COBEM 99
XV Congresso Brasileiro de Engenharia Mecânica
15th Brazilian Congress of Mechanical Engineering
22 - 26 de Novembro de 1999 / November 22 - 26, 1999 Águas de Lindóia, São Paulo,

TÉCNICA PARA SEPARAÇÃO DE CARACTERES SOBREPOSTOS

Silvana A. G. da Silva
Escola Politécnica da Universidade de São Paulo, Departamento de Engenharia Mecânica
Av. Prof. Mello Moraes, 2231, 05508-900, São Paulo, SP, Brasil, e-mail: silvanaa@usp.br
Iara K. Ike
Jun Okamoto Jr
Cinthia Itiki

Resumo. *O objetivo de um algoritmo de separação de caracteres é dividir a imagem de uma palavra em regiões, cada uma delas contendo um caracter isolado e completo. A maior dificuldade está nos caracteres sobrepostos, pois o método empregado deve distinguir com perfeição o que realmente faz parte de um caracter e não de outro. A complexidade é, também, devido à grande variedade de fontes, estilos de textos e características das imagens. Os métodos mais comuns de separação de caracteres são baseados em histogramas. Alguns utilizam o histograma propriamente dito enquanto outros utilizam um histograma vertical, que é uma representação gráfica do número de pixels pretos em cada coluna da imagem binária. Alguns, ainda, baseiam-se no histograma vertical envolvente, que é a forma gráfica de representar o conjunto dos pixels entre as letras na imagem. Outros métodos baseiam-se na análise da vizinhança dos pixels. Este artigo apresenta uma técnica para separação de caracteres sobrepostos ou não, baseada no histograma envolvente, no crescimento da região pela agregação de pixels e na análise da vizinhança dos pixels. Para comprovação dos resultados foi desenvolvido um algoritmo que apresentou resultados satisfatórios na separação dos caracteres sobrepostos em diversas situações.*

Palavras-chave: *Separação de caracteres, Reconhecimento de padrões, Histograma*

1. INTRODUÇÃO

O reconhecimento de caracteres impressos e manuscritos tem despertado o interesse de muitos pesquisadores devido à sua aplicação em sistemas automáticos de

reconhecimento de escrita. Estudos e técnicas diversas têm sido propostas e implementadas tanto no âmbito acadêmico como em aplicações práticas. Algumas destas técnicas baseiam-se no reconhecimento de caracteres através de imagens em níveis de cinza e outras em imagens binárias. Na maioria delas, primeiramente, faz-se a separação dos caracteres e após, o reconhecimento individual de cada um deles. Esta fase de separação é muito importante pois dela depende todo o sucesso do reconhecimento.

O objetivo do processamento automático de imagens de documentos é o reconhecimento de textos e figuras em imagens digitais e a extração de informações necessárias. Atualmente, utiliza-se um código de barras para reconhecimento de um código numérico que identifica um determinado produto ou componente. As técnicas de reconhecimento de caracteres têm a mesma função do código de barras, a vantagem desta é que o próprio número pode vir impresso no objeto a ser identificado, dispensando a decodificação das barras e facilitando a identificação visual, pois, os números, por si só, já identificam o produto, por exemplo, a empresa, o setor, o material, o tipo de produto, etc.

Uma aplicação do reconhecimento de caracteres é o direcionamento a ser dado para certas peças codificadas que se deslocam sobre uma esteira, num chão de fábrica. Suas imagens são capturadas por uma câmera e estas enviadas a um computador que rapidamente faz a leitura do código, a separação dos caracteres, o reconhecimento dos números e, com isto, decide qual o melhor destino a ser dado para estas peças.

As técnicas de reconhecimento de caracteres podem ser também aplicadas em correios, onde os códigos de endereçamento postal (CEPs) das correspondências são identificados e direcionados para um determinado local, onde serão posteriormente entregues ao destinatário.

Dependendo das posições dos caracteres na palavra, surgem algumas situações que dificultam a sua separação, são estas: a sobreposição e a conexão dos caracteres. Como área de um caracter define-se o espaço interior a um quadrado ou retângulo que envolve a letra. A Fig.1a mostra a área dos caracteres E, S, T, A ,V e A.

Um caracter sobrepõe a outro quando este ocupa parte da área do outro caracter. Isto é, um caracter está sobreposto a outro quando existe uma interseção entre as áreas, com a finalidade de aproximar as letras. Exemplo: Na palavra ESTAVA existe uma sobreposição entre as letras T, A e V, mostrado na Fig.1b.

Um caracter está conectado a outro quando existe algum tipo de ligação entre eles, uma linha por exemplo, caso comum em letras manuscritas. Neste caso, fica difícil identificar, o que faz parte, ou não, de um caracter ou de outro caracter. Na Fig.1c as letras A e V estão conectadas.

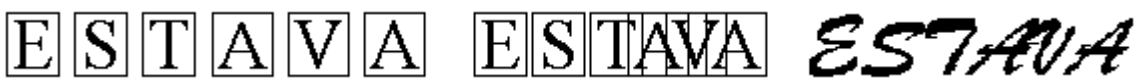

Figura 1 - Exemplo de caracteres. 1a) área de um caracter, 1b) caracteres sobrepostos, 1c) caracteres conectados.

Existem técnicas específicas para solucionar os problemas de sobreposição e de conexão. A técnica proposta soluciona o problema de caracteres maiúsculos sobrepostos. Para testar o algoritmo desenvolvido foram adquiridas imagens a partir de jornais e

revistas. Os tipos de letras usados foram com e sem serifa (pequeno traço, ou, às vezes, simples espessamento, que remata, de um ou de ambos os lados, os terminais das letras não lineais de caixa-alta e caixa-baixa, e que pode ter a forma de filete, barra, etc.).

O item 2 faz uma breve descrição de algumas técnicas de separação de caracteres, o item 3 descreve o método proposto. Seguem a análise dos resultados e as conclusões deste trabalho.

2. ALGUMAS TÉCNICAS EXISTENTES PARA SEPARAÇÃO DE CARACTERES

As técnicas de separação de caracteres diferem-se pelo tipo dos caracteres, estes podem ser escritos a máquina ou manuscritos. Yi Lu (1995) descreve algumas destas técnicas para caracteres impressos. Estas abrangem caracteres uniformemente espaçados, quebrados, sobrepostos ou conectados e baseiam-se nas características da imagem e nos resultados de reconhecimento. Existe uma técnica apropriada para cada um destes casos.

A técnica baseada na projeção vertical é aplicada para caracteres uniformemente espaçados. Projeção Vertical V(x) é o histograma obtido pela contagem do número de *pixels* pretos da linha para cada coluna. Se os caracteres estão bem separados, então V(x) terá valores zeros entre os caracteres. Yi Lu *et al* (1992) desenvolveu duas técnicas baseadas nas projeções multi-linhas: em ambas, o primeiro passo é agrupar as linhas do texto de acordo com suas alturas, estes grupos devem possuir espaçamento uniforme entre as letras. Então, a projeção vertical é feita em cada grupo de linhas. A projeção vertical multi-linha é obtida pela adição dos *pixels* de cada coluna. O intervalo de zeros nesta fase corresponde aos *gaps* entre os componentes horizontais na imagem binária.

Existem duas técnicas para separação de caracteres quebrados: a primeira, usa um processo de fusão baseado na largura e nos intervalos do caracter estimado. O algoritmo primeiro separa uma linha texto em regiões, baseado na projeção vertical. Através de estatísticas, faz o cálculo das larguras e dos intervalos sobre estas regiões. A decisão está baseada no número de regiões em que a largura destas e do intervalo são menores que um dado limiar. A segunda técnica utiliza a combinação dos componentes do caracter baseando-se nos resultados de reconhecimento.

A técnica para separação de caracteres sobrepostos consiste em detectar a área de sobreposição e analisar a estrutura do componente vizinho. Os dois componentes vizinhos são examinados por sobreposição vertical e por um possível agrupamento entre eles. Se o grau de sobreposição exceder a um dado limiar, então os componentes são possíveis candidatos a estarem sobrepostos.

Para o caso de caracteres conectados Yi Lu (1995) sugere que a largura pode ser dinamicamente estimada durante o processo de separação. Cada candidato é examinado pela comparação de sua largura com a largura do caracter estimado ou pela medida do seu *aspect ratio* (largura dividida pela altura do caracter*)*. A técnica é gerar modelos de perfis do caracter. Estes, são construídos pelo contorno suave de sua projeção vertical e esta, tem pelo menos dois picos significantes. Assim, existe somente uma opção para o caracter. Os fatores usados na construção do modelo são o número, a altura e a extensão dos picos, a largura e a profundidade dos vales e a simetria das projeções das curvas.

Yi Lu (1996) apresenta algumas técnicas para separação de caracteres manuscritos. No caso de letras de forma, os caracteres podem estar encaixotados, espaçados ou

conectados. Estes últimos apresentam maiores dificuldades para separação, pois o ponto de contato entre os caracteres pode estar em qualquer posição da letra e o limite de separação pode ser não linear. Outras referências para reconhecimento de caracteres podem ser encontradas em (Mori,1992), (Simon,1992) e (Suen,1980). Os algoritmos mais comuns baseiam-se no comportamento do contorno e na transformação de distância para encontrar o melhor caminho para separar os caracteres conectados.

No caso de palavras cursivas, os algoritmos consistem de dois passos: a pré-segmentação e a segmentação. Na pré-segmentação removem-se os ruídos e faz-se a correção da inclinação. Nesta fase assume-se duas aproximações: uma assegura que cada segmento contenha não mais do que um caracter e a outra que cada segmento contenha ao menos um ou mais caracteres. Na segmentação existe o algoritmo de R. M. Bozinovic e Srihari (1989) e a técnica de segmentação usando singularidades e regularidades proposta por M. Y. Chen (1992).

3. DESCRIÇÃO DO MÉTODO PROPOSTO

A idéia básica do método proposto é encontrar a região de separação dos caracteres utilizando para isto o histograma vertical envolvente. Para a construção deste histograma utilizou-se os conceitos descritos em Gonzales (1993): crescimento da região por agregação de *pixel* e análise da vizinhança e também o trabalho desenvolvido por Zhixin Shi (1997).

Crescimento da região por agregação de *pixel* segundo Gonzales (1993) é um procedimento para identificar regiões. Começa com uma série de sementes colocadas em pontos estratégicos na imagem, e a partir destas agregam-se *pixels* vizinhos que possuem propriedades similares (como nível de cinza, textura, cor, etc.). Assim há um crescimento da região até que os *pixels* vizinhos não possuam mais as mesmas características. Alguns problemas referentes a esta técnica dizem respeito à seleção do ponto onde serão colocadas as sementes. Este pode ser, por exemplo, um valor de nível de cinza, ou ainda, o *pixel* mais brilhante. Estes pontos devem estar relacionados às condições chaves características da imagem em questão.

Segundo Gonzales (1993), pode-se fazer uma análise dos *pixels* vizinhos a um *pixel* considerado para descobrir a conectividade, a indicação de componentes conectados, a relação binária e a equivalência, a distância entre as medidas e também fazer operações aritméticas e lógicas entre os *pixels*. Um *pixel* p de coordenadas (x, y) tem quatro vizinhos horizontais e verticais, cujas coordenadas são: (x + 1, y), (x - 1, y), (x, y + 1), (x, y - 1). Esta série de *pixels*, é chamada de 4-vizinhos de p. Cada *pixel* está distante 1 unidade a partir de (x, y). Os quatro vizinhos diagonais de p têm as seguintes coordenadas: (x + 1, y + 1), (x + 1, y - 1), (x - 1, y + 1), (x - 1, y - 1). Estes pontos, juntamente com os 4-vizinhos são chamados de 8-vizinhos de p.

O histograma vertical envolvente foi descrito por Zhixin Shi (1997). A técnica por ele proposta era encontrar uma linha ótima para separar a imagem em partes, cada uma delas contendo um único caracter. Nesta técnica, o histograma vertical envolvente é a forma gráfica de representar a distância entre os *pixels* pretos superiores e os inferiores das letras. Em outras palavras, é um gráfico no qual se visualiza a área interna a um perfil

envolvente das letras. Um exemplo deste histograma pode ser visto na Fig. 2 das letras A e M.

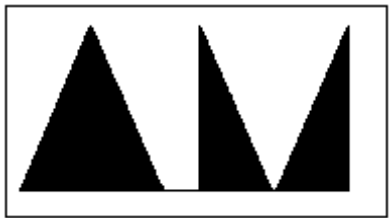

Figura 2 - Histograma Vertical Envolvente

A técnica proposta neste trabalho foi desenvolvida com base na definição deste histograma, só que neste caso, o histograma envolve o intervalo entre os caracteres, como pode ser visto na fig. 3.

Figura 3 - Histograma Vertical Envolvente do Intervalo entre os Caracteres

O objetivo do algoritmo proposto é encontrar a área de separação entre os caracteres. A lógica utilizada está descrita nos itens abaixo.

1. Fase de pré-separação dos caracteres: O objetivo desta fase é, através da projeção vertical, separar os caracteres que não estão sobrepostos e encontrar os blocos que contêm os caracteres sobrepostos.
2. Fase de separação dos caracteres sobrepostos: A condição imposta para que os blocos encontrados na fase anterior sejam considerados blocos de caracteres sobrepostos é que a largura do caracter seja maior que a sua altura. Passos:
   - encontrar uma linha horizontal imaginária posicionada na metade da altura dos caracteres;
   - encontrar, nesta linha, a seqüência de vetores formados por pixels de valor 1 (espaços em branco entre os caracteres ou no interior deste);
   - encontrar, nestes vetores, o elemento médio de cada um deles;
   - investigar o primeiro pixel superior a estes elementos médios. Se este pixel também for branco, analisar a partir dele, os pixels à direita e à esquerda até encontrar vizinhos pretos. As posições anteriores a estes pontos pretos à esquerda e à direita são armazenadas para posterior elaboração do histograma envolvente;
   - prosseguir pela mesma coluna do elemento médio inicialmente definido até encontrar a linha superior da imagem, isto ocorre se todos os pixels superiores

forem brancos. Caso seja encontrado um pixel preto nesta trajetória adota-se o primeiro pixel à direita, e se este também for preto, adota-se o primeiro pixel à esquerda como alvo de subida. Esta trajetória descrita é armazenada em um vetor caso consiga-se atingir a linha superior da imagem. Se ambos os pixels da esquerda e da direita forem pretos, então o algoritmo encontra-se no interior de uma letra e este vetor será ignorado. Este procedimento é idêntico para a investigação dos pixels vizinhos inferiores partindo da linha média imaginária até a linha inferior da imagem.

- Comparar cada vetor superior a todos os vetores inferiores. Se o primeiro elemento do vetor superior for igual ao primeiro elemento do vetor inferior então, unindo-se estes vetores temos a área de separação;

Foi estipulado um limite de h pixels para serem percorridos a esquerda e a direita caso não sejam detectados pixels pretos nestes intervalos, para tornar o sistema eficaz no caso de imagens levemente inclinadas.

4. RESULTADOS

Os resultados obtidos estão mostrados na Fig. 4 com a palavra AVALIAR que contém serifa (Courier New e Times New Roman) e na Fig. 5 com a palavra VALTER, sem serifa. Onde pode-se visualizar a palavra original binarizada, os caracteres segmentados pelo histograma inicial, nesta fase aparecem também os blocos que contêm os caracteres sobrepostos, os caracteres sobrepostos segmentados e finalmente o histograma vertical envolvente do intervalo entre os caracteres.

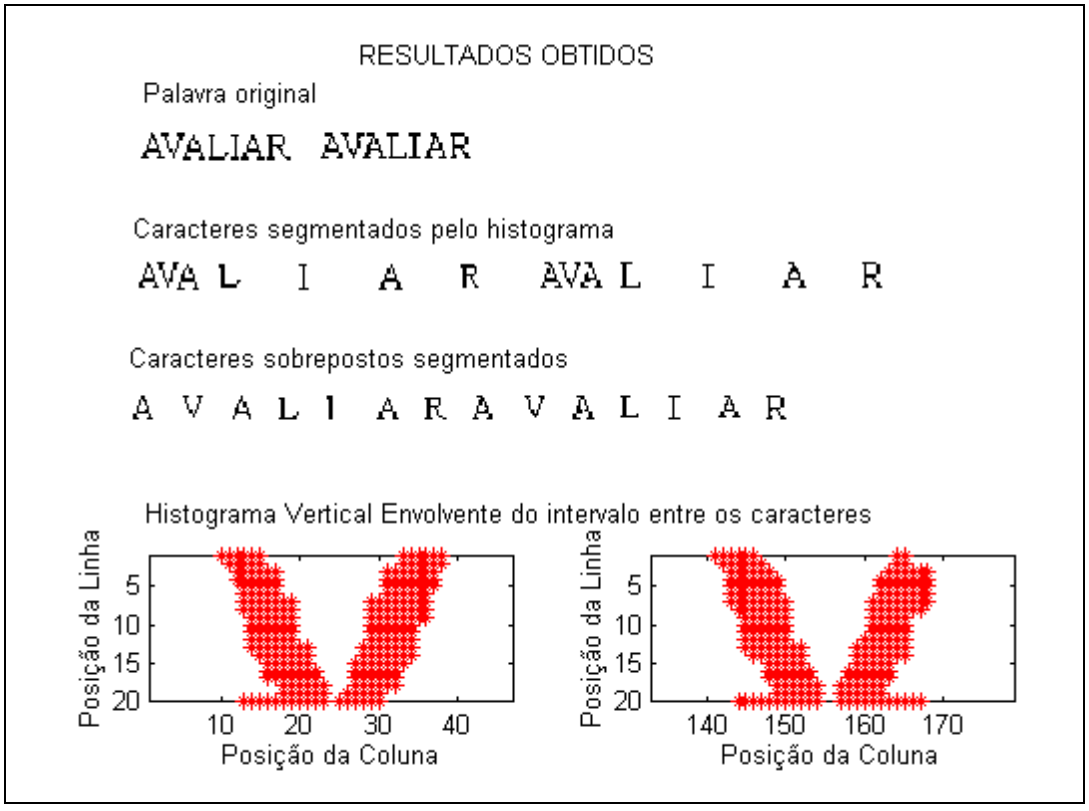

Figura 4 - Resultados Obtidos da Palavra AVALIAR

Na Fig. 4 aparecem dois histogramas verticais envolventes pois a palavra original AVALIAR aparece duas vezes e correspondem ao bloco AVA onde as letras A, V e A estão sobrepostas. O que visualiza-se nestes histogramas é o intervalo entre as letras A, V e A para a primeira e para a segunda palavra AVALIAR. Na linha 20 o histograma extravasa para as laterais, isto é devido a leve inclinação da palavra original.

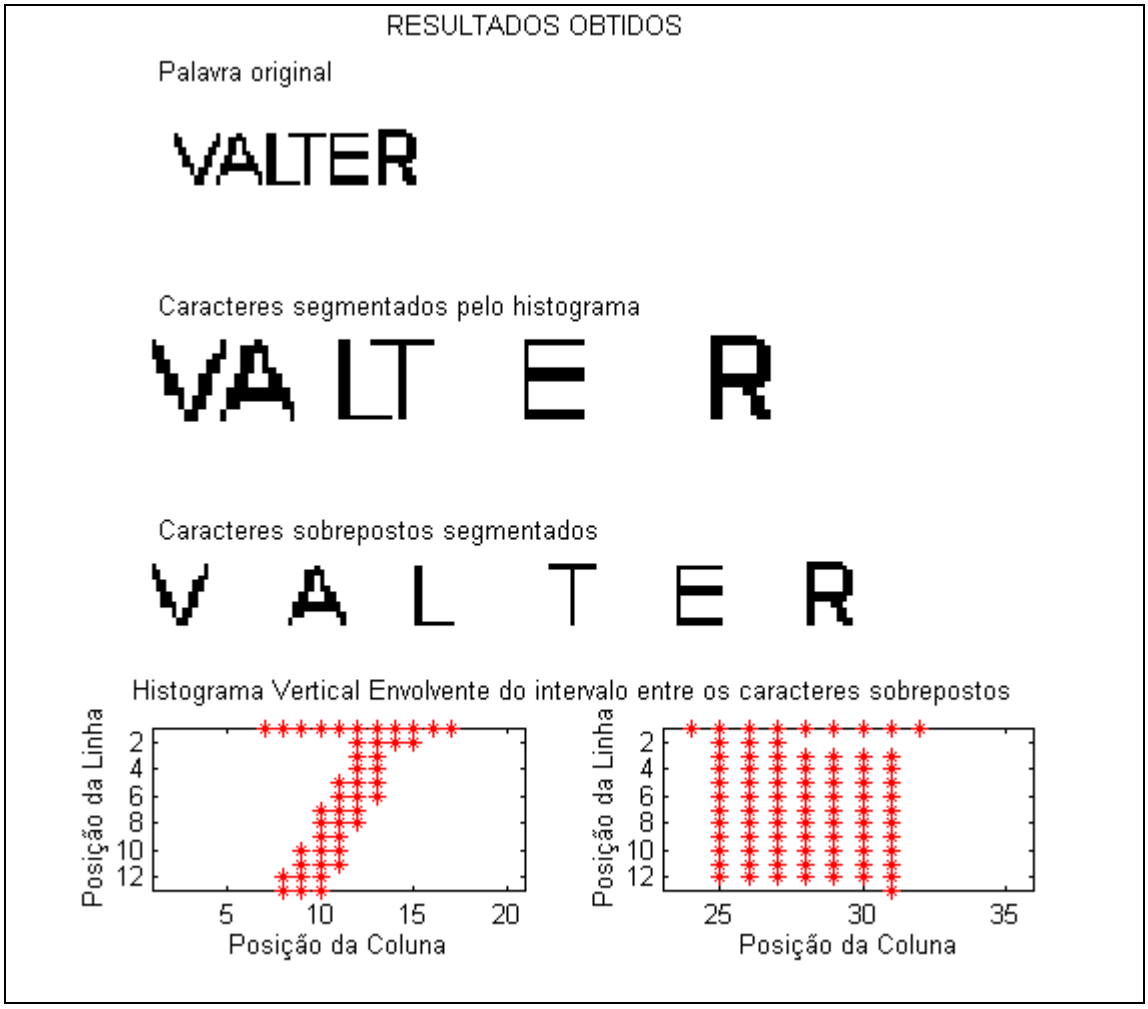

Figura 5 - Resultados Obtidos da Palavra VALTER

3. CONCLUSÃO

Foi mostrado neste artigo um novo algoritmo para separação de caracteres sobrepostos baseado no histograma envolvente, no crescimento da região pela agregação de pixels e na análise da vizinhança dos pixels. Foram apresentados exemplos de alguns

resultados obtidos que mostram o funcionamento do algoritmo proposto para caracteres sobrepostos maiúsculos com e sem serifa, os quais apresentam resultados satisfatórios.

O algoritmo apresentou bom desempenho quanto a rapidez devido a fase de pré-separação dos caracteres. Nota-se eficiência na fase de separação dos caracteres sobrepostos devido a lógica adotada de se começar a análise dos pixels vizinhos a partir da linha média da altura dos caracteres.

O sistema apresenta robustez para palavras levemente inclinadas, o que pode ser visualizado nos resultados obtidos. O algoritmo funciona também para palavras com fonte estilo itálico. Este trabalho será estendido para atender a separação de caracteres sobrepostos minúsculos. Algumas adequações também serão estudadas para solucionar casos de palavras emendadas e com graus de inclinação elevados.

Agradecimentos

Às instituições de fomento CAPES e CNPq por propiciarem condições financeiras aos pesquisadores para desenvolverem este trabalho. E, também, aos professores pelo incentivo e acompanhamento despendidos.

REFERÊNCIAS

C. Y. Suen, M. Berthod and S. Mori, Automatic recognition of handprinted characters – the state of the art, Proc IEEE, 68(4), 469-487 (April 1980).L. S. Frishopf and L. D. Harmon, Machine reading of cursive script, Information Theory, Symposium on Information Theory-London C. Cherry, ed., pp. 300-316, Butterworths, Washington (1961).

Gonzalez, R. C. and Woods, R. E., "Digital Image Processing", Addison Wesley, 1993.

J. –C. Simon, Off-line cursive word recognition, Proc. IEEE 80(7), 1150-1161 (1992).

L. S. Frishopf and L. D. Harmon, Machine reading of cursive script, Information Theory, Symposium on Information Theory-London C. Cherry, ed., pp. 300-316, Butterworths, Washington (1961).

M. Y. Chen, ª Kundu, J. Zhou and S. Srihari, Off-line handwritten word recognition using Hidden Markov Model, U. S. Postal Service 5th Adv. Technol. Conf. Pp. 563-577 Washington, DC (November 1992).R. M. Bozinovick and S. N. Srihari, Off-line cursive script word recognition, IEEE Trans. PAMI, 11, 68-83 (1989).

R. M. Bozinovick and S. N. Srihari, Off-line cursive script word recognition, IEEE Trans. PAMI, 11, 68-83 (1989).

S. Mori, C. Suen and K. Yamamoto, Historical review of OCR research and development, Proc. IEEE, 80(7), (July 1992).

Yi Lu and M. Shridhar, "Character segmentation in handwritten words - an overview", Pattern Recognition, Vol. 29, No. 1, pp. 77-96, 1996.Yi Lu, B. Haist, L. Harmon, J. Trenkle and R. Vogt, Na accurate and efficient system for segmenting machine-printed text, U. S. Postal Service 5th Advanced Techonology Conference, Washington D.C., November, Vol. 3, pp. A-93-A-105.

Yi Lu, B. Haist, L. Harmon, J. Trenkle and R. Vogt, Na accurate and efficient system for segmenting machine-printed text, U. S. Postal Service 5th Advanced Techonology Conference, Washington D.C., November, Vol. 3, pp. A-93-A-105 (1992).

Yi Lu, "Machine printed character segmentation - an overview", Pattern Recognition, Vol. 28, No. 1, pp.67-80, 1995.

Zhixin Shi and Venu Govindaraju, "Segmentation and recognition of connected handwritten numeral strings", Pattern Recognition, Vol. 30, No.9, pp. 1501-1504, 1997.

Technique for Separation of Overlapped Characters

Abstract: The objective of character separation is to split the words' image in many parts, each one containing a complete but isolated character. The most difficult issue is separating overlapped characters, because the method used has to discern perfectly one character from another. The complexity is founded in the diversity of text and image features. The most common methods of character separation are based on histograms. Some of them use vertical projections, others, vertical histograms. This is the graphical representation of the number of black pixels in each column of the binary image. Others are based on histogram of vertical extent, this is the graphical representation of the sun of the all pixels in the characters in the image. Moreover, it is based on the analysis of neighboring pixels. This paper presents a technique to separate overlapped characters based on a histogram of vertical extent, the growth of the region by pixel aggregation and the analysis of neighboring pixels. The results confirmed that the developed algorithm was effective in the separation of overlapped characters for several situations.

Keywords: Separation of characters, Pattern Recognition, Histogram