RESTAURAÇÃO DE DOCUMENTOS ELETRÔNICOS DEGRADADOS COM BINARIZAÇÃO ADAPTATIVA

Fabiano Amorim Vaz
Mestre em Ciência da Computação - Universidade Federal de Pernambuco – UFPE
e-mail:fav2@cin.ufpe.br

Camila Gonzaga de Araújo
Especialista em Gestão de Projetos em TI - Centro de Ensino Superior de Maceió – CESMAC
e-mail:cmilaaraujo@gmail.com

Filipe Rafael Gomes Varjão
Mestrando em Ciência da Computação - Universidade Federal de Pernambuco – UFPE
e-mail:varjaofilipe@gmail.com

RESUMO
Este trabalho tem como objetivo implementar a técnica proposta por Sauvola o artigo intitulado "*Binarização imagem Adaptive documento*", este artigo utiliza técnicas híbridas para binarização de documentos que possuem ambos os componentes de texto e não texto. No entanto, esta proposta foi adaptada e extrapolada, dando origem a um novo método híbrido com base no método de Sauvola e usando as vantagens dos Kmeans método. Esta abordagem tem o aspecto principal da classificação dos elementos de estrutura do documento e, portanto, a binarização deste documento, como um todo.
Palavras-chave: Adaptável. Binatization. Documento Híbrido. Método.

ABSTRACT
This work aims to implement the technique proposed by Sauvola the article entitled "Adaptive document image binarization", this article uses hybrid techniques for binarization of documents that have both text and non-text components. However, this proposal was adopted and extrapolated, giving rise to a new hybrid method based on the method of Sauvola and using the advantages of the method Kmeans. This approach has the main aspect of the classification of elements of document structure and, therefore, the binarization of this document as a whole.
Keywords: Adaptive. Binatization. Document. Method Hybrid.

INTRODUÇÃO

Por sua simplicidade de implementação e propriedade intuitiva, a limiarização de imagem possui uma importância relevante para aplicações que envolva em seu contexto a segmentação como base no procedimento. Esta limiarização é a escolha de um valor *T*, chamado de limiar, dentro de uma escala de tons que uma imagem possui. Os tons que forem maiores que este limiar receberam um valor, que geralmente é 1 (preto), indicando que aquele pixel corresponde a parte de um objeto e para os tons inferiores ao limiar serão atribuídos o valor 0 (branco), considerando a existência de um *background* (fundo). O resultado desta normalização é chamado de binarização, onde os tons que anteriormente variavam entre 0 e 255, agora possuem apenas valores binários (0 e 1).

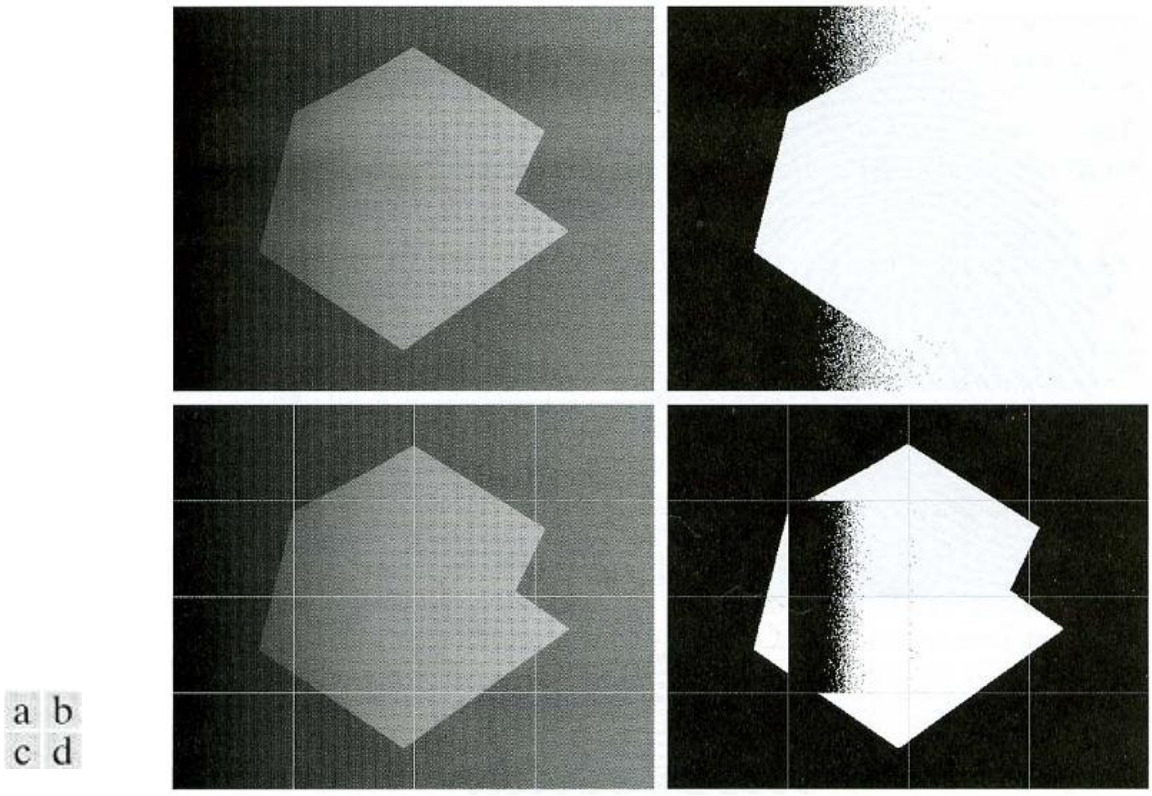
Muitos algoritmos de análise de documentos são feitos usando como base os resultados de uma imagem binarizada. O uso de informações binarizada, diminui o custo computacional e facilita o tratamento dos dados, simplificando a análise.

Fabiano Amorim Vaz
Camila Gonzaga de Araújo
Filipe Rafael Gomes Varjão

Análise e reconhecimento de imagens de documentos é uma área de grande relevância por suas inúmeras aplicações. Por possuir características próprias, cada documento naturalmente exige soluções computacionais específicas.

Os algoritmos de binarização utilizam dois tipos de limiar o global e o local. No primeiro caso é calculado um único limiar e aplicado a toda imagem indiferentemente, porém em algumas situações um único limiar não é suficiente, como no caso de mudanças ou intensidades diferentes de iluminação na mesma imagem, digitalização de má qualidade, falhas na captura do documento digital. O uso do algoritmo de binarização com limiar local consegue suprir uma grande parte dos problemas ocorridos no limiar global, pois utiliza um valor de limiar para cada região ou bloco da imagem, tornando possível um melhor efeito na imagem.

Figura 1. (a) Imagem original. (b) Resultado da binarização global. (c) Imagem dividida em subimagens. (d) Resultado da binarização local.



Alguns métodos híbridos têm sido propostos, estes métodos fazem uso tanto das informações do método global quanto do local, para obter um resultado satisfatório.

A maioria dos algoritmos propostos não trata dos objetos que compõe uma imagem de documento de forma a distinguir de qual classe pertence um determinado trecho da imagem. A técnica proposta por Sauvola (2000) indica que ao se fazer uma separação de elementos, a maneira como tal trecho deve ser tratado é simplificado e utilizando o método híbrido nesta imagem o resultado é mais eficiente que usar o algoritmo global ou local, de forma separada.

1 PROCESSAMENTO DE IMAGENS

As áreas de Computação Gráfica e Processamento de Imagens podem ser representadas a partir do processo utilizado, em ambos, para extrair ou entender de forma clara imagens, fazendo uso dos sinais obtidos nestes processos. Cada um ao seu modo, mas com o objetivo comum da informação (GONZALEZ e WOODS, 2000).

Processamento de Imagens (PI) pode ser conceituado como a transformação de imagens digitais para imagens digitais com dados compreensíveis ao humano ou o melhoramento visual destes dados. Contudo, os estudos de Processamento de Imagens estão relacionados a basicamente duas aplicações, sendo a primeira a melhoria de imagem, com o intuito de facilitar a visualização das informações para interpretação humana. O segundo foco de aplicação fundamenta-se na automação do reconhecimento de informações contidas nas imagens digitais.

Um exemplo de aplicação do Processamento de Imagens para se obter melhoria em imagens digitais está na restauração de imagens com ruído ou mesmo com informações deturpadas. Como exemplo de reconhecimento de padrões pode ser mencionado à identificação de pessoas, objeto e/ou eventos contidos nas imagens. Em alguns casos, ambas as aplicações se complementam, principalmente, quando se faz necessário o uso de melhoria da imagem para que posteriormente seja realizado o reconhecimento ou extração de sinais (informações).

Neste projeto, da área de Processamento de Imagens, as duas principais vertentes são aplicadas de modo complementar. Para entendimento dos recursos utilizados para realizar o projeto, um conhecimento básico dos componentes de um sistema de Processamento de Imagens é necessário. Os principais componentes do sistema podem ser classificados em: aquisição de imagens, armazenamento, processamento, comunicação, exibição de imagens.

Para aquisição de imagens digitais são necessários dois elementos básicos: dispositivo físico, que é sensível à energia irradiada pelo objeto a ser capturado e o digitalizador, que converte a saída do dispositivo físico em formato digital. Após ser adquirida (capturada) a imagem deve ser registrada em algum tipo de memória. Podendo este armazenamento ser realizado temporariamente (no processamento), on-line (retirada rápida) e permanente, onde necessita de maior poder de armazenamento.

O processamento é variável, de acordo com a aplicação, pois pode ser realizado por um computador de pequeno/médio porte ou até por grandes sistemas especializados. Entretanto, a forma de realizar o processamento é basicamente a mesma independente de tamanho, ou seja, a maioria das funções de processamento podem ser implementadas em software.

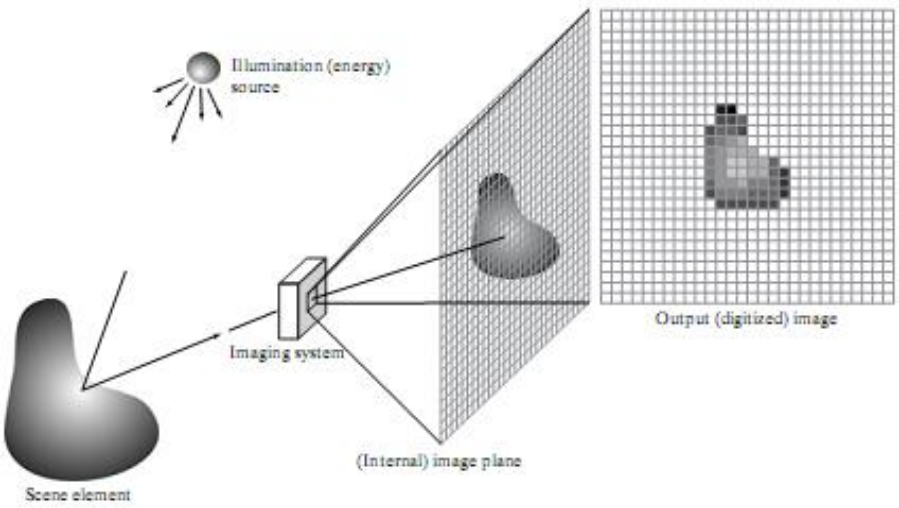
A comunicação é o modo de entrelaçar os componentes ou sistemas. Atualmente, é considerada uma função padrão em qualquer sistema de computador, em redes locais não existem tantos problemas, contudo, na comunicação remota (internet) nem sempre são tão eficientes. Por fim, a exibição é a forma como os resultados são apresentados. Nos dias atuais existem inúmeras formas de exibição, variando de uma simples TV ou monitor a óculos estéreos especiais.

Fabiano Amorim Vaz
Camila Gonzaga de Araújo
Filipe Rafael Gomes Varjão

1.1 REPRESENTAÇÃO DAS IMAGENS DIGITAIS

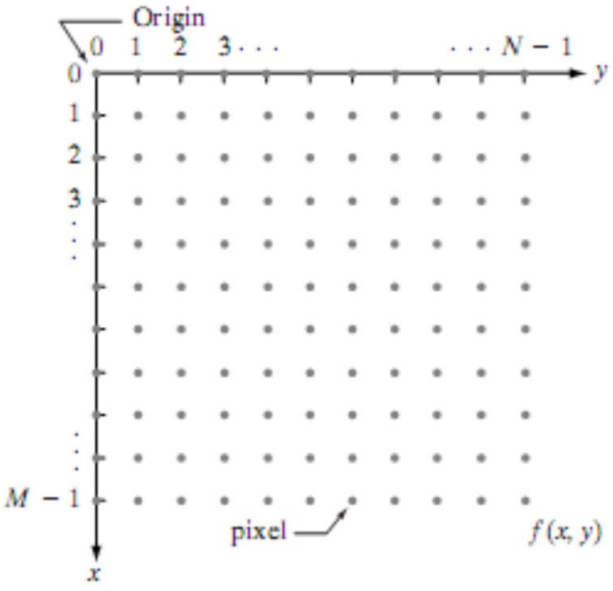
Quando uma cena é captura por um sensor de aquisição, sua representação é transformada e quantificada em uma matriz de números reais, conforme Figura 2.

Figura 2 – Exemplo de Aquisição de Imagens



Assumindo que uma imagem digital é função discreta *f(x, y)*, temos como resultado *M* linhas e *N* colunas (Figura 3), sendo o valor da coordenada (*x, y*) a representação numérica da cor correspondente a este pixel. Onde o primeiro pixel é, portanto, a coordenada (*0, 0*), conforme indicado na Figura 3.

Figura 3 – Matriz de Coordenadas de uma imagem



Esta notação permite representar genericamente qualquer imagem bidimensional. Salientado que cada pixel de uma imagem corresponde a uma cor, onde imagens coloridas possuem em cada pixel um vetor de três valores (RGB / CMY), imagens em escalas de tons cinza possui uma variação de 0 à 255 em cada pixel, por fim, imagens binárias possuem valores entre 0 e 1 por pixel.

1.2 SEGMENTAÇÃO

Para facilitar a extração e análise das imagens em sistemas computacionais à segmentação torna-se uma ferramenta eficaz, pois tem enfoque no conhecimento do problema, isto é, dada uma imagem de entrada, as informações lá contidas devem ser tratadas.

A etapa de segmentação de imagens está presente na maioria dos projetos de processamento de imagens, na qual é definida a existência, a localização e os tipos de estruturas procuradas. Pela grande variedade de "primitivas" ou "segmentos significativos", que contêm as informações semânticas e pelas inumeráveis aplicações a etapa de segmentação talvez seja um dos maiores desafios da área de processamento e análise de imagens.

Neste trabalho, em especial, a segmentação é crucial para conseguir receber dados da interação do usuário, através de *frames*. Por se tratar de um assunto amplo e relativamente complexo, apenas conceitos essenciais para o entendimento do projeto serão discorridos nesta seção.

Algoritmos de segmentação de imagens são, geralmente, baseados em duas propriedades básicas: descontinuidade e similaridade. Na primeira categoria, a busca é por mudanças bruscas de intensidade (contraste), na segunda, o que se busca é padrão ou similaridade entre partes da imagem, de acordo com critérios pré-definidos.

Uma região pode ser definida como sendo um conjunto de pontos que respeitam um mesmo predicado de homogeneidade. Assim, deve sempre existir pelo menos um caminho inteiramente contido nessa região ligando dois pontos.

1.2.1 Detecção de Descontinuidades

Descontinuidade é uma mudança brusca do nível de cinza entre duas regiões relativamente homogêneas. A segmentação por detecção de descontinuidade consiste, portanto, em localizar pontos nessas mudanças bruscas de níveis de cinza.

A literatura (GONZALEZ e WOODS, 2000; NIBLACK, 1986) adota três tipos principais de abordagem de descontinuidades: pontos, linhas e bordas. Uma maneira de procurar descontinuidades é através da varredura de uma imagem por uma máscara.

Figura 4 – Máscara de varredura de tamanho 3x3

A Figura 4 apresenta uma máscara de tamanho 3x3, onde cada ponto é representado por w_i que será aplicado a uma região de mesmo tamanho na imagem original, isto é, z_i é o nível de cinza do pixel associado com o coeficiente w_i da máscara. Logo, esta região pode ser representada por R, sendo este o somatório dos produtos de cada pixel da máscara (w_i) pela intensidade de cinza do ponto respectivo da imagem (z_i):

$$R = \sum_{i=1}^{9} w_i z_i \qquad (1)$$

Para exemplificar, para detecção dos pontos de uma imagem em tons de cinza, assume *T* como um valor limiar (entre 0 e 255), logo se $|R| > T$ o ponto é detectado. Neste procedimento é medida a diferença ponderada entre o ponto central e seus vizinhos.

1.2.2 Detecções de Similaridades

A proposta básica da detecção por similaridade está na binarização da imagem. A limiarização consiste em definir um ponto que seja divisor entre o fundo da imagem (*background*) e o ponto de interesse (*foreground*). Supondo uma imagem *f(x, y)* em tons de cinza, *T* sendo o *thresould* (limiar) que separa as duas regiões, temos:

$$g(x,y) = \begin{cases} 1 & \text{se } f(x,y) > T \\ 0 & \text{se } f(x,y) \leq T \end{cases} \qquad (2)$$

Contudo, esta normalização pode tornar-se custosa quando não se conhece a imagem, podendo este *T* ser definido diversas vezes até se chegar o mais próximo possível do que se almeja.

- **Binarização Global**

Na binarização global busca-se um único valor de limiar para toda a imagem. Assim, seja *T* o valor de limiar, para cada pixel *P(x,y)*. Portanto, um algoritmo para estimar o *T* apropriado é:

1. Selecionar uma estimativa inicial para T

2. Segmentar a imagem usando T

Isso irá produzir dois grupos de pixels:

G1 – todos os pixels com os níveis de cinza < T

G2 – todos os pixels com os níveis de cinza >= T

3. Para cada grupo calcule o nível de cinza médio (μ_1 e μ_2)

4. Calcule o novo limiar

$T = (\mu_1 + \mu_2)/2$

5. Repita os passos de 2 a 4 até que a diferença entre sucessivos T's seja menor do que um parâmetro T_0

A limiarização apresenta a desvantagem de nem sempre as imagens conterem intensidades de primeiro e segundo planos bem diferenciados. A maioria das técnicas buscam segmentar a imagem em duas classes, maximizando a variância inter-classes e minimizando a variância intra-classes. Isto se faz, empregando funções específicas para cada abordagem (OTSU, 1979).

- **Binarização Local Adaptativa**

Segundo Sauvola e Pietikainen (2000), devido à dificuldade em selecionar um limiar global, definir valores diferentes de limiar para regiões diferentes da imagem provou ser uma abordagem interessante. Este tipo de binarização é chamado de binarização adaptativa ou local. O problema principal deste tipo de abordagem é a escolha do tamanho da janela para a definição do limiar local. A janela é importante, pois define o tamanho da região para estimar o limiar.

- **Limiarização Global Ótima**

De acordo Gonzalez e Woods (2000), no cenário hipotético em que uma imagem possua uma região clara e outra região escura, a escolha da limiarização global ótima é satisfatória, pois esta considera o histograma (função discreta $h(r_k) = n_k$, onde r_k é o k-ésimo nível de cinza, e n_k é o número total de pixel com nível de cinza *k*) da imagem como uma estimativa da função de probabilidade do brilho *P(x)*.

Fabiano Amorim Vaz
Camila Gonzaga de Araújo
Filipe Rafael Gomes Varjão

Contudo, o objetivo é encontrar o limiar *T* cujo erro seja mínimo. Logo, o limiar global ótimo é aquele que a probabilidade total do erro seja mais próxima de *0*. Portanto, o limiar ótimo é expresso por:

$$T = \frac{\bar{u}_1 + \bar{u}_2}{2} + \frac{s^2}{\bar{u}_1 + \bar{u}_2} \ln \left(\frac{(P(x_1)=K_0)}{(P(x_2)=K_1)} \right) \qquad (3)$$

Onde $\bar{u}_1 + \bar{u}_2$ representa o somatório das médias dos pixels das duas regiões (*foreground* e *background*) e *s* o desvio padrão dos nível de cinza da imagem referência.

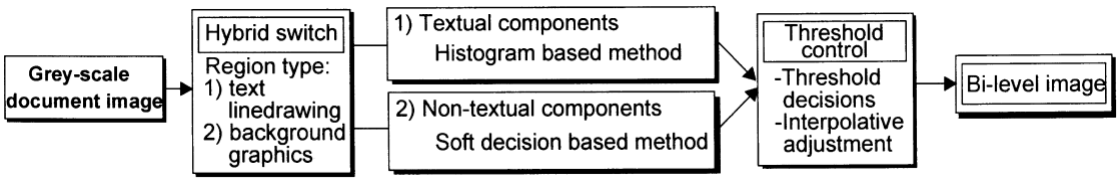
2 TÉCNICAS DE BINARIZAÇÃO

2.1 SAUVOLA

Devido ao fato que a identificação dos componentes de um documento digital não é algo trivial entre os algoritmos proposto por outros autores, Sauvola (2000) propõe em seu artigo um novo método que tem a característica de distinguir entre textos, imagens e fundo. Para tornar isso possível, é necessário aplicar duas técnicas, a SDM (*soft decision method*) utilizada para classificar e binarizar imagens e o fundo, a outra técnica utilizada é a TBM (*text binarization method*) especifica para binarização de componentes textuais.

A técnica indicada analisa as regiões locais, com o objetivo de identificar qual o método necessário para se fazer uso naquela determinada área, durante esta decisão um módulo de "chaveador híbrido" seleciona uma ou duas técnica especializada em binarização que será aplicada neste bloco. A finalidade é utilizar o melhor algoritmo para produzir um limiar ótimo para cada pixel. Como alternativa para uma custo computacional menor é utilizar os valores dos n-primeiros pixels para realizar uma interpolação para o resto da janela.

Figura 5. Algoritmo de Binarização



Essa técnica é utilizada na primeira etapa em um conjunto de análise entre vários documentos, processamento e em tarefas de recuperação ou melhoramento em imagens de documentos. A Fig. 2 demonstra, de forma resumida, as etapas de o algoritmo de binarização seguido por Sauvola (2000).

Sendo este abrangente aos casos das classes de textuais e não textuais. Entretanto, devido à ausência de informações mais precisas sobre a utilização do SDM em componentes não textuais, neste projeto, foi descartado os objetos que se encaixavam nesta classe. Restando, dessa maneira, a binarização dos elementos considerados como textuais, utilizada pelo autor.

No entanto, o método utilizado por Sauvola (2000) para realização da binarização é produto de uma modificação do algoritmo de Niblack (1986), que tem como ideia fundamental o uso da média local (m) e o desvio padrão (s), também local. O limiar (T) é calculado conforme mostrada na Eq. (4).

$$T = m + k*s \quad (4)$$

Onde *k* é um parâmetro definido pelo usuário. Porém este método foi modificado para uma otimização de resultados, onde o limiar é calculado conforme uma faixa dinâmica do desvio padrão (*R*), conforme mostrada na Eq. (3). Onde o *m(x,y)* e *s(x,y)* são como na Eq (4) de Niblack (1986). Em alguns exemplos do artigo o autor indica o uso do $K = 128$ e o $k = 0,5$.

2.2 KMEANS

O algoritmo *kmeans* tem sido amplamente usado na área de aprendizagem de máquina para clusterização e para segmentação em processamento de imagem, para binarização de imagem de forma global o *kmeans* não funciona muito bem para imagens degradadas e com variações de iluminação, já usado em binarização local o mesmo se comporta bem nos blocos onde há de fato texto, mas muito mal onde há apenas background, pois background é colocado parte para preto e em parte branco característica não desejada. Na nossa abordagem com *kmeans* consideramos regiões onde o desvio padrão é muito baixo como background e binarizamos para branco. A nossa implementação do algoritmo *Kmeans* funciona da seguinte forma:

- Dividir a imagem global em janelas de subimagens;
- Indicar que 50% dos pixels da região como pertencentes à Classe 1 e os demais pertencentes a Classe 2.
- Repetir o algoritmo até que os centros das regiões sejam encontrados;
- Após o centro ser localizado, atribui-se ao valor do *threshould* a média de todos os centros.

3 EXPERIMENTOS

Para se chegar a um resultado idêntico ou similar ao utilizado por Sauvola (2000), foi implementado o método que diferencia os elementos de um documento, entre imagem, texto e fundo (*background*), como também o método de binarização de textos em uma imagem proposta.

Nossos experimentos foram focados em duas principais etapas, sendo elas a análise das regiões da imagem e a binarização das regiões classificadas como texto. Nas subseções 4.1 e 4.2 estas etapas estão sendo detalhadas e na subseção 4.3 é exposto o método proposto para melhoria da técnica analisada, no âmbito da binarização de textos.

3.1 ANÁLISE DE REGIÕES

Como etapa inicial do processo de binarização, foi utilizado o método proposto por Sauvola (2000), onde é citado certa quantidade de regras a serem seguidas para este procedimento. Através destas regras, é possível classificar as regiões que contém informações do tipo textual e não textual (imagens e *background*). Onde foi utilizado o resultado dessa análise para a continuidade dos demais experimentos, sendo utilizados no escopo do projeto apenas os resultados classificados como textos e posteriormente aplicados um algoritmo de binarização.

3.2 BINARIZAÇÃO DE ELEMENTOS TEXTUAIS

Após a detecção dos componentes estruturantes da imagem analisada, tem-se o retorno textual e não textual. Nossos experimentos inicialmente focalizaram na técnica de binarização usada por Sauvola (2000) para produção dos resultados referentes à binarização dos componentes classificados como texto.

Entretanto, os resultados obtidos em algumas imagens não foram, visualmente, satisfatórios, dando inicio a elaboração de uma nova técnica proposta para otimização de resultados em imagens com componentes textuais ou não.

3.3 MÉTODO PROPOSTO

Nossos experimentos focalizaram na melhoria da técnica de binarização usada por Sauvola (2000) para produção dos resultados referentes à binarização dos componentes classificados como texto.

Este melhoramento consiste em utilizar o método Kmeans nas regiões em que o método Sauvola (2000) não tem bom rendimento.

4 RESULTADOS

Em análises feitas de forma visual e estatística, em ambas, o resultado obtido pelo método proposto por nós teve um rendimento melhor que o Sauvola (2000). Levando em consideração imagens que tenha em sua estrutura uma diversidade de componentes, entre textos, imagens e fundo.

Figura 6. Resultados visuais dos experimentos (a) Imagem original (b) Método Sauvola (2000) (c) Método proposto.

Fabiano Amorim Vaz
Camila Gonzaga de Araújo
Filipe Rafael Gomes Varjão

(a) (b) (c)

Na Figura 6 pode se observar que a imagem resultante do método proposto tem uma visualização mais agradável e nítida que a técnica abordada para o experimento. Este efeito se dá, pelo motivo de se utilizar uma técnica híbrida de dois métodos que tem, conhecidamente, bons resultados em determinados cenários e aplicados o melhor de cada técnica para um resultado otimizado.

4.1 AVALIAÇÃO DE DESEMPENHO DOS RESULTADOS

Para avaliar os resultados obtidos entres os algoritmos foi utilizada uma abordagem onde a imagem resultante é comparada com uma imagem ótima da binarização (*Ground Truth*). Nessa abordagem o valor de cada pixel das duas imagens são comparadas e associadas a três critérios: TP (*true positive*), FP (*false positive*) e FN (*false negative*), onde:

TP ocorre se os dados contidos na imagem ótima e na imagem resultante forem ambas 1 (preto) ou ligados (ON) o FP ocorre quando apenas na imagem resultante o pixel tem valor 1 e no caso do pixel da imagem resultante esta 0 (branco) e a imagem ótima estiver 1, esse é classificado como FN. Após percorrer toda a imagem acumulando a quantidade de ocorrências de cada fator, calcula-se a precisão (*PR*) e o erro (*RC*), conforme a Eq. (5) e Eq. (6):

$$RC = CTP / (CFN + CTP) \quad (5)$$

$$PR = CTP / (CFP + CTP) \quad (6)$$

Utilizados na Eq. (7), para um valor de aproximação da imagem ótima, sendo este o valor que representa a qualidade da binarização.

$$FM = (2 \times RC \times PR / (RC + PR)) \quad (7)$$

Tabela 1. Comparativo entre as técnicas de binarização abordadas

Técnica	Avaliação
Método proposto	95,5%
Sauvola	92,3%

CONSIDERAÇÕES FINAIS

De acordo com o levantamento bibliográfico utilizado para realização deste projeto pôde-se concluir que binarização de imagens de documentos é base das tarefas dos sistemas de análise de imagens. Bem como, um resultado de boa qualidade pode definir o sucesso das etapas seguintes do processo.

Foi observado que a técnica híbrida proposta teve um rendimento de qualidade maior ao método utilizado por Sauvola (2000). Com base nestes dados e nos dados coletados na literatura é possível concluir que o uso de técnicas híbridas tem si mostrado mais eficaz que o uso de métodos isolados já conhecidos.

REFERÊNCIAS BIBLIOGRÁFICAS

Gonzalez, R. C.; Woods, R. E.: "**Processamento de Imagens Digitais**", Edgard Blücher Ltda, (2000).

Hannah, I; Pastel, D and Davies, R. **The use of variance and entropic thresholding methods for image segmentation**. Pattern Recognition. 28(8). Pages: 1135-1143.

I.-K. Kim, D.-W. Jung, R.-H. Park, **Document image binarization based on topographic analysis using a water flow model**, Pattern Recognition 35 (2002) 265–277.

J. Sauvola, M. Pietikainen, **Adaptive document image binarization**, Pattern Recognition 33 (2000) 225–236.

N. Otsu, **A Threshold Selection Method from Gray-Level Histograms**, IEEE Transactions on Systems, Man, and Cybernetics, v. 9, n 1, pp. 62-66, 1979.

S. Nieminen, J. Sauvola (2000), T. SeppaKnen, M. Pietikäinen (1998), "***A benchmarking system for document analysis algorithms****"*, Proc. SPIE 3305 Document Recognition V 3305, p 100-111.

Trier, O.D.; Jain, A.K, **Goal-directed evaluation of binarization methods**, IEEE Transactions on Pattern Analysis and Machine Intelligence,Volume 17, Issue 12, Dec. 1995 Page(s):1191 – 1201

Trier, O.D.; Taxt, T., **Evaluation of binarization methods for document image**, IEEE Transactions on Pattern Analysis and Machine Intelligence,v.17, Issue 3, March 1995 Page(s):312 – 315

W. Niblack, **An Introduction to Digital Image Processing**, Prentice-Hall, Englewood Cliffs, NJ, 1986 pp. 115–116.