



XXV Congresso de Iniciação Científica da UNESP

I Fórum Internacional de Iniciação Científica da UNESP

Reconhecimento Óptico de Caracteres



Helder Cesar R. de Oliveira, Prof. Dr. Marco Antônio Piteri

Unesp/FCT Câmpus de Presidente Prudente

Introdução

O reconhecimento óptico de caracteres (*Optical Character Recognition - OCR*) é uma área atual, dinâmica e, em constante processo de renovação. Considerando a necessidade de passar para o computador toda informação textual existente em forma física, seja em papel ou qualquer outro material, principalmente livros antigos e documentos raros que estão sujeitos as mais diversas formas de deterioração, uma alternativa é digitalizar essas informações, para que possam ser guardadas com mais segurança sem qualquer risco e, também para que possam ser compartilhadas com outras pessoas fazendo com que todos possam ter acesso ao documento. O processo de digitalização de um material, faz com que o mesmo se torne um arquivo (de imagem) no computador, sendo necessário transformá-lo, "a posteriori", num arquivo de texto, de modo que o mesmo possa ocupar um menor espaço de memória, além de poder ser editável, ver Figura 1.



a) Página de jornal digitalizada (~1Mb).

b) Texto associado à página digitalizada (~10kb).

Figura 1: Imagens ocupam mais espaço de armazenamento se comparado ao texto ASCII correspondente.

É importante realçar ainda que reconhecimento de textos impressos não é uma tarefa trivial, na medida em que existem diferentes tipos de fontes. Por sua vez, o reconhecimento de textos manuscritos é mais difícil ainda, já que a caligrafia é uma característica biométrica. Em outras palavras, cada pessoa tem um estilo único de escrita.

Objetivos

O objetivo primário do projeto é estudar e implementar a metodologia que envolve o processo de OCR para o reconhecimento de caracteres impressos e o reconhecimento de dígitos decimais manuscritos, que tipicamente inclui etapas, como: suavização, binarização, normalização, segmentação, extração das características e classificação, conforme pode ser observado na Figura 2.

Material e Método

O desenvolvimento do protótipo está sendo feito na linguagem C++, juntamente com o apoio da biblioteca Qt que será usada para a construção da interface e algumas manipulações básicas sobre a imagem. Na implementação também será usada a biblioteca OpenCV, que é direcionada para aplicações que envolvem processamento digital de imagens e visão computacional, áreas em que o presente projeto se insere. Ambas as bibliotecas são *open source*.



Pré-processamento:

- Binarização, Segmentação, *Thinning*, *Thresholding*, Filtros Lineares/Não Lineares;
- Filtro Gaussiano, Filtro Laplaciano, Gradiente, ...

Análise do Layout:

- Detecção/Correção da Inclinação das Linhas;
- Normalização Baseada em Momento, Normalização Não Linear, Análise do Contorno, ...

Extração das Características:

- *Chain Code* (Fig. 3a), Matriz de Co-ocorrência (Fig. 3b), Descritores de Fourier, Curvas de Bézier (Fig. 3d);
- Transformada de Hough, *Template Matching* (Fig. 3c);
- Análise dos Componentes Principais (PCA), Histogramas (Fig. 3e) ...

Classificação dos Caracteres:

- K-Vizinhos Mais Próximos, Teoria Bayesiana, Classificadores Gaussiano, *Support Vector Machines*;
- Redes Neurais, Distância Euclidiana, ...

Figura 2: Pipeline associado a um sistema OCR.

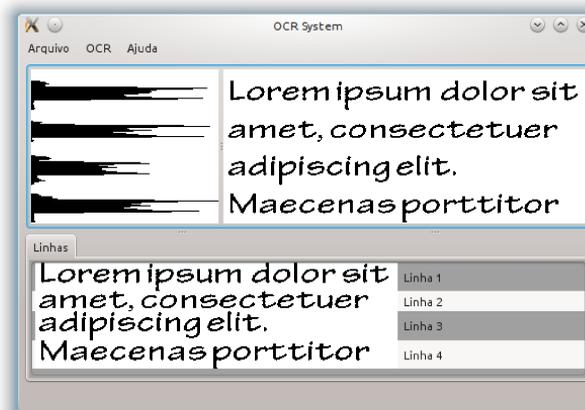
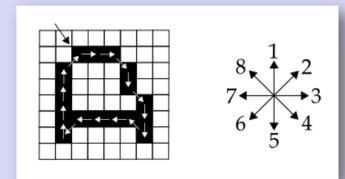
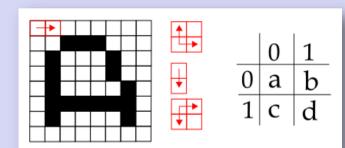


Figura 4: Screenshot do protótipo ilustrando a projeção horizontal das linhas de um texto.

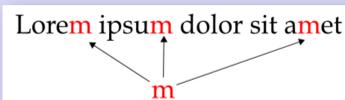


a) Chain Code.

Código associado: 33454558777611112.



b) Matriz de Co-ocorrência.



c) Template Matching.



d) Curvas de Bézier.



e) Histogramas.

Figura 3: Algoritmos de Extração de Características.

Conclusões

O projeto está sendo elaborado no contexto de um Trabalho de Conclusão de Curso (TCC) que tem previsão de ser concluído até o final do corrente ano letivo e encontra-se atualmente, na fase quatro, relativa à implementação do protótipo (Fig. 4). Para ser aceitável, espera-se, com a conclusão do trabalho, que a precisão no reconhecimento dos caracteres impressos seja superior a 95%.

Bibliografia

1. Gonzalez, R.; Woods, R.; *Digital image processing*. Prentice Hall, 2008.
2. Parker, J. R.; *Algorithms for image processing and computer vision*. Wiley Computing, 2010.
3. Cheriet, M.; Kharna, N.; Liu, C.; *Character recognition systems: a guide for students and practioners*. Wiley-Interscience, 2007.
4. Miranda, R. A.; Silva, F. A.; Artero, A. O.; Piteri, M. A.; *Handwritten Character Recognition based on Frequency, Character-edge Distances and Densities*; Anais do WVC 2013 - IX Workshop de Visão Computacional. Rio de Janeiro, 2013.