UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE TECNOLOGIA E GEOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

DISSERTAÇÃO DE MESTRADO

ACADEMUS: UMA PLATAFORMA PARA GERAÇÃO DE BIBLIOTECAS DIGITAIS DE TESES E DISSERTAÇÕES.

PAULO HUGO ESPÍRITO SANTO LIMA

RECIFE 2011

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE TECNOLOGIA E GEOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ACADEMUS: UMA PLATAFORMA PARA GERAÇÃO DE BIBLIOTECAS DIGITAIS DE TESES E DISSERTAÇÕES.

POR

PAULO HUGO ESPÍRITO SANTO LIMA

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco como parte dos requisitos para obtenção do grau de Mestre em Engenharia Elétrica.

ORIENTADOR: PROF. DR. RAFAEL DUEIRE LINS.

Recife, Julho de 2011.

Catalogação na fonte
Bibliotecária Raquel Cortizo, CRB-4 664

L732a Lima, Paulo Hugo Espírito Santo.
Academus: uma plataforma para geração de bibliotecas digitais de teses e dissertações / Paulo Hugo Espírito Santo Lima. - Recife: O Autor, 2011.
120 folhas, il., gráfs., tabs., figs.

Orientador: Prof. Dr: Rafael Dueire Lins
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia Elétrica, 2011.
Inclui Referências Bibliográficas e anexos.

1. Engenharia Elétrica 2 .Bibliotecas digitais 3. Análise de imagens de documentos I. Lins, Rafael Dueire (orientador). II. Título.

UFPE

621.3 CDD (22. ed.) BCTG/2011-220

A Deus

AGRADECIMENTOS

Gostaria agradecer primeiramente a Deus, pois nEle me movo, existo e sou (At.17,28). Agradecer especialmente à minha família, aos colegas do dia-a-dia do laboratório de Criptografia (LACRI) e do laboratório de Telemática, aos professores e funcionários do DES/CTG, e aos integrantes do laboratório LIBER, em especial ao Prof. Marcos Galindo.

Gostaria de lembrar os companheiros (alunos e professores) do grupo de comunicações que proveram o importante ambiente de pesquisa que me possibilitou cumprir esse trabalho.

De uma forma especial agradeço ao professor Rafael Dueire Lins, por sua orientação neste trabalho e nas atividades acadêmicas.

Por fim, agradeço o apoio financeiro prestado pela CAPES.

RESUMO DA DISSERTAÇÃO APRESENTADA À UFPE COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA.

ACADEMUS: Plataforma para geração de bibliotecas digitais de teses e dissertações.

Paulo Hugo Espírito Santo Lima.

JULHO DE 2011

ORIENTADOR: PROF. DR. RAFAEL DUEIRE LINS.

ÁREA DE CONCENTRAÇÃO: TELECOMUNICAÇÕES/ ENGENHARIA DE DOCUMENTOS.

PALAVRAS-CHAVE: BIBLIOTECAS DIGITAIS, RECONHECIMENTO DE CONTEÚDO, ANÁLISE DE IMAGENS DE DOCUMENTOS, OCR.

NÚMERO DE PÁGINAS: 120

Trabalhos de Pós-Graduação tais como teses de doutorado e dissertações de mestrado são documentos importantes para diversas áreas das ciências, que precisam ficar disponíveis para acesso remoto para qualquer pesquisador. A plataforma *ACADEMUS* foi desenvolvida com o propósito de gerar semi-automaticamente bibliotecas digitais desses documentos de pós-graduação, capturando as informações relevantes que o documento carrega.

ABSTRACT OF DISSERTATION PRESENTED TO UFPE AS A PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER IN ELECTRICAL ENGINEERING

ACADEMUS: A Platform for the Generation of Digital Libraries of M.Sc. and Ph.D. Theses.

Paulo Hugo Espírito Santo Lima.

JULY 2011

ADVISOR: PROF. RAFAEL DUEIRE LINS, PH.D.

AREA OF CONCENTRATION: TELECOMUNICAÇÕES/ ENGENHARIA DE DOCUMENTOS.

KEYWORDS: BIBLIOTECAS DIGITAIS, RECONHECIMENTO DE CONTEÚDO, ANÁLISE DE IMAGENS DE DOCUMENTOS, OCR.

NUMBER OF PAGES: 120

Postgraduate degrees are one of the most important propellers of all areas of science. M.Sc. and Ph.D. theses witness important developments and provide a solid and global account of research projects. The *ACADEMUS* platform is an environment developed with the aim of generating digital libraries of theses and dissertations, making explicit their relevant indexing information.

Sumário

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

1.1. CONTEXTUALIZAÇÃO

O acesso a documentos e obras é fundamental para a disseminação cultural e científica de uma sociedade. No passado, o acesso a tais documentos era restrito aos que fisicamente se deslocavam à biblioteca. Na atualidade, principal fonte de pesquisas é a Internet, onde qualquer individuo escreve o que pensa, com ou sem coerência, com ou sem responsabilidade. Isso ocorre pela quebra da barreira geográfica em que se acessam conteúdos de diferentes regiões do globo.

Para que se tenha uma fonte segura de informação, várias iniciativas foram tomadas para disponibilizar na Internet o conteúdo dos documentos científicos [1]: as grandes associações e organizadoras de simpósios disponibilizam artigos e publicações; editoras e provedores disponibilizam livros; iniciativas como Google Books [47] disponibilizam obras da atualidade e clássicas que ficaram na história das ciências.

Contudo, ainda hoje a maior parte das teses de doutorado e dissertações de mestrado realizadas em universidades e centros de pesquisa fica armazenada em suas bibliotecas locais, sem a possibilidade de acesso remoto. Novamente o acesso a essas obras é restrito a quem fisicamente pode se deslocar à biblioteca, e muitas vezes essas obras não podem ser manuseadas, o que restringe ainda mais o acesso a elas.

Felizmente, este cenário tende a melhorar. A CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) já começa a disponibilizar as teses e dissertações geradas digitalmente nos últimos anos. Iniciativas como a da UNICAMP[10] e USP[11] disponibilizam o acervo de teses e dissertações na WEB, e

as obras sem cópia digital são submetidas a digitalizações, para serem disponibilizadas. Também serviços de registros de patentes [3] utilizam ferramentas similares.

Na biblioteca física, para se procurar uma obra é necessário ter algumas informações sobre ela, como seu título, o autor, ou um código de arquivamento. Para saber se a obra trata do assunto de interesse é necessário ler a obra. Na WEB para que a obra possa ter "visibilidade" da mesma forma que em bibliotecas físicas, é necessário ter essas mesmas informações sobre a obra para poder encontrá-la no acervo.

O termo indexar designa o fato de associar as informações da obra a própria obra, para poder-se identificá-la no acervo. O processo de indexação na atualidade é manual, ou seja, uma pessoa lê a obra, coleta informações para identificá-la e depois armazena num tipo de banco de dados. Para automatizar o processo de indexação foi desenvolvida a plataforma ***ACADEMUS***[1], que gera uma biblioteca digital de teses e dissertações, com uma coleta semi-automática de informações sobre as obras e gerando o banco de dados do acervo.

Algumas obras não apresentam formato digital, e para isso é necessário que sejam digitalizadas. A fonte principal de digitalização é o *Scanner*, que captura uma imagem de cada página da obra para armazená-la no formato digital. Diretamente da imagem não é possível se capturar informações sobre a obra, como por exemplo, do que ela trata, sendo necessário empreender uma extração de caracteres num processo chamado de Reconhecimento Óptico de Caracteres (*Optical Character Recognition* – OCR). Após o OCR, obtém-se o texto contido naquela imagem, e a partir do texto realiza-se a captura de informações. Dissertações e teses mais

[1] Academus foi um local sagrado para Athena, a deusa da sabedoria na mitologia grega. Lugar onde Platão ensinou, fora dos muros da cidade de Atenas [9].

recentes em formato digital, geralmente em PDF, também necessitam ter o conteúdo de indexação extraído.

Esta dissertação apresenta a plataforma *ACADEMUS*, com todos os processos (características e problemas) envolvidos, aplicada a um programa de Pós-Graduação com teses e dissertações em papel e digital. A plataforma *ACADEMUS* foi usada com sucesso no acervo do Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) da Universidade Federal de Pernambuco (UFPE), onde esta dissertação foi desenvolvida.

1.2. MOTIVAÇÃO

ACADEMUS se propõe a coletar e gerar semi-automaticamente bibliotecas digitais, para que seja disponibilizado na WEB esse acervo de bibliotecas físicas. Diversos problemas aparecem no processo de coleta de informações, em especial pela falta de cópias digitais (algumas obras estão em papel, e precisam ser digitalizadas), e principalmente por causa da falta de padronização na apresentação e distribuição das informações sobre a obra em seu conteúdo.

A geração das bibliotecas digitais, a coleta de informações, e a automatização do processo é a motivação para a elaboração da plataforma *ACADEMUS*. A descrição da plataforma *ACADEMUS*, a definição dos problemas encontrados, e os métodos e técnicas para solucioná-los impulsionaram a realização deste trabalho.

Atualmente, não existe no PPGEE uma biblioteca digital sobre o conteúdo acadêmico já produzido, o que ocorre também em outros programas de pós-graduação. A criação de uma biblioteca digital se faz necessária, porque torna mais eficiente os trabalhos de buscas, pesquisas e o acesso aos trabalhos já desenvolvidos, o que motiva a realização desse projeto.

1.3. OBJETIVOS

O objetivo principal desta dissertação é apresentar a plataforma *ACADEMUS*, descrevendo suas técnicas para captura de informações e geração de biblioteca digital [4], e abordando seu desempenho quando aplicada ao banco de teses do PPGEE-UFPE. São discutidos e avaliados os formatos dos documentos, os problemas encontrados na manipulação desses documentos, principalmente dos documentos digitalizados, que foram submetidos a processo de OCR. É visto como a digitalização insere diferentes tipos de ruído na imagem, e como cada ruído influencia no desempenho do OCR. O tratamento das imagens para remoção desses ruídos efetuados na plataforma também é tratado nesta dissertação.

A proposta desta dissertação é apresentar, analisar e avaliar o desempenho da plataforma *ACADEMUS*, aplicada ao banco de teses e dissertações do PPGEE. Devem ser capturadas informações e classificá-las de acordo com seu conteúdo, sendo essa a forma de avaliação do desempenho da plataforma, a correta captura e classificação de conteúdo. Como objetivo secundário está a avaliação da influência da qualidade dos documentos digitalizados no processo de OCR, mas com objetivo final de capturar e classificar corretamente as informações. Cada técnica para remoção de ruído é avaliada, e são apresentados os resultados sobre o desempenho na captura de informações.

De uma forma geral, o processo do *ACADEMUS* ocorre segundo o fluxograma da Figura 1.1, em que cada tipo de documento (texto ou imagem) é submetido a diferentes processos antes de terem seus conteúdos analisados e capturadas suas informações.

Dentre os documentos utilizados nos testes, os que foram elaborados após o ano 2000 estão no formato Adobe® PDF [2] e são submetidas diretamente aos processos de captura de informações. As obras mais antigas, que só existiam em versão impressa, foram digitalizadas e suas imagens armazenadas como a versão digital da obra. Essas imagens são binarizadas de acordo com o módulo de binarização da plataforma LiveMemory [5], tratadas para remoção de ruído com

técnicas desenvolvidas na plataforma BigBatch [6] [7], e em seguida submetidas a um OCR para extração do texto. Os resultados desse processo são armazenados e submetidos aos processos de captura de informações.

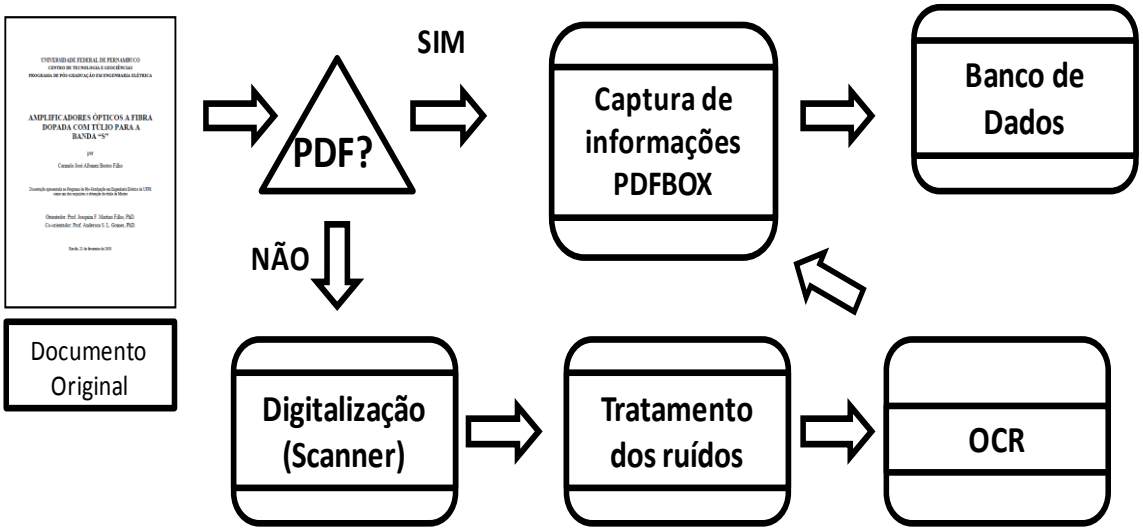


Figura 1.1: Fluxograma simplificado da plataforma *ACADEMUS*.

Nos processos de tratamento de imagens e captura de conteúdo são realizados testes para verificar a influência em processos posteriores e na plataforma geral. Esses testes têm seus resultados apresentados, avaliados e analisados nesta dissertação.

1.4. ORGANIZAÇÃO DO TRABALHO

A seguir é apresentada a estruturação e uma breve descrição do trabalho, sua divisão em capítulos e anexos.

Capítulo 2: Esse capítulo apresenta as características dos documentos usados pela plataforma *ACADEMUS*, isto é, o tipo do documento, a linguagem utilizada, os formatos em que estão armazenados.

Capítulo 3: Esse capítulo apresenta as características das imagens e os processos de tratamento aos quais são submetidas antes de terem capturadas suas informações. É descrito o processo de binarização, de tratamento de imagens para remoção de ruído e de reconhecimento de caracteres pelo OCR.

Capítulo 4: São abordadas as estratégias para captura e classificação das informações de acordo com as características dos documentos.

Capítulo 5: Nesse capítulo são apresentadas as interfaces da plataforma, descrevendo as etapas da plataforma.

Capítulo 6: A metodologia dos testes para avaliação geral da plataforma é apresentada nesse capítulo, juntamente com resultados e avaliações.

Capítulo 7: Esse capítulo condensa as conclusões do trabalho e apresenta futuras perspectivas para a continuidade da pesquisa.

Anexo 1: Estrutura do programa fonte da plataforma ***ACADEMUS***.

Anexo 2: Produção acadêmica fruto do projeto.

2. DOCUMENTOS ANALIZADOS

2.1. SOBRE OS DOCUMENTOS

No *ACADEMUS* a geração de um banco de dados sobre teses e dissertações a partir de informações capturadas dos documentos é o principal objetivo de sua constituição. Essas informações são adquiridas através de uma análise de conteúdo de cada documento, em especial através de palavras chaves, que sempre são utilizadas nesse tipo de documento.

A plataforma *ACADEMUS* foi projetada para documentos científicos do tipo tese de doutorado e dissertação de mestrado. Esses documentos têm linguagem específica e obrigatoriamente apresentam informações que ajudam a os caracterizar e identificar. A Figura 2.1 mostra a capa de uma dissertação de mestrado e seus constituintes. Para esse tipo de documento sempre são informados o programa de pós-graduação e a universidade a qual a obra foi submetida, o título e o autor da obra sempre são apresentados. Na maioria dos casos o nome do orientador da pesquisa fica explícito, em poucos casos o orientador é citado nos agradecimentos. O documento apresenta um texto que deve ser claro e objetivo o que é fundamental para o procedimento de captura de conteúdo, pois não é necessária uma “análise do contexto”. Outras informações que são capturadas sobre os documentos são tratadas em detalhes no capítulo 4.

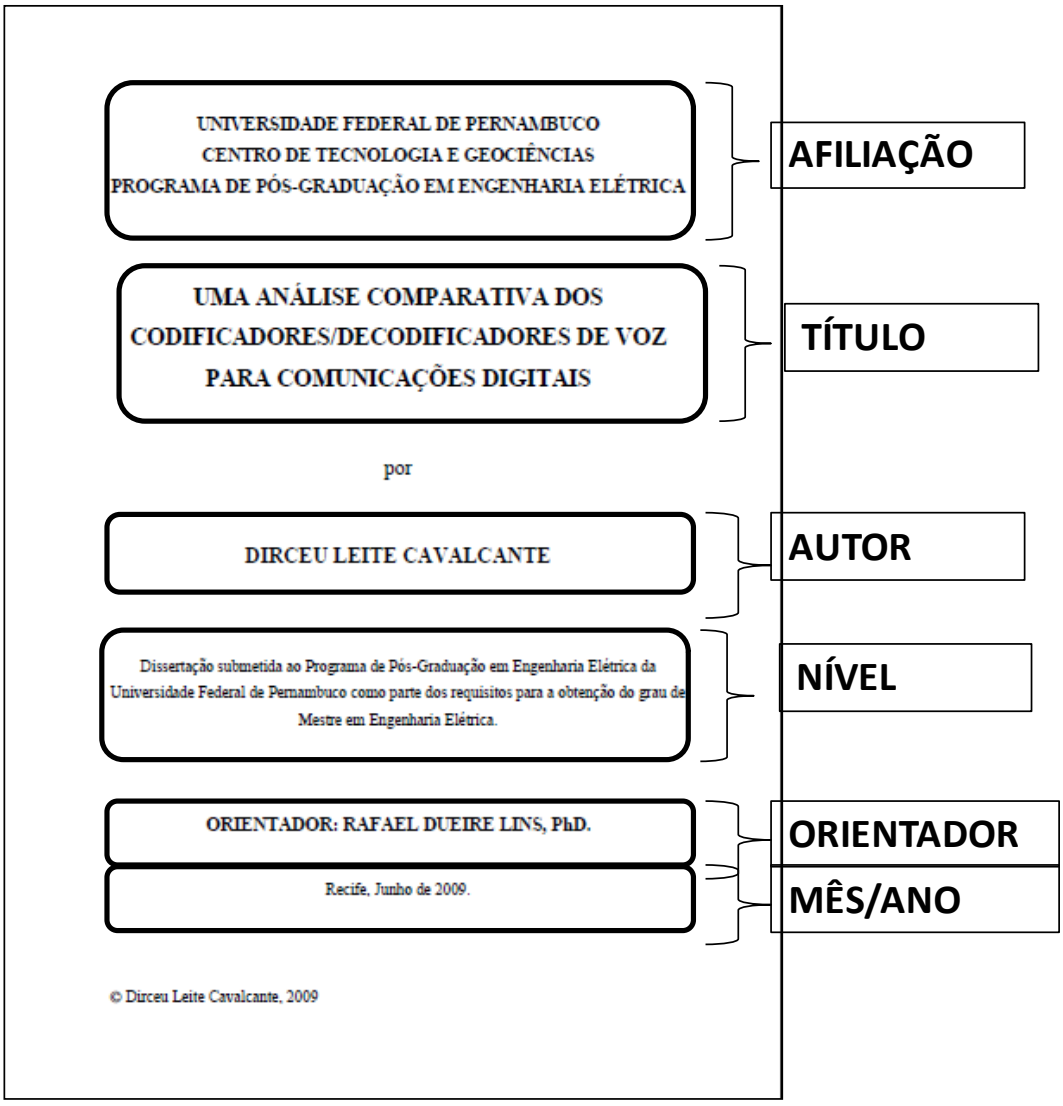


Figura 2.1: Capa de dissertação de mestrado defendida no PPGEE em 2009, que contém informações próprias desse tipo de documento, disponível em versão digital.

Apesar da obrigatoriedade de conter essas informações de forma clara, não há uma padronização na forma de fornecê-la. A Figura 2.2 mostra uma capa de dissertação que contém as mesmas informações que as da Figura 2.1, porém é evidente a diferença na forma que a página foi elaborada e as informações dispostas. Os tamanhos dos textos, a localização do nome do autor, a inexistência do nome do orientador são algumas características que diferenciam os documentos, e é nesse ponto que se encontra a maior dificuldade do processo de captura de conteúdo, a falta de padronização.

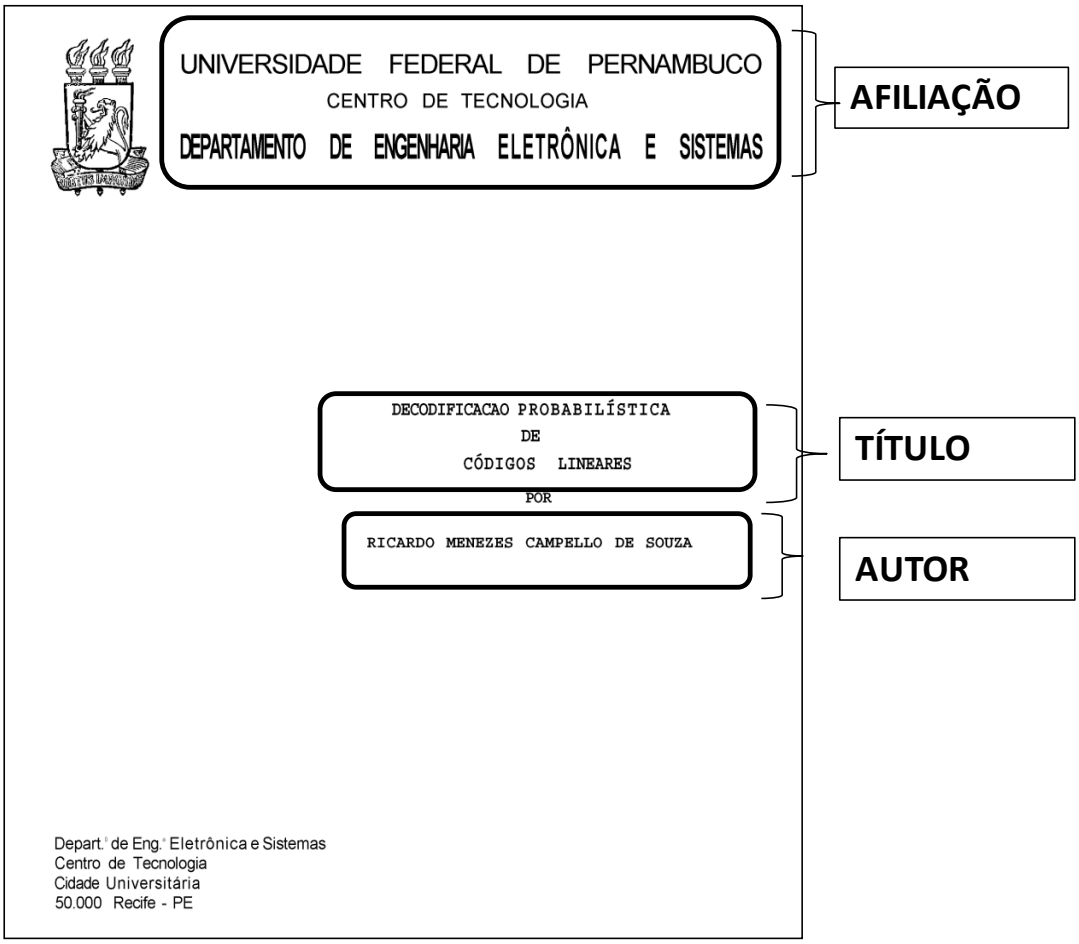
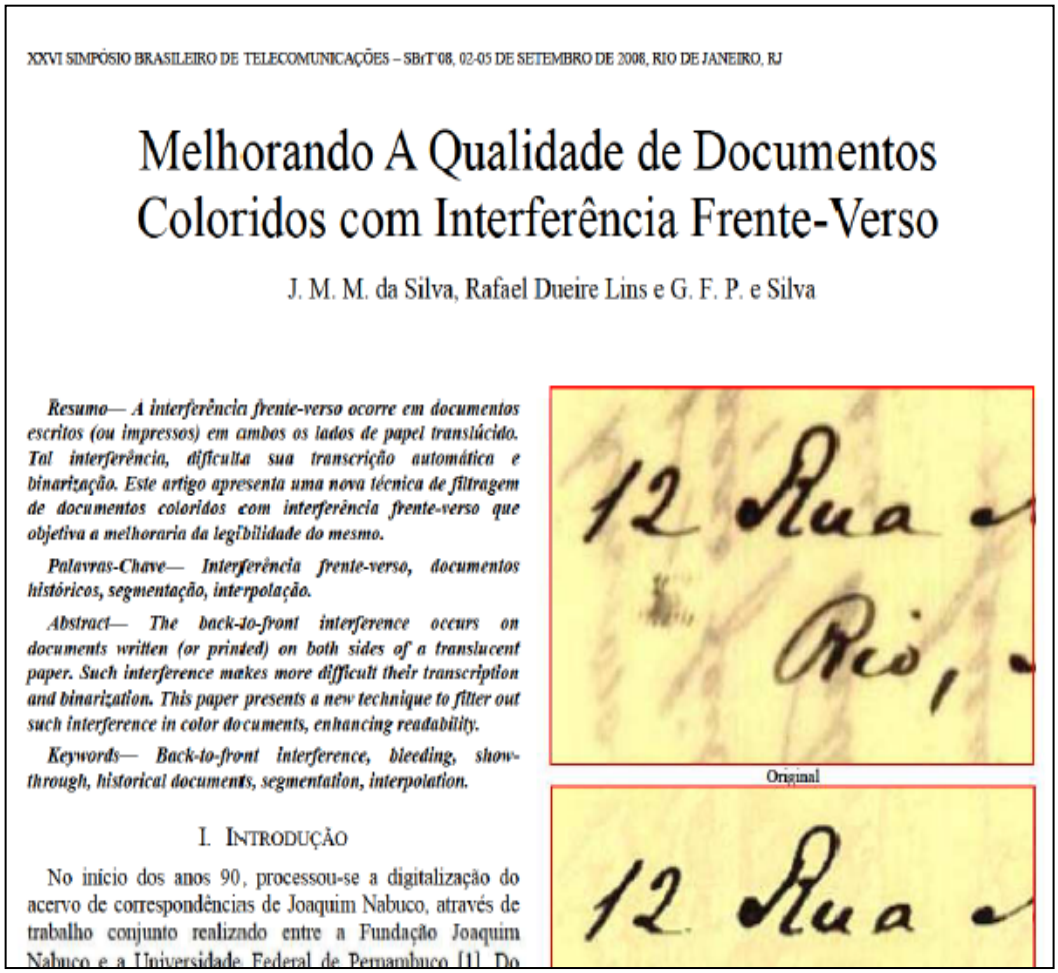


Figura 2.2: Capa da primeira dissertação de mestrado defendida no PPGEE-UFPE em 1979, existente na versão impressa.

Abordagens similares para geração de bibliotecas digitais foram elaboradas para documentos científicos de conferências na plataforma LiveMemory [5], para artigos técnicos. Nesses documentos há também problemas de padronização, com uma sequência de informações que variam entre anos e até mesmo dentro do mesmo volume.

LiveMemory é uma plataforma com o propósito de geração de bibliotecas digitais de anais de eventos científicos. De cada um dos artigos, ela captura informações sobre o título, os autores, resumos, e sobre a edição da conferência. A Figura 2.3 mostra uma cópia de artigo publicado no simpósio da Sociedade Brasileira de Telecomunicações (SBrT) no ano de 2008 e analisado pela plataforma LiveMemory. Percebe-se que na primeira linha está a edição e ano da conferência,

nas linhas seguintes, e com tamanho considerável, está o título do documento, e em seguida estão os nomes dos autores.



XXVI SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES – SBrT'08, 02-05 DE SETEMBRO DE 2008, RIO DE JANEIRO, RJ

Melhorando A Qualidade de Documentos Coloridos com Interferência Frente-Verso

J. M. M. da Silva, Rafael Dueire Lins e G. F. P. e Silva

Resumo— A interferência frente-verso ocorre em documentos escritos (ou impressos) em ambos os lados de papel translúcido. Tal interferência, dificulta sua transcrição automática e binarização. Este artigo apresenta uma nova técnica de filtragem de documentos coloridos com interferência frente-verso que objetiva a melhoraria da legibilidade do mesmo.

Palavras-Chave— Interferência frente-verso, documentos históricos, segmentação, interpolação.

Abstract— The back-to-front interference occurs on documents written (or printed) on both sides of a translucent paper. Such interference makes more difficult their transcription and binarization. This paper presents a new technique to filter out such interference in color documents, enhancing readability.

Keywords— Back-to-front interference, bleeding, show-through, historical documents, segmentation, interpolation.

I. INTRODUÇÃO

No início dos anos 90, processou-se a digitalização do acervo de correspondências de Joaquim Nabuco, através de trabalho conjunto realizado entre a Fundação Joaquim Nabuco e a Universidade Federal de Pernambuco [1]. Do

Original

Figura 2.3: Parte da folha-de-rosto de artigo publicado em simpósio, ilustrando a disposição das informações sobre o documento.

A padronização na disposição das informações auxilia na elaboração de estratégias para captura de informações, algo que não acontece com os documentos analisados pela plataforma ***ACADEMUS***. A determinação de onde começa e onde termina o título ou o autor no documento da Figura 2.2 não é o mesmo da Figura 2.3, na qual se sabe que as informações estão numa determinada região do documento.

2.2. CARACTERÍSTICAS DOS DOCUMENTOS EM PDF

A partir do ano 2000, o PPGEE-UFPE, bem como a maioria dos programas de Pós-Graduação passaram a exigir uma cópia em formato digital das teses e dissertações, que geralmente está no formato Adobe PDF [2], por causa de sua portabilidade, facilidade de acesso aos programas leitores do formato, e pela dificuldade de alteração do documento.

Pode-se fazer uma analogia desse formato como o XML [8], em que as informações são armazenadas em árvore. Por exemplo, para uma linha de texto, é armazenado em um nó da árvore o texto, noutro nó do mesmo ramo é armazenado a posição vertical, noutro é armazenado a formatação, etc., o mesmo acontecendo para figuras, tabelas.

A estrutura de um documento no formato PDF pode ser simplificada em quatro características: o texto, o tamanho do texto, a fonte do texto, a cor do texto, e a posição do texto no layout da página. Na verdade, essas são as características do formato analisadas pelas estratégias de captura de conteúdo da plataforma ***ACADEMUS***. Os documentos que originalmente estão no formato PDF são denominados neste trabalho como “pdf-texto”.

2.3. CARACTERÍSTICAS DOS DOCUMENTOS DIGITALIZADOS

No caso do PPGEE as teses e dissertações defendidas até o ano 2000 só existiam em versões impressas, portanto necessitaram ser digitalizadas. Para isso, foram capturadas imagens de cada página do documento e armazenadas, para terem seu texto extraído posteriormente.

A digitalização foi feita num *scanner* Ricoh Aficio modelo 1075 [32], usando o processo de ADF (*Automatic Document Feed*) [13], com uma resolução de 200 dpi [39] [42], em escala de cinza com 256 níveis, numa taxa de 8 *bits*/pixel, e armazenadas sem compressão no formato TIFF [14], cada documento correspondente a um volume (dissertação ou tese) foi armazenado em um diretório.

Com essas imagens não é possível fazer a captura direta de informações, e por isso é necessário submetê-las a um OCR (extração do texto). Alguns desses documentos antigos estão degradados, a digitalização gerou imagens com “ruído”, dificultando o processo de extração de texto [40]. No intuito de melhorar a “qualidade da imagem” e precisão da extração, as imagens devem ser antes binarizadas (fundo branco e texto preto), e ter removidos vários tipos de ruído advindos da digitalização [15] [41] [46], que dificultam o processo do OCR.

Após ter o texto transcrito pelo OCR são armazenados dois arquivos correspondentes ao documento analisado: o primeiro é uma versão no formato PDF que contém as características já escritas; a segunda, uma versão no formato txt que contém apenas os textos extraídos.

As duas versões são necessárias, pois no processo de OCR ocorre um incomodo erro, espaços em brancos são inseridos entre letras de uma mesma palavra quando convertidos em PDF, o mesmo não ocorrendo no formato txt. Assim, no processo de captura do conteúdo os dois arquivos são analisados e unidos, para melhorar o desempenho do sistema. A Figura 2.4 mostra o efeito de inserção de espaços em branco entre letras de uma mesma palavra quando o texto extraído pelo OCR é armazenado num arquivo do formato PDF.

Pela Figura 2.4, percebe-se a inserção de um “espaço em branco” entre as letras “r” e “o” da palavra “Centro”, ficando claro quando se compara com os espaços entre as letras da palavra “como”. Para uma inspeção visual esse fenômeno pode não se caracterizar como um “prejuízo”, porém, para sistemas baseados em reconhecimento de conteúdo o fenômeno é muito danoso. Por exemplo, caso se faça uma busca de qual centro acadêmico pertence o autor do documento, e a estratégia para a busca seja procurar a palavra “centro” no mesmo parágrafo que palavras como “tese”, “apresentada” e “pós-graduação”, a busca não teria êxito, pois não existiria a palavra “Centro”, existiria sim “C_e_n_t_r_o”, em que “_” denota um “espaço em branco”.

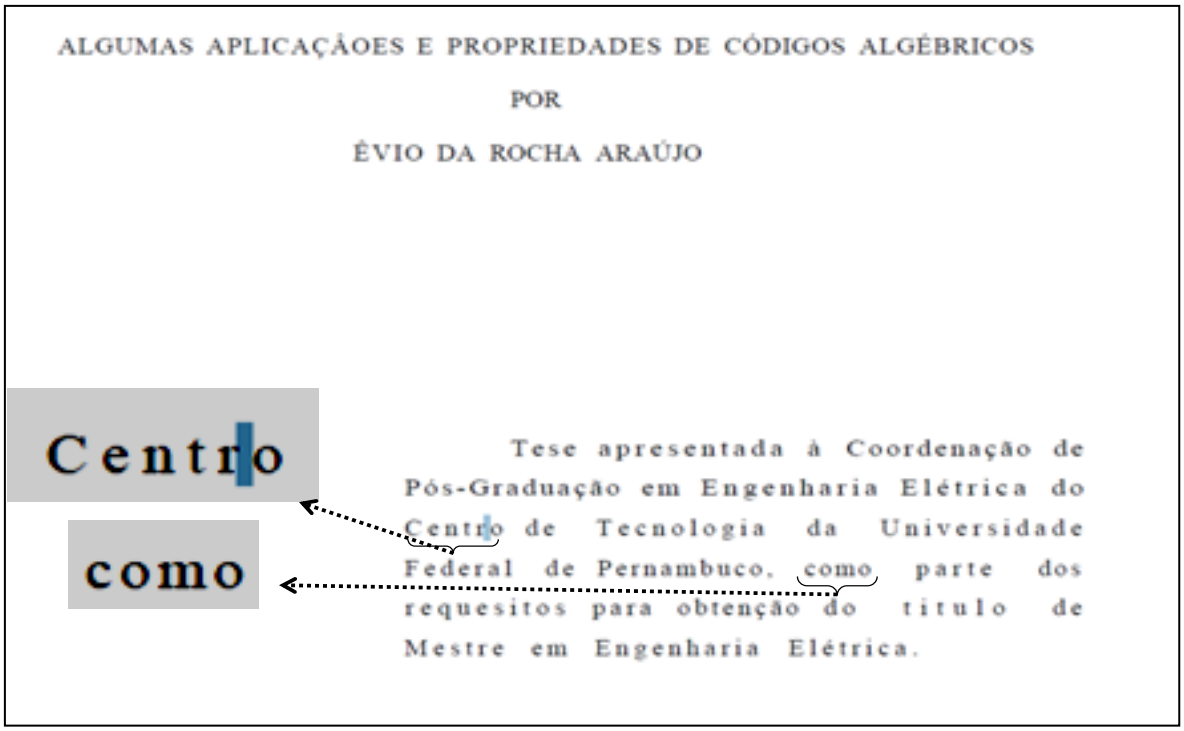


Figura 2.4: Inserção de espaços em branco entre letras da mesma palavra. O detalhe em azul indica um espaço em branco entre letras de uma mesma palavra.

3. IMAGENS DIGITALIZADAS

Como já foi dito, as dissertações e teses anteriores ao ano 2000 do PPGEE-UFPE só existiam em versão impressa e necessitaram de digitalização, que foi realizada por um equipamento chamado de *scanner*. O processo de digitalização de uma imagem é idêntico à digitalização de um sinal, em que a amostragem é feita pela câmera, a quantização é definida pela quantidade de tons de cinza, e a codificação é o formato digital no qual a imagem é armazenada. A Figura 3.1 ilustra com um diagrama de blocos o processo clássico de digitalização de um sinal associado a uma imagem.

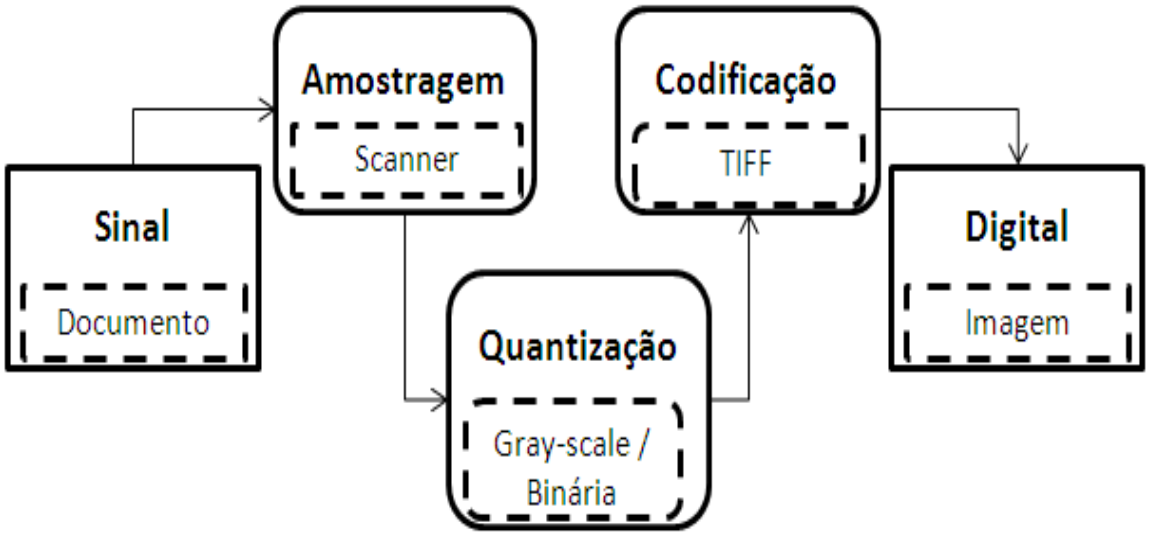


Figura 3.1: Etapas do processo de compressão de sinais.

3.1. DIGITALIZAÇÃO DE UMA IMAGEM

O processo de digitalização pode ser encarado como uma concatenação de três processos: amostragem, quantização e codificação.

3.1.1. AMOSTRAGEM

O processo de amostragem transforma um sinal contínuo, em um sinal discreto, tomando amostras do sinal com uma determinada freqüência de amostragem. No contexto de imagens é feita uma amostragem bidimensional, a freqüência de amostragem pode ser analogamente relacionada com a resolução da imagem, a quantidade de pixels por unidade de área [bib:amostragem].

Amostras da imagem são tomadas por pequenos sensores da câmera, sendo cada amostra denominada *pixel* (*picture element*). Cada *pixel* (amostra) tomado é um elemento de uma matriz bidimensional, em que cada linha é associada à posição vertical e cada coluna a posição horizontal da amostra tomada da imagem. A cada *pixel* é associado uma intensidade numa escala e a faixa do espectro (cor).

Em sistemas monocromáticos existe uma única cor como referência (preto) e variações de cinza até chegar ao branco [12], já em sistemas cromáticos existem diversas escalas de cinza para cada cor do sistema representativo. Para o sistema RGB (*Red, Green, Blue*), um pixel da imagem é composto por uma combinação dos valores da escala de cinza das cores vermelha, azul e verde, já para o sistema CMY (*Cyan, Magenta, Yellow*) há uma combinação das escalas de cinza das cores ciano, magenta e amarela.

Na literatura, o tratamento de imagens monocromáticas é feito diretamente na imagem em escala de cinza. Quando se trata de imagens coloridas, o tratamento é feito na imagem em escala de cinza representativa de cada cor e depois há uma composição. Para a plataforma ***ACADEMUS*** as imagens foram digitalizadas em escala de cinza, desconsiderando os aspectos de cor da imagem, em vista de uma menor complexidade computacional (apenas uma “imagem em escala de cinza” é processada e não há combinação).

A resolução da imagem é uma característica de grande influência para manipulação e tratamentos na imagem. Pode-se enxergar a resolução sobre duas perspectivas: a primeira é a resolução espacial, e é concernente às dimensões de comprimento e altura; a segunda é a resolução de saída, definida como o número de

pixels por unidade de área, tendo a métrica de pontos por polegada *(dots per inch - dpi*) como a mais conhecida.

A Figura 3.2 mostra quatro cópias de uma imagem com resoluções espaciais diferentes. Na primeira (superior esquerda), a imagem original foi digitalizada em 512x512 pixels e depois reduzida para 256x256 *pixels*, segunda imagem (superior direita). A terceira imagem (inferior esquerda) foi obtida originalmente em 256x256 *pixels* e depois ampliada para 512x512 *pixels*, quarta imagem (inferior direita). Nota-se uma degradação quando se passa de um nível de resolução mais baixo para um mais alto, distorcendo os contornos da imagem.

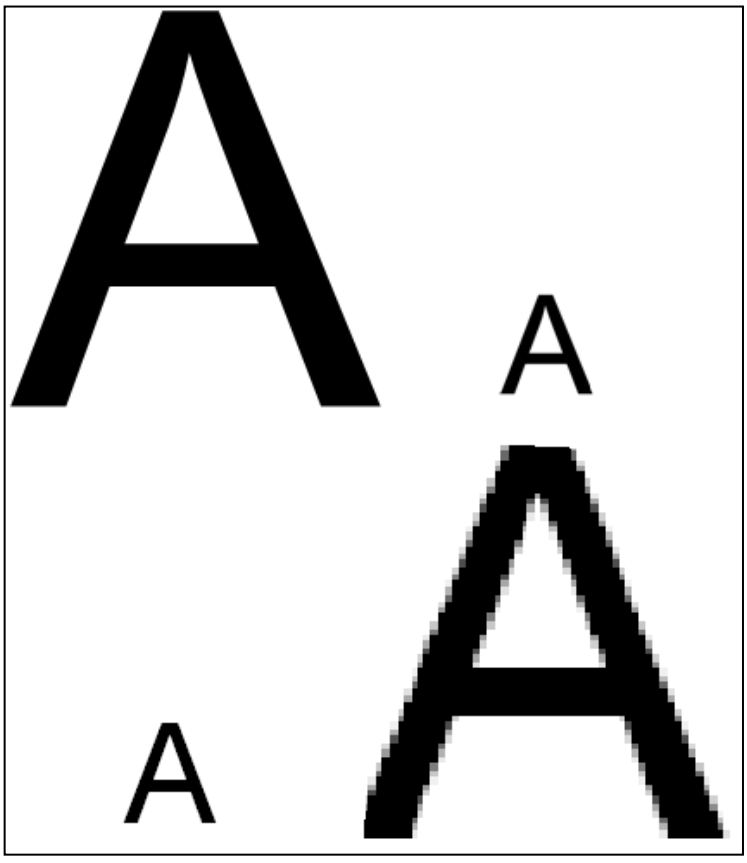


Figura 3.2: Resolução espacial das figuras

A Figura 3.3 apresenta duas cópias de uma imagem que foi digitalizada com 50 *dpi* e 200 *dpi*, respectivamente. Nota-se a baixa qualidade da imagem de 50 *dpi*, no tocante a pouca definição de contornos e formas na imagem. O quadro tracejado mostra que o “tamanho do *pixel*” da segunda imagem é maior que o “tamanho do

pixel" da primeira imagem, quadro com linhas contínuas. Isso significa que se utilizaram mais *pixels* por unidade de área, como se pode observar pelo quadro com linhas contínuas inseridos no quadro tracejado.

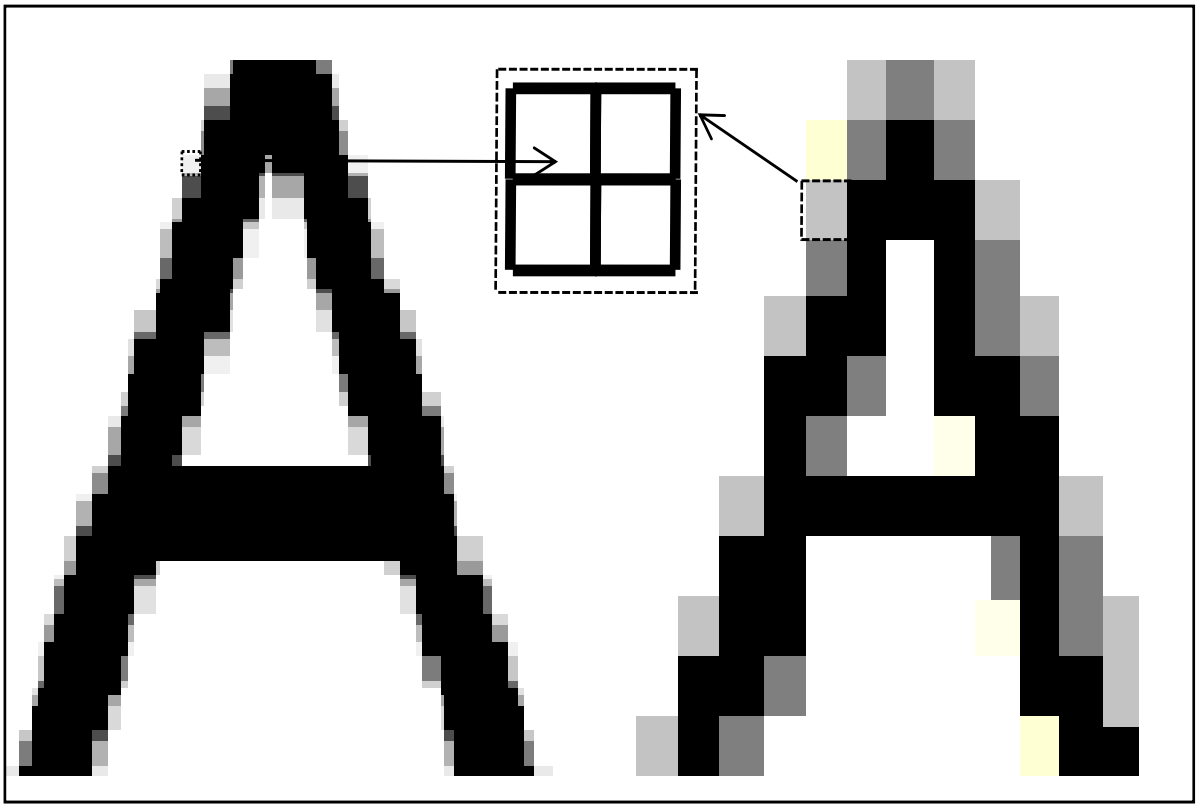
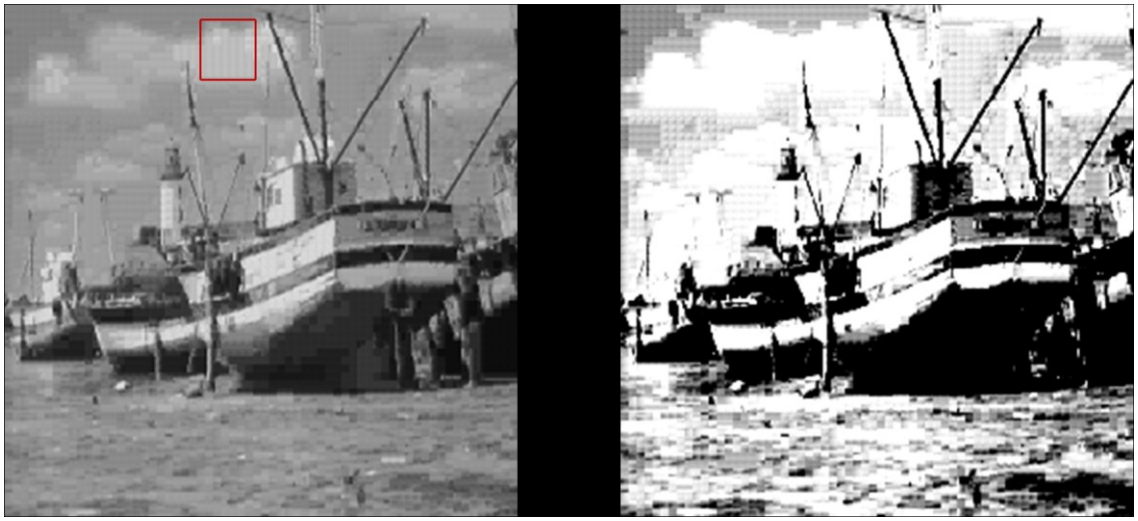


Figura 3.3: Resolução de saída.

3.1.2. QUANTIZAÇÃO

A quantização define o número de tons de cinza que representam uma imagem e o valor possível de um pixel assumir, isto é, a quantização define a quantidade de *bits* utilizada para representar um pixel. O número de *bits* é diretamente proporcional ao número de tons de cinza da imagem, ou seja, quanto mais tons da cor se utiliza para representar a imagem, mais *bits* são necessários. Nesta etapa é inserido o ruído de quantização, a Figura 3.4 ilustra como ocorre o ruído de quantização na digitalização de uma imagem.



(a) (b)

Figura 3.4: Imagem (a) original com 256 níveis de quantização (tons de cinza), (b) quantizada com 32 níveis.

O conjunto de pixels destacado na nuvem na Figura 3.4 (a) assume uma gama de valores de 20 a 100, já na Figura 3.4 (b) assumem o valor zero, pois foram aproximados para o valor mais próximo quando se limitou os níveis de quantização a 32. Essa pequena diferença em relação à imagem original é chamada de ruído de quantização. Quanto menor o número de tons de cinza (níveis de quantização), maior é o valor do ruído de quantização. Um documento que contém pequenos detalhes nas figuras, palavras com cores claras, um fundo mais escuro, pode obter sérios prejuízos com um processo de quantização mal projetado, o que leva a uma perda de informação.

3.1.3.CODIFICAÇÃO

A codificação da imagem é de fundamental importância para o processo de digitalização, pois representa a maneira como as informações obtidas foram armazenadas. À codificação está associado o formato do arquivo da imagem digital, com os valores representativos de cada pixel, a compressão utilizada, o cabeçalho, etc.

O formato utilizado neste trabalho é o TIFF® (*Tag-based File Format*) [14], em que são armazenadas imagens em escala de cinza e sem compressão.

3.2. EXTRAÇÃO DE TEXTO DAS IMAGENS

Na plataforma *ACADEMUS*, a captura de conteúdo (informações) para identificação do documento não é feito através das imagens dos documentos, mas sim através de análise do texto do documento. É necessário, portanto, que se extraiam os textos presentes na imagem e sobre eles capturar informações. O processo de extração de texto a partir de uma imagem é conhecido como Reconhecimento Óptico de Caracteres, OCR (*Optical Character Recognition*).

O reconhecimento óptico de caracteres é inserido na área de reconhecimento de padrões quando se fala em processamento digital de imagens ou visão computacional. Nesse sentido, para que a máquina identifique e classifique um padrão é necessário realizar um treinamento traçando conexões entre as imagens dos caracteres e os próprios caracteres (letras, números, símbolos especiais), para que cada imagem corresponda a um único caractere.

A conexão entre o caractere e a imagem é feita através da criação de uma descrição (conjunto de características da imagem) de cada imagem, que é associada a um caractere. Quando uma imagem é apresentada à máquina, sua descrição é feita (suas características são obtidas) e comparada com as descrições previamente obtidas, associando-a ao caractere que fornecer a melhor correspondência.

A Figura 3.5 mostra um diagrama de blocos do processo de reconhecimento óptico de caracteres.

3.2.1. LOCALIZAÇÃO DO TEXTO

A segmentação é um processo que determina as componentes da imagem. Para aplicações de extração de texto, pode-se distinguir dois processo de segmentação: área do texto; e caractere do texto.

No primeiro, pretende-se localizar as regiões da página que contêm texto das outras, separar a região de texto do fundo da página e de elementos gráficos (figuras, gráficos, logomarcas, etc.).

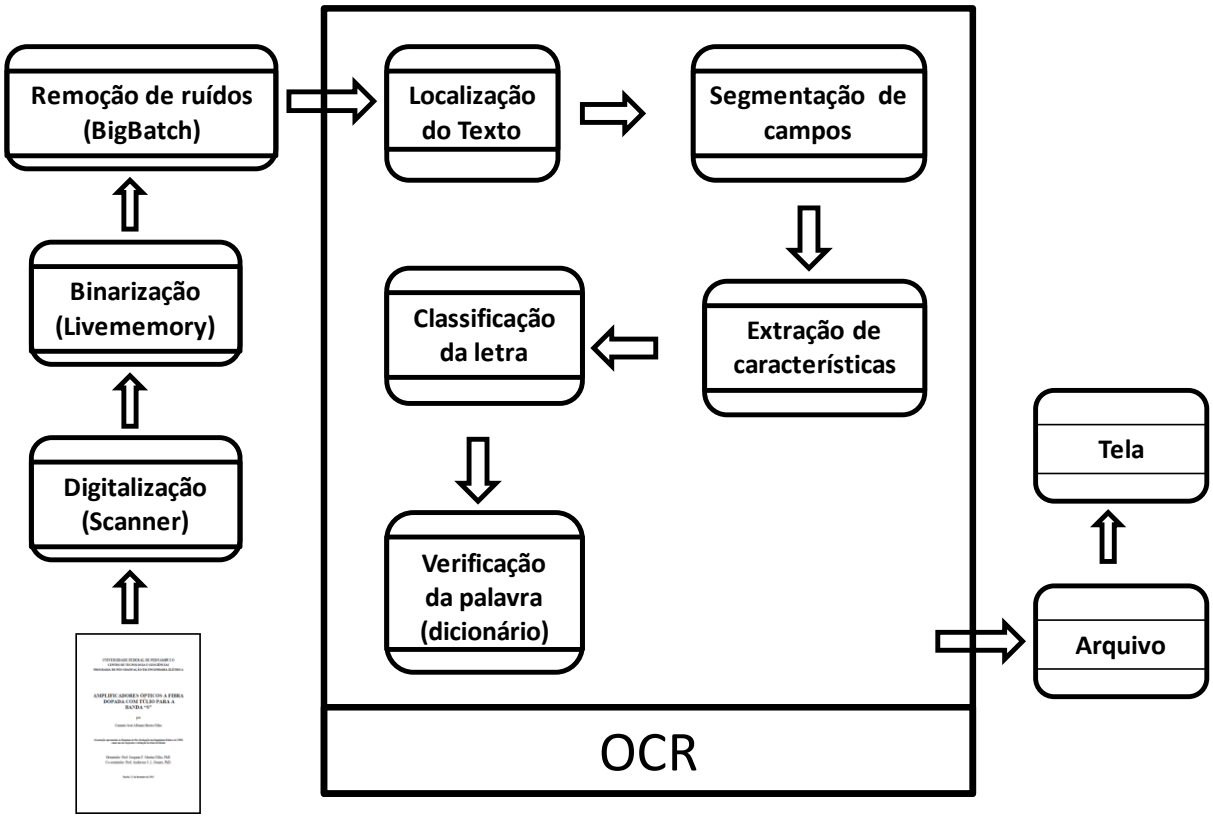


Figura 3.5: Diagrama de blocos da extração de texto por um OCR, detalhando os principais processo do OCR.

Na segunda, a aplicação da segmentação na área de texto pretende separar cada caractere da imagem, inicialmente detectando linhas e posteriormente separando os caracteres de cada palavra e linha. Isso se dá porque a comparação da imagem do texto é feita caractere a caractere, isoladamente, e não sobre um conjunto de caracteres (palavras). Uma forma de implementação da segmentação de caractere é a detecção de bordas, em que é aplicado um filtro laplaciano ou obtido o gradiente da imagem, e em seguida, procurado caminhos fechados.

Na etapa da segmentação alguns caracteres podem ser degradados [23], e podem introduzir erros no processo de extração de texto. Essas degradações podem ser por:

- Fragmentação de caractere: na aplicação dos algoritmos para detecção de bordas alguns caracteres, em especial os de baixa resolução, podem ser

fragmentados, produzindo dois caracteres. Na Figura 3.6 (a), a letra “m” após a aplicação de um filtro laplaciano foi fragmentada, gerando duas letras: “i” e “n”. Esse fenômeno ocorre para altos valores do limiar de binarização, sendo comum a compreensão da letra “E” como “F” + “.” / “_”.

- União de caracteres: de forma análoga, mas com o produto inverso, pode ocorrer a união de caracteres quando da aplicação de algoritmos de detecção de borda. Na Figura 3.6 (b), as letras “ffi” após a aplicação de um filtro laplaciano foram unidas, gerando uma letra: “m”. Esse fenômeno ocorre para baixos valores do limiar de binarização, sendo comum a compreensão da letra “i” como “l”, havendo união do ponto em cima do traço vertical dessa letra.
- Não compreensão de acentos e caracteres de pontuação: da mesma forma que acontece com a letra “i”, acentos podem ser compreendidos como um caractere pertencente a uma linha acima do que realmente está. Em alguns casos, a pontuação do texto é vista como ruído, em especial, o caractere “.”, é unido a outro caractere.

- Textos inseridos em imagens: como em artigos científicos que diagramam o documento em colunas, ou quando se coloca uma imagem na mesma linha que o texto, os dizeres e/ou textos das figuras também podem ser extraídos. Eles ficam na mesma linha que outros textos, inseridos na linha, ou inserem linhas extras num parágrafo, descontinuando o texto.

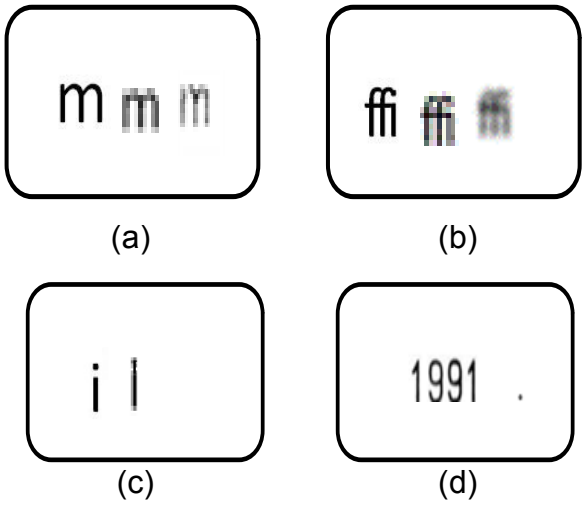


Figura 3.6: Ilustração de imagens degradadas pela manipulação e maus processos de digitalização.

A Figura 3.7 ilustra o fenômeno de união de caracteres que ocorreu num documento usado como teste da plataforma. Observa-se em (a) o documento original, enquanto em (b) a versão transcrita, e em (c), dentro dos quadros tracejados, o detalhe das letras “ri” que foram unidas e transcritas com a letra “n.”

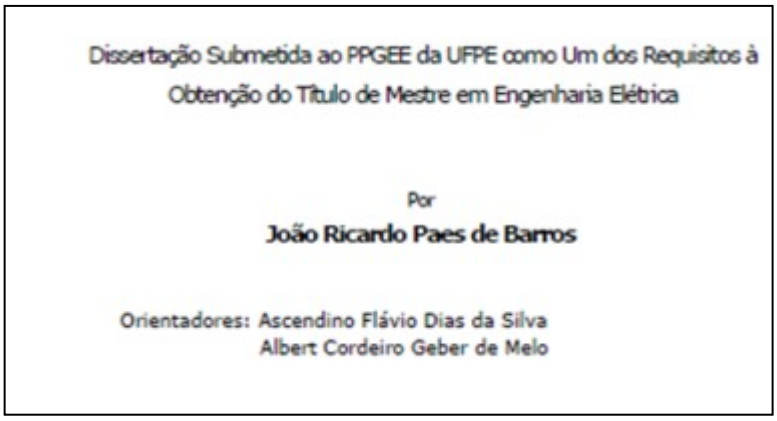
3.2.2. EXTRAÇÃO DE CARACTERÍSTICAS

Após a segmentação, é necessário saber com que caractere aquela imagem é parecida. Para isso, são extraídas características da imagem e comparadas com características de imagens já conhecidas para identificá-las.

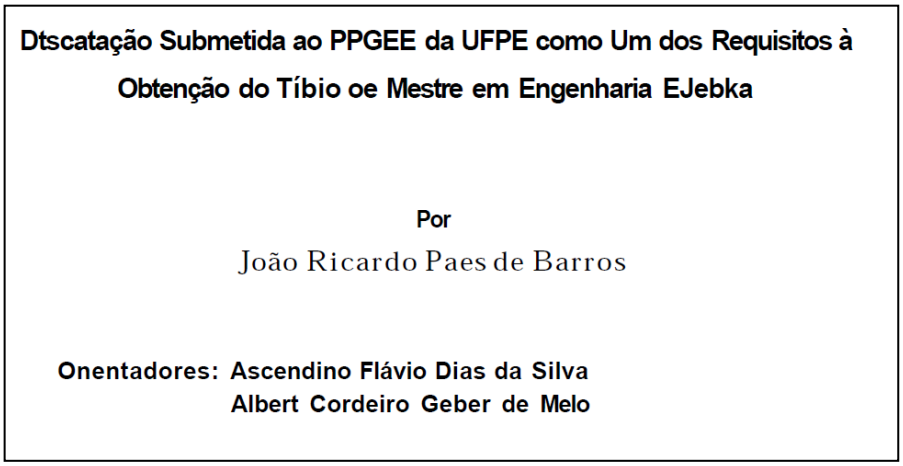
Essas características são comumente denominadas de descritores da imagem, e as mais empregadas com o propósito de reconhecimento de padrão são: estatísticas dos pixels; histogramas; posição de pixels específicos; coeficientes das transformadas bidimensionais; etc. Dentre esses descritores de imagens os momentos invariantes [24] [31] se destacam, pois resolvem o problema de escala e rotação do caractere, conforme se observa na Figura 3.8.

3.2.2.1.MOMENTOS INVARIANTES

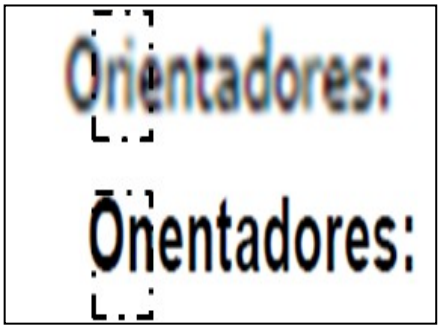
Conforme se observa na Figura 3.8, a distribuição de pixels das imagens na figura é diferente, porém se trata da mesma imagem inclinada. Algumas características independem da posição, tamanho e orientação da imagem, ou seja, das variações da mesma imagem. Momentos invariantes são características da imagem que independem de transformações espaciais, como escalonamento, rotação, translação, inclinação, etc. Características ligadas a cor e textura geralmente são invariantes a rotação e translação. Dentre os momentos invariantes, os baseados em distribuição estatística dos pontos como os momentos de Hu [31] e os descritores de Fourier [26] [27], conforme Gonzalez e Woods [12], apresentam-se como boas técnicas. Sinteticamente, essas técnicas se baseiam na forma, no contorno da imagem, em que um conjunto de funções matemáticas é aplicado e os coeficientes extraídos.



(a)



(b)



(c)

Figura 3.7: Ilustração de um documento que teve letras degradadas. (a) documento original; (b) versão transcrita; (c) detalhes do erro em que as letras “ri” da palavra “orientadores” foram substituídas por pela letra “n”.

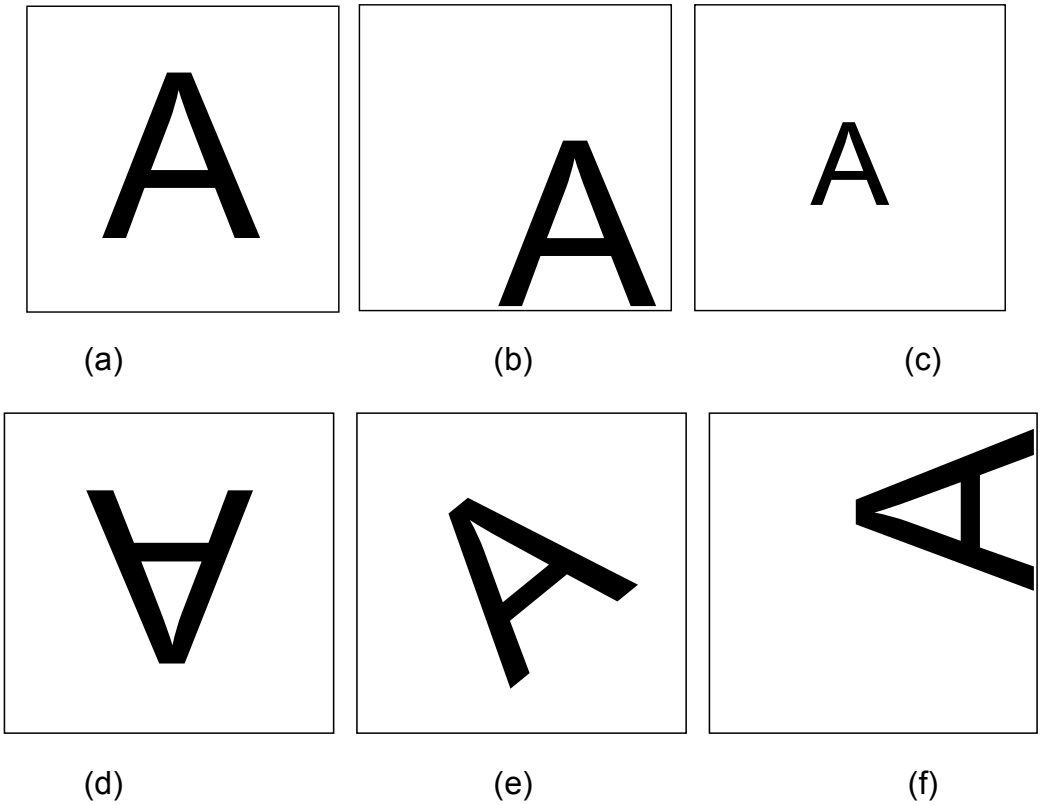


Figura 3.8: Imagens da letra “a” que sofreram transformações espaciais. (a) original; (b) deslocado; (c) escalonado, (d) invertido horizontalmente, (e) inclinado a 45º; (f) inclinado a 90º e deslocado.

3.2.3.CLASSIFICAÇÃO

Para a área de reconhecimento de padrões, um padrão pode ser definido como um arranjo de descritores da imagem, em que uma classe de padrões é um conjunto de padrões com algumas similaridades. Portanto, quando se fala em classificação propõem-se determinar quão similar uma classe de padrão de uma imagem é de uma classe de um conjunto de classes.

Para reconhecimento de caracteres cada símbolo do código ASCII (*American Standard Code for Information Interchange*) [21] existe uma classe de padrões, e cada imagem é comparada com as classes de referência.

Um padrão desconhecido é atribuído à classe mais similar de acordo com uma métrica predefinida. Por exemplo, assume-se que o descritor da imagem seja a ocorrência de um círculo e uma barra vertical em sua lateral, então as letras “p” e “q”

se encaixam nesse padrão, sendo necessários melhores descritores para definir a qual classe (caractere) pertence àquela imagem. É necessário definir a “lado” da barra, ou seja, melhorar o decisor.

3.2.3.1. MÉTODOS TEÓRICOS DE DECISÃO.

O classificador tipo *Matching* (casamento) por distância mínima é a abordagem mais utilizada para classificação, mas classificadores estatísticos e redes neurais também são muito utilizados.

No caso de uma classificação tipo *Matching*, cada classe usa um vetor descritor como referência, e a cada apresentação de uma imagem desconhecida seus descritores são obtidos, sendo calculada a distância euclidiana entre eles e os descritores de todas as classes. O método escolhe a menor distância para definir a qual classe pertence àquela imagem. O cálculo da distância mais comum é o da distância euclidiana, ou a euclidiana quadrática, mas pode-se de utilizar medidas de correlação.

Para classificadores estatísticos, a idéia é usar as probabilidades de um padrão pertencer a certa classe. Um erro (diferença entre a referência e a imagem) é associado a cada classe, e aquele que em média apresentar o menor valor de erros é definido como pertencente à classe.

As redes neurais artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural do cérebro humano [45]. O cérebro é um sistema biológico de processamento de informação complexo, não-linear e paralelo. Ele tem a capacidade de organizar suas componentes (células) básicas (neurônios), de forma a realizar os mais diversos tipos de processamento (controle motor, percepção, reconhecimento de padrões). Os neurônios estão conectados entre si por terminações sinápticas, e a capacidade de uma rede é determinada pelos pesos sinápticos, a forma como estão feitas as terminações sinápticas. Um neurônio é uma unidade de processamento de informação que é fundamental para a operação de uma rede neural. O diagrama da Figura 3.9 apresenta o modelo matemático do neurônio.

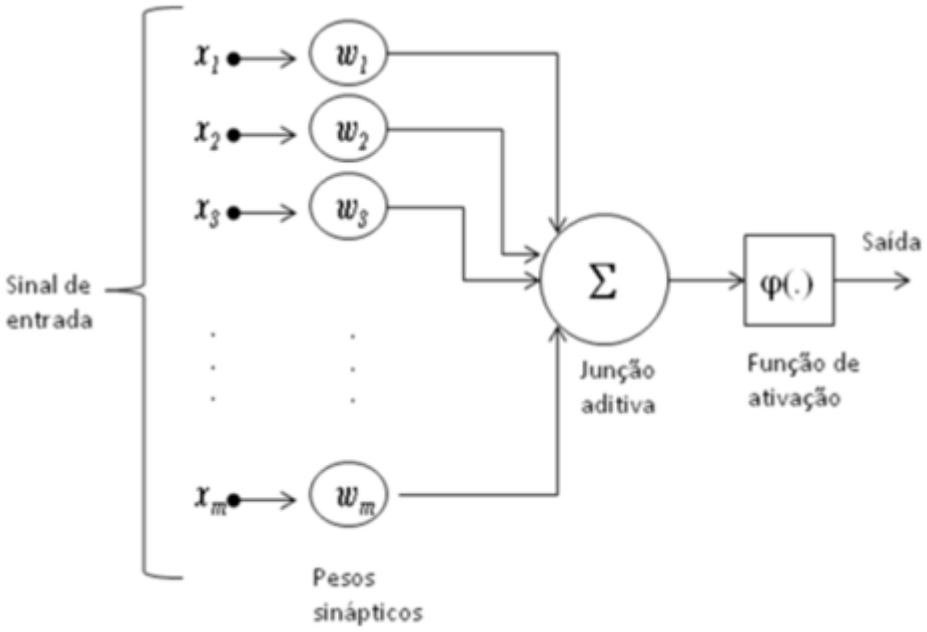


Figura 3.9: Modelagem matemática de um neurônio.

em que $x_i, i = 1,2,...,m$ são padrões de entradas já conhecidos, $w_i, i = 1,2,...,m$ são os neurônios cujos pesos sinápticos correspondem aos padrões de treinamento são ajustados, Σ é uma junção aditiva, e $\varphi(.)$ é uma função de ativação.

O poder computacional das redes neurais consiste em apresentar uma estrutura paralelamente distribuída como também na habilidade de aprender, e, portanto de generalizar. As generalizações se referem ao fato de a rede neural produzir saídas adequadas para entradas que não estavam presentes durante o treinamento (aprendizagem).

O aspecto mais importante das redes neurais é a habilidade de aprender e com isso melhorar seu desempenho. Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos, o treinamento. O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas. O reconhecimento de padrões constitui-se a área dominante da atuação das redes neurais.

A rede *back-propagation* é composta de uma série de camadas com neurônios interconectados, e essas conexões são modeladas por pesos sinápticos. Padrões são inseridos na rede até que a saída desejada seja obtida.

3.2.3.2. CONSIDERAÇÕES SOBRE A CLASSIFICAÇÃO

Os caracteres utilizados nas abordagens de OCR's são os definidos pelo sistema ASCII (*American Standard Character Information for Interchange*) [21]. Caracteres não inseridos nesse sistema não são reconhecidos por abordagens clássicas de OCR, mas alguns softwares mais recentes conseguem identificar esses símbolos. Quando não reconhecidos os caracteres são aproximados (classificados) para símbolos mais próximos e que estejam no sistema ASCII.

Esse fenômeno ocorre com expressões matemáticas, letras do alfabeto grego e de alfabetos orientais (cirílico, hindu, japonês, chinês, etc.). Para aplicações em que a transcrição é o resultado último, os erros não são tão significativos, porém em sistemas de reconhecimento e análise de conteúdo, os erros obtidos podem ser prejudiciais.

Outro problema característico de reconhecimento de caracteres trata do tipo de fonte utilizada no texto. Para cada tipo de fonte uma mesma letra apresenta formatos diferentes, a Figura 3.10 mostra a versões da letra "A" em normal e itálico, nas fontes "Times New Roman", "Arial", "Agency FB", "Blackadder ITC" e "Copperplate Gothic Light", e fica claro como o formato, a disposição dos traços e o contorno são diferentes para cada tipo de fonte.

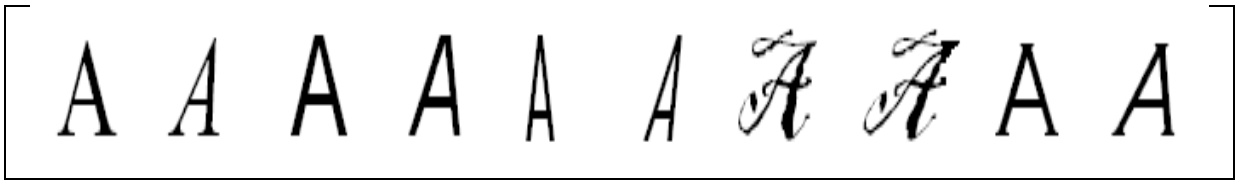


Figura 3.10: Versões da letra "A" em normal e itálico, nas fontes "Times New Roman", "Arial", "Agency FB", "Blackadder ITC" e "Copperplate Gothic Light", respectivamente.

Na tentativa de classificar letras de diferentes fontes, o resultado pode indicar a inexistência da letra ou cometer um erro de classificação (transcrição).

3.2.4. PÓS-PROCESSAMENTO

O resultado da extração de texto é um conjunto de caracteres, que individualmente não trazem informação, sendo necessário agrupá-los em palavras, números, sentenças, etc. Esse processo é denominado agrupamento em *Strings* (palavra em inglês que denota uma série de letras ou símbolos), sendo realizado através da localização do texto na página e do comprimento do “espaço em branco” entre caracteres sucessivos.

O comprimento do “espaço em branco” entre duas letras sucessivas de palavras distintas é maior do que o comprimento entre letras sucessivas de uma mesma palavra, sendo esse comprimento uma medida adaptativa. Essa abordagem apresenta um menor desempenho quando se trata de textos manuscritos e alguns datilografados em antigas máquinas de escrever, em que o espaço entre letras sucessivas de uma mesma palavra é maior do que em textos atualmente impressos.

Para solucionar os erros de agrupamento de *Strings*, pode-se utilizar abordagens adaptativas que devem ser treinadas previamente com o usuário, ou abordagens que utilizam a estatística da língua em que foi escrito o documento: letras proibidas de serem sucessivas (antes de “p” não se escreve “n”, então o caractere é “m” ou se trata de uma nova palavra); verificação da existência da palavra no dicionário da língua, mudando a palavra extraída, mas sem equivalente no dicionário para uma palavra mais similar.

Essas abordagens são para correção de erros, no entanto, sua precisão não é unitária. Se a palavra extraída apresenta um equivalente no dicionário, mas for diferente daquela que o autor escreveu será aceita, e o erro assumido como acerto.

3.2.5. OUTRAS CARACTERÍSTICAS DO TEXTO

De forma similar, pode ser determinada a fonte, cor, tamanho do caractere. São feitas comparações da imagem segmentada com imagens padrão, das características do caractere. Em processos de OCR mais modernos, quando não se encontra texto numa extensa área, essa área acaba sendo assumida como uma figura, uma imagem inserida junto ao texto, mas também alguns detalhes da imagem podem ser reconhecidos como texto, levando a uma transcrição incorreta do texto.

3.2.6. QUALIDADE DAS IMAGENS

A qualidade da imagem é fundamental para que o OCR transcreva corretamente os textos. Imagens de baixa resolução não “delineiam” bem a imagem da letra, enquanto que imagens de alta resolução tendem a “amplificar” o ruído existente na imagem. A Figura 3.11 apresenta um trecho da imagem de um documento e o texto extraído, o qual apresentou 10 erros na transcrição (detalhes em vermelho) uma união de duas palavras (detalhe em azul).

A Figura 3.12 mostra a transcrição em (a) e em (b) imagem original de um documento que teve suas folhas encadernadas com espiral, revelando erros na transcrição do texto, quando se considerou os “furos da espiral” como letras a serem transcritas. Por exemplo, os furos foram considerados como letras “O”, “C)”, “(I”, “U”, “ü”, “J” e outros caracteres, o símbolo em destaque indica que o furo da espiral foi transcrito como “(“I”.

A análise da Figura 3.12 ainda indica uma alta sensibilidade do OCR em relação a ruído na imagem, e que a remoção de ruído (“furos da espiral”) traria um ganho na transcrição. Por tanto, antes da transcrição do texto uma série de tratamentos na imagem se faz necessário para melhorar a precisão na transcrição.

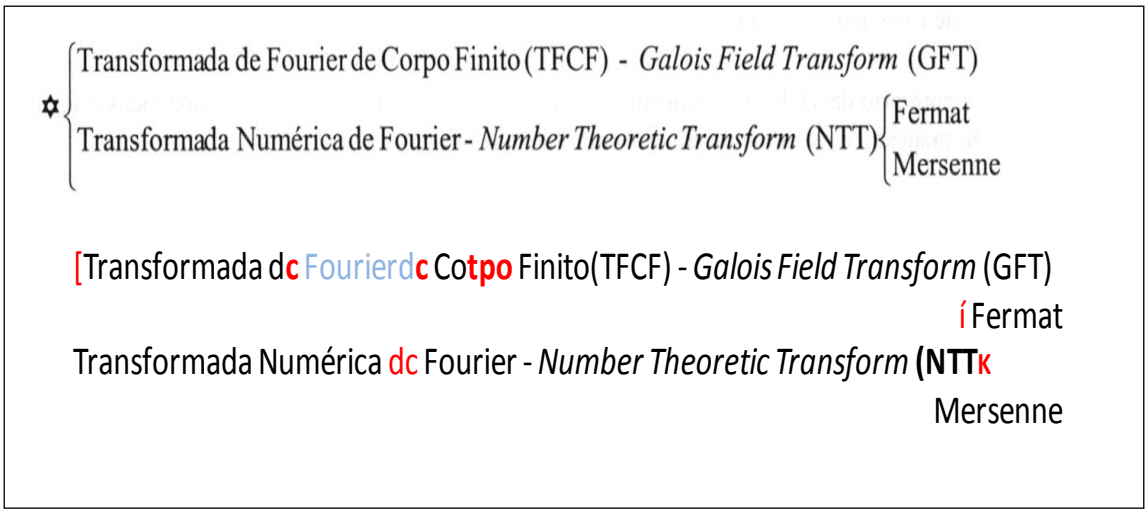


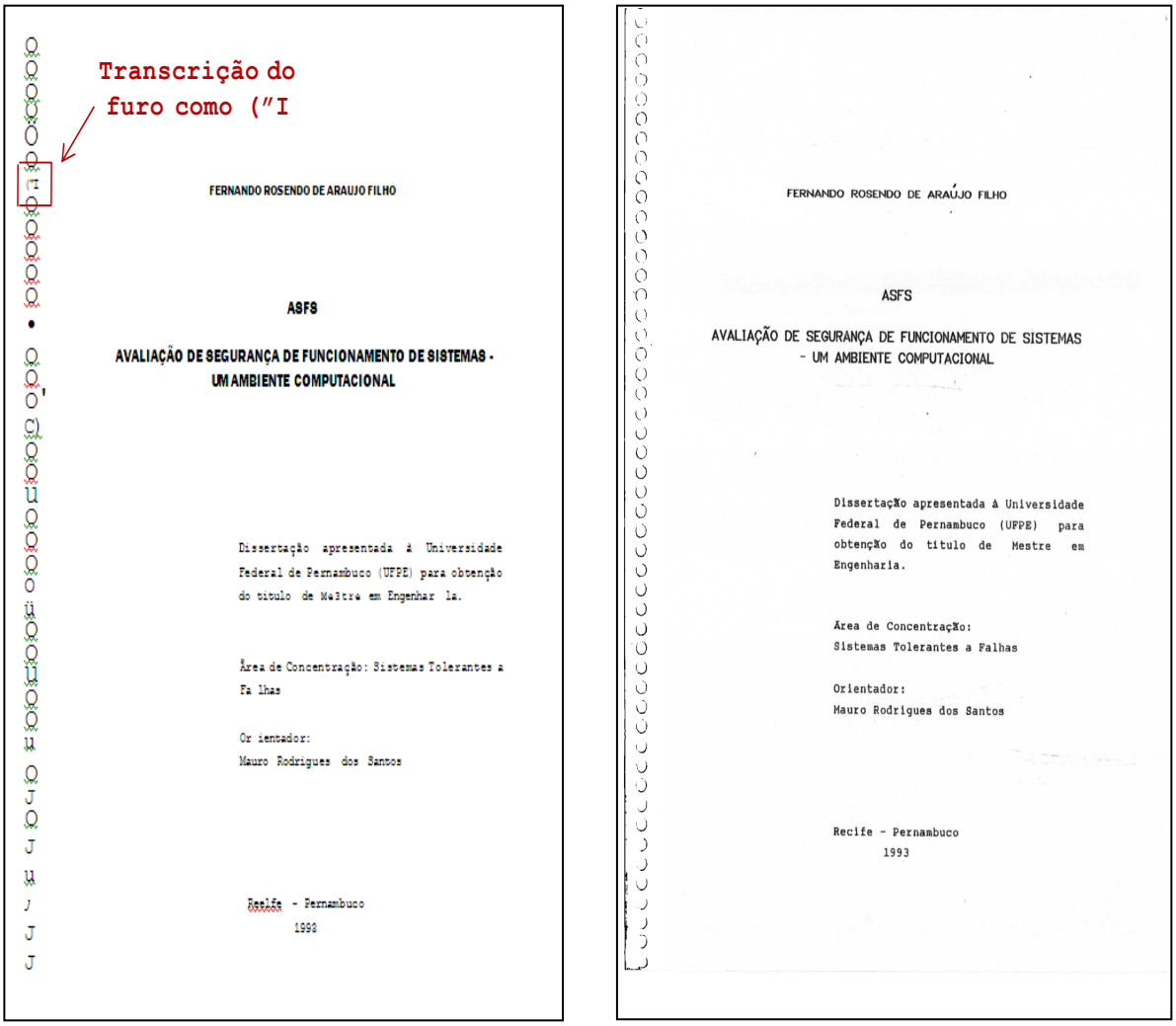
Figura 3.11: Erros de transcrição de texto de um trecho da imagem de um documento pelo OCR.

3.3. PRINCIPAIS TIPOS DE RUÍDO DAS IMAGENS

Em documentos impressos a ação do tempo, a exposição ao ar, o manuseio são danosos para conservação do papel e do conteúdo do documento, ao passo que há destruição das fibras do papel, as páginas ficam “coladas”, há um borramento de tinta de impressão, de sorte que o texto fica ilegível. Muitas vezes, a forma de arquivamento do documento físico insere características diferentes daquelas produzidas pelo autor, tais como espirais e furos guia, colagem de páginas, etc. No próprio processo de digitalização podem advir algumas falhas, inserindo características diferentes à imagem. Todas essas características inseridas são diferentes tipos ruído da imagem, que influenciam negativamente a transcrição do texto pelo OCR, devendo ser removidos.

Os furos de espiral ou de guia são considerados ruído de borda. No processo de digitalização, linhas pretas nas bordas do documento surgem quando não se alinha o documento com as bordas da área de digitalização do equipamento (*scanner*), ou quando todas as bordas do documento não entram em contato com a superfície do equipamento. Essas linhas pretas também são consideradas ruído de borda. Tanto os furos como as linhas na bordas são resultados da grande diferença de iluminação dessa área em relação à iluminação do restante do documento. A

Figura 3.13 ilustra uma imagem digitalizada de um documento que fora arquivado em espiral, e os furos da espiral se configuram como ruído de borda.



(a) (b)

Figura 3.12: (a) Transcrição do documento pelo OCR e
(b) imagem de um documento com ruído de borda.

Quando as bordas do documento digitalizado não estão alinhadas com as bordas do equipamento de digitalização, a imagem é adquirida com uma rotação ou inclinação. Na rotação, as bordas e como conseqüência as linhas do texto não estão alinhadas na imagem, enquanto que na inclinação há uma inserção de perspectiva, isto é, uma determinada parte da imagem parece estar maior que as outras. A Figura 3.13 ilustra um ruído de borda.

Alguns pontos na imagem aparecem por causa de sujeiras no papel, pelo próprio desvanecimento do papel, ou por falhas no processo de aquisição. Pequenos pontos isolados na área da imagem que “ocupam” poucos pixels não definem um caractere ou um elemento gráfico pertencente ao documento, e geralmente são tratados como ruído sal-e-pimenta. Esses pontos, apesar de não fazerem parte do texto do documento, podem ser transcritos: considerados como algum caractere num tamanho de fonte pequeno, ou se estiver próximo a um caractere pode “deformar” sua imagem, inserindo erro na transcrição do OCR. A Figura 3.14 apresenta a imagem de um documento com “sujeiras”, pontos que não fazem parte do texto.

O fundo do papel em cores menos claras, ou até mesmo letras com tons de cinza diferentes podem inserir erros no processo de transcrição. A binarização da imagem pretende colocar todos os pixels da imagem em duas cores possíveis, o fundo do documento em branco e o texto em preto. É interessante também porque o armazenamento, o processamento e a velocidade de tratamento são otimizados.

3.4. TÉCNICAS PARA REMOÇÃO DE RUÍDO

O ruído provoca grandes erros na extração dos textos das imagens, e sua remoção é fundamental para melhorar esse processo. A primeira etapa consiste na binarização da imagem, e logo após as imagens são submetidas a técnicas de remoção de ruído implementadas na plataforma BigBatch [7], que trata as imagens monocromáticas em lotes, ou seja, todas as páginas de uma documento são tratadas de uma só vez. As imagens tratadas pelas técnicas do BigBatch são submetidas ao OCR para extração do texto.

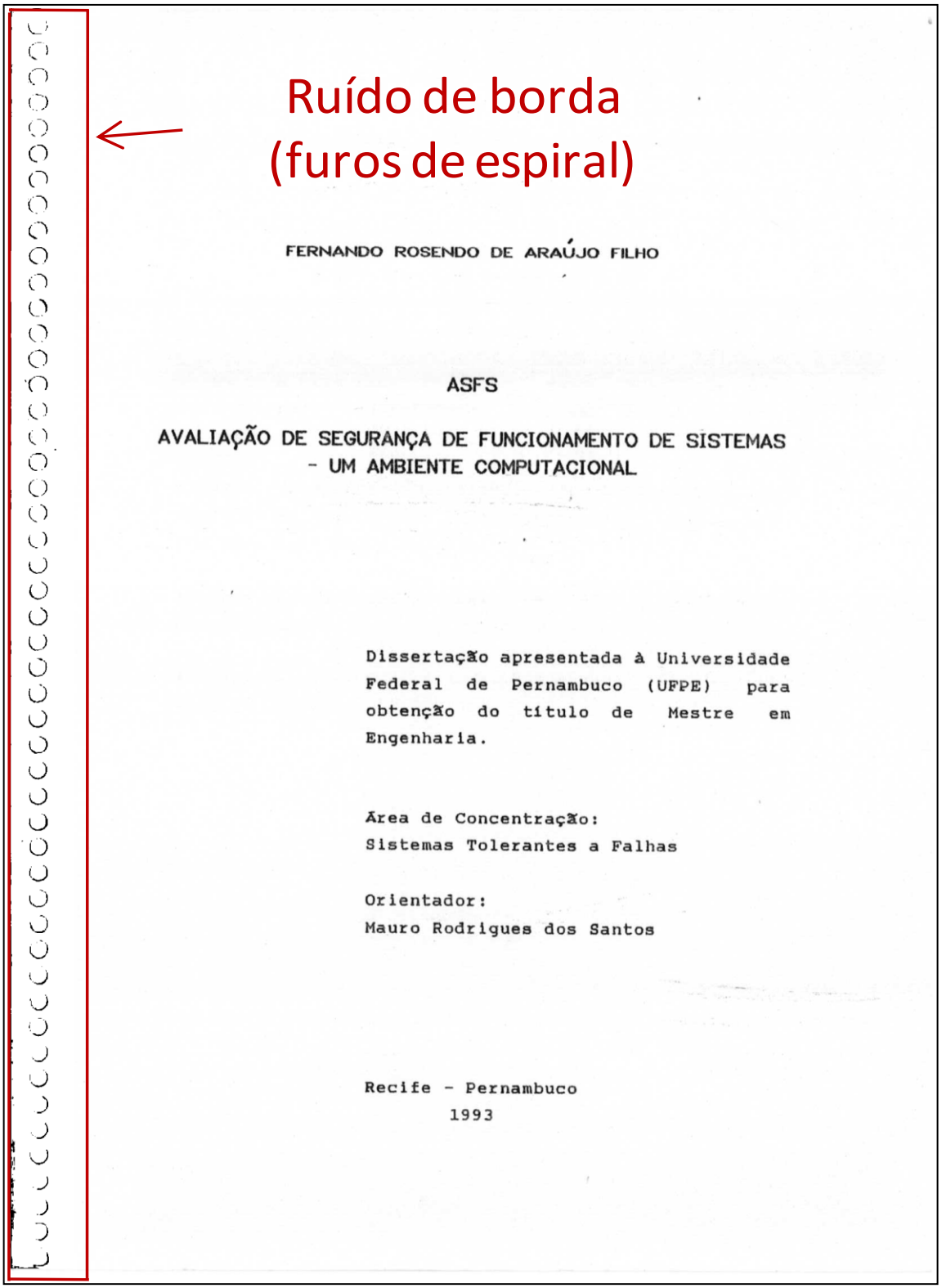
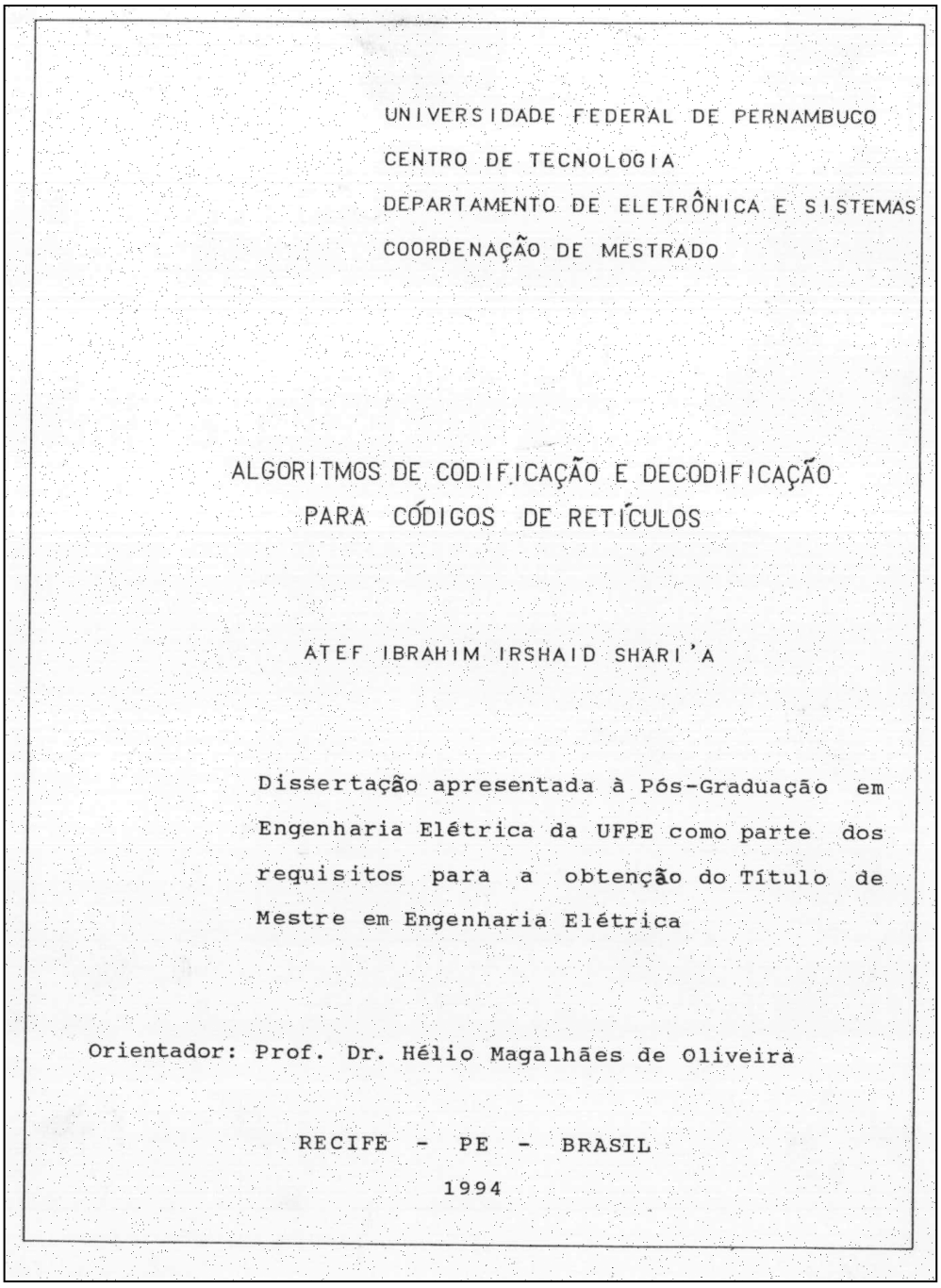


Figura 3.13: Ilustração da imagem de um documento com ruído de borda.



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ELETRÔNICA E SISTEMAS
COORDENAÇÃO DE MESTRADO

ALGORITMOS DE CODIFICAÇÃO E DECODIFICAÇÃO
PARA CÓDIGOS DE RETÍCULOS

ATEF IBRAHIM IRSHAID SHARI'A

Dissertação apresentada à Pós-Graduação em Engenharia Elétrica da UFPE como parte dos requisitos para a obtenção do Título de Mestre em Engenharia Elétrica

Orientador: Prof. Dr. Hélio Magalhães de Oliveira

RECIFE - PE - BRASIL
1994

Figura 3.14: Ilustração da imagem de um documento com ruído sal-e-pimenta.

A Figura 3.15 ilustra com um diagrama de blocos o processo de extração de ruído das imagens trabalhadas. Logo após serem digitalizadas, as imagens são binarizadas de acordo com o módulo de binarização do LiveMemory, que identifica e preserva os pixels de figuras na imagem do documento e binariza o texto e o fundo da página.

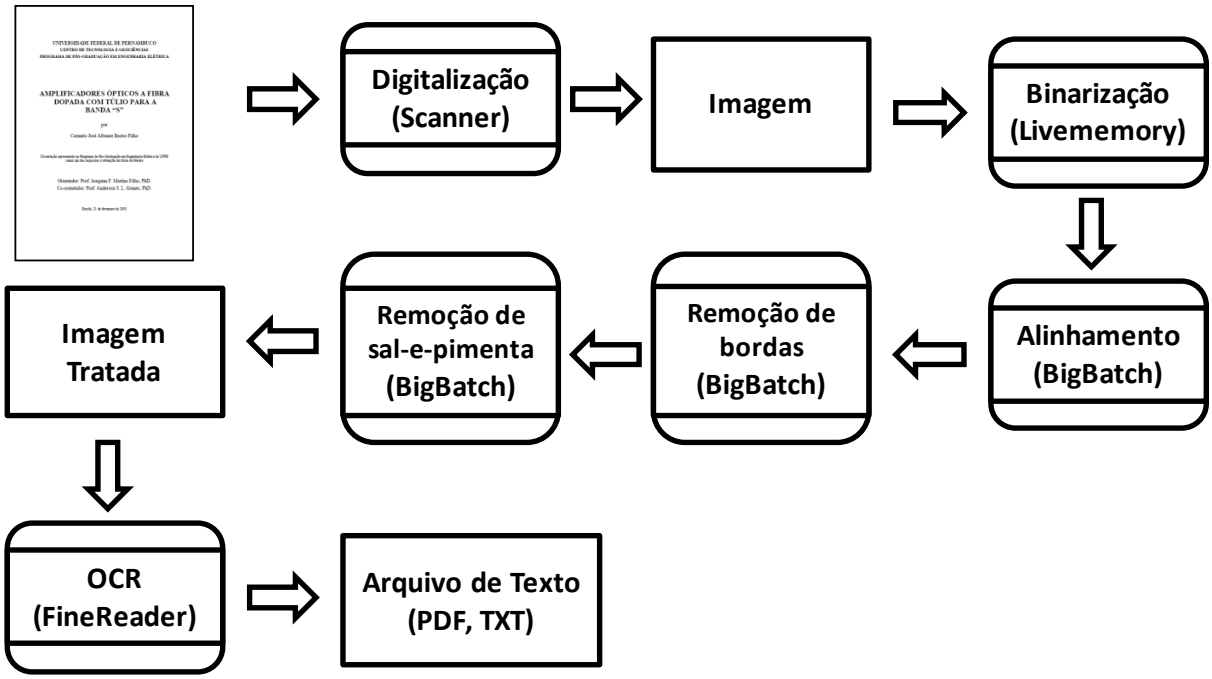


Figura 3.15: Fluxograma do processo de tratamento de ruído.

3.4.1. BINARIZAÇÃO

O objetivo principal da binarização aplicada ao processo de OCR é separar o texto do fundo do papel, ou seja, segmentar as imagens dos caracteres [37]. A binarização é uma quantização com dois níveis possíveis. Dessa forma, todos os pixels têm seus valores reais aproximados para os dois níveis (preto e branco). As vantagens da quantização como menor espaço para armazenamento, maior velocidade para transmissão e processamento são encontradas na binarização de forma otimizada.

A técnica de quantização com dois níveis é também conhecida como limiarização, pois sendo definido um limiar todos os pixels que apresentarem valores abaixo do limiar são aproximados para o branco, e os que apresentam valores acima do limiar são aproximados para preto. Existem várias técnicas para limiarização, e podem ser classificadas em: globais como o algoritmo de Otsu [30], o de Sauvola [36] e o da_Silva_Lins_Rocha [16], em que um único limiar é definido para toda a imagem; e locais como o algoritmo de Niblack [29], o de Bersen [33], e o

Oliveira_Lins [28], em que um limiar é definido por regiões da imagem. Diversos trabalhos foram propostos para avaliação do desempenho das técnicas [35] [38] de binarização de uma forma geral, e com o propósito de segmentação de texto [34].

Além do texto, os documentos trazem elementos gráficos como figuras, logomarcas, fotos e gráficos. Esses elementos “necessitam” ter um número maior de tons de cinza para serem representados do que os fornecidos pela binarização, caso contrário poderia acarretar a perda da informação desses elementos gráficos. É interessante, portanto, antes de binarizar todo o documento que se identifique o que é texto e o que é elemento gráfico, para preservar os elementos gráficos e só então binarizar os textos.

O módulo de binarização da plataforma LiveMemory é utilizado com esse propósito, sendo capaz de detectar os elementos gráficos, e só após binarizar os textos. A técnica de binarização é feita por um algoritmo baseado em entropia projetado para remoção da interferência frente-verso em documentos históricos [16]. A idéia da técnica é considerar a distribuição do histograma da imagem com 256 níveis de cinza como uma distribuição a priori dos pixels da fonte, e a partir da dessa distribuição calcular a entropia da fonte, que determina o limiar para binarização global.

A Figura 3.16 mostra a versão original da imagem de um documento digitalizado que foi binarizado com diferentes limiares:

- Figura 3.17 imagem binarizada com limiar 64;
- Figura 3.18 imagem binarizada com limiar 128;
- Figura 3.19 imagem binarizada com limiar 192;
- Figura 3.20 imagem binarizada com limiar 220.

Para essa imagem, percebe-se que um baixo valor do limiar auxilia na transcrição do texto, ao passo que o texto contido na tabela é preservado nos casos Figura 3.17 e Figura 3.18, enquanto que é perdido nos casos Figura 3.19 e Figura 3.20 com limiares altos 192 e 220, respectivamente.

transformados sobre corpos finitos em geral são completamente diferentes dos pares transformados discretos sobre corpos infinitos .

	Característica Zero	Característica p
Transformadas Discretas	**Transformada Discreta de Fourier**	**Transformada de Fourier em um Corpo Finito✡**
	Transformada Discreta do Cosseno	_____
	Transformada Discreta do Seno	_____
	Transformada Discreta de Hartley	**Transformada de Hartley em um Corpo Finito**
	Outras	**Outras**

✡ { Transformada de Fourier de Corpo Finito (TFCF) - *Galois Field Transform* (GFT)
Transformada Numérica de Fourier - *Number Theoretic Transform* (NTT) { Fermat / Mersenne }

3.2.2 A Transformada Discreta de Fourier

Sendo de fundamental importância no Processamento de Sinais, bem como na análise de Sistemas de Comunicações, a Transformada de Fourier é uma ferramenta constantemente utilizada em Engenharia Elétrica. Suas aplicações se destacam, por exemplo, no cálculo eficiente de convoluções, na obtenção da resposta em freqüência de sistemas, na análise espectral de sinais, entre diversas outras.

A Transformada Discreta de Fourier tem a seguinte lei de transformação

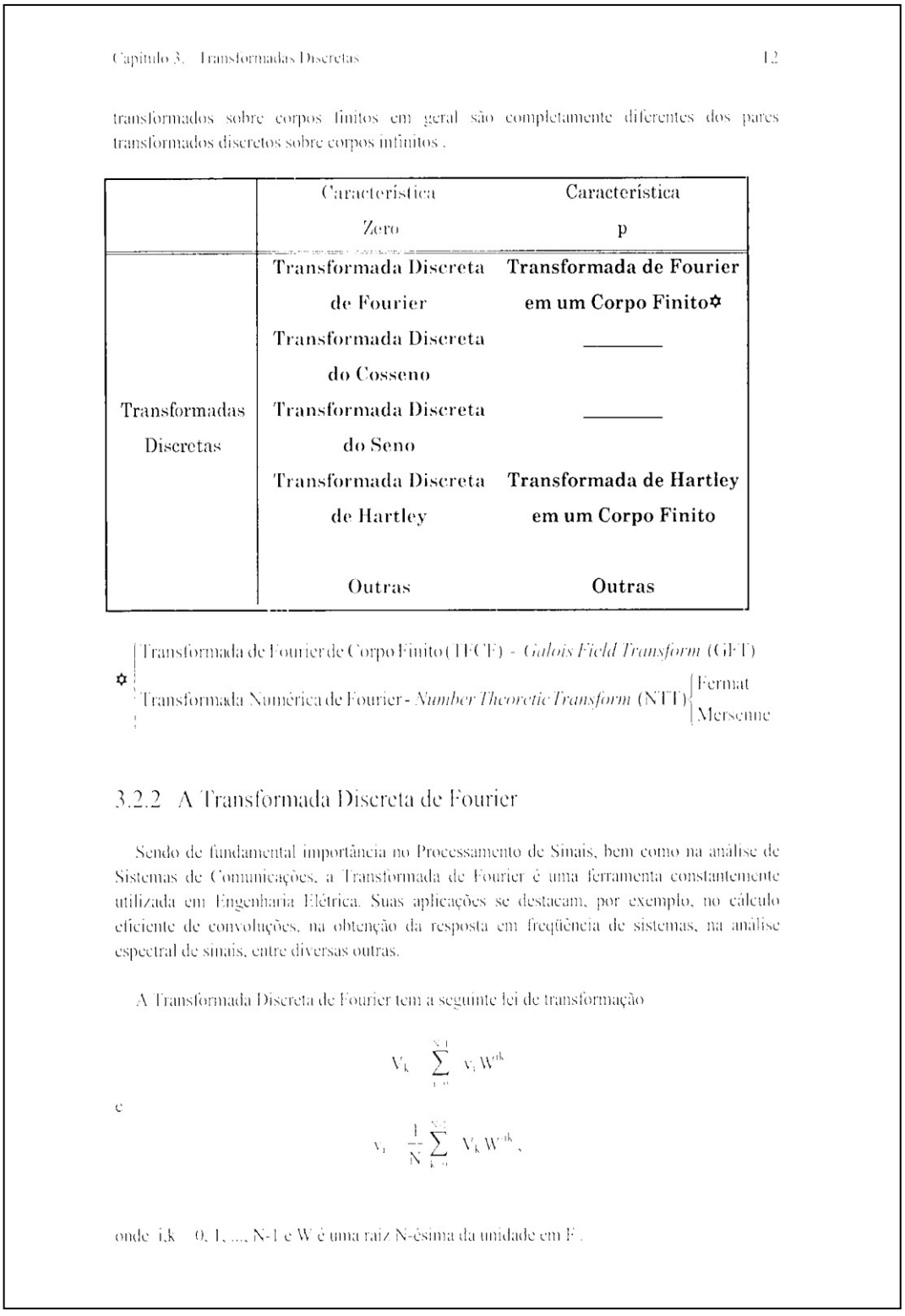
$$V_k = \sum_{i=0}^{N-1} v_i W^{ik}$$

e

$$v_i = \frac{1}{N} \sum_{k=0}^{N-1} V_k W^{-ik},$$

onde i,k = 0, 1, ..., N-1 e W é uma raiz N-ésima da unidade em F .

Figura 3.16: Imagem original de uma página de uma dissertação.



Capítulo 3. Transformadas Discretas 12

transformados sobre corpos finitos em geral são completamente diferentes dos pares transformados discretos sobre corpos infinitos .

	Característica Zero	Característica p
Transformadas Discretas	**Transformada Discreta de Fourier**	**Transformada de Fourier em um Corpo Finito✡**
	Transformada Discreta do Cosseno	_____
	Transformada Discreta do Seno	_____
	Transformada Discreta de Hartley	**Transformada de Hartley em um Corpo Finito**
	Outras	**Outras**

✡ { Transformada de Fourier de Corpo Finito (FFCF) - *Galois Field Transform* (GFT)
Transformada Numérica de Fourier - *Number Theoretic Transform* (NTT) { Fermat / Mersenne

3.2.2 A Transformada Discreta de Fourier

Sendo de fundamental importância no Processamento de Sinais, bem como na análise de Sistemas de Comunicações, a Transformada de Fourier é uma ferramenta constantemente utilizada em Engenharia Elétrica. Suas aplicações se destacam, por exemplo, no cálculo eficiente de convoluções, na obtenção da resposta em freqüência de sistemas, na análise espectral de sinais, entre diversas outras.

A Transformada Discreta de Fourier tem a seguinte lei de transformação

$$V_k = \sum_{i=0}^{N-1} v_i W^{ik}$$

e

$$v_i = \frac{1}{N} \sum_{k=0}^{N-1} V_k W^{-ik},$$

onde i,k = 0, 1, ..., N-1 e W é uma raiz N-ésima da unidade em F .

Figura 3.17: Versão binarizada da Figura 3.16 com limiar de 64, utilizando o algoritmo de binarização de documento históricos implementado na plataforma BigBatch [16].

transformados sobre corpos finitos em geral são completamente diferentes dos pares transformados discretos sobre corpos infinitos .

	Característica Zero	Característica p
Transformadas Discretas	**Transformada Discreta de Fourier**	**Transformada de Fourier em um Corpo Finito✡**
	Transformada Discreta do Cosseno	———
	Transformada Discreta do Seno	———
	Transformada Discreta de Hartley	**Transformada de Hartley em um Corpo Finito**
	Outras	**Outras**

✡ { Transformada de Fourier de Corpo Finito (TFCF) - *Galois Field Transform* (GFT)
Transformada Numérica de Fourier - *Number Theoretic Transform* (NTT) { Fermat
Mersenne

3.2.2 A Transformada Discreta de Fourier

Sendo de fundamental importância no Processamento de Sinais, bem como na análise de Sistemas de Comunicações, a Transformada de Fourier é uma ferramenta constantemente utilizada em Engenharia Elétrica. Suas aplicações se destacam, por exemplo, no cálculo eficiente de convoluções, na obtenção da resposta em freqüência de sistemas, na análise espectral de sinais, entre diversas outras.

A Transformada Discreta de Fourier tem a seguinte lei de transformação

$$V_k = \sum_{i=0}^{N-1} v_i W^{ik}$$

e

$$v_i = \frac{1}{N} \sum_{k=0}^{N-1} V_k W^{-ik},$$

onde i,k = 0, 1, ..., N-1 e W é uma raiz N-ésima da unidade em F .

Figura 3.18: Versão binarizada da Figura 3.16 com limiar de 128, utilizando o algoritmo de binarização de documento históricos implementado na plataforma BigBatch [16].

transformados sobre corpos finitos em geral são completamente diferentes dos pares transformados discretos sobre corpos infinitos .

Característica Zero
Transformada Discreta de Fourier
Transformada Discreta do Cosseno
Transformada Discreta do Seno
Transformada Discreta de Hartley
Outras

- Transformada de Fourier de Corpo Finito (TFCF) - *Galois Field Transform* (GFT)
- Transformada Numérica de Fourier - *Number Theoretic Transform* (NTT)
  - Fermat
  - Mersenne

3.2.2 A Transformada Discreta de Fourier

Sendo de fundamental importância no Processamento de Sinais, bem como na análise de Sistemas de Comunicações, a Transformada de Fourier é uma ferramenta constantemente utilizada em Engenharia Elétrica. Suas aplicações se destacam, por exemplo, no cálculo eficiente de convoluções, na obtenção da resposta em freqüência de sistemas, na análise espectral de sinais, entre diversas outras.

A Transformada Discreta de Fourier tem a seguinte lei de transformação

$$V_k = \sum_{i=0}^{N-1} v_i W^{ik}$$

e

$$v_i = \frac{1}{N} \sum_{k=0}^{N-1} V_k W^{-ik},$$

onde i,k = 0, 1, ..., N-1 e W é uma raiz N-ésima da unidade em F .

Figura 3.19: Versão binarizada da Figura 3.16 com limiar de 192, utilizando o algoritmo de binarização de documento históricos implementado na plataforma BigBatch [16].

transformados sobre corpos finitos em geral são completamente diferentes dos pares transformados discretos sobre corpos infinitos .

Característica Zero
Transformada Discreta de Fourier
Transformada Discreta do Cosseno
Transformada Discreta do Seno
Transformada Discreta de Hartley
Outras

- Transformada de Fourier de Corpo Finito (TFCF) - *Galois Field Transform* (GFT)
- Transformada Numérica de Fourier - *Number Theoretic Transform* (NTT)
  - Fermat
  - Mersenne

3.2.2 A Transformada Discreta de Fourier

Sendo de fundamental importância no Processamento de Sinais, bem como na análise de Sistemas de Comunicações, a Transformada de Fourier é uma ferramenta constantemente utilizada em Engenharia Elétrica. Suas aplicações se destacam, por exemplo, no cálculo eficiente de convoluções, na obtenção da resposta em freqüência de sistemas, na análise espectral de sinais, entre diversas outras.

A Transformada Discreta de Fourier tem a seguinte lei de transformação

$$V_k = \sum_{i=0}^{N-1} v_i W^{ik}$$

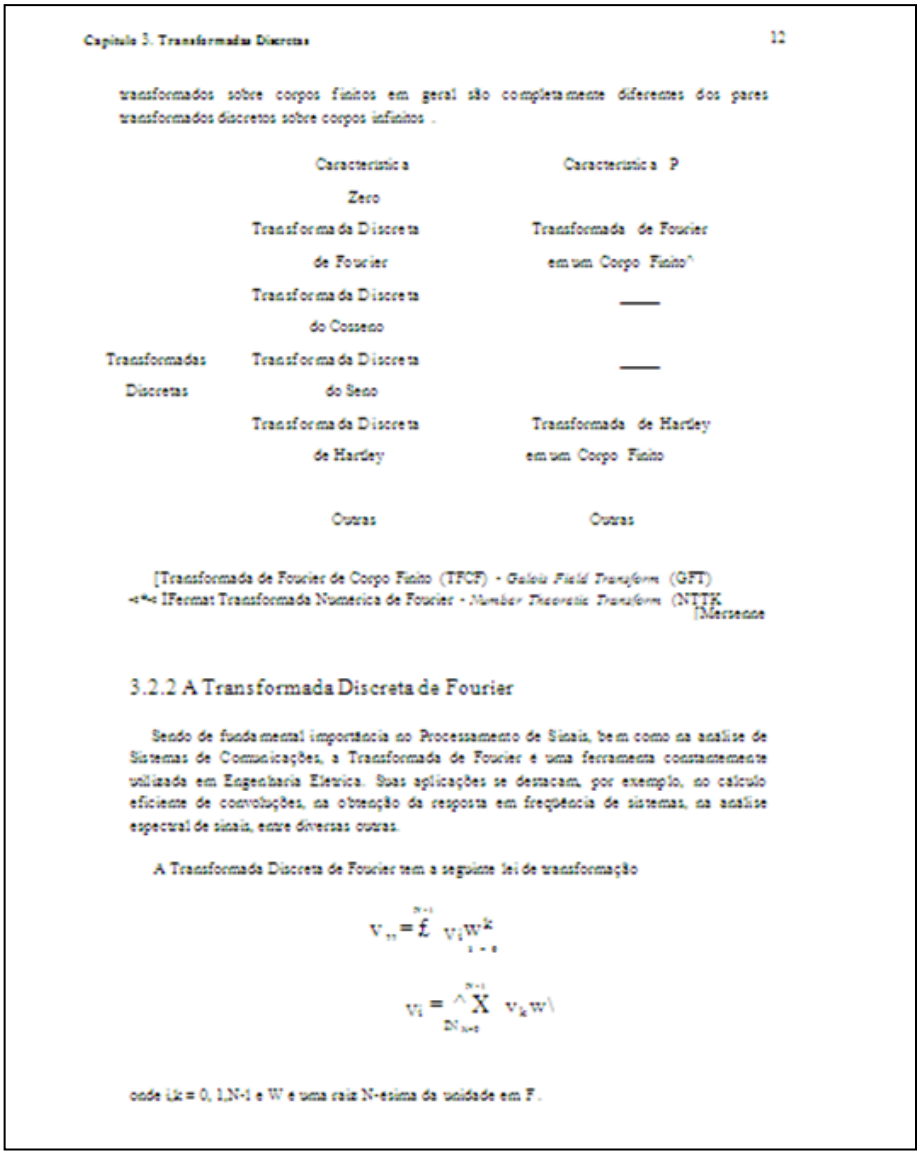
e

$$v_i = \frac{1}{N} \sum_{k=0}^{N-1} V_k W^{-ik},$$

onde i,k = 0, 1, ..., N-1 e W é uma raiz N-ésima da unidade em F .

Figura 3.20: Versão binarizada da Figura 3.16 com limiar de 220, utilizando o algoritmo de binarização de documento históricos implementado na plataforma BigBatch [16].

A Figura 3.21 e a Figura 3.22 apresentam duas versões da transcrição imagem original da Figura 3.16, de modo que na primeira foi feita a transcrição sem efetuar previamente uma binarização, e na segunda a transcrição após a binarização. Nota-se que a informação do texto contido numa tabela foi perdida na transcrição sem binarização prévia. É importante frisar, que o próprio software [17] executa uma binarização antes de transcrever o texto, a menos que a imagem já esteja binarizada.



Capítulo 3. Transformadas Discretas 12

transformados sobre corpos finitos em geral são completamente diferentes dos pares transformados discretos sobre corpos infinitos .

	Característica Zero	Característica P
Transformadas Discretas	Transformada Discreta de Fourier	Transformada de Fourier em um Corpo Finito\
	Transformada Discreta do Cosseno	——
	Transformada Discreta do Seno	——
	Transformada Discreta de Hartley	Transformada de Hartley em um Corpo Finito
	Outras	Outras

[Transformada de Fourier de Corpo Finito (TFCF) - *Galois Field Transform* (GFT)
<*<= IFermat Transformada Numerica de Fourier - *Number Theoretic Transform* (NTT)[Mersenne

3.2.2 A Transformada Discreta de Fourier

Sendo de fundamental importância no Processamento de Sinais, bem como na análise de Sistemas de Comunicações, a Transformada de Fourier é uma ferramenta constantemente utilizada em Engenharia Elétrica. Suas aplicações se destacam, por exemplo, no cálculo eficiente de convoluções, na obtenção da resposta em freqüência de sistemas, na análise espectral de sinais, entre diversas outras.

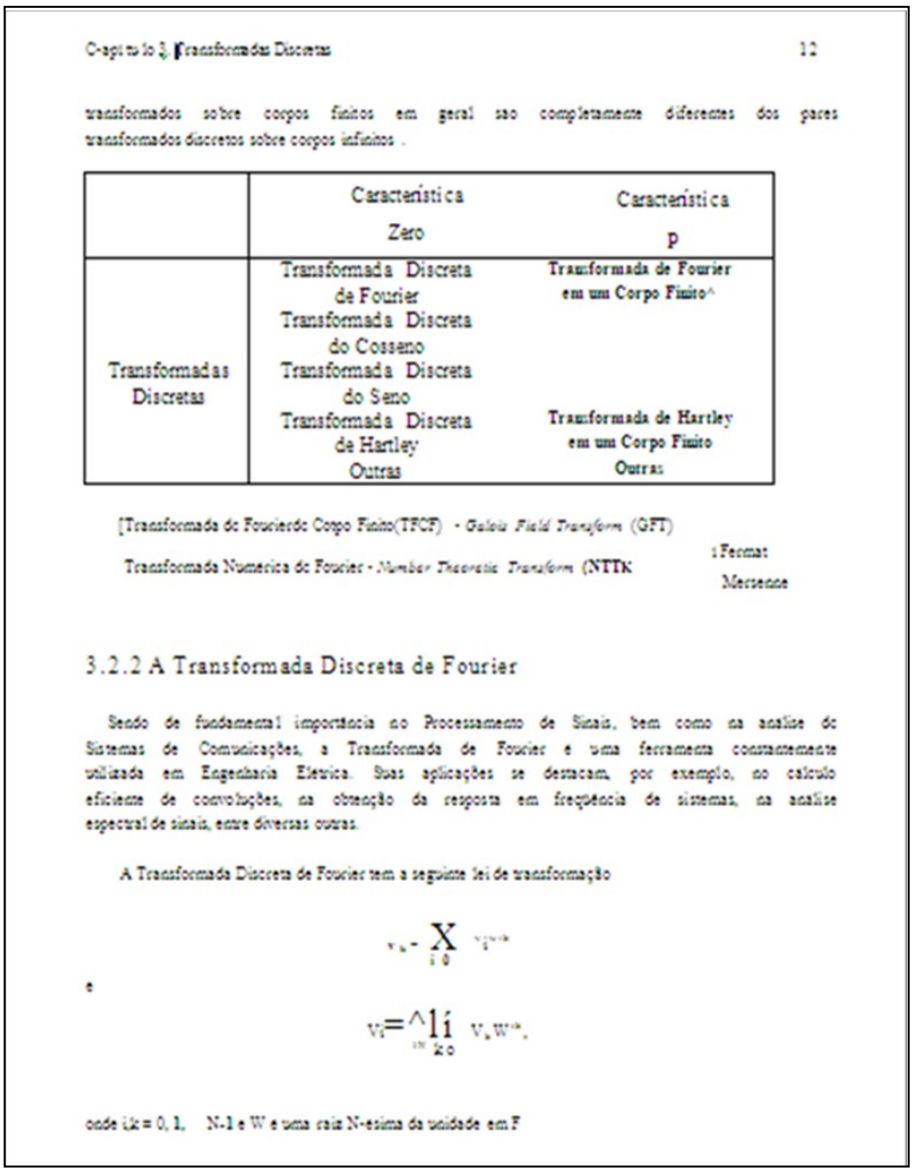
A Transformada Discreta de Fourier tem a seguinte lei de transformação

$$v_{..} = \hat{f} \; v_i W^{k}$$

$$v_i = \hat{X} \; v_k W\backslash$$

onde i,k = 0, 1,N-1 e W é uma raiz N-ésima da unidade em F .

Figura 3.21: Transcrição da imagem da Figura 3.16 sem efetuar previamente uma binarização.



C-apitulo 3. Transformadas Discretas 12

transformados sobre corpos finitos em geral são completamente diferentes dos pares transformados discretos sobre corpos infinitos .

	Característica Zero	Característica p
Transformadas Discretas	Transformada Discreta de Fourier Transformada Discreta do Cosseno Transformada Discreta do Seno Transformada Discreta de Hartley Outras	Tranformada de Fourier em um Corpo Finito^ Tranformada de Hartley em um Corpo Finito Outras

[Transformada de Fourierdo Corpo Finito(TFCF) - *Galois Field Transform* (GFT)

Transformada Numerica de Fourier - *Number Theoretic Transform* (NTTK ↓Fermat Mersenne

3.2.2 A Transformada Discreta de Fourier

Sendo de fundamental importância no Processamento de Sinais, bem como na analise de Sistemas de Comunicações, a Transformada de Fourier é uma ferramenta constantemente utilizada em Engenharia Eletrica. Suas aplicações se destacam, por exemplo, no calculo eficiente de convoluções, na obtenção da resposta em freqüência de sistemas, na analise espectral de sinais, entre diversas outras.

A Transformada Discreta de Fourier tem a seguinte lei de transformação

$$v_k = \sum_{j=0}^{N-1} v_j w^{jk}$$

e

$$v_j = \frac{1}{N} \sum_{k=0}^{N-1} v_k w^{-jk}$$

onde j,k = 0, 1, ... N-1 e W é uma raiz N-esima da unidade em F

Figura 3.22: Transcrição da imagem da Figura 3.16 após a binarização.

3.4.2. BORDAS

O ruído de borda pode ser visto como detalhes da imagem nas bordas do documento que não foram inseridas pelo autor e, portanto não fazem parte do documento. Esse ruído, freqüentemente chamado de ruído de margem [18] ocorre devido a forma de arquivamento da cópia física do documento.

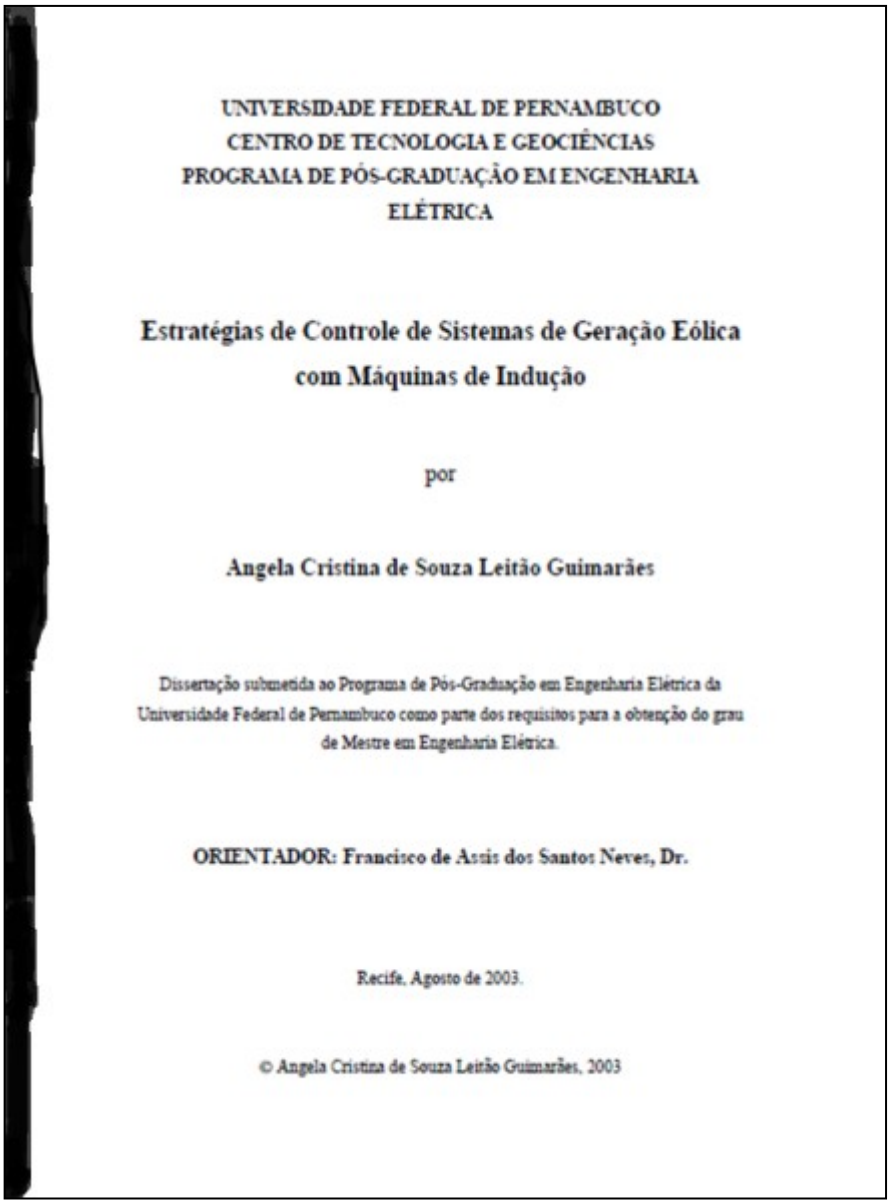
Quando a cópia física do documento é arquivada em espiral, as páginas precisam ser furadas nas bordas. Ao digitalizar o documento, os furos são processados, e passam a fazer parte da imagem do documento. A Figura 3.13 ilustra a imagem do documento com esse tipo de ruído. O prejuízo que esse tipo de ruído traz se refere a inserção de caracteres, freqüentemente a letra “o” e suas variantes, no inicio de cada linha, quebrando a seqüência da escrita do texto.

Quando a cópia física do documento é arquivada em brochura, as páginas são coladas nas bordas. Ao se digitalizar, a parte mais próxima da borda fica mais distante da superfície de aquisição de imagem do equipamento, causando um efeito de perspectiva nas letras iniciais, ou então, são inseridas linhas pretas ao longo da borda.

Para o caso das linhas pretas, o maior inconveniente se configura no maior espaço para armazenamento da cópia digital, e, em especial, para a impressão da cópia digitalizada. A Figura 3.23 ilustra a imagem de um documento com esse tipo de ruído de borda.

Para remoção de bordas foi utilizado o algoritmo ÁVILA-LINS [18] [22] projetada para documentos binarizados e que pertence ao módulo de tratamento da plataforma BigBatch.

No tocante ao efeito de perspectivas nas letras iniciais, existe uma deformação das letras. Para pequenas deformações, as técnicas de descritores de imagem que usam momentos invariantes nos módulos de classificação dos OCR’s auxiliam para minorar o efeito do ruído. No entanto, para grandes deformações, a imagem da letra fica cada vez mais diferente da imagem correta, levando a precisão na transcrição a diminuir.



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Estratégias de Controle de Sistemas de Geração Eólica com Máquinas de Indução

por

Angela Cristina de Souza Leitão Guimarães

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco como parte dos requisitos para a obtenção do grau de Mestre em Engenharia Elétrica.

ORIENTADOR: Francisco de Assis dos Santos Neves, Dr.

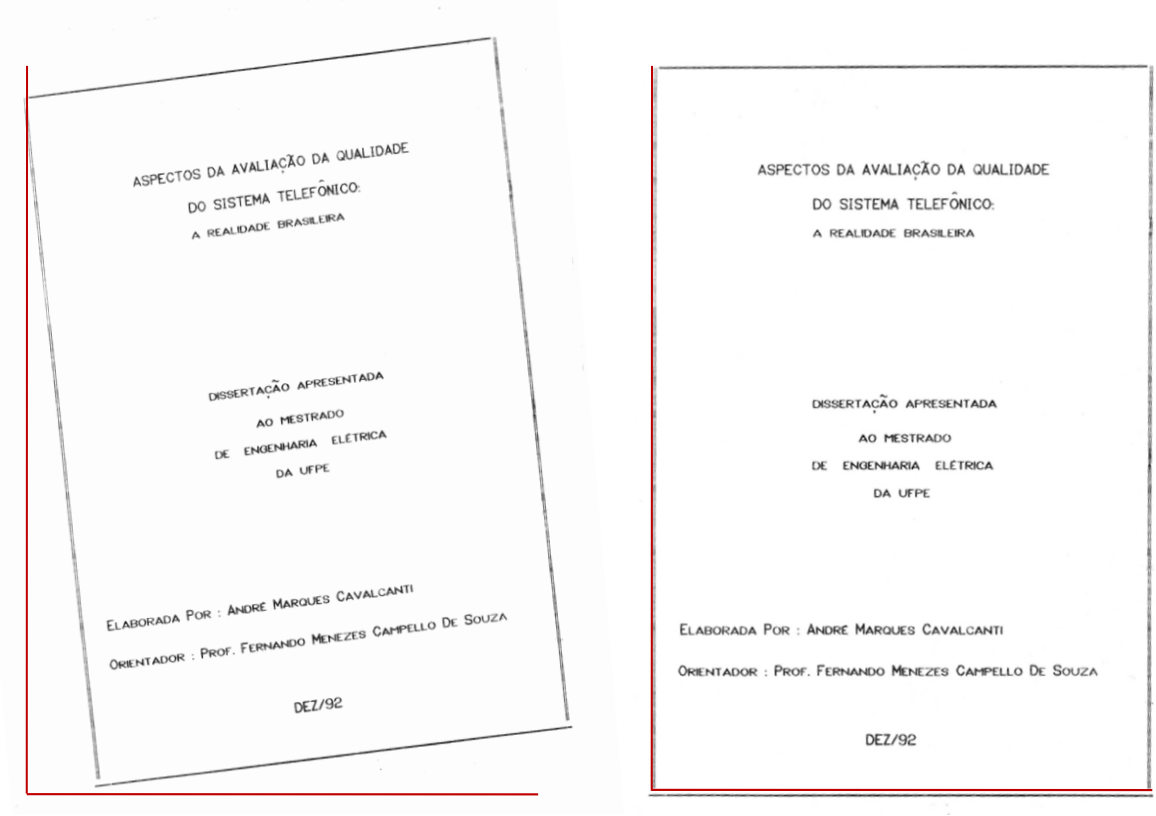
Recife, Agosto de 2003.

Figura 3.23: Ilustração de imagem com ruído de borda do tipo linhas pretas nas bordas.

3.4.3. INCLINAÇÃO E ORIENTAÇÃO

Para alinhamento, foi utilizada uma técnica do módulo de tratamento do BigBatch que consiste em calcular o ângulo de inclinação entre as partes brancas e sequências de pontos pretos (“linhas”) da imagem e depois realizar a rotação [19].

A Figura 3.24 mostra um exemplo de ruído de inclinação numa imagem de um documento (a) e a imagem após o ruído removido (b).



(a) (b)

Figura 3.24: Imagem (a) original com inclinação, (b) com o ruído removido.

Na Figura 3.25, Figura 3.26, Figura 3.27 e Figura 3.28 pode-se encontrar quatro versões de um mesmo documento que sofreram diferentes inclinações e suas respectivas transcrições feitas pelo OCR, para ilustrar a influência do ruído de inclinação no desempenho do OCR.

Na Figura 3.25 tem-se uma versão inclinada do documento em 15º e sua respectiva transcrição sem realizar nenhuma correção. Nota-se que diversas letras não foram “encontradas” como “DISSERTAÇÃO APRESENTADA”, e outras foram transcritas incorretamente “FERNANDO” foi transcrito como “FECHANDO”. De 207 caracteres apenas 49 foram transcritos corretamente, ao passo a disposição das letras (layout) foi profundamente prejudicada.

Na Figura 3.26 há a mesma condição que a anterior, porém para uma inclinação de 5º. Percebe-se um fraco desempenho na extração de texto, tanto na quantidade de caracteres detectados quanto na precisão da transcrição. Mesmo com

uma menor inclinação (7º) não houve uma melhora no desempenho do OCR, de 207 caracteres 39 foram corretamente transcritos.

Na imagem da Figura 3.27 está a versão original do documento sem nenhuma inclinação e sua respectiva transcrição. A precisão do OCR só é questionada sobre a formatação do texto, pois todos os caracteres foram transcritos corretamente.

Na imagem apresentada na Figura 3.28 está uma versão de (a) com o ruído de inclinação removido e sua respectiva transcrição. Com a correção foram o desempenho foi consideravelmente melhorado, visto que de 207 caracteres 186 foram corretamente transcritos e a formatação ficou similar aos casos da transcrição direta do documento original.

3.4.4. REMOÇÃO DO RUÍDO SAL E PIMENTA

Quando uma imagem binária apresenta áreas onde um *pixel* preto aparece circundado de *pixels* brancos, ou ao contrário, um *pixel* branco é circundado de *pixels* pretos, chama-se esse ruído de “sal-e-pimenta”. O ruído sal-e-pimenta [12] é danoso para imagem, pois pode inserir caracteres não existentes ou deformar os caracteres existentes.

Esse é, em geral, um “ruído de digitalização” [15] devido a poeira ou algum tipo de sujeira no scanner ou lente da câmera, por exemplo, mas pode também ser advinda de processamento da imagem como por exemplo na rotação de uma imagem binária.

(a)

(b)

Figura 3.25: Imagem (a) de documento inclinado a 15º e (b) sua transcrição pelo OCR.

ASPECTOS DA AVALIAÇÃO DA QUALIDADE

DO SISTEMA TELEFÔNICO:

A REALIDADE BRASILEIRA

DISSERTAÇÃO APRESENTADA

AO MESTRADO

DE ENGENHARIA ELÉTRICA

DA UFPE

ELABORADA POR : ANDRE MARQUES CAVALCANTI

ORIENTADOR : PROF. FERNANDO MENEZES CAMPELLO DE SOUZA

DEZ/92

(a)

DO SISTEMA TELEFÔNICO-

DEZ/92

(b)

Figura 3.26: (a) Imagem de documento inclinado a 7º e (b) sua transcrição pelo OCR.

ASPECTOS DA AVALIAÇÃO DA QUALIDADE

DO SISTEMA TELEFÔNICO:

A REALIDADE BRASILEIRA

DISSERTAÇÃO APRESENTADA

AO MESTRADO

DE ENGENHARIA ELÉTRICA

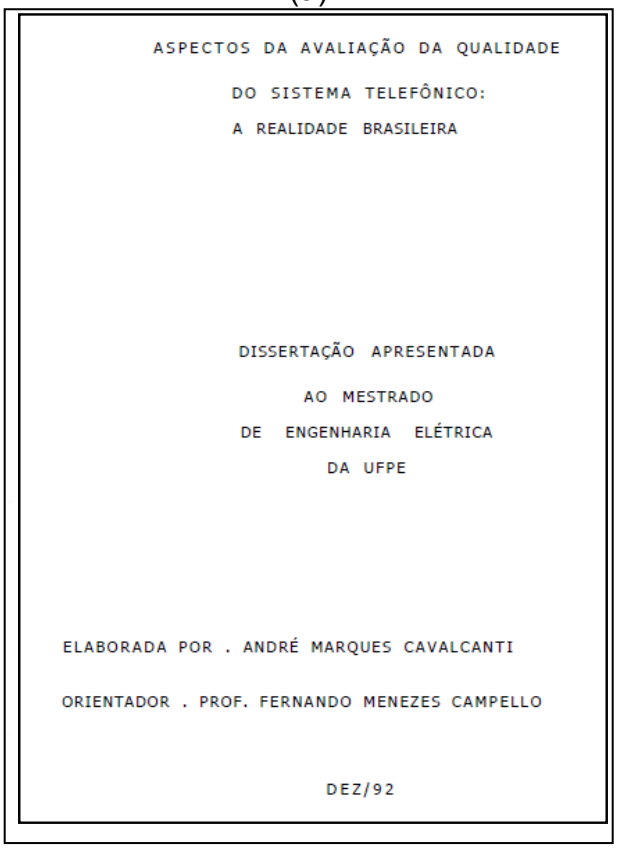
DA UFPE

ELABORADA POR : ANDRÉ MARQUES CAVALCANTI

ORIENTADOR : PROF. FERNANDO MENEZES CAMPELLO DE SOUZA

DEZ/92

(a)



ASPECTOS DA AVALIAÇÃO DA QUALIDADE

DO SISTEMA TELEFÔNICO:

A REALIDADE BRASILEIRA

DISSERTAÇÃO APRESENTADA

AO MESTRADO

DE ENGENHARIA ELÉTRICA

DA UFPE

ELABORADA POR . ANDRÉ MARQUES CAVALCANTI

ORIENTADOR . PROF. FERNANDO MENEZES CAMPELLO

DEZ/92

(b)

Figura 3.27: (a) Imagem de documento original sem inclinação e (b) sua transcrição pelo OCR.

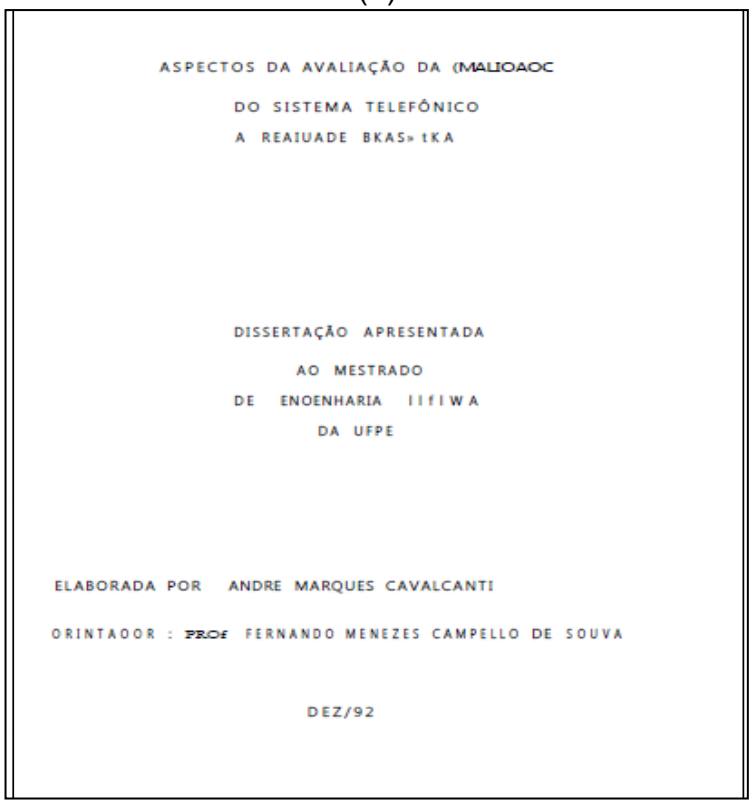
ASPECTOS DA AVALIAÇÃO DA QUALIDADE
DO SISTEMA TELEFÔNICO:
A REALIDADE BRASILEIRA

DISSERTAÇÃO APRESENTADA
AO MESTRADO
DE ENGENHARIA ELÉTRICA
DA UFPE

ELABORADA POR : ANDRÉ MARQUES CAVALCANTI
ORIENTADOR : PROF. FERNANDO MENEZES CAMPELLO DE SOUZA

DEZ/92

(a)



ASPECTOS DA AVALIAÇÃO DA (MALIOAOC
DO SISTEMA TELEFÔNICO
A REAIUADE BKAS» tKA

DISSERTAÇÃO APRESENTADA
AO MESTRADO
DE ENOENHARIA I I f I W A
DA UFPE

ELABORADA POR ANDRE MARQUES CAVALCANTI
ORINTAOOR : PROf FERNANDO MENEZES CAMPELLO DE SOUVA

DEZ/92

(b)

Figura 3.28: (a) Imagem de corrigida do documento inclinado a 15º e (b) sua transcrição pelo OCR.

Bons desempenhos são obtidos com o uso de filtros de média para remoção do ruído sal-e-pimenta [25]. A Figura 3.29 apresenta um trecho de uma imagem e sua respectiva transcrição, em que se encontram ruído tipo sal-e-pimenta.

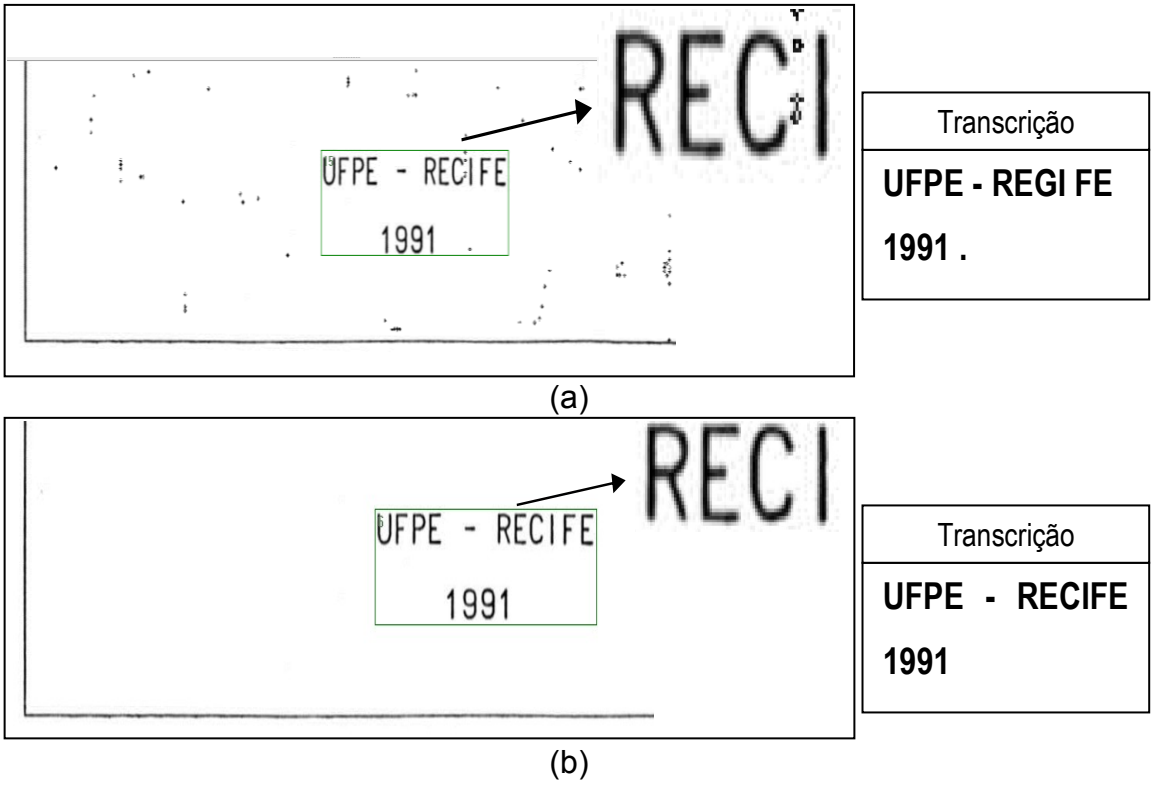


Figura 3.29: (a) Imagem original com ruído sal-e-pimenta e sua transcrição;
(b) imagem com o ruído removido e sua transcrição.

Ainda na Figura 3.29 em (b) visualiza-se o mesmo trecho da imagem após o tratamento e sua respectiva transcrição. Na imagem original, a letra “C” de “RECIFE” foi deformada pelo ruído, levando o OCR a transcrevê-la como a letra “G”. No entanto, para a imagem tratada a letra “C” não foi deformada, e, portanto, transcrita corretamente.

4. INFORMAÇÕES CAPTURADAS

4.1.O MÓDULO DE RECONHECIMENTO

O módulo de reconhecimento de conteúdo tem como entrada documentos em formato digital PDF. Um Documento neste formato contém informações de texto, tamanho e tipo de fonte, e posicionamento do texto no layout da página. Os documentos que foram digitalizados tiveram o texto e suas formatações extraídas pelo OCR e em seguida armazenadas como arquivos PDF. Para captura de conteúdo foram elaborados algoritmos para manipulação de arquivos PDF em linguagem de programação Java®, utilizando o PDFBOX [20].

O fluxograma da Figura 4.1 mostra os processos que os documentos são submetidos para a captura de informações e elaboração do banco de dados. Se o documento já estiver no formato PDF são utilizados os algoritmos do PDFBOX para manipulação, em seguida os textos e suas características são analisados e armazenados.

Caso o documento esteja em cópia impressa, há uma digitalização para obtenção das imagens. Em seguida as imagens são binarizadas na plataforma LiveMemory e tratadas com algoritmos da plataforma BigBatch. Após o tratamento, a imagem do documento tem seu texto extraído pelo OCR e armazenado no formato PDF e txt. Em seguida, o conteúdo desses dois arquivos é unificado e levado para ser analisado pelos algoritmos desenvolvidos na plataforma PDFBOX, que em seguida tem as informações capturadas e armazenadas no banco de dados.

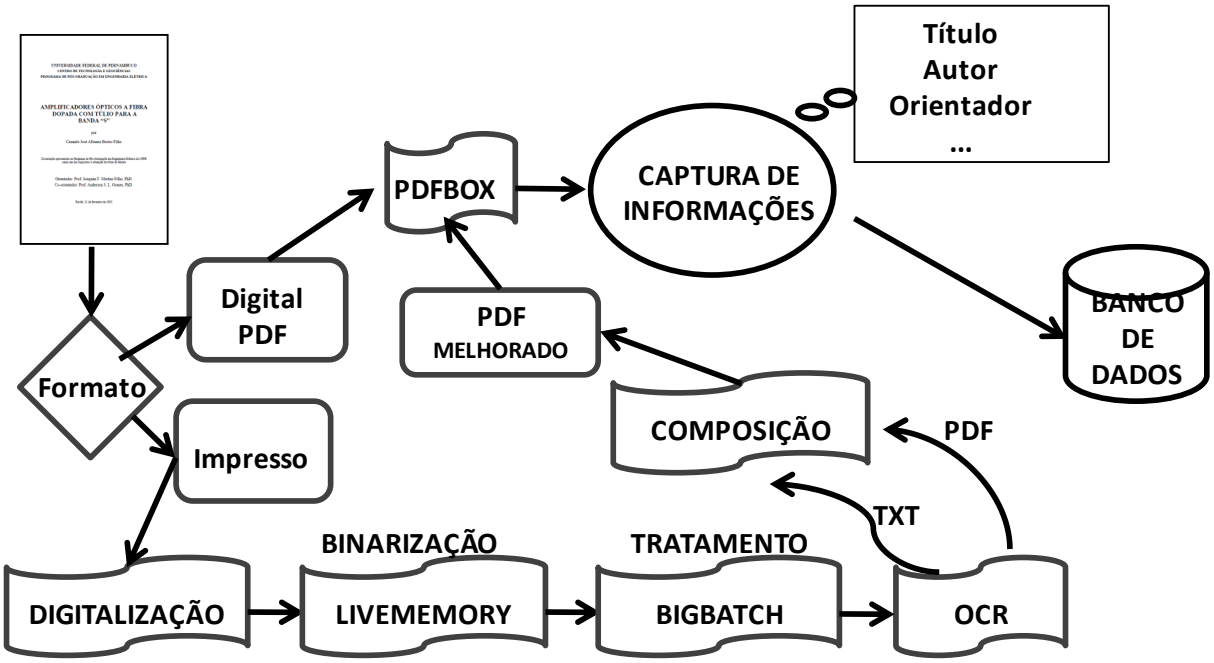


Figura 4.1: Fluxograma ilustrativo do processamento de um documento na plataforma *ACADEMUS* para captura de suas informações.

Para o texto reconhecido pelo OCR e armazenado no formato PDF, são extraídas e armazenadas informações textuais (as palavras) e a formatação do texto. Porém, em alguns casos a formatação não é bem precisa, isto é, o tamanho ou o tipo da fonte divergem do documento original. A Figura 4.2 mostra uma página do documento e uma versão em PDF após o OCR, na qual é possível ver que o tamanho da fonte do título do documento em PDF é menor que a do original. Isso é um problema para o reconhecimento de conteúdo da plataforma, pois essas informações são utilizadas para capturar as informações sobre o documento.

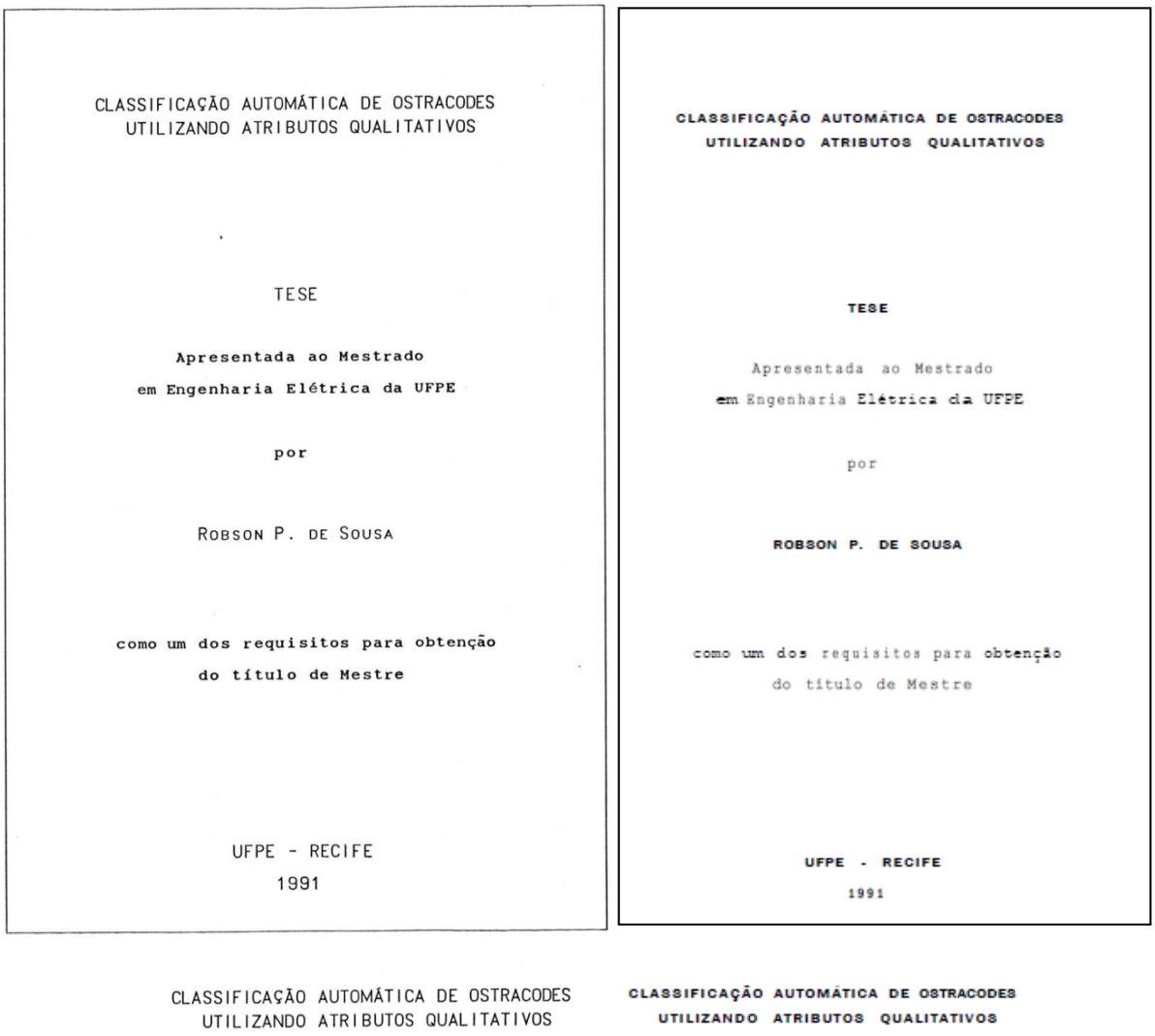


Figura 4.2: Exemplo de arquivo em pdf gerado à partir da transcrição do texto feita pelo OCR.

4.2.SOBRE AS INFORMAÇÕES

Na busca por um documento quanto mais informações sobre ele se tenha maiores são as chances de encontrá-lo, e ainda, com muitas informações sobre os documentos, é possível ainda agrupar, relacionar ou indexar diferentes documentos [44]. O título da obra sempre é a principal informação para identificá-la, como não se trata de um documento com versões ou edições, o título é único. É possível que se pretenda descobrir os trabalhos de um determinado autor, ou de quais obras essa pessoa foi orientadora. As principais informações, que auxiliam na identificação de

um documento, e que são capturadas pelo módulo de reconhecimento do *ACADEMUS* são:

- **Título:** O nome do documento que carrega a idéia principal ou assunto de que trata o texto**.**
- **Autor:** Pessoa ou instituição que elaborou o documento. Pode-se ter mais e um autor por documento, ou num banco de dados uma pessoa pode ser "autor" de diversos documentos.
- **Orientador (co-orientador):** Para os documentos da plataforma *ACADEMUS*, existe uma pessoa que orienta, auxilia na elaboração do documento, mas não é sua autora, essa pessoa é chamada de orientador, e caso haja mais pessoas que orientaram na elaboração do trabalho se fala em co-orientadores.
- **Palavras-chave:** São palavras que sintetizam o assunto que trata o documento. São especialmente fornecidas para auxiliar em sistemas de busca do documento pelo assunto.
- **Área de concentração:** É a área do conhecimento que o assunto trata. Diferentes faculdades ou cursos podem ter áreas de concentração afins, mesmo que as abordagens sejam diferentes.
- ***Link*:** Endereço da WEB ou da rede privada, em que está disponível uma cópia digital do documento.
- **Resumo e Abstract:** Antes de lê completamente uma obra, é importante saber do que ela trata, de onde parte, os resultados que se pretende obter e suas principais observações e conclusões. O conteúdo do resumo e do *abstract* é disponibilizado para os usuários, mesmo que não estejam itens nas ferramentas de busca.
- **Bibliografia:** Tem como objetivo apresentar a bibliografia usada no documento, fornecer os *links* de cada referência, e auxiliar na determinação de citações dos artigos e documentos (teses e dissertações).

A Figura 4.3 apresenta um documento com as informações que são capturadas.

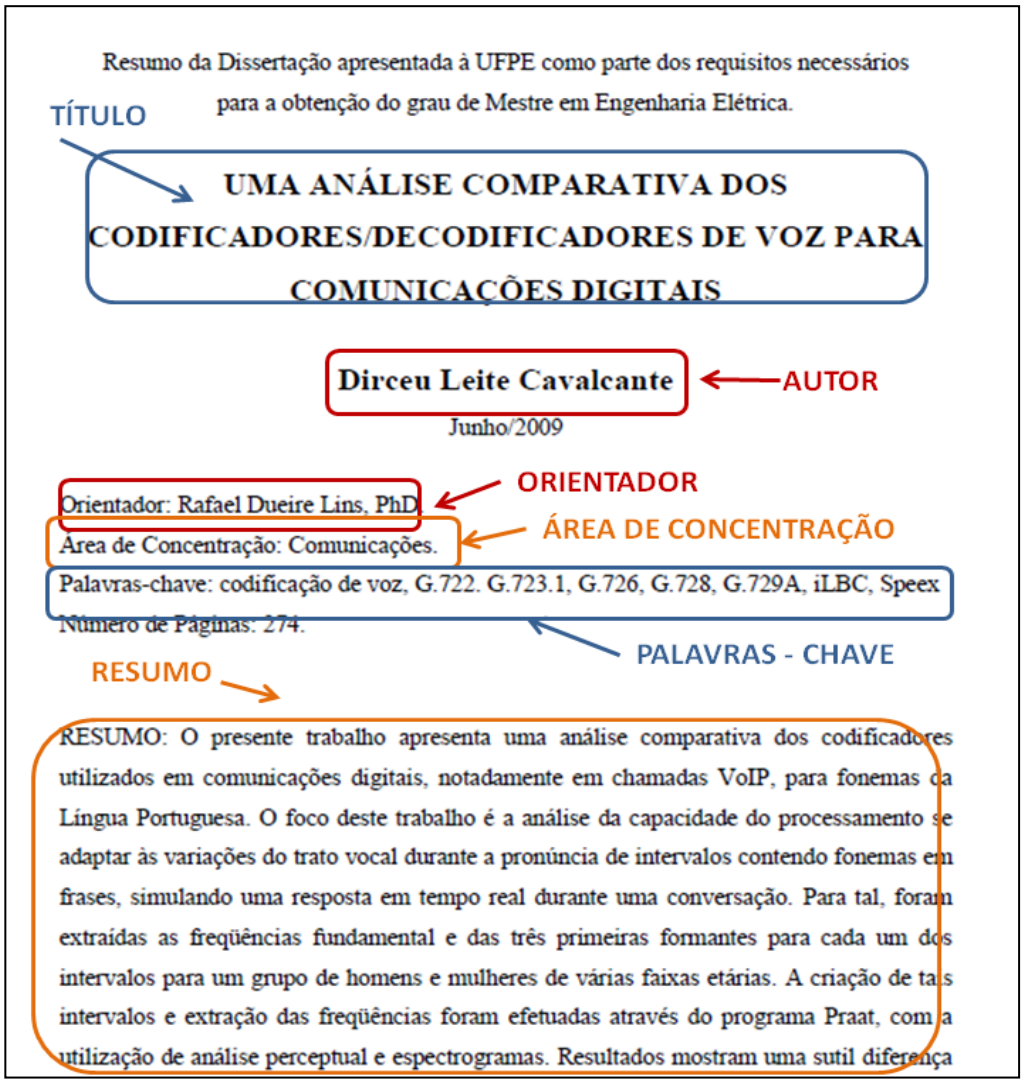


Figura 4.3: Informações capturadas do documento.

4.3. ARMAZENAMENTO DAS INFORMAÇÕES

As informações capturadas são armazenadas em um banco de dados como atributos do documento original. Somente o título é um atributo único e por isso é a principal informação para identificação de um documento. A indexação é uma atividade que visa agrupar ou classificar documentos com atributos afins. Por exemplo, pode-se indexar uma tese de doutorado e uma dissertação de mestrado de um mesmo autor, ou seja, o autor é o atributo de indexação.

Para padronizar a inserção de informações com vistas na fácil e rápida indexação, as informações são armazenadas em arquivos no formato XML. A Figura 4.4 ilustra a estrutura do arquivo XML do banco de dados do *ACADEMUS*, com o título sendo a informação raiz e as demais informações como nós de uma árvore.

```
<?XML version="1.0" encoding="UTF-8" standalone="no"?>
<Academus>
 <Paper>
        <title>Título do documento</title>
            <autor> Autor(a) </autor>
            <orientador> Prof. Orientador.</orientador>
            <coorientador> Prof. Co-rientador. </coorientador>
            <words> Palavras – chave </words>
```

Figura 4.4: Ilustração da estrutura de informações em árvore no banco de dados do *ACADEMUS*, em que se utiliza o formato XML.

4.4. AS REFERÊNCIAS BIBLIOGRÁFICAS

Os trabalhos desenvolvidos em programas de pós-graduação de universidades citam outros trabalhos de pós-graduação e artigos de conferências, mas essas citações não são computadas e os trabalhos não são interligados, ou não tem reconhecimento de citações entre as editoras. Uma das informações sobre o documento tratada pela plataforma *ACADEMUS* é a bibliografia, que consiste na busca por referências bibliográficas contidas em cada documento, segmentação e contabiliza cada citação.

Na citação existem diversas informações que auxiliam na indexação de documentos e em sua busca, como nome de autor, títulos, conferência de publicação, *link* da WEB onde tem uma cópia disponível, etc.

As referências dos documentos são inseridas no banco de dados da plataforma como atributos do documento original. Ao ser encontrada e segmentada, cada referência é um atributo do documento original, sendo armazenada no banco de dados de referência e recebendo um rótulo, um *link*. Esse *link* passa a ser o atributo da citação do arquivo original.

4.5. ESTRATÉGIAS PARA CAPTURA DAS INFORMAÇÕES

A falta de padronização na disposição das informações (texto) no documento (*layout*) dificulta a identificação de conteúdo. Em abordagens de reconhecimento de conteúdo e captura de informações para geração de bibliotecas digitais, como o LiveMemory, é possível identificar e extrair as informações desejadas com estratégias de captura de informações numa determinada região e área do documento. Por exemplo, nos artigos recentes do Simpósio Brasileiro de Telecomunicações, o evento técnico anual da SBrT (Sociedade Brasileira de Telecomunicações), na parte superior da primeira página do documento, a primeira linha contém a edição da conferência, enquanto que na segunda linha está o título do documento e nas linhas seguintes os nomes dos autores. Esse é o padrão, portanto todos os documentos estão assim. A Figura 4.5 mostra um exemplo com a disposição dessas informações.

A segmentação por área realizada apresentou bons resultados, evidenciando a importância da padronização. Apesar de ter modelos para elaboração da tese ou dissertação, os documentos abordados pela plataforma *ACADEMUS* não têm padronização, o que invalida o uso da estratégia de segmentação por área.

Com a indisponibilidade de utilizar a abordagem de área para extração de conteúdo, cada linha do documento foi analisada isoladamente, sendo tratada como um vetor de quatro componentes, que contem o texto e sua formatação: texto, tamanho da fonte das palavras, tipo da fonte das palavras, e posição do texto no layout da página.

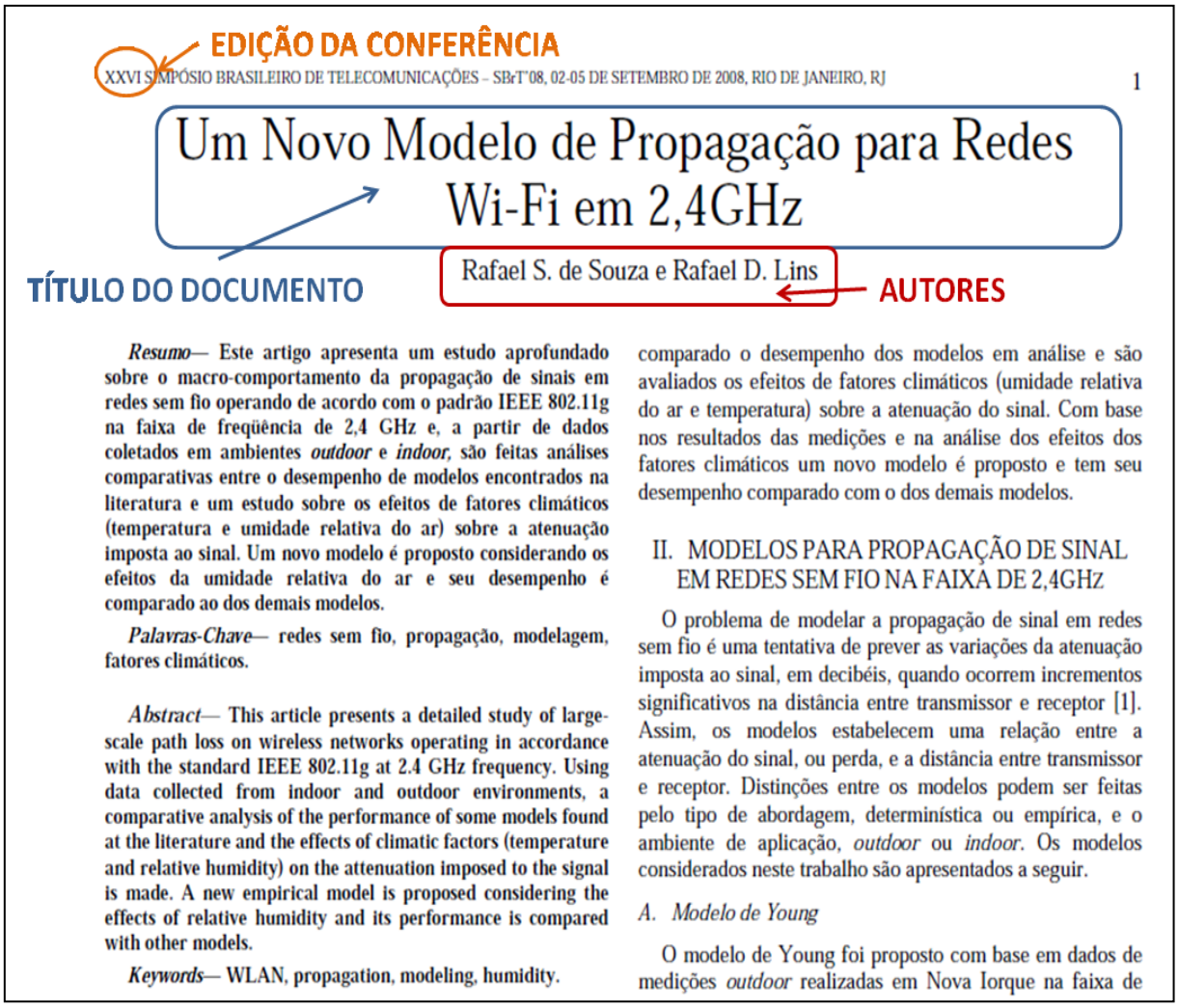


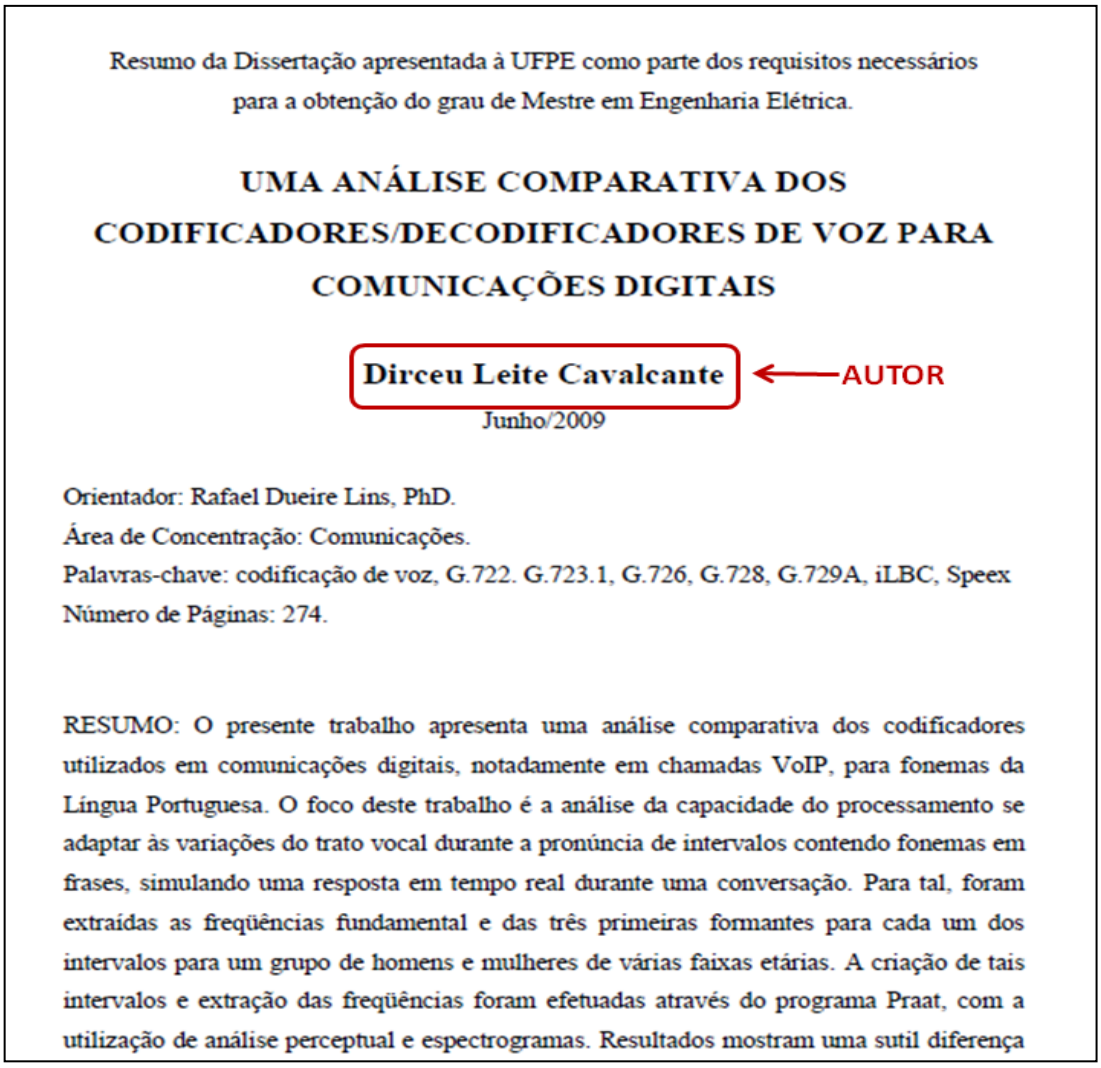
Figura 4.5: Ilustração da disposição de informações em trecho de artigo submetido ao SBrT.

Essas são as informações utilizadas para o reconhecimento de conteúdo e determinação das informações sobre o documento. De acordo com essas características e com o uso de estratégia de busca e análise de palavras “reservadas”, que sempre são encontradas neste tipo de documento, e de um dicionário controlado para o nome de autores, é feito o reconhecimento de conteúdo.

É importante lembrar que os documentos digitalizados são analisados no formato PDF, mas são obtidos por uma composição de dois arquivos obtidos do processo de OCR. Para esse tipo de documento, existe uma perda na informação do tipo e tamanho da fonte após a transcrição do texto, inviabilizando a estratégia que utiliza formatações para se capturar alguma informação.

4.5.1. AUTOR

A Figura 4.6 destaca a informação sobre o autor numa página do documento. Percebe-se que o nome do autor está logo abaixo do título do documento. Essa é uma estratégia para captura do nome do autor do documento.


Resumo da Dissertação apresentada à UFPE como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

UMA ANÁLISE COMPARATIVA DOS CODIFICADORES/DECODIFICADORES DE VOZ PARA COMUNICAÇÕES DIGITAIS

Dirceu Leite Cavalcante ← AUTOR

Junho/2009

Orientador: Rafael Dueire Lins, PhD.
Área de Concentração: Comunicações.
Palavras-chave: codificação de voz, G.722. G.723.1, G.726, G.728, G.729A, iLBC, Speex
Número de Páginas: 274.

RESUMO: O presente trabalho apresenta uma análise comparativa dos codificadores utilizados em comunicações digitais, notadamente em chamadas VoIP, para fonemas da Língua Portuguesa. O foco deste trabalho é a análise da capacidade do processamento se adaptar às variações do trato vocal durante a pronúncia de intervalos contendo fonemas em frases, simulando uma resposta em tempo real durante uma conversação. Para tal, foram extraídas as freqüências fundamental e das três primeiras formantes para cada um dos intervalos para um grupo de homens e mulheres de várias faixas etárias. A criação de tais intervalos e extração das freqüências foram efetuadas através do programa Praat, com a utilização de análise perceptual e espectrogramas. Resultados mostram uma sutil diferença

Figura 4.6: Ilustração das informações sobre “autor” de um documento.

Porém, não existe um padrão na disposição das informações, impossibilitando que essa seja a única estratégia. Em alguns casos o nome do autor é precedido da palavra “elaborado por”, ou seja, o texto após essas palavras contém o nome do autor.

Em outros casos nenhuma das situações anteriores acontecem, assim é utilizado um dicionário controlado, que contém o nome de todos os autores. Ao se

checar uma linha que contenha um nome dentro do dicionário, aquele texto é tido como o nome do autor.

4.5.2. TÍTULO

Para captura do título do documento uma característica é interessante: o texto do título tem o maior tamanho de fonte da página, em outros casos o texto do título é precedido pela identificação da universidade, centro acadêmico e curso do autor. A Figura 4.7 destaca a informação sobre o título numa página do documento.

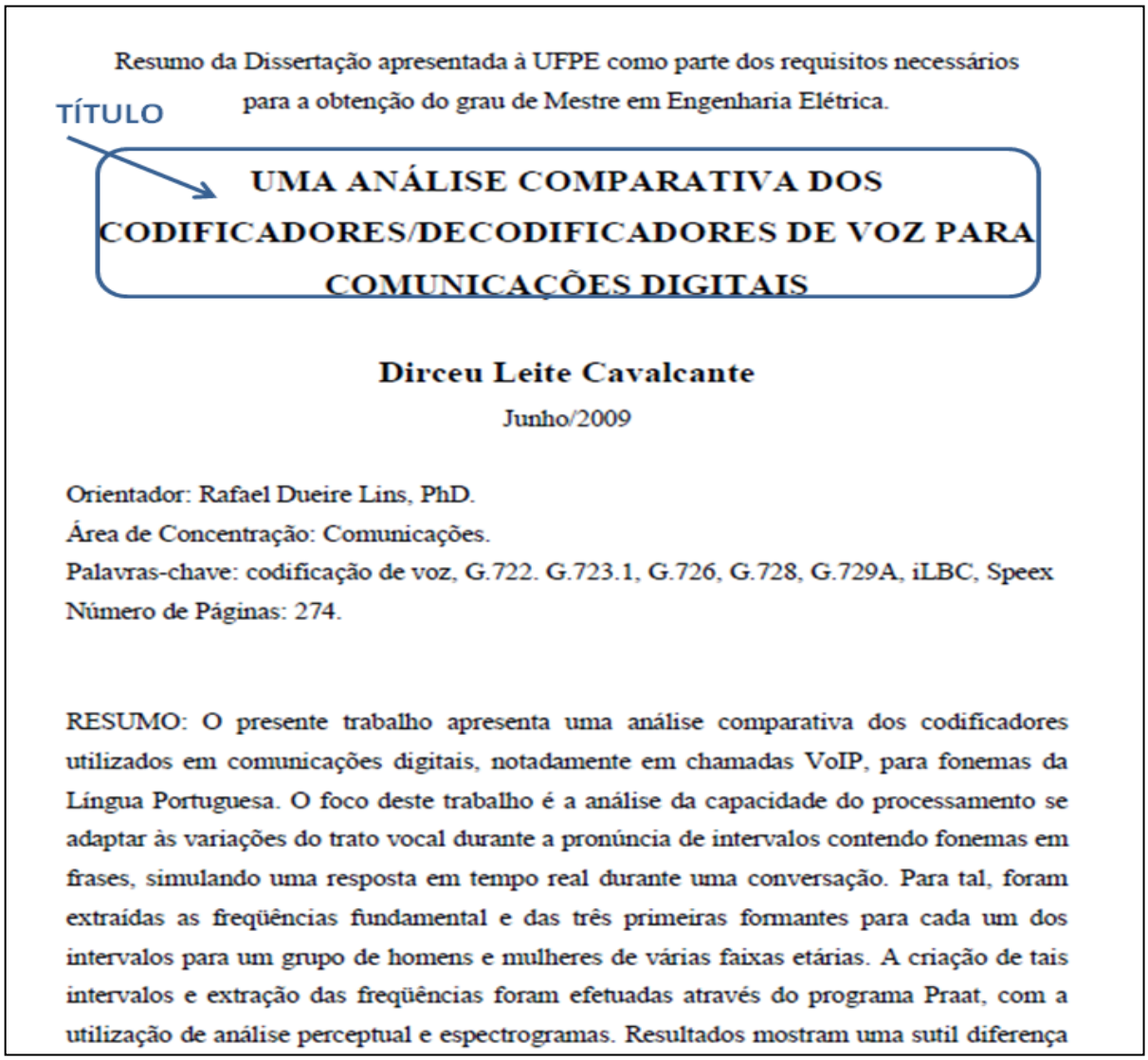
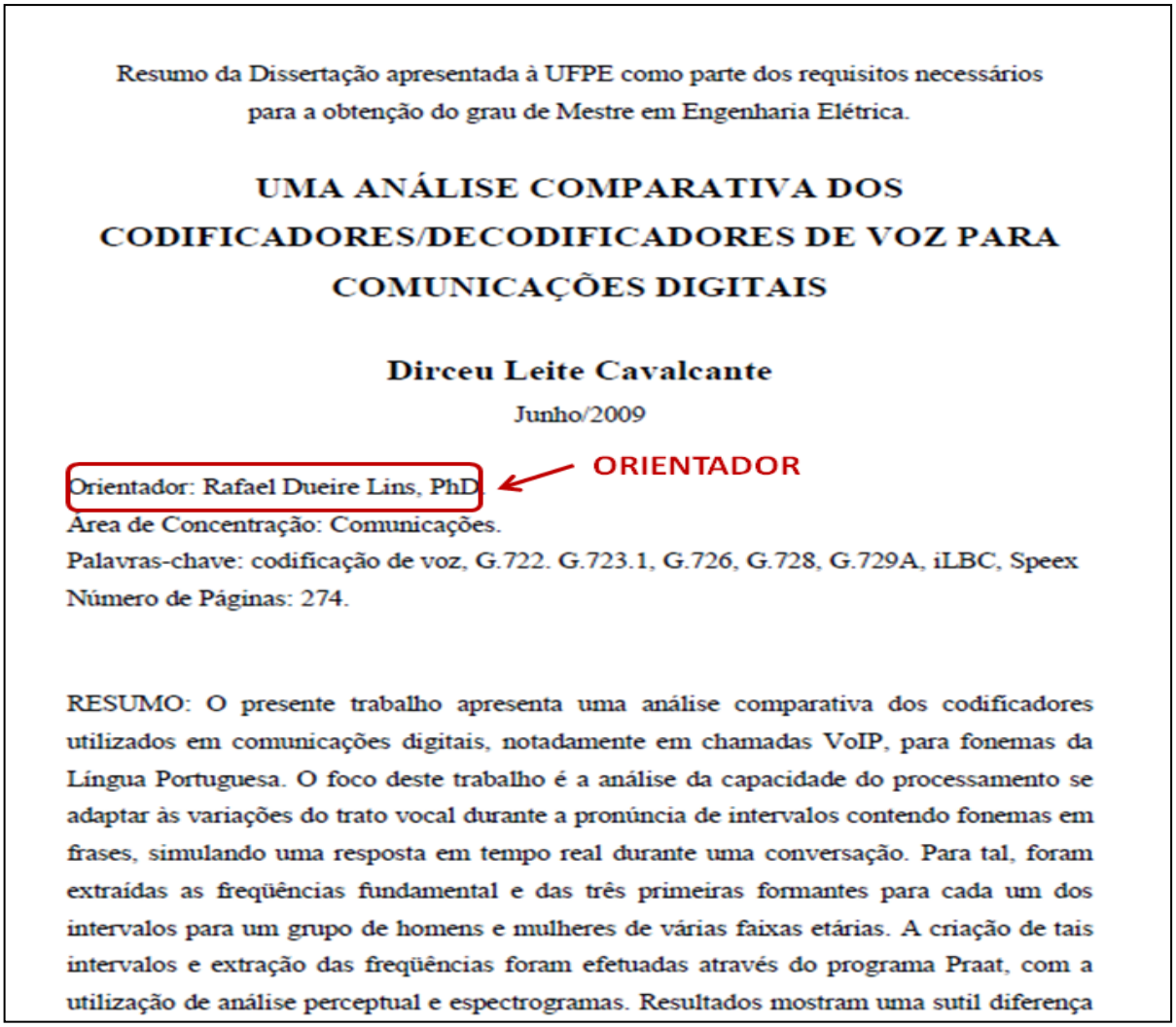


Figura 4.7: Ilustração das informações sobre “título” de um documento.

4.5.3. ORIENTADOR

Para se obter o nome do orientador do documento procurou-se por linha a palavra “orientador”, e suas variantes, extraindo o texto posterior dessa palavra e que esteja na mesma linha. A Figura 4.8 destaca a informação sobre o orientador numa página do documento.


Resumo da Dissertação apresentada à UFPE como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

UMA ANÁLISE COMPARATIVA DOS CODIFICADORES/DECODIFICADORES DE VOZ PARA COMUNICAÇÕES DIGITAIS

Dirceu Leite Cavalcante

Junho/2009

Orientador: Rafael Dueire Lins, PhD

ORIENTADOR

Area de Concentração: Comunicações.

Palavras-chave: codificação de voz, G.722. G.723.1, G.726, G.728, G.729A, iLBC, Speex

Número de Páginas: 274.

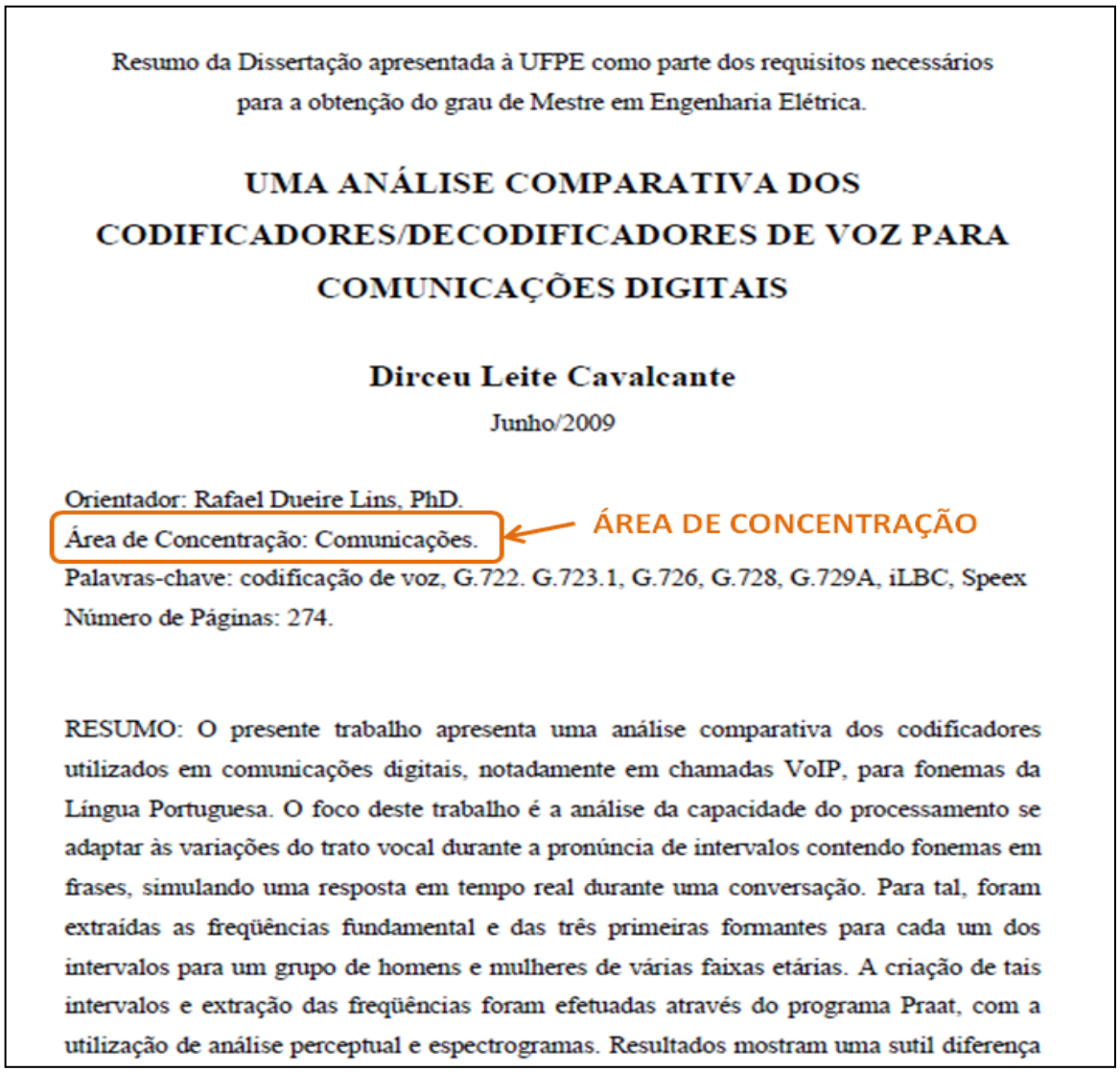
RESUMO: O presente trabalho apresenta uma análise comparativa dos codificadores utilizados em comunicações digitais, notadamente em chamadas VoIP, para fonemas da Língua Portuguesa. O foco deste trabalho é a análise da capacidade do processamento se adaptar às variações do trato vocal durante a pronúncia de intervalos contendo fonemas em frases, simulando uma resposta em tempo real durante uma conversação. Para tal, foram extraídas as freqüências fundamental e das três primeiras formantes para cada um dos intervalos para um grupo de homens e mulheres de várias faixas etárias. A criação de tais intervalos e extração das freqüências foram efetuadas através do programa Praat, com a utilização de análise perceptual e espectrogramas. Resultados mostram uma sutil diferença

Figura 4.8: Ilustração das informações sobre “orientador” de um documento.

Da mesma forma é feito na busca por co-orientadores do documento, procura-se em cada linha a palavra “co-orientador” e suas variantes extraindo o texto posterior dessa palavra e que esteja na mesma linha. É feita uma comparação com o texto do orientador, excluindo-o como co-orientador.

4.5.4. ÁREA DE CONCENTRAÇÃO

Para se obter a área de concentração do documento procurou-se por linha a palavra “área de concentração”, e suas variantes, extraindo o texto posterior dessa palavra e que esteja na mesma linha. A Figura 4.9 destaca a informação sobre a área de concentração numa página do documento.



Resumo da Dissertação apresentada à UFPE como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

UMA ANÁLISE COMPARATIVA DOS CODIFICADORES/DECODIFICADORES DE VOZ PARA COMUNICAÇÕES DIGITAIS

Dirceu Leite Cavalcante

Junho/2009

Orientador: Rafael Dueire Lins, PhD.

Área de Concentração: Comunicações.

ÁREA DE CONCENTRAÇÃO

Palavras-chave: codificação de voz, G.722. G.723.1, G.726, G.728, G.729A, iLBC, Speex

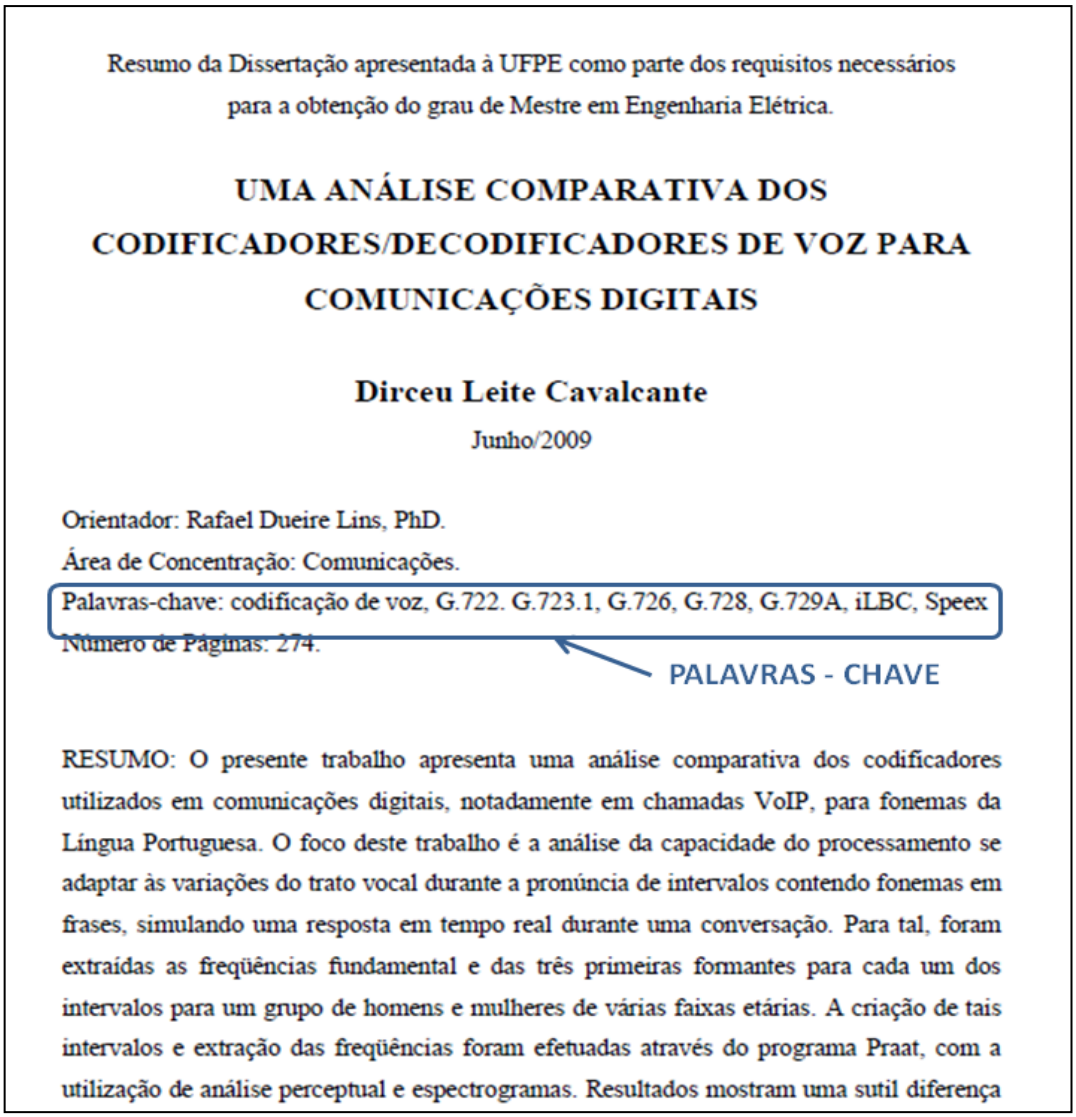
Número de Páginas: 274.

RESUMO: O presente trabalho apresenta uma análise comparativa dos codificadores utilizados em comunicações digitais, notadamente em chamadas VoIP, para fonemas da Língua Portuguesa. O foco deste trabalho é a análise da capacidade do processamento se adaptar às variações do trato vocal durante a pronúncia de intervalos contendo fonemas em frases, simulando uma resposta em tempo real durante uma conversação. Para tal, foram extraídas as freqüências fundamental e das três primeiras formantes para cada um dos intervalos para um grupo de homens e mulheres de várias faixas etárias. A criação de tais intervalos e extração das freqüências foram efetuadas através do programa Praat, com a utilização de análise perceptual e espectrogramas. Resultados mostram uma sutil diferença

Figura 4.9: Ilustração das informações sobre “área de concentração” de um documento.

4.5.5. PALAVRAS CHAVE

Para se obter as palavras chave do documento procurou-se por linha a palavra “palavras-chave”, “*keywords*”, e suas variantes, extraindo o texto posterior dessa palavra e que esteja na mesma linha. A Figura 4.10 destaca a informação sobre as palavras chave numa página do documento.



Resumo da Dissertação apresentada à UFPE como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

UMA ANÁLISE COMPARATIVA DOS CODIFICADORES/DECODIFICADORES DE VOZ PARA COMUNICAÇÕES DIGITAIS

Dirceu Leite Cavalcante

Junho/2009

Orientador: Rafael Dueire Lins, PhD.

Área de Concentração: Comunicações.

Palavras-chave: codificação de voz, G.722. G.723.1, G.726, G.728, G.729A, iLBC, Speex

Número de Páginas: 274.

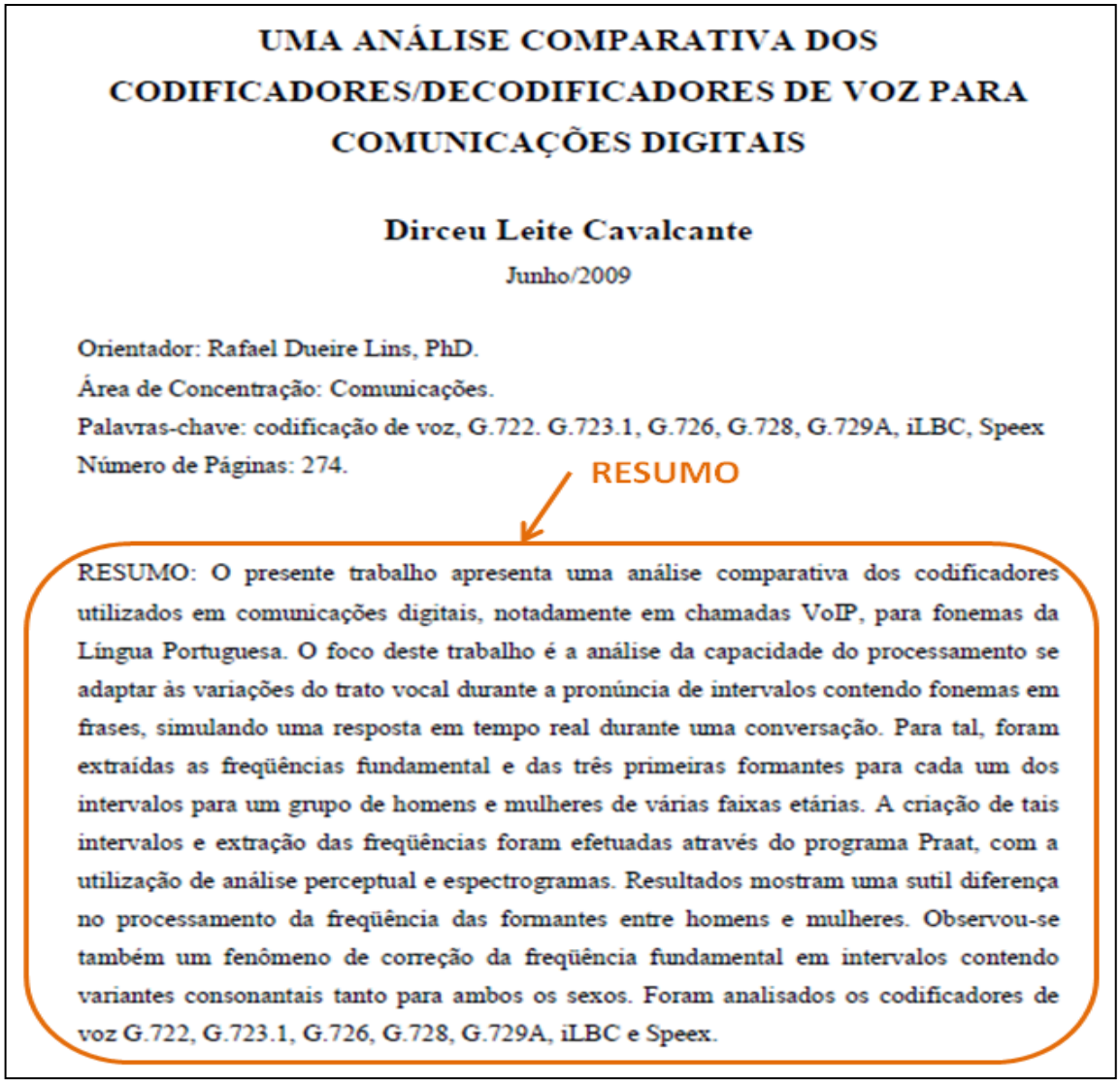
PALAVRAS - CHAVE

RESUMO: O presente trabalho apresenta uma análise comparativa dos codificadores utilizados em comunicações digitais, notadamente em chamadas VoIP, para fonemas da Língua Portuguesa. O foco deste trabalho é a análise da capacidade do processamento se adaptar às variações do trato vocal durante a pronúncia de intervalos contendo fonemas em frases, simulando uma resposta em tempo real durante uma conversação. Para tal, foram extraídas as freqüências fundamental e das três primeiras formantes para cada um dos intervalos para um grupo de homens e mulheres de várias faixas etárias. A criação de tais intervalos e extração das freqüências foram efetuadas através do programa Praat, com a utilização de análise perceptual e espectrogramas. Resultados mostram uma sutil diferença

Figura 4.10: Ilustração das informações sobre “palavras-chave” de um documento.

4.5.6. RESUMO

Para se obter o resumo do documento procurou-se por linha a palavra "resumo", e suas variantes, extraindo o texto posterior dessa palavra e que esteja na mesma página, ou até se encontrar a palavra "*abstract*" ou "palavras-chave". A Figura 4.11 destaca a informação sobre o resumo numa página do documento.


UMA ANÁLISE COMPARATIVA DOS CODIFICADORES/DECODIFICADORES DE VOZ PARA COMUNICAÇÕES DIGITAIS

Dirceu Leite Cavalcante

Junho/2009

Orientador: Rafael Dueire Lins, PhD.
Área de Concentração: Comunicações.
Palavras-chave: codificação de voz, G.722. G.723.1, G.726, G.728, G.729A, iLBC, Speex
Número de Páginas: 274.

RESUMO

RESUMO: O presente trabalho apresenta uma análise comparativa dos codificadores utilizados em comunicações digitais, notadamente em chamadas VoIP, para fonemas da Língua Portuguesa. O foco deste trabalho é a análise da capacidade do processamento se adaptar às variações do trato vocal durante a pronúncia de intervalos contendo fonemas em frases, simulando uma resposta em tempo real durante uma conversação. Para tal, foram extraídas as freqüências fundamental e das três primeiras formantes para cada um dos intervalos para um grupo de homens e mulheres de várias faixas etárias. A criação de tais intervalos e extração das freqüências foram efetuadas através do programa Praat, com a utilização de análise perceptual e espectrogramas. Resultados mostram uma sutil diferença no processamento da freqüência das formantes entre homens e mulheres. Observou-se também um fenômeno de correção da freqüência fundamental em intervalos contendo variantes consonantais tanto para ambos os sexos. Foram analisados os codificadores de voz G.722, G.723.1, G.726, G.728, G.729A, iLBC e Speex.

Figura 4.11: Ilustração das informações sobre "resumo" de um documento.

4.5.7. ABSTRACT

Para se obter o *abstract* do documento procurou-se por linha a palavra "*abstract*", e suas variantes, extraindo o texto posterior dessa palavra e que esteja na mesma página, ou até se encontrar a palavra "*keywords*". A Figura 4.12 destaca a informação sobre o *abstract* numa página do documento.

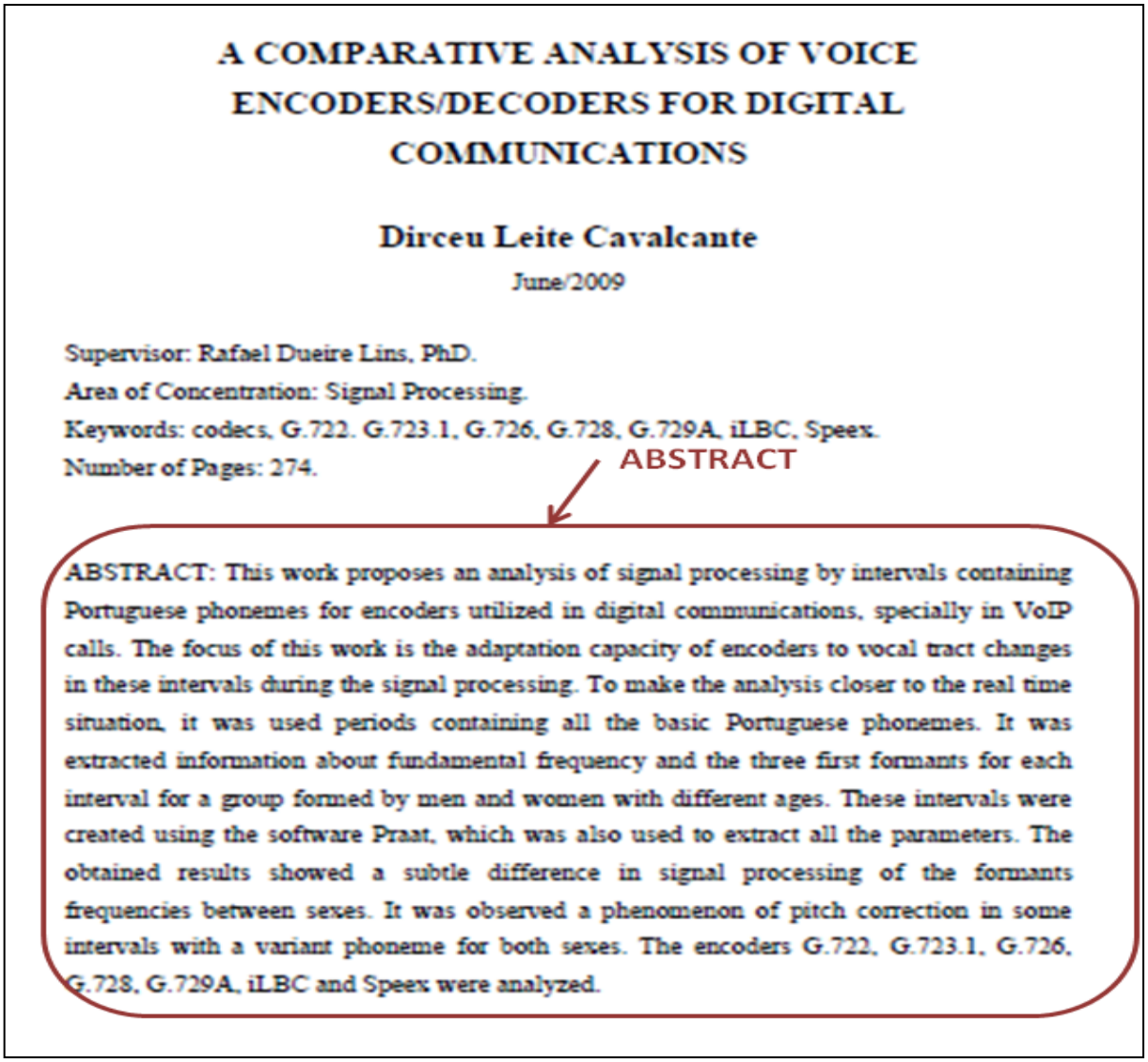
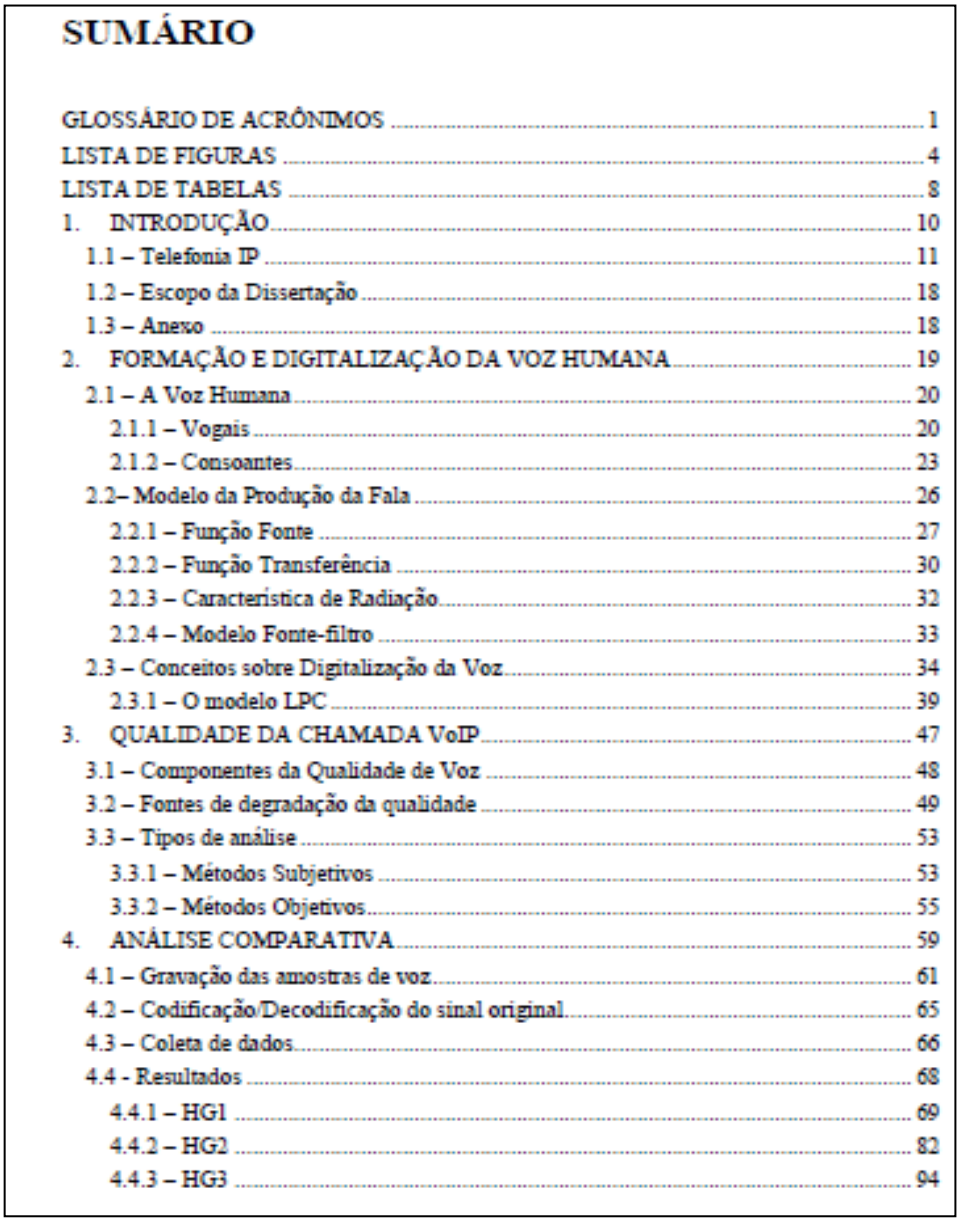


Figura 4.12: Ilustração das informações sobre "*abstract*" de um documento.

4.5.8. ÍNDICE

Para se obter o conteúdo do índice do documento procurou-se por linha a palavra “índice”, “sumário”, e suas variantes, extraindo o texto até encontrar palavras reservadas, como: “capítulo”, “lista de figuras”, “resumo”, “glossário”, indicam o término do texto sobre o Índice. Todo o texto foi extraído para ser segmentado em seguida por item. A Figura 4.13 destaca a informação sobre o índice numa página do documento.

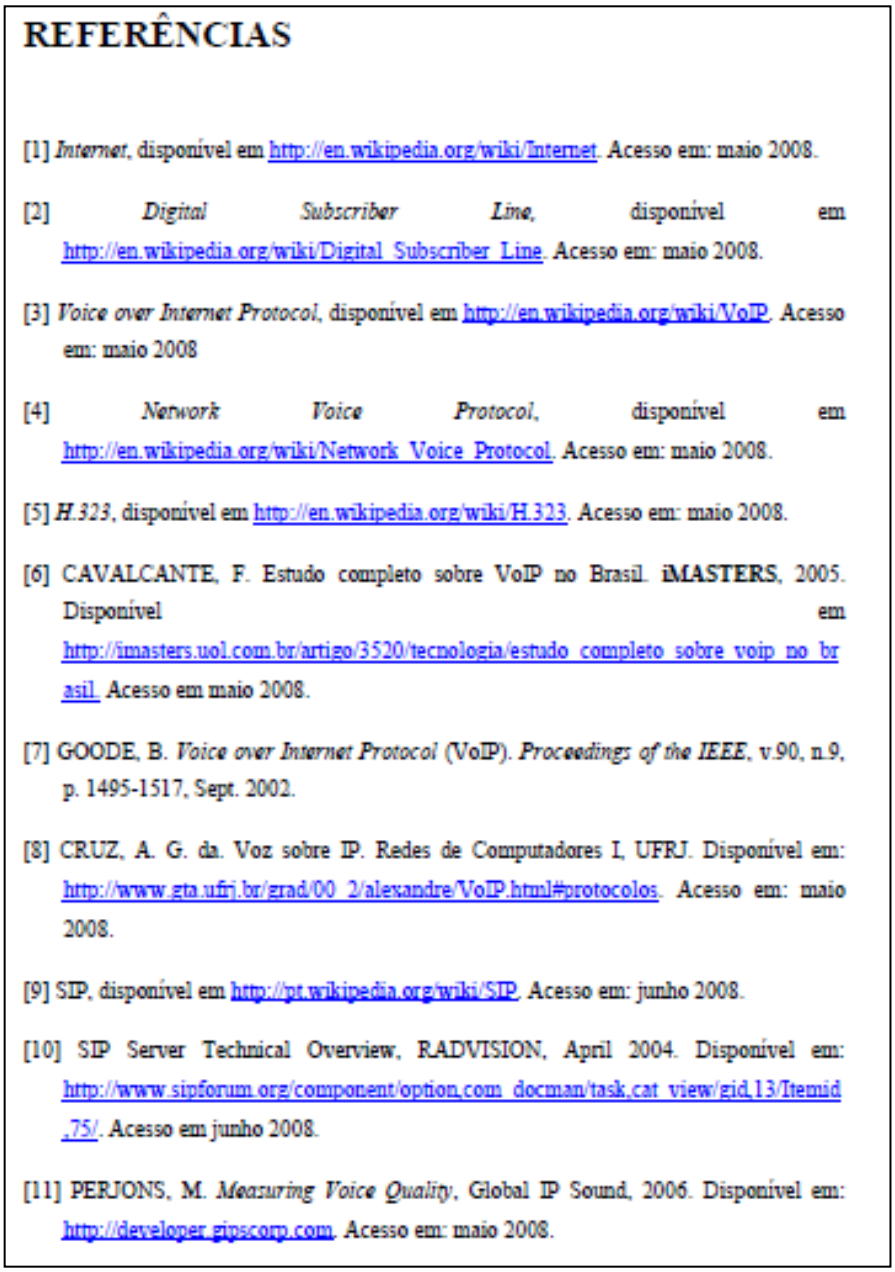


SUMÁRIO

Figura 4.13: Ilustração das informações sobre “índice” de um documento.

4.5.9. BIBLIOGRAFIA

Para se obter o conteúdo da bibliografia do documento procurou-se por linha a palavra “bibliografia”, “referências”, e suas variantes, extraindo o texto até ao final do documento ou até encontrar a palavra “anexo” ou “apêndice”, e variantes. Todo o texto foi extraído para ser segmentado em seguida por referência. A Figura 4.14 destaca a informação sobre a bibliografia numa página do documento.



REFERÊNCIAS

[1] *Internet*, disponível em http://en.wikipedia.org/wiki/Internet. Acesso em: maio 2008.

[2] *Digital Subscriber Line*, disponível em http://en.wikipedia.org/wiki/Digital_Subscriber_Line. Acesso em: maio 2008.

[3] *Voice over Internet Protocol*, disponível em http://en.wikipedia.org/wiki/VoIP. Acesso em: maio 2008

[4] *Network Voice Protocol*, disponível em http://en.wikipedia.org/wiki/Network_Voice_Protocol. Acesso em: maio 2008.

[5] *H.323*, disponível em http://en.wikipedia.org/wiki/H.323. Acesso em: maio 2008.

[6] CAVALCANTE, F. Estudo completo sobre VoIP no Brasil. **iMASTERS**, 2005. Disponível em http://imasters.uol.com.br/artigo/3520/tecnologia/estudo_completo_sobre_voip_no_brasil. Acesso em maio 2008.

[7] GOODE, B. *Voice over Internet Protocol* (VoIP). *Proceedings of the IEEE*, v.90, n.9, p. 1495-1517, Sept. 2002.

[8] CRUZ, A. G. da. Voz sobre IP. Redes de Computadores I, UFRJ. Disponível em: http://www.gta.ufrj.br/grad/00_2/alexandre/VoIP.html#protocolos. Acesso em: maio 2008.

[9] SIP, disponível em http://pt.wikipedia.org/wiki/SIP. Acesso em: junho 2008.

[10] SIP Server Technical Overview, RADVISION, April 2004. Disponível em: http://www.sipforum.org/component/option,com_docman/task,cat_view/gid,13/Itemid,75/. Acesso em junho 2008.

[11] PERJONS, M. *Measuring Voice Quality*, Global IP Sound, 2006. Disponível em: http://developer.gipscorp.com. Acesso em: maio 2008.

Figura 4.14: Ilustração das informações sobre “bibliografia” de um documento.

As referências não estão padronizadas, isto é, existem diversas formas de se fazer uma citação de trabalho externo como: colchetes+número [1111]; colchetes+texto [aaa]; colchetes+texto+número [aaaa, 111]; parênteses+número (1111); parênteses+texto (aaa); parênteses +texto+número (aaaa, 111). Erros na forma de descrever a referência são frequentes e isso atrapalha na identificação. A Figura 4.15 apresenta exemplos de citação existentes em um documento de dissertação de mestrado.

(31) *5-, 4-, 3- and 2-bits sample embedded Adaptative Differential Pulse Modulation (ADPCM)*, ITU-T Recommendation G.727, 1990.
(24) *Vocabulary of Digital Transmission and Multiplexing, and Pulse Code Modulation (PCM) Terms*. ITU-T Recomendation G.701. 1993.
(10) Maron, M.: Automatic Indexing: an experimental inquiry. Journal of the ACM (JACM) 8(3), 404-417 (1986).

Figura 4.15: Ilustração de citações que podem gerar erros na estratégia de segmentação de citações empregada.

Uma rápida análise revela que há duas potenciais formas de citação: parêntese+número (10), e parênteses+texto (JACM). Isso gera um conflito, pois de uma única referência real do documento podem ser extraídas duas referências virtuais e incorretas. Para resolver isso foi usada uma rede neural com uma única saída e seis neurônios, que identifica o tipo de indexação, aprende durante o processo e não reconhece outras formas de citação no atual documento. Portanto, a forma de citação parênteses+texto (JACM) não é computada como uma referência válida e faz parte a referência anterior parênteses+número (10).

5. INTERFACE

5.1. FUNCIONALIDADES DA PLATAFORMA

A plataforma *ACADEMUS* foi desenvolvida em linguagem de programação Java, linguagem selecionada por causa de sua portabilidade e facilidade de uso em diversos sistemas operacionais (Windows, MacOs, Unix). As técnicas de reconhecimento de conteúdo e para tratamento do documento PDF foram desenvolvidas em Java, enquanto que os algoritmos para tratamento de imagens foram implementados em linguagem C++.

Para a manipulação de documentos no formato PDF foi utilizada a plataforma PDFBOX [20], já para manipulação e tratamento das imagens as técnicas usadas foram implementadas na plataforma BigBatch. Para a extração do texto a partir das imagens foi utilizado o software Abby FineReader 9.0.6 [17].

Após realizar a captura das informações do documento a partir do reconhecimento de conteúdo, os resultados são armazenados em arquivos no formato XML. Esse formato para armazenamento do banco de dados é interessante, pois permite segmentar as informações, e principalmente, porque é um formato universal utilizado por diversos programas, plataformas, sistemas operacionais, e até por páginas da WEB.

A plataforma está dividida em dois módulos: busca e reconhecimento. Cada módulo tem uma interface específica, projetada para conter todas as informações necessárias para operação, e de modo intuitivo. No módulo de busca podem ser pesquisados documentos através de informações sobre o título, o autor, orientador, área de concentração e palavras-chave do documento. Já no módulo de

reconhecimento são disponibilizadas as rotinas para captura de informações e ferramentas para extração de texto de imagens.

5.2. MÓDULO DE BUSCA

Na plataforma foi desenvolvida uma interface de busca para se fazer consultas mais facilmente ao banco de dados, sem a necessidade de programas externos. O acesso aos arquivos do banco de dados não é restrito à plataforma, nessa interface, sendo apenas uma solução na própria plataforma.

A indexação de informação é um dos principais objetivos de bibliotecas digitais, em especial para interconexão de documentos afins. O módulo de busca *ACADEMUS* permite o acesso a informações sobre o título, autor, orientador, palavras-chave, e área de concentração dos documentos que integram sua base de dados.

São listados todos os títulos dos documentos que contêm o argumento da busca. Cada item da lista pode ter seu conteúdo acessado, disponibilizando todas as informações capturadas sobre o documento, e uma versão digital em PDF do documento.

A Figura 5.1 mostra a interface do módulo de busca do *ACADEMUS*, em que são feitas pesquisas no banco de dados das informações dos documentos. Para realizar uma busca deve-se digitar o texto e selecionar o tipo de busca. Há a opção da interface em inglês que pode ser selecionada nos botões seletores de idioma. Para acessar o módulo de reconhecimento para captura de informações está disponível um botão para abertura da interface.

A Figura 5.2 mostra a interface do módulo de busca em inglês como resultado da seleção desse idioma. Para seleção do idioma, basta clicar no botão com as iniciais do língua: “En” para inglês; e “Pt” para português.

Para exemplificar, foi realizada uma busca da palavra “estratégia” nos títulos dos documentos, e os resultados da busca são apresentados na Figura 5.3.

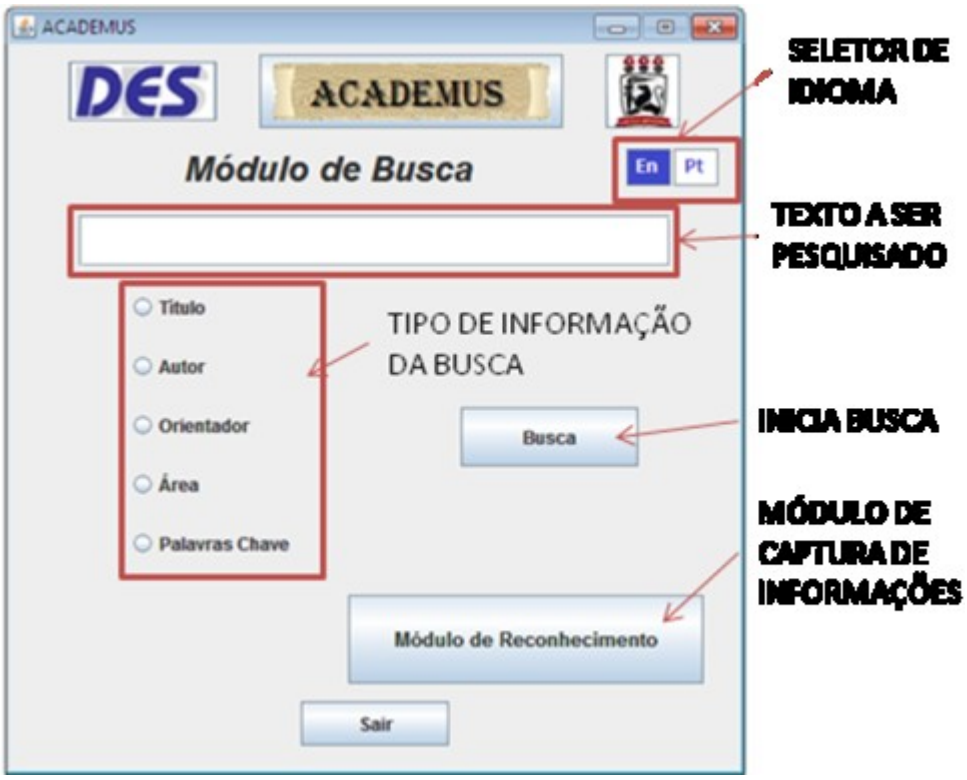


Figura 5.1: Interface inicial do módulo de busca do *ACADEMUS*.

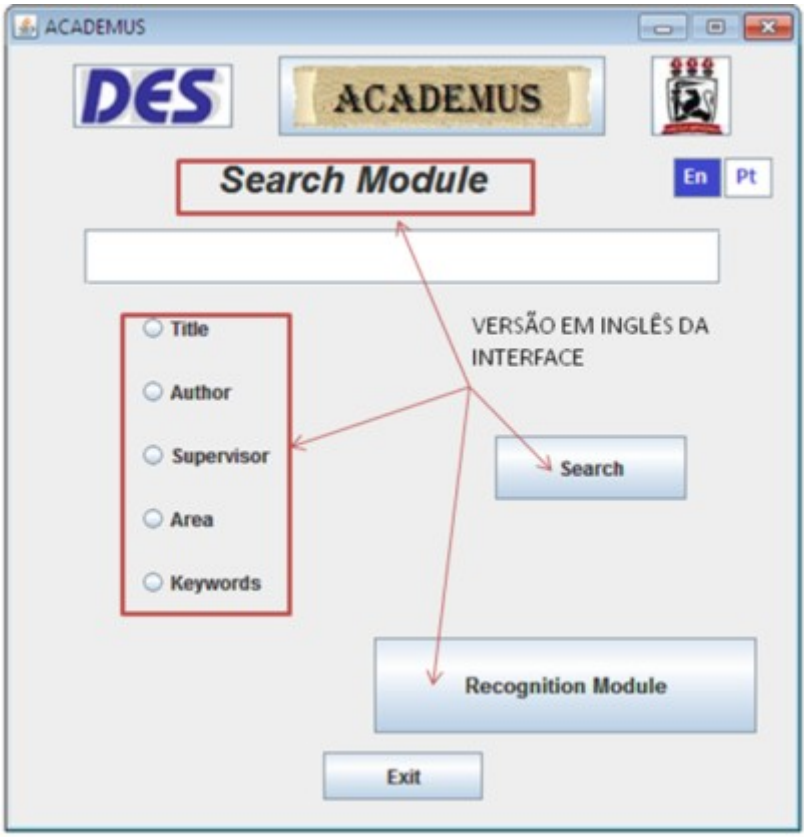


Figura 5.2: Versão em inglês da interface do módulo de busca do *ACADEMUS*.

A Figura 5.3 apresenta na interface os resultados obtidos a partir dos argumentos da busca. A caixa de texto no canto superior esquerdo da figura mostra a quantidade de documentos que contêm os argumentos da busca, logo abaixo são mostrados os títulos dos documentos encontrados. Para obter detalhes de cada documento, deve-se clicar em cima do ícone ao lado de cada título. Logo em seguida, são apresentadas as informações mais detalhadas sobre o documento como o título, autor, orientador, co-orientador, área de concentração e palavras-chave, conforme ilustra a Figura 5.4. Para o exemplo anterior de busca, foram encontrados 4 documentos que têm em seu título o texto “estratégia”.

Na interface da Figura 5.4, são disponibilizados botões para acesso a informações do resumo, Figura 5.5, do *abstract*, Figura 5.6, e da bibliografia, Figura 5.7, do documento encontrado.

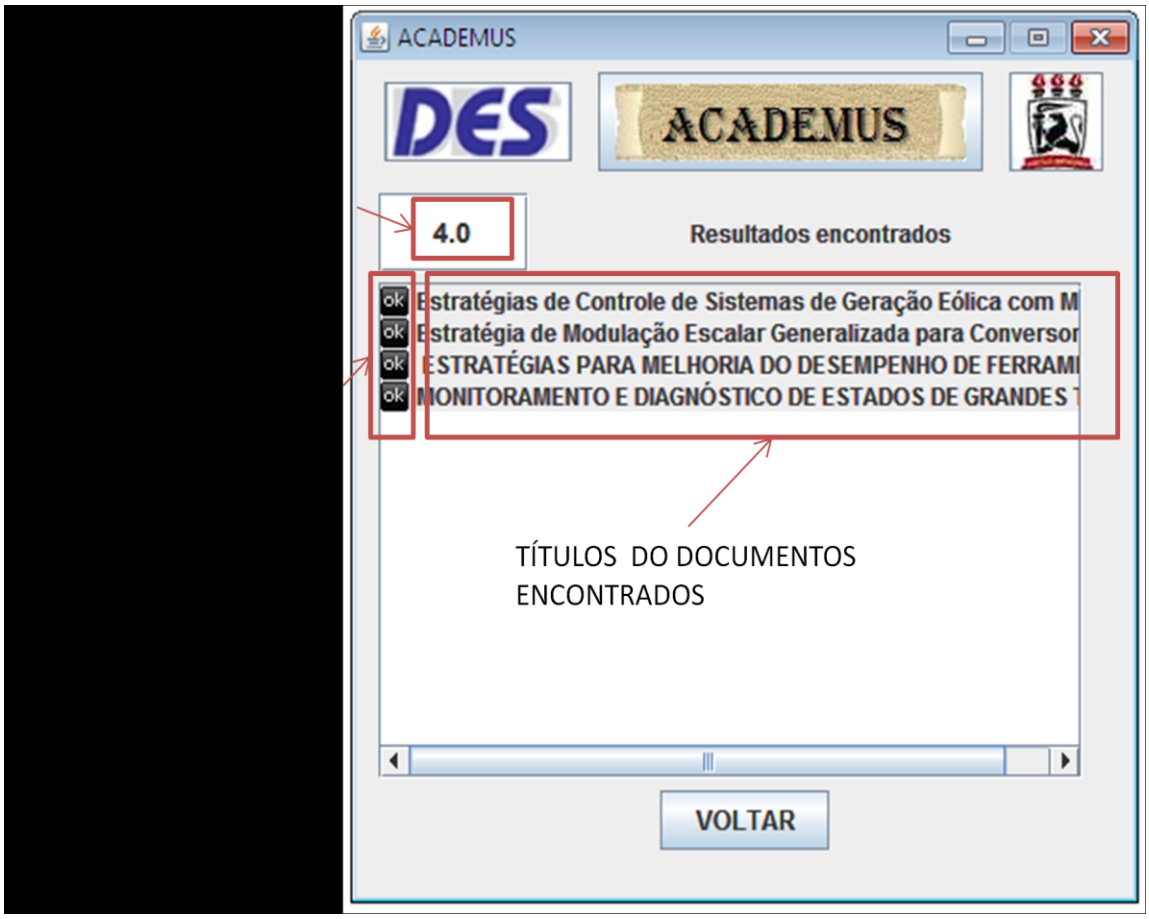


Figura 5.3 Interface de apresentação dos resultados da busca.

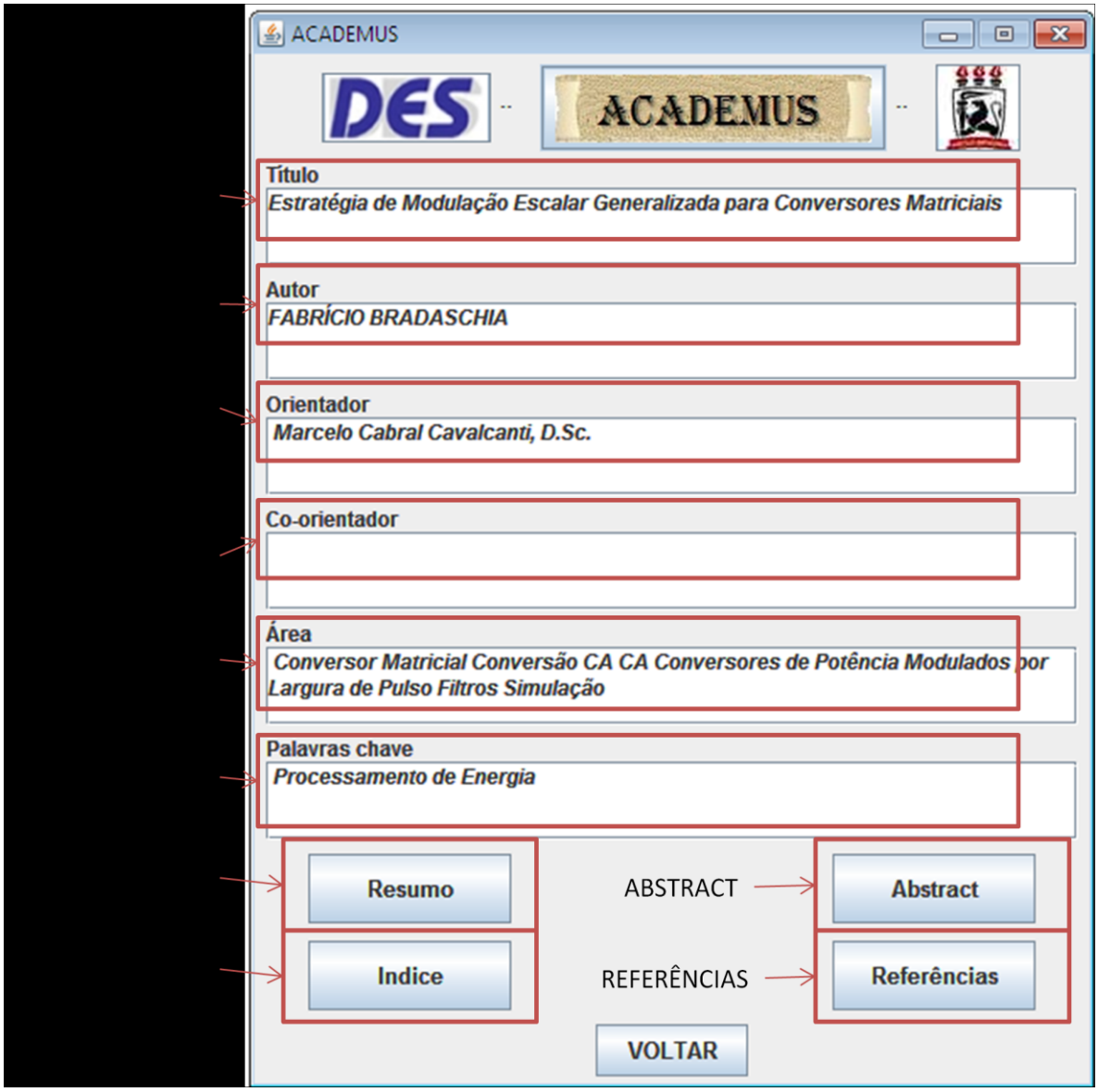


Figura 5.4: Interface de apresentação das informações sobre o documento selecionado.

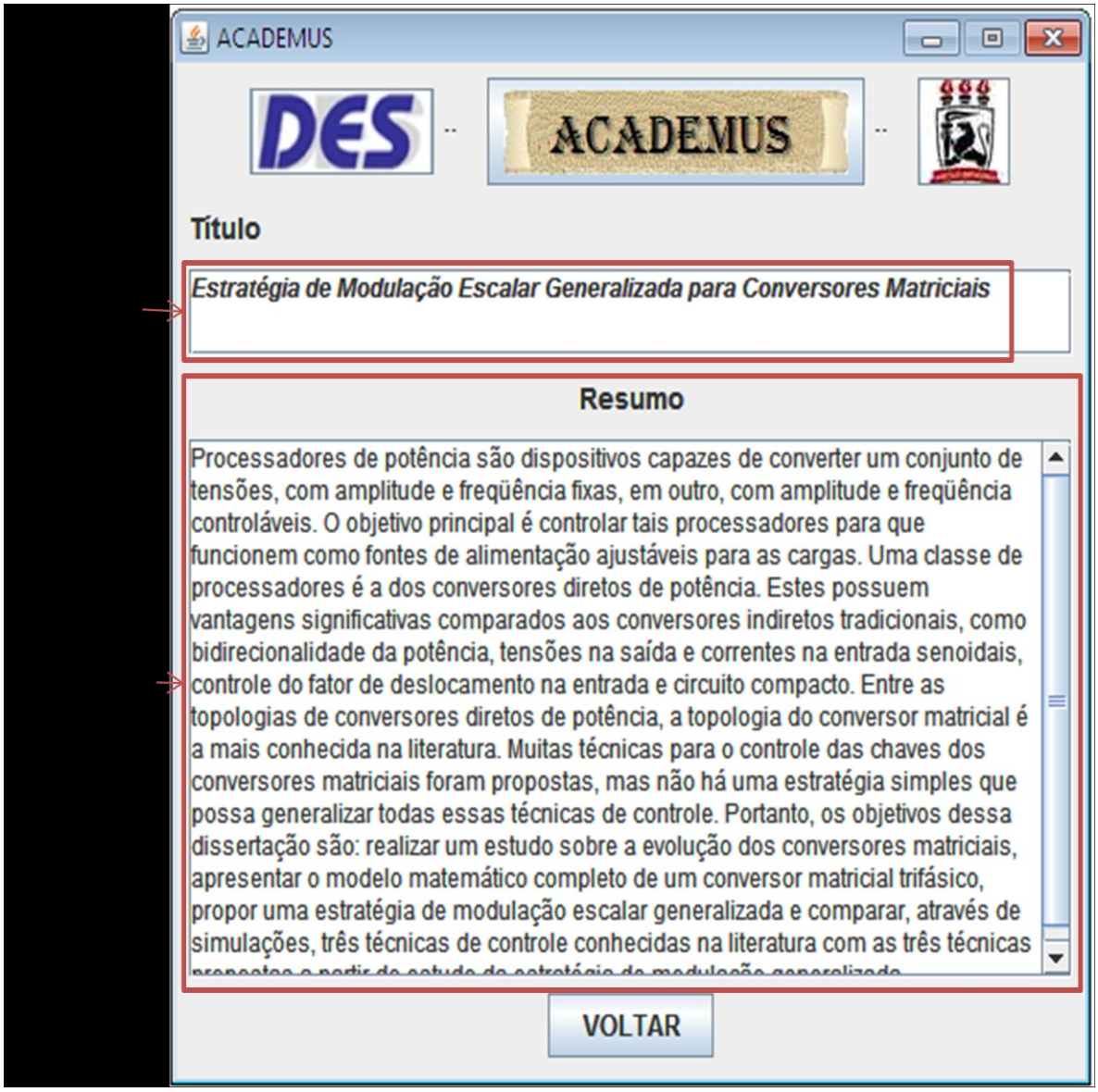


Figura 5.5: Interface de apresentação do resumo do documento selecionado.

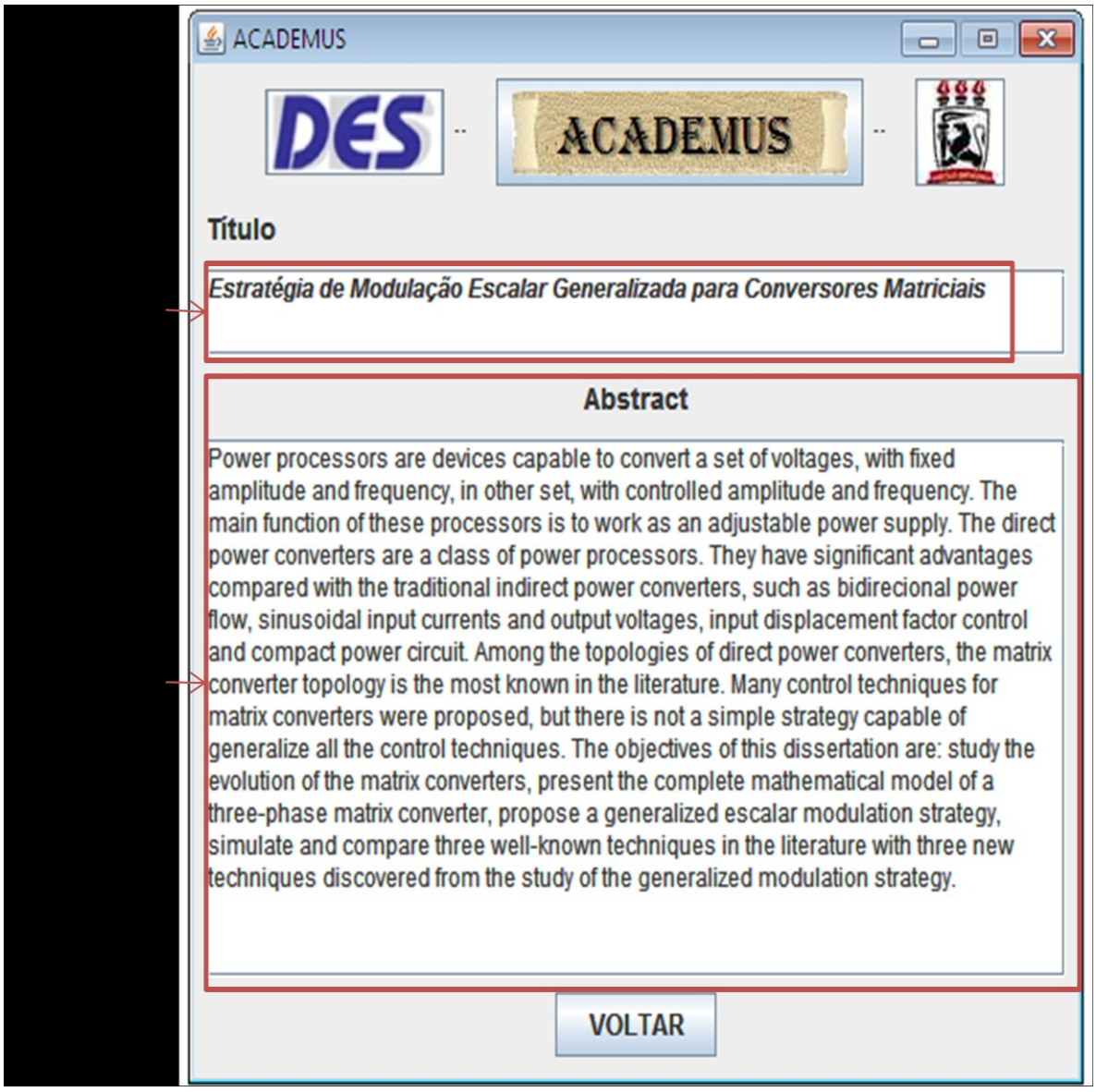


Figura 5.6: Interface de apresentação do *abstract* do documento selecionado.

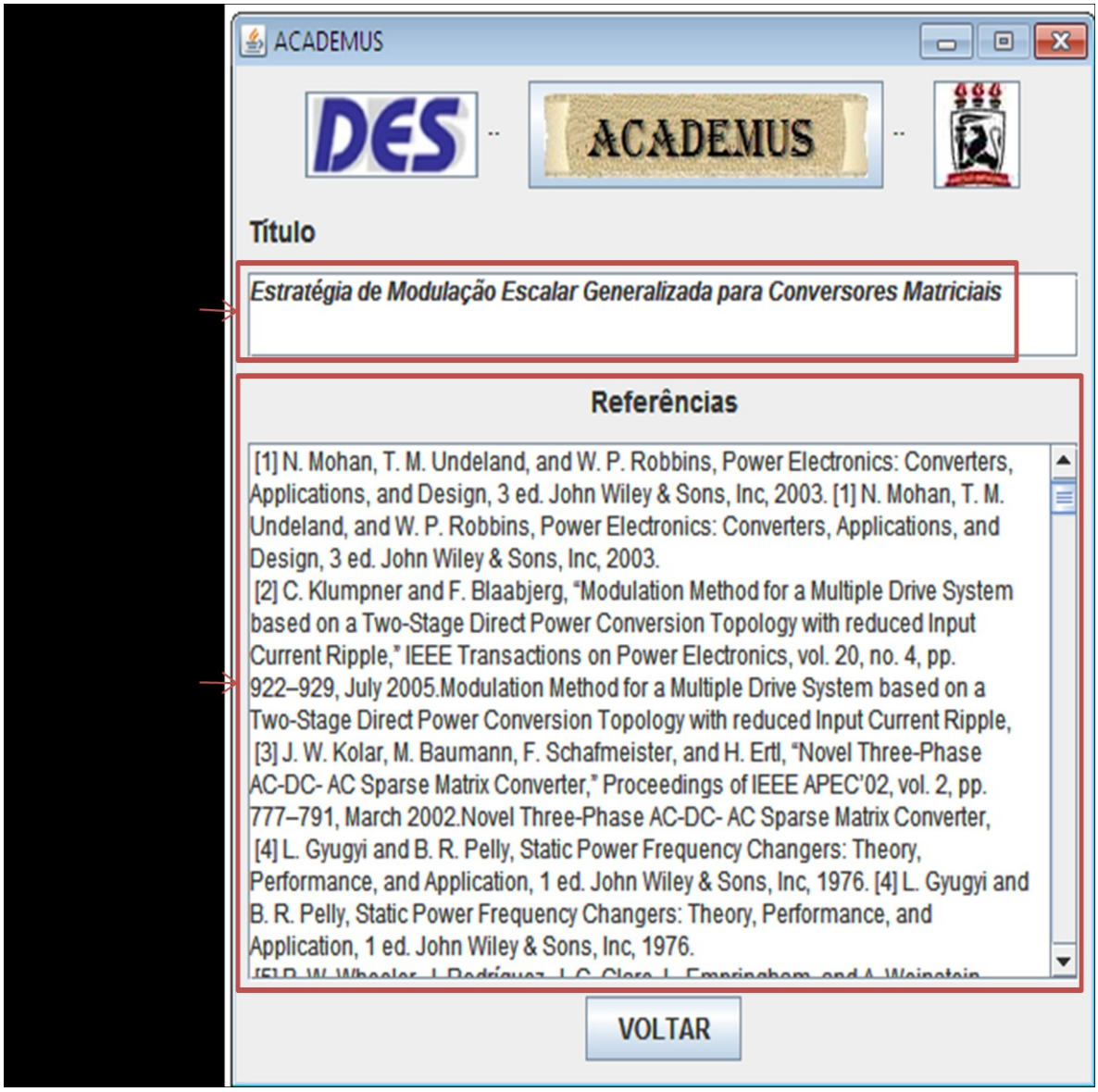


Figura 5.7: Interface de apresentação da bibliografia do documento selecionado.

5.3. MÓDULO DE AQUISIÇÃO

No módulo de aquisição são capturadas as informações sobre o documento, através de rotinas que fazem a análise de seu conteúdo. Quando o documento já se encontra no formato PDF, é possível realizar a captura diretamente. No entanto, quando o documento estiver como imagem é necessário abrir o aplicativo de OCR para extrair o texto e armazená-lo em arquivos no formato PDF. Só então, a versão PDF do documento é submetida a análise de conteúdo.

A Figura 5.8 mostra a interface do módulo de reconhecimento para captura de informações sobre o documento. Inicialmente se deve selecionar o documento a ser analisado, através da opção “abrir”, que apresenta uma interface para seleção do arquivo em PDF, Figura 5.9, dentro dos diretórios do computador. Após a seleção, o *link* do arquivo é mostrado na janela de texto e a opção “Analisar” é habilitada. Na opção “Analisar” inicia-se a análise do documento para captura de informações. A opção “OCR” inicializa a ferramenta de extração de caracteres de um conjunto de imagens.

As informações capturadas são armazenadas automaticamente no banco de dados da plataforma, para posteriormente estarem disponíveis para busca de informações.

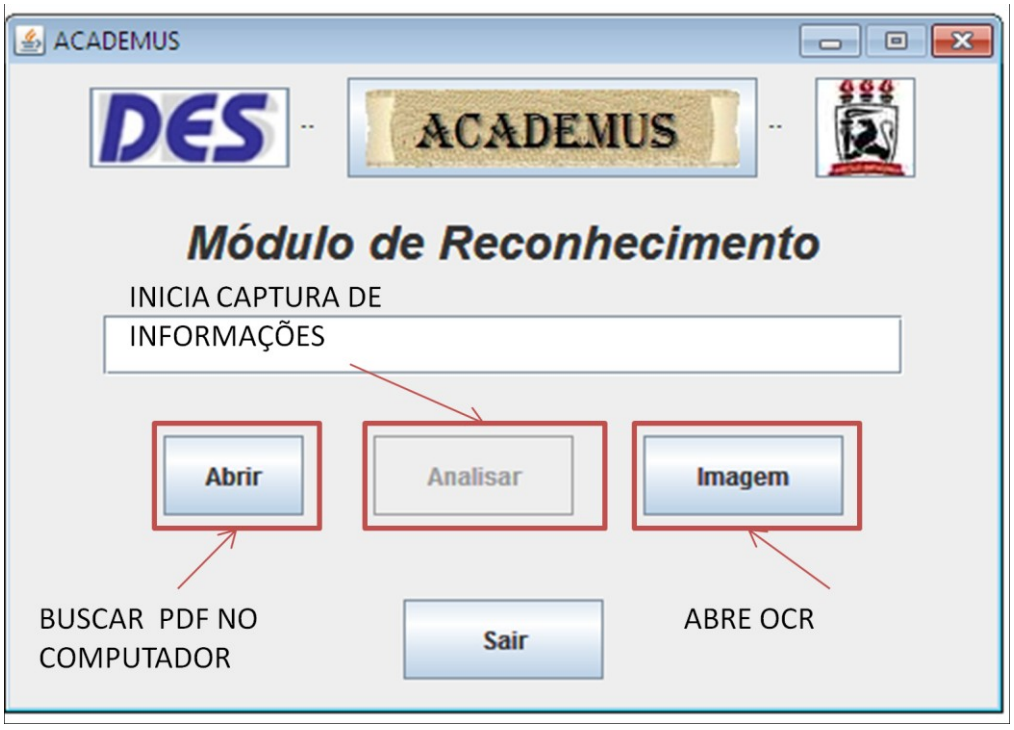


Figura 5.8: Interface do módulo de reconhecimento para captura de informações.

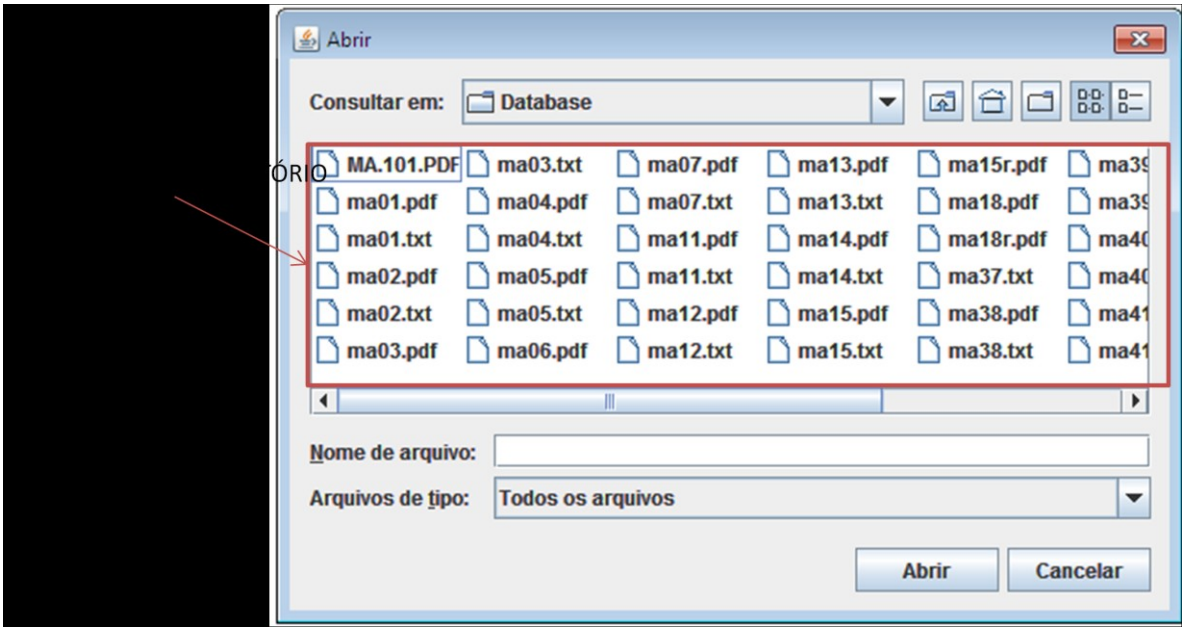


Figura 5.9: Interface para seleção do arquivo pdf a ser analisado pelas técnicas para captura de informações sobre o documento.

O módulo de reconhecimento de conteúdo tem como entrada documentos em formato digital PDF. Um documento nesse formato contém informações de texto, tamanho e tipo de fonte, e posicionamento do texto no layout da página. Os documentos que foram digitalizados tiveram o texto e suas formatações extraídas pelo OCR e em seguida armazenadas como arquivos PDF. Para captura de conteúdo foram elaborados algoritmos para manipulação de arquivos PDF em linguagem de programação Java®, utilizando o PDFBOX.

6. RESULTADOS

6.1. METODOLOGIA

O objetivo principal da plataforma *ACADEMUS* é capturar as informações que identificam um documento, para tanto, são utilizadas técnicas para extração de texto e reconhecimento de conteúdo. Nesse sentido, são feitos testes e análises de desempenho da plataforma e de seus módulos, e sobre a influência isolada de cada módulo, em termos captura exata da informação.

Para os testes de avaliação da plataforma *ACADEMUS* foi usado o banco de teses e dissertações do PPGEE (Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco) como conjunto de testes, com todas as teses e dissertações finalizadas no PPGEE até a presente dissertação. Algumas teses e dissertações defendidas não chegaram a ser finalizadas, pois os autores não cumpriram todas as exigências formuladas pela banca examinadora durante a defesa, não apresentando a versão final do trabalho, e por isso não estão disponíveis na plataforma.

Em relação a esses documentos e aos testes realizados, inicialmente, foi necessário classificar os documentos em dois grupos: grupo “A” relativo aos documentos que não foram digitalizados, pois já possuíam uma versão digital, com 120 unidades; e, grupo B relativo aos documentos digitalizados, com 50 unidades. Essa divisão foi necessária, pois o grupo B foi submetido a diferentes módulos de tratamento de imagem e extração de texto que introduzia erros quando comparados com os documentos originais.

Os resultados para avaliação da plataforma *ACADEMUS* podem ser classificados em 4 grupos: Verdadeiros Positivos (VP), quando **existe** a informação

documento e essa **é capturada**; Verdadeiros Negativos (VN), quando **não existe** a informação no documento e o texto **não é capturado**; Falsos Positivos (**FP**), quando **não existe** a informação no documento e algum texto **é capturado**; Falsos Negativos (**FN**), quando existe a informação documento e essa **não é capturada** (pode ser por ausência de texto ou por texto errado).

A influência do OCR no desempenho da plataforma *ACADEMUS*, foi avaliada através de testes em 10 documentos digitalizados, grupo B. Para tanto, foi utilizada a ferramenta ABBYY FineReader 9.0.7 [17]. Inicialmente foi efetuada a extração de texto nesses documentos sem tratamento e em seguida foi feita a extração de texto nos mesmos documentos só que tratados para remoção de ruído. Ainda sobre o processo de extração de texto, foi feita uma avaliação do desempenho do OCR com o tratamento de um tipo ruído por vez. Os resultados são dados em percentual de acertos, isto é, comparando os textos extraídos com os textos originais do documento que foram transcritos manualmente.

6.2. INFLUÊNCIA DO RUÍDO NO OCR

Com a adição de ruído na imagem o desempenho das técnicas de extração de texto tende a piorar, pois alguns tipos de ruído são tratados como imagens de letras, levando a segmentação das linhas a ficar incorreta, ou as imagens das letras são avariadas levando a erros na extração, conforme Item 3.3.

Os diferentes tipos de ruído, a forma com a qual ele se apresenta e até mesmo o grau ou a intensidade do ruído pode diminuir ou, até em casos extremos anular a correta transcrição do texto. Nos itens 6.2.1 a 6.2.3 são apresentados resultados de extrações de texto para imagens que foram submetidas diferentes tipos de ruído.

6.2.1. INFLUÊNCIA DA BINARIZAÇÃO

No item 3.4.1 desta dissertação foi descrita a importância da binarização para extração de texto e apresentadas algumas técnicas para determinação do limiar e

para binarização. A técnica usada pela plataforma é baseada na determinação de um limiar global através da entropia dos pixels e separação do texto e de elementos gráficos [2].

A Figura 6.1 apresenta em (a) uma imagem em escala cinza com 256 níveis e sua respectiva transcrição, e em (b) apresenta a imagem binarizada e sua respectiva transcrição. Na Tabela 6.1 estão os resultados da captura de informações na imagem não binarizada e binarizada.

Tabela 6.1: Precisão na captura de informações em imagens de documentos em escala de cinza (gray) e binarizada (bin). Testes com 10 documentos.

	Autor		Título		Orientador	
	gray	bin	gray	bin	gray	bin
VP	9	9	9	9	5	5
VN	0	0	0	0	3	3
FP	0	0	0	0	0	0
FN	1	1	1	1	2	2

Observa-se que a precisão na captura da informação mesmo sem a binarização prévia usando a técnica de limiarização via entropia, manteve-se a mesma. Isso ocorre porque o próprio software já dispõe de uma binarização prévia.

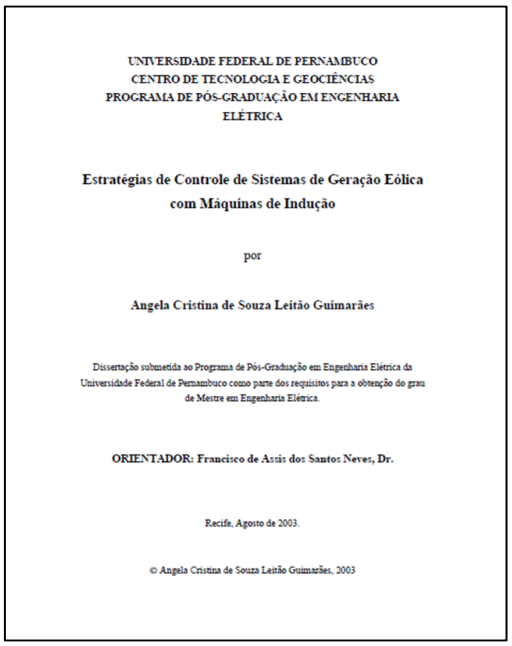
6.2.2. INFLUÊNCIA DO RUÍDO DE BORDA

No item 3.4.2 foram introduzidos os tipos de ruído de borda, que são devido a: espiral, linhas pretas, perspectiva.

6.2.2.1.INLFUÊNCIA DA ESPIRAL DE ENCADERNAÇÃO

A Figura 6.2 apresenta em (a) uma imagem com o ruído de borda tipo efeito espiral e sua respectiva transcrição, e apresenta em (b) a imagem sem o ruído e sua

respectiva transcrição. Na Tabela 6.2 estão os resultados da captura de informações na imagem com o ruído e sem o ruído.



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Estratégias de Controle de Sistemas de Geração Eólica com Máquinas de Indução

por

Angela Cristina de Souza Leitão Guimarães

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco como parte dos requisitos para a obtenção do grau de Mestre em Engenharia Elétrica.

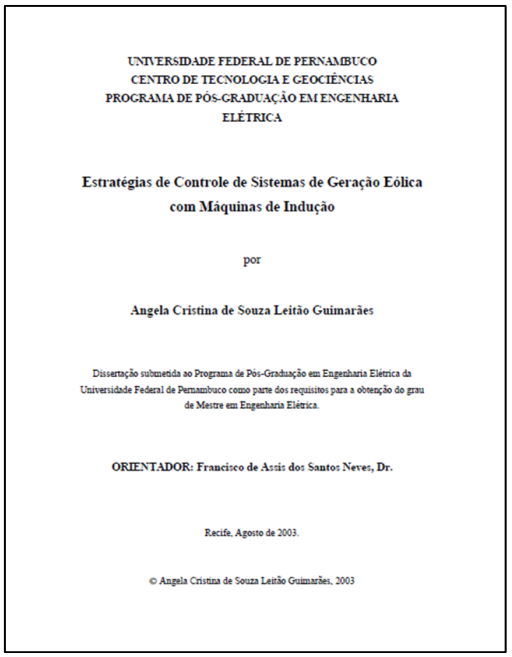
ORIENTADOR: Francisco de Assis dos Santos Neves, Dr.

Recife, Agosto de 2003.

© Angela Cristina de Souza Leitão Guimarães, 2003

Texto transcrito
UNRTRSIDADE FEDERAL DE PERNAMBVC O
CENTRO DE TECNOLOGIA E GEOCTÉNC1AS
PROGRAMA DE PÓS-GRADUAÇÀO EM ENGENHAREI
ELÉTRICA
Estratégias de Controle de Sistemas de Geração Eólica
corn Máquinas de Iuducào

(a)



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Estratégias de Controle de Sistemas de Geração Eólica com Máquinas de Indução

por

Angela Cristina de Souza Leitão Guimarães

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco como parte dos requisitos para a obtenção do grau de Mestre em Engenharia Elétrica.

ORIENTADOR: Francisco de Assis dos Santos Neves, Dr.

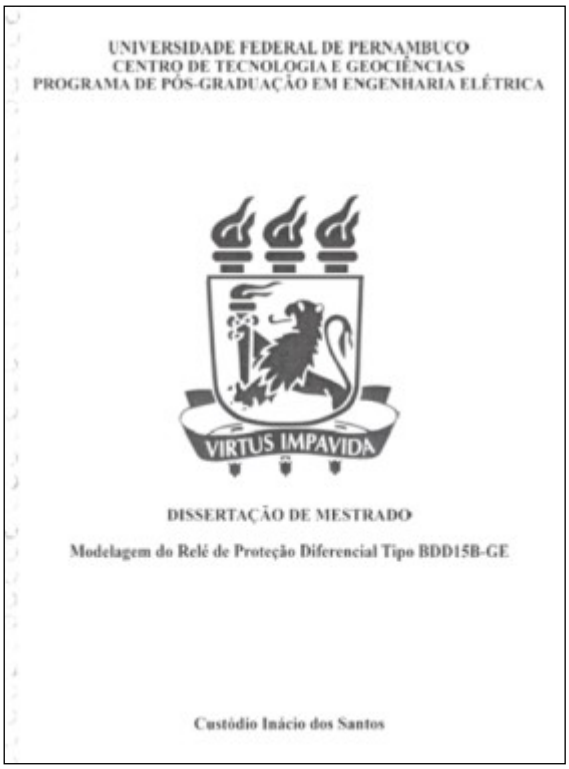
Recife, Agosto de 2003.

© Angela Cristina de Souza Leitão Guimarães, 2003

Texto transcrito
UNIVERSIDADE FEDERAL DE PERNAMBVCO
CENTRO DE TECNOLOGIA E GEOCIÉNCIAS
PROGRAMA DE PÓS-GRADUAÇÀO EM ENGANHARIA
ELÉTRICA
Estratégias de Controle de Sistemas de Geração Eólica
com Máquinas de Indução

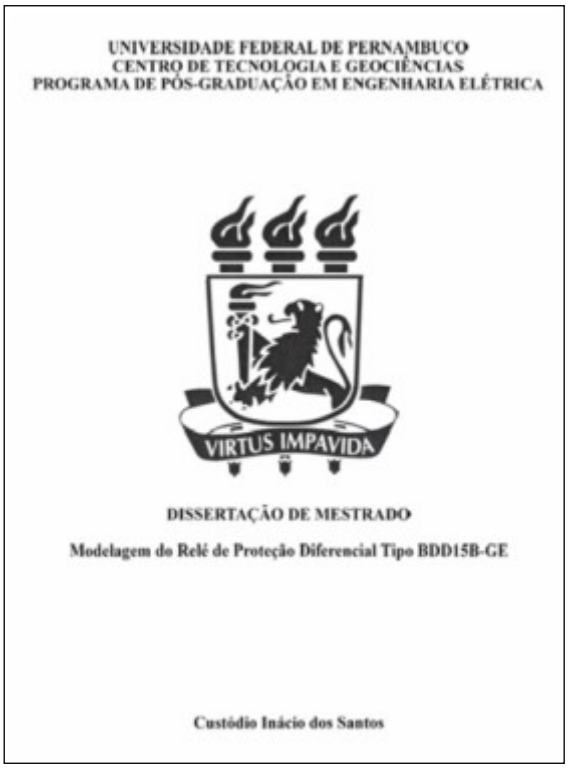
(b)

Figura 6.1: Imagem (a) de documento em escala de cinza com 256 níveis e sua transcrição; (b) binarizada e sua transcrição.



Texto transcrito
1 I UNIVERSIDADE FEDERAL DE PERNAMBUCO CENTRO DE TECNOLOGIA E GEOCIÊNCIAS PROGRAMA DF PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA **• *s*** ***i*** *ÚÚÚ* **•• /** ***i*** Custódio Inácio dos Santos

(a)



Texto transcrito
UNIVERSIDADE FEDERAL DE PERNAMBUCO CENTRO DE TECNOLOGIA E GEOCIÊNCIAS PROGRAMA DF PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA ¥ ¥ ¥ DISSERTAÇÃO DE MESTRADO Modelagem do Relê" de Proteção Diferencial Tipo BDD15B-GE Custódio Inácio dos Santos

(b)

Figura 6.2: Imagem (a) de documento com ruído de borda tipo espiral e sua transcrição; (b) com ruído removido e sua transcrição.

Tabela 6.2: Precisão na captura de informações em imagens de documentos com ruído de borda (c/R) tipo espiral e sem o ruído (s/R). Testes com 10 documentos.

	Autor		Título		Orientador	
	c/R	s/R	c/R	s/R	c/R	s/R
VP	7	9	7	9	4	5
VN	0	0	0	0	3	3
FP	0	0	0	0	0	0
FN	3	1	3	1	3	2

Observa-se uma precisão na captura da informação de 70% para o autor, título, e orientador. Para as imagens sem ruído a precisão foi de 90%, 90% e 70%, respectivamente.

6.2.2.2. INFLUÊNCIA DO EFEITO BORDAS PRETAS

A Figura 6.3 apresenta em (a) uma imagem com o ruído de borda tipo linhas pretas e sua respectiva transcrição, e apresenta em (b) a imagem sem o ruído e sua respectiva transcrição. Na Tabela 6.3 estão os resultados da captura de informações na imagem com o ruído e sem o ruído.

UNIVERSIDADE FEDERAL DE PERNAMBUCO - UFPE
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS – CTG
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA
PPGEE

MONITORAMENTO E DIAGNÓSTICO DE ESTADO DE DISJUNTORES DE ALTA E EXTRA ALTA TENSÃO COMO TÉCNICA DE MANUTENÇÃO PREDITIVA

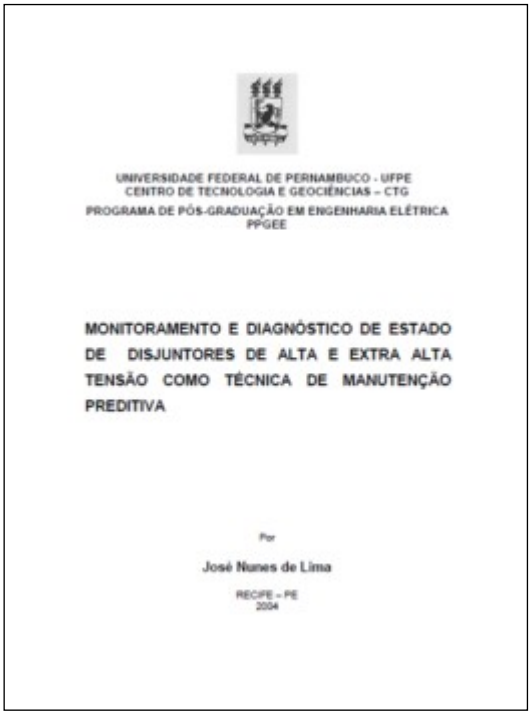
Por

José Nunes de Lima

RECIFE – PE
2004

Texto transcrito
UNIVERSIDADE FEDERAL DE PERNAMBUCO – UFPE CENTRO DE TECNOLOGIA E GEOCIÊNCIAS - CTG PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA PPGEE MONITORAMENTO E DIAGNÓSTICO DE ESTADO DE DISJUNTORES DE ALTA E EXTRA ALTA TENSÃO COMO TÉCNICA DE MANUTENÇÃO PREDITIVA Por José Nunes de Lima RECIFE - PE

(a)



UNIVERSIDADE FEDERAL DE PERNAMBUCO - UFPE
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS – CTG
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA
PPGEE

MONITORAMENTO E DIAGNÓSTICO DE ESTADO DE DISJUNTORES DE ALTA E EXTRA ALTA TENSÃO COMO TÉCNICA DE MANUTENÇÃO PREDITIVA

Por

José Nunes de Lima

RECIFE – PE
2004

Texto transcrito
UNIVERSIDADE FEDERAL DE PERNAMBUCO – UFPE CENTRO DE TECNOLOGIA E GEOCIÊNCIAS - CTG PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA PPGEE MONITORAMENTO E DIAGNÓSTICO DE ESTADO DE DISJUNTORES DE ALTA E EXTRA ALTA TENSÃO COMO TÉCNICA DE MANUTENÇÃO PREDITIVA Por José Nunes de Lima RECIFE - PE

(b)

Figura 6.3: Imagem (a) de documento com ruído de borda tipo linhas pretas e sua transcrição; (b) com ruído removido e sua transcrição.

Tabela 6.3: Precisão na captura de informações em imagens de documentos com ruído de borda (c/R) tipo linha preta e sem o ruído (s/R). Testes com 10 documentos.

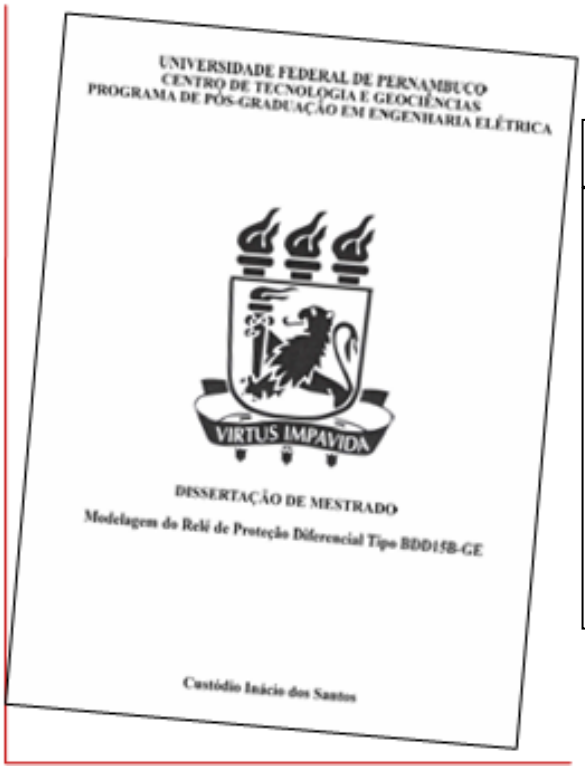
	Autor		Título		Orientador	
	c/R	s/R	c/R	s/R	c/R	s/R
VP	9	9	9	9	5	5
VN	0	0	0	0	3	3
FP	0	0	0	0	0	0
FN	1	1	1	1	2	2

Observa-se uma precisão na captura da informação de 90% para o autor e título, e de 70% para o orientador. Para as imagens sem ruído a precisão foi de 90%, 90% e 70%, respectivamente. Os resultados apontam a não influência do ruído, mas na verdade a segmentação do software “imunizou” o processo contra o ruído.

6.2.2.3. INFLUÊNCIA DA INCLINAÇÃO E ORIENTAÇÃO

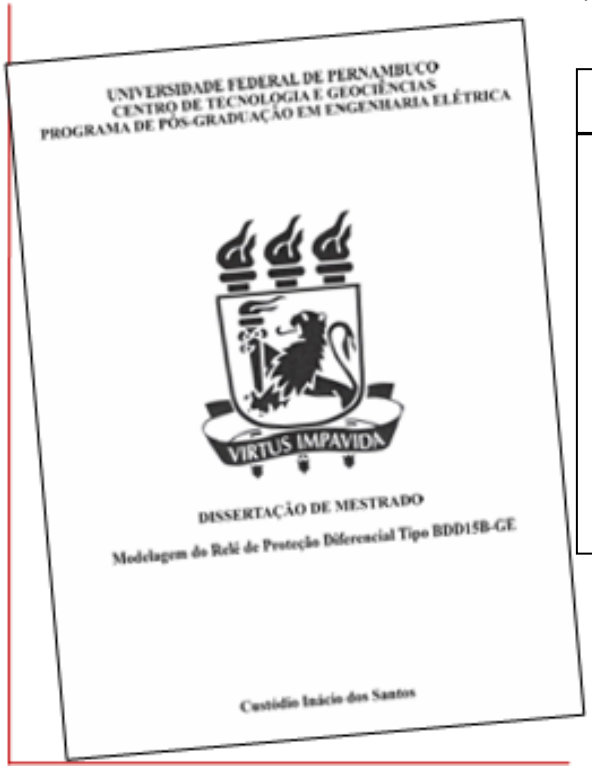
No item 3.4.3 desta dissertação foi apresentada a estratégia para minorar a influência de ruído de inclinação e orientação da página no desempenho do OCR. Comparação do desempenho na extração de texto de alguns ângulos de inclinação já foi tratada no item 3.4.3, sugerindo uma alta dependência para ângulos superiores a 5º.

A Figura 6.4 apresenta em (a) uma imagem com ruído de inclinação de 5º à direita e sua respectiva transcrição, e em (b) apresenta uma imagem com ruído de inclinação de 5º à esquerda e sua respectiva transcrição. A Figura 6.4 apresenta em (a) uma imagem com ruído de inclinação de 5º à esquerda e sua respectiva transcrição, e em (b) apresenta a imagem sem inclinação e sua respectiva transcrição. Na Tabela 6.4 estão os resultados da captura de informações na imagem com diversos ângulos de inclinação.



Texto transcrito
UNIVERSIDADE FEDERAL DE PERNAMBUCO CENTRO DE TECNOLOGIA E CEOCIÊNCIAS PROGRAMA DF PÓS-GRADI AÇÃO FM ENGENHARIA ELÉTRICA * f * DISSERTAÇÃO DE MESTRADO agem do Rclc dc Proteção Diferencial Tipo BDD15R-CE CUS,Ódio In«io dos Santos

(a)



Texto transcrito
I M\ EKS1DADE ILPLKAL DI. PLRNAMBI (O CENTRO DE TECNOLOGIA E GEOCIÊNCIAS PROGRAMA DE PÓS-GRADI AÇÃO EM ENGFNII \ RI \ ELÉTRICA DISSERTAÇÃO DE MESTRADO Diferencia» TipoBDDt?B-GF. Modelagem do Re»e de Proteção Custódio Inácio dos

(b)

Figura 6.4: Imagem (a) de documento com ruído de inclinação de 5º à direita e sua transcrição; (b) de documento com ruído de inclinação de 5º à esquerda e sua transcrição.

Tabela 6.4: Precisão na captura de informações em imagens de documentos com ruído de inclinação. Versões inclinadas nos ângulos 15º no sentido anti-horário (D) a 15º no sentido horário (E), em que “S” denota a captura da informação e “N” denota falha na captura.

Ângulo	Autor	Título	Ângulo	Autor	Título
15 D	N	N	15 E	N	N
14 D	N	N	14 E	N	N
13 D	N	N	13 E	N	N
12 D	N	N	12 E	N	N
11 D	N	N	11 E	N	N
10 D	N	N	10 E	N	N
9 D	N	N	9 E	N	N
8 D	N	N	8 E	N	N
7 D	N	N	7 E	S	N
6 D	N	N	6 E	S	N
5 D	N	N	5 E	S	N
4 D	N	N	4 E	S	N
3 D	S	N	3 E	S	S
2 D	S	S	2 E	S	S
1 D	S	S	1 E	S	S
0	S	S			

O ruído de inclinação é muito prejudicial, pois para pequenas inclinações (3° sentido horário) houve perda de informação. Note-se que a inclinação no sentido anti-horário é mais prejudicial que a inclinação no sentido horário, pois somente para inclinações superiores a 7° se encontra perdas de informação.

Apesar de a transcrição estar incorreta foi possível capturar a informação correspondente ao documento original na maioria dos casos. Isso se deveu às técnicas de segmentação do texto e à binarização feitas pelo software FineReader®.

6.3. RESULTADOS GERAIS.

O objetivo fundamental do módulo de reconhecimento é capturar as informações contidas no documento sobre o próprio, informações que o identificam como: título autor, e demais informações já citadas no item 4.2.

É apresentado o desempenho da plataforma na captura de informações para cada informação sobre o documento, em termos da precisão na captura da informação. A Tabela 6.5 mostra os resultados da captura de informações de 120 documentos do grupo A (documentos que não foram digitalizados).

Tabela 6.5: Precisão na captura de informações em documentos do grupo A.
Testes com 120 documentos.

	Autor	Título	Orien-tador	Co-ori-entador	Palavras Chave	Área	Resumo	Abstract
VP	118	110	116	14	95	88	82	81
VN	0	0	0	88	18	3	2	2
FP	0	0	0	16	1	1	0	0
FN	2	10	4	2	6	28	36	37

Esses resultados indicam um bom desempenho da plataforma, em especial na captura de informações sobre o autor, título, orientador do documento. Para as demais informações o desempenho cai, mas ainda se observa uma regularidade.

Com 98% de precisão, a captura de informações sobre o autor teve o melhor desempenho, isso se deve principalmente à estratégia de uso de um dicionário controlado com os nomes dos autores. Já para a captura do título a precisão foi menor, mas com resultados interessantes (92%), pois não se utilizou dicionários controlados.

A Tabela 6.6 mostra a precisão na captura de informações de documentos do grupo b (documentos digitalizados). Observa-se que a precisão foi bem menor que a precisão em documentos do grupo A, recuando razoavelmente na captura do título e do orientador.

Um pior desempenho na captura de informações se deve à degradação na transcrição do texto, tanto das letras quanto das formatações. É interessante lembrar que as estratégias se baseiam também na formatação do texto.

Tabela 6.6: Precisão na captura de informações em documentos do grupo B.
Testes com 50 documentos.

	Autor	Título	Orien-tador	Co-ori-entador	Palavras Chave	Área	Resumo	Abstract
VP	43	32	30	0	6	4	30	28
VN	0	0	1	0	12	12	1	3
FP	0	0	4	43	3	3	0	0
FN	7	18	15	7	29	31	19	19

Um fenômeno bastante prejudicial ocorre na transcrição em documentos que foram datilografados. O espaço entre letras freqüentemente é maior que em documentos impressos, sendo esse espaço transcrito como um espaço em branco real, como se as letras fizessem parte de palavras diferentes.

A imprecisão da transcrição também é problemática, pois da mesma forma que se inserir “espaço em branco”, inserir outros caracteres, ou trocá-los, provoca um desastre na análise de conteúdo. Por exemplo, caso a transcrição da letra “o” tenha gerado as letras “ci”, assim a palavra “orientador” seria trocada por “cirientadcir”, e a plataforma não encontraria a informação sobre o orientador, informando da não existência, ou cometendo o erro de “Falso Negativo”.

Para melhorar o desempenho, pode-se utilizar dicionários controlados para o título, orientador (co-orientador) e áreas de concentração. Como também, antes do reconhecimento de conteúdo e logo após a extração de texto, pode-se implementar um módulo para verificação da existência da palavra na língua. Por exemplo, o caso anterior em que trocou a letra “o” pela letra “ci”, com o uso de um verificador, poder-se-ia constatar a inexistência da palavra, trocando-a por uma palavra mais semelhante.

7. CONCLUSÕES E TRABALHOS FUTUROS

A plataforma *ACADEMUS* foi desenvolvida para geração semi-automática de bibliotecas digitais, através da captura de informações sobre documentos de teses e dissertações. As estratégias para reconhecimento de conteúdo e captura de informações foram desenvolvidas neste projeto para a plataforma *Academus*, especificamente para trabalhar com teses e dissertações. Uma nova abordagem de análise das características do texto (tamanho e tipo de fonte, e posição do texto no *layout* da página) para captura de informações foi uma importante contribuição deste trabalho. Além disso, este projeto propiciou a criação da biblioteca digital de teses e dissertações já defendidas no PPGEE.

Foram digitalizadas 59 dissertações do PPGEE-UFPE, dentre as quais 50 foram analisadas. Antes da conversão de imagem para PDF, as imagens foram submetidas a rotinas para remoção de ruído, em algoritmos que já tinham sido desenvolvidos na plataforma BigBatch. Posteriormente, o texto transcrito foi armazenado no formato PDF e no formato txt para a análise de conteúdo. A composição dos arquivos em formatos diferentes foi uma nova abordagem que conferiu uma melhora no desempenho geral da plataforma.

Foi verificado um bom desempenho na tarefa de captura de informações de teses e dissertações, principalmente em documentos que já estavam no formato PDF. Bons resultados também foram obtidos nos documentos mais recentes, devido à padronização na apresentação e distribuição das informações.

Para documentos digitalizados (geralmente, os mais antigos) o desempenho foi comprometido, levando a um menor desempenho. Em especial, isso ocorreu devido a erros introduzidos pelas ferramentas de OCR, fundamentalmente por causa dos erros na transcrição do texto e formatação. A menor precisão da transcrição pelo

OCR ocorreu pela existência de ruído que não puderam ser removidos no pré-processamento. Esse ponto se mostrou crítico, pois a falha na transcrição de uma única letra pode comprometer a captura de uma informação.

De forma a melhorar o desempenho da plataforma algumas estratégias podem futuramente ser implementadas, como:

- Aumento do número de expressões regulares e palavras chaves, para ampliar o universo de busca. Isso implica também no aumento de processamento, e, por conseguinte tempo de execução.
- Uso de dicionários controlados para outros tipos de informação além do autor (como o título e orientador).
- Uso de ferramentas para completa eliminação de ruído, implicando também no aumento do processamento.
- Uso de ferramentas de OCR que obtenham maior precisão.
- Implementação de um módulo para verificar se o texto extraído pelo OCR já existe no dicionário da língua do documento.

Em se tratando da captura de informações da bibliografia, a plataforma apresentou um bom desempenho, porém a segmentação de cada citação, ou seja, o reconhecimento de cada citação e a indexação não apresentaram resultados bons, indicando que melhorias precisam ser feitas nas estratégias de segmentação. Além disso, a conexão dos documentos e das referências bibliográficas extraídas da base de dados das editoras é uma área ainda não explorada, sendo objetivo de futuros trabalhos.

Da mesma forma que para a bibliografia, o “texto” do sumário foi capturado, mas a segmentação não apresentou bons resultados, o que prejudica na indexação da informação e da página do documento.

A plataforma *ACADEMUS* aqui descrita é o foco do artigo “Academus-Generating Digital Libraries of M.Sc. and Ph.D. Theses” publicado nos anais do GREC 2011, cuja cópia encontra-se no Anexo 2, desta dissertação.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] LAWRENCE, S.; LEE G. C.; BOLLACKER, K. “Digital libraries and autonomous citation indexing”. IEEE Computer Society, vol. 32, n. 6, pp 67-71, 1999.

[2] Adobe® Supplement to the ISO 32000. Edição 3. Disponível em http://www.images.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/adobe_supplement_iso32000.pdf, visitado em 04/06/2011.

[3] LARKEY, L.S. “A patent search and classification system”. Proceedings of DL 1999, 4th ACM Conference on Digital Libraries, Berkeley, CA, pp. 179–187 1999.

[4] VAN BEUSEKOM,J.; KEYSERS, D.; SHAFAIT, F.; BREUEL, T.M. “Example-Based Logical Labelling of Document Title Page Images”. ICDAR 2007, pp. 919–924. IEEE Press, Los Alamitos 2007.

[5] LINS, R. D.; TORREÃO, G.; SILVA, G. F. P. “Content Recognition and Indexing in the LiveMemory Platform”. GREC 2009. LNCS. p.220 – 230, Springer Verlag, 2010.

[6] LINS, R. D.; ÁVILA, B.T.; FORMIGA, A. A. “BigBatch: An Environment for Processing Monochromatic Documents”. Campilho, A., Kamel, M.S. (eds.) ICIAR 2006. LNCS, vol. 4142, pp. 886–896. Springer, Heidelberg 2006.

[7] LINS, R. D.; ÁVILA, B. T. “BigBatch: A toolbox for monochromatic documents”. ACM International Conference on Document Engineering, 2005, Bristol. ACM Document Engineering 2005. New York: ACM Press, pp.239 – 240, 2005.

[8] Extensible Markup Language (XML) 1.0 (Fifth Edition) W3C Recommendation, 26 November 2008. Disponível em http://www.w3.org/TR/2008/REC-XML-20081126/, visitado em 04/06/2011.

[9] The Ancient Library. Disponível em http://www.ancientlibrary.com/smith-bio/0014.html, acessado em 20/06/2011.

[10] Biblioteca Digital da UNICAMP. Disponível em http://cutter.unicamp.br/, acessado em 01/07/2011.

[11] Biblioteca Digital de Teses e Dissertações da USP. http://www.teses.usp.br/, acessado em 01/07/2011.

[12] GONZALEZ, R. C.; WOODS, R. C. "'Processamento digital de imagens". 3º Ed. Pearson Prentice Hall. São Paulo, 2009.

[13] Techterms.Com "Automatic Document Feeders". Disponível em http://www.techterms.com/definition/adf, acessado em 10/01/2011.

[14] Adobe Developers Association. "Technical Specification of Tiff image Format". Revision 6.0, June 3, 1992 Disponível em http://partners.adobe.com/ public/developer/en/tiff/TIFF6.pdf, acessado em 02/07/2011.

[15] LINS, R. D. "A Taxonomy for Noise in Images of Paper Documents - The Physical Noises". ICIAR '09 Proceedings of the 6th International Conference on Image Analysis and Recognition. Springer-Verlag Berlin, Heidelberg, 2009.

[16] DA SILVA, J. M.; LINS, R. D.; DA ROCHA, V.C. "Binarizing and filtering historical documents with back-to-front interference". ACM SAC'06: Proceedings of the 2006 ACM Symposium on Applied Computing, pp. 853-858, ACM Press, 2006.

[17] ABBY Fine Reader version 9.0.7. Disponível em http://finereader.abbyy.com/, acessado em 04/06/2011.

[18] ÁVILA, B. T.; LINS, R. D. "A New Algorithm for Removing Borders from Monochromatic Documents". ACM-SAC'2004, pp 1219-1225, ACM Press, 2004.

[19] ÁVILA, B. T.; LINS, R. D. "A New Fast Orientation and Skew Detection Algorithm for Monochromatic Document Images". ACM DocEng 2005, ACM Press, 2005.

[20] PDF-Box Home Page. Disponível em http://www.pdfbox.org, acessado em 04/06/2011.

[21] American Standard Code for Information Interchange, ASA X3.4-1963, American Standards Association, June 17, 1963

[22] ÁVILA, B. T; LINS, R. D “Efficient Removal of Noisy Borders from Monochromatic Documents”. International Conference on image Analysis and Recognition, 2004, Porto. Proceedings of ICIAR 2004. Berlin: Springer Verlag, 2004. v.3212. p.249 – 256.

[23] EIKVIL, L. “Tutorial OCR Optical Character Recognition”. Oslo, 1993. Disponível em http://www.nr.no/~eikvil/OCR.pdf, acessado em 10/06/2011

[24] CHEN, Q. “Evaluation of OCR Algorithms for Images with Different Spatial Resolutions and Noises”. Master Thesis, Ottawa-Carleton Institute for Electrical Engineering, Ottawa, 2003.

[25] JAIN, R., KASTURI, R.; SCHUNCK, B. G. “Machine Vision”. McGraw-Hill, 1995.

[26] ZAHN, C. T.; ROSKIES, R. Z. “Fourier descriptors for plane closed curves”. IEEE Transaction on Computer, C-21 (1), pp.269-281, 1972.

[27] PERSOON, E.; FU, K. S. “Shape discrimination using Fourier descriptors”. IEEE Transactions on Systems, Man Cybernetics, Vol.7, No.2, pp.170-179, 1977.

[28] OLIVEIRA, D.M; LINS, R.D. “A New Method for Shading Removal and Binarization of Documents Acquired with Portable Digital Cameras”. International Workshop on Camera-Based Document Analysis and Recognition, 2009, Barcelona. Proceedings of CBDAR 2009. New York: IAPR Press, 2009. pp. 98-105.

[29] NIBLACK, W. "An Introduction to Image Processing". pp. 115-116, Prentice-Hall, 1986.

[30] OTSU, N. "A Threshold Selection Method for Gary_level Histograms”. IEEE Trans. System, Man, and Cybernetics, v. 9, n. 1, p. 62-66. 1979

[31] HU, M. K. "Visual Pattern Recognition by Moment Invariants". IEEE Trans. Info. Theory, v. IT-8, p 179-187.

[32] Operating Instructions of Aficio 1060/1075. Disponível em http://www.ricoh-usa.com/downloads/popup/popup_manuals_drivers_download.aspx?path=/downloads /local/manuals/multifunction_bw/aficio1075.pdf, acessado em 15/05/2011.

[33] BERNSEN, J. "Dynamic thresholding of gray level images". ICPR'86: Proc. Intl. Conf. Patt. Recog. pp. 1251-1255, 1986.

[34] KHASHMAN, A.; SEKEROGLU, B. "A Novel Thresholding Method for Text Separation and Document Enhancement". Proceedings of the 11th Panhellenic Conference on Informatics (PCI 2007), Patras, Greece, pp. 18-20, May 2007.

[35] PALUMBO, P. W.; SWAMINATHA,P.; SRIHARI, S. N. "Document image binarization: Evaluation of algorithms". Proc. SPIE 697, 278-286, 1986.

[36] SAUVOLA, J.; PIETAKSINEN, M. "Adaptive document image binarization". Pattern Recognition. 33, 225-236, 2000.

[37] WHITE, J. M.; ROHRER, G. D. "Image thresholding for optical character recognition and other applications requiring character image extraction". IBM J. Res. Dev. 27(4), 400-411, 1983.

[38] O'GORMAN, L. “Experimental comparisons of binarization and multithresholding Methods on document images”. Proceedings of the IAPR International Conference on Pattern Recognition, vol. 2, IEEE, pp. 395-398, 1994.

[39] BAKER, S.; KANADE, T. “Limits on super-resolution and how to break them”. IEEE Trans. Pattern Anal. Mach. Intell., 24(9), pp. 1167-1183, 2002.

[40] DRIRA, F. “Towards restoring historic documents degraded over time”. DIAL '06, pages 350.357. IEEE Computer Society, 2006.

[41] LINS, R.D.; GUIMARÃES NETO, M.S.; FRANÇA NETO, L.R.; ROSA, L.G. “An Environment for Processing Images of Historical Documents. Microprocessing & Microprogramming”. pp. 111-121, North-Holland, January 1995.

[42] LINS, R.D.; MACHADO D.S.A. “A Comparative Study of File Formats for Image Storage and Transmission”. vol 13(1), pp 175-183, 2004, Journal of Electronic Imaging, Jan/2004.

[43] MELLO, C.A.B.; LINS, R.D. “Image Segmentation of Historical Documents”. Visual 2000, Aug. 2000, Mexico.

[44] MARON, M. “Automatic indexing: an experimental inquiry”. Journal of the ACM (JACM) 8(3), 404–417 1961.

[45] HAYKIN, Simon. Redes Neurais: Princípios e Práticas. Editora: Bookman, 2007.

[46] LINS, R. D. “Projeto Nabuco: Processamento de Imagens de Documentos Históricos”. PROC. OF PANEL'95, XXI Conferência Latino Americana de Informática, 1995. p.111-122.

[47] “Google Books”. Disponível em http://books.google.com/intl/pt-BR/googlebooks/about.html, acessado em 17/08/2011.

ANEXO 1 – SOFTWARE DESENVOLVIDO

A execução das rotinas e o software desenvolvido para a plataforma estão no DVD em anexo ao presente documento. As figuras a seguir ilustram o fluxograma funcional dos vários componentes de software desenvolvidos, e uma breve explanação sobre cada componente é fornecida.

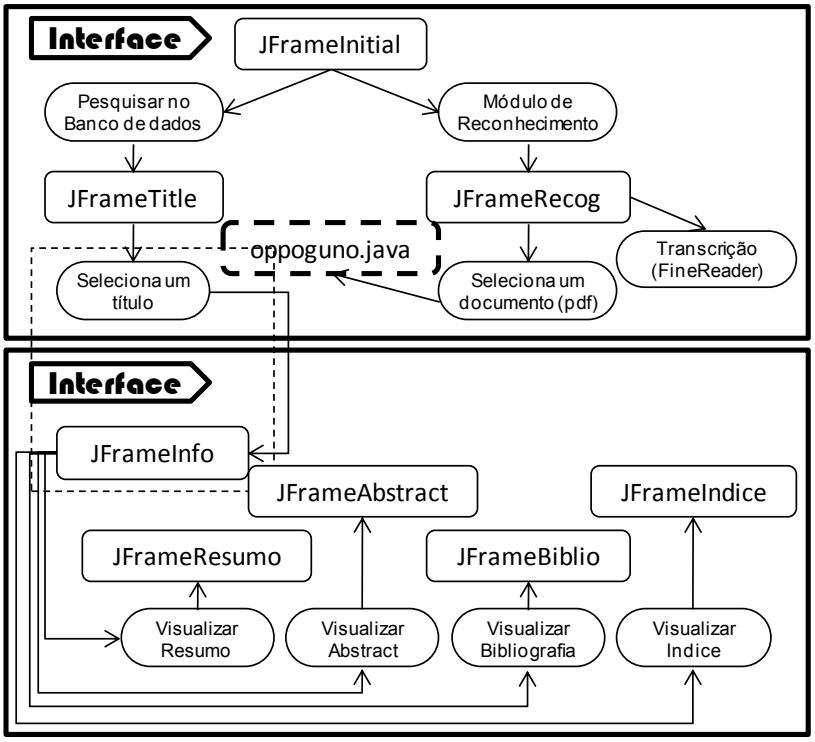


Figura 0.1: Fluxograma de operação da interface da plataforma *ACADEMUS*.

Interface: JFrameInitial

****

Tela inicial da plataforma, em que se escolhe o idioma, pode-se fazer buscas e se tem acesso ao módulo de aquisição.

****

Interface: JFrameRecog

****

Tela do módulo de aquisição, em que se insere um documento para a captura de informações, e acessa o software para transcrição das imagens.

****

Interface: JFrameInfo

****

Tela de apresentação das principais informações do documento selecionado (título, autor, orientador, palavras-chave,área de concentração.).

****

Interface: JFrameTitle

****

Tela de apresentação dos documentos (títulos) em resposta à busca na tela inicial.

****

Interface: JFrameResumo

****

Tela de apresentação do resumo do documento selecionado.

****

Interface: JFrameAbstract

****

Tela de apresentação do abstract do documento selecionado.

****

Interface: JFrameBiblio

****

Tela de apresentação da bibliografia do documento selecionado.

****

Interface: JFrameIndice

****

Tela de apresentação do Índice do documento selecionado.

****

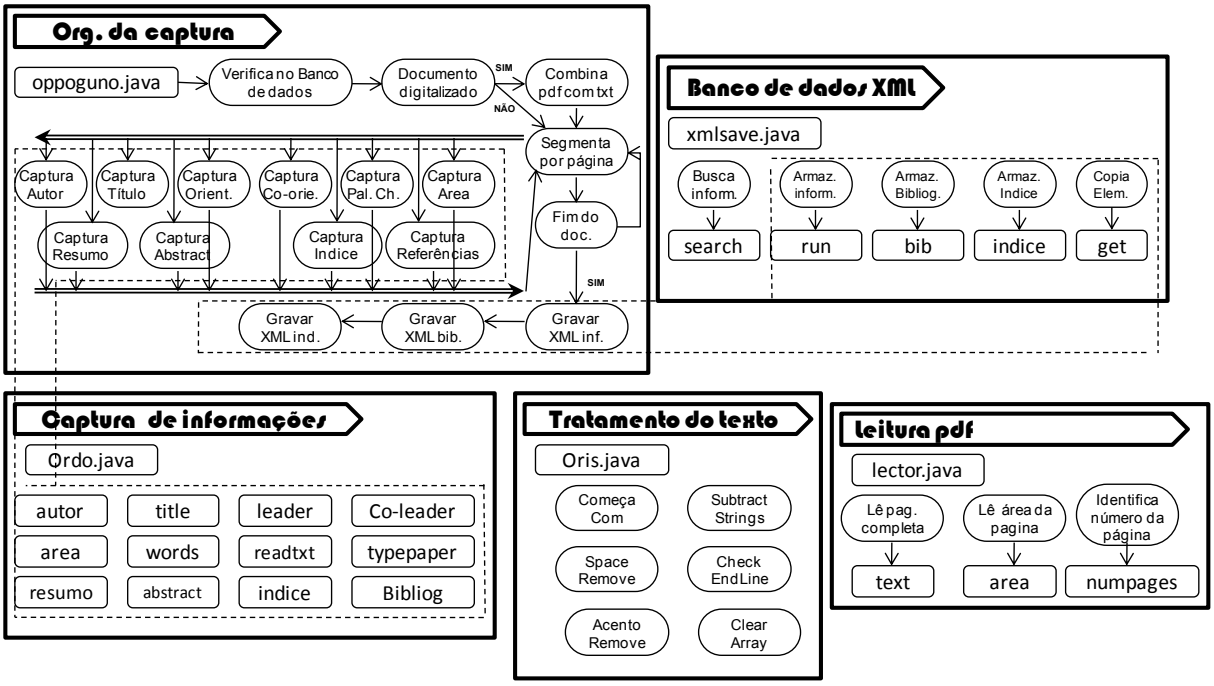


Figura 0.2: Fluxograma de operação da análise e captura de informações da plataforma ***ACADEMUS***.

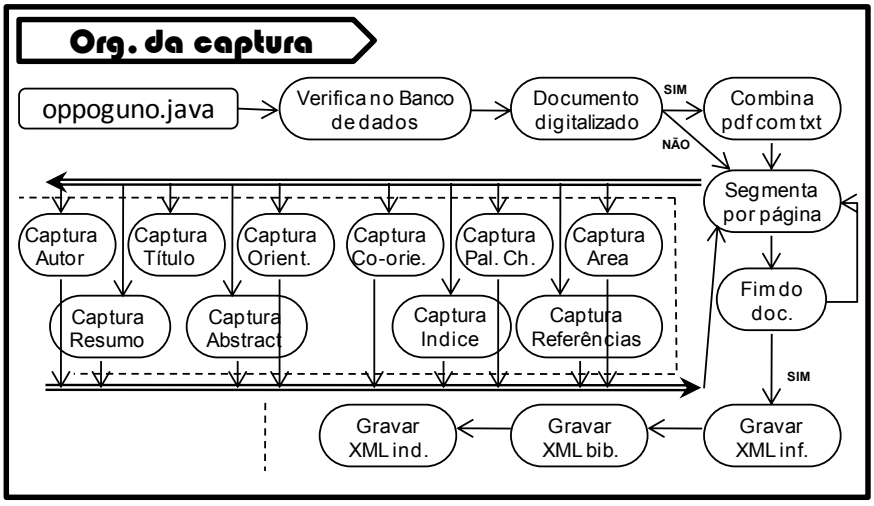


Figura 0.3: Fluxograma de operação do componente de organização da captura de informações.

Organização da Captura: oppoguno.java

****

Rotina para organização da busca, reconhecimento e armazenamento das informações. Verifica se o documento foi digitalizado, faz a composição de arquivos txt e PDF, habilita a leitura e reconhecimento de uma página por vez, e chama rotina de análise do conteúdo.

****

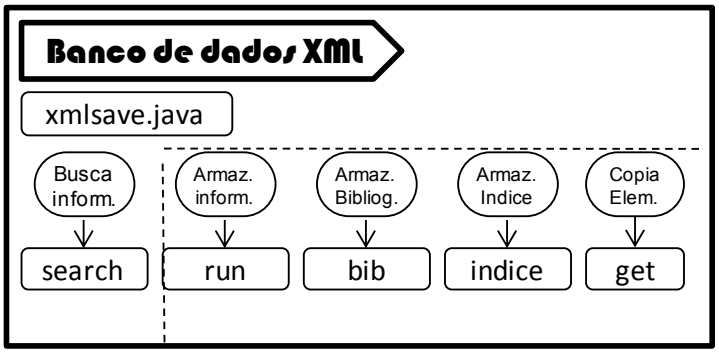


Figura 0.4: Fluxograma de operação do componente de manipulação do banco de dados.

Banco de dados: XMLsave.java

****

Rotina para leitura e gravação no banco de dados. Pode-se pesquisar no banco de dados (search), verificar se o arquivo já existe (get), armazenar a informações capturadas (run) (bib) (indice).

****

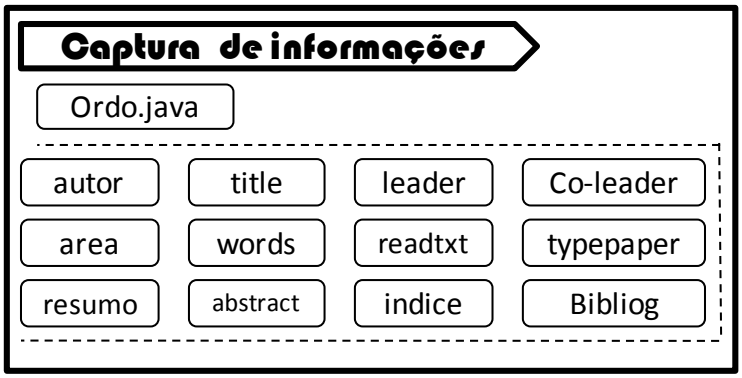


Figura 0.5: Fluxograma de operação do componente de captura de informações.

Captura de informações: ordo.java

****

Rotina para captura de todas as informações sobre o documento: autor (autor), título (title), leader (Orientador), co-leader (Co-orientador), área de concentração (area), palavras-chave (words), resumo (resumo), abstract (abstract), bibliografia (Bibliog), índice (indice), tipo de documento (typepaper). Também apresenta um componente para leitura de dicionários controlados (readtxt).

****

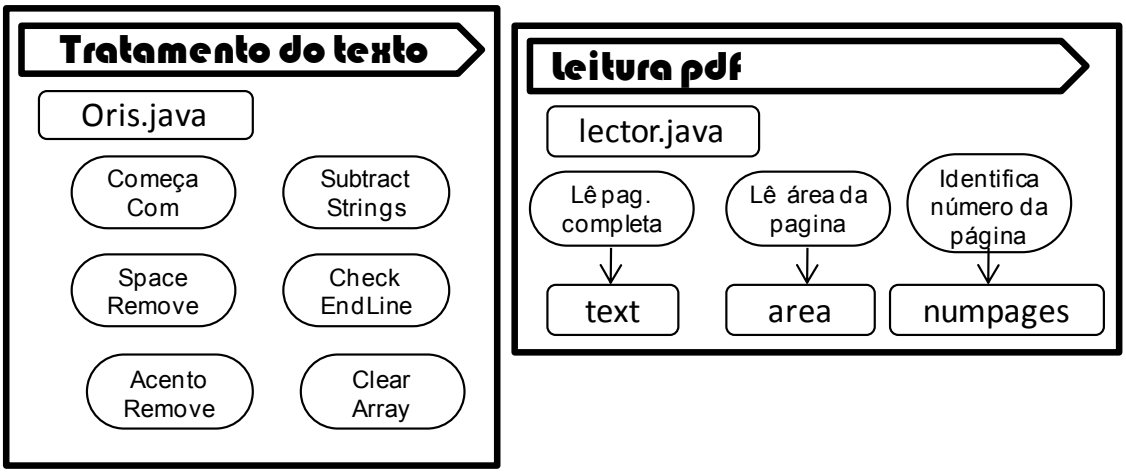


Figura 0.6: Fluxograma de operação dos componentes de tratamento do texto e leitura do pdf, respectivamente.

Tratamento de texto: oris.java

****

Rotina para normalização do texto, como remoção de excesso de espaços em branco, remoção de acentos, segmentação de linha, comparação de strings.

****

Leitura PDF: lector.java

****

Rotina para manipulação do documento PDF, podendo realizar uma leitura completa da página, leitura de determinadas regiões da página, e identificação da página em que se encontra uma informação.

****

ANEXO 2 – ARTIGOS

Rafael Dueire Lins, Gabriel Pereira e Silva, e Paulo Hugo Espírito Santo.

Academus - Generating Digital Libraries of M.Sc. and Ph.D. Theses.

Proceeding of GRREC'2011 – International Workshop on Graphics Recognition. IAPR-Press, September 2011.

*Academus**

Generating Digital Libraries of M.Sc. and Ph.D. Theses

Rafael Dueire Lins
DES – CTG – UFPE
Recife, PE, BRAZIL
rdl@ufpe.br

Paulo Hugo Espírito Santo
DES – CTG – UFPE
Recife, PE, BRAZIL
paulohugos@gmail.com

Gabriel Pereira e Silva
DES – CTG – UFPE
Recife, PE, BRAZIL
gfps@cin.ufpe.br

Abstract **— Postgraduate degrees are one of the most important propellers of all areas of science. M.Sc. and Ph.D. theses witness important developments and provide a solid and global account of research projects. The *Academus* platform is an environment developed with the aim of generating digital libraries of theses and dissertations, making explicit their relevant indexing information.**

Keywords **—** ***Digital libraries, content extraction, M.Sc. thesis, Ph.D. thesis.***.

I. Introduction

The advent of the Internet made digital libraries widely available and to cover all areas of knowledge. Initiatives such as Google Books have as aim to generate a digital library with all published books ever, covering all branches of knowledge and regardless of language and geographical barriers. In general, the first substantial body of research one produces is associated with a postgraduate degree and documented in a *thesis*. Very seldom, M.Sc. or Ph.D. theses become a published book. In general, the access to them remains restricted to the university libraries. Such physical restriction can be overcome by making theses available in digital libraries, but to do so it is not straightforward. Making the volume available in a digital library is only really so if some basic information, such as title, author and supervisor names, keywords, abstract, are made explicit.

Most theses and dissertations are available only in their printed version in the library of the university the degree was read.

This paper presents the *Academus* platform, an environment developed to generate digital libraries of theses. The platform is able to semi-automatically collect information from theses that are either scanned or already in digital format (pdf).

II. Features of Documents

The *Academus* platform automatically generates a database of information extracted from the Ph.D. thesis and M.Sc. dissertations, performing searches by keywords.

(*) *Academus* site sacred to Athena, the goddess of wisdom in Greek mythology. Place where of Plato taught, outside the walls of Athens.
Source: http://en.wikipedia.org/wiki/Akademeia

The documents for which only printed copies were available were digitized in gray scale (256 levels), 200 d.p.i. resolution, using a Ricoh Aficio scanner model 1075, and stored in uncompressed tiff® format. Each volume (Ph.D or M.Sc. thesis) is stored in a single directory.

Binary files claim for much less space for storage and transmission time through computer networks. Often printed theses have photos or figures. The direct binarization of pages with such graphical elements yields in the loss of their content. *Academus* used the binarization module of the LiveMemory platform [2]. Pages with photos or figures are "disassembled" into text and graphical elements. The text part of the page is binarized, while the graphical element, which remains in gray scale, is later copied to the page. Although such page is kept in gray scale, its size is smaller than the original one and the reader has no visual discontinuity as the text part is binarized.

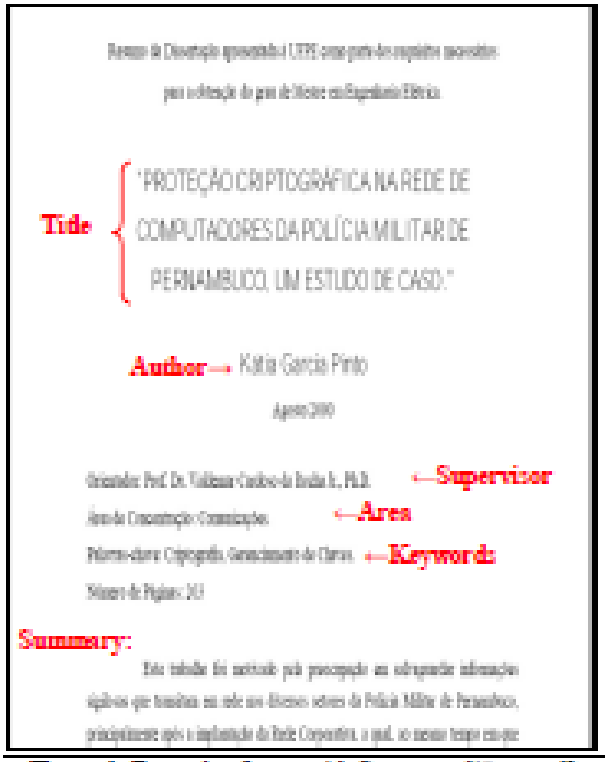


Figure 1. Example of page with Summary ("*Resumo*") of a scanned M.Sc. dissertation with its search elements.

The binary image files are processed through BigBatch [2], which performs noisy border removal, skew and orientation detection and correction, and salt-and-pepper noise filtering. Some of the dissertations were bound using spiral or coil binding. In such case the volume was unbound and the pages were scanned. The marginal holes used for binding need to be removed from the image prior to OCR processing, otherwise the holes are either interpreted as the letter "o" and often cause lay-out identification problems. Figure 1 presents an example of a page of a scanned M.Sc. dissertation after all the image processing performed, with the areas of information highlighted.

The most recent documents are already in Adobe PDF® [1] format. Some pdf files allow extracting textual information (format called *pdf-text*), while some others (known as *pdf-image*) do not. The pdf-text files also store color, font type and size, as well as lay-out information of the page-plan [1]. In *Assistmem*, the pdf-image files go through OCR processing for transcription. The commercial software ABBY FineReader version 9.0.7 [2] is used as it provides the possibility of generating a pdf-text file from an image file, which tries to keep the lay-out of the original page. Besides the pdf-text file another file is generated for each page image with only textual information in it. The two files are cross checked during information extraction. Figure 2 presents the main functionality of the modules of the *Assistmem* platform.

The complete OCR transcription of the documents which are originally in printed format only is a processing intensive and error prone task. The aim of *Assistmem* is to get enough reliable information of the main features of each document and to generate a good quality as small as possible pdf-image file.

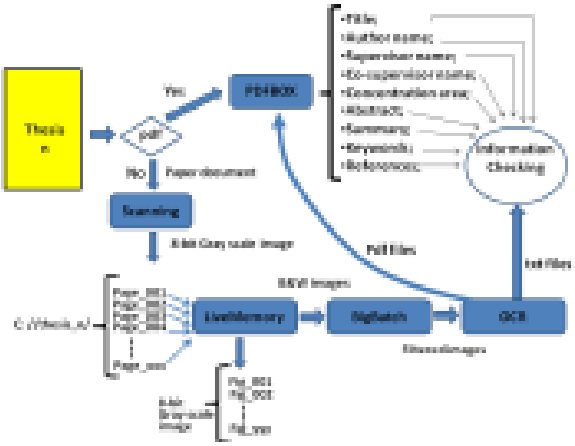


Figure 2. Functional scheme of the *Assistmem* platform

III. Content Recognition.

The content recognition module takes as input the pdf-text version of all documents, either the generated by the OCR or a pdf original document. This module automatically scans the document and searches the information keys to automatically store them in the database, such as:

- Title
- Author
- Supervisor Name
- Co-supervisor name (if any)
- Concentration area
- Keywords
- Summary (in Portuguese)
- Abstract (in English).

The *Assistmem* platform offers the user an interface which allows the correction of the information extracted. The degree of correct automatic information extraction is presented in the section "Results and Analysis", below.

The collected information is stored in the database targeting at two criteria: rapid information retrieval and writing standardization to allow other tools to read the data. The XML [6] format was chosen as it meets such requirements. Each line of the pdf document is stored as a vector with its properties and has four components: the text, the font size, the font type, and the position of the text. To improve the performance of the recognition module at present the text information in the pdf-text file is replaced by the transcription in the txt file generated by the OCR. This replacement may be further improved by some sort of comparisons of the two text information with the help of dictionaries built for the application with the feedback of the entries made by the user in the correction module.

The lack of a strict content and lay-out pattern overall in the older documents (which needed to be scanned) makes hard the correct information extraction. In the case of the papers and scientific articles in proceedings as handled by the LiveMemory platform [5] the layout is better defined and even for scanned documents the information rests in specific areas. For instance, the title or an article is always followed by the authors' names and that information appears on top of the first page of the article. This allows area segmentation to work satisfactorily in that case. For theses the features used were based for text segmentation were based on line segmentation, font size and type, the position of the text in the page layout, "reserved" words always found in the documents, and by a dictionary of supervisors.

Besides the information keys listed above, *Assistmem* play special attention to the bibliography of the thesis. Thus, the list of references is individually transcribed with the aim of cross referencing within the database, as often a thesis makes references to the preceding ones from the same group, as well as building hyperlinks to web documents available, to make easier its use.

The lack of a standard in referencing is also a problem in the implementation of this feature. Some documents adopt the pattern "[number]" as used in this paper, others "[text]", "[text, number]", "(number)", "(text)", and "(text, number)". The bibliography (or list of References) is processed and one of the reference patterns listed above is

detected. The references of the thesis are inserted in the *Analisador* database and become an attribute of the original document. The document text is scanned looking for references which when found are replaced by a link to its entry in the Bibliography.
In a similar fashion the "Table of Contents" of the document is processed to allow the reader to more easily navigate in the document, jumping directly into chapters and sections.

IV. The Search Module

Analisador allows the user to make searches through a user-friendly interface. Several queries are possible: searching a thesis or dissertation by title, author name, supervisor name, concentration area, and keywords. The search module presents a list of all entries that meet the query in full details. The reader may read the Summary, Abstract, Table of Contents and Bibliography in the search module or request to open the corresponding pdf file with the whole document.
Figure 3 (top) exemplifies the use of the data-search module, in which the user asks to the database for the documents with "Rafael" as "Supervisor". The result is exhibited in middle part of the same figure, which lists 19 documents. Clicking on the document number one obtains its details as shown in the bottom part of Figure 3.

V. Results and Analysis

The set of 200 Ph.D theses and M.Sc. dissertations from the Post-graduate Program in Electrical Engineering (PPGEE) (PPGEE) of Universidade Federal de Pernambuco was used as test set in the development of the *Analisador* platform. These documents were split into two groups: (A) Documents in pdf-text; (B) Scanned documents.
One of the aims of the platform is to automatically find and store the search keys and information for each document, such as: title, level, author, supervisors' names, area, keywords, summary (in Portuguese), and abstract (in English).

Table 1 presents the results obtained for the set of 120 documents in Group A (pdf-text). TP stands for "True-Positive" when the correct information is found. TN (True-Negative) is when the recognition module finds information that does not exist. "False Negative" (FN) occurs when the platform does not find existing information in the document.

Table 1. Performance of the Recognition Module. Group (A)

	Author	Title	Supervisor	Co-supervisor	Key words	Area	Summary	Abstract
TP	118	110	116	14	95	88	82	81
TN	0	0	0	88	18	3	2	2
FP	0	0	0	16	1	1	0	0
FN	2	10	4	2	6	28	36	37

The results shown in Table 1, which was obtained with the Pdf-Box script is acceptable, overall for the Author, Title and Supervisor Keys. The performance is lower for the other entries.
Table 2 presents the performance of the platform in the recognition of 50 documents which were scanned and processed as detailed above.

Table 2. Performance of the Recognition Module. Group (B)

	Author name	Title	Supervisor names	Co-supervisor	Key words	Area	Summary	Abstract
VP	43	32	30	0	6	4	30	28
VN	0	0	1	0	12	12	1	3
FP	0	0	4	43	3	3	0	0
FN	7	18	15	7	29	31	19	19

As one would expect, the results obtained are less good than the one for Group A documents, due to the OCR process. One of the problems most often found in the transcription was the insertion of blank space within words. The documents in this group were often typed, most of them with mechanical typewriters, only a few were produced using electric typewriter, and even fewer of them were printed. There is a lot of room for further improvement in this case. For instance, the strategy to find the "Summary" of the document ("Resumo" in Portuguese) is to look for the word in the transcribed version of the front page. But sometimes that word appears in the text itself and in this happens the search tool gets lost. In some other cases the incorrect transcription of the search key causes it to lose track. The inclusion of a dictionary and a word correction strategy may make the process more efficient.
The *Analisador* platform must be seen as a semi-automatic tool which helps the user in collecting and correcting such information, it is certainly a valuable platform. After the data/information key extraction process the data is displayed in a data-correction interface that the user may easily check and make corrections to the information.

VI. Conclusions and Lines for Further Work

This paper presented the *Analisador* platform, a system to semi-automatically generate digital libraries of Ph.D. thesis and M.Sc. dissertations. The platform is able to work both with documents in pdf format and also to efficiently handle documents that are printed. In the latter case the dissertations should be scanned in grayscale at 200 dpi resolution. A number of tools integrated to the platform will generate good quality compressed images. For documents that were originally in pdf-format a PDF-Box routines were generated to automatically extract information such as title, author, supervisor, summary, abstract, etc. In the case of the scanned documents, after a number of image filtering procedures which besides generating good quality

compressed images also improve the quality of transcriptions, a pdf-text file is generated. Such file undergoes the PDF-Box routines developed in the *Anaflower* platform to try to extract the same set of information which is extracted from the pdf-text documents. The OCR routine also generates a text file that is used for double-checking the result of the features extracted from the pdf-text generated file. The tests performed so far with the content retrieval module of the *Anaflower* platform has shown to be very promising. In the case of pdf-text original documents the correct feature extraction rate was very high. As one may expect, in the case of the scanned documents the feature extraction presented much lower correct rate due to OCR imprecision.

Improvements may be made to the Feature Recognition Module in two ways: increasing the number of regular expressions and keywords, causing an increase in the feature extraction time and the use of better OCR tools.

The automatic extraction of references is also performed in *Anaflower*, but there is room for improvements on this feature as well. The target to be pursued is to be able to automatically find enough information to automatically check for the *doi* (digital object identifier) of each reference.

If this paper were accepted all the code for the algorithms and test images will be made publically available.

References

[1] Adobe® Supplement to the ISO 32000. Edição 3. http://www.images.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/adobe_supplement_iso32000.pdf, visited on 04/06/2011.

[2] LINS, Rafael Dueire, TORREAO, Gabriel, SILVA, Gabriel de França Pereira. Content Recognition and Indexing in the LiveMemory Platform. GREC 2009. LNCS. p.220 – 230,Springer Verlag, 2010.

[3] ABBY Fine Reader version 9.0.7. http://finereader.abbyy.com/ visited on 04/06/2011.

[4] LINS, Rafael Dueire, ÁVILA, Bruno Tenório, FORMIGA, Andrei de Araújo. BigBatch: An Environment for Processing Monochromatic Documents, ICIAR 2006, Springer Verlag, 2006. v.41142. p.886 – 896

[5] PDF-Box Home Page. Extracted from http://www.pdfbox.org , visited on 04/06/2011.

[6] Extensible Markup Language (XML) 1.0 (Fifth Edition) W3C Recommendation 26 November 2008. http://www.w3.org/TR/2008/REC-xml-20081126/, visited on 04/06/2011.

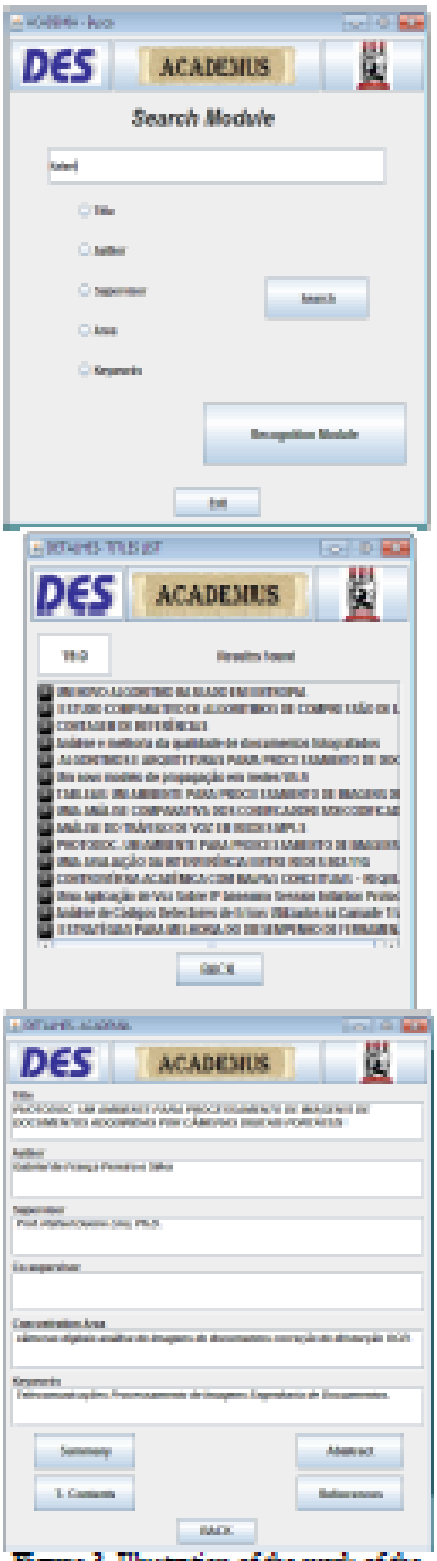


Figure 3. Illustration of the work of the search module of *Anaflower*