Desenvolvimento de uma ferramenta computacional para recuperação e correção de textos digitalizados

Marlon Marcon [1]
André Luiz Brun [2]
Jorge Bidarra [2]

Resumo: Atualmente, existem muitas ferramentas voltadas para a identificação e o reconhecimento de palavras em arquivos digitais. Não obstante, um dos maiores desafios enfrentados pelos projetistas desses sistemas tem sido em relação ao tratamento de ruídos, normalmente produzidos durante o processo de digitalização. Essas inconveniências fazem com que os algoritmos de reconhecimento digital apresentem resultados abaixo do esperado. Este trabalho implementa uma solução para um reconhecimento mais adequado de palavras. Para tanto, são aplicados algoritmos de remoção de ruído e de melhoria no contraste. Para que esse objetivo seja alcançado pelos algoritmos, foi associado um lexico ao módulo de reconhecimento de caracteres. Os resultados obtidos a partir dos testes de validação e correção das palavras vêm apresentando resultados satisfatórios, com taxas de aproveitamento, se não ideais, pelo menos em níveis aceitáveis para um sistema com esse tipo de complexidade.

***Palavras-chave*:** Correção ortográfica. Histogramas de projeção. Léxicos. OCR.

Abstract: *Nowadays, there are many tools aimed at identification and word recognition in digital files.Nevertheless, one of the biggest challenges faced by designers of these systems have been in relation of the noise processing, usually produced during the scanning process. These inconveniences make the digital recognition algorithms performing below expectations. This work implements a solution to a more adequate word recognition. Therefore, algorithms are applied to remove noise and improve the contrast. To achieve this goal, we associate to the algorithms a lexicon module of character recognition. The results from the validation tests and correction of words have been showing satisfactory results, with success rates, that if are not ideal, at least at have acceptable levels for a system with this kind of complexity.*

***Keywords*:** *Lexicons. OCR. Projection histograms. Spelling correction.*

1 Introdução

Muitos sistemas de reconhecimento de caracteres (Optical Character Recognition- OCR), quando trabalham com imagens nítidas, limpas e com bom contraste, tendem a produzir resultados efetivos e bastante satisfatórios. Quando, no entanto, se deparam com imagens ruidosas ou danificadas, isso provoca um forte descontentamento junto ao usuário do aplicativo. Para que esses sistemas possam produzir, cada vez mais, resultados eficientes, quando da captura e do registro das palavras, várias técnicas de processamento de imagens para remoção de ruídos (como os filtros da média, mediana, kfill ou erosão) e operações de contraste (como a equalização do histrograma) podem ser utilizadas, no intuito de contribuir com o processo de análise do texto.

Os problemas de reconhecimento encontrados em material impresso de baixa qualidade podem prejudicar o desempenho da aplicação e produzir resultados inadequados ou indesejáveis. Quando o sistema realiza o reconhecimento dos caracteres desconsiderando a palavra composta por eles, podem ocorrer ambiguidades nos resultados, pois as letras poderão ser trocadas e até mesmo não reconhecidas ([1]).

[1]Instituto Federal de Minas Gerais (IFMG), Bambuí (MG) - Brasil
{marlon_marcon@hotmail.com}
[2]Colegiado de Ciência da Computação, Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel (PR) - Brasil
{andre.brun, jorge.bidarra@unioeste.br}

http://dx.doi.org/10.5335/rbca.2013.2719

Para sanar esses problemas, sistemas de reconhecimento aliam OCR e léxicos como parte importante do processo. Enquanto a função do OCR é capturar das imagens digitalizas os caracteres e tentar reconhecer as palavras presentes na imagem, cabe ao léxico auxiliar na seleção e disponibilização de palavras candidatas a resultado e, a partir de suas microcaracterísticas, escolher quais são as palavras propostas como solução ([1], [2], [3], [4]).

O léxico computacional é uma base de dados estruturada, na qual estão armazenados os itens lexicais de uma língua e informações gramaticais e de significados a eles correspondentes. Dentre tais informações lexicais destacam-se as referentes às categorias gramaticais, além de outras características como gênero, número, grau, pessoa, tempo, modo etc ([5]).

De forma a contribuir para o processo de reconhecimento e validação de palavras em textos digitalizados, bem como a recuperação de textos em documentos com presença de ruídos, neste artigo apresentamos o desenvolvimento de uma ferramenta computacional qua alia técnicas de processamento de imagens com analisadores léxicos para a língua portuguesa.

Para essa apresentação, o artigo assim se organiza: A seção 2 discorre sobre a metodologia aplicada na solução do problema, descrevendo as decisões adotadas no projeto da ferramenta. Na Seção 3 é apresentada a ferramenta obtida e a Seção 4 apresenta os principais resultados obtidos até o momento com a aplicação do sistema desenvolvido. Por fim, na Seção 5 são apresentadas as considerações finais da pesquisa.

2 Materiais e métodos

Visando desenvolver uma metodologia para melhoria das caraterísticas e posterior interpretação das imagens de entrada, propõe-se aqui um sistema de processamento e análise de imagens. O sistema foi projetado para ser executado em cinco etapas, de acordo com a sugestão fornecida em Gonzales e Woods [6]. A Figura 1 ilustra o esquema de execução da ferramenta proposta, cujas etapas de processamentos são, na sequência, explicadas.

2.1 Etapa de Captura do Texto

Embora a etapa de captura do texto de entrada faça parte do processamento geral, essa parte não foi implementada, cabendo ao usuário fornecer ao sistema a imagem já digitalizada (seja por scanner, câmera ou outro dispositivo óptico de captura de caracteres), levando em consideração a importância da qualidade da imagem para o processo como um todo. Como entrada, são aceitas imagens nos formatos bmp, png e jpg.

2.2 Pré-processamento

Para remover os ruídos presentes no texto digitalizado, um módulo de pré-processamento foi implementado. As operações realizadas nessa fase são abaixo descritas.

A primeira etapa do pré-processamento corresponde à quantifização da imagem para tons de cinza. O objetivo com essa transformação é reduzir ao máximo a quantidade de cores, o que facilita a identificação e a recuperação das informações digitalizadas, para o processamento que se segue.

Para a transformação das cores originais do texto para tons de cinza foi utilizada a seguinte equação ([7],[8]):

$$g(x, y) = 0,114R(x, y) + 0,587G(x, y) + 0,299B(x, y). \quad (1)$$

Uma vez a imagem já equalizada em tons de cinza, constrói-se o histograma e, então, efetua-se sua equalização, adotando-se a função da frequência acumulada (CDF) ([6], [9]). Essa operação é realizada buscando maior contraste entre o texto e o fundo, pois, dependendo da imagem, esses podem ser confundidos.

Obtida uma distribuição mais uniforme das cores, realiza-se então o processo de remoção de ruídos. Essa etapa adota os filtros da média e mediana, que são aplicados com finalidade de remover ruídos representados por pontos isolados na imagem. Nesse ponto, vale mencionar que ao usuário é dada a opção de escolher o tamanho desejado para a dimensão da máscara. Como valor padrão para os métodos, foi fixado um tamanho de máscara

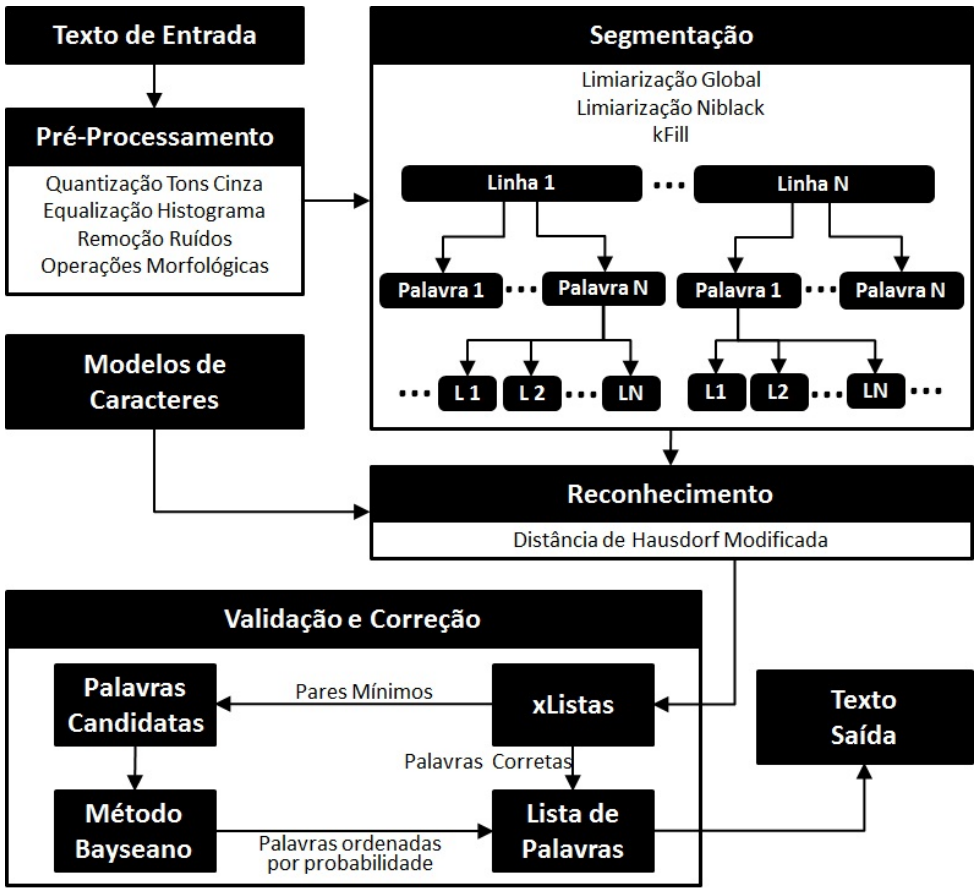

Figura 1: Estrutura do funcionamento do software

de $3x3$ pixels, o que se justifica em razão de que um filtro desse tamanho permite a remoção de pequenos ruídos preservando as bordas dos objetos ([10],[11]).

A última etapa do pré-processamento corresponde à aplicação das operações morfológicas. As opções apresentadas na ferramenta desenvolvida quanto à morfologia dos objetos compreendem os quatro operadores básicos (dilatação, erosão, abertura e fechamento), sendo que em cada uma delas é possível a definição, via interface, do elemento estruturante que realizará a operação ([6], [9]). A aplicação dessa etapa é facultativa, cabendo ao usuário a opção por adotá-la. Essa abordagem foi utilizada em razão de que não há, em imagens onde a separação das palavras e caracteres é nítida, a necessidade da adoção da morfologia matemática. Entretanto, em imagens onde os caracteres podem estar conectados, a utilização da morfologia pode prover melhor interpretação das palavras.

2.3 Segmentação

O objetivo principal da fase de segmentação é a divisão da imagem em sub-imagens contendo os elementos estruturais do problema, no caso, caracteres. Um segmento é definido como um par de coordenadas (x, y) que determinam onde este começa (ponto superior direito) e termina (ponto inferior esquerdo).

Para iniciar o processo de segmentação, a realização da limiarização é de ampla utilidade, visto que esta visa separar o texto do fundo, gerando uma dicotomia na imagem, onde o fundo receberá a coloração branca e os objetos a cor preta. Para essa solução, duas variações para o método são apresentadas, a limiarização global clássica (onde todos os pixels da imagem são submetidos a um mesmo limiar) e a limiarização local e adaptativa chamada Limiarização Niblack proposta por Niblack [12].

A aplicação da limiarização global é indicada quando a imagem apresenta objetos (caracteres) bem defi-

nidos em relação ao fundo. Quando o usuário adota essa abordagem de limiarização, a ferramenta apresenta, em sua interface, a opção de especificar o limiar a ser aplicado no processo de acordo com o resultado apresentado imediatamente, de forma que a discrepância entre a região de interesse e background seja segregativa.

Por outro lado, a limiarização Niblack é indicada para imagens que apresentam um baixo nível de contraste, podendo ainda remover alguns ruídos da imagem. Como geralmente os caracteres em um documento são relativamente pequenos, o tamanho da janela para o método de Niblack foi definido inicialmente para dimensões $15x15$ pixels ([13]), no entanto, o usuário tem a possibilidade de modificar esses valores para que se adequem à sua imagem. Outro valor que pode ser modificado é o limite inferior para o desvio padrão de uma determinada janela, ou seja, quando os níveis de cinza de uma imagem possuem alta similaridade, o desvio padrão entre eles apresenta valores próximos de zero, e isso faz com que o resultado da limiarização seja menos efetivo. Essa alteração foi realizada buscando contornar este problema, tornando, portanto, a imagem mais limpa, dando prioridade a altos níveis de contraste na operação.

Efetuada a limiarização, a imagem pode ainda conter ruídos (pixels pretos que não constituem os caracteres). Para tentar reduzir ao máximo o nível de perturbações na imagem, é aplicado o algoritmo kFill proposto por O'Gorman [14]. O autor propõe a adoção de uma máscara de $5x5$ pixels, porém, optou-se por permitir ao usuário o efetuar, via interface, o dimensionamento desta. A adoção do filtro kFill se deu, segundo Parker [15], em razão de que este permite a remoção de ruídos em imagens que contenham objetos pequenos como caracteres.

A etapa seguinte consiste na segmentação da imagem em linhas, palavras e caracteres. Uma abordagem interessante para este contento é a adoção dos histogramas de projeção ([16]). Realizou-se este processo em três fases, apresentadas a seguir.

O primeiro passo consiste na segmentação das linhas. Partindo do princípio de que as imagens devam estar alinhadas ao eixo x, a segmentação das linhas é realizada levando em consideração a presença de baixas frequências horizontais. Faz-se a contagem dos pixels horizontalmente e constrói-se sua projeção, que é então analisada. Nessa fase verifica-se a presença de descontinuidades que indicarão o ponto de separação entre duas linhas. Tal cenário pode ser melhor compreendido a partir da Figura 2. A operação número 1 indica onde foram identificados os pontos de descontinuidade no histograma de projeção e, com base nesses, os locais onde há o limite entre duas linhas (demarcados pelas linhas horizontais em amarelo).

Quando a altura de uma linha é muito pequena, é verificado se esta linha está próxima de outra. Se estiver, assume-se que tal linha faz parte da outra e as duas são agrupadas, formando uma linha única. Entretanto, se uma linha for pequena e estiver a uma distância que não demonstre que ela pode fazer parte de outra linha do texto (acima ou abaixo), o sistema interpretará que essa linha isolada é causada por ruídos e, portanto, será excluída.

Identificadas as linhas que compõem a imagem, o passo seguinte é a segmentação das letras. Cada linha encontrada na etapa anterior é submetida à construção do histograma de projeção vertical. O processo de análise é semelhante à fase anterior: são verificados os pontos onde há descontinuidade da projeção para determinar o local de divisão dos caracteres. Essa abordagem é apresentada na Figura 2 (operação número 2).

A divisão dos caracteres, porém, não trabalha de forma independente. Esse fato ocorre devido ao fato de que a descontinuidade no histograma vertical pode indicar tanto uma separação entre caracteres como uma divisão entre duas palavras. Dessa forma, a segmentação faz uma análise da dimensão das descontinuidades do histograma. Nessa análise, o processo identifica dois tipos de descontinuidades: as pequenas, que caracterizarão a separação entre caracteres de um mesmo vocábulo, e as grandes, que indicarão a separação de duas palavras dentro de uma linha. Importante destacar que não foram empregados valores estáticos nesse processo, pois as distâncias entre caracteres e palavras podem variar entre dois textos distintos. Dessa forma, os valores de classificação para caractere e palavra são obtidos pela análise do histograma de projeção vertical de cada texto especificamente.

2.4 Reconhecimento de caracteres

Uma das etapas mais importantes para esse tipo de tratamento diz respeito ao reconhecimento dos caracteres. O objetivo dessa fase é identificar a correspondência correta de cada segmento obtido na etapa de segmentação, ou seja, descobrir qual caractere está representado em cada um dos segmentos. Como métrica para esse processo adotou-se a distância de Hausdorff modificada, proposta por Dubuisson e Jain [17]. O método consiste em calcular a distância euclidiana média entre dois conjuntos sobrepostos de pontos, de forma que, quanto mais similares

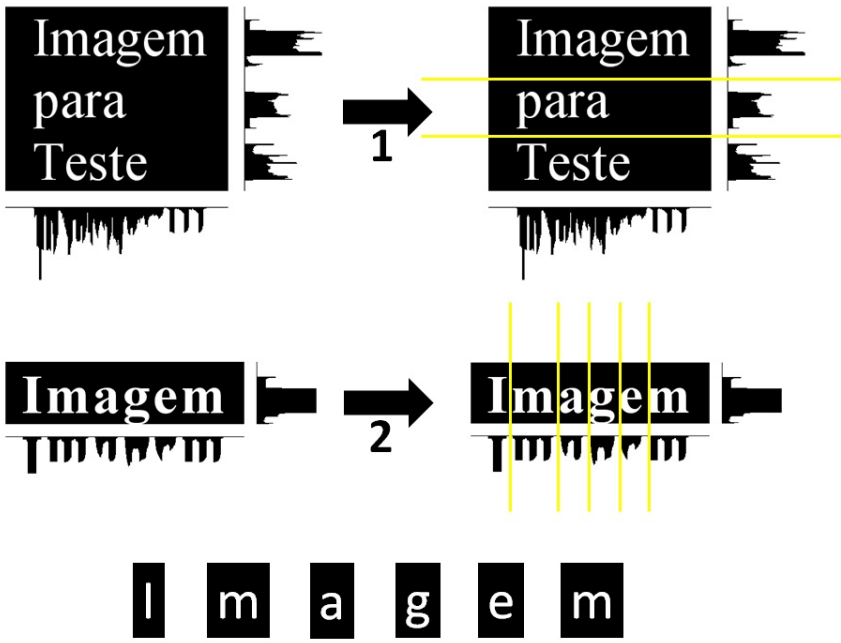

Figura 2: Histogramas de projeção na segmentação do texto

forem os conjuntos, menor será o valor médio obtido.

Visando identificar os caracteres presentes nos segmentos é preciso efetuar-se a comparação desses com modelos pré-definidos. Assim, foi necessária a construção de modelos que pudessem representar, de forma genérica, os caracteres da língua portuguesa. Foram definidos modelos individuais para cada letra do alfabeto (considerando distinção entre maiúsculas, minúsculas e acentuadas) e também para números cardinais (0 a 9). Cada modelo consiste em uma sub-imagem com altura de 14 pixels (tal valor foi obtido empiricamente) e largura variável de acordo com o caractere representado. Adotou-se a fonte Arial como tipo de fonte padrão para a construção dos modelos.

Para reconhecimento de um segmento, calcula-se a distância Hausdorff modificada entre a sub-imagem segmentada e cada modelo definido e vice-versa. A letra reconhecida será aquela que apresentar o menor valor, dentre os caracteres comparados. Calcula-se as distâncias em duas vias, entre o objeto e o modelos e entre os modelos e o objeto, e, após isso, computa-se a média.

Para efetuar o reconhecimento de um segmento, calcula-se a distância de Hausdorff modificada entre a sub-imagem e o modelo definido de forma bidirecional, ou seja, inicialmente calcula-se a distância entre os pixels do segmento e o modelo e, em seguida, computa-se a distância entre os pixels do modelo e do segmento. Obtidos os dois valores, faz-se a divisão de forma a obter um valor médio. Essa estratégia foi adotada devido a erros de reconhecimento causados na comparação unidirecional do objeto com os modelos. Caso essa abordagem não fosse adotada, o sistema poderia, por exemplo, retornar erroneamente uma letra “C” ao invés de retornar o caracter “O”. Esse fato ocorreria em virtude de que os pixels que representam uma letra “C” podem estar contidos num conjunto de pixels que compõem a letra “O”. A recíproca, porém, não é valida, visto que o conjunto de pixels da letra “C” não pode abranger todos os elementos necessários para compor uma letra “O”.

2.5 Validação e correção

Mesmo com o processamento de reconhecimento dos caracteres, é possível que alguns não sejam reconhecidos. Para verificar a correção dos resultados e buscar corrigir erros do reconhecimento, aplica-se o xListas, que, aliado ao método bayesiano, será o responsável pela correção da ortografia.

O xListas é uma base lexical de dados que armazena algumas características fonético-fonológicas da língua

portuguesa. Foi desenvolvido pelo Grupo de Inteligência Aplicada (GIA) da Unioeste, utilizando como base o Listas, um léxico eletrônico desenvolvido pela equipe do Laboratório de Fonética e Psicolinguística (LAFAPE) do Instituto de Estudos da Linguagem (IEL) da Unicamp.

Os dados tratados no xListas correspondem ao mesmo conjunto de dados contido no Listas, porém, a estruturação dos dados e o ambiente de execução foram modificados, visando à redução do volume de dados e à adequação do sistema para padrões atuais de desenvolvimento, como a portabilidade do sistema ([18]).

A base de dados lexical conta com uma quantidade significativa de vocábulos, de aproximadamente 40.000 verbetes, no entanto, somente as formas infinitivas dos verbos são armazenadas, não contendo as flexões verbais, tão pouco palavras no plural, no particípio, no gerúndio etc ([18]). Essa base contém características fonético-fonológicas das palavras, tais como:

- Transcrição ortográfica: corresponde à palavra em si;
- Transcrição fonêmica: os fonemas que compõem uma palavra;
- Máscara ortográfica: a disposição das vogais e consoantes na palavra;
- Máscara da transcrição fonêmica: classificação dos fonemas em vogal, consoante inicial e consoante final de cada sílaba;
- Número de sílabas;
- Classe gramatical: a classe gramatical de uma palavra, como substantivo, adjetivo, verbo etc;
- Acentuação: classificação em oxítonas, paroxítonas, proparoxítonas e átonas;
- Pares mínimos: uma palavra é considerada par mínimo de outra quando estas se diferem apenas por uma letra.

Os passos para a validação são os seguintes:

- Para verificar se o resultado está correto, é realizada uma consulta à base do xListas. Se a consulta à palavra retornar sucesso, considera-se que essa é a palavra correta;
- Se a consulta inicial não retornar resultados, assume-se que a palavra consultada está incorreta;
- Inicialmente a correção oferecida pelo sistema é realizada considerando apenas erros causados somente por substituição (ou seja, um caractere incorreto no lugar de outro correto). Se a palavra não ocorrer na base e todos os caracteres da palavra forem reconhecidos pelo sistema, outra consulta deverá ser realizada na tentativa de se encontrarem os pares mínimos da palavra;
- Se nenhuma palavra retornar da consulta por pares mínimos, considera-se a palavra como errada ou não reconhecida, no entanto, se os pares mínimos são retornados, realiza-se a classificação das palavras pelo método bayesiano e uma lista de palavras ordenadas por probabilidade é retornada para que posteriormente o usuário verifique se a proposta de correção é correta.

O objetivo do método bayesiano é classificar uma dada informação. Para erros ortográficos, o dado a ser classificado é um conjunto de letras que formam uma palavra possivelmente errada. Nesse contexto, o ideal é que uma palavra errada seja corrigida, independente do erro encontrado ([19]).

Segundo o autor, para correção de uma palavra, o primeiro passo consiste em descobrir em um vocabulário as palavras candidatas à correção do erro e, então, dentre estas, escolher a palavra com maior probabilidade de ser a correta.

Para simplificar a compreensão do método de busca pelas palavras candidatas, o exemplo a seguir realiza a consulta para a palavra “cosa”. Admitindo-se que qualquer caractere da palavra é um possível erro, obtém-se, realizando a consulta ao xListas, os resultados apresentados na Tabela 1.

Tabela 1: Palavras candidatas à correção da palavra “cosa”

Palavra Incorreta	Correção	Classe Gramatical	Posição do Erro	Erro	Correção
cosa	rosa	substantivo	1	c	r
cosa	rosa	adjetivo	1	c	r
cosa	tosa	substantivo	1	c	t
cosa	casa	substantivo	2	o	a
cosa	roca	substantivo	3	s	c
cosa	coça	substantivo	3	s	ç
cosa	cola	substantivo	3	s	l
cosa	coma	substantivo	3	s	m
cosa	copa	substantivo	3	s	p
cosa	cota	substantivo	3	s	t
cosa	cova	substantivo	3	s	v
cosa	coxa	substantivo	3	s	x

3 A ferramenta desenvolvida

A implementação da ferramenta foi feita com base na orientação a objetos, com o seu código escrito em Java. Para execução de alguns dos métodos, foi utilizada a API Java chamada Java Advanced Imaging (JAI). Os métodos utilizados que são implementados pela API são: quantização para tons de cinza; limiarização global; filtro da média; filtro da mediana; dilatação; erosão e escala.

Os outros métodos, por seu uso não ser tão comum, não estão presentes na API e, portanto, foram implementados. São eles: equalização de histograma; limiarização Niblack; kFill e remoção de pontos.

Além dos métodos de pré-processamento das imagens, foram também implementadas as etapas de segmentação, reconhecimento, validação e correção das palavras.

Nas subseções abaixo, com o intuito de ilustrar não só a execução da ferramenta, mas também o que o usuário tem à sua disposição quando da manipulação desta, algumas telas do sistema são mostradas na sequência.

3.1 Tela principal

A tela principal apresenta as opções referentes ao pré-processamento da imagem para que após isso o sistema realize as outras opções. A tela inicial é representada pela Figura 3.

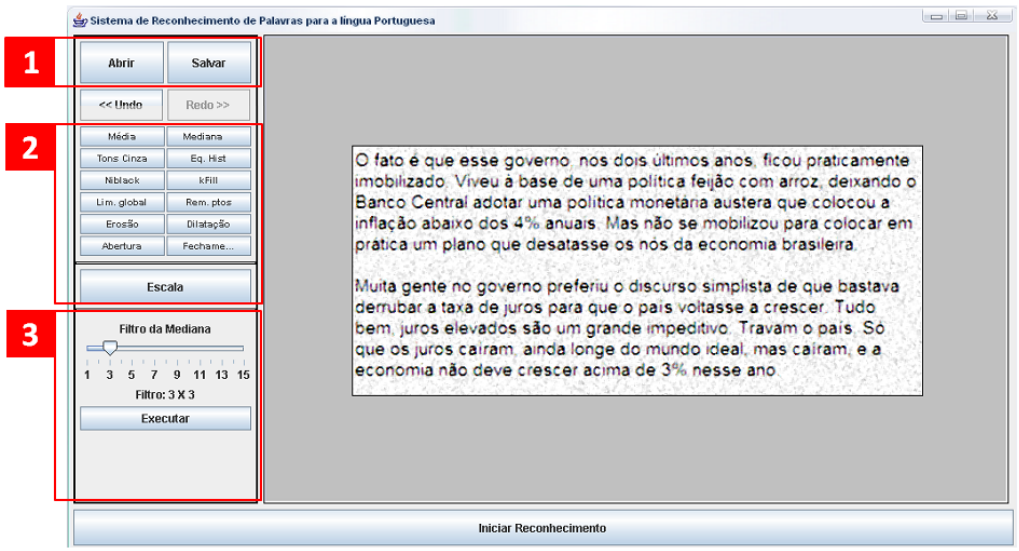

Figura 3: Opções de processamento oferecidas ao usuário

Através da tela inicial, o usuário pode escolher a imagem que deseja reconhecer, bem como pode salvar uma imagem que já passou pela fase de pré-processamento, para que posteriormente essa seja utilizada (marcação 1). Outras opções são os botões de Undo (Desfazer) e Redo (Refazer), que avançam ou retrocedem para as imagens resultado, de operações já realizadas.

Abaixo das opções de desfazer e refazer, são exibidas as opções de pré-processamento, representados pela Figura 3 (marcação 2).

Ao selecionar uma opção, imediatamente aparecem os parâmetros que devem ser definidos para o algoritmo juntamente com o botão de executar, ou, se este não possuir parâmetros de entrada, somente o botão é exibido. A Figura 3 (marcação 3) mostra esse processo para a aplicação de um filtro da mediana, sendo, para tal processo, exibidos componentes da interface responsáveis pela definição do tamanho da máscara utilizada e o botão de executar o processo.

Após todas as operações de pré-processamento realizadas, o usuário seleciona a opção de iniciar reconhecimento e, com isso, a ferramenta inicia a execução das outras operações necessárias para a finalização do processo.

3.2 Tela de resultados

Após realizada as operações de segmentação, reconhecimento e validação, inicia-se o processo, com supervisão do usuário, de correção das palavras. A tela que exibe o resultado preliminar é representada pela Figura 4, que também oferece a opção de salvar o texto que está sendo exibido em arquivo (marcação 1).

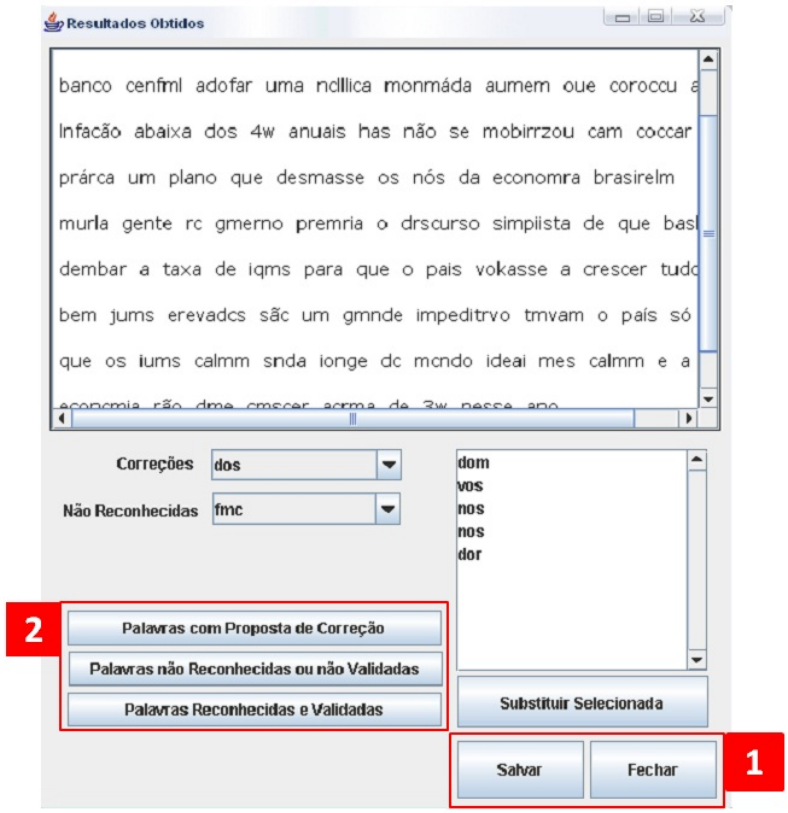

Figura 4: Tela que apresenta os resultados do reconhecimento

Os botões destacados na Figura 4 (marcação 2), quando acionados, permitem a seleção das palavras do texto que estão compreendidas nas três possíveis classes: palavras com proposta de correção, palavras não reconhecidas ou não validadas e palavras reconhecidas e validadas. A Figura 5 mostra um exemplo de seleção das palavras validadas pelo sistema, sendo estas destacadas no texto exibido.

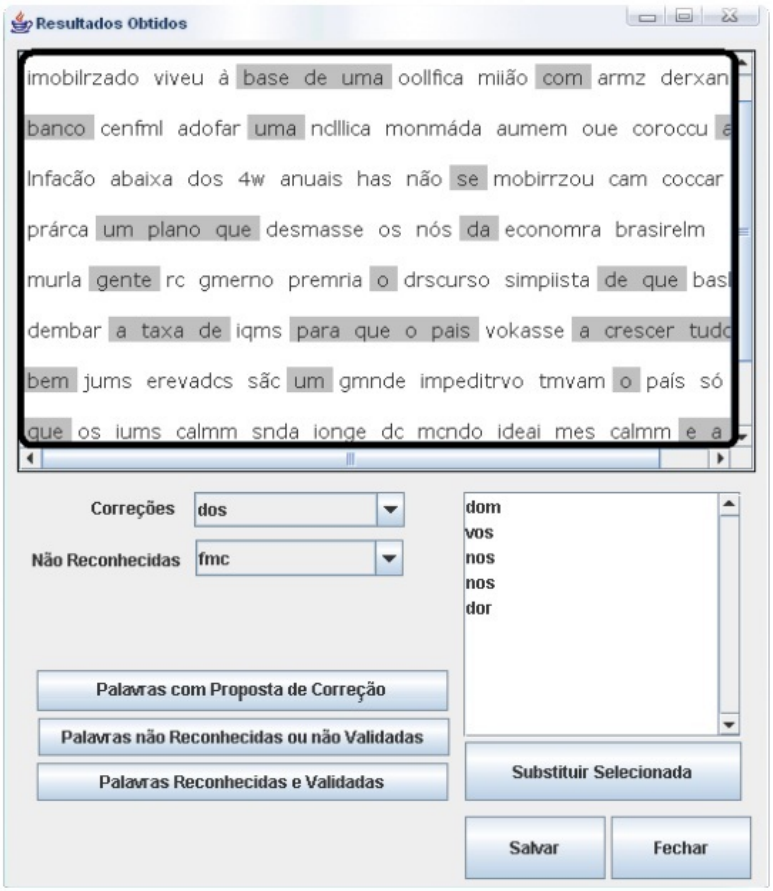

Figura 5: Seleção das Palavras Validadas pelo sistema

Outra opção oferecida é a de correção. As palavras que apresentam propostas de correção são inseridas na caixa de seleção e, quando uma é selecionada, suas possíveis correções são exibidas em uma lista, posicionada à direita e destacada na Figura 6.

Quando o usuário desejar corrigir uma palavra eventualmente não reconhecida, mas que, no entanto, o sistema propôs uma correção, este poderá selecionar a palavra correta e então acionar a opção "substituir selecionada", e o sistema realiza a substituição da palavra incorreta pela selecionada pelo usuário. Por exemplo, a Figura 6 mostra tal operação para correção da palavra "mcndo", suas possíveis correções são as palavras "mundo" e "mando", no contexto do texto reconhecido, a palavra "mundo" é a correta, portanto, esta é escolhida e substituída no texto.

Quando a palavra estiver incorreta e o sistema não oferecer correções válidas, o usuário poderá selecionar a palavra na caixa de seleção correspondente às palavras não reconhecidas. O sistema marcará a palavra escolhida e o usuário pode, então, editar tal palavra no texto, corrigíndo-a.

4 Resultados e discussões

Visando avaliar a robustez do sistema perante diferentes cenários, foram realizados testes sobre diversos textos digitalizados, com intensidade de ruídos variável. O conjunto teste era composto três grupos de vinte imagens, sendo cada um composto de mais de 2000 palavras e possuindo um nível de presença de ruídos (nenhum ruído, presença moderada e alta presença). Os vocábulos presentes nos textos compreendiam variações de gênero, número e flexão verbal, de forma que a base fosse representativa da linguagem normalmente encontrada na prática.

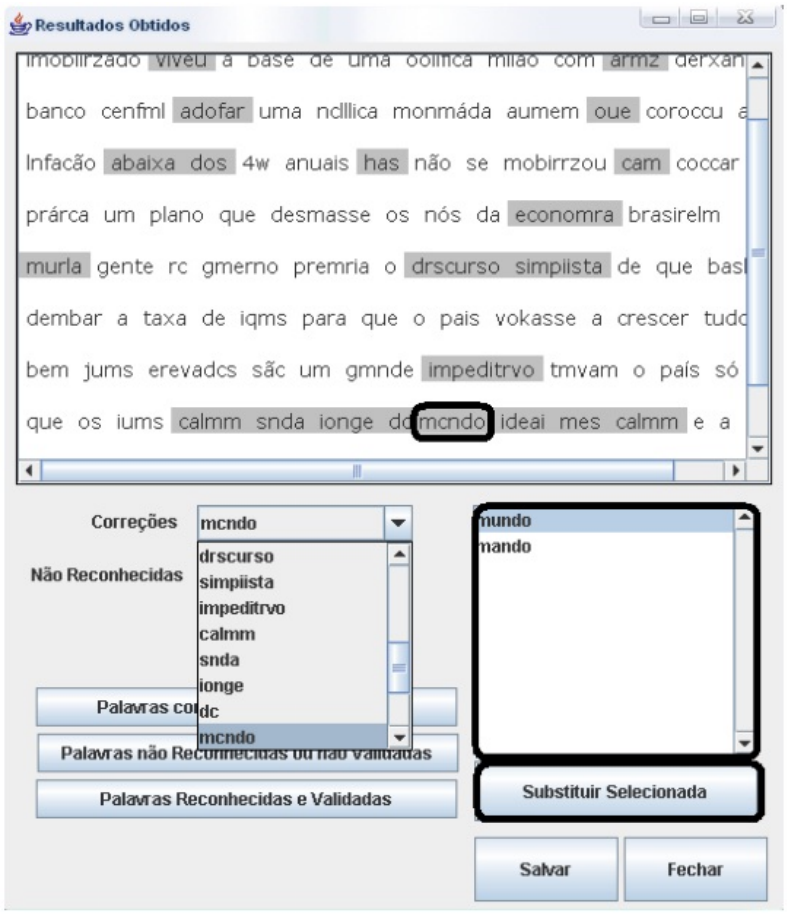

Figura 6: Exibição das possíveis correções para a palavra selecionada

Durante a realização dos testes, com o objetivo de ressaltar as características dos objetos componentes da imagem para melhor interpretação pelo sistema, todas as imagens foram submetidas ao mesmo conjunto de operações de pré-processamento

Nessa etapa, buscou-se quantizar o desempenho da ferramenta, tanto com relação ao reconhecimento pela distância de Hausdorff modificada (DHM), quanto à validação das palavras, através da união do léxico com o método bayesiano. Dessa forma, fez-se levantamento do percentual de palavras reconhecidas de forma correta (através da DHM), do percentual de palavras validadas e também do percentual de falsa validação das palavras não reconhecidas (quando o sistema indica uma palavra diferente da que consta no texto).

Inicialmente foram realizados testes concernentes à proposta de reconhecimento de palavras pela DHM, levando em consideração o reconhecimento de palavras inteiras, foram obtidos os dados apresentados na Tabela 2.

Como observado na Tabela 2, quando o nível de ruído apresentado nas imagens não existia ou era moderado, a taxa de reconhecimento das palavras ficou próxima a 69%, porém, quando este nível aumentou, a taxa caiu

Tabela 2: Dados obtidos com os testes de reconhecimento com a aplicação de DHM nas imagens com diferentes níveis de ruído

Nível de presença de ruído	Palavras reconhecidas (%)	Palavras não reconhecidas (%)
nenhum	68.9	31.1
moderado	69.3	30.7
alto	31.7	68.3

Tabela 3: Dados obtidos com os testes de validação das palavras reconhecidas para nenhum, moderado e alto nível de ruído

Nível de Presença de Ruído	Validadas (%)	Não Validadas (%)
nenhum	70.0	30.0
moderado	75.2	24.8
alto	82.5	17.5

Tabela 4: Percentual de vocábulos validados de forma incorreta (falsa validação)

Nível de Presença de Ruído	Validadas (%)	Não Validadas (%)
nenhum	2.0	98.0
moderado	3.8	96.2
alto	7.2	92.8

significativamente, para 31.7%. Esse fato evidencia que nas imagens com menor quantidade de ruídos ocorre um reconhecimento mais eficiente de suas palavras, aplicando-se apenas a DHM como métrica. Entretanto, em imagens com alta incidência de ruído o resultado é o oposto disso.

Um fator que pode ter influenciado nessa baixa taxa de acerto diz respeito aos algoritmos utilizados na fase de pré-processamento, que, ao tentar remover os ruídos presentes, podem, algumas vezes, remover parte componente dos caracteres, diminuindo a taxa de reconhecimento das palavras. A segmentação também é um fator pertinente na obtenção de tais resultados, pois, em muitos casos, mesmo em imagens isentas de ruídos, ela pode ocasionar a junção de duas letras, resultando em apenas uma, ou, ainda, pode separar uma letra em dois ou mais segmentos diferentes.

As palavras reconhecidas, para que sejam consideradas corretas, necessitam de validação (por meio do analisador léxico). Para isso, cada palavra obtida pelo módulo de reconhecimento do sistema foi verificada, sendo considerada validada ou não. Em caso negativo, a palavra deve ser submetida à aplicação do método bayesiano, o qual lista suas possíveis correções. A Tabela 3 apresenta o percentual de palavras que foram validadas pelo léxico considerando os verbetes presentes na base do xListas.

Como pode ser visto na Tabela 3, a taxa de acerto do xListas para validação das palavras reconhecidas, para os três níveis de ruído, é superior a 70%, sendo este um valor satisfatório se considerarmos que sua base lexical não apresenta verbos conjugados, palavras no gerúndio, particípio, plural etc.

Do ponto de vista da aplicabilidade, contudo, o xListas carece de enriquecimento de sua base lexical de forma a obter maior sucesso com a sua utilização, visto que, durante os testes, verificou-se que a maioria dos vocábulos reconhecidos que não foram validados, eram palavras apresentadas no feminino, no plural ou eram flexões verbais.

Em alguns casos, as palavras eram reconhecidas de forma incorreta (falsa validação), uma vez que o sistema encontrava palavras que estavam presentes na base do xLista mas que não consistiam na palavra presente no texto. A Tabela 4 apresenta o percentual de palavras que foram reconhecidas e validadas de forma equivocada para cada um dos níveis de ruído.

Apesar das taxas de erros relacionadas às falsas validações de palavras serem de no máximo 7.2% para imagens com alto nível de ruído, essas podem ser prejudiciais a sistemas de leitura automática de textos, por exemplo, pois a palavra poderá estar totalmente fora do contexto do assunto, mesmo que perante ao analisador ela esteja correta.

Nos casos em que as palavras não foram reconhecidas, o sistema propunha sugestões de vocábulos que poderiam corresponder à palavra não reconhecida. A ferramenta efetua a verificação na base do xListas em busca de pares mínimos relacionados à palavra em análise. Caso seja encontrado algum par mínimo, este é sugerido como opção para substituição ao termo não reconhecido e validado. Quando há mais de um par mínimo, cabe ao usuário definir qual melhor se enquadra ao contexto.

Acredita-se que os problemas identificados durante o processo podem ser resolvidos, se não no todo, pelo menos em boa parte, pela implementação de um método mais eficiente para a etapa de segmentação. Para tanto,

deve-se realizar um estudo mais aprofundado acerca de métodos de dicotomização e segmentação que podem implicar taxas mais efetivas na separação das palavras e caracteres, evitando principalmente separação em posições incorretas de palavras e a junção de dois caracteres.

5 Conclusão

Este trabalho apresentou o desenvolvimento de uma ferramenta para o reconhecimento de vocábulos que, aliando técnicas de processamento de imagens a um analisador léxico, busca reconhecer palavras em imagens digitalizadas de textos e, em casos de reconhecimento negativo, procura sugerir possíveis correções para a palavra encontrada.

Verificou-se que o método baseado em histogramas de projeção não apresentou resultados satisfatórios para imagens de baixa qualidade, pois baseia-se na descontinuidade dos caracteres, o que pode não vir a ocorrer. Já a distância Hausdorff modificada obteve bons resultados em imagens de melhor qualidade, onde o método de segmentação se mostrou mais efetivo.

A proposta de utilização do léxico, mesmo com todas as suas limitações encontradas, mostrou-se eficiente para a validação dos resultados do reconhecimento. No entanto, para correção, este não apresentou taxas efetivas de sucesso, não só devido aos problemas no léxico (como por exemplo a falta de flexões verbais), mas também à falta de outros métodos de correção de palavras para a ortografia. Uma alternativa para aperfeiçoar o processo de correção seria adotar o tratamento a outros padrões de erros encontrados na ortografia, como, por exemplo, um conjunto de palavras que são comumente escritas erradas, além de apenas analisar a inversão de caracteres.

O baixo desempenho do sistema em fatores, como validação e correção, se deve também ao fato de não ser realizada nem uma análise gramatical nem uma análise contextual. Isso, além de reduzir o número de palavras possíveis para uma correção (no caso de, por exemplo, reconhecimento de textos médicos), reduz significativamente o tempo de resposta e também, as propostas de correção da ferramenta podem ser mais condizentes com o contexto.

O software apresentou a restrição para o tamanho médio dos caracteres. Um conjunto maior de caracteres possibilitaria o reconhecimento de imagens com características variadas, com relação ao estilo ou ao tamanho das letras.

Visando à obtenção de resultados melhores, poderiam ser desenvolvidas futuramente novas adequações ao sistema, tais como o aprimoramento do método utilizado para segmentação, o tratamento dos outros erros propostos para a correção da ortografia, além do enriquecimento da base do xListas, o que aumentaria significativamente a taxa de validação e taxa de correção de palavras, dada a maior proposição de palavras candidatas à correção de um determinado erro.

Referências

[1] CHEN, C.H. Lexicon-driven word recognition. In: Internacional Conference on Document Analysis and Recognition (ICDAR'95), I, 1995, Montreal, Canada *Proceedings...* Montreal: IEEE Computer Society, 1995, p. 919-922.

[2] TULYAKOV, S. e GOVINDARAJU, V. Probabilistic model for segmentation based word recognition with lexicon. In: Internacional Conference on Document Analysis and Recognition (ICDAR'01), VI, 2001, Seattle, United States. *Proceedings...* Seattle: International Association for Pattern Recognition, 2001, p. 164-167.

[3] LUCAS, S.M.; PATOULAS, G. e DOWNTON, A.C. Fast lexicon-based word recognition in noisy index card images. In: Internacional Conference on Document Analysis and Recognition (ICDAR'03), VII, 2003, Edinburgh, Scotland. *Proceedings...* Edinburgh: IEEE Computer Society, 2003, p. 462-466.

[4] BERMAN, B.P. e FATEMAN, R. J. Optical character recognition for typeset mathematics. In: International symposium on Symbolic and algebraic computation (ISSAC'94), Oxford, England. *Proceedings...* Oxford: Association for Computing Machinery, 1994, p. 348-353.

[5] MUNIZ, M. C. M. ; NUNES, M. G. V. ; LAPORTE, E. . Unitex-PB, a set of flexible language resources for Brazilian. In: III Workshop em Tecnologia da Informação (TIL), São Leopoldo, Brasil. *Proceedings...* Sociedade Brasileira de Computação, 2005. p. 2059-2068.

[6] GONZALEZ, R.C. e WOODS, R.E. *Processamento de Imagens Digitais*. São Paulo: Edgard Blucher Ltda, 2000.

[7] GOMES, J. e VELHO, L. *Computação Gráfica: Imagem*. Rio de Janeiro: IMPA/SBM, 1994.

[8] PITAS, I. *Digital Image Processing Algorithms and Applications*. New York: Wiley-Interscience, 2000.

[9] PEDRINI, H. e SCHWARTZ, W.R. *Análise de Imagens Digitais*. São Paulo: Thomson Pioneira, 2007.

[10] ROSA, M.A.; BRUN, A.L. e KIEL, G. Ferramenta Multiplataforma para Construção Automática de Dendogramas a partir de Imagens de Eletroforese. *Revista de Exatas e Tecnológicas - RETEC*, Rondonópolis, v. 2, p.1-10, 2011.

[11] QUEIROZ, J.E.R. e GOMES, H.M. Introdução ao Processamento Digital de Imagens. *Revista de Informática Teórica e Aplicada - RITA*, Porto Alegre, v. 8, n. 1, p. 1-31, 2001.

[12] NIBLACK, W. *An Introducing to Digital Image Processing*. São Paulo: Prentice Hall, 1986.

[13] TRIER, D. e TAXT, T. Evaluation of binarization methods for document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Los Alamitos, v. 17, n. 3, p. 312-315, 1995.

[14] O'GORMAN, L. Image and document processing techniques for the rightpages electronic library system. In: International Conference on Pattern Recognition (ICPR), XI, 1992, Den Haag, Netherlands. *Proceedings...* Den Haag: International Association for Pattern Recognition, 1992, p. 260-263.

[15] PARKER, J.R. *Algorithms for Image Processing and Computer Vision*. New Jerssey: John Wiley & Sons, 1996.

[16] SILVA, E. e THOMÉ, A.C. Reconhecimento de caracteres manuscritos utilizando time de redes neurais. In: Congresso da Sociedade Brasileira de Computação, XXIII, 2003, Campinas, Brasil. *Anais...* Campinas: Sociedade Brasileira de Computação, 2003, p. 1-8.

[17] DUBUISSON, M.P. e JAIN, A.J. A modified hausdorff distance for object matching. In: Internacional Conference on Pattern Recognition, I, 1994, Jerusalem, Israel. *Proceedings...* Jerusalem: International Association for Pattern Recognition, 1994, p. 566-569

[18] BIDARRA, J.; FOLADOR, E.L.; CAVASIN, R. e MARCON, M. Um léxico eletrônico para a língua portuguesa: xListas. In: Encontro Paranaense de Computação, I, 2005, Cascavel, Brasil. *Anais...* Cascavel: Unioeste, 2005, p. 1-9.

[19] JURAFSKY, D.S. e MARTON J.H. *Speech and Language Processing*. São Paulo: Prentice Hall, 2000.