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

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

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:

_____

Profª. Aura Conci – IC/UFF (Presidenta)

_____

Profª. 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 Iltom.

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 handwritted 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

Lista de Tabelas

Lista de Algoritmos