

Universidade Federal Fluminense

LINCOLN FARIA DA SILVA

**Distinção Automática de Texto Impresso e
Manuscrito em uma Imagem de Documento**

NITERÓI
2009

LINCOLN FARIA DA SILVA

**Distinção Automática de Texto Impresso e Manuscrito em
uma Imagem de Documento**

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre. Área de concentração: Computação Visual e Interfaces.

Orientadora:
Aura Conci

Universidade Federal Fluminense

NITERÓI
2009

Ficha Catalográfica elaborada pela Biblioteca da Escola de Engenharia e Instituto de Computação da UFF

S586 Silva, Lincoln Faria da.

Distinção automática de texto impresso e manuscrito em uma imagem de documento. / Lincoln Faria da Silva. – Niterói, RJ : [s.n.], 2009.

100 f.

Orientador: Aura Conci.

Dissertação (Mestrado em Computação) - Universidade Federal Fluminense, 2009.

1. Mineração de dados (Computação). 2. Análise de documento. 3. Visão computacional. 4. Computação visual. I. Título.

CDD 005.74

Lincoln Faria da Silva

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre. Área de concentração: Computação Visual e Interfaces.

Aprovada por:

Prof^a. Aura Conci – IC/UFF (Presidenta)

Prof^a. Débora Christina Muchaluat Saade – TET/UFF

Prof. Creto Augusto Vidal – LIA/UFC

Niterói, 20 de março de 2009

Ao Deus soberano, criador de todo o universo. Esteve sempre ao meu, deu-me capacidade e sabedoria necessária para o desenvolvimento deste trabalho.

Tudo o que tenho e tudo o que sou devo a Ele.

Agradecimentos

A Aura Conci, que depositou em mim sua confiança e compartilhou comigo sua experiência e conhecimento, profissional e acadêmico, na tarefa de orientar-me. Tenho aprendido muito com ela.

A minha noiva, Bianca dos Santos Maciel, pela compreensão, paciência e apoio durante todo o tempo, e pela ajuda na revisão do texto.

Quero agradecer também a minha mãe, Maria das Graças Faria da Silva, e minha avó, Vitória Ghiotti Faria (em memória), que me educaram, superando situações adversas.

E, por fim, quero agradecer a todos os familiares, amigos e professores que, direta ou indiretamente, me ajudaram e me apoiaram, em especial a minha tia Glacy e ao meu tio Ilton.

Resumo

As metodologias de reconhecimento de texto manuscrito e texto escrito por máquina são totalmente diferentes. Por isso, é importante separar esses dois tipos de texto, em imagens de documentos nas quais eles aparecem juntos, antes de enviá-los aos seus respectivos sistemas de reconhecimento. Documentos que apresentam texto manuscrito e texto impresso, concomitantemente, não são poucos, sendo alguns deles: formulários, cartas, requerimentos, memorandos, envelopes postais e cheques bancários.

O trabalho aqui desenvolvido executa essa separação por meio de regras de classificação mineradas, na fase de treinamento, de um conjunto de dados, os quais representam as características de cada tipo de texto. Primeiramente, a imagem é pré-processada por várias técnicas com a finalidade de: eliminar ruídos, separar o texto do fundo, retirar linhas horizontais e suavizar os contornos verticais das letras das palavras. Em seguida, é realizada a extração de componentes conectados e cada um desses é cercado pelo menor retângulo capaz de contê-lo. Retângulos próximos ou sobrepostos são unidos de modo a formarem palavras. Dos retângulos já unidos são calculadas as características pré-definidas. Essas características têm a função de representar a palavra dentro de cada retângulo e apresentam valores diferentes para as impressas e as manuscritas. Os valores são usados nas regras de classificação, as quais decidem se um retângulo contém uma palavra impressa ou manuscrita. O sistema desenvolvido é testado em duas bases de imagens: na AIM off-line Database 3.0 e na base de imagens de formulários cadastrais criada durante este trabalho e disponível na Internet. Na primeira, a acurácia e a precisão do sistema foi de 100% em 45% das imagens, com acurácia média de 97,55% e precisão média de 96,70% em relação às palavras impressas e com acurácia média de 98,09% e precisão média de 98,10% em relação às manuscritas. Na base criada como parte do trabalho, a acurácia e a precisão do sistema foi

de 100% em 33,33% das imagens, com acurácia média de 97,17% e precisão média de 98,85% em relação às palavras impressas e com acurácia média de 99,46% e precisão média de 98,75% em relação às manuscritas. O trabalho desenvolvido apresenta vantagens quando comparado com outro trabalho, que também utilizou para testes a base de imagens AIM off-line Database 3.0. O sistema foi implementado em C++ e compilado usando o GCC. Ele foi executado em uma máquina equipada com o processador AMD Athlon™ MP 900Mhz consumindo 74 segundos, em média, para realizar 184,04 bilhões de instruções no processamento de cada imagem.

Palavras-chaves: Mineração de Dados, análise de documento, identificação de texto, reconhecimento óptico de caractere, Visão de Máquina.

Abstract

The printed text and handwriting recognition methods are totally different. That is why, it is important to separate those two text types, which appear together in a document image, before sending them to their respective recognition systems. The number of documents which present printed text and handwriting, simultaneously, is significant, for example: Forms, letters, requirements, memorandums, envelopes you post and bank checks.

The separation process proposed in this work uses classification rule mining, on the training phase, of a data set, which represent the characteristics of each type of text. Initially, the image is preprocessed by applying different techniques aimed at: Eliminating noises, separating the text from the background, removing horizontal lines, and smoothing the vertical contours of the words' characters. Then, the extraction of connected components is performed and, for each connected component identified, a bounding rectangle is defined. Neighboring or overlapping bounding rectangles are united in order to form words. Predefined characteristics are computed from the already united rectangles. Those characteristics have the function of representing the word within each rectangle and they present values which are different for printed and handwriting words. Those values are used in the classification rules which decide if a given rectangle contains a printed or a handwritten word. The developed system is applied to two image databases: AIM off-line Database 3.0 and cadastral forms image database constructed simultaneously with this work and available on the Internet. On first, the system's accuracy and precision was of 100% in 45% of the images, with average accuracy of 97.55% and average precision of 96.70% in relation to printed words and with average accuracy of 98.09% and average precision of 98.10% in relation to handwritten. On database constructed as part of this work, the system's accuracy

and precision was of 100% in 33.33% of the images, with average accuracy of 97.17% and average precision of 98.85% in relation to printed words and with average accuracy of 99.46% and average precision of 98.75% in relation to handwritten. This work presents advantages when compared with another work, which also used the AIM off-line Database 3.0 as its test database. The system was implemented using C++ and compiled with GCC. It was carried out in a machine equipped with the AMD Athlon™ MP 900Mhz processor consuming 74 seconds, on the average, in order to perform 184.04 billions of instructions in the process of each image.

Keywords: Data Mining, document analysis, text identification, optical characters recognition, Machine Vision.

Glossário

AIM – Institute of Computer Science and Applied Mathematics

API – Application Programming Interface

ARFF – Attribute-Relation File Format

DCBD – Descoberta de Conhecimento em Bancos de Dados

GCC – GNU Compiler Collection

IDE – Integrated Development Environment

OCR – Optic Character Recognition

VC – Visão Computacional

WEKA – Waikato Environment for Knowledge Analysis

Sumário

Lista de Figuras	xi
Lista de Tabelas	xiii
Lista de Algoritmos	xiii
1. INTRODUÇÃO	1
1.1 Motivação	1
1.2 Metodologia proposta	2
1.3 Organização da dissertação	4
2. FUNDAMENTAÇÃO TEÓRICA	5
2.1 Visão Computacional	5
2.1.1 Principais etapas	6
2.1.1.1 Aquisição da imagem	7
2.1.1.2 Pré-processamento	8
2.1.1.3 Segmentação	11
2.1.1.4 Extração de característica	13
2.1.1.5 Classificação e reconhecimento	13
2.1.1.6 Decisão	14
2.2 Filtros espaciais de mediana e de Prewitt	15
2.3 Limiarização	18
2.4 Extração de componentes conectados	22
2.4.1 Vizinhanças de um pixel	23
2.4.2 Conectividade entre pixels	23
2.4.3 Rotulação de componentes conectados	24
2.5 Morfologia Matemática em imagens binárias	25
2.5.1 Operações morfológicas	27

2.5.1.1	Algumas definições da Teoria dos Conjuntos	27
2.5.1.2	Erosão	28
2.5.1.3	Dilatação	30
2.5.1.4	Abertura	32
2.6	Descoberta de Conhecimento em Bancos de Dados e Mineração de Dados	34
2.6.1	Preparação de dados	35
2.6.2	Mineração de Dados	36
2.6.3	Pós-processamento	39
2.6.4	Classificação	39
2.6.5	Avaliadores da classificação	41
2.6.6	Validação Cruzada	43
3.	TRABALHOS ANTERIORES	45
4.	METODOLOGIA PROPOSTA	57
4.1	Tipo documento analisado pelo sistema	57
4.2	Pré-processamento da imagem	58
4.2.1	Redução de ruídos por filtragem espacial	58
4.2.2	Binarização	60
4.2.3	Linhas horizontais no formulário	62
4.2.4	Eliminação de ruídos e suavização de contornos verticais por abertura morfológica	63
4.3	Extração de componentes conectados.....	64
4.4	União dos componentes conectados em palavras	65
4.5	Características extraídas	67
4.5.1	Desvio da Largura, Desvio da Altura e Desvio da Área	68
4.5.2	Densidade	69
4.5.3	Variância da Projeção Vertical	70

4.5.4	Maior Diferença Encontrada na Projeção Horizontal	71
4.5.5	Distribuição de Pixels	72
4.5.6	Divisão da Linha Inferior de Pixels pela Largura	74
4.5.7	Soma das Divisões de Pixels de Cada Linha pela Largura	75
4.5.8	Divisão do Maior Contorno Vertical pela Altura	75
4.5.9	Divisão da Soma dos Comprimentos dos Contornos Verticais pela Área	77
4.6	Classificação do sistema	78
5.	TREINAMENTO, TESTES E RESULTADOS	80
5.1	Treinamento do sistema	80
5.2	Bases de dados utilizadas para testes	83
5.3	Resultados	87
6.	CONCLUSÕES	91
6.1	Trabalhos futuros	94
	Referências Bibliográficas	96
	Apêndice A	
	Apêndice B	

Lista de Figuras

Figura 2.1. Etapas de um sistema de VC genérico [Conci et al., 2008]	6
Figura 2.2. Filtragem espacial de uma imagem digital	9
Figura 2.3. Etapas no processamento de imagens no domínio da frequência.....	10
Figura 2.4. Segmentação baseada em descontinuidade	12
Figura 2.5. Segmentação baseada em similaridade.....	12
Figura 2.6. Aplicação do filtro de mediana	16
Figura 2.7. Imagem com regiões de tonalidades diferentes	17
Figura 2.8. Perfil de tonalidade da imagem na Figura 2.7.....	17
Figura 2.9. Perfil da derivada da imagem na Figura 2.7	17
Figura 2.10. Filtros de Prewitt vertical e horizontal	18
Figura 2.11. Filtragem por filtros de Prewitt	18
Figura 2.12. Imagem em tons de cinza e seu respectivo histograma	19
Figura 2.13. Resultado da limiarização	20
Figura 2.14. Vizinhanças de pixels	23
Figura 2.15. Extração de componentes conectados.....	25
Figura 2.16. Imagem binária	27
Figura 2.17. Erosão de uma imagem binária	30
Figura 2.18. Dilatação de uma imagem binária.....	32
Figura 2.19. Abertura de uma imagem binária	34
Figura 2.20. Fases e etapas da DCBD	35
Figura 2.21. Modelo de classificação	38
Figura 2.22. Construção e aplicação do modelo de classificação	41
Figura 3.1. Etapas de metodologias de distinção de texto impresso e manuscrito	45
Figura 4.1. Formulário com manuscrito separado do texto impresso	58

Figura 4.2. Formulário de cadastro	58
Figura 4.3. Filtragem espacial por filtro de mediana	59
Figura 4.4. Binarização de uma imagem de formulário	61
Figura 4.5. Linhas horizontais extraídas	63
Figura 4.6. Linha horizontal ignorada	63
Figura 4.7. Eliminados ruídos por abertura morfológica.....	63
Figura 4.8. Suavização de bordas verticais	64
Figura 4.9. Elemento estruturante	64
Figura 4.10. Distância entre dois retângulos envoltórios	66
Figura 4.11. União de retângulos envoltórios da Figura 4.10.....	66
Figura 4.12. União de retângulos envoltórios com interseção de áreas.....	67
Figura 4.13. Largura, altura e área de um retângulo envoltório	68
Figura 4.14. Coordenadas de um retângulo envoltório	69
Figura 4.15. Retângulos envoltórios envolvendo uma palavra impressa e outra manuscrita ..	69
Figura 4.16. Projeção vertical de uma palavra impressa	70
Figura 4.17. Projeção vertical de uma palavra manuscrita	70
Figura 4.18. Perfil da projeção vertical de uma palavra manuscrita	71
Figura 4.19. Projeção horizontal de uma palavra impressa	71
Figura 4.20. Projeção horizontal de uma palavra manuscrita	71
Figura 4.21. Divisão de retângulo envoltório com palavra manuscrita	72
Figura 4.22. Divisão de retângulo envoltório com palavra impressa	73
Figura 4.23. Linha inferior de pixels de uma palavra manuscrita	74
Figura 4.24. Linha inferior de pixels de uma palavra impressa	74
Figura 4.25. Contornos verticais em palavras impressas	75
Figura 4.26. Contornos verticais em palavras manuscritas	75
Figura 4.27. Filtros espaciais para detecção dos contornos verticais	76

Figura 4.28. Maior contorno vertical em uma palavra impressa	76
Figura 4.29. Maior contorno vertical em uma palavra manuscrita	76
Figura 5.1. Exemplo de arquivo ARFF	81
Figura 5.2. Título e identificação de formulário da base AIM	84
Figura 5.3. Porção de texto impresso de formulário da base AIM	84
Figura 5.4. Porção de texto manuscrito de formulário da base AIM	84
Figura 5.5. Identificação do escritor do formulário na base AIM	84
Figura 5.6. Linha base para manuscritos	85
Figura 5.7. Exemplo de arquivo <i>ground truth</i>	87

Lista de Tabelas

Tabela 2.1. Tarefas e técnicas da Mineração de Dados	37
Tabela 4.1. Resumo das características extraídas	77
Tabela 5.1. Regras mineradas de um determinado conjunto de treinamento	82
Tabela 5.2. Características mais significativas na classificação	83
Tabela 5.3. Resultados do teste na base de dados AIM-DB v.3	88
Tabela 5.4. Resultados do teste na base de imagens de formulários de cadastro	99

Lista de Algoritmos

Algoritmo 4.1. Método de Otsu	61
Algoritmo 4.2. Eliminando linhas horizontais	62
Algoritmo 4.3. Extração de componentes conectados	65