_L$ , is obtained by the first application of the Silva-Lins-Rocha algorithm and the blocks threshold,  $T_H$ , by the second. The pixels for which their gray-levels are less than  $T_L$  are classified as ink of the front face. The pixels with gray-level greater than  $T_H$  are classified as belonging to the paper. Pixels with gray-levels between  $T_L$  and  $T_H$  are discriminated as interference. The difference here is the application on each image block of a fine tuning between the thresholds  $T_L$  and  $T_H$  taking into account local image information. It is also worth stressing that this new filtering procedure has the advantage of reducing the risk of “damaging” image areas in which there is no interference, once the segmentation algorithm is not applied on them.



**Figure 04.** Views of a document with back-to-front interference: (a) ink of the front face and (b) paper with interference. Image segments of Figure 3b: (c) interference and (d) paper.

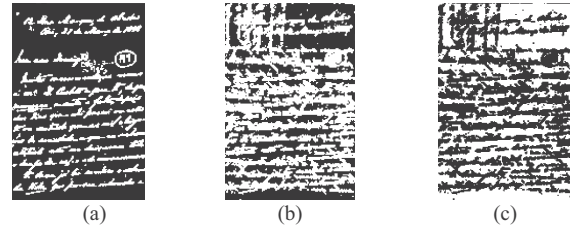
### 2.3. Document Reconstruction

The process proposed here makes use of a “linear” interpolation to fill in the blank pixels that originally corresponded to the interference area. Two binary masks are defined: TEXT and INTERF. The first one identifies the pixels from the ink of the front text (see Figure 5a); the second one highlights the interference area (see Figure 5b). One could assume that only the INTERF mask would be sufficient to the fulfillment process, because the pixels to be replaced “are known already”. Some difficulties appear, however. The key idea is to replace the colors of the noisy pixels with colors as close as possible to the paper in their neighborhood. This is achieved by interpolation, using the colors of the pixels that surround the area to be filled in. There is still the need to remove some of the vestigial shades surrounding the ink pixels in the resulting image; otherwise those pixels will “damage” the interpolation process, bringing in noisy dark colors to the interference area. To solve this problem, one should apply the dilate morphological expansion operation to both masks, with that, the text and

interference contours will be properly classified as “text” and “interference”, respectively (see Figure 6a and Figure 6b). As mentioned earlier on, the pixels that are used in the interpolation process are surrounding the interference area and with the pixels belonging only to the paper. This mask, PAPER, is obtained by the complement of the logical OR operation between the TEXT and INTERF dilated masks (see Figure 6c). Equation 1 calculates a weighed mean, where the intensity of the nearest pixel from the pixel P has the greatest weight. This is reasonable, because in a neighborhood, generally, the closer a pixel is from another, the more alike they should look. Figure 7b shows the result of the application of the proposed filtering strategy applied to the image in Figure 7a.



**Figure 05.** Masks that identify (a) the text and (b) the interference.



**Figure 06.** Dilated masks: (a) text (T) and (b) interference (I) (c) T or I.

Now, the interpolation process is presented. Let the coordinates be as depicted in Figure 7b:

- $(x_0, y_0)$  of a pixel  $P$  from the interval to be interpolated;
- $(x_0, y_1)$  of pixel  $P_N$  – first pixel north  $P$ ;
- $(x_0, y_2)$  of pixel  $P_S$  – first pixel south  $P$ ;
- $(x_1, y_0)$  of pixel  $P_W$  – first pixel west  $P$ ;
- $(x_2, y_0)$  of pixel  $P_E$  – first pixel east  $P$ ,

Where  $i_C(x, y)$  is the value of the component  $C$  (R, G or B) of the pixel  $(x, y)$ . The intensity of the interpolated pixel ( $P$ ) is given by

$$i_C(x_0, y_0) = \frac{d_4 \times i_1 + d_3 \times i_2 + d_2 \times i_3 + d_1 \times i_4}{d_4 + d_3 + d_2 + d_1}, \quad (1)$$

where the  $i_k$  and  $d_k$  ( $k = 1, \dots, 4$ ) represent the intensities and the distances from the pixels –  $P_N$ ,  $P_S$ ,  $P_W$  and  $P_E$  – to  $P$ , sorted by increasing distances. For example, the closest pixel to  $P$  has distance  $d_1$  and intensity  $i_1$ , the second closest one has distance  $d_2$  and



intensity  $i_2$ , and so on. The distance between any two pixels A e B with coordinates  $(x_a, y_a)$  and  $(x_b, y_b)$ , is the standard Euclidian distance:

$$d_{A,B} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}. \quad (2)$$

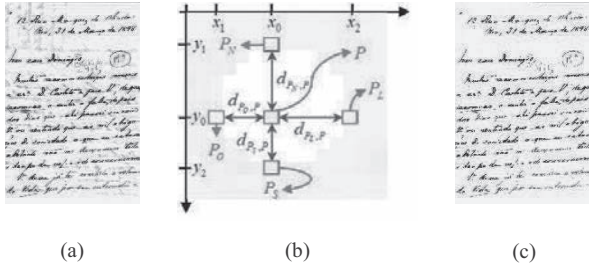


Figure 7. Images: (a) original, (b) Interpolation process and (c) filtered by the new strategy proposed here.

### 3. Classification Features

The choice of the features to be extracted from each image is of paramount importance to the success of the classifier. The following set of features, based a combination on the classifiers described in references [9]. Image binarization is performed by using Otsu [10] algorithm. The height and width stand for the number of pixels in the image. RGB size stands for the true color size of the image (if it is a color image). 8-bits size is either the size of the original image if in grey scale or the size of the grey-scale converted from true-color. #Black\_pixels stands for the number of black pixels in the monochromatic converted image. The combination of the features presented in [9] and the two new features (Local Power Spectrum Slope and Maximum Saturation) in [11] yielded a relative gain to the performance of the classifier. Each of these features was taken on nine blocks of the image (see Figure 5).



Figure 8. Areas of interest for extract features

### 4. Back-To-Front Noise Detection

Back-to-front noise, depending on its strength, may make document binarization unviable. As most OCRs take a binary image as input, thus documents with such noise may not be automatically transcribed. Researchers [12][13] have pointed out that no algorithm in the literature is good enough to remove bleeding noise in all sorts of documents. Depending on the strength of the noise, some algorithms may perform

better than others. Unfortunately, the back-to-front noise appears more often in the digitalization of documents than one may assume to start with. The test set of documents we used with show-though had 260 real-world documents (no synthetic ones) which were obtained either from historical files (as shown in Figure 1). The images were divided into nine blocks totaling 2,340 blocks and were hand labeled according to four levels of interference as: strong (773), medium (856), light (524) and none (187). The classifiers for this noise were cascaded, as shown in Figure 2. The strong-classifier was trained with the blocks tagged as strong in the training set, against all the remaining images (Medium-Light-None) from the training set (strong 150; medium 150; light 100; and 20 none). Similarly, the medium-classifier was trained with the blocks labeled as medium, against the others with a lighter or no interference. The classification results obtained are shown in Table 2.

Back-to-front	Strong	Medium	Light	None	Accuracy %
Strong	703	58	11	1	90.94
Medium	27	816	4	9	95.32
Light	5	9	96	12	92.93
None	1	2	11	187	92.51

Table 2. Confusion matrix of the back-to-front noise classifier with sub-sampled block images

### 5. Blur Detection

The presence of blur may be an indicator of low quality digitalization, but can also be associated with other problems such as the scanning of hard-bound volumes. The blur noise is seldom global. In general, it affects some areas of a document. In the case of the documents studied here, the blur noise is originated from the spreading out of the ink in the verso face of the document. While the issues related to the analysis of image-blur have attracted much attention from researchers in recent years, the work reported in the literature focus mainly on solving the problem of deblurring. Blur detection is a complex task. The work reported in [7], points at blur as one of the greatest difficulties for the filtering out of the back-to-front noise in historical documents. Bluer detection is solved here by using the classifier presented in subsection 2.1 and the characteristics described in section 3. The image blocks were classified manually, 124 blocks with and 2216 without blur. The training set used 20 blurred blocks and 150 unblurred ones. Besides those, 500 blocks with synthetic blur were used to validate the

classifier. The result is shown by the myo classifier confusion matrix (see Table 3).

Orientation	With	Without	Accuracy %
With	615	9	98.55
Without	3	2,213	99.86

**Table 3.** Confusion matrix of the blur noise classifier with sub-sampled images in proposed architecture

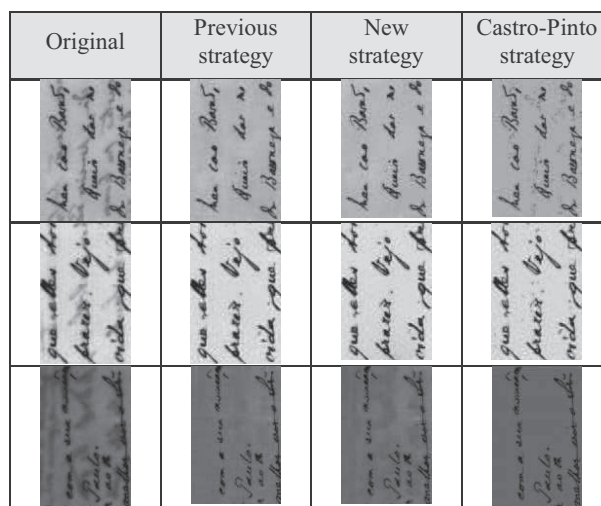
## 6. Results and Analysis

The proposed algorithm was tested in a set of 260 images from the Joaquim Nabuco bequest of digitalized documents [2], yielding good results. Evidences of the efficiency of the new filtering technique are shown in Figure 9, as the back-to-front interference was removed yielding a more readable document with a “natural” look. Figure 9 provides the results of using different strategies, amongst them using as fulfillment for the blanks the result of the interpolation based on Laplace’s equation (the MATLAB function “*roifill*” was used). The third alternative is one of the strategies proposed by Castro and Pinto [16] that uses the algorithm by Salvola and Pietikainen [17] which defines a mask that identifies the pixels of the foreground and background objects. The final image is obtained through keeping the object pixels and replacing the background pixels with the average of the colors of the pixels in that class. The latter strategy yielded the best results. The two strategies proposed herein yielded very similar quality results. However, the one based on Laplace interpolation leaves the filled-in area look undesirably uniform with a “flat” color. On the other hand, the linear interpolation yields a residual pattern of vertical/horizontal stripes. The strategy proposed by Castro and Pinto [16] aims to yield a uniform paper surface with unchanged text, while the ones presented here try to remove only the interference, keeping the pixels from the paper and text unchanged. However, in the very few images in the Nabuco file that the back-to-front interference looks very “blurred” (see last segments of Nabuco file in Figure 9), the proposed algorithm did not perform well. The detection of the whole back-to-front interference area is far from being a trivial task.

## References

[1] J. M. M. da Silva; *et al.* A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference". *J. Universal Computer Science*, v. (14):299-313, 2008.  
 [2] FUNDAJ: [www.fundaj.gov.br](http://www.fundaj.gov.br).

[3] R. D. Lins. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. *ICIAR 2009*. LNCS v. 5627. p. 844-854, Springer Verlag, 2009.  
 [4] G. F. P. e Silva and R. D. Lins. PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. *Proceedings of CBDAR 2007*. IAPR Press, 2007. p. 107-115.  
 [5] Weka 3: Data Mining Software in Java, website <http://www.cs.waikato.ac.nz/ml/weka/>.  
 [6] L. Breiman, "Random Forests", *Machine Learning*, 45(1), pp. 5-32, 2001.  
 [7] J. M. M. da Silva; *et al.* Enhancing the Quality of Color Documents with Back-to-Front Interference. *ICIAR 2009*. LNCS, v. 5627, p. 875-885, Springer Verlag, 2009.  
 [8] N. Abramson, "*Information Theory and Coding*", McGraw-Hill Book Co, 1963.  
 [9] R.D Lins; G.F.P. Silva; S. Banergee; A. Kuchibhotla and M. Thielo. "Automatically Detecting and Classifying Noises in Document Images", *ACM-SAC '2010*, ACM Press, March 2010.  
 [10] N. Otsu. "A threshold selection method from gray level histograms". *IEEE Trans. Syst. Man Cybernetics SMC-9*, 62-66, 1979.  
 [11] L. Renting, L. Zhaorong, and J. Jiaya, Image Partial Blur Detection and Classification, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.  
 [12] R.D Lins; J. M. M. da Silva; F. M. J. Martins. Detailing a Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents. *Journal of Universal Computer Science*, v. 14, pp. 299-313, 2008.  
 [13] P. Stathis; E. Kavallieratou; N. Papamarkos. An Evaluation Technique for Binarization Algorithms. *Journal of Universal Computer Science*, v. 14, pp. 3011-3030, 2008.  
 [14] R. Kasturi, L. O’Gorman and V. Govindaraju, "Document image analysis: A primer", *Sadhana*, (27):3-22, 2002.  
 [15] G. Sharma, "Show-through cancellation in scans of duplex printed documents", *IEEE Trans. Image Processing*, v10(5):736-754, 2001.  
 [16] P. Castro, J. R. C. Pinto: "Methods for Written Ancient Music Restoration". *ICIAR 2007*: 1194-1205.  
 [17] J. Sauvola, M. Pietikainen: "Adaptive document image binarization", *Patt. Recognition* 33(2):225-236, 2000.



**Figure 9** - Parts of documents from the Nabuco file: original and filtered.

## **A.6 Publicações sobre remoção de ruído Especular**

(MARIANO et al., 2011) E. Mariano, R. D. Lins, G. F. P. Silva and J. Fan. Correcting Specular Noise in Multiple Images of Photographed Documents. In: International Conference on Document Analysis and Recognition, pp: 915-919.

(LINS et al., 2013) R. D. Lins; G. F. P. Silva; E. Mariano, F. Fan, P. Majewicz and M. Thielo. Removing Shade and Specular Noise in Images of Objects and Documents Acquired with a 3D-Scanner. In: Lecture Notes in Computer Science, vol.1, pp: 299-307.

## Correcting Specular Noise in Multiple Images of Photographed Documents

Ednardo Mariano

Rafael Dueire Lins

Gabriel de França Pereira e Silva

Universidade Federal de Pernambuco

Recife, Brazil

{rdl, gabriel.psilva}@ufpe.br

Jian Fan

Peter Majewicz

HP Labs

Palo Alto, USA

{jian.fan, peter.majewicz}@hp.com

Marcelo Thielo

HP Labs.,

Porto Alegre, Brazil

marcelo.resende.thielo@hp.com

**Abstract** — Portable digital cameras have become omnipresent. Their low-cost, simplicity to use, flexibility, and good quality images have widened their applicability far beyond their original purpose of taking personal photos. Every day people discover new uses for them from photographing teaching boards to documents. One of the difficulties of using cameras is the occurrence of specular noise whenever the photographed object is glossy. This paper presents an efficient algorithm for removing the specular noise of photographed documents by taking multiple images with different illumination sources.

**Keywords**— *specular noise, photographed documents, multiple images.*

### I. INTRODUCTION

Portable digital cameras may be considered as a pervasive good. The quality and resolution of the cameras embedded in cell phones today are as good as the ones of dedicated devices from a not distant past. Such omnipresence has widened its applicability into unforeseen domains. One of them is using portable digital cameras for digitizing documents. People now use those devices as a fast way to acquire document images, avoiding photocopying, take photos of teaching boards and bill boards instead of taking notes. Such application gave rise to new research area [3], which is evolving fast in many different directions and claims for new algorithms, tools and processing environments that are able to provide users in general with simple ways of visualizing, printing, transcribing, compressing, storing and transmitting through networks such images. Reference [6] points out some particular problems that arise in this document digitization process using portable digital cameras: the first of all is background removal. Very often the document photograph goes beyond the document size and incorporates parts of the area that served as mechanical support for taking the photo of the document. The second problem is due to the skew often found in the image in relation to the photograph axes, as documents have no fixed mechanical support very often there is some degree of inclination in the document image. The third problem is non-frontal perspective, due to the same reasons that give rise to skew. A fourth problem is caused by the distortion of the lens of the camera. This means that the perspective distortion is not a straight line but a convex line, depending on the quality of the lens and

the relative position of the camera and the document. The fifth difficulty in processing document images acquired with portable cameras is due to non-uniform illumination. A even more complex situation is faced when the paper of the document is glossy: often the photo has areas in which the in-built strobe flash or intense lighting from the environment “erases” parts of the document presenting what is called the “specular noise” [1]. Figure 1 presents an example of a photo with specular noise. The photo from two bound pages of a magazine printed on glossy paper exhibits two areas of specular noise. On the right hand page one may observe an “erased” area to the left of the leg of the chair. On the opposite page, there is another damaged area on the pillows (which erases part of the stripes). Red arrows point at the specular noises in Figure 1. The photo was obtained with the built in strobe flash on.

This paper proposes an algorithm to remove the specular noise by taking three pictures of documents or 3D-objects with a fixed camera and varying the position of the light sources. results obtained yield to conclude that the presented scheme is valid and provided much better quality results than any other previous report in the technical literature [7][10].



**Figure 1.** Photo of a two page magazine printed on glossy paper with specular noises.

## II. MOTIVATION

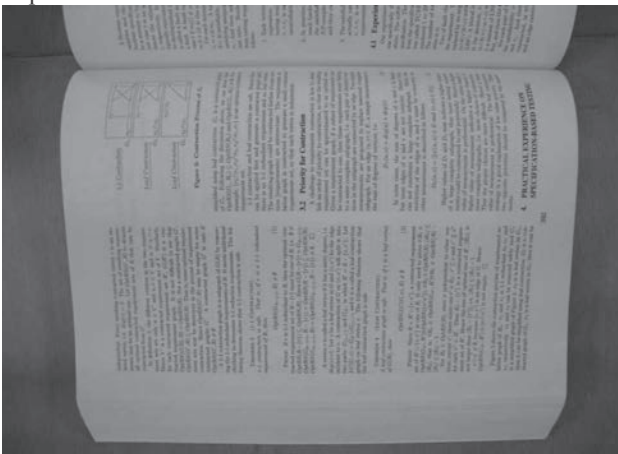
Besides the aforementioned omnipresence of digital cameras, there are situations in which their use is more adequate and easier than flatbed scanners. Since scanners have been integrated with printers in “all-in-one” devices (scanner, printer and copier) the price for an A4 scanner became marginal. But larger-size scanners have kept their high prices unchanged and are not easily available to buy. Scanning documents larger than A4 size, such as maps, is a difficult task. The same happens with bound books (either hard or soft) for which one either damages the volume or has to address the difficult problem of geometrical warping and even blur in some areas. In the case of oversized old books or documents, using standard A4 flatbed scanners is unviable.

A simple digitalization platform for such documents based on portable digital cameras is shown in Figure 2.



**Figure 2** – Digitization platform with two light sources and portable digital camera (*UFPE-Planetarium*).

The platform presented in Figure 2 was built at UFPE (Brazil) and it has shown to be a suitable test bed for document and book digitization for a low-cost and flexible platform, capable of digitizing oversized documents such as maps.



**Figure 3** – Example of document digitized at the *UFPE-Planetarium* Both lamps on and strobe flash off.

The features of the *UFPE-Planetarium* are:

- Camera: Digital Sony Cyber-Shot W220, 12.1 Mpixels, in-built strobe flash.
- Camera height (lens) to support plan: 12.2 cm;
- Lamp: Compact fluorescent – 20 W;
- Right-hand side lamp height to support plan: 13.7 cm;
- Left-hand side lamp height to support plan: 13.5 cm;
- Distance between camera lens and right-lamp: 10.8 cm;
- Distance between camera lens and left-lamp: 11.2 cm.

A second platform was also built with the following features:

- Camera: HP 8 Mpixels with fixed resolution and zoom. Lens distortion is negligible;
- Camera height (lens) to support plan: 21 cm;
- Lamp: high-power phosphor-converted white LEDs;
- The center lamp is positioned as close to the lens as possible.
- Lamp offset left: 7.5 cm
- Lamp offset right: 7.5 cm

A total of 4 frames of the scene with different illumination are available. They are:

- ambient lighting only;
- left lamp on, right lamp off, center lamp off;
- left lamp off, right lamp on, center lamp off;
- left lamp off, right lamp off, center lamp on.

The correction of the geometrical warp caused by book binding was studied in a series of papers [4][9] with satisfactory results. The other problem faced is being addressed here: the removal of the specular noise.

## III. METHODOLOGY

Removing the specular noise from images such as the one presented in Figure 1 is a difficult task. One possibility is to take several photos under different illumination set-ups to try to “re-assemble” a new image replacing the parts “erased” by the specular noise from one image with the “information” from another shot. This paper follows exactly such an alternative taking three photos of a given document under three different illumination patterns and generating a new image with “parts” of the three “tributaries”. Figure 4 presents an example of three images of the same document. Although the solution proposed here seems to be a simple one, in reality to automatically extract the information from each image and yield a “natural looking” result is a complex task. After a large number of attempts that ranged from functor analysis of color variation to splitting the image into small blocks and using a neural classifier on each image to spot where the specular noise could be found in each of them [5], the scheme that is presented below was the one that provided the most consistent best results in terms of resulting image quality and also in processing time performance.

A. The algorithm.

The photos taken are true-color RGB 24-bits. Each document photo is split in their RGB components, which each of them is an 8-bit grayscale image (*Left-R*, *Left-G*, *Left-B*; *Center-R*, *Center-G*, *Center-B*; *Right-R*, *Right-G*, *Right-B*).



Figure 4 – Document photographed with the second platform. **Top:** Right lamp only. **Center:** Central lamp only. **Bottom:** Left lamp only.

The main idea of the algorithm proposed is to compare the intensity of the pixels of the image between components.

The pixel with the lowest intensity of the same component of the three images is copied into the component of the final image. The image with the specular noise filtered out is formed by the RGB components, following the scheme depicted in Figure 5.

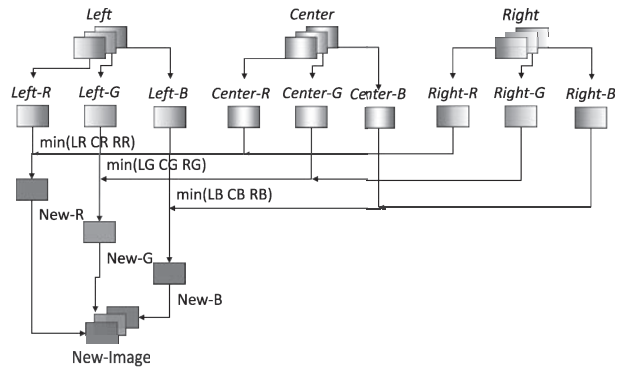


Figure 5 – Scheme adopted for removing the specular noise by analyzing three images simultaneously.

The images illuminated by the left and right hand side lamps (strobe flashes) tend to exhibit a circular halo at opposite sides to the lamp position causing a large variation to the pixel intensities and covering a large part of the image area. For better results, and such halo may have its contrast increased [1] by lowering the variation of the intensity of the pixels, through the equalization of the global histogram of each RGB component both of the left and right illuminated images.

The resulting image presents a significant reduction of the specular noise as may be observed in Figures 6 and 7.



Figure 6 – Image of Figure 4 with specular noise removed

Due to the choice of pixels with the lowest intensity in each component, the resulting image has its brightness reduced in comparison with the original ones. But all the information, colors and shapes are recovered, making possible the application of techniques of histogram correction to enhance the brightness and contrast of the resulting image [1].



Figure 7 – Image of Figure 1 with specular noise removed

In glossy paper, commonly used in magazines and color printed books the reflection of the light of the strobe flash may be so strong such as to cause saturation of the sensor of the camera, causing a “chroma noise” or “confetti” [1][2] at areas surrounding the specular noise. Figure 8 top zooms into part of Figure 1 and exhibits some of the chroma noise in the image. The algorithm proposed is not able to remove such noise as it may also be observed in the bottom part of the same figure.



Figure 8 – Zoom into parts with chroma noises in Figures 1 and 7.

### B. Compensating Chroma Noises in Images

The direct application of the algorithm presented in the last section does not filter out the confetti noise from the resulting image. A more careful study of the different images taken shows that such noise appears close to the specular halo and at the outskirts of images. Thus, it is possible to identify the regions where the chroma noise may appear and eliminate them before comparing the RGB components in the algorithm presented. This was done by creating three sub-images of the original image. The width of each image depends on the incidence angle of the lamp which determines the position of the halo and the chroma noise. For the angles of the current configuration of the second platform the images were split into 25%, 50%, and 25% of the original width. Regions will be labeled as “A”, “B”, and “C”. Figure 9 illustrates such an image splitting in relation to the halo and specular noise. If one considers that each object is photographed three times under different illumination patterns and each has 3 RGB-components, then one has 27 sub-images in total.

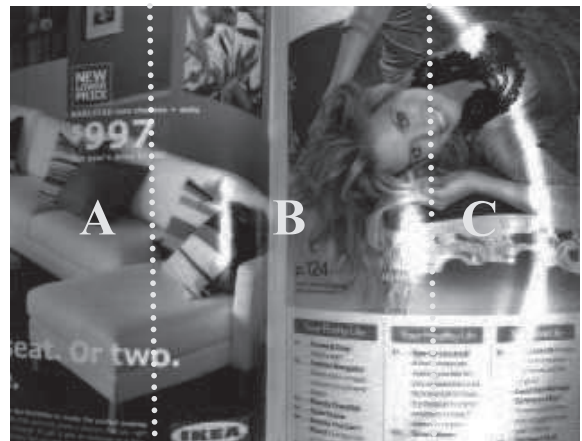


Figure 9 – Left illuminated image split into three regions

Instead of comparing the intensity of pixels directly of the three images following the scheme presented in Figure 5, one groups images in three sets:

- A-Left and A-Center;
- B-Left, B-Center, and B-Right;
- C-Right and C-Central;

Where “A-Left” means the region “A” of the Left-illuminated region, “A-Center” means the region “A” of the image with the centre lamp on and the other lamps off, etc. The sets above avoid that the chroma noise is taken into the final image. The images in each set have their intensity of the RGB-components compared in a similar fashion to the original algorithm, yielding one resulting image per region. The final image is obtained by merging the three images in their position. Figure 10 shows the resulting image of the images of Figure 4 with the specular and chroma noises reduced.



Figure 10 – Result of Figure 4 with specular and chroma noises removed

Figure 11 zooms into the image of Figure 1 with the specular and chroma noises reduced.



Figure 11–Zoom into part of Figure 1 with specular and chroma noises removed

One may still observe that the resulting image has lower brightness level than the original images.

The implementation of the algorithm was done using Matlab® of The MathWorks, running on Microsoft Windows Seven Professional®.

#### IV. CONCLUSIONS

Portable digital cameras are a low-cost option for digitizing large documents such as maps or even bound books. These are new applications that are far away from their original purpose that is to take souvenir photos of people and places. Whenever taking pictures of documents printed on glossy paper the specular noise may arise due to the short distance between the camera and the photographed object [7][10][8].

This paper presents an efficient method to remove the specular noise by taking multiple images under different illumination patterns. It was tested in images obtained in two different platforms and performed well in both of them. In the case of the second platform, besides the specular noise, the intensity of the strobe flash lamps also gave rise to chroma noises. Splitting the image in regions and analyzing them in a way not to take into account the noisy areas yielded good results.

The simple yet efficient method proposed here also provided suitable removal for the specular noise in grayscale images and also in images of 3-D objects.

#### REFERENCES

- [1] R. Gonzalez, R. Woods. Digital Image Processing, 3<sup>o</sup> Edition, Prentice Hall, 2008, pp. 68-102,pp. 259-300.
- [2] G. V. Landon, Y. Lin, and W. B. Seales. Towards Automatic Photometric Correction of Casually Illuminated Documents., IEEE Conference on Computer Vision and Pattern Recognition, 2007.
- [3] J. Liang, D. Doermann and H. Li. Camera-Based Analysis of Text and Documents: A Survey. International Journal on Document Analysis and Recognition, 2005.
- [4] J. Liang, D. DeMenthon, and D. Doermann. Geometric Rectification of Camera-Captured Document Images. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol 30 (4):591-605.
- [5] R.D. Lins, G. F. P. e Silva, S. Banergee, S Kuchibhotla, and M. Thielo. Automatically Detecting and Classifying Noises in Document Images. ACM-SAC 2010. ACM Press, 2010. v.1. p.33 – 39.
- [6] R. D. Lins, A. R. G e Silva, G. F. P. e Silva. Enhancing Document Images Acquired Using Portable Digital Cameras. ICIAR 2007. Springer Verlag, 2007. v.LNCS. p.1229 - 1241.
- [7] C.-A. Saint-Pierre, J. Boisvert, G. Grimard, F. Chriet, Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images, Machine Vision and Applications. V22(1): 171-180. Springer-Verlag, 2007.
- [8] D. M. Oliveira, R. D. Lins. Improving the Border Detection and Image Enhancement Algorithms in Tableau. ICIAR'2008. Springer Verlag, 2008. LNCS v.5112. p.1111 - 1121
- [9] D. M. Oliveira, R. D. Lins, G. Torreão, J. Fan, M. Thielo. A New Method for Text-line Segmentation for Warped Documents ICIAR 2010. Springer Verlag, 2010. LNCS 6112. p.398 - 408.
- [10] L. B. Wolff, L. B., "On the relative brightness of specular and diffuse reflection. In: Proceedings of CVPR", 1994, pp. 369–376.



# Removing Shade and Specular Noise in Images of Objects and Documents Acquired with a 3D-Scanner

Rafael Dueire Lins<sup>1</sup>, Gabriel França Pereira e Silva<sup>1</sup>, Ednardo Mariano<sup>1</sup>, Jian Fan<sup>2</sup>, Peter Majewicz<sup>2</sup>, and Marcelo Thielo<sup>3</sup>

<sup>1</sup> Universidade Federal de Pernambuco, Recife, Brazil

<sup>2</sup> HP Labs., Palo Alto, USA

<sup>3</sup> HP Labs., Porto Alegre, Brazil

**Abstract.** This paper presents an efficient algorithm for removing the specular noise and undesired shades in images of objects and documents acquired with a 3D-Scanner. The basic principle of such device is to photograph objects by taking multiple images with different illumination sources.

**Keywords:** Specular noise, 3D-Scanner, illumination sources.

## 1 Introduction

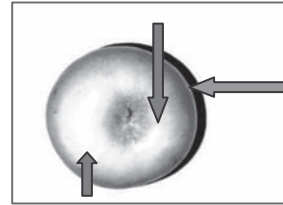
The fast pace of technological evolution today opens new horizons and brings new challenges every day. Digital cameras appeared not long ago and caused the death of "traditional" over a century-old photography. Portable digital cameras are omnipresent today and even cell-phones embedded ones yield images of quality and resolution unforeseen in not distant past, widening its applicability into new domains such as digitizing documents with an ease not allowed by scanners. People now photograph documents, avoiding photocopying, and take photos of teaching boards and bill boards, instead of taking notes. A new and fast evolving research area [4] was born and claims for new algorithms, tools and processing environments that are able to provide users in general with simple ways of visualizing, printing, transcribing, compressing, storing and transmitting through networks such images. Some particular problems arise in document digitization using portable digital cameras [7]. One of them is due to non-uniform illumination. A even more complex situation is faced when the paper of the document is glossy: often the photo has areas in which the in-built strobe flash or intense lighting from the environment "erases" parts of the document presenting what is called the "specular noise" [2]. The same phenomenon may also occur with 3D-objects if it reflects part of the incident illumination.

Figure 1 presents an example of a photo with specular noise. The photo from two bound pages of a magazine printed on glossy paper exhibits at least three areas of specular noise, pointed at by red arrows. The photo was obtained with a portable digital camera with the built in strobe flash on.

The situation of taking photos of 3D-objects is even more difficult as besides the specular noise the illumination shaded areas often appear. An example of both phenomena may be observed in the image of the apple shown in Figure 2, where the red arrows point at the specular noise areas and the blue arrow to the undesired shade one.



**Fig. 1.** Photo of a two page magazine printed on glossy paper with specular noises



**Fig. 2.** Photo of an apple with areas of specular noises (red arrow) and shade (blue arrow)

## 2 The Multiple Image Platform

To solve such problems of filtering out the specular noise and undesired shades in 3D-object photographs the Hewlett-Packard Company developed a simple and low-cost platform, which was released as the HP TopShot laser printer with “3D scanner”, which, besides targeting at removing such noises, it also allows the fast and efficient digitalization of documents, included bounded books. It is important to remark that there are situations in which the use of cameras is more adequate and easier than flatbed scanners. Such 3D-scanner was integrated with a laser printer as an “all-in-one” device (scanner, printer and copier), thus the price for the 3D-scanner became marginal. Larger than A4 flatbed scanners have kept their high prices unchanged and are not “off-the-shelf” available to buy. Scanning larger documents, such as maps, is a difficult task. The same happens with bound books (either hard or soft) for which one either damages the volume or has to address the difficult problem of geometrical



**Fig. 3.** Photo of the HP TopShot laser printer with “3D-scanner”



**Fig. 4.** Detail of the camera and illumination set-up of the “3D-scanner”

warping and even blur in some areas. In the case of oversized old books or documents, using standard A4 flatbed scanners is unviable. The correction of the geometrical warp caused by book binding was studied in a series of papers [9] [10] with satisfactory results.

Figure 3 presents a photo of the HP TopShot laser printer with "3D scanner" and Figure 4 zooms into the camera and illumination part. The experiments reported here made use of a slightly different platform from the one already released as a product by HP. The platform takes three pictures of documents or 3D-objects with a fixed camera and varying the position of the light sources. The platform used here has the following technical features:

- Camera: 8 Mpixels with fixed resolution and zoom. Lens distortion is negligible; True color RGB 24-bits.
- Camera height (lens) to support plan: 21 cm;
- Lamp: high-power phosphor-converted white LEDs;
- The center lamp is positioned as close to the lens as possible.
- Lamp offset (left and right): 7.5 cm

In the platform under study a total of three frames of the scene with different setups are obtained. Figure 5 provides an example of an image of a 3D-object acquired with the platform, in which one may clearly observe shaded areas and specular noises. The transparency of the object (bottle) highly increases the complexity of the image filtering.



Left lamp on, right lamp off, center lamp off;



Left lamp off, right lamp off, center lamp on.



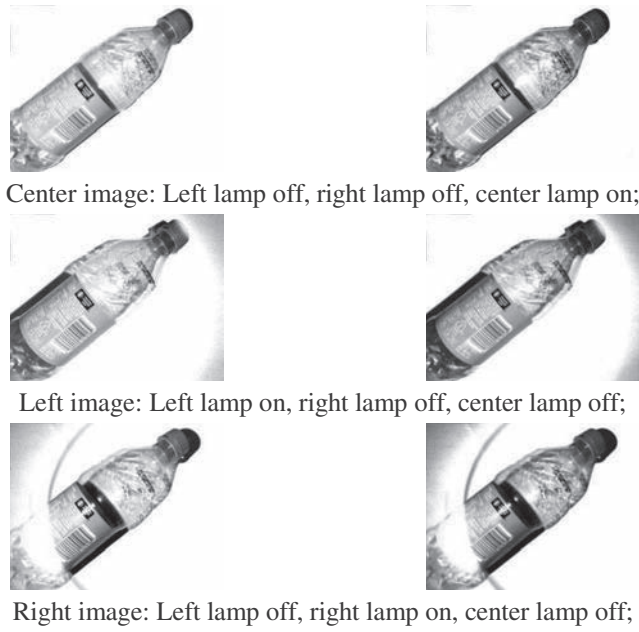
Left lamp off, right lamp on, center lamp off;

**Fig. 5.** Example of images of a 3D-object obtained with the 3D-Scanner

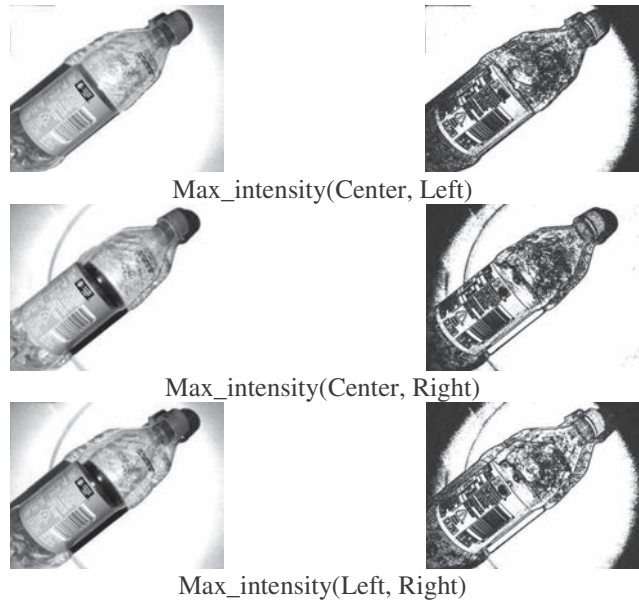
Two problems are addressed in this paper: the removal shades and filtering-out the specular noise. The first one is attacked by using the multiple images for developing a mask that contains only the area of the object. The second processing step is performed to find the areas "erased" by the specular noise and filtering it out. The two strategies are presented in the next sessions.

### 3 Removing Shades

The removal of shades of 3D-objects is performed by using the three images obtained with the HP-Platform. The key idea here is to obtain an image "mask" which is restricted to the object area. The three images are enhanced, binarized, composed in a special way to generate a final image which yields the final mask.



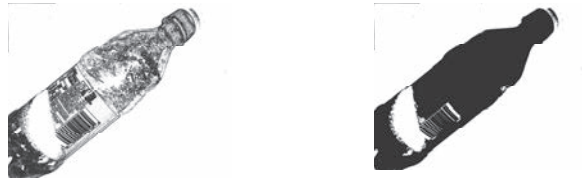
**Fig. 6.** Images of a 3D-object shown in Figure 5 and corresponding edge enhanced images with Canny filter



**Fig. 7.** Composition of the images in Figure 6 using "highest intensity" pixels and corresponding binarized image using Otsu algorithm

The first step is to enhance the image borders by using a Canny edge detector [1] and then performing histogram equalization, both implemented in ImageJ [13]. The application of the filter to the images shown in Figure 5 may be seen in Figure 6. The next step is to stress the contour of the object by adding the images obtained by edge enhancement. Several different possibilities were tested and the best results obtained are shown in Figure 7, where the resulting image is obtained by getting the pixels of the highest intensity from the two input images. Then, the image is made monochromatic by using Otsu [11] binarization algorithm. Figure 8 presents the composed images and the result of binarization.

The three masks obtained in Figure 7 are now combined into a single mask through “majority voting” to find an image that serves as the “contour” of the final mask that is obtained through the “fill holes” algorithm implemented in the ImageJ [13] open-source tool. The majority voting image and the final mask obtained through fill holes may be found in Figure 8.



**Fig. 8. Left:** Result of applying the “majority vote” algorithm to the three images in Figure 5 – Right column. **Right:** Final mask obtained by the application of the “fill holes” algorithm to the image to its left.

The mask obtained is used to select the “Region of Interest” in the three original images.

#### 4 Filtering-Out the Specular Noise

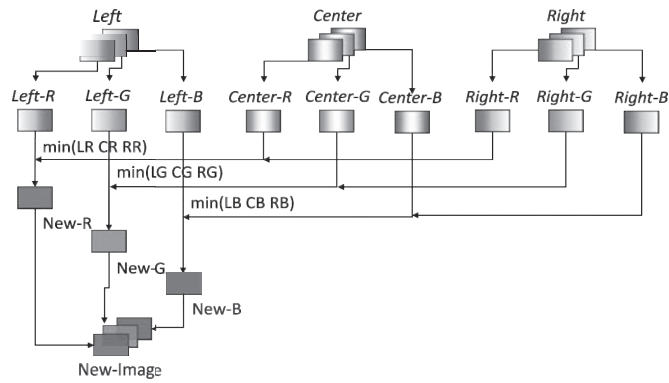
In the case of 3D-objects such as the ones shown in Figure 2 and 5, after the shade areas were removed by applying the mask developed in the previous section, now the problem of removing the specular noise is addressed. The case of the image shown in Figure 1 has no shade area, thus the “region of interest” encompasses the whole image. In any case, removing the specular noise from images such as the one presented in Figure 1, Figure 2 and Figure 5 is a difficult task. The basic idea is to try to “re-assemble” a new image replacing the parts “erased” by the specular noise from one image with the “information” from another shot. The solution presented here was found after an exhaustive number of attempts of many different strategies and techniques because to automatically extract the information from each component image and yield a result that reaches a reasonable quality standard is a complex task. The unsuccessful attempts made ranged from local ones such as splitting the image into small blocks and using a neural classifier [6] to spot where the specular noise could be found in each of them, to global ones such as the functor analysis of color variation on each image. The algorithm that is presented below was the one that yielded the best and most consistent results in image quality and, besides that, is also fast in

processing time, which is an aspect of paramount importance as it must be executed in an embedded device with an acceptable response-time and throughput.

*A. The algorithm.*

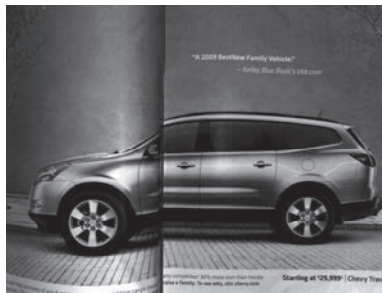
The main idea of the algorithm proposed is to compare the intensity of the pixels of the image between components. Each image is split in its RGB components, which corresponds to an 8-bit grayscale image (*Left-R*, *Left-G*, *Left-B*; *Center-R*, *Center-G*, *Center-B*; *Right-R*, *Right-G*, *Right-B*).

The final image with the specular noise filtered out is assembled by choosing the lowest intensity of the same RGB-component of the three images for each pixel, as depicted in Figure 9.

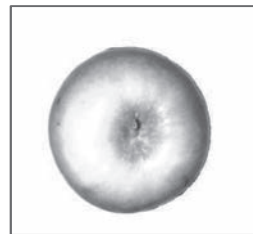


**Fig. 9.** Removing the specular noise by simultaneous analysis of the RGB components of the three images obtained with the HP-Platform

A circular halo may be found in the images illuminated by the right and left hand strobe flashes (LED lamps) as they appear at the opposite sides to the lamp position, causing the pixel intensities to vary widely and covering a large part of the image area. The equalization of the global histogram of each RGB component both of the left and right illuminated images causes an increase [1] in the contrast of such halo. The resulting image presents a significant reduction of the specular noise as may be observed in Figures 10 and 11.



**Fig. 10.** Figure 1 after processing



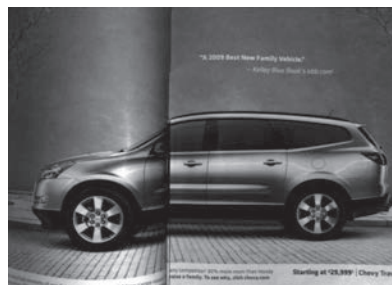
**Fig. 11.** Figure 2 after processing

As the algorithm presented chooses the pixels in each component with the lowest intensity, the resulting image has its brightness reduced in relation to the original ones. All the information, colors and shapes are recovered, making possible the application of techniques of histogram correction to enhance the brightness and contrast of the resulting image [1]. If one closely observes Figure 10, one may see that the filtered image is slightly darker than the original one and that there appeared a brownish area to the left of the soft bound area, not present in the original picture. This problem is addressed further on.

The image of the apple presented in Figure 11 minimizes the shade area leaving only some residual pixels in the contour. The specular noise was attenuated but is still present in the final image, as the three images obtained with the HP-Platform exhibit specular noises in the same areas.

*B. Enhancing the Results Obtained*

The removal of the specular noise may be made faster and yield better quality images both in 3D-objects and in document images acquired with the HP Platform if instead of applying the algorithm presented above to the whole image, one applied it only to the regions where the specular noise is found in the image. This can be done by



**Fig. 12.** Image of Figure 1 with the specular noise removed by the application of the algorithm proposed in the areas affected by the specular noise only



**Fig. 13.** Image of Figure 1 with the specular noise removed by the application of the algorithm proposed in the areas affected by the specular noise only

looking for areas within a 3D-object or an image where the intensity of pixels are saturate, the area is surrounded by lower intensity pixels and the saturated pixels are

not so in the other two images. This allows one to find the regions “erased” by the specular noise to which the algorithm presented is applied to replace such pixels with the information received from the other images.

The same strategy applied to the image in Figure 5, after shade removal using the mask presented in Figure 8 (right) is shown in Figure 13, where one may observe that the results obtained are very satisfactory, despite the very high degree of complexity of the original image due to the transparency of the object (plastic bottle).

## 5 Conclusions

The strategy of using multiple illuminated images photographed with a single camera is a simple way to develop a low-cost 3D-scanner for objects and documents. There are some problems that arise from such set-up, however. Shaded areas and specular noises may arise in the digitalization of documents printed on glossy paper or of 3D-objects, due to uneven illumination sources and the short distance between the camera and the photographed object [7][10][8].

This paper presents an efficient method to remove the shaded areas and the specular noise in such kind of platform. It was tested in images obtained in the HP-Platform and performed well. In some document images, besides the specular noise, the intensity of the strobe flash lamps also gave rise to chroma noises, also known as confetti [2]. The removal of chroma noise may be performed by using the scheme presented in reference [12].

The implementation of the algorithm presented here was done in Java, running on Microsoft Windows Seven Professional® in a platform Intel quad core 2.5 with 4 Gb RAM. The average processing time per image set (3 images with the illumination patterns described) was originally of 10 seconds. The enhancement strategy presented above of applying the specular noise removal algorithm only to the areas “erased” by the noise yielded a reduction of processing time to 1.8 seconds per image set.

## References

- [1] Canny, J.: A Computational Approach To Edge Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence* 8(6), 679–698 (1986)
- [2] Gonzalez, R., Woods, R.: *Digital Image Processing*, 3rd edn., pp. 68–102, 259–300. Prentice Hall (2008)
- [3] Landon, G.V., Lin, Y., Seales, W.B.: Towards Automatic Photometric Correction of Casually Illuminated Documents. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2007)
- [4] Liang, J., Doermann, D., Li, H.: Camera-Based Analysis of Text and Documents: A Survey. *International Journal on Document Analysis and Recognition* (2005)
- [5] Liang, J., DeMenthon, D., Doermann, D.: Geometric Rectification of Camera-Captured Document Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(4), 591–605



- [6] Lins, R.D., Pereira e Silva, G.F., Banergee, S., Kuchibhotla, S., Thielo, M.: Automatically Detecting and Classifying Noises in Document Images. In: ACM-SAC 2010, vol. 1, pp. 33–39. ACM Press (2010)
- [7] Lins, R.D., Gomes e Silva, A.R., Pereira e Silva, G.: Enhancing Document Images Acquired Using Portable Digital Cameras. In: Kamel, M.S., Campilho, A. (eds.) ICIAR 2007. LNCS, vol. 4633, pp. 1229–1241. Springer, Heidelberg (2007)
- [8] Saint-Pierre, C.-A., Boisvert, J., Grimard, G., Cheriet, F.: Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images. In: Machine Vision and Applications, vol. V22(1), pp. 171–180. Springer (2007)
- [9] de Oliveira, D.M., Lins, R.D.: Improving the Border Detection and Image Enhancement Algorithms in Tableau. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2008. LNCS, vol. 5112, pp. 1111–1121. Springer, Heidelberg (2008)
- [10] Oliveira, D.M., Lins, R.D., Torreão, G., Fan, J., Thielo, M.: A New Method for Text-line Segmentation for Warped Documents. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6112, pp. 398–408. Springer, Heidelberg (2010)
- [11] Otsu, N.: A threshold selection method from gray level histograms. *IEEE Trans. Syst. Man Cybern.* v(9), 62–66 (1979)
- [12] Mariano, E., Lins, R.D., Pereira e Silva, G.F., Fan, J., Majewicz, P., Thielo, M.: Correcting Specular Noise in Multiple Images of Photographed Documents. In: ICDAR 2011 - International Conference on Document Analysis and Recognition, pp. 111–116. IEEE Press, Pequim (2011)
- [13] ImageJ: <http://rsbweb.nih.gov/ij/> (accessed March 20, 2013)

## **A.7 Publicação sobre remoção de Embaçamento**

(OLIVEIRA et al., 2011) – D. Oliveira; G. F. P. Silva e R. D. Lins. Deblurring Textual Document Images. In: The Ninth International Workshop on Graphics Recognition, vol.1, pp: 154-157.

# Deblurring Textual Document Images

Daniel M. Oliveira, Rafael Lins, Gabriel P. Silva  
Departamento de Eletrônica e Sistemas - UFPE  
Recife - Brazil  
{daniel.moliveira, rdl, gabriel.psilva}@ufpe.br

Jian Fan<sup>1</sup>, Marcelo Thielo<sup>2</sup>  
Hewlett-Packard Labs.  
<sup>1</sup>Palo Alto - USA, <sup>2</sup>Porto Alegre - Brazil  
{jian.fan, marcelo.resende.thielo}@hp.com

**Abstract**—Documents digitized by portable cameras or flatbed scanners may exhibit some blurred areas. Most deblurring algorithms are hard to implement and slow. Often they try to solve the problem for any kind of image. In the case of text document images, the transition between characters and the paper background has a high contrast. With that in mind, a new algorithm is proposed for deblurring of textual documents; there is no need to estimate the PSF and the filter can be directed applied to the image.

*Deblurring; blur; camera documents; scanner documents;*

## I. INTRODUCTION

The recent paper [9] presents a taxonomy for noises in document images and, besides providing an explanation of how such noise appeared in the final image, may provide pointers to the literature that show ways of avoiding or removing it. Noise is defined here as any phenomenon that degrades document information. In the classification proposed [9], there are four kinds of noise:

1. *The physical noise – whatever “damages” the physical integrity and readability of the original information of a document. The physical noise may be further split into the two sub-categories proposed in as internal and external.*
2. *The digitization noise – the noise introduced by the digitization process. Several problems may be clustered in this group such as: inadequate digitization resolution, unsuitable palette, framing noises, skew and orientation, lens distortion, geometrical warping, out-of-focus digitized images, motion noises.*
3. *The filtering noise – unsuitable manipulation of the digital file may degrade the information that exists in the digital version of the document (instead of increasing it). The introduction of colors not originally present in the document due to arithmetic manipulation or overflow is an example of such a noise.*
4. *The storage/transmission noise – the noise that appears either from storage algorithms with losses or from network transmission. JPEG artifact is a typical example of this kind of undesirable interference.*

The blur noise has the effect of unsharpening images. Depending on how it arises it may be included in any of the four categories above. The physical blur may be the result of document “washing”, for instance, in which a document, printed with water soluble ink, gets wet. Blur may also be the result of unsuitable digitization, due to several reasons: non-flat objects, digitization errors, out of focus, motion etc. The presence of blur may be an indicator of low quality digitization, but can also be associated with other problems such as the scanning of hard-bound volumes. Blur may be the result of unsuitable filtering, such as a Gaussian or low-pass filter. And finally, blur may appear as the result of storing images in a file format with losses that perceptually degrades the image.

The technical literature points at several approaches proposed for deblurring images in general. To list a few of them: Demoment [2] uses statistics, Neelamani, Choi, and Baraniuk [3] use Fourier and wavelet transforms, references [4] and [5] apply variational analysis, and Roth and Black [6] use total variation and Field of experts. Most times the computational complexity is prohibitively high and can yield undesirable artifacts such as ringing [7] as presented in Figure 1.



Figure 1. Ringing artifact [7]

The most successful approaches to blur removal point at focusing at one specific kind of blur. For instance, the literature presents several algorithms [11, 12, 13, 14, 15, 16, 17] that address the problem of motion blur, an specific kind of digitization noise.

In this paper, to increase the chances of better deblurring, the application domain is restricted to monochromatic

scanned documents in which the blur is a digitization noise originated from the unsuitable document placing on the scanner flatbed due to a number of factors, one of which is book binding warping [10]. The document images treated here are basically constituted by text and plain paper background. The transition between them in the original physical document is sharp. Using this fact a new algorithm is proposed by using nearby pixels to increase the difference between them, no Point Spread Function (PSF) [18] estimation is done and blur is minimized into a direct application of the image.

## II. THE NEW METHOD

### A. Blur effect

The study performed here on the compensation of the blur effect was done for scanned images. Patterns were arranged in an elevated plane model [1] as illustrated in Figure 2 with  $\psi = 0$ . A HP PhotoSmart C4280 and HP ScanJet 5300c scanners were used. Some of the results are shown in Figure 3, where one may observe that in this case blur kernel can be approximated to a 1D horizontal function.

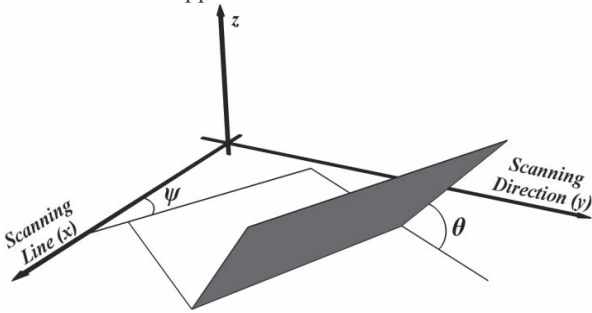


Figure 2. Elevated plane [1]

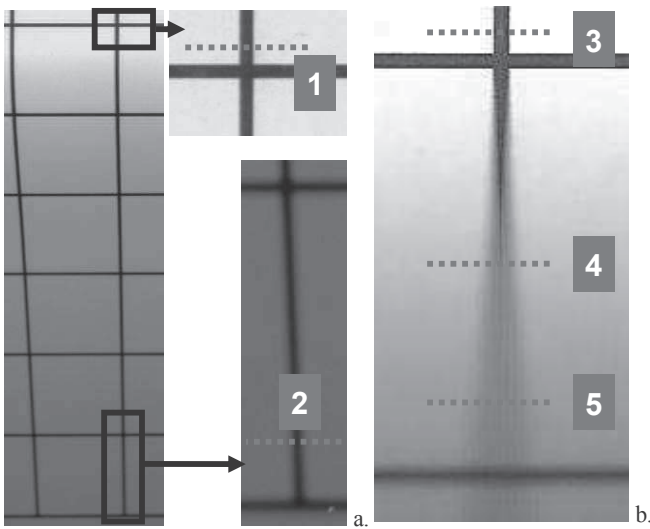


Figure 3. Grid image: ScanJet,  $\theta = 45^\circ$  (left), PhotoSmart C4280,  $\theta = 30^\circ$  (right)

In both images shown in Figure 3, as the paper is farther away from the scanner flatbed the blur increases and illumination is weaker; this happens due to the scanning

device being calibrated to digitize documents at a pre-defined distance that is exactly the flatbed surface.

Figure 4 presents two cross sections of the left image of Figure 3 on areas with and without blur; in this case the blurred stroke can be recovered.

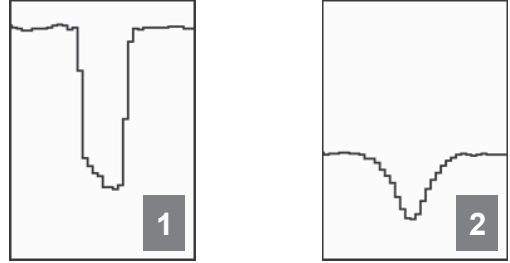


Figure 4. Cross sections of Figure 3.a: unblurred (left) and blurred (right).

The image on the right hand side of Figure 3 shows strong blur noise; three cross sections are presented in Figure 5, where the “4” one is the beginning when the blur kernel turns wider than the grid thickness. In such case the blur noise cannot be removed.

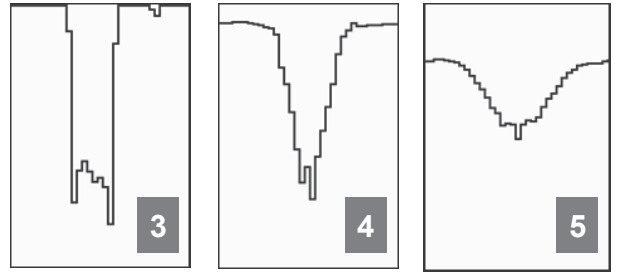


Figure 5. Cross sections of Figure 3.b: unblurred (left); kernel equal to stroke width (center); totally blurred (right)

### B. Reconstruction function

Figure 4 shows that if a signal has a blur kernel smaller than the grid thickness it is possible to improve or even recover the un-blurred signal. In the case of characters, the scenario is similar in a smaller scale.

For the values that belong to the paper background, the blurred intensity value is closer to the unblurred ones. In the same way, for blurred strokes values are closer to undistorted part. In this way a S-function can be built, with input and output varying from 0 to 1, whereas the output is below the line of the identity function between 0 and 0.5, and above it between 0.5 and 1.0.

This work proposes function  $S(t)$  defined by eq. 1 with the fixed parameter  $p$  that varies between 0 and 1.0, which controls how strong the correction will be. For  $p$  values closer to 0, the function shape looks similar to a step function with higher transitions; for values closer to 1.0, the shape gets closer to a sin function scaled by  $\pi$ . Figure 6 shows two plots for  $p = 0.06$  and  $p = 1.00$ .

$$S(t) = 0.5 - 0.5 \times \text{sign}(\cos t\pi) \times |\cos t\pi|^p \quad (1)$$

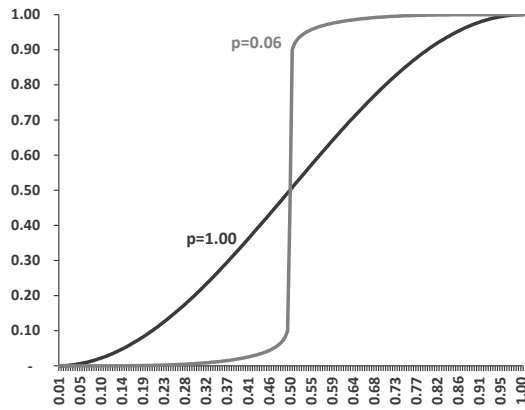


Figure 6. S-function plots for  $p$  equal to 1.00 and 0.06

To apply the S-shape function, two reference values must be determined for the paper background and character stroke. This is done by looking out in a window for the pixel with largest and lowest intensity, then to obtain the unblurred value eq. (2) is applied, where  $I_b$  is the blurred intensity value (i.e. original image),  $min$  and  $max$  are the lowest and the highest intensity values in the given window, respectively.

$$I_u = \left[ S \left( \frac{I_b - min}{max - min} \right) \times (max - min) \right] + min \quad (2)$$

One may notice that using large windows and low values for  $p$  is more intrusive than the other way round. The choice of this parameter will depend on the blur level and kernel size.

### III. RESULTS AND ANALYSIS

Figure 7 shows the results of applying the proposed algorithm to Figure 3.b using a  $7 \times 7$  window. One may observe that vertical line grid was recovered until blur kernel got larger than a  $7 \times 7$  window; although the blurred horizontal line on the bottom part could be partially recovered.

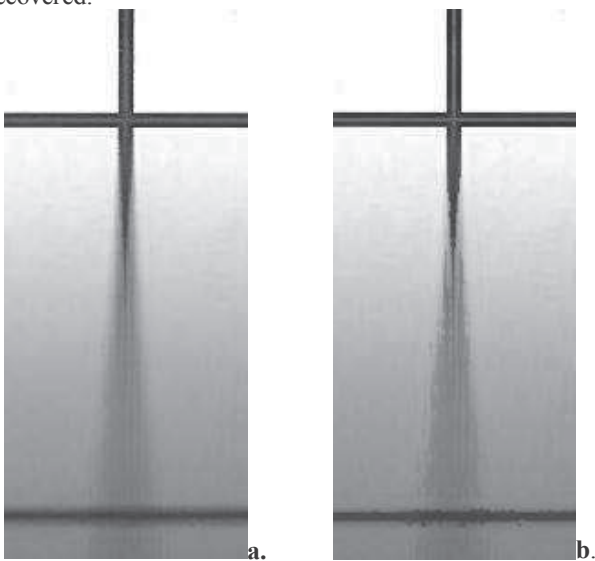


Figure 7. Output with  $7 \times 7$  window with  $p$ : 1.0 (a); 0.25(b)

Increasing the window size is possible to recover the area where the blur is larger, which is shown in Figure 8 with windows of sizes  $11 \times 11$  and  $19 \times 19$  and  $p = 0.25$ .

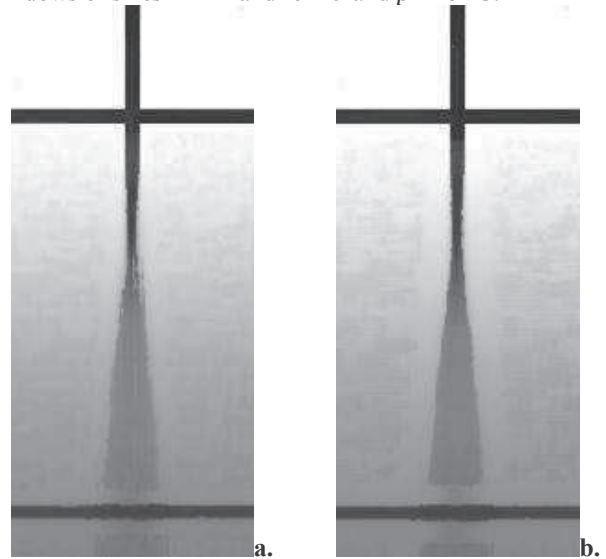
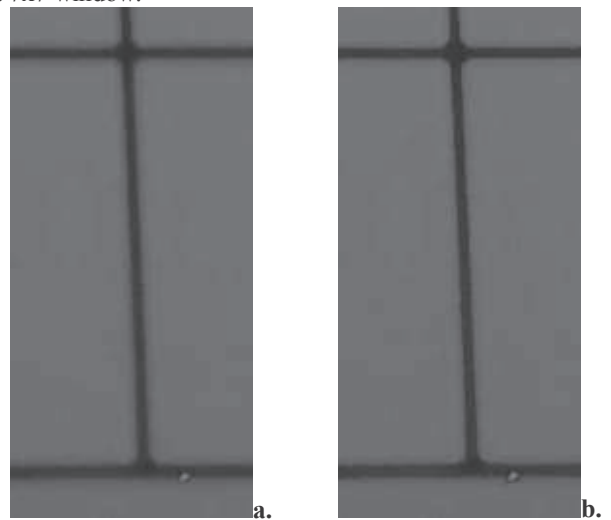
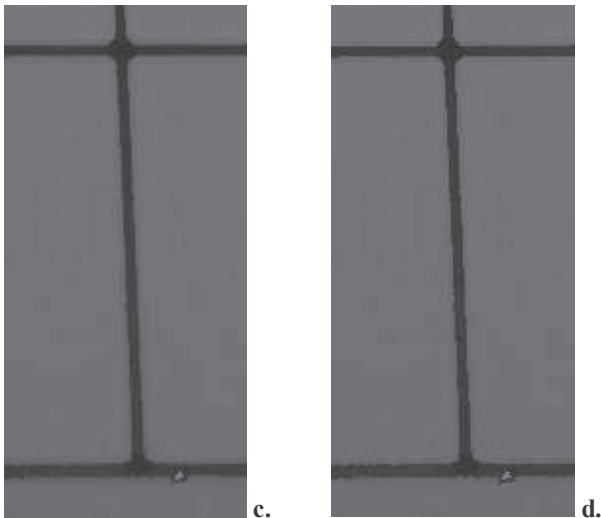


Figure 8. Results with  $p = 0.25$ :  $11 \times 11$  window (a);  $19 \times 19$  window (b);

For figure 2.a, in which the blur is weaker than in 2.b, Figure 9 provides the results for same parameters. One may note that satisfactory results were obtained for  $p = 0.5$  with a  $7 \times 7$  window.





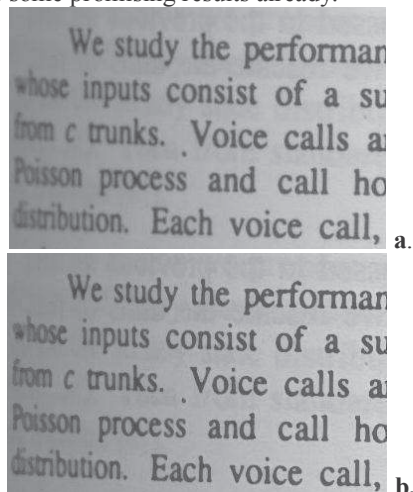
**Figure 9.** Results for 7x7 window:  $p = 1.00$  (a);  $p = 0.5$  (b);  $p = 0.25$  (c);  $p = 0.06$  (d)

Finally, Figure 10 shows some examples of de-blurring applied to real document images.

#### IV. CONCLUSIONS

The study performed here shows that focusing the scope of application of deblurring algorithms stand a better chance of better solving such a complex noise that may appear due to several sources: physical, digitization, filtering and storage. This paper presents an algorithm to compensate the digitization non-constant blur that appear in scanning bound grayscale documents, for instance.

The automatic inference of the parameters of the algorithm through the use of blur intensity classifier such as the one described in reference [8] is under implementation and shows some promising results already.



**Figure 10.** Document processed with 5x5 window: original image (a);  $L L r w f(b)$ ;  $L L r r x(c)$

#### REFERENCES

- [1] Ukida, H. and Konishi, K. . 3D Shape Reconstruction Using Three Light Sources in Image Scanner. IEICE Trans. on Inf. & Syst., Vol.E84-D, No. 12, pp.1713-1721, Dec. 2001.
- [2] Demoment, G.. Image reconstruction and restoration: Overview of common estimation structures and problems. IEEE Transactions on Acoustics, Speech, & Signal Processing, 37(12), 2024–2036, 1989.
- [3] Neelamani, R., Choi, H., and Baraniuk, R. G.. Wavelet-based deconvolution for ill-conditioned systems. Proc. of IEEE ICASSP, Vol. 6, pp. 3241–3244, March 1999.
- [4] Chambolle, A., and Lions, P. L.. Image recovery via total variation minimization and related problems. Numerische Mathematik, 76(2), 167–188, 1997.
- [5] Rudin, L. I., Osher, S., and Fatemi, E.. Nonlinear total variation based noise removal algorithms. Physica D, 60, 259–268, 1992.
- [6] Roth, S., and Black, M. J.. Fields of experts: a framework for learning image priors. CVPR, Vol. 2, pp. 860–867, 2005.
- [7] Joshi, N. S.. Enhancing photographs using content-specific image priors. Phd thesis, University of California, San Diego, 2008.
- [8] Lins, R.D, Silva, G.F.P., Banerjee, S., Kuchibhotla, A. and Thielo, M. Automatically Detecting and Classifying Noises in Document Images, ACM-SAC'2010, ACM Press, v.1. p.33 – 39, March 2010.
- [9] Lins, R.D. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. ICIAR 2009. LNCS v. 5627. p. 844-854, Springer Verlag, 2009.
- [10] Lins, R.D., Oliveira, D. M., Torreão G., Fan, J., and Thielo, M., Correcting Book Binding Distortion in Scanned Documents. ICIAR 2010. LNCS 6112, pp. 398-408, Springer Verlag, 2010.
- [11] Chang, M. M., Tekalp, A. M., and Erdem, A. T., Blur identification using the bispectrum, *IEEE Trans. Signal Process.*, Vol. 39, No. 10, 1991, pp. 2323-2325.
- [12] Mayntz, C., Aach T., and Kunz D., Blur identification using a spectral inertia tensor and spectral zeros, *Proc. of IEEE ICIP*, 1999.
- [13] Cannon M., Blind deconvolution of spatially invariant image blurs with phase, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 24, No. 1, 1976, pp. 56-63.
- [14] Biemond, J., Lagendijk, R. L., and Mersereau R. M., Iterative methods for image deblurring, *Proc. of the IEEE*, 1990, pp. 856-883.
- [15] Rekleities, I. M., Optical flow recognition from the power spectrum of a single blurred image, *Proc. of IEEE ICIP*, 1996.
- [16] Moghaddam, M.E. and Jamzad, M., Motion blur identification in noisy images using fuzzy sets, *Proc. IEEE ISSPIT*, Athens, 2005.
- [17] Lokhande, R., Arya, K.V., Gupta, P. Identification of parameters and restoration of motion blurred images, ACM-SAC'2006, Dijon, 2006.
- [18] Jain, A.K. , Fundamentals of digital image processing, Prentice-Hall, Inc., Upper Saddle River, NJ, 1989.

## **A.8 Publicação sobre remoção de Bordas**

(SILVA et al., 2013) G. F. P. Silva ; R. D. Lins ; A. R. Silva. A New Algorithm for Background Removal of Document Images Acquired Using Portable Digital Cameras. In: Lecture Notes in Computer Science, Ed. Springer, vol.1, pp: 290-298.

# A New Algorithm for Background Removal of Document Images Acquired Using Portable Digital Cameras

Gabriel França Silva<sup>1,2</sup>, Rafael Dueire Lins<sup>1</sup>, and André Ricardson Silva<sup>1</sup>

<sup>1</sup> Universidade Federal de Pernambuco, Recife, Brazil

<sup>2</sup> Universidade Federal Rural de Pernambuco, Garanhuns, Brazil  
{gfps,rdl}@cin.ufpe.br

**Abstract.** Document images digitalized with cameras are framed with parts of the background where the document lied on. In the case of images acquired with portable digital cameras such background may be complex, of non-uniform colors and texture. This paper presents a new algorithm designed to remove the background of document images acquired using portable digital cameras. Comparative tests performed show that the new algorithm largely outperforms its predecessor.

**Keywords:** portable digital cameras, background removal, image processing.

## 1 Introduction

Digital cameras are a pervasive good nowadays. Portable models have already reached professional standards and offer resolutions over 12 Mpixels, while several mobile phones have good quality cameras embedded of resolution reaching 5 Mpixels. At the same time, the prices of digital cameras are dropping fast. This improvement in price-performance greatly expanded the number of users of digital cameras in such a way that conventional analogical cameras have almost disappeared. Not long ago, document digitalization using cameras was restricted to very fragile illumination-sensitive historical documents; it is becoming more popular every day. Professionals from different areas now use these devices as a quick way to acquire images of documents, taking advantage of its availability, low weight, portability, low cost, small size, etc. A new research area of [5] was opened and it is rapidly unfolding in different directions.

This article focuses on the automatic removal of the frame that often appears when a photo from a document is taken, as parts of the background where lies the document tend to appear in the document photos. Due to the diversity of backgrounds, which may be non-uniform in color and texture, the automatic removal of the framing border is far from being a trivial task. An algorithm for such purpose must be as general as possible and should impose as few restrictions as needed. This problem has some similarities with the removal of the framing border that appears in automatically fed scanned documents [1] [2] [3] [4] [6] [9]. Depending on a number of factors as the size of the documents their conservation state, their physical integrity, the presence or absence of dust in documents and scanner flatbed, etc. very often the generated image



is not framed by a black border either solid or stripped. This undesirable artifact, also known as marginal noise, not only drops the quality of the resulting image whenever seen on the screen of a CRT display, but also consumes storage space, network bandwidth for transmission, and a large amount of ink or toner for printing. However, the challenge of processing images of documents digitalized using digital cameras is somewhat more complex than those acquired using a scanner. Figure 1 illustrates the difference between document images obtained using the two devices.

Some specific problems [11] arise in the digitalization of documents using portable digital cameras without mechanical support. The first of all is the removal of background. As already mentioned, often the photo paper goes beyond the size of the document and incorporates parts of the area that served as a mechanical support to take the picture of the document. The second problem is due to the slope found in the image relative to the photo axes; there is often some degree of tilt in the document image. The third problem is the non-frontal view, due to the same reasons that give rise to slope, which makes more complex the complete removal of the document edges. A fourth problem is caused by the distortion of the camera lenses. This means that the distortion of perspective is not a straight line but a convex line, depending on the quality of the lenses and the relative position of the camera and the document. The fifth difficulty is due to the lack of uniform illumination. While in scanners the illumination of a document is uniform, in cameras, besides the unevenness of the strobe flash illumination one may have other light sources from the environment. The removal of the border frame manually is not practical because of the need for a specialized user and the time consumed in the operation.

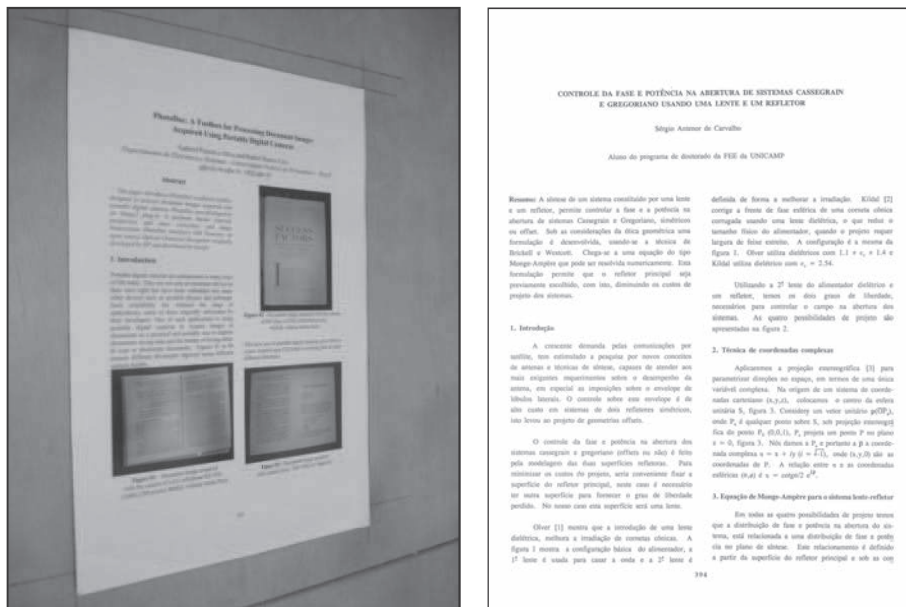


Fig. 1. Example of a photo and scanned document

This paper is organized as follows. The section that comes after this Introduction describes the features of the documents acquired with digital cameras and used in testing the solution proposed herein. Section 3 introduces the new algorithm for automatic border removal. In Section 4, the results obtained are presented and compared with the other solution described in the technical literature. The Conclusions and Lines for Further Work close this paper.

## 2 Document Features

One thousand images of documents were used in the development of the algorithm presented here. They were acquired using four different resolutions (3.1, 4.2, 5.1 e 7.2 Mpixels), the original document was printed or written in translucent paper in which only weak back-to-front interference [8] was observed, and were stored using the camera standard Jpeg (1% loss) compression algorithm.

Two hundred of the one thousand documents in the test set are formed by freely hand held documents, whose content range from bureaucratic documents, pages extracted from magazines and phone directories. Amongst the bureaucratic documents one may find typewritten ones, forms and handwritten documents. The conservation state of documents also varied widely. Some documents may have their physical borders torn off, thus making the border detection more difficult.

The remaining eight hundred documents were obtained by photographing the pages of the Proceedings of Second International Workshop on Camera-Based Document Analysis and Recognition (CBDAR 2007) using the mechanical support called the "Planetarium", a picture of which is shown in Figure 2. The "Planetarium" allows the creation of an image database in which one knows the angle the camera makes with

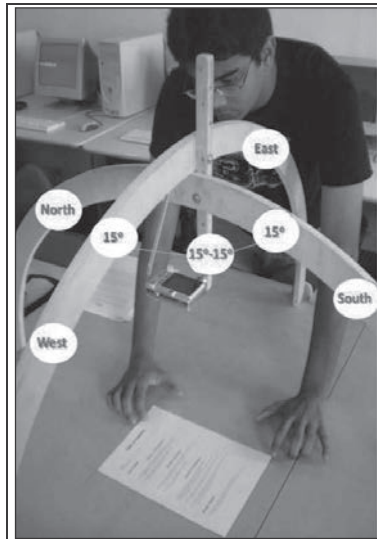


Fig. 2. The "Planetarium"

the normal line to the plane where the physical document lies on. For the experiments described here images of the following positions were chosen: angle of  $0^\circ$  with height Low (40 cm) and (60 cm),  $15^\circ$  and  $30^\circ$  at Low height. The same documents were also photographed without the mechanical support of the "Planetarium".

### 3 The New Algorithm

The new algorithm developed to be part of PhotoDoc [12], a software platform for processing images acquired with digital cameras, is an improvement of the algorithm presented in reference [9], in such a way to solve some of the drawbacks in the previous one, as may be observed in Table 1.

**Table 1.** Comparison between the Silva-Lins [11] and the New Algorithm proposed

Problems	Silva-Lins [11]	New Algorithm
<b>Fixed Parameters</b>	Yes (set to a given document database)	No (parameters vary for each image)
<b>Applicable to Low Resolution Documents</b>	No	Yes
<b>Documents with complex layout</b>	Low performance	High performance
<b>Paper and frame border of similar colors</b>	Low performance	Medium performance

Before detailing the mechanism of the new algorithm, one needs to clarify the notation adopted in it:

- Source pixel block: stands for the set of pixels to be classified either as being border or not.  $S$
- Support pixel block: the set of pixels that is used to compare with the source pixel block.
- Border block: stands for a source block that is formed only by pixels that do not contain document information.

Most of the classical methods for border detection are based on computations between pixels and its neighbors. The new scheme is based on Shannon entropy [13] [14] calculated for each component of the pixel (RGB pattern) between adjacent blocks. The algorithm starts with five pre-defined support blocks defined the following way: the central block (1/9 of the image) and the lateral blocks (10% of the image height and 10% of the image width) left, right, upper and lower. Figure 3 shows the position of those blocks. The first source block corresponds to 1/81 of the image in its central part.

#### Step 1 - Calculus of the Entropy of the Blocks of Pixels

One calculates the entropy for each R, G and B component of the blocks of pixels. Such entropy will define a threshold value for each component. The calculus is performed by using the following formula:

$$H_{component} = -\sum_{i=0}^{255} p_i \log_2(p_i)$$

where  $\{p_0, p_1, \dots, p_i\}$  is a priori distribution given by:

$$p_i = \frac{\text{the\_number\_of\_pixels\_in\_a\_given\_comp\_in\_a\_block}}{\text{the\_total\_number\_of\_pixels\_in\_a\_block}}$$

For each  $t$  one calculates the distribution a posteriori  $\{P_t, 1-P_t\}$ , while the entropy follows the distribution:

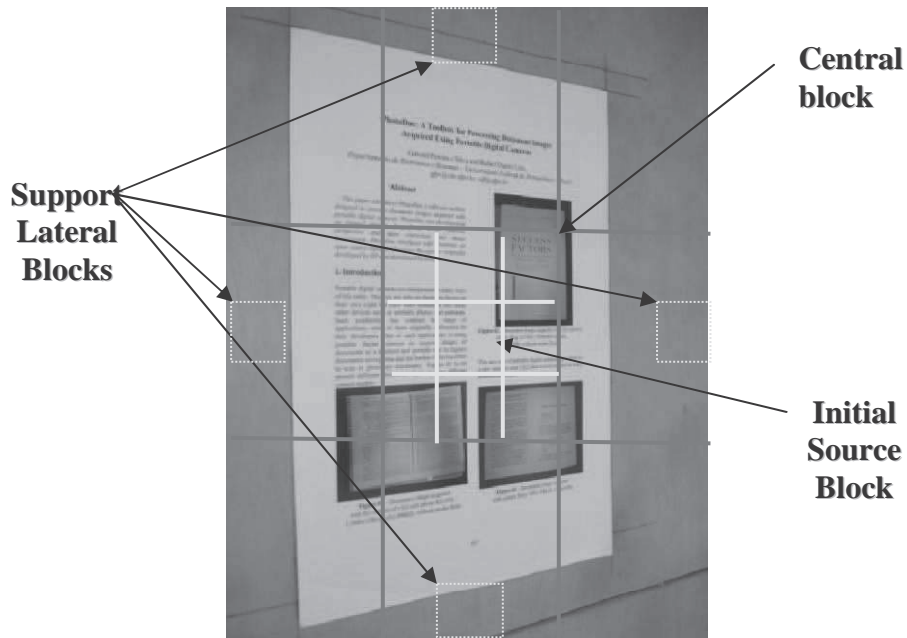
$$H'(t) = -P_t \log(P_t) - (1 - P_t) \log(1 - P_t)$$

Finally, one determines the optimal limit that minimizes the  $|e(t)|$  value given by:

$$|e(t)| = \left| \frac{H'(t)}{H/\log(256)} - \alpha(H/\log(256)) \right|$$

where  $\alpha$  is a loss factor, experimentally determined, given by:

$$\alpha(H/\log(256)) = \begin{cases} -(3/7)(H/\log(256)) + 0.8 & \text{if } H/\log(256) < 0.7 \\ H/\log(256) - 0.2 & \text{if } H/\log(256) \geq 0.7 \end{cases}$$



**Fig. 3.** Start-up support pixel blocks. The initial source central block corresponds to 1/81 block of pixels in the center of the image.

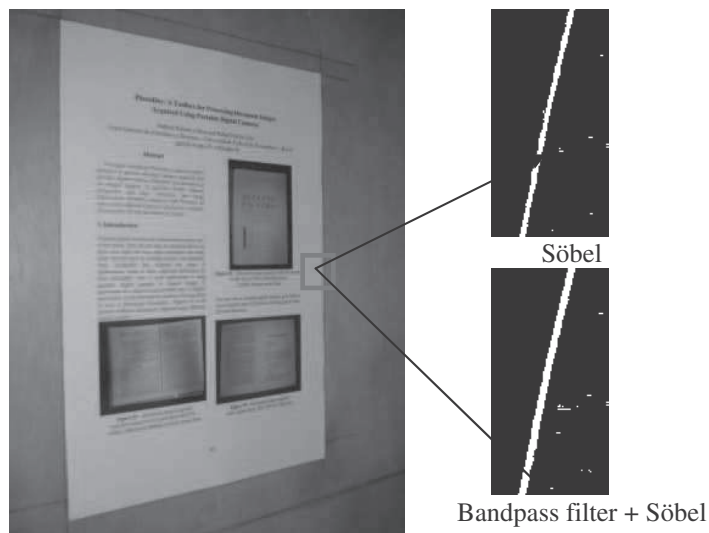
One starts calculating the threshold for the four support lateral blocks of pixels (left, right, top, bottom) and the central block (which corresponds to 1/9 of the image).

### Step 2: Searching the Boundaries of a Document

Once the thresholds were calculated, one looks for the limits of the paper of the document, starting from the center outwards, until it reaches the document borders. One defines a new source block corresponding to 1/81 of the central starting the image scanning, where the new blocks are defined corresponding to half of the height and of the previous block in the case of the horizontal scan, and half of the width of the previous block in the case of the horizontal scan. For each new block, one calculates the threshold of the R, G, B component. Then one compares them with the threshold of the support blocks. If more than one threshold of the source block threshold exceed in over 25% its corresponding threshold in the support central block and less than 35% in relation to at least two support lateral blocks, this block is considered a border block. Otherwise, it is compared with the next source block.

### Step 3: Calculating the Tolerance of the Algorithm

The variation in illumination due to the unevenness of the strobe flash and the interfering light sources in the environment, the Step 2 described above may not be precise in detecting the border blocks. Thus, one needs to estimate the tolerance values to be applied. Each "candidate" to become a border block is filtered with a bandpass filter and then the border detection algorithm of Söbel [15] is used to strengthen the contour of the document, thus allowing a more precise analysis of the features of the pixels of the content and background. Figures 4 and 5 show the results of the application of filters onto a block "candidate" to become a border block. All border blocks are merged together forming the contour of the document.



**Fig. 4.** Document image (7.2 Mpixels 15°S - 0°O) with similar color of paper and framing background

#### 4 Results

The algorithm presented above was programmed in JAVA and its executable is available under e-mail request to one of the authors. It was executed on an Intel Pentium

**Table 2.** Algorithm comparative validation test

Resolution Mpixels	Flash on	Acquisition method	Angle	Number of images	Detection algorithm 1	Detection algorithm 2
3.1	YES	Free hand	Not applicable	50	64%	88%
3.1	NO	Free hand	Not applicable	50	62%	88%
4.2	YES	Free hand	Not applicable	50	70%	91%
4.2	NO	Free hand	Not applicable	50	70%	91%
5.1	YES	Free hand	Not applicable	100	58%	98%
5.1	NO	Free hand	Not applicable	100	51%	95%
5.1	YES	"Planetarium"	0° Low	30	63.33%	100%
5.1	YES	"Planetarium"	0° High	30	60%	100%
5.1	YES	"Planetarium"	15°S - 0°O	35	34.28%	97.14%
5.1	YES	"Planetarium"	30°S - 0°O	35	28.57%	97.14%
5.1	NO	"Planetarium"	15°S -15°O	35	48.57%	97.17%
5.1	NO	"Planetarium"	15°S - 30°O	35	35.71%	100%
7.2	YES	Free hand	Not applicable	100	51%	98%
7.2	NO	Free hand	Not applicable	100	53%	98%
7.2	YES	"Planetarium"	0° Low	30	60%	100%
7.2	YES	"Planetarium"	0° High	30	60%	100%
7.2	YES	"Planetarium"	15°S - 0°W	35	34.28%	100%
7.2	YES	"Planetarium"	30°S - 0°W	35	28.57%	100%
7.2	NO	"Planetarium"	15°S -15°W	35	34.28%	100%
7.2	NO	"Planetarium"	15°S - 30°W	35	28.57%	100%
Average accuracy:					61%	96%

Core 2 Duo, 2.8 GHz clock, 4GB RAM, elapsed on average 100ms per image, in old algorithm is developed in C and elapsed 79 ms per document on average. The result is shown in Table 2.

## 5 Conclusions and Lines For Further Work

Border detection of either scanned or photographed documents is the first and fundamental step in its processing. This paper presented an automatic algorithm for border detection that is able to work with documents of different resolutions and in perspective. The new method was compared with the other algorithm for similar purpose in the literature and presented an accuracy gain of 35% in the number of documents for which the borders were correctly detected rising from 61% to 96%.

Along the lines for further work the authors plan to focus on low resolution images (under 3 Mega pixels), and also in improving the robustness of the detection of the borders of images of (book) bound documents in which the book spine often presents a blurred region. If this paper were accepted all the code for the algorithms and test images will be made publically available.

**Acknowledgments.** Research reported here was partly funded by CNPq – Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico (Brazilian Government).

## References

- [1] Ávila, B.T., Lins, R.D.: A New Algorithm for Removing Noisy Borders from Monochromatic Documents. In: ACM-SAC 2004, pp. 1219–1225. ACM Press (2004)
- [2] Baird, H.S.: Document image defect models and their uses. In: Proc. 2nd Int. Conf. on Document Analysis and Recognition, Japan, pp. 62–67. IEEE Comp. Soc. (1993)
- [3] Fan, C., Wang, Y.K., Lay, T.R.: Marginal noise removal of document images. *Patt. Recog.* 35, 2593–2611 (2002)
- [4] Kanungo, Haralick, R.M., Phillips, I.: Global and local document degradation models. In: Proc. 2nd Int. Conference on Document Analysis and Recognition, pp. 730–734 (1993)
- [5] Liang, L., Doermann, D., Li, H.: Camera-Based Analysis of Text and Documents: A Survey. *International Journal on Document Analysis and Recognition* (2005)
- [6] Lins, R.D., Guimarães Neto, M.S., França Neto, L.R., Rosa, L.G.: An Environment for Processing Images of Historical Documents. *Microprocessing & Microprogramming*, North-Holland, pp. 111–121 (1995)
- [7] Lins, R.D., Machado, D.S.A.: A Comparative Study of File Formats for Image Storage and Transmission. *Journal of Electronic Imaging* 13(1), 175–183 (2004)
- [8] da Silva, J.M.M., Lins, R.D., da Rocha, V.C.: Binarizing and filtering historical documents with back-to-front interference. In: ACM SAC 2006, pp. 853–858. ACM Press (2006)
- [9] Pe Silva, G.F., Lins, R.D., da Silva, J.M.M., Banergee, S., Kuchibhotla, A., Thielo, M.: Enhancing the Filtering-Out of the Back-to-Front Interference in Color Documents with a Neural Classifier. In: ICPR 2010, pp. 2415–2419. IEEE Press (2010)

- [10] Shapiro, L.G., Stockman, G.C.: Computer Vision (March 2000), <http://www.cse.msu.edu/~stockman/Book/book.html>
- [11] Gomes e Silva, A.R., Lins, R.D.: Background removal of document images acquired using portable digital cameras. In: Kamel, M.S., Campilho, A.C. (eds.) ICIAR 2005. LNCS, vol. 3656, pp. 278–285. Springer, Heidelberg (2005)
- [12] Silva, G.F.P., Lins, R.D.: PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. In: CBDAR 2007. IAPR Press (2007)
- [13] Shannon, C.: A mathematical theory of communication. Bell System Technology Journal 27, 370–423 and 623–656 (1948)
- [14] Abramson, N.: Information Theory and Coding. McGraw-Hill Book Co. (1963)
- [15] Söbel, I.E.: Camera Models and Machine Perception. Ph.D. dissertation. Stanford University. Palo Alto. USA (1970)



## **A.9 Publicações sobre Transcrição automática de imagens de Documentos Históricos**

(SILVA e LINS, 2011) - G. F. P. Silva; R. D. Lins. An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. In: International Conference on Document Analysis and Recognition, vol.1, pp: 553-557.

(ALMEIDA et al., 2011) - A. B. S. Almeida; R.D. Lins; G. F. P. Silva. Thanatos: automatically retrieving information from death certificates in Brazil. In: Workshop on Historical Document Imaging and Processing, vol.1, pp: 146-153.

(SILVA e LINS, 2012b) - G. F. P. Silva; R. D. Lins. Generating Training Sets for the Automatic Recognition of Handwritten Documents. In: Advances in Character Recognition, 1ed, New York: InTech, 2012, pp: 155-174.

(SILVA e LINS, 2014) G. F. P. Silva; R. D. Lins. Automatic Training Set Generation for Better Historic Document Transcription and Compression. In: International Workshop on Document Analysis Systems, vol.1. pp: 20-31.

## An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents

Gabriel Pereira e Silva  
DES – CTG – UFPE  
Recife, PE, BRAZIL  
[gfps.cin@gmail.com](mailto:gfps.cin@gmail.com)

Rafael Dueire Lins  
DES – CTG – UFPE  
Recife, PE, BRAZIL  
[rdl.ufpe@gmail.com](mailto:rdl.ufpe@gmail.com), [rdl@ufpe.br](mailto:rdl@ufpe.br)

**Abstract** — Automatic optical character recognition is an important research area in document processing. There are several commercial tools for such purpose, which are becoming more efficient every day. There is still a lot to be improved, in the case of historical documents, however, due to the presence of noise and degradation. This paper presents a new approach for enhancing the character recognition in degraded historical documents. The system proposed consists in identifying regions in which there is information loss due to physical document degradation and process the document with possible candidates for the correct text transcription.

**Keywords** — historical documents, OCR, character recognition, physical noises.

### 1. Introduction

There is today a huge quantity of paper legated historical documents. Such documents, accumulated over the centuries, contain important information that is the memory of mankind. Most of those documents were not digitized yet, thus the access to them is limited to very few specialists in libraries or museums. There is a wide effort throughout the world in making historical documents [4], books, and even conference proceedings [2] available in the world-wide-web. Such effort is only possible using automatic processing for image enhancement and transcription. The OCR systems available today do not yield satisfactory results for historical documents, as they are very sensitive to a wide range of noises [1]. The taxonomy for noises in paper documents proposed in reference [3], asserts that in historical documents there is a prevalence of physical noises originated either by the natural paper degradation in unsuitable storage conditions. Historical documents often exhibit back-to-front interference [4] (also known as bleeding or show-through [9]), paper aging, washed-out ink, folding marks, stains, and torn-off parts, as one may observe in the document shown in Figure 1, a hand written letter from Joaquim Nabuco, a Brazilian statesman, writer, and diplomat, one of the key figures

in the campaign for freeing black slaves in Brazil (b.1861-d.1910), kept by the Joaquim Nabuco Foundation [18] a social science research institute in Recife, Brazil.

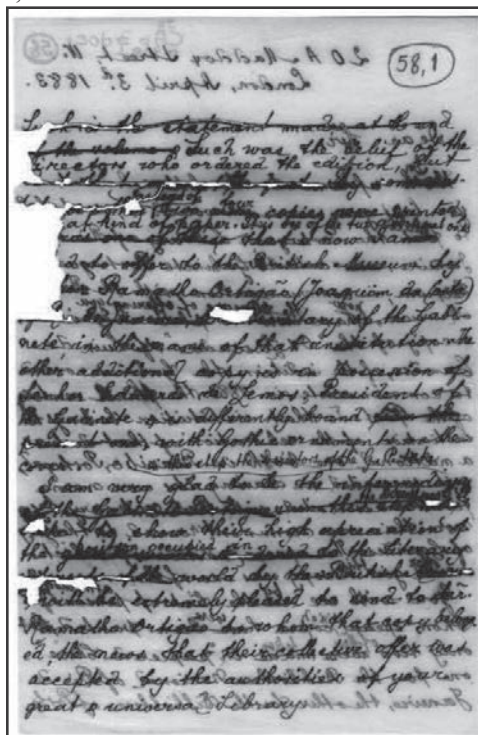


Figure 1. Historical document from Joaquim Nabuco's bequest

If a document has the physical noises one finds in the document of Figure 1, one stands very little chances of having any success in automatic image-into-text transcription with any commercial system today [6][7][8]. The idea of the system proposed here is to look for torn-off regions or holes and try to “complete” such areas with possible images in such a way as to maximize the probability of the correct transcription of the word as a whole. This way, instead of performing character-to-character recognition as used in conventional OCR tools, the system proposed here

infers a set of possible words and chooses the one with the highest probability to occur. Incomplete words in holes or torn off areas are completed taking into account the parts left of characters completing them with characters that may possibly “fit” the remaining parts. The OCR drives the choice of the most suitable part to fill in the holes. The choice of the most probable word may be helped by using a dictionary of terms already recognized in the document or in the file as a whole.

The structure of this paper is as follows. Section 2 describes the pre-processing stages of historical document handling. Section 3 presents the automatic image into text transcription system. Section 4 discusses the results obtained and draws lines for future work.

## 2. Image Pre-processing

The direct application of OCR to noisy images of documents tends to yield poor quality transcription [1]. Thus, to enhance the quality of the images of historical documents for transcription the pre-processing scheme presented in Figure 2 is applied.

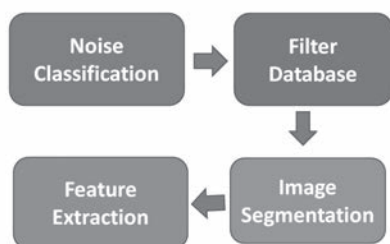


Figure 2. Pre-processing scheme

The pre-processing scheme presented in Figure 2 encompasses four modules. The first module performs “Noise Classification” it analyses the document and a neural classifier similar to the one described in reference [5], but specially tailored for historical documents detects which noises are present in a document. Besides that, the “intensity” of the back-to-front interference (also known as bleeding or show-through) is measure. Determining the intensity of the back-to-front interference is an important step for the adequate filtering out of such noise in the second module, the “Filter Database”. The third module, the “Image Segmentation” is responsible for segmenting the document in lines of text, and those, at their turn, into characters. The last module of the pre-processing scheme performs “Feature Extraction” in which the features of the characters spotted in the previous step are classified.

## 2.1 Noise Classification and Filtering

The first pre-processing phase is responsible for the automatic generation of “noise maps” that may be present in the document image. The noise classifier described [5] was tuned and retrained to work with the kinds of noises more often found in historical documents such as detection of noisy borders, skew, incorrect image orientation, blur and back-to-front interference. In the case of the last noise the global classification is the result of three cascaded classifiers that detect the strength of the interference and classifies it in “light”, “medium”, and “strong”. In the case of “blur” detection, the classifier works in a similar fashion to the back-to-front one. Figure 3 presents sketches the noise classifier “architecture”.

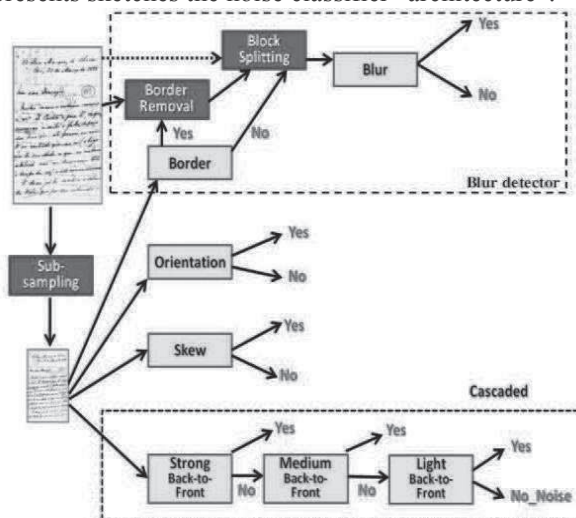


Figure 3. Noise classifier “architecture”

Besides those noises presented in the architecture above a new classification feature was added to detect the presence of holes and thorn-off regions, such as the noises shown in Figure 04.

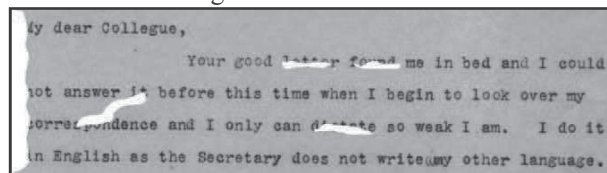


Figure 04. Image from Nabuco’s bequest with thorn-off regions and holes.

The new classifier besides making use of the set of features presented in reference [5], needs two new ones for better classification performance:

- Mean edge value (sharper edges have higher values);
- Mean image saturation.

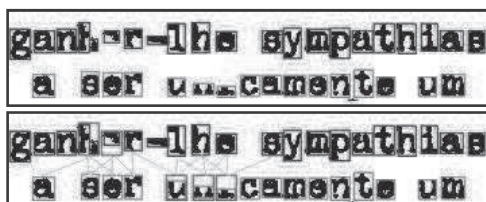
One hundred images with thorn-off regions and holes were used in the tests performed here (10 original and 90 synthetic images). Besides those images another hundred images without such noises were also tested. The 200 images were split into blocks totaling 2,053 “damaged” blocks and 18,000 “perfect” ones. The result of block classification appears on Table 1, where the entry “Holes” stands for blocks that cover thorn-off regions and holes altogether. One may observe that the accuracy of the classifier is higher than 98%, which may be considered excellent.

**Table 1** – Confusion Matrix for hole detection.

Classes	Holes	No holes	Accuracy
Holes	2,017	36	98.2%
No holes	78	17,922	99.5%

### 2.3 Segmentation

The character segmentation algorithm is based on the one in reference [10], suitably modified to handle degraded areas (thorn-off and with holes). The size of the characters in a block classified as having a hole takes into account the size the surrounding characters of the block under observation. Figure 05 top presents an example of the direct segmentation and at the bottom part the result of segmentation taking into account block classification and the size of the segmented characters in the surrounding areas.



**Figure 05.** Character segmentation. **(Top)** Direct. **(Bottom)** Taking into account block classification.

### 2.3 Feature Extraction

One of the most steps in the development of systems for automatic character recognition is the choice of the feature set to be compared. Such features should attempt to be as disjoint as possible to allow good discrimination. The technical literature [20][21][22][23] points at several successful feature sets for character recognition. The choice in this research was for extracting features in specific areas (zones) of the image of characters of the Latin alphabet (A, B, C..., Z). The set of features selected is generated by:

- Geometric Moments [20][23];
- Concavity Measurements [22];
- Shape Representation of Profile [21].

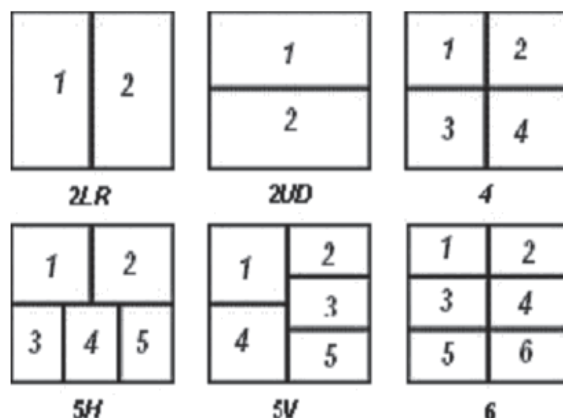
In this research 26 classes of characters were used, each of them with 3,100 character images, from which 1,500 are from NIST database [17], 100 were originated from the set of documents used for testing and the remaining 1,500 were obtained by random “erasure” of parts of characters. Tests were performed using three subsets for training, validation and benchmarking, which correspond to: 25%, 25%, and 50%, of the total, respectively.

## 3. The Automatic Transcription System

The transcription system presented here is based on how the human brain identifies and processes visual information. One of the most accepted theories about character recognition is that the brain looks only at specific parts of [11]. Thus, the literature presents several papers which adopt zoning as a classification strategy. From the features obtained in each image “zone” it is possible to organize meta-classes to generate sets of words that may “fit” the degraded areas of a document.

### 3.1 Zoning Mechanism

“Zoning” may be seen as splitting a complex pattern in several simpler ones. In the case of degraded texts, the concern of this paper, this becomes an important discrimination basis amongst classes, as the “real” information is limited only to some classes. Some researchers propose only the “empirical” zoning [11][12][14], in which each character is represented by a rectangle Z, that may assume several different formats, such as the ones presented in Figure 6. Other researchers propose methods of automatic zoning [13]. This work adopted the strategy of empirical zoning, looking at the best combination of zones targeting at obtaining the best meta-class formation for the degraded characters.



**Figure 6.** Z = 2LR, 2UD, 4, 5H, 5V and 6.

### 3.3 The Creation of Meta-classes

Meta-classes are called the “classes of classes”. Such approach targets at reducing the complexity of character recognition. In the specific case of this work, the term meta-class is used to express the set of possible words that may be generated from the block of characters with information loss. For such a purpose, a SOM [15] network was trained with different configurations and the Euclidean distance was used for group and map formation. After the SOM was trained another technique, the treeSOM, is applied to find the best cluster [16]. Such algorithm is adjusted through the choice of randomly chosen thresholds in the interval [0,1] to form different clusters for a given network. The experiments performed in this work took the following values of threshold: 0.6, 0.4, 0.3, 0.25 and 0.1. Each map of the SOM network builds a different cluster and the best amongst them is the one that minimizes the value of  $\mathcal{E}$  in Equation 01.

$$\mathcal{E} = m \frac{\sum_{C \in E} X_C}{\sum_{A, B \in E} \delta_{AB}} \quad (01)$$

The distance between the groups A and B, denoted by  $\delta_{AB}$  is equal to the average between the pairs of all elements of group A and all elements of group B. The density of the group denoted by  $X_C$ , is the average between the distances of all elements in the same group. The number of groups is  $m$ .

For all tests performed the best SOM network was the one that presented a 4x5 mesh configuration and threshold of 0.3. The clusters formed from this network are presented in Table 02.

Table 02 – The best clusters obtained	
Clusters	Meta-classes
G 01	m, n, r, v, y, w
G 02	c, e, o, p, u
G 03	b, d, l, k, s
G 04	f, i, j, t, z
G 05	a, g, h, q, x

For the formation of the “candidate” words one observes for each character the two clusters with the highest activation. This way it is possible to build a graph with which one generates all possible combinations of the characters in the defined clusters, as shown in the graph of Figure 7. After the graph completion, one sweeps the graph and forms all possible words. From those one sieves the valid ones by dictionary look-up

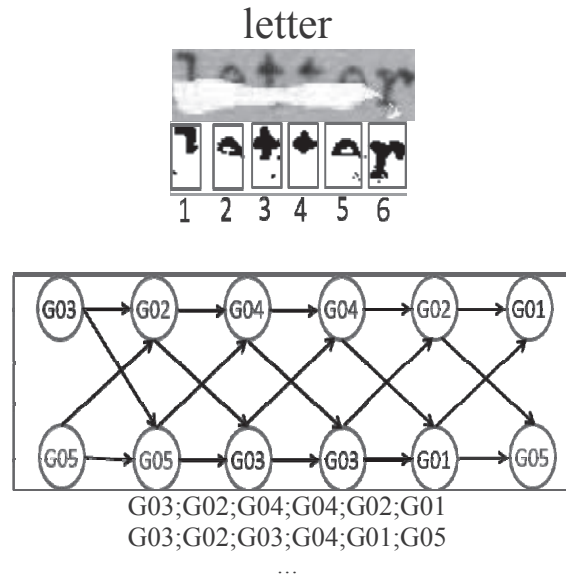


Figure 7. Example of the generation of graphs for the generation of words.

## 5. Experimental Results

The proposed method was tested with a file of 100 historical documents kept by the Fundação Joaquim Nabuco [18]. The documents were digitized using a flatbed scanner with 200 dpi. Such letters were transcribed and corrected yielding 13,833 words. From those, 3,814 present physical noises (holes and thorn-off parts) that implied in information losses reaching a maximum of 70% of the original word. To access the result of the improvement in recognition rate after word correction the test documents were transcribed with the commercial OCR tool ABBYY FineReader 10 Professional Editor. Figure 8 presents a sample of the results obtained. In part (c) of Figure 8 one sees in green the inserted characters, in red the ones used as models for substitution and blue is no model found in the document area.

A quantitative analysis was performed to evaluate the improvement in transcription rate presented in this work. The same one hundred documents were transcribed twice by the tool ABBYY FineReader 10 Professional Editor and the results for the 3,582 with physical noises is shown in Table 03, classified between “Correct”, “Incorrect” and “Not-recognized” words.

Table 03 – Transcription results in degraded areas.		
	FineReader	New Method
Correct	1949	2752
Incorrect	895	646
Not-recognized	738	184
Accuracy	54.5%	78.2%

<p>My dear Colleague,</p> <p>Your good letter found me in bed and I could not answer it before this time when I begin to look over my correspondence and I only can dictate so weak I am. I do it in English as the Secretary does not write my other language.</p>
<p><b>(a) Original text binarized</b></p>
<p>My dear Colleague,</p> <p>Your good letter found me in bed and I could not answer it before this time when I begin to look over my correspondence and I only can dictate so weak I am. I do it in English as the Secretary does not write my other language.</p>
<p><b>(b) Transcription of (a) using ABBYY FineReader 10</b></p>
<p>my dear Colleague,</p> <p>Your good letter found me in bed and I could not answer it before this time when I begin to look over my correspondence and I only can dictate so weak I am. I do it in English as the Secretary does not write my other language.</p>
<p><b>(c) Text after processing with the proposed scheme.</b></p>
<p>my dear Colleague,</p> <p>Your good letter found in bed and I could not answer it before this time when I begin to look over my correspondence and I only can dictate so weak I am. I do it in English as the Secretary does not write my other language.</p>
<p><b>(d) Transcription of (c) using ABBYY FineReader 10</b></p>
<p><b>Figure 8.</b> Transcription of the document shown in Figure 4 without and with the scheme presented here.</p>

## 6. Conclusions

This paper presents a scheme for improving the correct transcription rate of historical documents damaged by physical noises such as thorn-off regions and holes originated by punches (i.e. filing and staples) and worms, etc. The scheme automatically detects the damaged areas and when the text is segmented they are associated with blocks that are “replaced” with characters of the same group of features that increase the probability of being the original “damaged” character. Such replacement yields a class of possible lexemes that stand as “candidates” for the original word. Dictionary look-up decides which word is to be chosen as the most likely transcription.

The processing strategy presented here was tested in a batch of one hundred typewritten historical documents totaling over 12,000 words, of which 3,500 were “damaged”. The results obtained with ABBYY FineReader 10 Professional Editor and a gain in the correct transcription rate of about 25% was observed.

## References

- [1] R.D. Lins, G. F. P. Silva. Assessing and Improving the Quality of Document Images Acquired with Portable Digital Cameras. In: ICDAR 2007, Curitiba. IEEE Press, 2007. v. 2. p. 569-573.
- [2] R.D. Lins, G. F. P. Silva, G. Torreao. Content Recognition and Indexing in the Livememory Platform. LNCS, v. 6020, p. 224-230, 2010.
- [3] R. D. Lins. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. In: ICIAR 2009. Springer Verlag, 2009. v. 5627. p. 844-854.
- [4] R. D. Lins, *et al.* An Environment for Processing Images of Historical Documents. Microprocessing & Microprogramming, pp. 111-121, North-Holland, 1994.
- [5] R. D. Lins, G. F. P. Silva, S. Banergee, A. Kuchibhotla, M. Thielo. Automatically Detecting and Classifying Noises in Document Images. SAC 2010, ACM Press, 2010. v. 1. p. 33-39.
- [6] R. Farrahi, R. Moghaddam and M. Cheriet. Application of multi-level classifiers and clustering for automatic word spotting in historical document images. ICDAR 2009, IEEE Press, 2009. p. 511-515.
- [7] S. Pletschacher, J. Hu and A. Antonacopoulos. A New Framework for Recognition of Heavily Degraded Characters in Historical Typewritten Documents Based on Semi-Supervised Clustering. ICDAR 2009, IEEE Press, 2009. p. 506-510.
- [8] V. Kluzner, A. Tzadok, Y. Shimony, E. Walach, and A. Antonacopoulos. Word-Based Adaptive OCR for Historical Books. ICDAR 2009, IEEE Press, 2009. p. 501-505.
- [9] G. F. P. e Silva; R.D. Lins, S. Banergee, A. Kuchibhotla, M. Thielo. A Neural Classifier to Filter-out Back-to-Front Interference in Paper Documents, ICPR 2010, Istanbul. 2010.
- [10] D. M. Oliveira, R. D. Lins, G. Torreao, J. Fan, M. Thielo. A New Algorithm for Segmenting Warped Text-lines in Document Images, ACM-SAC 2011, ACM Press, 2011.
- [11] C.Y. Suen, J. Guo, Z.C Li, Analysis and Recognition Of Alphanumeric Handprints by parts, IEEE Transactions on Systems, Man, and Cybernetics, N. 24, p. 614-631, 1994.
- [12] Z.C. Li, C.Y. Suen, J. Guo, A Regional Decomposition Method for Recognizing Handprinted Characters, IEEE Transactions on Systems, Man, and Cybernetics, N. 25, p. 998-1010, 1995.
- [13] P. V. W. Radtke, L.S. Oliveira, R. Sabourin, T. Wong, Intelligent Zoning Design Using Multi-Objective Evolutionary Algorithms, ICDAR2003, p.824-828, 2003.
- [14] C. O. A. Freitas, L.S. Oliveira, S.B.K. Aires, F. Bortolozzi, Zoning and metaclasses for character recognition. ACM-SAC 2007. P. 632-636, 2007.
- [15] T. Kohonen, Self-Organizing Maps, Springer Series in Information Sciences, Springer, second edition, vol. 30, 1997.
- [16] E.V. Samsonova, J.N. Kok, A.P. IJzerman, TreeSOM: cluster analysis in the self-organizing map, Neural Networks, N. 19, p. 935-949, 2006.
- [17] NIST Scientific and Tech. Databases <http://www.nist.gov/data/>.
- [18] Fundação Joaquim Nabuco Fundaj, <http://www.fundaj.gov.br/>.
- [19] ABBYY FineReader 10 Professional Editor, <http://finereader.abbyy.com/>.
- [20] C. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art techniques", Pattern Recognition, 36(10):2271-2285, 2003.
- [21] C. Liu, Y. Liu, and R. Dai, "Preprocessing and statistical/structural feature extraction for handwritten numeral recognition", Progress of Handwriting Recognition, A.C. Downton and S. Impedovo eds., World Scientific, 1997.
- [22] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, "Automatic recognition of handwritten numerical strings: A recognition and verification strategy", IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(11):1438-1454, 2002.
- [23] M. Hu, "Visual pattern recognition by moment invariants", IEEE Transactions on Information Theory, 8(2):179-187, 1962.

# Thanatos

## Automatically Retrieving Information from Death Certificates in Brazil

Alessandra B. S. Almeida  
PPGEE - U.F.PE.  
Recife-Pernambuco-Brazil  
+55 81 9422-4537  
alessandrabsa@gmail.com

Rafael Dueire Lins  
U.F.PE.  
Recife-Pernambuco-Brazil  
+55 81 8896-0698  
rdl.ufpe@gmail.com

Gabriel de F. Pereira e Silva  
PPGEE - U.F.PE.  
Recife-Pernambuco-Brazil  
+55 81 8803-8715  
gfps.cin@gmail.com

### ABSTRACT

Death certificates provide important data such as *causa mortis*, age of death, birth and death places, parental information, etc. Such information may be used to analyze not only what caused the death of the person, but also a large number of demographic information such as internal migration, the relation of death cause with marital status, sex, profession, etc. Thanatos is a platform designed to extract information from the Death Certificate Records in Pernambuco (Brazil), a collection of “books” kept by the local authorities from the 16<sup>th</sup> century onwards. The current phase of the Thanatos project focus on the books from the 19<sup>th</sup> century.

### Categories and Subject Descriptors

I.4.9 [Image Processing and Computer Vision]: Applications.

### General Terms

Algorithms, Document image analysis.

### Keywords

Document processing, death certificates, image processing, historical documents.

## 1. INTRODUCTION

The Mormon Church teaches that their members are responsible to be baptized for dead loved ancestors. If a person dies having never been baptized in this life, a Mormon relative can be baptized in his place. Then the dead person may have a chance after death to believe the gospel, repent, and be saved. Joseph Smith, the founder of Mormonism, taught that seeking the dead in this manner is a Mormon’s greatest responsibility [1]. Such duty motivates the Mormon Church to have genealogical records all over the world. In the state of Pernambuco (Brazil), the Mormon Church celebrated an agreement with the Judiciary Power of the State (Tribunal de Justiça de Pernambuco - TJPE) to digitize all

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HIP’2011, Sep. 16-17, 2011, Beijing, China.  
Copyright 2011 ACM 1-58113-000-0/00/0010...\$10.00.

“books” of Civil Records kept by the local authorities. The complete Mormon-TJPE file encompasses over one million records and includes birth, wedding, and death certificates from all over the state of Pernambuco. The oldest records are from the 16<sup>th</sup> century just after Trento Council (1543 to 1563), when the Catholic Church that it was mandatory for people to have birth, christening, wedding, and death certificates. Figure 1 presents the platform used in the digitization of such records. It is a camera-based platform with controlled illumination. The height of the camera is adjustable to allow for larger volumes. The images were made available in grey-scale (8-bits) with an (approximate) resolution of 200 dpi. They were obtained between 1998 and 2000.



**Figure 1** – Digitalization platform used by the Mormon Church to acquire the images of the death certificates in Pernambuco (Brazil).

Figure 2 presents an example of a document from the Mormon-TJPE file. It is a civil wedding certificate dated from 1948 that took place at the city of Recife.

The information contained in Death Certificates goes far beyond genealogical data. It may be used to analyze not only the *causa mortis* of an individual, what caused the death of the person, but also a large number of demographic information such as inter and extra-regional migration, the relation of death cause with the person’s age, marital status, sex, profession, the way diseases were transmitted, the quality health assistance, sanitary

conditions, infant mortality, etc. The correlation of such information may provide an invaluable testimony of a society in many different aspects.

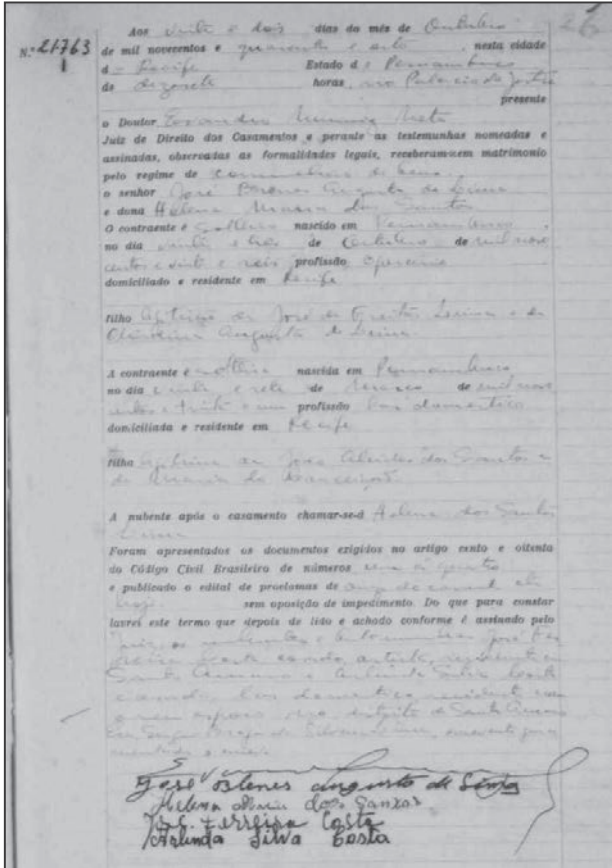


Figure 2 – Wedding certificate from the Mormon-TJPE file.

In Greek mythology, *Thanatos* (in Greek, *Θάνατος*—"Death") was the personification of death [2]. This paper presents the *Thanatos* platform, designed to extract data from the Death Certificate Records in Brazil. The current phase of the *Thanatos* project focuses on the books from the 19<sup>th</sup> century. The platform pre-processes the image to perform image binarization, border removal, skew correction, and content extraction.

## 2. The *Thanatos* Platform

The last two decades has witnessed an important growth in all fields of document engineering. Several new techniques, algorithms, tools, and platforms have been developed for document acquisition, processing, enhancement, and content extraction. New challenges appear every day in this area as the digitalization pace moves faster.

The analysis of the documents in the Mormon-TJPE bequest shows that several of them present several kinds of noise, which according to the classification in reference [3], are either physical (back-to-front interference, paper aging, faded ink, stains, folding marks, torn-off regions, etc.) or digitalization (borders, skew, salt-and-pepper, etc.). In the specific case of the death certificates from the file the kinds of noises found were similar, as one may

observe in the image shown in Figure 3, which presents an example of a two-page image with four death certificates of such a book from the 19<sup>th</sup> century. The book of records had little change over the centuries. The predominant ones from late 19<sup>th</sup> and the 20<sup>th</sup> centuries up to the late 1990s, when notaries were informatized, were pre-printed with fields to be filled in (that is also the case of the wedding certificate in Figure 2).

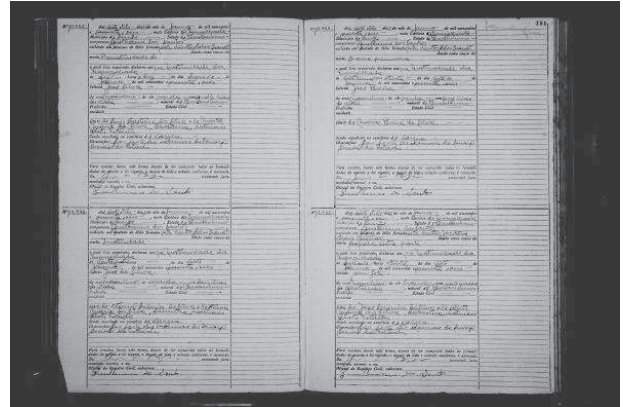


Figure 3 – Two-page image of a “book” showing four Death Certificates acquired by the platform of Figure 1

Although hard binding warp was observed in many of the images of documents the authors opted not to address such problems for two reasons:

- The margin between the binding and the start of textual information is wide enough not to warp significantly textual areas.
- In most of cases, only certificate number was affected.

## 2.1 Preprocessing

As already mentioned, the images from the Mormon-TJPE file were acquired with a digitalization platform such as the one presented in Figure 1. One may observe that there is a fixed basis in Black background plane where lays the document to be photographed. The camera is held at a column perpendicular to the basis in such a way as to move up and down as parallel as possible to the background plan. Before the start of the image acquisition process there is a camera calibration phase which performs diaphragm aperture control to check the intensity of the lightning that reaches the sensor, gray scale checking, focus adjustment, heating of illumination lamps, etc.

In spite of all the care taken in image acquisition and initial calibration, there are variations from one set of images to another that yield significant distortions that complicate image processing, segmentation, information extraction and automatic recognition. As one may observe in Figure 3, the image shows the presence of black borders (that correspond to the mechanical support where the book rested), a small skew angle (generally of less than 3 degrees), and some salt-and-pepper noise possibly due to dust and small stains in the paper.

The closer analysis of the documents of the death certificates showed that many of the documents exhibit a weak back-to-front interference (also known as bleeding [4] or show-through [5]). The new version [22] of the HistDoc platform [6] was used to



remove such noise by applying the algorithm described in reference [7], which assesses the intensity of the interference for tuning the global threshold algorithm. After the removal of the back-to-front interference the image is binarized. Besides the problems already listed neither the illumination of the document is completely even nor is the resolution uniform as there is a variation of the height of the camera. Both factors cause difficulties for document processing and information extraction. A general scheme of the *Thanatos* platform and its integration with the HistDoc platform is shown in Figure 4.

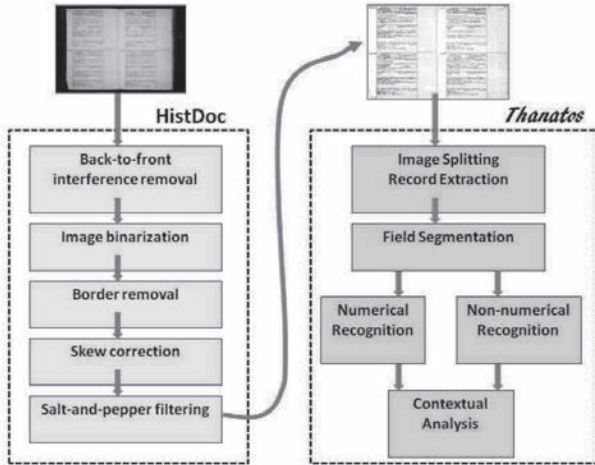


Figure 4 – Block diagram of the *Thanatos* platform and its integration with HistDoc.

The removal of the black surrounding border is performed by using the algorithm presented in reference [8]. Skew correction is performed by the algorithm described in reference [9], and finally salt-and-pepper filtering is applied. The new version of HistDoc works similarly to BigBatch [10] in either operator-driven mode, or stand-alone batch mode that also makes possible working in clusters and grids. Figure 5 shows the document in Figure 3 after being processed with the filters of HistDoc [6], [22] and BigBatch, followed by cropping.

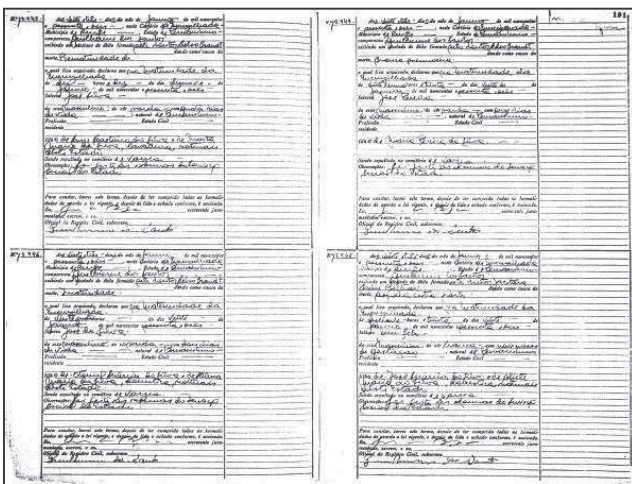


Figure 5 – Image after processed by HistDoc and BigBatch.

The next step in the document processing is performed by the core of the *Thanatos* platform that splits the image either in two (for the first page of the book) or in four (for the remaining pages) death certificates. This operation is performed by automatically seeking the central line of the image, which corresponds to the volume spine, yielding the left and right page images with two certificates each. Then, a search is performed horizontally to find the central line of the document. Due to the image acquisition process, such line does not correspond to the median line in the image. This task is performed by getting a central area from the page and performing a horizontal projection profile in it. The desired line is slightly thicker than the other blank lines. This method works satisfactorily and yields the upper and lower images, an example of which is shown in Figure 6.

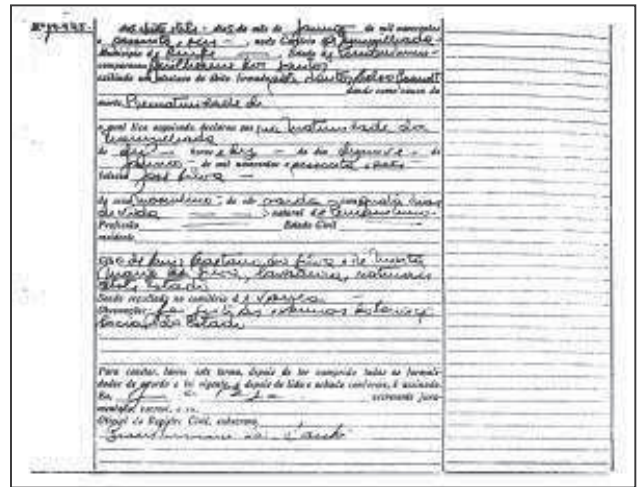


Figure 6 – Death Register after segmentation by the *Thanatos* platform.

The direct segmentation of the certificate image such as the one in Figure 6, has been unsuccessful in automatic content extraction due to book binding warp that increases the complexity of such a task. Image uniformization has to be performed because the four certificates on an image (top/bottom-left, top/bottom-right) have different “shapes”. Image pre-processing is performed to automatically search for “field” delimiters, as shown in Figure 7.

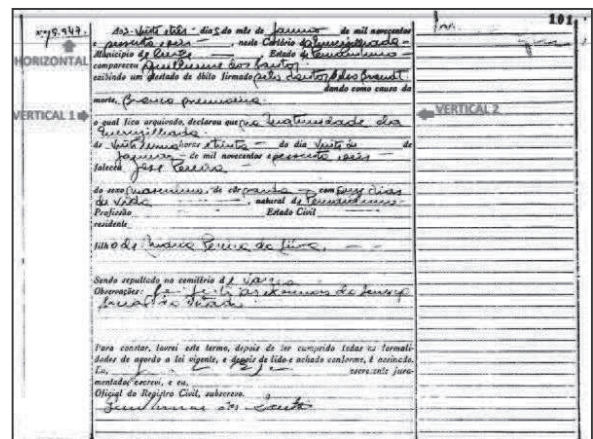


Figure 7 – Finding reference points for content segmentation.

Several problems were found in the development of the standardization process. As one may observe in Figure 7, the vertical lines have several missing points due to bad quality typographical printing and also during image binarization. Book binding warp in some cases also caused skew in the horizontal lines and this made boundary line detection difficult. Besides those factors there is an additional problem: certificate books also varied in size. Some of them were smaller than the 1,900 by 1,470 pixels as the horizontal and vertical dimensions adopted as the minimum for a register. In such cases, white pixel padding was performed at the bottom and right-hand side margins of the images. After the register standardization process all death registers had the same dimensions (1,900 x 1,470 pixels) and the same relative positions between different images for the same information fields, allowing segmentation to take place.

## 2.2 Segmentation

The aim of this phase is to automatically extract the information from each field in the death certificate form. Fields were hand-written and often they overlapped the fields. The ten first fields of the death certificate are described below. (The underlined text corresponds to the information the notary wrote in the field of the form):

- **Nº** (Register number) – placed at the top of the left margin of the register. It conveys numerical information only. Example: Nº 19.945.
- **Data** (Date) – the date is written in words and the information is filled in three fields for *day*, *month*, and *year* in this sequence. Example: Aos vinte e três dias do mês de janeiro de mil novecentos e sessenta e seis (At the twenty three days of the month of January of one thousand nine hundred and sixty six).
- **Nome do cartório** (Notary name) – this field holds the name of the place where the notary office was found. Example: neste cartório da Encruzilhada (at this notary office at Encruzilhada).
- **Município do Cartório** (City of the notary office) – Example: município de Recife (at the city of Recife).
- **Estado do Cartório** (State of the notary office) - Example: Estado de Pernambuco (State of Pernambuco).
- **Nome do Declarante** (Name of declarer) – Name of who attended the office to inform the death. Example: compareceu Guilherme dos Santos (attended Guilherme dos Santos).
- **Nome do Médico** (Name of the Medical Doctor) – Name of the M.D. who checked the death. Example: exibindo um atestado de óbito firmado pelo doutor José Ricardo (showing a death declaration signed by doctor José Ricardo).
- **Causa mortis** – Specifies the reason of the death in the declaration from the M.D. Example: dando como causa da morte edema pulmonar, o qual fica arquivado (that states as causa mortis lung edema, which is archived).

For better understanding of the correspondence of the information above the fields, they were translated on a word-to-word basis between (Brazilian) Portuguese and English.

The automatic information extraction was performed by masks for each of the fields of the death certificate. Each field mask has variable size and is placed onto a matrix of masks (MM) as a two-dimensional array, with the following lay-out:

$$MM = \begin{bmatrix} X0_1 & Y0_1 & W_1 & H_1 \\ X0_2 & Y0_2 & W_2 & H_2 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X0_n & Y0_n & W_n & H_n \end{bmatrix}$$

In which ‘n’ is the n<sup>th</sup> element from MM matrix ‘X0’ is the initial coordinate of the X axis, ‘Y0’ is the initial coordinate of the Y axis, ‘W’ is the width of the n<sup>th</sup> field and ‘H’ is the height of the n<sup>th</sup> field.

The creation of the MM matrix allows cropping of each image field. As a result of the application of the matrix of masks to a (standardized) death register one obtains the image for the fields of the death certificate as shown in Table 1 (the images correspond to the item by item examples provided above).

Mask nº	Field Name	Image
01	Registry Number	
02	Date	
03	Month	
04	Year	
05	Name/Place of notary	
06	City	
07	State	
08	Name of declarer	
09	Name of the M.D. who certified the death.	
10	Cause of death	

**Table 1** - Segmented fields of death records.

The direct recognition of the handwritten fields above has shown to provide very bad results. Several problems were identified in the images that had deleterious effects in the accuracy of character recognition. They were: the horizontal lines of the gaps to be filled in, the vertical lines that set the margins of the record, remaining stains and physical noise within the image, and too

narrow characters. To solve such problems, each of the records has to go through a new preprocessing stage. These problems can be observed in the column “image” in Table 1.

### 2.3 Recognition Preprocessing

The first preprocessing step for the segmented images is to remove the printed horizontal lines. To solve this problem, an algorithm was developed to sweep the field image from left to right and from top to bottom, scanning it in blocks of 8x8 pixels to identify regions which resemble a horizontal line, even if it has “holes” in it as found in several images. If the first and last horizontal lines of the block are all white the block under the mask is replaced by white pixels. If most pixels under the mask are white and the first column has at least one black pixel the area must remain unchanged to avoid erasing areas that are the starting points of character strokes, as shown Figure 8.

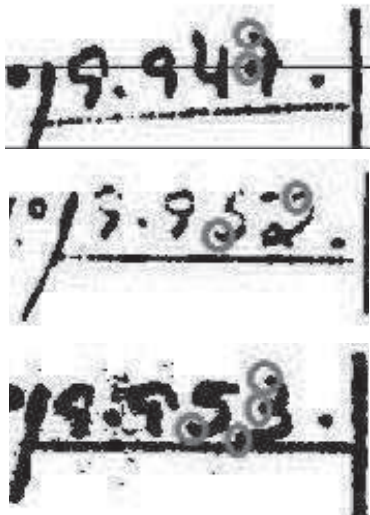


Figure 8 - Areas of the image that may be erased if the first column of the block is not checked.

Figure 9 illustrates the algorithm for removing horizontal lines, displaying the 8x8 block of pixels, comparing their size and the image of a block being processed and the example of a block that will be erased with this technique.

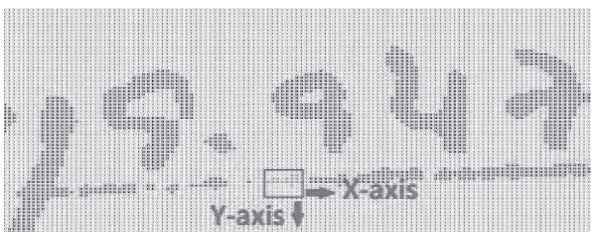


Figure 9 – Example image under the line removal algorithm.

The result of the same image of the Register Number field (N° 19.947) before and after being processed by the algorithm proposed here can be seen in Figure 10.



Figure 10 – Register number before and after horizontal line removal.

The second step of the recognition preprocessing is noise reduction. Such noise is either physical (stains, mould, paper aging, dust, etc) or introduced during digitalization (dust on the camera lens or on the book during image acquisition), or even introduced during binarization. Although the whole image had the salt-and-pepper noise filtered out in the last phase of preprocessing performed by HistDoc, field-specific filtering enhances the quality of the image and increases the correct recognition rate. Figure 11 presents an example of such a field image in the original gray-scale document and after binarization.

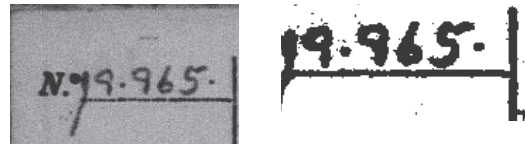


Figure 11 - Noise present in the original and binarized images.

As one may observe in the image of Figure 11 the size of the noise “grains” is much coarser than the salt-and-pepper one. To remove such noise the whole image is scanned from left to right and from top to bottom, with a 4x4 pixel mask. The pixels under the mask are painted white if either the first and last lines are white, if over 50% of the number of pixels under the mask is white, or if the first column of the mask has at least one black pixel. After the application of coarse-grain noise removal algorithm, the image is re-processed with the algorithm to remove horizontal lines to improve the results.

Before Algorithm	After Algorithm

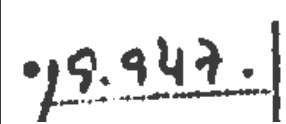
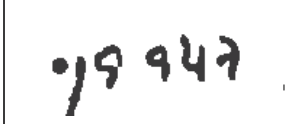
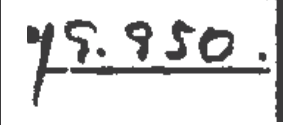
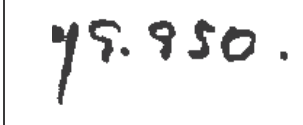

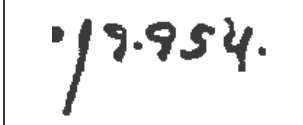
Table 2 – Examples of Registry number before and after applying the noise reduction algorithm.

The third and last step of the recognition preprocessing aims to remove the vertical line in the images of the record number. The algorithm developed for this purpose is based on the automatic evaluation of the distribution of black pixels along the projection profile of the whole figure. It was possible to

observe that vertical lines have a greater number of black pixels per column than the areas that correspond to characters. In this context, we use, for each image, the maximum value of black pixels per column ( $maxColumnSum$ ) and the mean value ( $meanColumnSum$ ), per column, for the entire image in both cases to generate a specific threshold deletion ( $thresholdDel$ ) for each image processed. The variable threshold is important to differentiate the clearer images (or with fine lines), from the ones with more dilated characters. Additionally, the columns immediately anterior and posterior to the columns with the number of pixels greater than the threshold value are also deleted. The equation below represents the calculation of the deletion threshold value.

$$thresholdDel = maxColumnSum - 2 * meanColumnSum$$

The results for this algorithm are shown in the images of Table 2, where one may observe images before and after preprocessing process (vertical and horizontal line removal and noise reduction), the last step before the information recognition step.

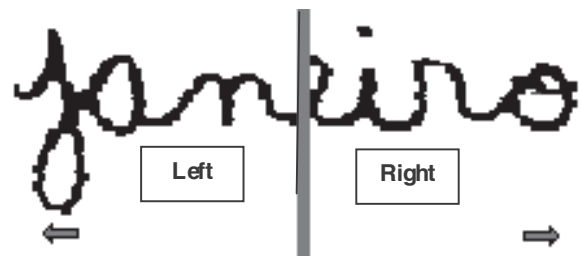
Original Image	Image after the recognition preprocessing stage
	
	
	

**Table 3** – Examples of Registry number before and after applying the entire Recognition Preprocessing Stage.

## 2.4 Recognition and Classification

The first strategy used for information recognition was to transcribe the fields using the commercial OCR tool ABBYY FineReader 10 Professional Editor [11]. The results obtained were zero correct recognition for all fields, including even the numerical ones. Such disappointing results forced the authors to develop a recognition tool for the *Thanatos* platform.

Recognizing handwritten symbols is much harder when they are connected. The alternatives one finds to increase the global recognition rate is either to split symbols or to try to recognize the whole word (or the connected part of it). The classifier developed for the *Thanatos* platform assumes that numerical symbols form a contiguous block and no further segmentation is needed. In the case of non-numerical fields, there is the need of applying the preprocessing techniques described above, in which the image is split into two parts as presented in Figure 12.



**Figure 12** – Non-numerical field recognition

The majority of studies using pattern recognition have as central theme the selection of a set of features capable of representing and discriminating between the different shapes to be classified. To find such set of features is far from being an easy task. The technical literature presents several techniques for such a purpose [12][13][14][15] which when applied to handwritten text recognition may be summarized into three different classes:

- Primitives based on global transforms and expansion series, such as Fourier, Walsh, Harr, etc provide invariants to some global transformations such as rotation and translation. These techniques require greater processing power and are time consuming.
- Primitives based on the statistical distribution of the points. They include moments, n-tuples, crossing and distances. They allow for shape distortion and take into account hand written style variation, in some cases. They have low implementation complexity.
- Geometrical and perceptual primitives. These are the primitives more widely used to represent global and local properties of characters. In this class one finds: ascending and descending strokes, loops, line-segment intersection, ending points, angular properties, relations between strokes, etc. These primitives have a high tolerance to distortions, style variation, translation and rotation.

The current study follows the approach in reference [16] and makes use of a set of geometrical and perceptual features extracted from “zoning” the image. This technique counts the number of loops, concavities, horizontal and vertical strokes, etc. “Zoning” may be seen as splitting a complex pattern in several simpler ones. In the case of degraded texts, the concern of this paper, this becomes an important discrimination basis amongst classes, as the “real” information is limited only to some classes. Some researchers propose only the “empirical” zoning [17][18][19], in which each character is represented by a rectangle  $Z$ , that may assume several different formats, such as the ones presented in Figure 13. Other researchers propose methods of automatic zoning [20].

This work adopted a very pragmatic approach to recognition and used the different techniques to analyze which one provided the best result for each information field.

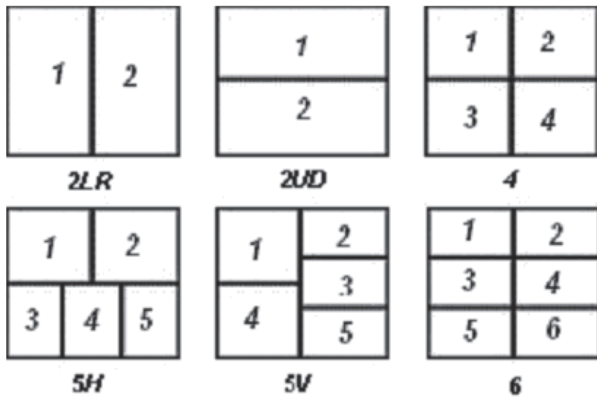


Figure 13 – Automatic character zoning

### A. Numerical Field Classifier

The *Thanatos* numerical field classifier made use of 10 classes of numerical digits, each of them with 2,500 images per class, 2,000 of them were obtained from the NIST database [21], and the other 500 were extracted from the documents from the The Church-TJPE files. The recognition used three subsets for testing, training and validation that correspond to: 50%, 25% and 25%, of the total data respectively. Two classifiers were tested: a MLP (Multi-layer perceptron) and a RBF (Radial Basis Function). The latter presented a slightly better performance, as may be observed in the results presented in Table 4 for isolated character recognition and in Table 5 for the complete recognition of the field.

Digit	MLP	RBF
0	100%	100%
1	100%	100%
2	98%	99%
3	100%	100%
4	95%	98%
5	96%	96%
6	98%	98%
7	100%	100%
8	100%	100%
9	96%	95%

Table 4 – Correct recognition rates for digits in field Register Number.

Field	MLP	RBF
Numerical Register	92%	95%
Day	93%	96%

Table 5 – Correct recognition rates for fields Register Number and Day of a month.

The diagram for the *Thanatos* platform shown in Figure 4 includes a block for context analysis. The Register Number follows a sequential order and this information is taken into account to increase the correct recognition rate of the information in this field. The “Day” field is either the same as the previous Register or is (28/29/30/31 cyclically) incremented. The use of such information allowed a 100% correct data recognition for these fields, for both classifiers.

### B. Non-Numerical Field Classifier

Character recognition for non-numerical fields makes use of two strategies that take into account the variance of the information in the field. The first approach is used for fields where the maximum variance is four words in which the extraction of geometrical and perceptive information is used. This is the case for the field “estado civil” (marital status) that presents only four options per gender: “solteiro” (single), “casado” (married), “viúvo” (widower), and “separado” (divorced); the latter one found only in the most recent registers. In the case of other fields that have a wider range of possibilities, such as “Estado” (State), “Cidade” (city), numbers in writing, name of the month, etc. the use of geometrical and perceptive features yielded unsatisfactory results and was replaced by the zoning mechanism.

The results obtained for 300 death certificates are presented in Table 6.

Field	Correct recognition rate
Name of Notary	98%
City of the Notary	71%
State of the Notary	98%
Place of death	31%
Numbers in writing: (Time of obit, date of death, date of birth)	69%
Color of skin	100%
Marital status	100%

Table 6 – Recognition rate for non-numerical fields

It is important to stress that the automatic contextual analysis of the *Thanatos* platform was also used here to Grant such good results. For instance, in the case of the Name of Notary, each office has at maximum four officers in such position who remain active for long periods of time (in Brazil that service is a concession of the State and that is a life-long position). One may observe that in the case of the field “Place of death” the recognition rate was low. The contextual analysis was not implemented. The addition of a dictionary of all the cities of the State of Pernambuco, together with the information of jurisdiction (the obits within a region must be registered in the closest possible notary office) are yet to be implemented. The other non-numerical fields were trained using the information in the database of words used in bank cheques in Brazil. Similarly, one may implement a dictionary of medical doctors that worked at a certain region for better data recognition.

### 3. Conclusions and lines for further work

Death certificates provide important anthropological, sociological, and medical information of populations.

This paper presents the *Thanatos* platform, a platform designed to extract information from death certificates from the Mormon-TJPE files from Pernambuco, Brazil. The platform introduced several new strategies for content extraction and recognition that was able to retrieve information at a very high accuracy rate. It is important to stress that even if the correct recognition rate is lower than 100% the information may be useful for demographic studies.

The current version of the platform focused on the books from late 19<sup>th</sup> century to close to the year 2000, when notaries used pre-printed books to fill information in the fields. Notary books from the earlier periods have a much higher degree of difficulty for processing and information extraction as there are no patterns to guide the register and field extraction. Addressing such documents is left for further work.

### 4. Acknowledgements

The authors are grateful to the The Church of Jesus Christ of Latter-day Saints (Family Search International) for the initiative of digitizing the death certificate records of Pernambuco (Brazil) and to Tribunal de Justiça de Pernambuco (TJPE) to allow the use of such data for research purposes.

Research presented here is partly sponsored by CNPq- Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico, Brazilian Government.

### 5. REFERENCES

- [1] Marvelous Work and a Wonder, p. 189.
- [2] <http://en.wikipedia.org/wiki/Thanatos>, visited on 18/06/2011.
- [3] R. D. Lins. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. ICIAR 2009, LNCS v.5627, p.844 - 854 Springer Verlag, 2009.
- [4] R. Kasturi, L. O'Gorman and V. Govindaraju, "Document image analysis: A primer", Sadhana, (27):3-22, 2002.
- [5] G.Sharma, "Show-through cancellation in scans of duplex printed documents", IEEE Trans. Image Processing, v10(5):736-754, 2001.
- [6] G. F. P e Silva, R. D. Lins, J. M. M. da Silva. HistDoc - A Toolbox for Processing Images of Historical Documents, ICIAR 2010, LNCS v.6112, p.1 - 11. Springer Verlag, 2010.
- [7] G. F. P e Silva, R. D. Lins, J. M. M. da Silva, S. Banergee, A. Kuchibhotla, M. Thielo. Enhancing the Filtering-Out of the Back-to-Front Interference in Color Documents with a Neural Classifier. ICPR 2010. pp: 2415-2419. IEEE Press.
- [8] A. de A. Formiga and R. D. Lins. Efficient Removal of Noisy Borders of Monochromatic Documents. International Conference on Image Analysis and Recognition, 2009, LNCS v.5627. p.158 - 167, Springer Verlag, 2009.
- [9] B. T. Ávila and R. D. Lins. A Fast Orientation and Skew Detection Algorithm for Monochromatic Document Images. ACM International Conference on Document Engineering, 2005. ACM Press, 2005. p.118 - 126
- [10] G. G. de Mattos, A. de A. Formiga, R. D. Lins and F. M. J. Martins. BigBatch: a document processing platform for clusters and grids. Proceedings of ACM-SAC 2008. v.I. p.434 - 441, ACM Press, 2008.
- [11] ABBYY FineReader 10 Professional Editor, <http://finereader.abbyy.com/>.
- [12] C. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art techniques", Pattern Recognition, 36(10):2271-2285, 2003.
- [13] C. Liu, Y. Liu, and R. Dai, "Preprocessing and statistical/structural feature extraction for handwritten numeral recognition", Progress of Handwriting Recognition, A.C. Downton and S. Impedovo eds., World Scientific, 1997.
- [14] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, "Automatic recognition of handwritten numerical strings: A recognition and verification strategy", IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(11):1438-1454, 2002.
- [15] M. Hu, "Visual pattern recognition by moment invariants", IEEE Transactions on Information Theory, 8(2):179-187, 1962.
- [16] G. F. P. e Silva and R. D. Lins. An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. ICDAR 2011, Beijing, September, IEEE Press, 2011.
- [17] C.Y. Suen, J. Guo, Z.C Li, Analysis and Recognition of Alphanumeric Handprints by parts, IEEE Transactions on Systems, Man, and Cybernetics, N. 24, p. 614-631, 1994.
- [18] Z.C. Li, C.Y. Suen, J. Guo, A Regional Decomposition Method for Recognizing Handprinted Characters, IEEE Transactions on Systems, Man, and Cybernetics, N. 25, p. 998-1010, 1995.
- [19] C. O. A. Freitas, L.S. Oliveira, S.B.K. Aires, F. Bortolozzi, Zoning and metaclasses for character recognition. ACM-SAC 2007. P. 632-636, 2007.
- [20] P. V. W. Radtke, L.S. Oliveira, R. Sabourin, T. Wong, Intelligent Zoning Design Using Multi-Objective Evolutionary Algorithms, ICDAR 2003, p.824-828, 2003.
- [21] NIST Scientific and Tech. Databases <http://www.nist.gov/data/> .
- [22] R. D. Lins, G.F.P e Silva, A. de A. Formiga, HistDoc v. 2.0 - Enhancing a Platform to Process Historical Documents. HIP 2011, ACM Press, 2011.

---

# Generating Training Sets for the Automatic Recognition of Handwritten Documents

---

Gabriel Pereira e Silva and Rafael Dueire Lins

Additional information is available at the end of the chapter

<http://dx.doi.org/10.5772//52074>

---

## 1. Introduction

Handwritten character recognition is a task of high complexity even for humans, sometimes. People have different writing “style”, which may vary according to psychological state, the kind of document written, and even physical elements such as the texture of the paper and kind of pencil or pen used. Despite such wide range of variation possibilities, some elements tend to remain unchanged in a way that other people, in general, can recognize one’s writing and even identify the authorship of a document. Very seldom one is unable to identify one’s own writing. Very seldom someone is unable to identify his own writing.

The basis for pattern recognition rests on two corner stones. The first one is to find the minimal set of features that presents all maximum diversity within the universe of study. The second one is to find a suitable training set that also covers all possible data to be classified. Due to the variation of writing styles between people, one should not expect that a general classifier yields good recognition performance in a general context. Thus, one tends to either have general classifiers for very specific restricted vocabularies (such as digits), or to have personalized recognizers for general contexts. The scope of the present work is the latter. In such context it is a burden and very difficult to generate a good training set to allow the classifier to reach a reasonable recognition rate.

This paper proposes a new approach for the automatic generation of the training set for the handwritten recognizer of a given person. The first step for that is to select a set of documents representative of the author’s style. In the Internet one may find several public domain sites with font sets. In particular the site Fontspace [21] offers 282 different cursive font sets for download (e.g Brannboll Small, Jenna Sue, Signerica Fat, The Only Exception, Homemade Apple, Santos Dumont, etc). Figure 1 presents an example of some of them. The key idea presented here is “approximating” the author writing by a cursive typographical font, which is skeletonized and a “standard” training set is generated. Such strategy,

detailed as follows, was adopted with success with documents of the Nabuco bequest [12] and of the Thanatos Project [1].

<p> <i>                     ABCDEFGHIJKLMNOP                      QRSTUVWXYZ abcdefg                      hijklmnopqrstuvwxy                      0123456789                 </i> </p>	<p> <i>                     ABCDEFGHIJKLMNOP                      QRSTUVWXYZ abcdefg                      hijklmnopqrstuvwxy                      0123456789                 </i> </p>
<p>Santos Dumont by Billy Argel - 2007</p>	<p>Eutemia I by Bolt Cutter Design - 2008</p>
<p> <i>                     ABCDEFGHIJKLMNOP                      QRSTUVWXYZ abcdefg                      hijklmnopqrstuvwxy                      0123456789                 </i> </p>	<p> <i>                     ABCDEFGHIJKLMNOP                      QRSTUVWXYZ abcdefg                      hijklmnopqrstuvwxy                 </i> </p>
<p>Olho de Boi by Billy Argel - 2011</p>	<p>Wedding Nightmare by Billy Argel - 2011</p>
<p> <i>                     ABCDEFGHIJKLMNOP                      QRSTUVWXYZ abcdefg                      hijklmnopqrstuvwxy                      0123456789                      .,:;"'!@#%&amp;*('/\)                 </i> </p>	<p> <i>                     ABCDEFGHIJKLMNOP                      QRSTUVWXYZ abcdefg                      hijklmnopqrstuvwxy                      0123456789                      .,:;"'!@#%&amp;*('/\)                 </i> </p>
<p>Bernard by Philing - 1999</p>	<p>Discipuli Britannica by Peter Wiegel - 2009</p>
<p> <i>                     ABCDEFGHIJKLMNOP                      QRSTUVWXYZ abcdefg                      hijklmnopqrstuvwxy                      0123456789                      .,:;"'!@#%&amp;*('/\)                 </i> </p>	<p> <i>                     ABCDEFGHIJKLMNOP                      QRSTUVWXYZ abcdefg                      hijklmnopqrstuvwxy                      0123456789                 </i> </p>
<p>Gerards Gold by David Kerkhoff - 2010</p>	<p>Kathleenie by Robotic Attack Fonts - 1997</p>

Figure 1. Examples of cursive font sets extracted from Fontspace [1]



## 2. The proposed method

The choice of a representative training set together with a good set of features is fundamental for the success of automatic pattern recognition. These two factors are tightly linked to each other and in such a way as to grant good recognition results. To obtain a good training set for handwriting recognition of a given author during a period of time in which the writing features are stable (they changes with age, psychological and health factors, etc.) one has to group together the documents that have similar properties. A subset of them that is representative of the set of documents (in general the size of the training set is about 10% of the size of the whole data "universe") is chosen in such a way as to cover the whole diversity of the documents to be transcribed.

The development of the proposed method starts by using a set of cursive fonts. In the Internet one may find several public domain sites with font sets. In particular the site Fontspace [21] offers 282 different cursive font sets for download.

The central difficulty in generating the training set for handwritten documents is to have a "font set" of a specific author to extract the convenient features for patten matching.

The strategy adopted here is:

1. Select a number of documents that is representative of the author.
2. Process the documents to:
  - a. remove digitalization borders that may frame the image,
  - b. correct skew,
  - c. filter-out back-to-front interference (bleeding),
  - d. binarize the document.
3. Transcribe the document into text by a human reader.
4. The user should select a number of a cursive font set that bears "some resemblance" to the original author handwriting.
5. The text version of the document is typeset in each of the cursive font sets chosen in step 4 (above).
6. All the typeset versions of the document are converted into image.
7. The image of the original document is skeletonised and then dilated.
8. Segment the image in boxes around each letter (font cases) of the skeletonised and dilated version of the original image and the synthetically generated images.
9. Apply a deformation transform to make each font case in the synthetic images coincide with the font case of the skeletonised-dilated version of the original document.
10. Extract the features from each document and place in a vector.
11. Take the Hamming distance between the feature vectors of the synthetic and original images.
12. The font set used to generate the synthetic document which provides the smallest Hamming distance is the one to be used as the training set.

The structural features used for pattern recognition, mentioned in step 10 above, are:

- Geometric Moments [15] [9];

- Concavity Measurements [16];
- Shape Representation of Profile [14];
- Distance between barycentre points between two consecutive characters;
- Maximum and minimum heights of two consecutive characters.
- Maximum and minimum distance between concavities of two consecutive characters.

The image filtering operations listed in step 2 were performed using HistDoc v.2.0 environment [13] which offers a wide number of tools for historic document image processing including the several algorithms for the removal of back-to-front interference and binarization. The skeletonization and dilation processes in this work were performed using the filters available in ImageJ [20].

Step 9, performing image vectorization, is important to increase the likelihood between the synthetic and the original documents. Such operation is applied to each character in each synthetically generated image by deforming the bounding-box and the strokes until there is a perfect match between the synthetic and the original one. In this “deformation” process some statistical analysis is performed to infer data about inter character and inter word spacing, line and character skew, inter line separation, etc.

The feature vector of a document brings an account of the basic features of the author calligraphy. The Hamming distance between the feature vectors of the synthetic and original images, which is part of step 11, brings an account of their similarity, and is calculated using the formula:

$$H_w = \sum_{n=1}^{N \text{ features}} |f_{on} - f_{sn}|$$

where  $f_{on}$  and  $f_{sn}$  are the components of the feature vectors of the original and synthetic images, respectively. The choice of a vector of features such that one could extract “information” about the calligraphic pattern of the author shares some ideas with the work in reference [5]. The font set that provides the smallest Hamming distance to the original set is chosen to synthetically generate the whole training dictionary to the classifier.

In what follows the steps described above are detailed in two files of historical documents: the handwritten letters of Joaquim Nabuco that are about one century old and the hand filled information on the books of pre-printed forms of civil certificates from the state of Pernambuco-Brazil, from mid 20<sup>th</sup> century.

### 3. Results

The strategy presented above for developing the training set was tested in two sets of documents: letters from Nabuco bequest and death certificates from the Thanatos project [1].

#### 3.1. Transcribing Nabuco’s letters

The Nabuco Project [12] was an initiative of the second author of this paper. It started in 1991 with the aim to preserve the file of letters of Joaquim Nabuco, a Brazilian statesman,

writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil (b.1861-d.1910). The Nabuco file encompasses over 6,500 documents and about 30,000 pages of active and passive correspondence (including postcards, typed and handwritten letters), a bequest of historical documents of paramount importance to understand the formation of the political and social structure of the countries in the Americas and their relationship with other countries. The letters of Nabuco were catalogued and some of them summarized [2] [4], but the bequest was never fully transcribed. The Nabuco Project is acknowledged as being the pioneering initiative in Latin America to attempt to generate a digital library of historic documents. Figure 2 presents an example of letter in Nabuco bequest, written in a blank sheet of paper without lines, which presents a textured background due to paper aging, a horizontal folding mark in its central part, a light back-to-front interference (bleeding) as the letter was written on both sides of the sheet of paper. The image was acquired with an operator driven flatbed scanner, using 200 dpi resolution, in true color. There is no marginal noise (borders) framing the image and its skew is negligible.

To automatically generate the training set for recognizing the handwritten letters from Nabuco file a visual inspection was made to find letters that could represent the whole universe of letters. From the Nabuco file 50 letters were chosen and transcribed by historians, yielding 50 text files, totaling 3,584 words. Twenty-five letters (1,469 words) were used to develop the feature set used for training the classifier, and the remaining ones for ground-truth testing. All the selected documents were processed performing the steps listed in step 2 above, which encompasses marginal border removal, image de-skew, removal of back-to-front interference and binarization. An example of resulting document after filtering and binarization using the HistDoc v.2.0 environment [13] may be found in Figure 3. The image in Figure 2 is skeletonized and then dilated using the filters in ImageJ [20]; the resulting image is presented in Figure 4.

The synthetic image generation is performed by choosing a subset of the cursive fontsets available that resemble the writing of the original document. In the case of Nabuco, the subset selected encompassed 15 of the 282 cursive font sets available in Fontspace [21] (e.g Brannboll Small, Jenna Sue, Signerica Fat, The Only Exception, Homemade Apple, Santos Dumont, etc) that were closer to the author's writing style during the period of interest. The text of the original document was (human) transcribed into a text file, which was typeset using the choosen cursive fontsets. The text of document in Figure 2 typeset using the fonts in "Signerica Fat" type font is shown in Figure 5. The image shown in Figure 5 is now vectorized and "approximated" to the original skeletonized and dilated image by "deforming" each "letter case" and strokes until matching, as much as possible. The resulting image is presented in Figure 6.

The feature vector of each of the synthetic images "deformed" in such a way to the character case to match the original font case was extracted and the Hamming distance of each of them to the skelotonized-dilated original image was calculated. The image that exhibited the minimum distance was the "Signerica Fat" font set presented in Figure 6.

104,1

para as altas nomeações de  
me dispõe ou venha a dispôr.  
Muitas saudações a' Baroneza,  
Carlota, lembranças ao  
Burton e para si um abraço,  
apertado do seu  
d. J.  
Joaquim Nabuco

Figure 2. Letter from Nabuco bequest.

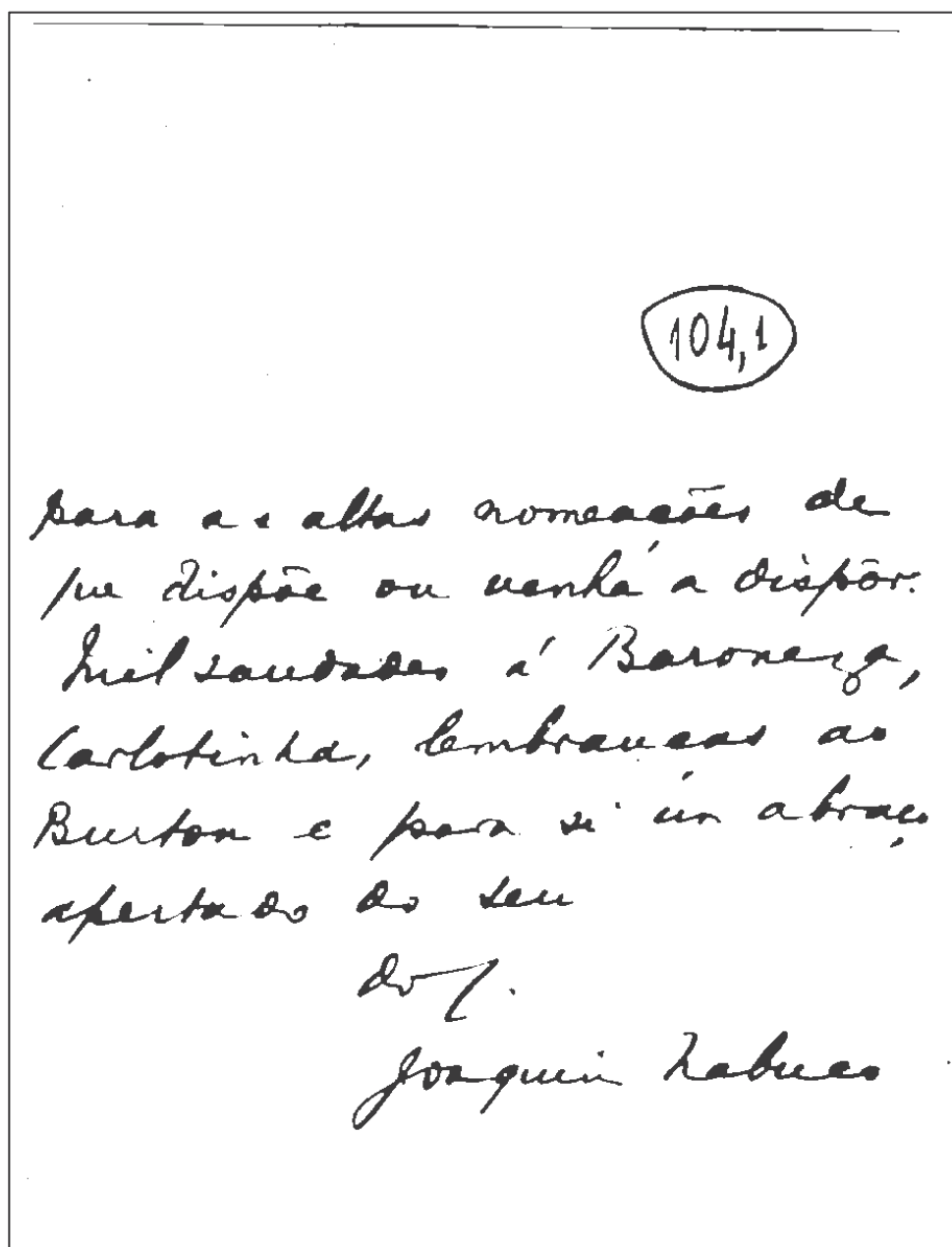


Figure 3. Document shown in Figure 2 after filtering and binarization.

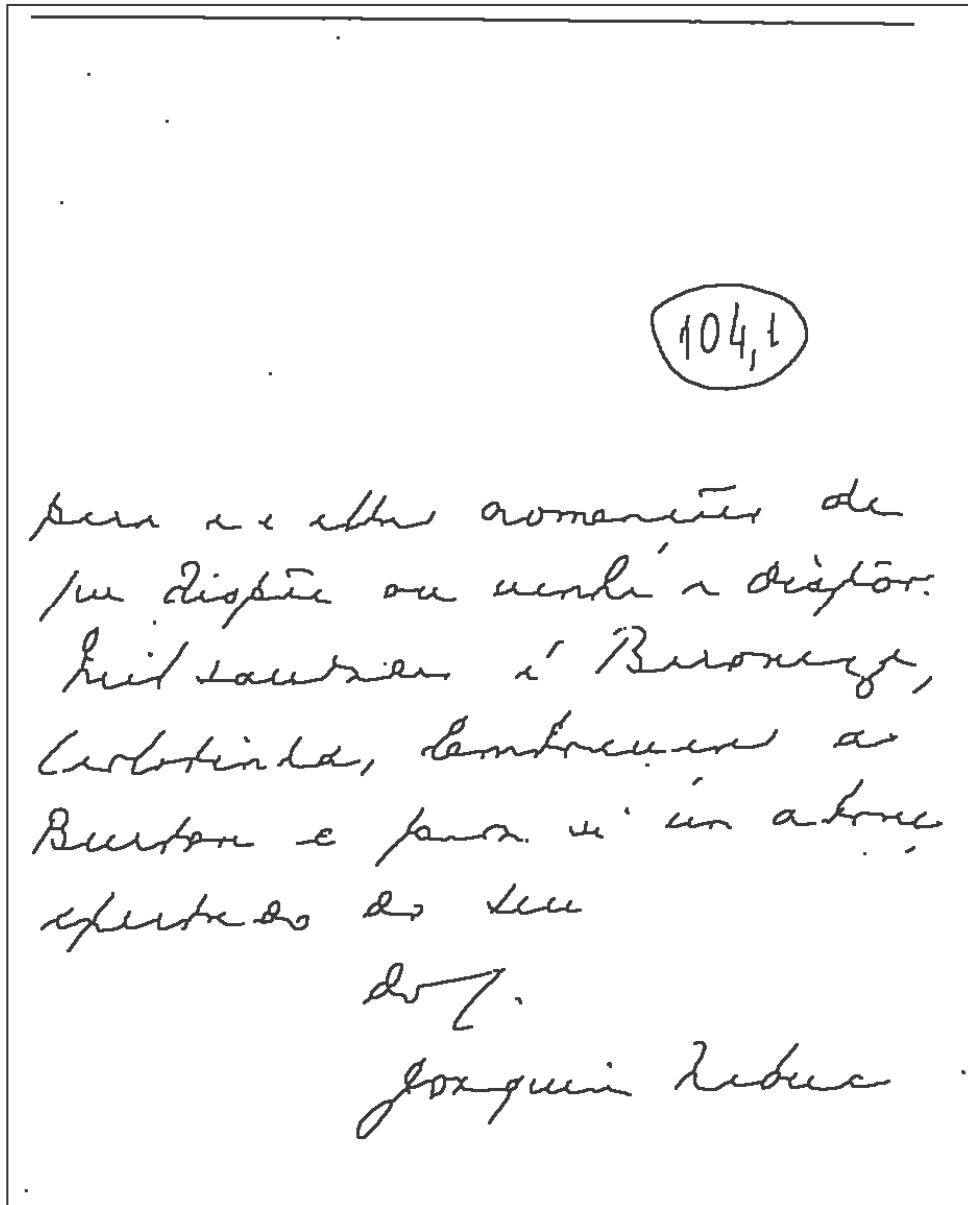
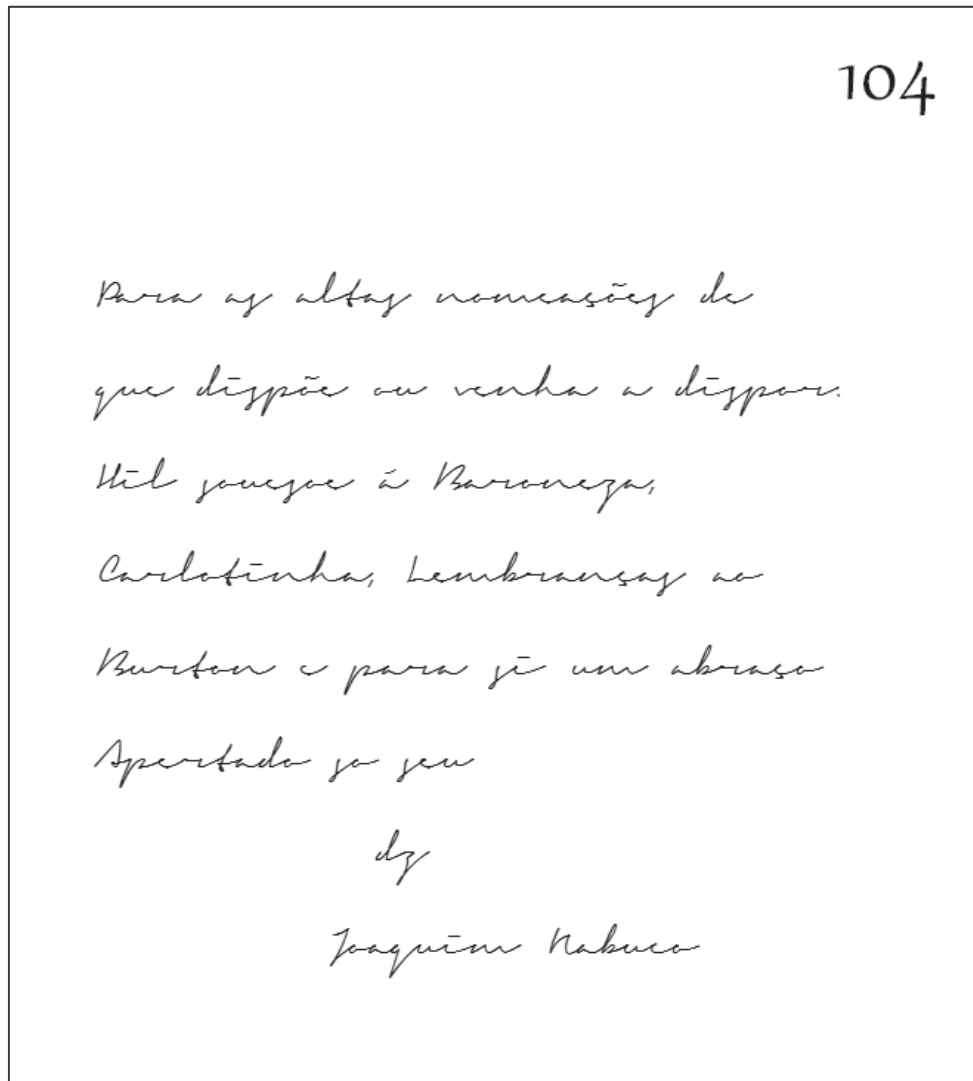
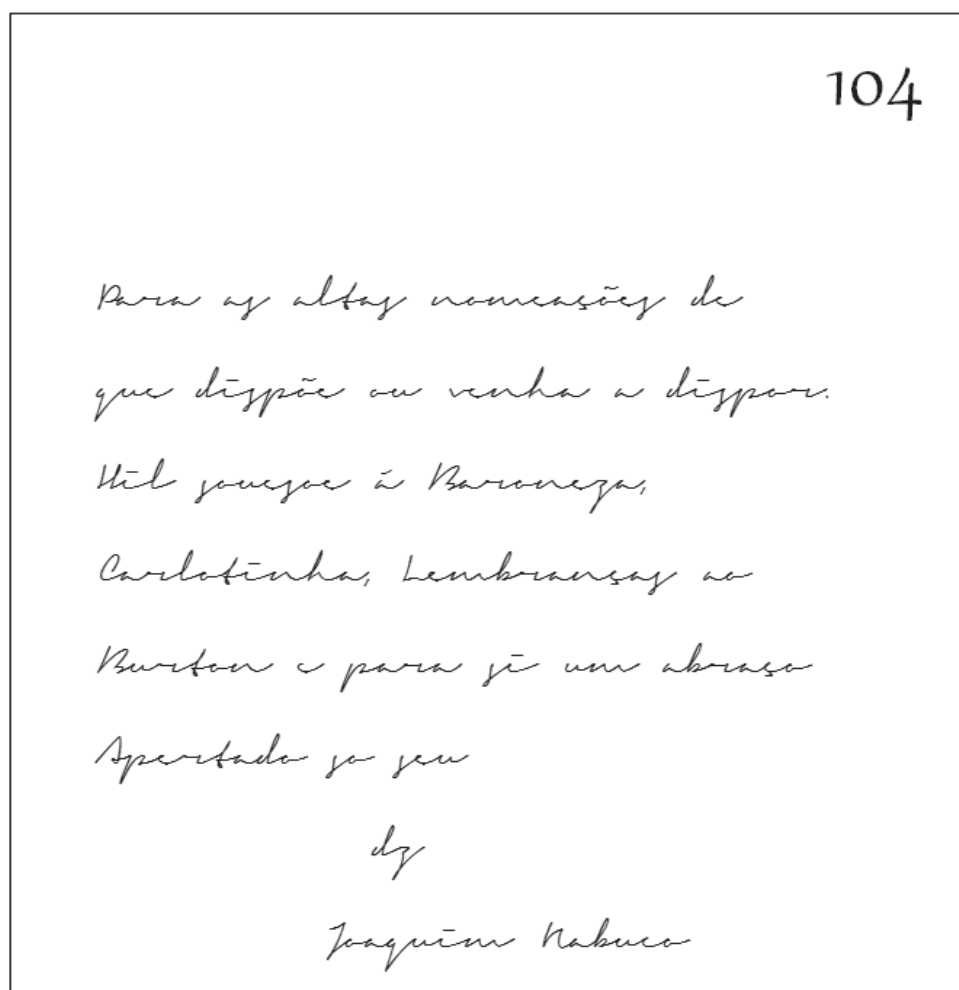


Figure 4. Skeletonized and dilated version of letter shown in Figure 3



**Figure 5.** Synthetic skeletonized image generated from typesetting the text of the original document with the “Signerica Fat” font set.



**Figure 6.** Image of Figure 5 after vectorization and deformation to make a font case matching to the original document after skeletonization and dilatation (Figure 4).



The comparison of the images of Figure 2 and Figure 5 shows several small differences, but there is a mapping path between each letter in the original text (ASCII character) and a “font” that resembles the author calligraphic pattern, which allows the automatic generation of a dictionary of patterns to be used as a training set for the recognizer.

The comparison of the images of Figure 2 and Figure 5 shows several small differences, but there is a mapping path between each letter in the original text (ASCII character) and a “font” that resembles the author calligraphic pattern, which allows the automatic generation of a dictionary of patterns to be used as a training set for the recognizer.

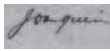
A MLP [8] and two SOM [10] fuzzy classifiers were used in parallel and the majority vote is taken for the transcription of the 25 letters in the document test set, totaling 2,115 words (with at least three letters), both trained with the same dictionary of synthesized words. The result obtained was of 61% (1,294 words) correctly transcribed and 17% (364 words) mismatched into (incorrect) valid words. Testing the whole set of fifty letters (3,584 words), that include the 25 letters used to develop the training set the results were of 67% words correctly transcribed and 15% of “false-positive” words. The result of the classifier applied to the remaining 25 letters yielded a precision and recall of 72%. Table 1 shows the significance of the recognition rate reached may be seen if one attempts to automatically recognize the document in Figure 2 with the classifier trained using the approach presented here and three of the best OCR softwares available today in the market: the Abby FineReader version 12 [19], Omnipage [23] and OCRopus 0.3.1 (alpha3) [22] that calls Tesseract.

Table 1 witnesses the suitability of the method proposed here. It is interesting to notice that even the human reader does not know what Joaquim Nabuco meant with the symbol (?) just before his signature. The transcription automatically made using the methodology proposed here may be considered very successful, overall if compared with the transcriptions obtained by the commercial OCRs tested (Tesseract produced no output at all). One interesting fact to observe is that although the grammatically correct accent in the third line of the text is “à”, Nabuco’s writing was very calligraphically “imprecise” and looks as “á”, as automatically transcribed. One may not consider that an error or even that he misspelled the lexeme “à”, because the “á” in isolation does not exist in Portuguese. The addition of a dictionary may solve such a problem as well as some other as for instance the transcribed word “Hil” does not exist in Portuguese and the only possible valid candidate is the correct word “Mil” (one thousand).

### 3.2. Word recognition in death certificates

Death certificates provide important data such as *causa mortis*, age of death, birth and death places, parental information, etc. Such information may be used to analyze not only what caused the death of the person, but also a large number of demographic information such as internal migration, the relation of death cause with marital status, sex, profession, etc.

Images were acquired by The Family Search International Institute using a camera-based platform.

104 Para as altas nomeações de que dispõe ou venha a dispor. Mil saudades à Baroneza, Carlotinha, Lembranças ao Burton e para si um abraço Apertado do seu ??? Joaquim Nabuco		104 Para as altas nomeações de que dispõe ou venha a dispor. Hil souesoe á Baroneza, Carlotinha, Lembranças ao Burton e para si um abraço Apertado do seu dz Joaquim Nabuco
<b>Human transcribed text</b>		<b>Proposed Method</b>
1*^	_* * ' ^ ^ ^ ^ f* ^CjL-iU As À-«-\$tjt_Jt-C 	
<b>Omnipage</b>	<b>Abby FineReader Professional Edition 12</b>	<b>OCRopus/Tesseract</b>

**Table 1.** Human transcribed text of the document in Figure 2 and the automatic transcriptions by the Proposed Method, Omnipage, Abby FineReader and Tesseract.

Thanatos [1] is a platform designed to extract information from the Death Certificate Records in Pernambuco (Brazil), a collection of “books” kept by the local authorities from the 16th century onwards. The current phase of the Thanatos project focuses on the books from the 19th century. During such period, registration books were pre-printed with blank spaces to be filled in by the notary, as shown in Figure 7. Pre-processing is performed to remove noisy borders using the algorithm described in reference [6] incorporated in the

HistDoc Platform as this step influences all the result of the other subsequent algorithms, the result of which is shown in Figure 8. Image processing continues on the border-removed image (Figure 8) to make image-size (resolution) uniform, binarize, correct skew (using the algorithm Ávila and Lins [3], 2005 also implemented in HistDoc [13]), remove salt-and-pepper and clutter noises, and finally splitting an image in two images each of them corresponding to one death certificate as shown in Figure 9.

Notaries in Brazil are a concession of the State. They are a permanent position many people exercise throughout their lives. Thus, most record books are written by a single person, allowing one to use the strategy proposed here to train the classifier to recognize the content of the different fields. Masks are then applied to extract the content of each of the fields filled in by notaries to extract the content.

They are:

- N<sup>o</sup> (Register number) – placed at the top of the left margin of the register. It conveys numerical information only. Example: N<sup>o</sup> 19.945.
- Data (Date) – the date is written in words and the information is filled in three fields for day, month, and year in this sequence. Example: Aos vinte e três dias do mês de janeiro de mil novecentos e sessenta e seis (At the twenty three days of the month of January of one thousand nine hundred and sixty six).
- Nome do cartório (Notary name) – this field holds the name of the place where the notary office was found. Example: neste cartório da Encruzilhada (at this notary office at Encruzilhada).
- Município do Cartório (City of the notary office) – Example: município de Recife (at the city of Recife).
- Estado do Cartório (State of the notary office) - Example: Estado de Pernambuco (State of Pernambuco).
- Nome do Declarante (Name of declarer) – Name of who attended the office to inform the death. Example: compareceu Guilherme dos Santos (attended Guilherme dos Santos).
- Nome do Médico (Name of the Medical Doctor) – Name of the M.D. who checked the death. Example: exibindo um atestado de óbito firmado pelo doutor José Ricardo (showing a death declaration signed by doctor José Ricardo).
- Causa mortis – Specifies the reason of the death in the declaration from the M.D. Example: dando como causa da morte edema pulmonar, o qual fica arquivado (that states as cause of death lung edema, which is filed).

The first strategy reported in reference [1] for information recognition in the Thanatos platform was to transcribe the fields using the commercial OCR tool ABBYY FineReader 12 Professional Editor [19]. The results obtained were zero correct recognition for all fields, including even the numerical ones. Such disappointing results forced the development of a recognition tool for the Thanatos platform based in the approach in reference [17] that makes use of a set of geometrical and perceptual features extracted from “zoning” the image.

“Zoning” may be seen as splitting a complex pattern in several simpler ones [18] [11] [7]. The original Thanatos strategy used dictionaries to analyze the possible “answers” to the blank fields. The original results of tests performed with 300 death certificates extracted from the same book of death records [1] were already considered reasonable and are shown in the first column of Table 2.

The adoption of the strategy presented here to generate the features of the writer through the modification of a cursive type font text was adopted. The list of all cities and places (villages, neighborhoods, etc) in the state of Pernambuco was collected from IBGE (the Brazilian Geographic and Statistical Institute) a social science research institute responsible for demographic and economic statistics and data collection in Brazil. Another list of family names was also generated having as basis the local phone directory. Those lists were “typeset” using the synthetic set of features extracted and then used to train the classifiers. The results obtained adopting this strategy is presented in the New column in Table 2. It is important to stress that the same parallel architecture (MLP + 2 SOM) fuzzy classifiers with majority vote was used in both cases, only with different training sets.

Field	Thanatos	New
Name of Notary	98.0%	98.5%
City of the Notary	71.0%	94.0%
State of the Notary	98.0%	100.0%
Place of death	31.0%	73.0%
Numbers in writing - Time of obit	69.0%	91.0%
Numbers in writing - Date of death	69.0%	91.0%
Numbers in writing - Date of birth	69.0%	91.0%
Color of skin	100.0%	100.0%
Marital status	100.0%	100.0%

**Table 2.** Recognition rate for non-numerical fields in 300 certificates.

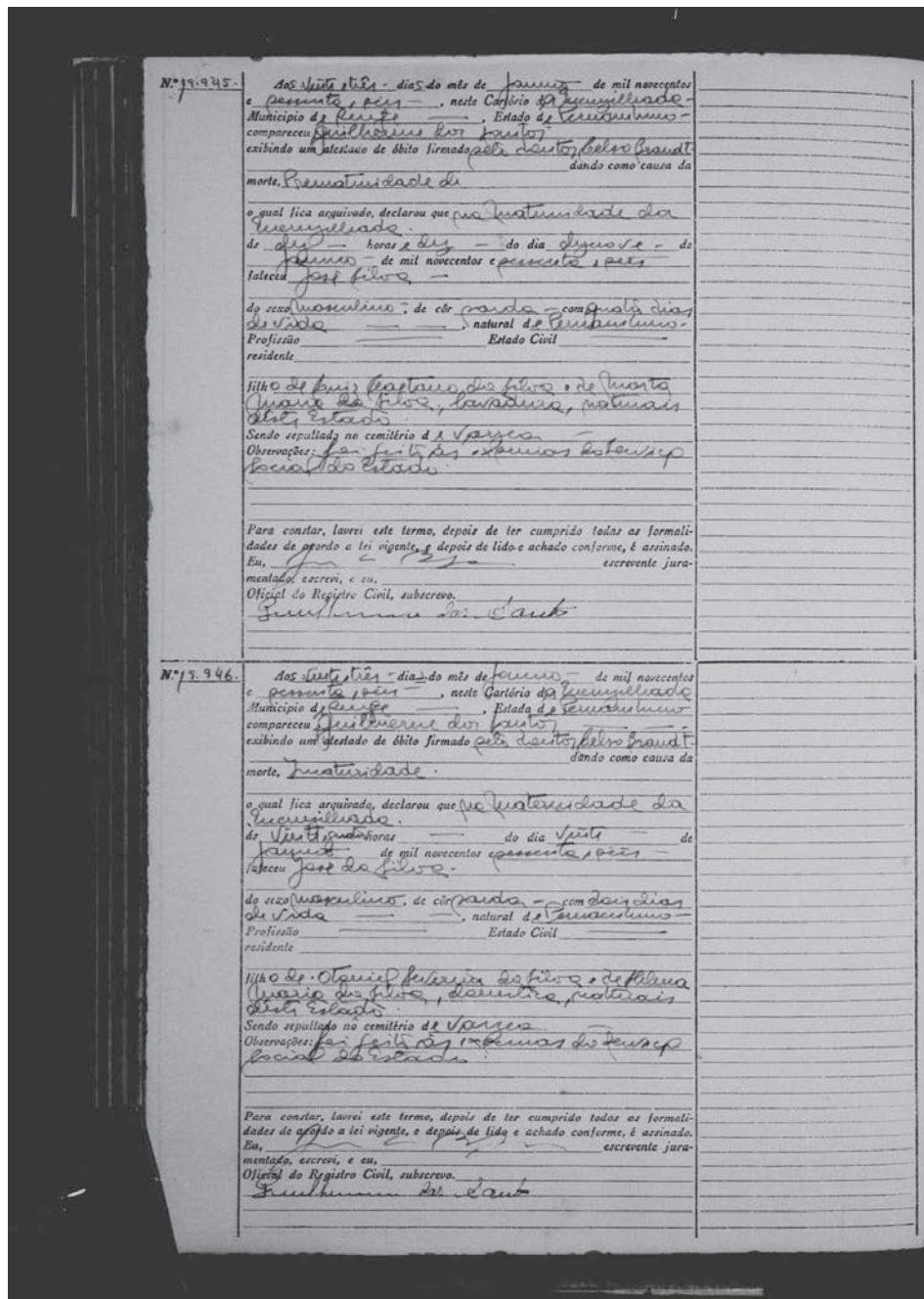


Figure 7. Original image from a book of printed forms of death certificates in Pernambuco (Brazil) – 1966.

N.º 945. Dos vinte e três dias do mês de Janeiro de mil novecentos e cinquenta e seis, neste Cartório de Registro Civil do Município de Recife, Estado de Pernambuco compareceu José Gomes dos Santos exibindo um atestado de óbito firmado pelo doutor Celso Brandt dando como causa da morte, Prematuridade de

o qual fica arquivado, declarou que por intermédio da

do dia vinte e três do mês de Janeiro de mil novecentos e cinquenta e seis faleceu José Filho.

do sexo masculino, de cor branca, com quatro dias de vida, natural de Pernambuco. Profissão residente Estado Civil

filho de Luiz Costares dos Filhos e de Maria Maria dos Filhos, brasileiros, naturais deste Estado.

sendo sepultado no cemitério de Várzea.

Observações: foi feita a exposição do corpo social do Estado.

Para constar, lavrei este termo, depois de ter cumprido todas as formalidades de acordo a lei vigente, e depois de lido e achado conforme, é assinado. Eu, escrevente juramentado, escrevi, e eu, Oficial do Registro Civil, subscrevo.

Frederico dos Santos

---

N.º 946. Dos vinte e três dias do mês de Janeiro de mil novecentos e cinquenta e seis, neste Cartório de Registro Civil do Município de Recife, Estado de Pernambuco compareceu José Gomes dos Santos exibindo um atestado de óbito firmado pelo doutor Celso Brandt dando como causa da morte, Intoxicação.

o qual fica arquivado, declarou que por intermédio da

do dia vinte e três do mês de Janeiro de mil novecentos e cinquenta e seis faleceu José do Filho.

do sexo masculino, de cor branca, com dois dias de vida, natural de Pernambuco. Profissão residente Estado Civil

filho de Otávio Ferreira dos Filhos e de Helena Maria dos Filhos, brasileiros, naturais deste Estado.

sendo sepultado no cemitério de Várzea.

Observações: foi feita a exposição do corpo social do Estado.

Para constar, lavrei este termo, depois de ter cumprido todas as formalidades de acordo a lei vigente, e depois de lido e achado conforme, é assinado. Eu, escrevente juramentado, escrevi, e eu, Oficial do Registro Civil, subscrevo.

Frederico dos Santos

Figure 8. Filtered version of the image in Figure 7.

Nº 19.945.

dos vinte e três - dias do mês de Januário de mil novecentos e sessenta e seis - neste Cartório de Recém-nascidos - Município de Recém-nascidos - Estado de Pernambuco - compareceu Antônio Carlos de Azevedo exibindo um atestado de óbito firmado pelo Dr. Carlos Brandt dando como causa da morte, Prematuridade da

o qual fica arquivado, declarou que na prematuridade da recém-nascido.

de dois - horas e dois - do dia dois - de Januário de mil novecentos e sessenta e seis faleceu José Filho -

do sexo masculino; de cor branca - com quatro dias de idade natural de Pernambuco - Profissão --- Estado Civil --- residente ---

filho de Antônio Carlos de Azevedo e de Martha Maria de Azevedo; brancos, naturais do Estado.

Sendo sepultado no cemitério de Verde.

Observações: for feitas as observações estabelecidas no Estado.

Para constar, lavrei este termo, depois de ter cumprido todas as formalidades de acordo a lei vigente, e depois de lido e achado conforme, é assinado. Eu, --- escrevente juramentado, escrevi, e eu, --- Oficial do Registro Civil, subcrevo.

Guilherme de Azevedo

Figure 9. Monochromatic version of Death Certificate after filtering and splitting the image in Figure 8.

The column Thanatos refers to the results obtained in reference [1], while **New** presents the results of the strategy presented in this paper. Table 2 shows that the new strategy presented here presented either no loss or gains in the recognition rate of all fields recognized in relation to the results presented in reference [1]. In the case of the field "Place of death" the increase in recognition rate reached 42%.

#### 4. Conclusions and lines for further work

Handwritten recognition to gain any degree of success either chooses a limited dictionary and allows a large number of writers or widens the vocabulary and largely restricts the numbers of writers. In both cases, the choice of the training set is of central importance for the success of any classification strategy and must be representative of the whole "universe" one wants to correctly recognize. This paper presents a new way of automatically generating the training set for the recognition of a large set of words written by a single user. It has as starting point different sets cursive type fonts, which are modified and compared to the original writing to "match their features". Once the "matching path" is found it is applied to a large dictionary in that encompasses the vocabulary of the document, generating the training set to be used for the whole batch of documents to be transcribed.

The strategy presented here was used with success in two sets of documents. In the case of the transcription of the handwritten letters in the bequest of Joaquim Nabuco it reached the correct rate of 67% transcribed words (of more than three letters), a result that may be considered successful at least for keyword indexing of such historical documents. In the case of death certificates of the Thanatos project, whose vocabulary is far more restricted the results presented either no loss or gains in the recognition rate of all fields recognized in relation to the previous results, reaching an average of 93.79% correct field recognition.

The statistical data collected inter character and inter word spacing, line and character skew, inter line separation were not used to enrich the generation of entries in the dictionary of the training set. Its use is left as a possibility for further work.

### Author details

Gabriel Pereira e Silva and Rafael Dueire Lins  
*Universidade Federal de Pernambuco, Brazil*

### Acknowledgement

The authors are grateful to the organizers of the Fontspace site for setting such a useful site, fundamental for the development of this work. The authors also thank the Family Search International Institute for the initiative of digitizing the death certificate records of Pernambuco (Brazil) and to Tribunal de Justiça de Pernambuco (TJPE) to allow the use of such data for research purposes.

Research presented here is partly sponsored by CNPq-Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico, Brazilian Government.

### 5. References

- [1] A. Almeida, R.D.Lins, and G.F. Pereira e Silva. Thanatos. Automatically Retrieving Information from Death Certificates in Brazil. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pp. 146-153, ACM Press, 2011.
- [2] A. I. de S. L. Andrade, C. L. de S. L. Rêgo, T. C. de S. Dantas, Catálogo da Correspondência de Joaquim Nabuco 1903-1906, volume I 1865-1884, volume II 1885-1889, volume III 1890-1910, Editora Massangana, ISBN 857019126X, 1980. (Available at: [www.fundaj.gov.br/geral/2010anojn/catalogo\\_nabuco\\_v2.pdf](http://www.fundaj.gov.br/geral/2010anojn/catalogo_nabuco_v2.pdf))
- [3] B. T. Ávila and R. D. Lins. A Fast Orientation and Skew Detection Algorithm for Monochromatic Document Images. 2005 ACM International Conference on Document Engineering, p.118 - 126. ACM Press, 2005.
- [4] L. Bethell, J. M. De Carvalho. Joaquim Nabuco, British Abolitionists, and the End of Slavery in Brazil: Correspondence 1880-1905, Institute for the Studies of the Americas, 2009. ISBN-13: 978-1900039956.



- [5] M. Bulacu and L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *IEEE Trans. on PAMI*, 29(4):701-717, April 2007.
- [6] A. de A. Formiga and R. D. Lins. Efficient Removal of Noisy Borders of Monochromatic Documents. *International Conference on Image Analysis and Recognition, 2009, LNCS v.5627*. p.158 – 167, Springer Verlag, 2009.
- [7] C. O. A. Freitas, L.S. Oliveira, S.B.K. Aires, F. Bortolozzi, Zoning and metaclasses for character recognition. *ACM-SAC 2007*. P. 632-636, 2007.
- [8] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Hardcover (2nd Edition), 1998.
- [9] M. Hu, "Visual pattern recognition by moment invariants", *IEEE Transactions on Information Theory*, 8(2):179-187, 1962.
- [10] T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, Springer, second edition, vol. 30, 1997.
- [11] Z.C. Li, C.Y. Suen, J. Guo, A Regional Decomposition Method for Recognizing Handprinted Characters, *IEEE Transactions on Systems, Man, and Cybernetics*, N. 25, p. 998-1010, 1995.
- [12] R.D.Lins. Nabuco - Two Decades of Document Processing in Latin America, *Journal of Universal Computer Science*, v. 17(1), pp. 151-161, 2011.
- [13] R.D.Lins, G.F. Pereira e Silva, A.de A. Formiga. HistDoc v. 2.0: enhancing a platform to process historical documents. *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pp. 169-176, ACM Press, 2011.
- [14] C. Liu, Y. Liu, and R. Dai, "Preprocessing and statistical/structural feature extraction for handwritten numeral recognition", *Progress of Handwriting Recognition*, A.C. Downton and S. Impedovo eds., World Scientific, 1997.
- [15] C. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art techniques", *Pattern Recognition*, 36(10):2271-2285, 2003.
- [16] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, "Automatic recognition of handwritten numerical strings: A recognition and verification strategy", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(11):1438-1454, 2002.
- [17] G. F. Pereira e Silva and R. D. Lins. An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. *ICDAR 2011, Beijing, September*, IEEE Press, 2011.
- [18] C.Y. Suen, J. Guo, Z.C Li, Analysis and Recognition of Alphanumeric Handprints by parts, *IEEE Transactions on Systems, Man, and Cybernetics*, N. 24, p. 614-631, 1994.
- [19] ABBYY FineReader 12 Professional Editor, <http://finereader.abbyy.com/> , last visited on 13.04.2012.
- [20] IMAGEJ. <http://rsbweb.nih.gov/ij/> , last visited on 13.04.2012.
- [21] <http://www.fontspace.com/category/cursive?p=19> , last visited on 13.04.2012.

[22] OCROPUS 0.3.1 (alpha3): <http://code.google.com/p/ocropus/>

[23] Nuance OminiPage Professional 16: <http://www.nuance.com/for-individuals/by-product/omnipage/index.htm>

INTECH

INTECH

# Automatic Training Set Generation for Better Historic Document Transcription and Compression

Gabriel França Pereira e Silva<sup>1,2</sup>, Rafael Dueire Lins<sup>1</sup> and Cesar Gomes<sup>2</sup>

<sup>1</sup>Universidade Federal de Pernambuco, Recife, Brazil

<sup>2</sup>Universidade Federal Rural de Pernambuco, Garanhuns, Brazil

{gfps, rdl}@cin.ufpe.br

**Abstract** — The more complete the training set of an optical character recognition platform, the greater the chances of obtaining a better precision in transcription. The development of a database for such purpose is a task of paramount effort as it is performed manually and must be as extensive as possible in order to potentially cover all words in a language. Dealing with historic documents either handwritten, typed, or printed is even a harder effort as documents are often degraded by time and storage conditions. The recent work of Silva-Lins showed how to automatically generate training sets of isolated characters for cursive writing of one specific person. This is particularly important in the transcription of historic files of important people. The present work improves that strategy by analyzing letter ligature patterns. The improvement in OCR transcription accuracy both of printed, typed and handwritten documents is borne out by experimental evidence.

**Keywords** — *OCR, training sets, font sets, documents.*

## I. INTRODUCTION

Someone's hand writing is part of his personality and carries some individual elements. Each person has a proper writing "style", which may vary according to psychological state, the kind of document written, and even physical elements such as the texture of the paper and kind of pencil or pen used. Despite such wide range of variation possibilities some elements tend to remain unchanged in a way that other people, in general, can recognize one's writing and even identify the authorship of a document. Although written character recognition is a task of high complexity even for humans, sometimes, very seldom one is unable to identify one's own writing.

The automatic transcription of historic documents, printed, typed or hand written, is still a challenging problem in pattern recognition. The basis for pattern recognition rests on two pillars. The first one is to find the minimal set of features that presents all maximum diversity within the universe of study. The second one is to find a suitable training set that also covers all possible data to be classified. In the case of hand written cursive documents, due to the variation of writing styles between people, one should not expect that a general classifier yields good recognition performance in a general context. Thus, one tends to either have general classifiers for very specific restricted vocabularies (such as digits), or to have personalized recognizers for general contexts.

The recent work of Silva-Lins [3] showed how to automatically generate training sets of isolated characters for cursive writing of one specific person. In such context it is a burden and very difficult to generate a good training set to allow the classifier to reach a reasonable recognition rate. The approach

followed for that is first to select a set of documents representative of the author's style. As in the Internet one may find several public domain sites with font sets, the key idea presented by Silva-Lins [3] is "approximating" the author writing by a cursive typographical font, which is skeletonized and a "standard" training set is generated. Such strategy was adopted with success with documents of the Nabuco bequest [1] and of the Thanatos Project [17].

This paper improves the scheme by Silva-Lins by analyzing letter ligature patterns. The proposed strategy was tested in two scenarios. The first is in automatic character recognition by transcribing the historic documents of Joaquim Nabuco [1], a Brazilian statesman, thinker and writer, the first Brazilian ambassador to the U.S.A., and one of the pioneers in the campaign in freeing the black slaves in Brazil. Three different types of documents were analyzed: hand written cursive, typed, and printed. Joaquim Nabuco, as many of his contemporaries, one may say had a calligraphy (from the Greek, beautiful cursive hand writing) that did not vary much throughout his life. He was also "loyal" to his typing machine that accompanied him all his life long. The printed documents in Nabuco bequest also used the same typographic fonts. Such a standard in the three different kinds of documents of interest allows the development of specific training sets for each kind of document. Thus it happens as if an OCR had been developed and tuned for each of those kinds of documents. The second contribution of the strategy for the automatic generation of training sets presented here is providing better results in the synthesis of historic documents as a compression strategy, as proposed by reference [14].

This paper is organized as follows: Section 2 presents the method developed for the generation of the training set. Section 3 presents the results of the experimental evaluation of the method proposed. The final section of this paper presents its conclusions and draws lines for further work.

## II. THE PROPOSED METHOD

The present section explains the proposed method on how to automatically generate a training set. The choice of a representative training set together with a good set of features is fundamental for the success of automatic pattern recognition. These two factors are tightly linked to each other and in such a way as to grant good recognition results. As the recognition of cursive hand writing is by far the most complex of the three kinds of documents analyzed here, the explanation of how the training set is developed focuses in such kind of documents. The strategy adopted for the development of training sets for the other two kinds of documents (typed and printed) is exactly the same.

To obtain a good training set for handwriting recognition of a given author during a period of time in which the writing features

are stable (they changes with age, psychological and health factors, etc.) one has to group together the documents that have similar properties. A subset of them that is representative of the set of documents (in general the size of the training set is about 10% of the size of the whole data “universe”) is chosen in such a way as to cover the whole diversity of the documents to be transcribed. Once the set of documents representative of the file is chosen it be transcribed by a human reader and will undergo five phases that are explained as follows: image pre-processing, word normalization, selection of the closest cursive font set, ligature extraction, and finally word generation.

### A. Image Pre-Processing

Document images acquired either using scanners or digital cameras almost always encompass noises that degrade the performance of OCRs [3]. Thus, some pre-processing must be performed to filter out the noises, increasing the chances of a good document segmentation and transcription. This phase made use of the HistDoc v2.0 tool presented in reference [4]. The first step consists in detecting the different kinds of noises and evaluating the parameters suitable to filter them out. The images undergo skew correction and the back-to-front interference (bleeding) removal. At last, the image of the document is binarized using a global [5] and a local [6] algorithm. The local algorithm makes use of a 50x50 pixel window and k-parameter equal to 0.4. Figure 1 presents a screen-dump of the tool processing a document from Nabuco bequest, while Figure 3 presents an example of a part of a historic document and its binarized version using HistDoc v2.0 [4].

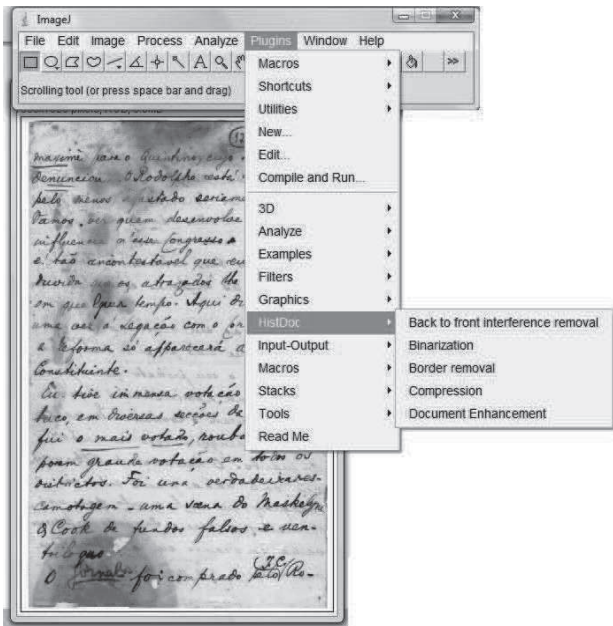


Figure 1. Screen dump of the HistDoc environment processing the image of a historic document.

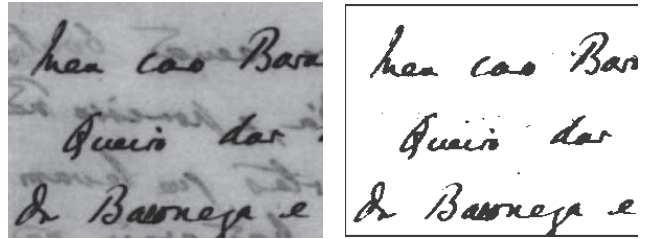


Figure 2. Zoom into a document from Nabuco’s file (left) and the result of its binarization (right).

### B. Word Normalization

This step has as objective to minimize the variations in the writing of the documents in the training set, such as different writing instrument (pen, pencil, etc.), different paper texture, variations in the writing speed, etc. It consists in skeletonization of the text (to make uniform the width of the writing), followed by image dilatation. Figure 3 presents an example of such process.



Figure 3. Example of the word normalization step applied to the word “Joaquim”.

### C. Choice of the Cursive Font Set

The development of the method proposed by Silva-Lins [3] starts by using a set of cursive fonts. In the Internet one may find several public domain sites with font sets. In particular the site Fontspace [8] offers 282 different cursive font sets for download. The central difficulty in generating the training set for handwritten documents is to have a “font set” of a specific author to extract the convenient features for pattern matching. The strategy adopted is:

- The user should select a number of a cursive font set that bears “some resemblance” to the original author handwriting.
- The text version of the document is typeset in each of the cursive font sets chosen.
- All the typeset versions of the document are converted into image.
- The image of the original document is skeletonised and then dilated.
- Segment the image in boxes around each letter (font cases) of the skeletonised and dilated version of the original image and the synthetically generated images.

Apply a deformation transform to make each font case in the synthetic images coincide with the font case of the skeletonised-dilated version of the original document. Image vectorization, is important to increase the likelihood between the synthetic and the original documents. Such operation is applied to each character in each synthetically generated image by deforming the bounding-box and the strokes until there is a perfect match between the synthetic and the original one. In this “deformation” process some statistical analysis is performed to infer data about inter character and inter word spacing, line and character skew, inter line separation, etc.

The process of determining the degree of likeness between the cursive font sets and the author writing is performed in two phases. In the first phase, the structural (global) features used for pattern recognition are:

- Geometric Moments [15] [9];
- Concavity Measurements [16];
- Shape Representation of Profile [14];
- Distance between barycentric points between two consecutive characters;
- Maximum and minimum heights of two consecutive characters.
- Maximum and minimum distance between concavities of two consecutive characters.

The feature vector of a document brings an account of the basic features of the author calligraphy. The Hamming distance between the feature vectors of the synthetic and original images, brings an account of their similarity, and is calculated using the formula:

$$H_w = \sum_{n=1}^{N \text{ features}} |f_{on} - f_{sn}|$$

where  $f_{on}$  and  $f_{sn}$  are the components of the feature vectors of the original and synthetic images, respectively. The choice of a vector of features such that one could extract “information” about the calligraphic pattern of the author shares some ideas with the work in reference [9].

The second phase, which is not performed in the Silva-Lins scheme [3], consists in applying region zoning on the images of words to obtain local features. The zone samples are presented in Figure 4 and consist of rectangular matrices which correspond to different shapes that are associated with the number of characters in each word. Thus, for a word with 2 to 4 characters one uses ( $Z = 2LR, 2UD, \text{ and } 4$ ). For words with more than 4 characters the other patterns are used.

The local and global features are combined to obtain the degree of likelihood between the original author handwriting and the synthetically generated words using the cursive font sets. The degree of likelihood is given by:

$$H_{wf} = |H_{wG} + H_{wL}|$$

where  $H_{wf}$  is the final distance,  $H_{wG}$  is the global Hamming distance and  $H_{wL}$  is the local Hamming distance.

The font set that provides the smallest  $H_{wf}$  to the original author writing is chosen to the “matching phase”.

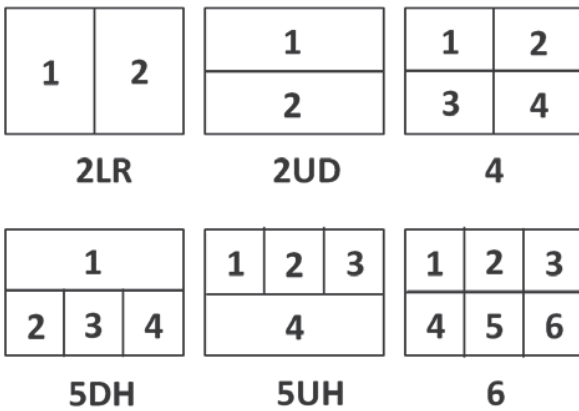


Figure 4.  $Z = 2LR, 2UD, 4, 5DH, 5UH$  and  $6$ .

### Ligature extraction

One of the main problems in the transcription of large files of handwritten documents in general is being able to correctly segment words and characters [10] [11] [12]. This is a crucial step for OCR performance. The work of Silva-Lins [3] is a preliminary step towards the successful character segmentation and automatically generating training sets of characters. This paper moves another step forward and attempts to automatically analyze the ligature between characters that form words to try to improve the quality of the training set of OCR’s.

A set of images of words in cursive fonts is generated. Such images are vectorized and normalized by calculating the number of pixels in the synthetic and original images and making them as close as possible. This process is performed by overlapping the original and synthetic image of each selected word and interactively “deforming” the synthetic image in a morphing process. During the vectorization process the ligatures between characters that form a word are marked.

The technique proposed in reference [13] is improved here as in the ligature detection process an additional resource is incorporated: the end points of ligature between characters are defined during the approximation step described above. This allows for a refinement in the neural classifier. The direct application of the result obtained in the approximation is to allow extracting the root, prefixes, and suffixes of words, besides splitting it in syllabic blocks. This technique allows finding the kind of ligature between one syllabic block and the following character or a character and the following syllabic block.

### D. Word Generation

The step presented above yields a better set of generated synthetic fonts that are closer to the author original writing, which can be used as a training set for an OCR or to synthetically generate documents. The pattern granularity evolved from isolated characters (as in Silva-Lins [3]) to syllabic blocks of characters. This allows having as training set not only words extracted from the original documents, but also synthetically generated words that are created by transcribing words in a dictionary, through the syllabic or character matching and their composition. Deformations such as skew, widening and narrowing tokens are allowed combining synthetic and original tokens to form new cursive images of words in the dictionary.

## III. OCR TRANSCRIPTION RESULTS

The experiments reported in Silva-Lins [3] show that the classifier trained with the automatically developed training set largely outperformed three of the best OCR softwares available today in the market: the ABBYY FineReader® version 11 [18], Omnipage® [19] and OCRopus® 0.3.1 (alpha3) [7] that calls Tesseract®. ABBYY FineReader® version 11 [18] without dictionary outperformed the same software running with the aid of its dictionary and the other two OCRs working in any configuration. Thus, in this paper only the best methods for automatic training set generation are compared. The performance of these algorithms is measured by the number of correctly transcribed words.

Similarly to reference [3], a MLP [15] and two SOM [16] fuzzy classifiers were used in parallel and the majority vote is taken.

The nets analyze the whole words and their segments, applying the algorithm proposed in reference [13].

The experiments reported here are split into three different parts depending on the nature of the documents: cursive handwritten, typed and printed ones. In what follows each of them is detailed.

### A. Cursive Handwritten Documents

Two databases were used. The first one is formed by 50 handwritten letters of Joaquim Nabuco [1], totaling 3,584 words, representative of the whole universe of the almost 6,000 documents in Nabuco's bequest. From those documents, 20 letters were selected for the training set, yielding 1,469 words. The execution of the method proposed here generated 5,000 different words used as pattern for the training set. The 3,631 new word patterns were synthetically generated by using the character, syllabic blocks, and ligatures patterns on top of a dictionary, which was generated though the analysis of the most frequent words in the complete works of Machado de Assis [2]. Joaquim Nabuco and Machado de Assis were contemporaries and the latter is considered one of the greatest Brazilian writers of all times. Besides that, a set of 8,000 syllabic block images was also generated by sieving them from a dictionary of Brazilian Portuguese. The method of cross validation was used with 10 folders to generate the classifier, with 82.68% of correctly classified instances. The result of the classification on the 2115 words that were not used as instances for training the classifier is presented in Table 01.

	<b>ABBYY FineReader® 11 Professional</b>	<b>Silva-Lins Ref. [3]</b>	<b>New Method</b>
<b>Nabuco test set</b>	---	<b>75.00%</b>	<b>87.87%</b>

### B. Typed Documents

The experiment performed with typed documents made use of 200 letters from Nabuco bequest such as the one presented in Figure 5. In such documents only the typed part was considered, as the handwritten parts were later annotations made by people other than Joaquim Nabuco. They are related either to the document classification of the bequest made by historians, such as the ones on the top right hand side margin or by some secretary or other people. Handmade corrections and all different sorts of physical noises in the original document increases the difficulty in OCR.

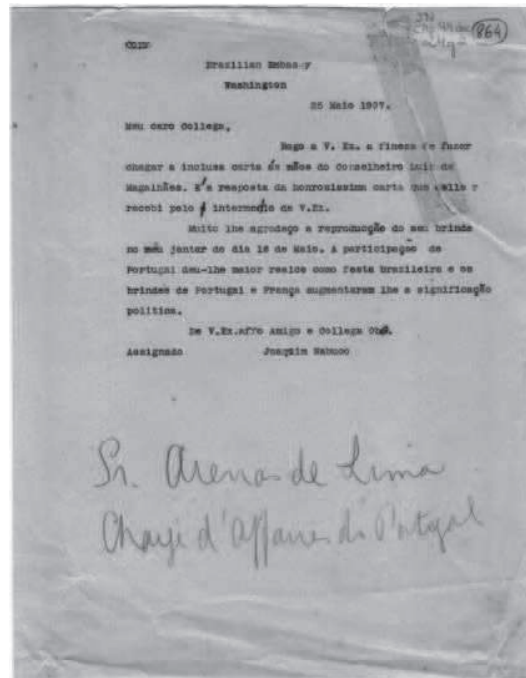
In total 7,785 printed words (length greater or equal to 2 characters) are present in the 200 documents. The training of the recognition set was made with 680 words randomly chosen from that set. The number of correctly transcribed words using the new method is shown in Table 02.

	<b>ABBYY FineReader® 11 Professional</b>	<b>Silva-Lins Ref. [3]</b>	<b>New Method</b>
<b>Nabuco test set</b>	<b>66,37%</b>	<b>74.3%</b>	<b>89.83%</b>

The results obtained show that the gains obtained are even more significant than in the previous one.

### C. Printed Documents

The experiment performed with typed documents made use of 239 pages from a book by Joaquim Nabuco. In total 8,057 printed words (length greater or equal to 2 characters) are present in those 239 printed pages. The training of the recognition set was made with 680 words randomly chosen from that set. The number of correctly transcribed words using the new method is shown in Table 03.



**Figure 5.** Typed letter from Nabuco file (JPG 1% - 390 kB)

As one would expect, printed documents are more easily recognized than the other two sets of documents studied, as the text to be recognized is by far the most uniform one. The performance of ABBYY FineReader® is very good and reached 80.12 % of correct word recognition, but the accuracy of the method proposed here is almost 20% better, reaching 94.47% correct word recognition.

	<b>ABBYY FineReader® 11 Professional</b>	<b>Silva-Lins Ref. [3]</b>	<b>New Method</b>
<b>Nabuco test set</b>	<b>80.12%</b>	<b>88.02%</b>	<b>94.47%</b>

#### IV. CONCLUSIONS AND FURTHER WORKS

This paper significantly improves the recently proposed scheme by Silva-Lins to automatically generate training sets for cursive handwritten documents by analyzing letter ligature patterns. The proposed strategy was tested in the historic documents of the Joaquim Nabuco, one of the most outstanding politician, thinker and social writers from the late 19<sup>th</sup> century, one of the key figures in the freeing of black slavery in Brazil with worldwide repercussion. The documents in Nabuco's bequest were categorized in three groups: handwritten, typed and printed documents, split as shown in Table 04.

Document Type	#Docs	#Words	#Training
Handwritten	50	3,584	1,469
Typed	200	7,785	680
Printed	239	8,057	680
<b>Total</b>	<b>489</b>	<b>19,426</b>	<b>2,829</b>

The automatic method introduced here for the generation of training sets, separated by document type, yielded the results presented in Table 05 compared with ABBYY FineReader® v.11 and the previous version of the automatic strategy proposed for generating training sets described in reference [3].

Document Type	ABBYY FineReader® 11 Professional	Silva-Lins Ref. [3]	New Method
Handwritten	---	75.00%	87.87%
Typed	66.37%	74.30%	89.83%
Printed	80.12%	78.02%	94.47%

The data presented in Table 05 yield to conclude that the presented method largely improves the accuracy of the transcription of the three classes of documents, performing better than the automatic training set generation strategy presented in reference [3] and ABBYY FineReader® v.11.[18].

The automatic training set generation scheme presented here was also advantageously used as part of the compression scheme presented in reference [14] that, through the synthetic document generation, resembles the original one. Table 06 presents the results obtained in file compression using the scheme described.

File type	Size
Original – JPG 1% loss	393 KB
Synthesized – MS Word	183 KB
Pdf text (editable)	91 KB
Synthetic – JPG 1% loss	39 KB

The results shown in Table 06 allow concluding that the proposed scheme is viable for document compression, being particularly suitable for the typed documents.

One may envisage several lines for further works. In the specific line of document synthesis as a compression strategy one may improve the method proposed by factoring out some layout

information in the documents. For instance, all pages in a book have approximately the same margins and texture, thus the compression scheme may store that only once and generate all pages in one batch. Another possibility is to save as image not only words and their ligature, but also to include in training set some special patterns, such as signatures.

#### V. REFERENCES

- [1] Lins, Rafael Dueire. "Nabuco - Two Decades of Document Processing in Latin America", Journal of Universal Computer Science, v. 17(1), pp. 151-161, 2011.
- [2] Machado de Assis. Complete Works. Available at <http://machado.mec.gov.br/> last visited on 13.02.2013.
- [3] Silva, Gabriel de França Pereira e; LINS, Rafael Dueire. "A New Strategy to Generate Training Sets for the Automatic Recognition of Handwritten Documents". In: Xiaoqing Ding (Org.). Character Recognition. 1<sup>st</sup> edition. New York: InTech, 2012, v. 1, p. 155-174.
- [4] Lins, Rafael Dueire. "A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises" ICIAR 2009, Springer Verlag, 2009. v.5627. p.844 - 854.
- [5] Lins, Rafael Dueire; Silva, Gabriel de França Pereira e ; Formiga, Andrei de Araújo. HistDoc v. 2.0: enhancing a platform to process historical documents. In: Workshop on Historical Document Imaging and Processing, 2011, Beijing. HIP '11. New York: ACM, 2011. v. 1. p. 169-176.
- [6] Silva, João Marcelo Monte da ; Lins, Rafael Dueire ; Martins, Fernando Mário Junqueira ; Wachenchauer, Rosa Graziela . "A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference". Journal of Universal Computer Science, v. 14, p. 299-313, 2008.
- [7] Nuance OmniPage Professional 16: <http://www.nuance.com/for-individuals/by-product/omnipage/index.htm>.
- [8] <http://www.fontspace.com/category/cursive?p=19>, last visited on 10.01.2013.
- [9] Bulacu, M. and Schomaker, L. Text-independent writer identification and verification using textural and allographic features. IEEE Trans. on PAMI, 29(4):701.717, April 2007.
- [10] Casey, R. G. and Lecolinet, E. "A Survey of Methods and Strategies in Character Segmentation", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 18, pp. 690-706, 1996.
- [11] Y. Lu and M. Shridhar, "Character Segmentation in Handwritten Words – An Overview", Pattern Recognition, Vol. 29, pp. 77-96, 1996.
- [12] G.Vamvakas, N. Stamatopoulos, B.Gatos, S.J.Perantonis, "Automatic Unsupervised Parameter Selection for Character Segmentation" , 9th IAPR International Workshop on Document Analysis Systems (DAS'10), pp 409-416, June 9-11, Boston, USA, 2010.
- [13] Chun Ki Cheng and Michael Blumenstein, "Improving the Segmentation of Cursive Handwritten Words using Ligature Detection and Neural Validation". The 4th Asia Pacific International Symposium on Information Technology. APIS, 2005.
- [14] Silva, João Marcelo Monte da; Lins, Rafael Dueire. "Color Document Synthesis as a Compression Strategy". ICDAR 2007, Curitiba. ICDAR 2007. IEEE Press, 2007. v.I. p.466 – 470.
- [15] S. Haykin, Neural Networks: A Comprehensive Foundation. Hardcover (2<sup>nd</sup> Edition), 1998.

- [16] T. Kohonen, Self-Organizing Maps, Springer Series in Information Sciences, Springer, second edition, vol. 30, 1997.
- [17] A. Almeida, R.D.Lins, and G.F. Pereira e Silva. Thanatos. Automatically Retrieving Information from Death Certificates in Brazil. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pp. 146-153, ACM Press, 2011.
- [18] ABBYY FineReader 11 Professional Editor, <http://finereader.abbyy.com/>, last visited on 13.02.2013.
- [19] OCROPUS 0.3.1 (alpha3): <http://code.google.com/p/ocropus/>, last visited on 13.02.2013.