Orientação de Documentos Digitalizados

Proposta do Trabalho de Graduação em Computação Visual

Junho de 2004

Universidade Federal de Pernambuco
Ciências da Computação

Orientador: Rafael Dueire Lins
Aluno: Bruno Tenório Ávila

Índice

1. Introdução

A quantidade de informações geradas nas empresas multiplica-se a cada dia e surge a necessidade de organizá-las de maneira que se possa localizar a informação de forma eficiente e eficaz. Uma grande porcentagem do capital intelectual está na forma de documentos físicos, ou seja, informações desestruturadas que dificilmente podem ser utilizadas para gerar conhecimento. Além de estarem sujeitos ao desgaste e a perda do papel.

A digitalização é uma solução comum para organizar os documentos físicos. Contudo, é a fase mais lenta do processo de estruturação das informações. Os papéis são convertidos em imagens, e então, são organizados em caixas e indexados em prateleiras. As imagens são salvas em dispositivos de armazenamento. Um sistema de Gerenciamento Eletrônico de Documentos (GED) é utilizado para controlar o acesso às informações podendo ser disponibilizado tanto na Intranet da empresa quanto na Internet.

Uma vez que o volume de papel é gigantesco, faz-se uso de imagens em preto e branco para diminuir o tamanho da imagem, além de uma resolução reduzida. Scanners de alta produção são utilizados para o processo de digitalização. Por causa disto, a qualidade das imagens são afetadas e podem aparecer ruídos, documentos desalinhados e com bordas. Isto prejudica não só a visibilidade quanto o tamanho de armazenagem da imagem.

A melhoria da qualidade e da produtividade do projeto de digitalização de média e alta produção melhora a satisfação do cliente e da empresa prestadora do serviço. O processamento do documento digitalizado melhora a qualidade da imagem, bem como, reduz o tamanho de armazenamento. O reconhecimento dos caracteres torna-se viável e, a partir disso, permite a indexação automática dos documentos, além de agilizar a busca por palavras no conteúdo.

2. Objetivo

O principal objetivo deste trabalho é estudar o estado da arte de algoritmos para orientação de documentos digitalizados e propor novos métodos, além de realizar um teste comparativo entre os procedimentos existentes.

A orientação de documentos digitalizados envolve a análise de documentos de diferentes formatos, línguas, cores, estado de degradação e resoluções da imagem e então, localizar características na imagem que possibilite a detecção automática da orientação do documento.

Esta é uma fase crítica para o reconhecimento de caracteres (OCR). Estudos realizados [26] mostram que os softwares de OCR são sensíveis ao desalinhamento do documento.

Este é um problema clássico na análise de documentos digitalizados e possui uma vasta literatura [1-25]. Vale a pena ressaltar que o documento pode estar rotacionado de 0º a 360º graus e um método de orientação deve ser capaz de detectar a rotação do documento em qualquer ângulo.

3. Metodologia

A velocidade e a precisão são fatores críticos a serem levados em conta para a formulação dos algoritmos, isto devido ao volume de imagens a serem processados e devido a sensibilidade dos softwares de OCR.

A execução do trabalho desta pesquisa envolverá o desenvolvimento de uma biblioteca com os algoritmos implementados. Para isto será necessário:

- Estudar algoritmos existentes para pré e pós-processamento;
- Estudar algoritmos existentes de orientação de documentos digitalizados;
- Desenvolver novos algoritmos de orientação de documentos digitalizados;
- Realizar benchmark dos algoritmos desenvolvidos e já existentes para validação e comparação dos métodos.

4. Equipe Técnica e Respectivas Atividades

Os membros da equipe técnica para o desenvolvimento deste trabalho são:

- Rafael Dueire Lins – Professor do Departamento de Eletrônica e Sistemas da UFPE.

  Atividades: Orientação do estudo.

- Bruno Tenório Ávila – Aluno do curso de graduação de Bacharelado em Ciências da Computação da UFPE.

  Atividades: Estudo e desenvolvimento de algoritmos;

  Implementação e verificação dos algoritmos;

  Desenvolvimento de softwares específicos;

  Elaboração de relatórios.

5. Cronograma de Atividades

O cronograma das atividades obedecerá ao especificado na tabela a seguir:

Descrição	**Junho**				**Julho**				**Agosto**			
	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**
Estudo do estado da arte												
Desenvolvimento de novos algoritmos												
Implementação de algoritmos existentes												
Realização do benchmark												
Desenvolvimento do relatório												
Preparar apresentação oral												

6. Resultados Esperados

O desenvolvimento e validação de novos algoritmos de detecção automática de orientação de documentos digitalizados. A construção de uma biblioteca com os algoritmos estudados e propostos de modo a possibilitar um benchmark dos diversos métodos de orientação de documentos digitalizados segundo a performance e precisão.

7. Referências Bibliográficas

[1] W. Postl. Detection of linear oblique structures and skew scan in digitized documents. In *Proc. of the 8th International Conference on Pattern Recognition*, páginas 687-689, Paris, França, 1986.

[2] H.S. Baird. The skew angle of printed documents. In Proc. of the Conference Society of Photographic Scientists and Engineers, volume 40, páginas 21-24, Rochester, Nova Iorque, Maio, 20-21 1987.

[3] G. Ciardiello, G. Scafuro, M.T. Degrandi, M.R. Spada, e M.P. Roccotelli. An experimental system for office document handling and text recognition. In *Proc. of the 9th International Conference on Pattern Recognition,* volume 2, páginas 739-743, Roma, Italy, Novembro, 14-17 1988.

[4] Y. Ishitani. Document skew detection based on local region complexity. In *Proc. of the 2nd International Conferece on Document Analysis and Recognition*, páginas 49-52, Tsukuba, Japão, outubro de 1993. IEEE Computer Society.

[5] A. Bagdanov e J. Kanai. Projection profile based skew estimation algorithm for JBIG compressed images. In *Proc. of the 4th International Conference on Document Analysis and Recognition*, páginas 401-405, Ulm, Alemanha, agosto de 1997.

[6] S.N. Srihari e V. Govindaraju. Analysis of textual images using the Hough Transform. *Machine Vision and Applications*, 2(3):141-153, 1989.

[7] S. Hinds, J. Fisher, e D. D'Amato. A document skew detection method using run-length encoding and the Hough Transform. In *Proc. of the 10th International Conferente on. Pattern Recognition,* páginas 464-468, Atlantic City, NJ, junho, 17-21 1990.

[8] A.L. Spitz. Skew determination in CCITT Group 4 compressed document images. In *Proc.. of the Symposium on Document Analysis and Information Retrieval*, páginas 11-25, Las Vegas, 1992.

[9] D.S. Le, G.R. Thoma e H. Wechsler. Automated page orientation and skew angle detection for binary document images. *Pattern Recognition*, 27(10):1325-1344, 1994.

[10] Y. Min, S.-B. Cho e Y. Lee. A data reduction method for efficient document skew estimation based on Hough Transformation. ln *Proc. of the 13th International Conference on Pattern Recognition*, páginas 732-736, Vienna, Áustria, agosto de 1996. IEEE Press.

[11] U. Pal e B.B. Chaudhuri. An improved document skew angle estimation technique. *Pattern Recognition Letters*, 17(8):899-904, julho de 1996.

[12] B. Yu e A.K. Jain. A robust and fast skew detection algorithm for generic documents. *Pattern Recognition*, 29(10):1599-1629, 1996.

[13] A. Hashizume, P.S. Yeh e A. Rosenfeld. A method of detecting the orientation of aligned components. *Pattern Recognition Letters*, 4:125-132, 1986.

[14] L. O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 15(11):1162-1173, 1993.

[15] R. Smith. A simple and efficient skew detection algorithm via text row accumulation. In *Proc. of the 3th Interrmational Conference on Document Analysis and Recognition*, páginas 1145-1148, Montreal. Canadá, agosto de 1995.

[16] T. Akiyama e N. Hagita. Automated entry system for printed documents. *Pattern Recognition*, 23(11):1141-1154, 1990.

[17] H. Yan. Skew correction of document images using interline cross-correlation.

CVGIP: Graphical Modela and Image Processing. 55(6):538-543, novembro de 1993.

[18] B. Gatos, N. Papamarkos e C. Chamzas. Skew detection and text line position determination in digitized documents. *Pattern Recognition*. 30(9):1505-1519, 1997.

[19] J. Sauvola e M. Pietikäinen. Skew angle detection using textura direction analysis. In *Proc. of the 9th Scandinavian Conference on Image Analysis*, páginas 1099-1106, Uppsala, Suiça, junho de 1995.

[20] C. Sun e D. Si. Skew and slant correction for document images using gradient direction. In *Proc. of the 4th International Conference on Document Analysis and Recognition*, páginas 142-146, Ulm, Alemanha, agosto de 1997.

[21] S. Chen e R.M. Haralick. An automatic algorithm for text skew estimation in document images using recursive morphological transforms. In *Proc. of the 1st IEEE International Conference on Image Processing*, páginas 139-143, Austria, Texas, 1994.

[22] H. K. Aghajan, B. H. Khalaj, e T. Kailath. Estimation of skew angle in text-image analysis by SLIDE: subspace-based line detection. *Machine Vision and Applications*, 7:267-276, 1994.

[23] A. Amin e S. Fischer. A Document Skew Detection Method Using the Hough Transform. Pattern Analysis & Applications. Springer-Verlag London Ltd, volume 3, número 3, setembro de 2000, páginas 243-253.

[24] J. Liu, C-M. Lee e R-B. Shu. An efficient method for the skew normalization of document image.

[25] H.S. Baird. Anatomy of a versatile page reader. In *Proc. of the IEEE*, volume 80, número 7, páginas 1059-105, 1992.

[26] N. F. Alves. Estratégias para melhoria do desempenho de ferramentas comerciais de reconhecimento óptico de caracteres. Tese de mestrado do Departamento de Eletrônica e Sistemas da UFPE, Recife/PE, Brasil, 2003.

8. Assinaturas

Recife, 7 de junho de 2004

_____

Rafael Dueire Lins

_____

Bruno Tenório Ávila