Reconhecimento de dígitos cursivos – um método de segmentação por histogramas

ROBERTO J. RODRIGUES [1]
cracky@nce.ufrj.br
ANTONIO CARLOS GAY THOMÉ [1]
thome@nce.ufrj.br
[1] NCE- Núcleo de Computação Eletrônica/UFRJ, Caixa Postal 2324, Ilha do Fundão, Rio de Janeiro, RJ, Brasil

Abstract

This paper reports the results of a study on a first sight decision tree algorithm for cursive script recognition based on the use of histogram as a projection profile technique. A postal code image data scanned is converted in a 2-dimension matrix representation to be used with a set of algorithms to provide full range segmentation. The results, based on this approach, are quite satisfactory for first stage classifier.

Keywords

Character segmentation, Character Recognition, Image Processing

Resumo

Este artigo demonstra a eficiência da técnica de histogramas de projeção para o processo de segmentação de caracteres cursivos. O método relatado é aplicado na segmentação de caracteres numéricos, obtidos com a digitalização do código de endereçamento postal, a partir de envelopes postais. O desempenho do método é mostrado com experimentação de dados reais.

Palavras-chave

Segmentação de caracteres, Reconhecimento de caracteres, Processamento de imagens.

1. Introdução

O uso de sistemas de manipulação de documentos vem se tornando cada vez mais comum. Aplicações como processadores de texto, sistemas de publicações, e programas gráficos, são usadas frequentemente em diversas organizações, e mesmo escritórios domésticos. A base instalada desta tecnologia cresceu muito nos últmos anos, mas apesar de todo este avanço, os métodos de utilização destes utilitários, requerem a digitação do material textual diretamente no computador, representando uma tarefa cansativa, sujeita a falhas e consumidora de tempo.

O reconhecimento de caracteres, mais conhecido como OCR (Optical Character Recognition), é um sub-conjunto da área de reconhecimento de padrões e, foi ele que estabeleceu as bases e a motivação para tornar o reconhecimento de padrões e a análise de imagens campos individuais de interesse da ciência. A importância das pesquisas nesta área não se baseia apenas em aplicações comerciais mas também no reconhecimento de padrões complexos em 2D e 3D para sistemas de visão robótica e computacional.

A origem do reconhecimento de caracteres provavelmente vem de meados do ano de 1870, quando foi criado o *scanner* de retina, que é um sistema de transmissão de imagem utilizando um mosaico de fotocélulas. O *scanner* sequencial, inventado em 1890, foi um avanço para o desenvolvimento da TV moderna e das máquinas de leitura ótica. O reconhecimento de caracteres propriamente dito apareceu primeiro como um auxílio para deficientes visuais em 1900. A versão moderna do OCR apareceu em meados de 1940 com o desenvolvimento do computador digital. Inicialmente o OCR foi desenvolvido apenas para processamento de informações com uma aplicação limitada ao mundo dos negócios.

Os sistemas de reconhecimento se vêem face a um paradoxo: Como reconhecer sem segmentação, e como segmentar sem reconhecimento? A estratégia usada frequentemente é a geração de muitas hipóteses de segmentação, seguidas de testes para todas as combinações possíveis. Entretanto, esta estratégia leva constantemente a uma explosão de combinações que representa um óbvio impacto negativo no tempo de resposta do sistema (L.Duneau [5]). Uma proposta que vem ganhando muitos adeptos e que já representa um novo segmento de pesquisa, o ICR (Inteligent Character Recognition), implica no uso de redes neurais.

As primeiras idéias no campo das redes neurais remontam aos trabalhos de Norbert Wiener e John von Neuman nos idos de 1940 [6]. O interesse nesta área emergente sofreu uma forte queda nos anos 60, face a descoberta da limitações apresentadas pelos modelos de rede até então conhecidos (perceptron). No final dos anos 60, o desenvolvimento de teoremas matemáticos e os algorítmos de busca heurística para jogos de xadrez, mostraram um forte potencial para métodos simbólicos. Durante os anos 70 e começo dos 80, os métodos simbólicos, como os sistemas especialistas, consistiram nas únicas metodologias usadas para a construção de sistemas inteligentes. Entretanto, já nos 80, tornou-se claro que as soluções simbólicas não eram assim tão brilhantes: apresentavam ótimo desempenho sob condições bem definidas, mas falhavam quando estas condições não eram totalmente conhecidas.

A maior vantagem apresentada pelas redes neurais provém de sua capacidade de aprendizado, ou seja, a capacidade de auto ajustar-se na tentativa de reconhecer padrões a partir das informações dadas. Isto pode ser obtido através de uma base de dados contendo relacionamentos previamente conhecidos. A capacidade das redes para aprender e generalizar tais relacionamentos as torna menos sensíveis ao ruído que outros sistemas. A capacidade de representar relacionamentos não-lineares as torna adequadas a um grande número de aplicações, como controle de sistemas industriais, visão computacional etc.

A proposta deste trabalho é de conceber um sistema inteligente para o reconhecimento automático de algarismos em um Código de Endereçamento Postal (CEP). O processo do reconhecimento, conforme a investigação que está sendo conduzida, compõe-se de 5 fases distintas: digitalização do material a ser analizado – envelopes; pré-processamento para isolamento da faixa que contém o CEP; segmentação dos dígitos; reconhecimento via rede neural (MLP) e finalmente, a geração do código de barras.

2. Reconhecimento – dificuldades e restrições

Segundo J. Mantas [2], um sistema de reoconhecimento de escrita pode ser enquadrado em uma das seguintes categorias:

1. **Reconhecimento de caracteres com fontes fixas:** é o reconhecimento de caracteres escritos com fontes de tipos gráficos específicos como Pica, Courier e etc.
2. **Reconhecimento *on-line*:** é o reconhecimento de caracteres escritos à mão, onde é levado em conta não só o traço do caracter mas também o tempo e a pressão que o autor impõe no processo de escrita do mesmo.
3. **Reconhecimento de caracteres manuscritos:** é o reconhecimento de caracteres à mão, porém com letra de forma e não conectados.
4. **Reconhecimento de caracteres cursivos:** é o reconhecimento de caracteres manuscritos sem restrição, isto é, são cursivos e podem estar conectados.

Das classes acima, definitivamente a mais complexa e difícil de ser implementada é a última, que trata da escrita cursiva com caracteres conectados. Até hoje ainda não foi concebida qualquer técnica de desempenho garantidamente satisfatório, uma vez que não se pode limitar os inúmeros parâmetros, formas e características individuais da escrita cursiva.

O desempenho de um sistema automático de reconhecimento depende da qualidade dos documentos nas suas formas original e digital. Usa-se processos para compensar uma qualidade pobre nos originais e/ou nas imagens após a captura e digitalização. Estes processos incluem realces, remoção de linhas, de sublinhados e de ruídos, além de outros. Os problemas geralmente relacionados com a qualidade e a dificuldade de tratamento da imagem e do texto são: [8]

1. *Ruído* – Causa segmentos de linha desconectados, saltos, pontos, curvas e etc.
2. *Distorção* – Variações locais, arredondamentos de cantos, saliências e reentrâncias indevidas, dilatações e etc.
3. *Variação de estilo* – Uso de formas diferentes para representar o mesmo caractere, como as serifas (“frescuras”) no tipo de fonte tipográfico, inclinações e etc.
4. *Translação* – Posição relativa ou movimentação do caractere inteiro ou de partes.
5. *Escala* – Tamanho relativo do caractere.
6. *Rotação* – Mudanças na orientação.
7. *Textura* – Variações no tipo do papel utilizado e do instrumento de escrita.
8. *Traço* – Variações na espessura do traço e falhas no mesmo.

As estratégias de reconhecimento geralmente abordam o problema na forma de um reconhecimento individual ou de um reconhecimento por contexto. Neste último, faz-se uso de uma vasta biblioteca de contexto específico para validar a imagem a ser classificada. No reconhecimento individual, a imagem é segmentada e a seguir cada segmento é validado comparando-o com uma biblioteca de padrões previamente estabelecidos. As pesquisas em redes neurais têm se mostrado bastante promissoras, particularmente nesta forma de abordagem.

O processo de reconhecimento da escrita-a-mão (cursiva ou não) ou texto impresso, inclui diversas fases, geralmente cobrindo desde o reconhecimento de padrões até a aplicação de alto nível do conhecimento. O reconhecimento de padrões está baseado em análises do padrão físico da linguagem escrita, e o conhecimento de alto nível é aplicado aos resultados do reconhecimento para avaliar a ambiguidade e propiciar melhorias. O conhecimento de alto nível inclui o uso de informações léxicas, sintáticas, semânticas e contextuais. O trabalho também baseia-se nas análises de estruturas de documentos e conteúdo. A integração de

diferentes estratégias e estruturas de conhecimento é um fato importante, e uma tendência na concepção de modelos automatizados.

Uma das tarefas mais complexas e mais relevantes neste processo de reconhecimento individual de caracteres é a segmentação. Hoje, das várias iniciativas divulgadas, a desenvolvida no CEDAR (Center of Excellence for Document Analysis and Recognition) da Universidade de Buffalo, é uma das mais recentes e interessantes. A pesquisa trata o reconhecimento do endereçamento completo, incluindo o código de endereçamento postal americano e a geração do código de barras correspondente. Parte das pesquisas relaciona-se com caracteres orientais como chinês, japonês e coreano. (http://www.cedar.buffalo.edu/)

Alguns grupos de pesquisa nesta área:

IBM Pen Technology (http://www.research.ibm.com/handwriting/)

NICI Handwriting Group, Nijmegen University, The Netherlands. (http://www.nici.kun.nl/)

Script and Pattern Recognition Group at the Nottingham Trent University (http://152.71.57.102/Research/recog.html)

CEDAR, Document Recognition Group at SUNY Buffalo. (http://www.cedar.buffalo.edu/)

CENPARMI, Centre for Pattern Recognition and Machine Intelligence at Concordia, Montreal. (http://www.cenparmi.concordia.ca/)

IDIAP, Artificial Intelligence group in Switzerland. (http://www.idiap.ch/)

DIMUND, University of Maryland. (http://documents.cfar.umd.edu/)

OSCAR, Handwriting Recognition at Essex University (http://hcslx1.essex.ac.uk/)

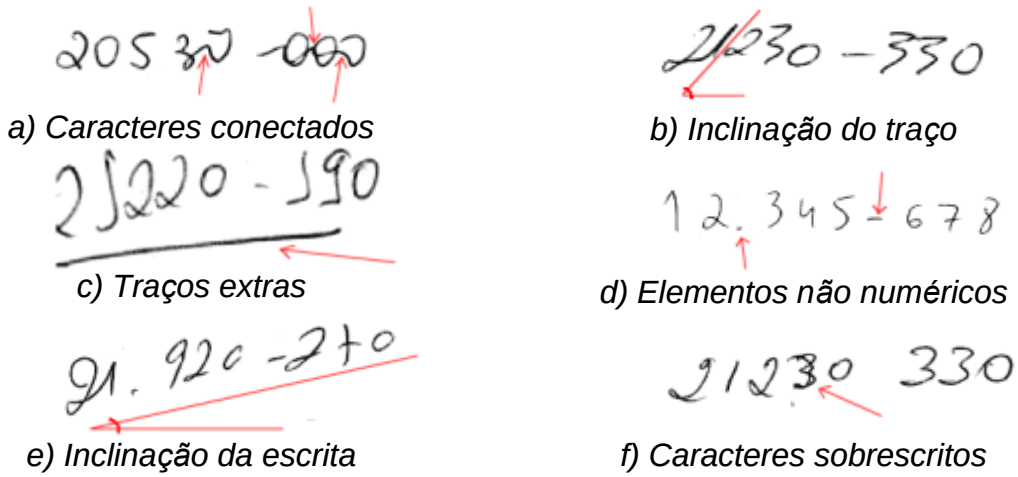
A segmentação de caracteres cursivos encontra diversos problemas ainda não solucionados, como a segmentação de caracteres inclinados, sobrescritos e com traços conectados. (Figura 1).

Dentre as técnicas mais utilizadas para a segmentação está a extração de pontos e linhas na forma de análise de contorno. Verschueren [7] usa uma matriz de 3x3 como uma máscara de direções, gerando um vetor de características para cada linha. Este método apresenta resultados satisfatórios para dígitos tipográficos mas falha ao usar dígitos cursivos. Outra técnica pesquisada define um plano geométrico ou topológico, baseado na distribuição geométrica dos pontos. Esta técnica define direção de segmentos, pontos terminais, intersecções e ciclos. A maior vantagem desta técnica é sua alta tolerância à distorções e variações de estilo quando comparado com outras técnicas.

Srikantan [9] adotou a técnica de gradiente onde as observações sobre o contorno e a estrutura dos objetos são codificadas como gradientes a partir de uma vetorização da direção e magnitude para cada pixel da imagem. Os estudos fisiológicos revelaram que as céluas visuais do cérebreo humano são mais sensíveis aos contornos dos objetos que ao seu interior com partes homogêneas. Os resultados são satisfatórios, entretanto o método mostra-se complexo com parametrização extremamente interdependente.

A escrita cursiva de padrões numéricos pode ser vista como um caso mais simples do cursivo genérico, uma vez que, por sua natureza, os dígitos em geral não são escritos de forma conectada. A investigação relatada neste artigo se concentra unicamente no reconhecimento de padrões cursivos de CEP (Código de Endereçamento Postal), usado pelo correio brasileiro e o processo de

reconhecimento que vem sendo adotado nesta investigação, inclui diversas etapas conforme mostra o diagrama da figura 2.



a) Caracteres conectados *b) Inclinação do traço*

c) Traços extras *d) Elementos não numéricos*

e) Inclinação da escrita *f) Caracteres sobrescritos*

Figura 1: Problemas na escrita cursiva

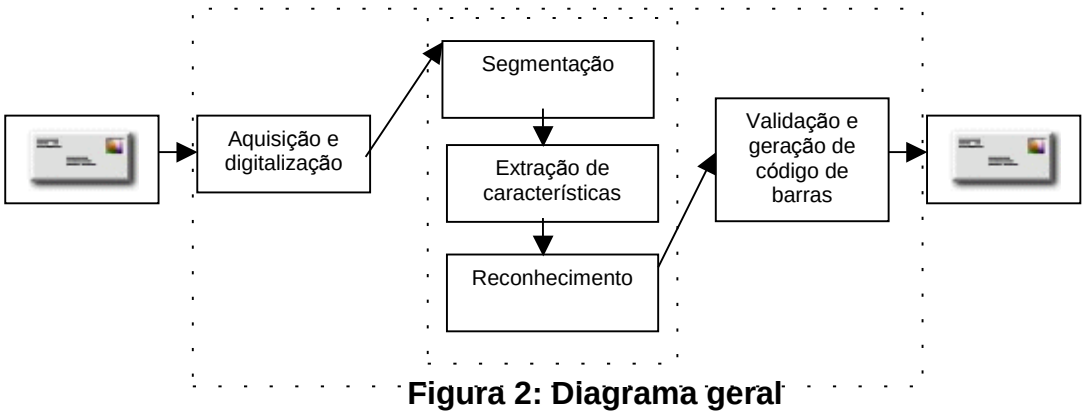


Figura 2: Diagrama geral

3. processo de segmentação

O método de segmentação proposto e relatado neste artigo tem uma lógica baseada em uma árvore de decisão por refinamento sucessivo e no uso de histogramas de projeção numérica.

O histograma de projeção é uma estrutura que armazena o resultado da projeção da contagem de pixels para cada dimensão da imagem (equação 1). Cada célula do vetor é associado à soma dos pixels (contagem) diferentes de um limiar (geralmente a cor de fundo) (equação 2) e (equação 3). Um histograma de projeção alternativo toma a média da intensidade dos pixels, ao invés da contagem dos mesmos.

$$X, Y \rightarrow M(x, y) \quad (1)$$

$$X_n = \sum_{i=0}^{h} Y_i, n \in [0, v] \quad (2)$$

$$Y_n = \sum_{i=0}^{v} X, n \in [0, h] \quad (3)$$

Onde *X* e *Y* representam os vetores de varredura horizontal e vertical, *h* representa a altura da imagem (comprimento vertical) para *X* ou largura (comprimento horizontal) para *Y* e *n* representa o comprimento da imagem

A idéia básica do método consiste em aplicar uma heurística de refinamentos sucessivos, a partir dos dados obtidos nos histogramas até alcançar que seja alcançado um desemplenho satisfatório. O método se divide em 3 etapas: compensção da qualidade da imagem; segmentação inicial e refinamentos sucessivos.

1. ***Compensação da qualidade da imagem***: Esta etapa busca corrigir problemas na imagem original, reduzindo ou realçando certos detalhes como ruídos ou contrastes:
   a) *Identificação e remoção de ruídos de fundo*: Uma imagem digitalizada com baixa resolução ou a partir de um original manchado ou com envelope colorido, irá resultar em um original com ruídos. Para esta tipo de imagem, será necessário remover o fundo com um filtro tendo como parâmetro um fator de limiar com base na densidade apresentada pelo ruído.

A figura 4 mostra a eliminação do ruído de fundo em uma imagem adquirida em papel branco com resolução de 200 dpi. Como pode ser observado, o processo de remoção com base no histograma de projeção, não degenera ou distorce a imagem como em alguns outros métodos de filtragem.

   b) *Remoção de ruídos que não pertencem a imagem*: Molduras, parte de outros objetos digitalizados dentro da área e linhas de sublinhados (figura 4) em geral também podem ser eliminados ou reduzidos com base nos histogramas de projeção (figuras 6 e 7).

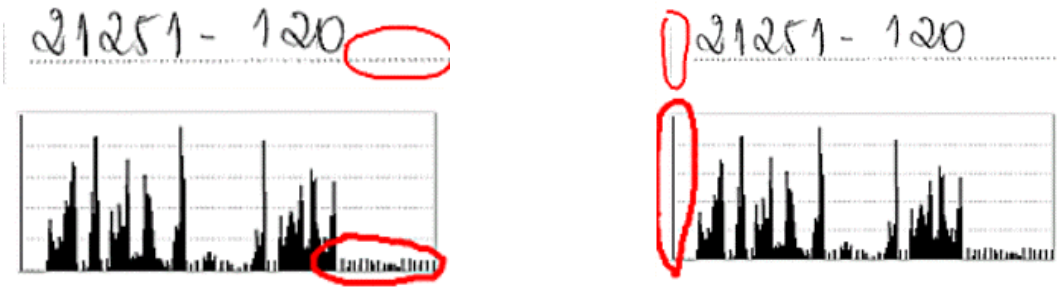
a) imagem original. *b) imagem filtrada.*

Figura 3: Remoção do ruído de fundo

   c) *Realce de contraste*: Imagens muito claras podem ser realçadas intensificando-se o contraste a partir da intensidade do fundo. (Figura 3)

a) Imagem original *b) Após a intensificação do contraste*

Figura 4: Realce de contraste



a) Ruídos horizontais ***b) Ruídos verticais***

Figura 5: Histogramas com níveis de ruído.

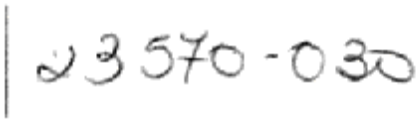
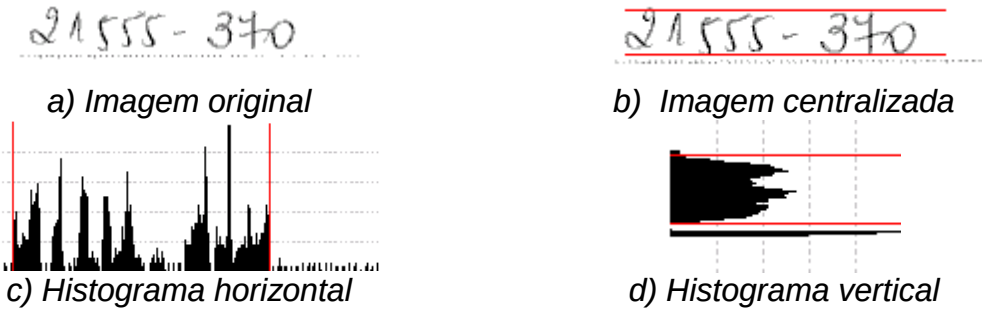


Figura 6: Remoção do sublinhado.

d) *Recorte*: O passo seguinte é destacar da imagem. A parte central onde se encontram os objetos de interesse, neste caso, os dígitos do CEP. Para isto, é necessário o cálculo de valores de limiar para um limite razoável de ruído à esquerda/direita da imagem assim como acima/abaixo. Estes valores garantem que a imagem principal estará sempre dentro de um retângulo destacado da cor de fundo e com pouca interferência dos ruídos não detectados nas etapas anteriores. (Figura 6)



a) Imagem original *b) Imagem centralizada*

c) Histograma horizontal *d) Histograma vertical*

Figura 7: Recorte da imagem

2. ***Segmentação inicial:*** *Ne*sta etapa, a imagem sofre uma primeira segmentação com base nas informações armazenadas na estrutura dos histogramas. A técnica usada baseia-se na densidade de pixels projetada. Uma vez isolada a imagem central, os dígitos são segmentados a partir do limite definidos pelo valor de limiar de ruído. Este valor não é necessariamente similar aos valores de limiar usados para centralizar a imagem. Este parâmetro, chamado de taxa de refinamento, define um valor mínimo para a separação dos dígitos. Nesta etapa inicial, todos os dígitos visivelmente separados são segmentados com sucesso.

Baseado na estatística do histograma, é possível prever possíveis erros como traços e dígitos conectados. Os pontos ou traços podem ser removidos usando as informaçoes do histrograma horizontal. Uma vez que a altura dos dígitos irão influenciar a média de alturas, qualquer elemento com uma altura inferior a uma porcentagem desta média será automaticamente removido. Os dígitos conectados, também previstos com este método, serão tomados como entrada para a segunda etapa da segmentação.

3. ***Refinamento:*** Esta etapa repete as etapas definidas na segmentação inicial diferenciando apenas na variação dos parâmetros de refinamento. Aumentando o limiar, é possível desconectar possíveis dígitos conectados. (Figura 8).

a) Imagem segmentada *b) Dígito 3 desconectado* *c) Dígito 0 desconectado*

Figura 8: Segmentação refinada.

Os elementos que não tiverem resultado satisfatório na sucessão dos refinamentos serão tratados com outros métodos de segmentação na sequência da árvore de decisão.

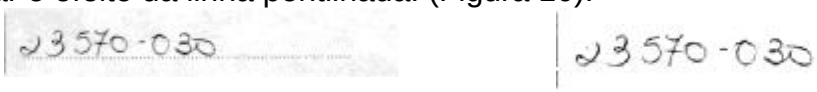
4. Resultados obtidos

O experimento foi realizado com base nas informações geradas por imagens coletadas na comunidade acadêmica da Universidade Federal do Rio de Janeiro. Para a coleta de dados, foi criado um formulário onde cada escritor preenche seus dados pessoais (faixa etária, escolaridade, sexo, etc.) e escreve 5 exemplos de CEP (2 iguais e 3 diferentes). Cada formulário gerou 5 tiras de CEP. Ao todo foram utilizadas 540 tiras de CEP.

As imagens usadas nos testes foram adquiridas a partir um digitalizador (*scanner*) de mesa nas resoluções de 100 e 200 dpi, entretanto só foram utilizadas as imagens de 200 dpi. A profundidade de cores para os experimentos relatados neste artigo, foi definida no padrão RGB com 24 bits/pixel em escala de cinza. A dimensão final de uma tira de CEP (figura 10) possui, aproximadamente, 500x120 pixels (200 dpi).

Para os testes foram usadas 540 tiras contendo CEPS escritos sobre uma linha pontilhada. Esta linha pontilhada interferiu negativamente no desempenho global. Como a digitalização também capta a linha, esta infere um elemento de ruído em todas as imagens.

As tiras digitalizadas apresentavam manchas, eventuais traços e sublinhados, visto que não foi utilizado formulário especial com cores nas delimitações das caixas de escrita. O primeiro passo da sequência foi remover o fundo manchado, ruídos e minimizar o efeito da linha pontilhada. (Figura 10).

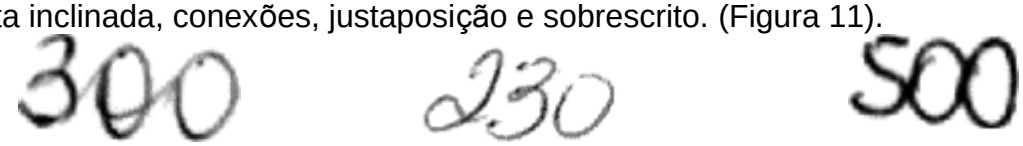


a): Uma tira usada para teste. *b) Primeiro resultado.*

Figura 10: Remoção de ruídos

A seguir foi realizado o recorte da imagem. De um total de 4320 dígitos, assumindo 8 em cada uma das 540 tiras, o algorítmo extraiu 3788 segmentos, incluindo os dígitos corretamente segmentados, os segmentos com múltiplos dígitos e erros de segmentação, conforme a estatística apresentada na tabela 1. Para esta etapa, foram usados os valores de 3 para refinamento e de $E0E0E0 (RGB hexadecimal) para o limiar de fundo.

Os segmentos com dígitos múltiplos incluem aglutinações geralmente resultantes de escrita inclinada, conexões, justaposição e sobrescrito. (Figura 11).



a) Conexão simples *b) Inclinação* *c) Aglutinação*

Figura 11: Problemas em segmentos com múltiplos dígitos.

Tabela 1: Primeira segmentação.

	Quantidade	% do extraído	% do esperado
Total de tiras	540	-	-
Dígitos esperados	4320	-	-
Segmentos extraídos	3788	-	-
Segmentos corretos	3286	86,71	76,06
Segmentos com múltiplos dígitos	389	10,26	9,00
Erros	113	3,00	2,60

Os erros mais comuns incluem traços e pontos que passaram pela filtragem inicial, e dígitos mal segmentados. Observou-se que o tipo de erro mais comum foi o corte de conexão indevida, como pode ser visto na figura 12.

a) Dígito original *b) Parte segmentada* *c) Parte segmentada*

Figura 12: Erro de segmentação.

Com base em uma heurística simples, o algorítmo separa os segmentos dentre dois conjuntos: supostamente corretos e supostamente múltiplos. Estes últimos são selecionados para refinamento. Após o refinamento com os valores iniciais alterados, 5 para o grau de refinamento, chegou-se aos resultados apresentados na tabela 2.

Tabela 2: Segunda segmentação.

	Quantidade	% do extraído	% do esperado
Segmentos corretos	320	86,26	-
Segmentos com múltiplos dígitos	48	12,33	-
Erros	21	5,41	-

Com este resultado, obteve-se o resultado final:

Tabela 3: Resultado final.

	Quantidade	% do extraído	% do esperado
Total de tiras	540	-	-
Dígitos esperados	4320	-	-
Segmentos corretos	3606	95,20	83,47

5. Conclusões e futuro

O uso de histogramas de projeção para o processo de segmentação de caracteres não resolve todos os problemas apresentados, como inclinação de traço, porém

soluciona mais de 70% dos casos com dígitos conectados. Uma vez que os algorítimos utilizados têm como base uma implementação extremamente simples, incluindo os filtros convencionais, o processo, no todo, resulta em um desempenho funcional ótimo, levando em conta o tempo necessário para realizar todas as etapas do processo.

A utilização de envelopes especiais, impressos com delimitadores coloridos, também é um elemento que certamente elevaria consistentemente o desempenho do método, pois os delimitadores seriam facilmente eliminados da imagem digitalizada por meio de filtragem simples, tendo como resultado dígitos praticamente separados.

Como o método de segmentação por histograma de projeção proposto neste artigo possui toda heurística e decisões formadas com base na densidade apurada da imagem, torna-se bastante intuitivo a concepção de novos algorítmos como folhas a serem inseridas na árvore. Assim sendo, os próximos passos incluem o aprimoramento do método, mais precisamente os algorítmos de análise de contorno para enumerar os caracteres conectados e o desenvolvimento de novos algorítmos para solucionar os problemas aqui relatados.

Uma outra forma de aumentar o desempenho deste método, assim como dos algorítmos da sucessão, é ter o suporte de uma rede neural para a validação dos caracteres segmentados. A melhoria do desempenho dar-se-ia na redução de passos significantes na sequência do processo a partir do resultado da consulta à rede.

6. Referências bibliográficas

[1] D.G. Elliman, I. T. Lancaster, *A review of segmentation and contextual analysis techniques for text recognition,* Pattern Recognition, Vol. 23, No. 3/4, pp 337-346, 1990

[2] J. Mantas, *An Overview Of Character Recognition Methologies,* Pattern Recognition, Vol. 19, No. 6, pp 425-430, 1986

[3] W.H. Abdula, A.O.M Saleh, A. H. Morad, *A preprocessing algorithm for hand-written character recognition*, Pattern Recognition Letters 7 (1988) 13-18

[4] C. Y. Suen, M. Berthod, S Mori, *Atomatic Recognition of Handprinted Characters,* Proceedings of the IEEE, Vol. 68, No. 4, April 1980

[5] L. Duneau, *Étude et réalisation d'un système adaptatif pour la reconnaissance en ligne de mots manuscrits,* Thèse de doctorat, Université Technologique de Compiègne, France, 1994

[6] J. Hertz, A. Krogh and R. Palmer, *An introduction to the Theory of Neural Computation*, ISBN 0-201-50395-6 and 0-201-51560-1 (1991) .

[7] W. Verschueren, B. Schaeken, Y. R. de Cotret, A. Hermanne, *Structural Recognition of Handwritten Numerals*, CH2046-1/84/0000/0760$01.00@1984 IEEE

[8] C.Y. Suen, M. Berthold, S. Mori, *Automatic Recognition of Handprinted Characters – The State of The Art,* Proceedings of the IEEE, Vol. 68, No. 4, April 1980

[9] G, Srikantan, S. W. Lam, S, N, Srihari, *Gradient-Based Contour Encoding For Character Recognitio*n, Pattern Recognition, Vol. 29, No. 7, pp. 1147-1160, 1996