

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

**PRÉ-PROCESSAMENTO DE IMAGENS
NA PLATAFORMA *Thanatos***

Elaborado por:

Alessandra Bárbara Santos de Almeida

Recife, outubro de 2011.

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

PRÉ-PROCESSAMENTO DE IMAGENS NA
PLATAFORMA *Thanatos*

por

ALESSANDRA BÁRBARA SANTOS DE ALMEIDA

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: Prof. Dr. RAFAEL DUEIRE LINS

Recife, outubro de 2011.

© Alessandra Bárbara Santos de Almeida, 2011



Universidade Federal de Pernambuco

Pós-Graduação em Engenharia Elétrica

PARECER DA COMISSÃO EXAMINADORA DE DEFESA DE
DISSERTAÇÃO DO MESTRADO ACADÊMICO DE

ALESSANDRA BARBARA SANTOS DE ALMEIDA

TÍTULO

**“PRÉ-PROCESSAMENTO DE IMAGENS
NA PLATAFORMA *THANATOS*”**

A comissão examinadora composta pelos professores: RAFAEL DUEIRE LINS, CIN/UFPE, VALDEMAR CARDOSO DA ROCHA JÚNIOR, DES/UFPE e MARIA LENCASTRE PINHEIRO DE MENEZES CRUZ, POLI/UPE sob a presidência do primeiro, consideram a candidata **ALESSANDRA BARBARA SANTOS DE ALMEIDA APROVADA.**

Recife, 31 de outubro de 2011.

CECÍLIO JOSÉ LINS PIMENTEL
Vice-Coordenador do PPGEE

RAFAEL DUEIRE LINS
Orientador e Membro Titular Interno

**MARIA LENCASTRE PINHEIRO DE
MENEZES CRUZ**
Membro Titular Externo

VALDEMAR CARDOSO DA ROCHA JÚNIOR
Membro Titular Interno

Agradecimentos

Agradeço primeiramente a Deus, por ter me iluminado, me dado força e me abençoado em cada momento dessa trajetória.

À minha mãe, Risonete Santos, pela paciência e apoio nos momentos mais difíceis dessa caminhada, por nunca ter me deixado desistir e, principalmente, por ter me dado todo o alicerce, com seu exemplo de vida, para chegar até aqui e ter conquistado cada uma das vitórias da minha vida.

À minha tia, Conceição Lima, que tanto me motivou, me cobrou, me aconselhou e me orientou durante a construção desse trabalho.

Ao meu grande orientador, Rafael Dueire, que acreditou em mim quando me aceitou como orientanda, mesmo com tantos alunos; que me cobrou nos momentos certos e que foi genial na sua orientação, nos momentos em que eu não sabia como continuar, cumprindo com maestria o seu papel.

Ao meu grande amigo, Aldsmayths Pinheiro, por ter me motivado a entrar no mestrado, por ter me acompanhado durante cada dificuldade e alegria enfrentada, sempre com uma palavra sábia e uma sensibilidade apurada para me entender e me motivar.

Ao meu amigo, Roberto Arteiro, meu mestre, que tem sido uma grande referência de profissional pra mim, que foi tão decisivo na definição desse trabalho, na motivação constante e na confiança que sempre demonstrou por mim.

Aos amigos do TJPE, Ana Luísa, minha chefe, pela compreensão e confiança dispensadas; aos meus amigos da equipe de Segurança da Informação, Carlos Henrique, Gliner Alencar e Marcelo Lima, que conduziram o Núcleo nas minhas ausências, pela amizade, compreensão e apoio que tiveram por mim nesse período.

Aos amigos do PPGEE, Prof. Valdemar Rocha, por ter sido um exemplo de docente que quero levar pra minha vida toda, pelo seu conhecimento profundo e humildade transcendente. E aos amigos de convívio, que mesmo eu não sendo tão presente devido ao trabalho, me acolheram como uma família; eu nunca vou esquecer de vocês: Gabriel, Danielle, Andréa, Eduarda, Lizandra, Neide, Ednardo, Caio, Daniel, Paulo Martins, Paulo Freitas, Gilson, Brenno, Roberto, Eurico e Victor.

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

PRÉ-PROCESSAMENTO DE IMAGENS NA PLATAFORMA

Thanatos

Alessandra Barbara Santos de Almeida

outubro/2011

Orientador: Prof. Dr. Rafael Dueire Lins, Ph.D.

Área de Concentração: Engenharia Elétrica

Palavras-chave: Processamento Digital de Imagens, Engenharia de Documentos, Documentos Históricos, Registro de óbito, Pernambuco.

Número de Páginas: 108

RESUMO

Registros de óbito possuem importantes informações demográficas, pois além dos dados tais como informações genealógicas, *causa mortis*, idade do óbito, locais de nascimento e morte, permitem analisar correntes migratórias internas, a relação entre a *causa mortis* e o estado civil, sexo, profissão, etc.

Thanatos é uma plataforma desenvolvida para extrair informações das imagens dos Registros de óbito digitalizadas dentro do convênio celebrado entre o *Family Search International* e o Poder Judiciário do Estado de Pernambuco. Esta dissertação está inserida no contexto dessa plataforma e foca no pré-processamento dessas imagens históricas, preparando-as para a fase de reconhecimento automático da informação.

Abstract of Dissertation presented to UFPE as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering.

IMAGING PREPROCESSING IN *Thanatos* PLATFORM.

Alessandra Bárbara Santos de Almeida

October / 2011

Supervisor: Prof. Dr. Rafael Dueire Lins.

Concentration Area: Documents Engineering.

Keywords: Digital Image Processing, Document Engineering, Historical Documents, Death Certificates, Pernambuco.

Number of Pages: 108

ABSTRACT

Death Certificates convey important demographic information, besides genealogical information, such as *causa mortis*, age of death, birth and death places. They also allow to analyze internal migration currents and the relationship between cause of the death and marital status, sex, profession, etc.

Thanatos is a platform developed to extract information from images of death certificates that were digitalized within the agreement celebrated between the *Family Search International* and the Judiciary Power of the State of Pernambuco (Brazil). This M.Sc. dissertation was developed in the context of such platform and describes the image preprocessing phase used for those historical documents, to prepare them to automatic information retrieval.

Lista de Figuras

Figura 1 - Imagem física e sua correspondente imagem digital.....	10
Figura 2 - Exemplo de Plataforma de digitalização utilizada pelo Family Search.....	11
Figura 3 - Padrão de Linha para Ajuste de foco da câmera.....	13
Figura 4 - Realização de Ajuste do foco e sua visualização em tela.....	13
Figura 5 - Mesa branca para calibração do nível de branco da imagem	14
Figura 6 - Modelo de escala de cinza para calibração.....	14
Figura 7 - Plaquetas indicativas do estado de conservação dos registros.....	15
Figura 8 - Exemplo de Registro Totalmente Manuscrito	18
Figura 9 - Exemplo de Registro Parcialmente Manuscrito.....	19
Figura 10 - Imagem de Abertura - nº de ordem.....	20
Figura 12 - Imagem de Abertura – Referências	21
Figura 13 – Exemplo de Imagem de Abertura do Livro.....	21
Figura 11 - Imagem de Abertura - página em branco.....	21
Figura 14 - Exemplo de Capa de Livro de Registro	22
Figura 15 - Exemplo de Termo de Apresentação e Termo de Abertura do Livro de Registro.....	23
Figura 16 – Exemplo de Imagem de Fechamento do Livro	25
Figura 17 - Exemplo de Registro de Casamento	26
Figura 18 - Exemplo de Registro de Óbito.....	29
Figura 19 - Exemplo de Imagem com rotação (skew).....	36
Figura 20 - Diagrama em blocos da Plataforma Thanatos	37
Figura 21 - Imagem após ser processada pelo HistDoc (SILVA et al., 2010)	40
Figura 22 - Registro de óbito após a segmentação realizada pela Plataforma Thanatos.....	41
Figura 23 - Pontos de referência encontrados para segmentar a imagem.	43
Figura 24 - Áreas da imagem que podem ser apagadas se a	50
Figura 25 - Exemplo de imagem durante a execução do algoritmo.....	51
Figura 26 - Número de Registro antes e depois da remoção da linha horizontal.....	51
Figura 27- Ruído apresentado na imagem original e binarizada.....	52
Figura 28 - Reconhecimento de Campo não-numérico.....	56
Figura 29 - Zoneamento automático de caracteres	58
Figura 30 - Resultado da detecção de linhas fortemente descontinuas.	63

Sumário

Capítulo 1. Introdução	1
1.1. Problema e Proposta	3
1.2. Organização da Dissertação	4
Capítulo 2. Descrição do Acervo	7
2.1. Origem do Acervo	8
2.2. Aquisição das Imagens	9
2.3. Validade Jurídica	16
2.4. Tipos de Registros	17
2.5. Composição Geral do Acervo	20
2.6. Registros de Casamentos	25
2.7. Registros de Óbitos	29
Capítulo 3. Pré-Processamento e Segmentação	33
3.1. Escopo	33
3.2. Metodologia	34
3.3. HistDoc	35
3.4. Plataforma Thanatos	36
3.5. Segmentação dos Registros	38
3.5.1 Características	38
3.5.2 Algoritmo	38
3.6. Normalização dos Registros	41
3.6.1 Características	41
3.6.2 Algoritmo	42
3.7. Segmentação dos Campos de Informação	44
3.7.1 Características	44
3.7.2 Algoritmo	45
Capítulo 4. Remoção de Ruídos e de Linhas	48
4.1. Remoção de Linhas Horizontais	48
4.1.1 Características	48
4.1.2 Algoritmo:	49
4.2. Remoção de Ruídos	51
4.2.1 Características	52

4.2.2	Algoritmo	53
4.3.	Remoção de Linhas Verticais.....	54
4.4.	Reconhecimento e Classificação.....	56
4.4.1	Reconhecimento.....	56
4.4.2	Classificador de Campo Numérico	58
4.4.3	Classificador de Campo Não-Numérico.....	60
4.5.	Trabalhos Relacionados	61
Capítulo 5. Conclusão e Trabalhos Futuros.....		64
5.1.	Contribuições.....	64
5.2.	Trabalhos Futuros.....	65
Referências Bibliográficas.....		67
Anexo A – Artigo Publicado		72
Apêndice A - Código da Segmentação dos Registros		81
Apêndice B - Código da Uniformização das Imagens dos Registros		83
Apêndice C - Código para geração das Máscaras e Extração dos Campos dos Registros		86
Apêndice D - Código para Remoção de Linhas Horizontais		90
Apêndice E - Código para Remoção de Ruídos.....		92
Apêndice F - Código para Remoção de Linhas Verticais.....		94
Apêndice G - Código Reverso para Remoção de Linhas Horizontais.....		95
Apêndice H - Código para Segmentação dos Dígitos Numéricos		97

Capítulo 1.

I

ntrodução

As duas últimas décadas do século XX inauguraram a “era da Informação”, com o crescimento do acesso à Internet, o que fez com que informações de todas as naturezas e em todas as áreas da atividade humana estivessem cada vez mais disponíveis ao acesso de todos, sem fronteiras de qualquer natureza, inclusive geo-política. Entretanto, esse processo é muito recente, fazendo com que estejamos ainda em uma fase de transição em que muitas informações legadas encontram-se ainda em papel, ao mesmo tempo em que há necessidade de busca e acesso em tempo real a essas informações, mantendo-se a mesma disponibilidade e celeridade que é exigida hoje aos documentos digitais. Devido a isso, torna-se cada vez mais necessária a digitalização desses documentos legados, para que, de forma avançada, se permita também o armazenamento seguro com redundância e disponibilidade das imagens dos documentos digitalizados, a publicação dessas imagens, tratamento/preprocessamento, extração automática de dados e pesquisa e estudos nessas informações. Daí surge a Engenharia de Documentos, que consiste numa disciplina da ciência da computação que investiga os sistemas de documentos, sob qualquer forma e em todos os meios de comunicação; ela está preocupada com os princípios, ferramentas e processos que melhoram a capacidade de criar, gerenciar e manter documentos.

A Engenharia de Documentos pode ser aplicada a diversas áreas do conhecimento. Neste trabalho, é realizado o tratamento de imagens dos registros de óbitos do Estado de Pernambuco. Trata-se de um acervo histórico de documentos manuscritos, composto por mais

de um milhão de imagens que foram obtidas através do convênio estabelecido entre o Tribunal de Justiça de Pernambuco e a então denominada Sociedade Genealógica de Utah, atualmente chamada *Family Search International*. Esse Convênio tem como objetivo digitalizar os livros de registro civil do Estado de Pernambuco. Tais livros estão de posse dos cartórios de registro civil localizados ao longo de todo o território do Estado de Pernambuco. O *Family Search International* é ligado à Igreja de Jesus Cristo dos Santos dos Últimos Dias, também conhecida como Igreja Mormón. Eles ensinam que seus membros são responsáveis pelo batismo dos seus antepassados mortos. Se uma pessoa morre sem nunca ter sido batizada nesta vida, é possível que seu parente Mórmon seja batizado em seu lugar, sendo assim, salva. Joseph Smith, fundador do Mormonismo, ensinou que buscar os mortos desta forma é a grande responsabilidade dos Mórmons (RICHARDS, 1976).

A importância desse acervo se deve ao elevado número de documentos históricos com valor legal. O fato desses documentos não se encontrarem em meio eletrônico acarreta uma série de limitações, as quais podem ocasionar perdas irreparáveis nessas informações, algumas delas são:

- Dificuldade de manuseio, transferência e compartilhamento das informações entre os interessados;
- Dificuldade na realização de buscas e consultas no conteúdo do documento;
- Falta de segurança na manipulação ou acesso aos registros;
- Dificuldade para realização de cópias de segurança dos documentos e o seu armazenamento em ambiente seguro;
- Sujeição à perda irreparável de informação em caso de desastres, perda, roubo ou até mesmo devido a acidentes naturais;

Devido a esses e outros fatores existentes em documentos em papel, se torna iminente a necessidade de digitalização e armazenamento em formato eletrônico desses documentos, com o objetivo de preservar, organizar e facilitar o acesso a essas informações.

Para essa dissertação, foram selecionadas as imagens dos registros de óbito do acervo Mórmon-TJPE, pois a partir desses documentos é possível extrair importantes informações demográficas, de forma automatizada, sendo possível realizar estudos populacionais, avaliando as informações de *causa mortis* mais frequentes, estatísticas de idade do óbito, locais de nascimento e morte, informações genealógicas, etc. Essas informações podem ser usadas para analisar não só o que causou a morte da pessoa, mas também para gerar um grande número de informações demográficas, tais como correntes de migração inter e intraestaduais, a relação entre a *causa mortis* e o estado civil, sexo, profissão, etc. dentre outras aplicações que podem ser iniciadas a partir da disponibilização de uma base de dados que podem ser extraídas de forma automática dos registros civis que compõem o acervo disponibilizado.

1.1. Problema e Proposta

Esta dissertação de mestrado foca o tratamento de imagens de registros de óbitos, datados do ano de 1960 ao ano 2000, do cartório da 4ª zona de Recife e do Cartório de Palmares, os quais apresentam problemas e complexidades para o seu processamento, devido tanto à forma de aquisição dos Registros, quanto à forma de armazenamento e conservação desses livros ao longo do tempo.

Nesses tipos de imagens, podem-se destacar alguns problemas observados na etapa de aquisição das imagens, como bordas perimetrais à imagem (*noisy border*) (FORMIGA e LINS, 2009), rotação do documento (*skew*) (AVILA e LINS, 2005) e deformação central relativa à encadernação (*book binding warp*) (LINS *et al.*, 2010).

Por se tratar de documentos históricos e das características intrínsecas a esse tipo de base de imagens, é possível observar também nessas imagens, características que dificultam o trabalho de extração de informações, como:

- Ruído de sal-e-pimenta;
- Interferência frente-e-verso;
- Linhas verticais, horizontais e inclinadas;
- Ruídos dilatados;
- Acervo ainda não estudado pela comunidade acadêmica;
- Informações manuscritas.

O objetivo geral desta dissertação é realizar o pré-processamento desses documentos históricos, a partir de imagens digitalizadas de registros de óbito do Estado de Pernambuco. Esses documentos se encontram em sua maioria, manuscritos aumentando a complexidade do seu reconhecimento e conseqüentemente da extração das informações neles contidas.

Os objetivos específicos da presente dissertação são:

- Abordar o tema de processamento digital de imagens, detalhando as etapas constituintes do processo específico de pré-processamento, que trata desde a etapa de retirada de borda e tratamento de interferências, até a segmentação e extração de informações;
- Apresentar a metodologia utilizada nesta pesquisa específica de pré-processamento de imagens, a partir das imagens dos registros civis de óbito, detalhando todas as suas etapas;
- Descrever as ferramentas utilizadas e algoritmos desenvolvidos para este trabalho, especificamente para segmentação das imagens, remoção de linhas e de ruídos granulares; os quais podem ser utilizados, com algumas adaptações, em outros trabalhos de natureza semelhante.

1.2. Organização da Dissertação

Além deste capítulo introdutório, esta dissertação é composta pelos seguintes capítulos:

Capítulo 2 – Descrição do Acervo

Nesse capítulo é apresentada a forma de obtenção das imagens dos registros civis do Estado de Pernambuco, hardware e softwares utilizados, bem como os tipos, formatos e composição dos registros que compõem o Acervo a ser estudado.

Capítulo 3 – Pré-processamento e Segmentação dos Registros

Nesse capítulo é apresentado a metodologia adotada no tratamento das imagens dos Registros de óbitos (foco desta dissertação) bem como todas as etapas seguidas para a realização do pré-processamento dessas imagens históricas e os resultados obtidos. Também são detalhadas as técnicas utilizadas para uniformização das imagens segmentadas e extração dos campos de informação.

Capítulo 4 – Remoção de Ruídos e de Linhas

Nesse capítulo são detalhadas as etapas realizadas para remoção de ruídos dilatados e das linhas horizontais, inclinadas e verticais presentes em cada campo do Registro, com o objetivo de melhorar os resultados obtidos na etapa posterior de reconhecimento dos manuscritos, a qual é sucintamente explicada também neste capítulo.

Capítulo 5 – Conclusões e Trabalhos Futuros

Nesse capítulo são apresentados, os resultados, as contribuições e os benefícios alcançados em toda a pesquisa, bem como a comparação dessa dissertação com outros trabalhos recentes publicados na área, bem como as sugestões para trabalhos futuros para que seja dada continuidade à pesquisa.

Anexo A

O Artigo intitulado “*Thanatos – Automatically Retrieving Information of Death Certificates in Brazil*” publicado nos anais *I Workshop on Historical Document Imaging and Processing*, realizado em setembro de 2011 em Pequim, na China, descrevendo a Plataforma *Thanatos*, encontra-se no Anexo A desta dissertação.

Apêndices

Os códigos desenvolvidos pela autora para a Plataforma *Thanatos* estão incluídos nos Apêndices A a H desta dissertação.

Capítulo 2.

Descrição do Acervo

As primeiras imagens digitais foram transmitidas no início da década de 1920 pela indústria jornalística, utilizando um cabo submarino que interligava as cidades de Nova Iorque e Londres, o que reduziu de mais de uma semana para menos de três horas o tempo para transportar uma fotografia pelo oceano atlântico. As técnicas de processamento de imagens então utilizadas, vêm sendo estudadas e evoluídas até hoje (GONZALEZ e WOODS, 2010).

A Engenharia de Documentos abrange as subáreas de compressão e processamento de imagens de documentos. Esta dissertação foca na linha de pesquisa de processamento digital de imagens, que dentre outros objetivos, busca melhorar a qualidade de imagens possibilitando a visualização adequada do seu conteúdo, seja por um humano ou por uma máquina. Mais especificamente, esse processo de realce das imagens pode ter diversos objetivos, mas todos se concentram, de forma abrangente, na extração de atributos de uma imagem e até o reconhecimento de objetos individuais.

Para a realização do processamento das imagens digitais, algumas etapas precisam ser seguidas, sejam as de mais baixo nível, de pré-processamento, que objetivam a redução de ruídos, interferências, realce de contraste, etc. Outra etapa, de nível mais alto, é a de segmentação, que visa separar a imagem em regiões ou objetos com o intuito de reduzi-las a uma forma adequada a permitir o processamento computacional e a classificação desses

objetos. E por fim, a etapa de nível mais alto, que trata de caracterizar e classificar propriamente a informação extraída da imagem.

Neste capítulo é apresentada uma descrição do acervo utilizado como fonte de estudo para este trabalho. É detalhada a sua origem, forma de captura e os tipos, formatos e conteúdos dos registros existentes.

2.1. Origem do Acervo

O acervo, objeto de estudo deste trabalho, é composto por imagens dos registros civis do Estado de Pernambuco, o qual é formado por mais de um milhão imagens de registros de nascimento, casamento e óbito de cidadãos de todo o território do Estado de Pernambuco.

Esses registros estão originalmente localizados nos cartórios de registro civil do Estado e suas imagens pertencem ao Tribunal de Justiça de Pernambuco (TJPE). As fotos dos Registros foram obtidas e disponibilizadas pelo *Family Search International*, através do convênio de cooperação nº 30/2008 firmado entre esta Sociedade e o TJPE.

O Family Search International é uma sociedade sem fins lucrativos, fundada em 1864, sediada em Salt Lake City - Estados Unidos da América, e representada no Brasil pela Associação Brasileira D'A Igreja de Jesus Cristo dos Santos dos Últimos Dias. (GENEALOGICAL, 2008). Também é conhecida como Igreja dos Mórmons, e consiste em uma instituição que possui fins educacionais com o propósito de ajudar pessoas a se conectar com seus ancestrais através do acesso aos seus registros históricos. A Igreja de Jesus Cristo dos Santos dos Últimos Dias ensina que seus membros são responsáveis pelo batismo dos seus antepassados mortos. Se uma pessoa morre sem nunca ter sido batizado nesta vida, é possível que seu parente Mórmon seja batizado em seu lugar. Tal dever motiva a Igreja Mórmon a ter registros genealógicos em todo o mundo.

O Acervo digitalizado foi entregue ao TJPE, em discos rígidos (*Hard Disks* - HDs) e encontra-se de posse do Memorial de Justiça de Pernambuco, inclusive disponível para consulta pública. O TJPE autorizou a utilização de uma amostra dessas imagens para fins acadêmicos. A amostra autorizada para esta Pesquisa é composta de Registros de óbitos dos Cartórios da 4ª Zona de Recife e do Cartório de Palmares. Este último foi escolhido devido às enchentes ocorridas nesta região em junho de 2010, na qual houve perda irreversível de aproximadamente 15% do acervo do Cartório.

2.2. Aquisição das Imagens

Esta seção apresenta os dispositivos e técnicas utilizados para digitalização das imagens dos registros de óbito. O processo de digitalização, de uma forma geral, consiste em dividir a imagem física em pequenas regiões chamadas de *picture elements* ou *pixels*. O esquema mais comum de subdivisão é uma amostra retangular, como mostrado na figura 1, . O valor representativo de cada pixel da imagem indica o nível de cinza correspondente na imagem física. Esse processo de conversão da imagem é chamado de digitalização. O valor do nível de cinza de cada *pixel* capturado da imagem é amostrado e quantizado; esse valor final é um número inteiro que representará cada *pixel* da imagem digital (CASTLEMAN, 1996)

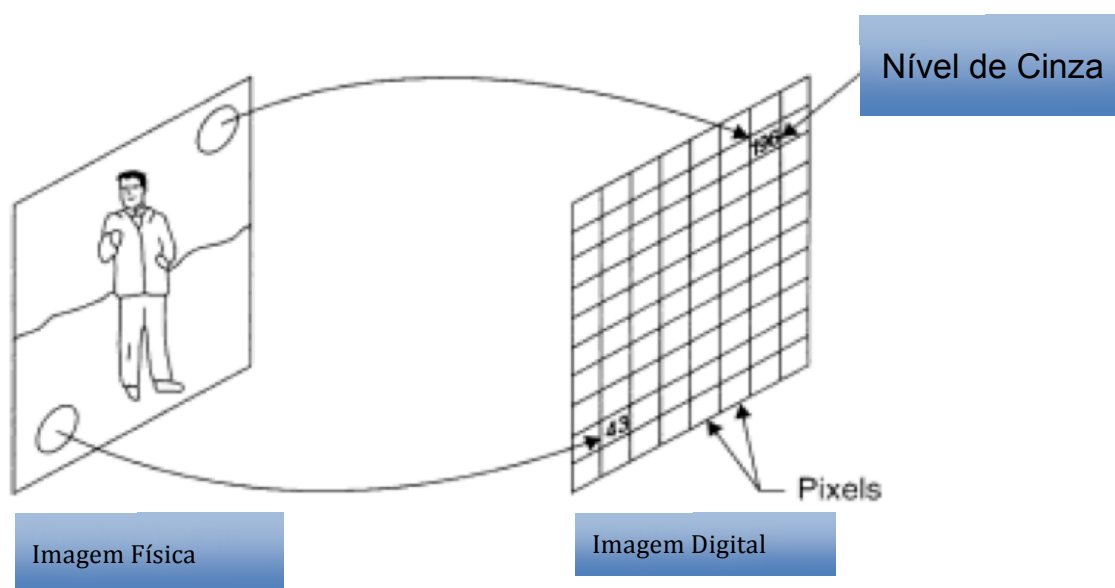


Figura 1 - Imagem física e sua correspondente imagem digital.

Na etapa de aquisição das imagens, os registros civis foram inicialmente microfilmados e disponibilizados em escala de cinza com intensidade de 8 bits e resolução de 200 pontos por polegada (dpi – *dots per inch*). As imagens adquiridas mais recentemente, utilizadas como objeto desta dissertação, foram digitalizadas utilizando câmeras fotográficas com resolução de 16,5 Mega Pixels.

A Figura 2 apresenta a plataforma utilizada atualmente para a digitalização desses registros, na qual pode ser observada uma câmera e seu suporte com altura ajustável para permitir fotografias de volumes de número de páginas variáveis, além de iluminação controlada, além dos demais detalhes que são descritos abaixo.

Como pode ser visto Figura 2, as imagens dos documentos foram capturadas utilizando uma mesa como base, com uma plataforma negra fixada à mesa. Perpendicularmente à mesa, foi fixado um suporte para a câmera com regulagem de altura, para ajuste da distância do objeto (livro de registros) à câmera. Na mesa também foram ligados 2 suportes laterais, com três lâmpadas cada, com o objetivo de iluminar a região a ser fotografada, com disposição de tal forma a evitar reflexos indesejáveis nas imagens; também não pode haver outros tipos de

iluminação, nem natural, nem artificial nessa região. Além desses itens estruturais, no ambiente de captura também foram observados aparatos de hardware e software, como um CPU e um monitor para processamento, armazenamento e visualização das imagens, e cabo serial para sua interligação.

Os softwares utilizados para captura, visualização e auditoria das imagens são os seguintes:

dCamII – esse aplicativo exibe a imagem em tamanho original e com tamanho ampliado, para permitir a verificação em tempo de aquisição das imagens que estão sendo geradas, antes do seu armazenamento definitivo.

Daims – consiste em um software para visualização das imagens, com o objetivo de se realizar uma auditoria visual e evitar erros grosseiros na imagem como um todo, como: parte da mão ou dos dedos compondo a imagem, objetos estranhos na área capturada, sombras, etc.

AuditTool – ferramenta que realiza auditoria automática dos metadados incluídos no fechamento de um lote de imagens (caixa).



Figura 2 - Exemplo de Plataforma de digitalização utilizada pelo Family Search para captura das imagens dos Registros Cíveis de Pernambuco.

Existe um procedimento padronizado pelo Family Search International que é seguido antes de se dar início à captura propriamente dita das imagens. Primeiramente, é realizada uma calibração da câmera para se ter uma aquisição das imagens com alguma padronização na qualidade, após se realizar o aquecimento das lâmpadas por 15 minutos, para se ter homogeneidade na iluminação. Nesta etapa de calibração é ajustada a abertura do diafragma; o ajuste do foco da lente, utilizando um padrão de linhas (exibido na Figura 3), a qual na imagem a ser capturada não pode ocupar mais de três linhas de pixels em todas as direções, como pode ser observado na tela da Figura 4.

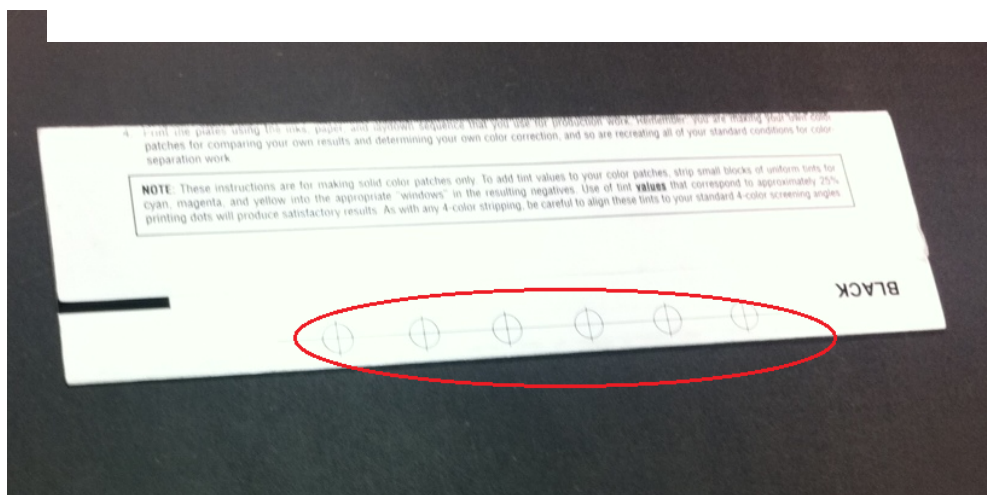


Figura 3 - Padrão de Linha para Ajuste de foco da câmera.



Figura 4 - Realização de Ajuste do foco e sua visualização em tela.

Também é realizada uma calibração para o nível de branco da imagem, para evitar saturação, para isso é utilizada uma mesa branca para este ajuste, como pode ser observado na Figura 5.

Também são calibrados os limiares de cinza, preto e negro, a partir da identificação de níveis pré-estabelecidos e padronizados seguindo a escala exibida na Figura 6 que varia do branco absoluto ao preto absoluto.

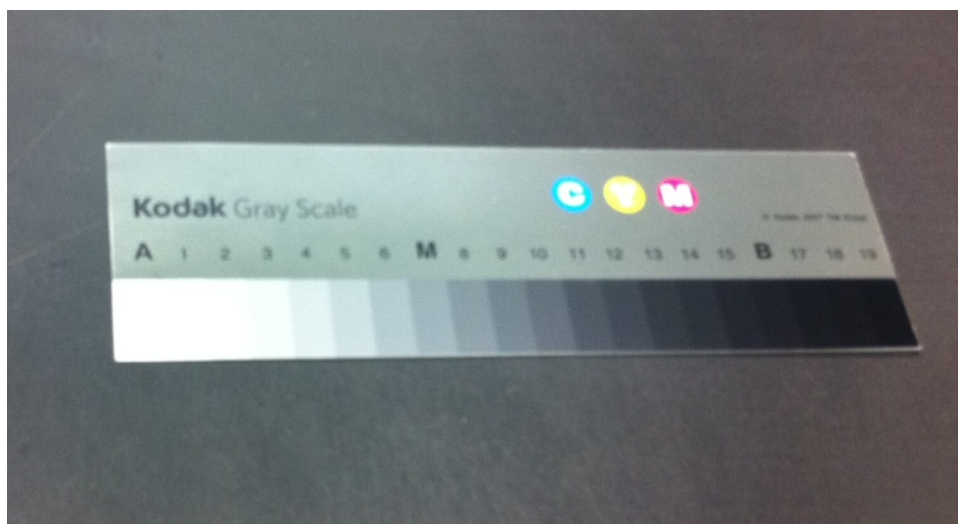


Figura 5 - Mesa branca para calibração do nível de branco da imagem

Para cada grupo de imagens (caixa) capturado são gerados macro metadados com as seguintes informações: nome da caixa, nome do primeiro registro do lote como referência, ano do acervo, modelo da câmera utilizada, número serial da câmera, código de identificação do fotógrafo, local de origem do acervo e comentários gerais sobre o estado de conservação dos documentos originais (ex.: folhas em mau estado de conservação, manchas d'água, folhas rasgadas, etc.). Alguns desses dados são também incluídos em plaquetas que são apensadas ao documento no momento da captura, como pode ser observado nas placas indicativas exibidas na Figura 7.

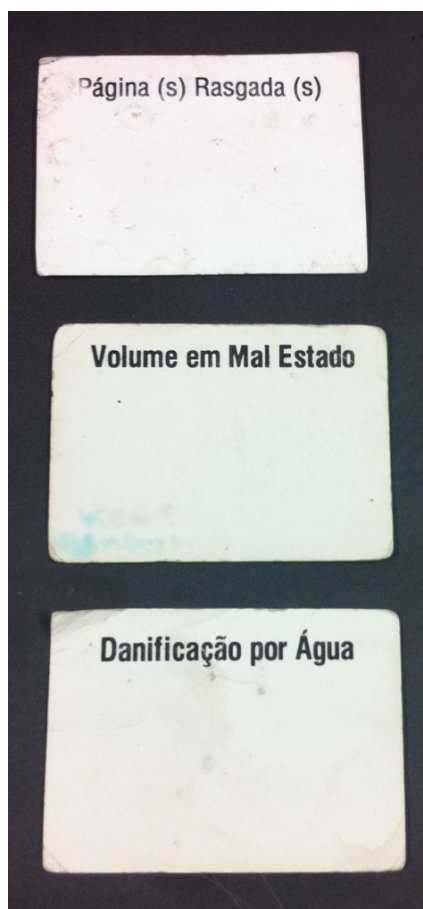


Figura 7 - Plaquetas indicativas do estado de conservação dos registros

São gerados arquivos de registro (*log*) que replicam esses metadados para cada imagem gerada, alterando apenas a data e hora da obtenção daquela imagem.

Após toda a etapa de digitalização das imagens, elas são gravadas em mídia magnética (HD – *Hard Disk*) e após uma etapa de Auditoria são entregues para o conveniado, no caso o TJPE. Segundo informações fornecidas por representante do Family Search, os HDs entregues ao TJPE são iguais aos que foram enviadas para os EUA, não possuindo metadados específicos para cada imagem (objeto deste trabalho). Essa inserção de informações, seria realizada manualmente, nas etapas pré-publicação no site do *Family Search International*, nos EUA.

2.3. Validade Jurídica

No Brasil, os microfilmes de documentos possuem a mesma validade jurídica que os seus originais, desde que atenda aos requisitos previstos na lei que os autorizam.

A lei nº 5.433 de 8 de maio de 1968 é a primeira regulamentação brasileira que autoriza e dá validade jurídica à utilização de microfilmes na reprodução de documentos, em substituição dos seus originais; conforme descrito no seu artigo 1º:

Art 1º É autorizada, em todo o território nacional, a microfilmagem de documentos particulares e oficiais arquivados, estes de órgãos federais, estaduais e municipais.

§ 1º Os microfilmes de que trata esta Lei, assim como as certidões, os traslados e as cópias fotográficas obtidas diretamente dos filmes produzirão os mesmos efeitos legais dos documentos originais em juízo ou fora dele.

(PRESIDENCIA, 1968)

O Decreto nº 1.799 de 30 de janeiro de 1996, regulamenta a Lei nº 5.433, de 8 de maio de 1968, detalhando como deve ser realizada a microfilmagem e suas restrições, conforme artigos 1º, 2º e 5º como se segue:

Art. 1º A microfilmagem, em todo território nacional, autorizada pela Lei nº 5.433, de 8 de maio de 1968, abrange os documentos oficiais ou públicos, de qualquer espécie e em qualquer suporte, produzidos e recebidos pelos órgãos dos Poderes Executivo, Judiciário e Legislativo, inclusive da Administração indireta da União, dos Estados, do Distrito Federal e dos Municípios, e os documentos particulares ou privados, de pessoas físicas ou jurídicas.

Art. 2º A emissão de cópias, traslados e certidões extraídas de microfilmes, bem assim a autenticação desses documentos, para que possam produzir efeitos legais, em juízo ou fora dele, é regulada por este Decreto.

·
·
·

Art. 5º A microfilmagem, de qualquer espécie, será feita sempre em filme original, com o mínimo de 180 linhas por milímetro de definição, garantida a segurança e a qualidade de imagem e de reprodução.

§ 1º...

§ 2º Fica vedada a utilização de filmes atualizáveis, de qualquer tipo, tanto para a confecção do original, como para a extração de cópias.

(PRESIDENCIA, 2006)

2.4. Tipos de Registros

O Acervo Mórmon-TJPE completo em estudo é composto por mais de 01 (um) milhão de imagens entre registros civis de nascimento, casamento e óbito do Estado de Pernambuco. O período de obtenção das imagens compõe registros que vão desde o ano de 1889 até o ano de 2000. O acervo é composto por imagens de registros que foram gerados de forma totalmente manuscrita, conforme Figura 8 ou parcialmente manuscritas, conforme Figura 9, as quais possuem os campos não variáveis previamente tipografados e os campos variáveis escritos de forma manual.

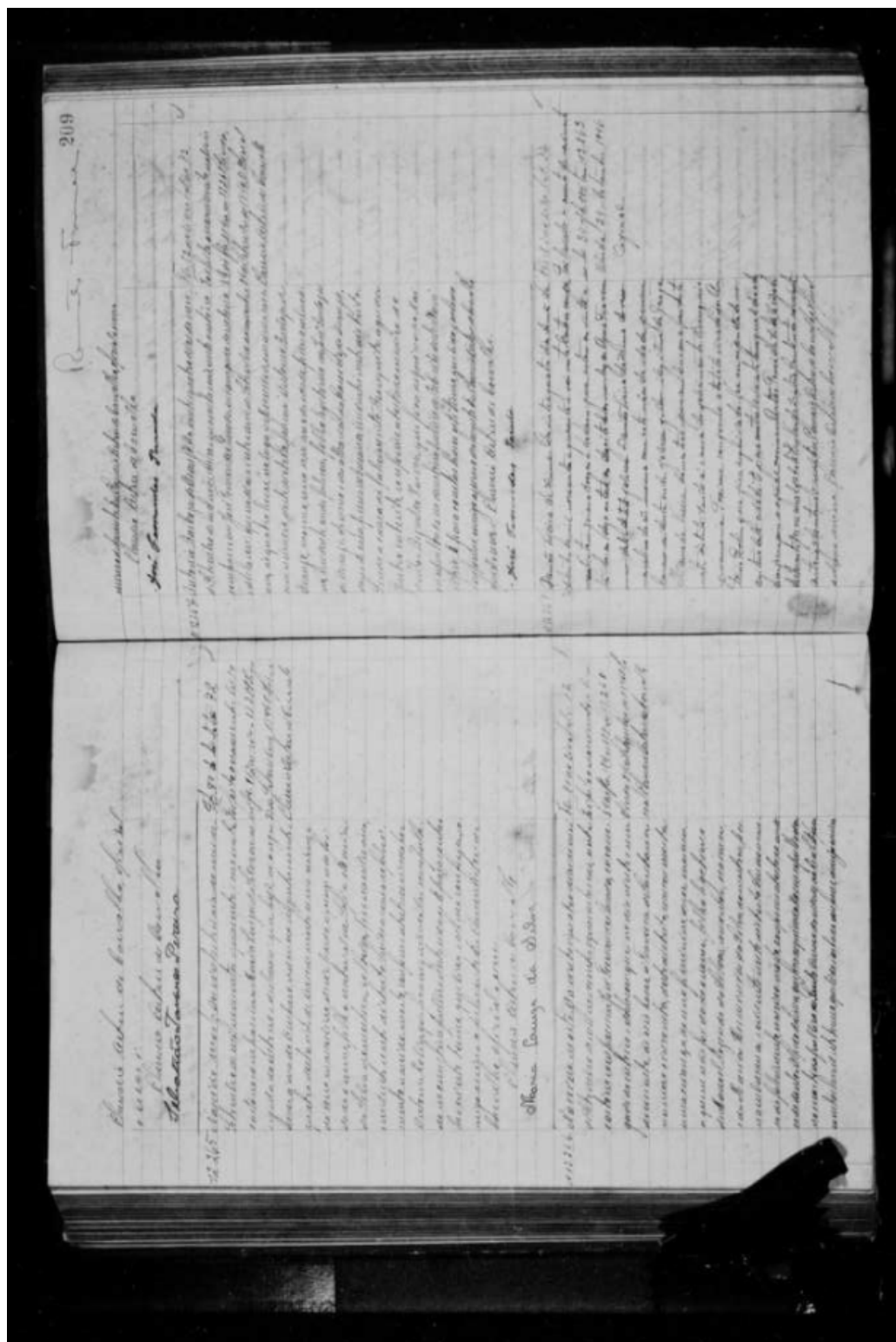


Figura 8 - Exemplo de Registro Totalmente Manuscrito

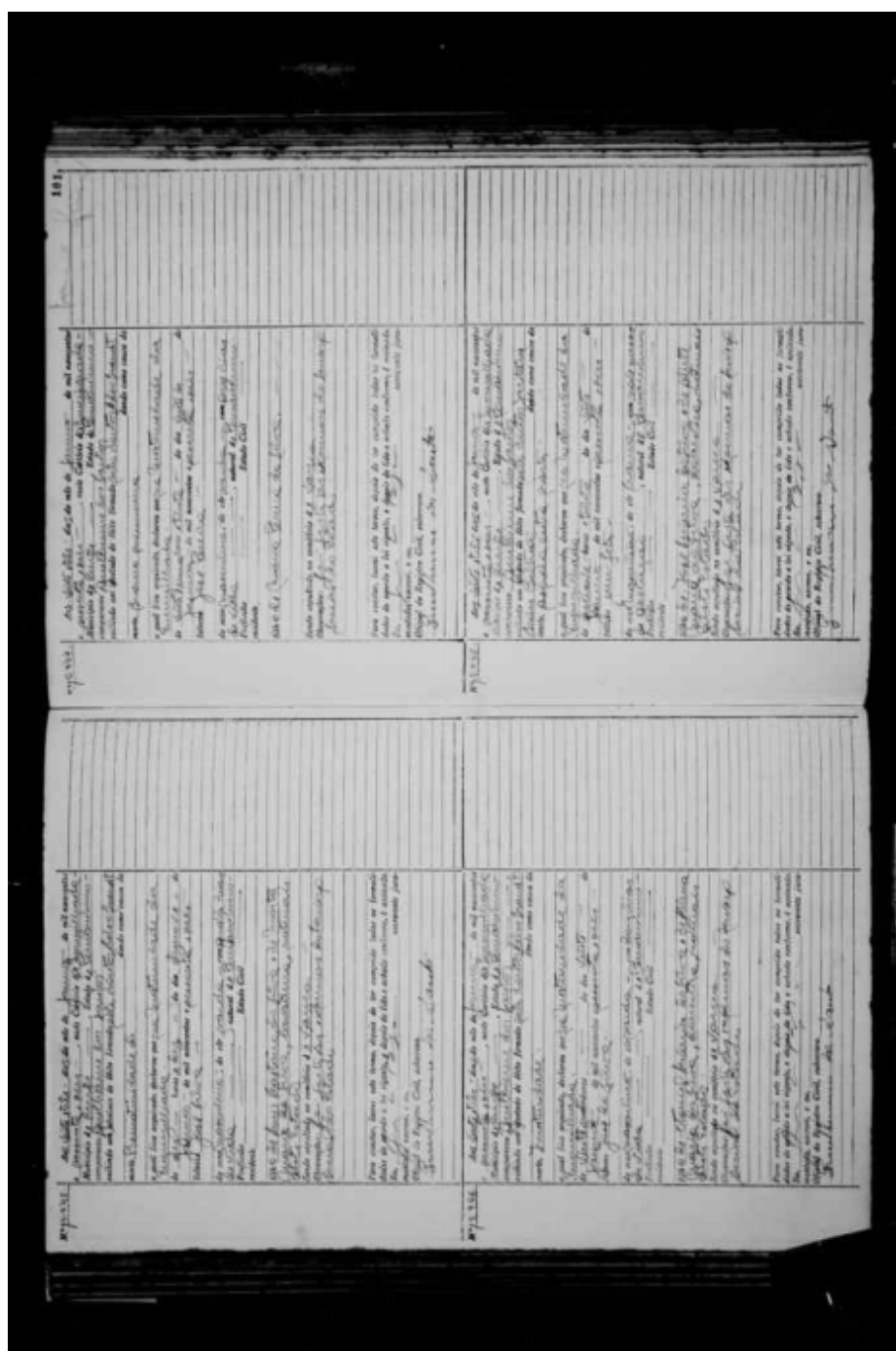


Figura 9 - Exemplo de Registro Parcialmente Manuscrito

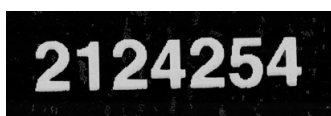
Na seção a seguir é feita uma descrição detalhada de cada uma das partes que compõem o Acervo em questão.

2.5. Composição Geral do Acervo

As imagens disponibilizadas estão agrupadas em diretórios, organizados não somente com as imagens dos registros propriamente ditos, mas também com imagens que foram adicionadas para organização e identificação do Acervo, sendo composto por imagens de abertura, fechamento e classificação do Acervo. O detalhamento dessa sequência é relevante para o tratamento automatizado das imagens.

O diretório tem como primeira imagem um número de ordem identificando cada diretório (*Figura 10*), seguida de uma imagem em branco (*Figura 11*), depois por uma imagem que contém referências de escala, direção, posicionamento e de preto e branco (*Figura 12*). Logo após, é adicionada uma imagem que descreve as características de obtenção da imagem (fotógrafo, data da filmagem, número da emulsão e da máquina, número do projeto e do rolo de microfilme) e da origem e conteúdo das imagens (localidade do registro, título do registro – informando óbito, nascimento ou casamento – e um número de ordem dentro daquele diretório), é a última imagem que antecede o início do Livro propriamente dito (*Figura 12*).

Figura 10 - Imagem de Abertura - n° de ordem



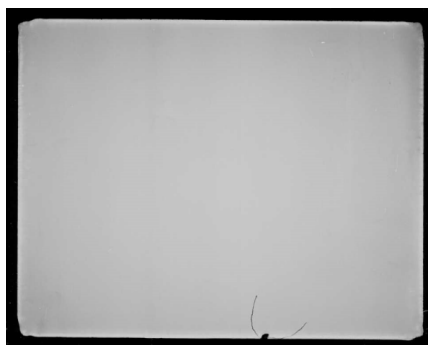


Figura 11 - Imagem de Abertura - página em branco.

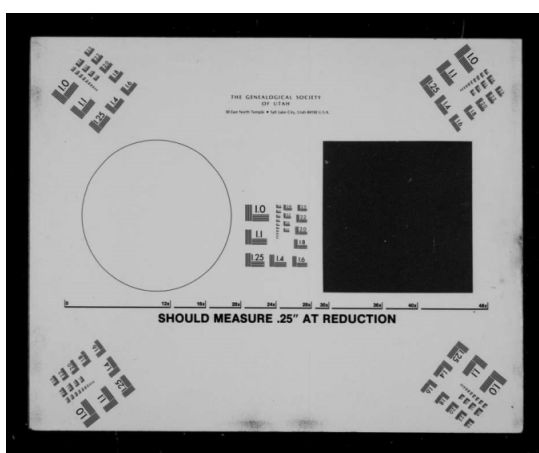


Figura 12 - Imagem de Abertura – Referências

2		FILMADO PELA SOCIEDADE GENEALÓGICA DE UTAH EM:		LOCALIDADE QUE ABRANJE O REGISTRO	
		REGISTRO CIVIL DO EST. DE PERNAMBUCO		CARTÓRIO DA 2ª ZONA DE CARUARU	
FOTOGRAFO	REDUÇÃO	TÍTULO DO REGISTRO			
JOSÉ C.F. CAMPOS 560	42	ÓBITOS			
DATA DE FILMAGEM	EXPOSIÇÃO	01			
20 OUT 99	23				
NÚMERO DA EMULSÃO	NÚMERO DA MÁQUINA				
AGFA 50330103	HRP 50875				
NÚMERO DO PROJETO	NÚMERO DO ROLO				
BRZC 2-017	343				

Figura 13 – Exemplo de Imagem de Abertura do Livro

Após a imagem de Abertura do Livro, são iniciadas as imagens do livro de registro propriamente ditas, começando pela sua capa, a qual contém informações do tipo de registro e o número de referência do Livro para o Cartório, conforme Figura 14, seguida pela página que contém o Termo de Apresentação (Figura 15) e de Abertura do Livro e, posteriormente, por todas as imagens dos registros civis.



Figura 14 - Exemplo de Capa de Livro de Registro

Figura 15 - Exemplo de Termo de Apresentação e Termo de Abertura do Livro de Registro.

No Termo de Apresentação é informada a quantidade de páginas que compõem o Livro, seu número de ordem, localidade do cartório e data de início do livro. Um exemplo de texto de um termo de Apresentação é o seguinte:

Termo de Apresentação
Ao Exmo Sr. Dr. Juiz de Direito da 2ª vara desta comarca, apresento este livro
contendo trezentas folhas, sob número de ordem c-21 para registro de óbito do
cartório da 2ª zona judiciária desta comarca, a fim de ser autenticado.
Caruaru, 17 de dezembro de 1961
<assinatura>
Oficial de Registro Civil

O Termo de Abertura contém, basicamente, as mesmas informações do Termo de Apresentação, segue um exemplo:

Termo de Abertura
Servirá este livro, contendo trezentas (300) folhas, tipograficamente numeradas
de um (1) a trezentos (300), sob número de ordem c-21, que levarão a rubrica
<rubrica>, de que faço uso para registro de óbitos do cartório de registro civil da
2ª zona judiciária do primeiro distrito desta comarca.
Caruaru, 17 de dezembro de 1961.

Na última página do livro de registro é descrito o Termo de Encerramento do Livro, o qual informa o número de páginas, o tipo de registro cadastrado e cartório equivalente aos registros. Segue um exemplo de Termo de Encerramento:

Termo de Encerramento
Fica encerrado este livro que contém trezentas (300) folhas, tipograficamente
numeradas e rubricadas com a rubrica , rubrica> de que uso e que nele serão
lavrados os termos de óbito, do Cartório de registro civil do distrito de Afogados
desta cidade.
Recife, 28 de junho de 1957
<assinatura>

Após a imagem da última página do Livro, é adicionada uma figura de encerramento (contendo o número do projeto e do rolo, localidade do registro, título do registro, o número de ordem dentro daquele diretório e a inscrição “FIM”, conforme Figura 16. Todas essas imagens adicionais servem para auxiliar a detecção e identificação dos registros de forma automatizada.



Figura 16 – Exemplo de Imagem de Fechamento do Livro

2.6. Registros de Casamentos

O Acervo completo é composto por registros de casamento, nascimento e óbito. Apesar do foco deste trabalho ser em registros de óbitos, é descrito a seguir, também, os campos e conteúdos dos registros de casamento para auxiliar em trabalhos futuros. Diferentemente dos registros de óbitos, os de casamento ocupam uma (01) folha inteira do Livro, com isso, cada imagem possui dois (02) registros de casamento. A Figura 18 demonstra um exemplo de registro de óbito.

N.º 21763

Aos dois dias do mês de Outubro de mil novecentos e quarenta e oito, nesta cidade de Pacife Estado de Pernambuco às dez horas, no Palácio da Justiça presente

o Doutor Evandro Lemos Neto Juiz de Direito dos Casamentos e perante as testemunhas nomeadas e assinadas, observadas as formalidades legais, receberam-se em matrimônio pelo regime de comunhão de bens, o senhor José Brenes Augusto de Lima e dona Helena Maria dos Santos. O contraente é Solteiro nascido em Pernambuco no dia dois de Outubro de mil novecentos e dois e sua profissão Operário domiciliado e residente em Pacife.

filho legítimo de José de Freitas Lima e de Olíndina Augusta de Lima.

A contraente é Solteira nascida em Pernambuco no dia dois de Março de mil novecentos e trinta e um e sua profissão Prof. Elementar domiciliada e residente em Pacife.

filha legítima de José Almeida dos Santos e de Marina de Conceição.

A nubente após o casamento chamar-se-á Helena dos Santos Lima.

Foram apresentados os documentos exigidos no artigo cento e oitenta do Código Civil Brasileiro de números um e quatro e publicado o edital de proclamas de um e quatro dias sem oposição de impedimento. Do que para constar lavrei este termo que depois de lido e achado conforme é assinado pelo Juiz, as testemunhas e as nubentes, José Brenes Augusto de Lima, casado, e Helena Maria dos Santos, casada, ambos de Pernambuco, residentes em Pacife, após o registro de dois dias, em Pacife, no dia dois de Outubro de mil novecentos e quarenta e oito.

José Brenes Augusto de Lima
Helena Maria dos Santos
José Ferreira Costa
Helinda Silva Costa

Figura 17 - Exemplo de Registro de Casamento

Os campos que compõem um registro de casamento são os seguintes:

- Número do registro – localizado no campo superior esquerdo do registro, de forma destacada das outras informações que o compõem e grafada em formato numérico e manuscrito, acompanhada da informação “Nº” digitada. Esse campo pode ter de 1 a 5 números. Ex.: Nº **21.763**.

Os campos descritos a seguir seguem a mesma ordem do seu posicionamento do registro, sendo a data de registro a primeira informação, até a assinatura dos envolvidos, como última informação. As informações grafadas em negrito correspondem àquilo que é dado variável do registro, as outras informações são fixas no livro.

- Data do registro – a data é grafada por extenso. Ex.: Aos **vinte e dois** dias do mês de **outubro** de mil novecentos e **quarenta e oito**.
- Cidade do Cartório – Ex.: *nesta cidade do Recife;*
- Estado do Cartório – Ex.: *Estado de Pernambuco;*
- Horário do Casamento – a hora é grafada por extenso e não são citados os minutos – Ex.: *às dezessete horas;*
- Nome do Fórum/Cartório de realização do casamento – Ex.: *no Palácio da Justiça;*
- Nome do Juiz responsável pela oficialização/legalização do Matrimônio – Ex.: presente o Doutor **Evandro Muniz Neto**, Juiz de Direito dos Casamentos e perante as testemunhas nomeadas e assinadas, observadas as formalidades legais;
- Regime de divisão dos bens – podendo ser comunhão de bens, comunhão parcial de bens ou separação de bens – Ex.: *receberam em matrimônio pelo regime de comunhão de bens;*
- Nomes dos cônjuges – Ex.: o senhor **J. B. A. de Lima** e dona **H. M. dos Santos**;
- Informações do Contraente:

- Estado Civil – Ex.: O contraente é **solteiro**;
- Estado do nascimento – Ex.: *nascido em **Pernambuco***;
- Data de Nascimento – por extenso – Ex.: no dia **vinte e três** de **outubro** de **mil novecentos e vinte e seis**;
- Profissão – Ex.: profissão **operário**;
- Cidade da Residência – Ex.: domiciliado e residente em **Recife**;
- Filiação – campo composto da expressão “legítimo(a) de” e dos nomes completos do pai e da mãe. Ex.: *filho **legítimo de J. F. Lima e O.A. de Lima***.
- Informações da Contraente:
- Estado Civil – Ex.: A contraente é **solteira**;
- Estado do nascimento – Ex.: *nascida em **Pernambuco***;
- Data de Nascimento – por extenso – Ex.: no dia **vinte e sete** de **março** de **mil novecentos e trinta e um**;
- Profissão – Ex.: profissão **lar doméstica**;
- Cidade da Residência – Ex.: domiciliada e residente em **Recife**;
- Filiação – campo composto da expressão “legítimo(a) de” e dos nomes completos do pai e da mãe. Ex.: *filha **legítima de J. A. dos Santos e M. da Conceição***.
- Novo nome da esposa – Ex.: A nubente após o casamento chamar-se-á **H. dos Santos Lima**;
- Validade do Ato – Neste campo é referenciado o código civil e o edital de proclamas – Ex.: Foram apresentados os documentos exigidos no artigo cento e oitenta do Código Civil Brasileiro de números **um a quatro** e publicado o edital de proclamas **de onze do corrente até hoje** sem oposição de impedimento.
- Nomes e assinaturas dos envolvidos no Ato – são descritos também o estado civil, profissão e residência das testemunhas - Ex.: Do que para contar lavrei este termo

que depois de lido e achado conforme é assinado pelo juiz, os nubentes e testemunhas J. F. Costa, casado, artista, residente em Santo Amaro e A. S. Costa, casada, doméstica, residente com o seu esposo no distrito de Santo Amaro. Eu, S. B. da Silva, escrevente juramentado escrevi. <assinaturas >

2.7. Registros de Óbitos

O foco desta dissertação é o tratamento das imagens de registros de óbitos do Estado de Pernambuco, devido a isso, é detalhada a seguir a formação dos campos desse tipo de registro e sua composição. A imagem original que foi disponibilizada é formada por quatro imagens de registro de óbito, sendo duas por folha, a qual precisa ser segmentada conforme mostrado na Figura 18 traz um exemplo de registro de óbito, referente a um quarto da imagem obtida.

N.º 19.945.

Aos vinte e três dias do mês de janeiro de mil novecentos e sessenta e seis, neste Cartório de Registro Civil do Município de Recife, Estado de Pernambuco compareceu Guilherme dos Santos exibindo um atestado de óbito firmado pelo doutor Celso Brandt dando como causa da morte, Prematuridade de

o qual fica arquivado, declarou que no Intermunidade de da

de dez horas e dez do dia de nove de de mil novecentos e sessenta e seis faleceu José Filipe

do sexo masculino, de cor branca, com qualis dias de vida, natural de Pernambuco

Profissão Estado Civil

residente

filho de Luiz Baptista dos Filhos e de Maria Maria de Fátima, casados, naturais do Estado

Sendo sepultado no cemitério de A. Vaz

Observações: Em falta de exames de toxicologia

Para constar, lavrei este termo, depois de ter cumprido todas as formalidades de acordo a lei vigente, e depois de lido e achado conforme, é assinado. Eu, escrevente juramentado, escrevi, e eu, Oficial do Registro Civil, subcrevo.

Guilherme dos Santos

Figura 18 - Exemplo de Registro de Óbito

Os campos que compõem esse tipo de registro são os seguintes:

- Número do registro – localizado no campo superior esquerdo do registro, de forma destacada das outras informações que o compõem e grafadas em formato numérico. Ex.: Nº **19.945**.

Os campos descritos a seguir seguem a mesma ordem do seu posicionamento do registro, sendo a data de registro a primeira informação, nome do cartório a segunda informação e assim sucessivamente até o nome do oficial de registro, como última informação. As informações grafadas em negrito correspondem àquilo que é dado variável do registro, as outras informações são fixas no livro.

- Data do registro – a data é grafada por extenso. Ex.: Aos **vinte e três** dias do mês de **janeiro** de mil novecentos e **sessenta e seis**.
- Nome do cartório – nesse campo é informada a localidade onde está situado o cartório. Ex.: *neste cartório da **Encruzilhada***.
- Município do Cartório – Ex.: *município de **Recife***.
- Estado do Cartório - Ex.: *Estado de **Pernambuco***.
- Nome do Declarante – Nome de quem compareceu ao Cartório para informar o óbito. Ex.: *compareceu **Guilherme dos Santos***.
- Nome do Médico – Nome do médico responsável pelo atestado de óbito. Ex.: *exibindo um atestado de óbito firmado **pelo doutor Celso Brandt***.
- Causa mortis – Especifica a causa da morte declarada no atestado de óbito. Ex.: dando como causa da morte **prematuridade**, o qual fica arquivado.
- Local da Morte – Nome do local onde foi identificada a morte. Ex.: *declarou que **na maternidade da Encruzilhada***.
- Horário do óbito – hora e minuto por extenso. Ex.: às **dez horas e dez**;

- Data do óbito – dia, mês e ano por extenso. Ex.: Aos **dezenove** dias do mês de **janeiro** de mil novecentos e **sessenta e seis**.
- Nome do de cujus - Ex.: faleceu **J. Silva**.
- Sexo do de cujus – campo binário, masculino ou feminino. Ex.: *do sexo masculino*.
- Cor do de cujus – É um campo controlado, foram observadas as cores: parda, branca, morena, , além de alguns estarem em branco. Ex.: *de cor parda*.
- Idade do de cujus – Idade na data da morte grafada por extenso. Pode ser descrita em dias, meses ou anos, seguida das expressões: de vida, de gestação, de idade ---. Ex.: *com quatro dias de vida*.
- Local de nascimento do de cujus – é informado o Estado de nascimento. Ex.: *natural de Pernambuco*.
- Profissão do de cujus – É informada a profissão que exercia o falecido, quando não há é adicionado um travessão (–). Ex.: *profissão –*.
- Estado Civil do de cujus –Este é um campo controlado, podendo conter um travessão (–), além de existirem campos em branco ou a expressão “ignorado”, quando não se aplicar ou não for informado ou as seguintes expressões: solteiro(a), casado(a), viúvo(a), --- . Ex.: *estado civil –*.
- Endereço do de cujus – Também é informado um travessão (–) quando não se aplicar. Ex.: *residente –*.
- Filiação do de cujus – São informados os nomes dos pais, profissão da mãe e do pai (mas em alguns casos só é informado a da mãe), e naturalidade de ambos; todos separados apenas por vírgulas. Ex.: *filho de L. C. da Silva e de M. M. da Silva, lavadeira, naturais deste Estado*.

- Local do Sepultamento – nome do cemitério onde foi sepultado o corpo. Ex.:
Sendo sepultado no cemitério da Várzea.
- Observações – São feitas observações quaisquer, sem padrão definido - Ex.:
observações: *Foi feito às expensas do serviço social do Estado.*
- Nome do escrevente - Ex.: Eu, **José da Silva**, escrevente juramentado, escrevi.
- Nome do oficial de registro – Esse campo pode estar em branco, quando só houver o escrevente. Ex.: *e eu, ____, Oficial de Registro Civil, subscrevo.*
- Assinatura do Declarante – Na última linha consta apenas a assinatura do declarante.

Em resumo, este capítulo apresentou todas as etapas realizadas para aquisição das imagens, bem como uma descrição detalhada da composição do acervo com informações das características necessárias à etapa que se segue de pré-processamento dos Registros.

Capítulo 3.

P

ré-Processamento e Segmentação

Este capítulo explica a parte inicial do pré-processamento dos Registros de Óbitos, abrangendo tanto a metodologia quanto os procedimentos realizados para retirada de borda, correção de alinhamento, remoção de ruído de sal-e-pimenta, remoção de interferência frente-e-verso (SILVA *et al.*, 2010), segmentação dos registros, extração dos campos e uniformização das imagens. Etapas essas necessárias para preparar as imagens para etapa seguinte de extração e tratamento automático dos campos dos registros, a serem apresentados no próximo capítulo.

Todos os algoritmos utilizados nesta dissertação foram desenvolvidos em MathWorks - MATLAB R2009a e estão disponíveis nos Apêndices de A a H dessa dissertação.

3.1. Escopo

Como já dito, este trabalho tem como restrição de escopo o tratamento dos registros de óbito do estado de Pernambuco. Esse tipo de Registro foi escolhido devido a possibilidade de extração de um maior número de informações para estudos populacionais, como através: do conteúdo da informação da *causa mortis*, idade do óbito, locais de nascimento e morte, informações genealógicas; além de correntes migratórias, pela avaliação do campo de naturalidade versus local do falecimento, a relação entre a *causa mortis* e o estado civil, sexo, profissão, etc.

Os registros de óbitos a serem estudados foram cedidos pelo TJPE para fins desta pesquisa, compreendem os Cartórios da 4ª Zona de Recife, datados de 1966 a 1998; e do Cartório de Palmares, no período de 1960 a 2000. Nesse período, os documentos apresentam-se parcialmente manuscritos, garantindo assim, o posicionamento dos campos em regiões previamente definidas, possibilitando assim a segmentação e busca textual. Para o pré-processamento são utilizadas quatrocentas imagens para testar s resultados dessa etapa.

Após restringido o escopo das imagens que serão estudadas, precisa-se definir o escopo desta etapa de pré-processamento. Nesta fase, são realizados diversos tratamentos na imagem com o objetivo de prepará-la, tratando e realçando as imagens, para a seção seguinte, de reconhecimento. Para isso, é necessário realizar os seguintes processamentos: retirada de borda, correção de rotação, segmentação individual dos registros, uniformização das imagens e extração dos campos de informação. Segundo (GONZALEZ e WOODS, 2010) o “realce de imagens” é o processo de manipular uma imagem de forma que o resultado seja mais adequado do que o original para uma aplicação específica. O termo “específica” é importante nesse contexto, pois estabelece que as técnicas utilizadas são selecionadas de acordo com o problema que se deseja solucionar, no caso, reconhecimento de caracteres manuscritos. Portanto, é esse o escopo do capítulo de pré-processamento.

3.2. Metodologia

A metodologia utilizada nesta etapa de pré-processamento e segmentação dos registros foi o estudo e a aplicação de técnicas e ferramentas existentes, para as primeiras etapas, mais comuns a maioria das imagens de documentos históricos que precisam de tratamento, como: binarização da imagem, retirada de borda, remoção de interferência frente-e-verso, remoção de ruído de sal-e-pimenta e correção de inclinação. Depois dessa primeira etapa, foi realizado o desenvolvimento de técnicas mais específicas para o caso particular, não restritivo, do

acervo a ser trabalhado, como: segmentação dos registros, uniformização das imagens, extração de cada um dos campos dos registros históricos e remoção de ruídos granulares e de linhas horizontais e verticais.

3.3. HistDoc

Na etapa inicial se utilizou a nova versão da plataforma HistDoc (LINS *et al.*, 2011) para remover os ruídos provocados por uma fraca interferência frente-e-verso, também conhecido como *bleeding* (KASTURI *et al.*, 2011) ou *show through* (SHARMA, 2001), através da aplicação do algoritmo descrito na referência (SILVA *et al.*, 2010), que avalia a intensidade da interferência para ajustar o algoritmo de limiarização global.

Após a remoção da interferência frente-e-verso (SILVA *et al.*, 2010), é realizada a etapa de remoção da borda (AVILA e LINS, 2004) preta presente ao redor da imagem utilizando o algoritmo apresentado na referência (FORMIGA e LINS, 2009). Após essa etapa, foi realizada a binarização da imagem, pois percebeu-se que realizando a etapa de binarização antes da retirada de borda, estava ocorrendo significativa perda de informação da imagem por se utilizar a limiarização global, com isso a média de pixels pretos aumentava o valor do limiar de apagamento, prejudicando a legibilidade das informações textuais.

A próxima etapa consiste na correção de inclinação (*skew*) existente na imagem, como pode ser observado na Figura 19, a correção de inclinação (*skew*) a qual é realizada pelo algoritmo descrito na referência (AVILA e LINS, 2005) e, finalmente, filtragem do ruído de sal-e-pimenta é aplicada.

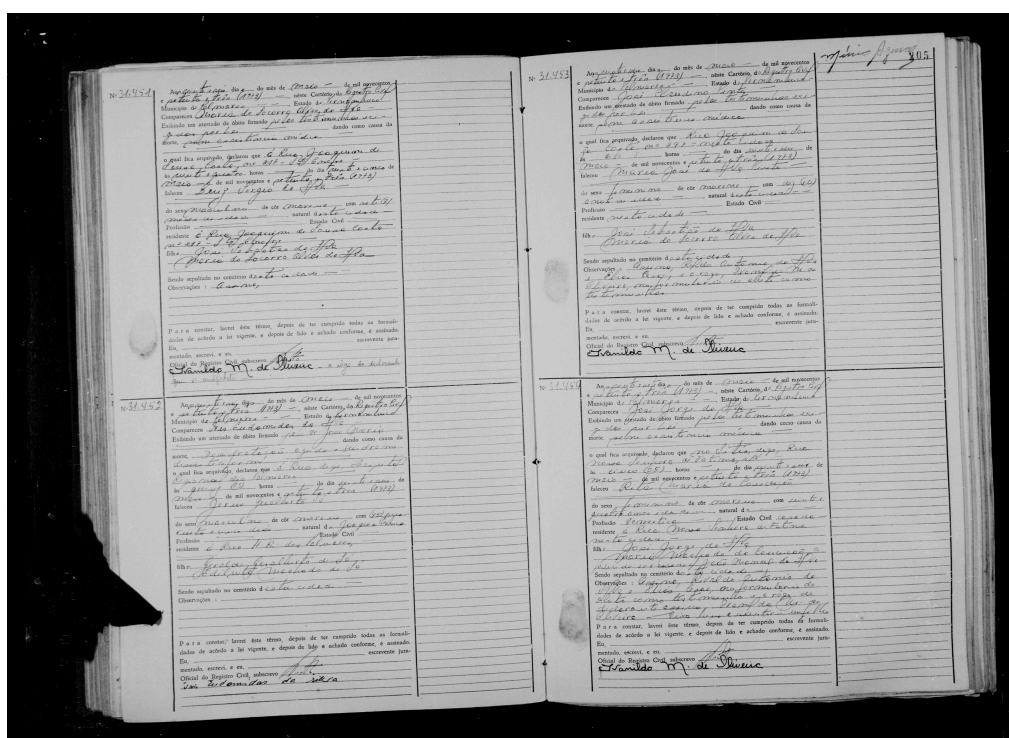


Figura 19 - Exemplo de Imagem com rotação (skew).

3.4. Plataforma *Thanatos*

A pós esta etapa de tratamento inicial e mais genérico na imagem, descrita acima, que é realizada pelo HistDoc (LINS *et al.*, 2011), se inicia a utilização da primeira parte da Plataforma *Thanatos*, que busca tratar mais especificamente, e não restritivamente, os registros de óbito, para realçar o seu conteúdo e retirar informações desnecessárias, como linhas e ruídos mais granulares da imagem, contribuindo para um melhor resultado na segunda etapa da Plataforma *Thanatos*, de reconhecimento dos caracteres, descrita de forma mais suscinta no final do Capítulo 4. A plataforma *Thanatos* é mais bem detalhada ao longo desse capítulo e do capítulo 4 e também pode ser observada no diagrama da Figura 20, onde são exibidas as etapas percorridas tanto na Plataforma HistDoc, quanto na *Thanatos*.

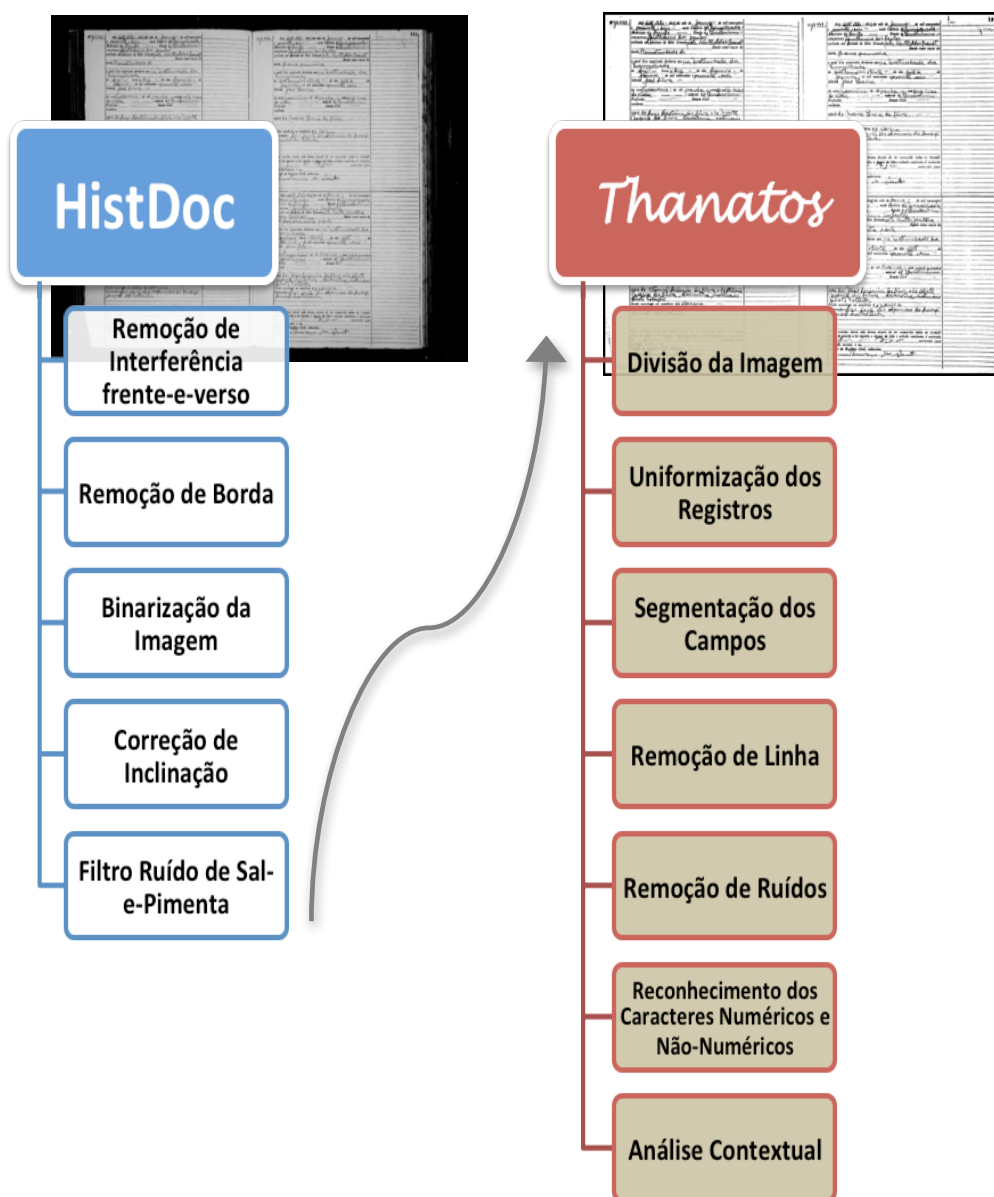


Figura 20 - Diagrama em blocos da Plataforma *Thanatos* e sua integração com o *HistDoc*.

3.5. Segmentação dos Registros

Essa etapa de segmentação dos registros é realizada pelo núcleo da plataforma *Thanatos*, a qual tem como objetivo segmentar a imagem pré-processada pela plataforma HistDoc em quatro registros de óbito, permitindo a extração de cada campo de informação dos registros na etapa posterior.

3.5.1 Características

A Figura 21 exibe uma imagem dos registros após todas as etapas executadas pela plataforma HistDoc. Nela é possível observar duas páginas do livro de registro aberto e com a presença de quatro registros de óbitos.

Esses registros apresentam características diferentes entre eles, dificultando essa fase de segmentação. Podem ser percebidas distorções no registro superior direito, devido a encadernação do livro; diferenças de tamanho entre as imagens da direita e da esquerda, superiores e inferiores; diferenças entre os tamanhos dos registros, devido à distância entre a mesa e a câmera fotográfica; entre outros.

3.5.2 Algoritmo

Esta operação é realizada automaticamente buscando a linha vertical central da imagem, o que corresponde à “coluna vertebral” do volume, sendo a imagem segmentada verticalmente neste ponto, produzindo as imagens da página esquerda e direita com dois certificados presentes em cada página.

Após isso, uma pesquisa é realizada horizontalmente para encontrar a linha central do documento. Devido ao processo de aquisição de imagem, tal linha não corresponde à linha mediana na imagem. Esta tarefa então é realizada fixando a análise numa área central da página e executando a técnica de *projection profile* (HA *et al.*, 1995) horizontal na imagem. A

linha a qual se está à procura é ligeiramente mais espessa do que as outras linhas. Este método funciona satisfatoriamente e produz as imagens superiores e inferiores, a exemplo do que é mostrado na Figura 22, na qual é possível ver o registro do canto superior esquerdo já segmentada. O algoritmo utilizado para a realização da segmentação dos Registros está disponível no Apêndice A desta dissertação.

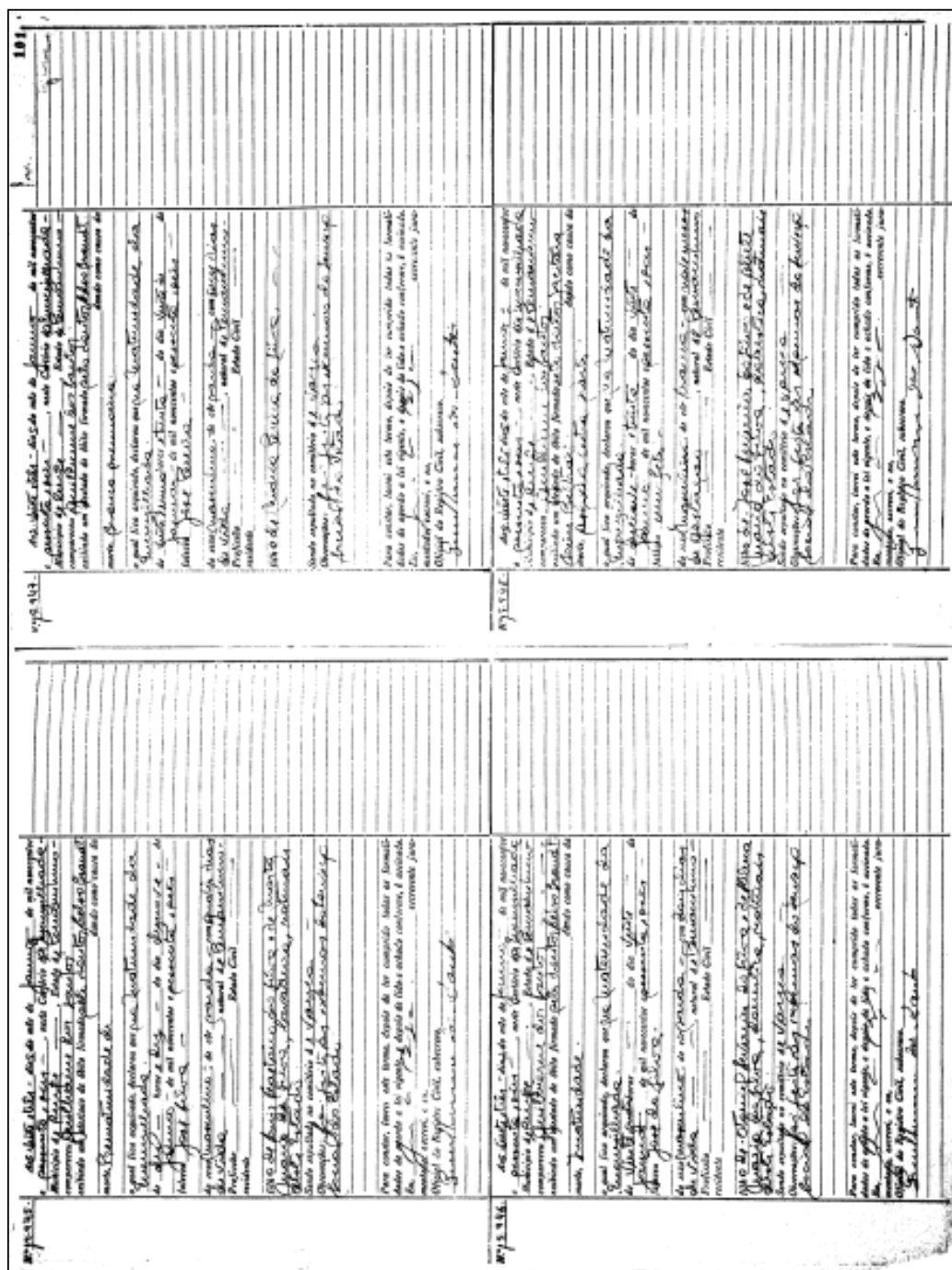


Figura 21 - Imagem após ser processada pelo HistDoc (SILVA et al., 2010) e pelo BigBatch (MATTOS et al., 2008).

N.º 967 - Dos dias - dias do mês de Setembro - de mil novecentos e sessenta e seis - neste Cartório da Municipalidade - Município de Recife, Estado de Pernambuco - compareceu Genilherme dos Santos exibindo um atestado de óbito firmado pelo doutor Genil Ferreira da Costa dando como causa da morte, Delirium sanguis.

o qual fica arquivado, declarou que ma naturalidade da mãe zelanda de quarta - horas - do dia quinta - de sete de mil novecentos e sessenta e seis - faleceu Marina Alice das Neves do sexo feminino - de cor branca - com três horas de vida - natural de Pernambuco - Profissão residente Estado Civil residente

filha de João Francisco das Neves e de Marina Alice das Neves, domiciliada, natural de Recife Sendo sepultado no cemitério de Nossa Observações: for feita as exéquias do funeral de Recife

Para constar, lavrei este termo, depois de ter cumprido todas as formalidades de acordo a lei vigente, e depois de lido e achado conforme, é assinado. Eu, Genil Ferreira da Costa escrevente juramentado, escrevi, e eu, Genil Ferreira da Costa Oficial do Registro Civil, subscrevo.

Figura 22 - Registro de óbito após a segmentação realizada pela Plataforma Thanatos.

3.6. Normalização dos Registros

Nesta etapa as imagens são normalizadas, ou seja, são uniformadas para permitir que imagens diferentes passem a possuir características semelhantes, facilitando assim, a identificação dos campos de forma padronizada e sua posterior extração.

3.6.1 Características

Nas imagens segmentadas dos registros, que foram resultado da etapa de segmentação anteriormente descrita, apresentam distorções que precisam ser corrigidas para permitir a posterior extração dos campos, são elas: a deformação causada pela encadernação dos livros, o que aumenta a complexidade de tal tarefa; o posicionamentos relativos de cada um dos registros no livro, em relação às suas margens; o tamanho do livro de registro; sua

quantidade de páginas; o seu posicionamento na mesa de captura e, até mesmo, da altura em que a Câmera fotográfica foi posicionada em relação à mesa.

Devido a isso, a uniformização da imagem precisa ser realizada, pois os quatro certificados presentes na imagem (superior/inferior esquerdo e superior/inferior direito) possuem diferentes formatos, dificultando assim a localização e segmentação dos campos individuais de informação.

As imagens tiveram sua normalização fixada em 1.900 pixels de largura por 1.470 pixels de altura, por ser o mínimo possível para manter os registros uniformizados, considerando os fatores apresentados acima, de diferenças de tamanho e por se perceber que a maioria dos registros possuíam dimensões em torno desse valor, sendo um pouco maiores ou um pouco menores, o que fez com que fosse possível, o corte das imagens maiores sem que houvesse perda de informação e o enchimento das imagens menores sem que houvesse um grande crescimento nas imagens.

3.6.2 Algoritmo

Para realizar a normalização de cada um dos Registros, se faz necessário um pré-processamento da imagem, o qual é realizado através da busca automática de "campos" delimitadores, como mostrado na Figura 23, na qual são destacadas as duas linhas verticais principais da imagem e a linha horizontal posicionada abaixo do número do Registro. Essas linhas serviram como base para delimitar e uniformizar os registros.

Vários problemas foram encontrados no desenvolvimento desse processo de normalização da imagem. Como pode ser observado na Figura 23, as linhas verticais têm vários trechos de descontinuidade devido à má qualidade de impressão tipográfica e também devido à binarização realizada na imagem. A encadernação do livro, em alguns casos,

também causou inclinação nas linhas horizontais, o que tornou a detecção da linha uma tarefa difícil.

Além desses fatores, outros também trouxeram uma carga extra de análise e processamento: livros de registro também variam de tamanho. Alguns dos registros, inclusive, eram menores do que 1.900 (horizontal) por 1.470 (vertical) pixels; dimensões essas que foram definidas como mínimas para um registro. Em tais casos, foi realizado o enchimento com pixels brancos na parte superior e margem direita das imagens.

Após o processo de normalização das imagens, todos os registros de óbitos tinham as mesmas dimensões (1.900 x 1.470 pixels) e as mesmas posições relativas entre diferentes imagens para os campos de mesma informação, permitindo assim o processo de segmentação.

O algoritmo utilizado para a realização da uniformização das imagens dos Registros está disponível no Apêndice B desta dissertação.

Handwritten text on the form:

Nº 19.947. Aos vinte e três dias do mês de Janeiro de mil novecentos e sessenta e seis, neste Cartório da Municipalidade do Município de Curitiba, Estado do Paraná, compareceu Antônio dos Santos exibindo um atestado de óbito firmado pelo doutor Antônio dos Santos dando como causa da morte Prisão prematura.

O qual fica arquivado, declarou que a testemunha de fato de Antônio dos Santos do dia 23 de Janeiro de mil novecentos e sessenta e seis faleceu Jose Pereira de sexo masculino, de cor branca, com seis dias de vida, natural de Paraná, Estado Civil residente.

Filho de João Pereira da Silva.

Sendo sepultado no cemitério de Varig.

Observações: foi feito o exame de corpo

Para constar, laurei este termo, depois de ter cumprido todas as formalidades de acordo a lei vigente, e depois de lido e achado conforme, é assinado. Eu, João Pereira da Silva, escrevo este juramentado escrevi, e eu, João Pereira da Silva Oficial do Registro Civil, subcreno.

Figura 23 - Pontos de referência encontrados para segmentar a imagem.

3.7. Segmentação dos Campos de Informação

O objetivo desta etapa é segmentar automaticamente cada um dos campos do registro de óbito, gerando uma imagem isolada para cada informação que será posteriormente processada e reconhecida.

3.7.1 Características

O conteúdo desses campos são manuscritos e frequentemente ultrapassam os espaços em branco, sobrescrevendo outras palavras ou ocupando as linhas adjacentes.

Um “Campo de Informação” abrange qualquer um dos campos que foram descritos no Capítulo 2. Nesta etapa foram tratados os oito primeiros campos do registro, os quais abrangem as informações mais relevantes do acervo, como: *causa mortis*, município do óbito, número e data do registro, etc. Podemos observar abaixo um exemplo desses dez primeiros campos são (os textos em negrito são os escritos nos espaços pelo escrivão do cartório):

- Número do registro – localizado no campo superior esquerdo do registro, de forma destacada das outras informações que o compõem e grafadas em formato numérico. Ex.: Nº **19.945**.
- Data do registro – a data é grafada por extenso e é formada por três (03) campos distintos. Ex.: [dia] Aos **vinte e três** dias [mês] do mês de **janeiro** [ano] de mil novecentos e **sessenta e seis**.
- Nome do cartório – nesse campo é informada a localidade onde está situado o cartório. Ex.: *neste cartório da **Encruzilhada***.
- Município do Cartório – Ex.: *município de **Recife***.
- Estado do Cartório - Ex.: *Estado de **Pernambuco***.

- Nome do Declarante – Nome de quem compareceu ao Cartório para informar o óbito. Ex.: *compareceu **Guilherme dos Santos***.
- Nome do Médico – Nome do médico responsável pelo atestado de óbito. Ex.: *exibindo um atestado de óbito firmado **pelo doutor Celso Brandt***.
- Causa mortis – Especifica a causa da morte declarada no atestado de óbito. Ex.: *dando como causa da morte **prematuridade**, o qual fica arquivado*.

3.7.2 Algoritmo

Para realizar a extração automatizada das informações, foram criadas máscaras para cada um dos Campos de Informação do registro de óbito a ser segmentado. Cada máscara de campo possui um tamanho variável, e para sua geração foi desenvolvido um algoritmo para identificação das dimensões de cada máscara e geração de uma Matriz de Máscaras, doravante denominada MM, de dimensão nx4, onde ‘n’ é o número de máscaras criadas ou o número de campos a ser trabalhado; a MM possui a seguinte formação:

$$MM = \begin{bmatrix} X0_1 & Y0_1 & W_1 & H_1 \\ X0_2 & Y0_2 & W_2 & H_2 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X0_n & Y0_n & W_n & H_n \end{bmatrix}$$

No qual ‘n’ é o enésimo elemento da matriz MM, ‘X0’ é a coordenada inicial do eixo dos X, ‘Y0’ é a coordenada inicial do eixo dos Y, ‘W’ é a largura do enésimo campo e ‘H’ é altura dessa máscara.

A criação de uma matriz MM permite segmentar cada um dos campos da imagem. Como um resultado da aplicação da matriz de máscaras nos registros de óbito é possível obter

cada uma das imagem com o conteúdo dos campos dos formulários como mostrado na Tabela 1. O algoritmo utilizado para a geração da matriz de máscaras (MM) realização da extração dos campos dos Registros de óbito está disponível no Apêndice C desta dissertação.

Tabela 1 - Campos do registro de óbito segmentados.

Máscara nº	Nome do Campo	Imagem
01	Número do Registro	09.947.
02	Data	24 de Maio de 1981 -
03	Mês	Junho
04	Ano	1981
05	Município do Cartório	Queluzópolis
06	Cidade	Queluzópolis
07	Estado	Paraná
08	Nome do Declarante	Queluzópolis
09	Nome do Médico que emitiu o atestado de óbito.	Dr. Carlos José Ricardo
10	Causa Mortis	Edema pulmonar

O reconhecimento direto dos campos manuscritos observados na coluna "imagem" na Tabela 1 não apresentaram resultados satisfatórios. Muitos dos problemas detectados nas imagens geraram efeitos prejudiciais na precisão do reconhecimento dos caracteres, são eles: as linhas horizontais das lacunas a serem preenchidas, as linhas verticais que definem as margens do registro, as manchas remanescentes, ruídos físicos da imagem, e caracteres muito

estreitos. Para resolver ou mitigar tais problemas, cada um dos registros passaram por um novo estágio de pré-processamento a ser apresentado no Capítulo 4 desta dissertação.

Capítulo 4.

Remoção de Ruídos e de Linhas

Nesse capítulo são apresentadas as técnicas desenvolvidas para a remoção de ruídos e de linhas horizontais e verticais dos campos segmentados dos registros que foram detalhados no capítulo anterior, com o objetivo de melhorar os resultados obtidos na etapa de reconhecimento das informações, a qual também é apresentada neste capítulo.

4.1. Remoção de Linhas Horizontais

A primeira etapa desta seção é remover as linhas horizontais pré-impressas no registro, presentes em cada campo da imagem segmentada no capítulo anterior. Para resolver este problema, um algoritmo foi desenvolvido para percorrer a imagem do campo segmentado à procura de regiões que se assemelham a uma linha horizontal.

4.1.1 Características

Esta técnica foi desenvolvida com o objetivo, também, de não ser sensível a discontinuidades na linha, ou seja, mesmo se a linha apresentar erosões ou até longas discontinuidades ela será removida mesmo assim, problema este que é encontrado em quase todas as imagens analisadas. Essa discontinuidade se deve principalmente à baixa qualidade das imagens dos documentos, devido ao processo de conservação dos livros de registros originais, bem como ao processo de binarização.

Em relação às características do algoritmo, é importante ressaltar que ele foi desenvolvido de forma a ser sensível aos caracteres que cruzam as linhas pré-impressas no documento, sendo assim, não realiza o apagamento dos caracteres de informação durante o seu processamento.

Também se tomou o cuidado para que o algoritmo não apresentasse restrições morfológicas relativas ao paralelismo e posicionamento das linhas, bem como a distância existente entre elas. Restrições essas que podem ser observadas, por exemplo, no artigo (ZHENG et al., 2003)

4.1.2 Algoritmo:

O algoritmo desenvolvido consiste na realização de uma varredura da imagem, utilizando um bloco de 8x8 pixels, o qual analisa a imagem percorrendo-a da esquerda para a direita, e de cima para baixo, na busca de identificar regiões similares à uma linha horizontal. O bloco analisado é preenchido com pixels brancos (apagamento do bloco) quando ocorrem as seguintes situações:

- (i) se a primeira e última linha do bloco for toda branca;
- (ii) se o bloco é predominantemente branco; e
- (iii) se primeira coluna possui pelo menos um pixel preto.

O critério ‘i’ se faz necessário para evitar o apagamento de partes dos caracteres que cruzam a linha. O critério ‘ii’ foi escolhido para evitar a exclusão de regiões que são predominantemente de informação – pixels pretos. O critério ‘iii’ foi definido para evitar o apagamento de regiões que representam entradas/início dos caracteres.

A Figura 24, apresenta exemplos de regiões de extremidade das imagens nas quais seriam apagadas se não fosse apagada a terceira condição acima, de verificar a existência de pelo menos 1 pixel preto na primeira coluna do bloco.

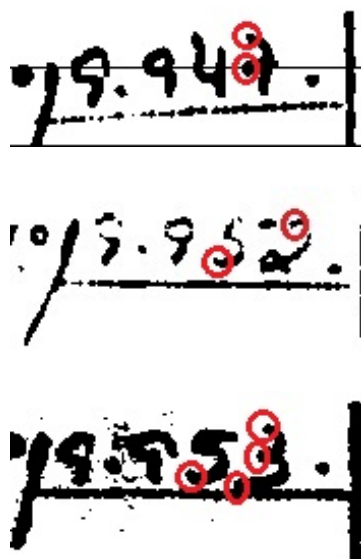


Figura 24 - Áreas da imagem que podem ser apagadas se a primeira coluna do bloco não for testada.

A Figura 25 ilustra de forma gráfica o algoritmo para remover linhas horizontais descrito. Nela é possível observar a execução do algoritmo de remoção de linha horizontal, sendo exibido um bloco de 8x8 pixels, percorrendo a figura da esquerda para direita e de cima para baixo, conforme as orientações de varredura exibidas na figura. Dessa forma, é possível realizar uma comparação entre o tamanho do bloco de varredura e o tamanho do campo a ser analisado.

Nesse exemplo o bloco será apagado por atender a todos os critérios de apagamento anteriormente descritos. O algoritmo utilizado para a realização da remoção das linhas horizontais está disponível no Apêndice D desta dissertação.

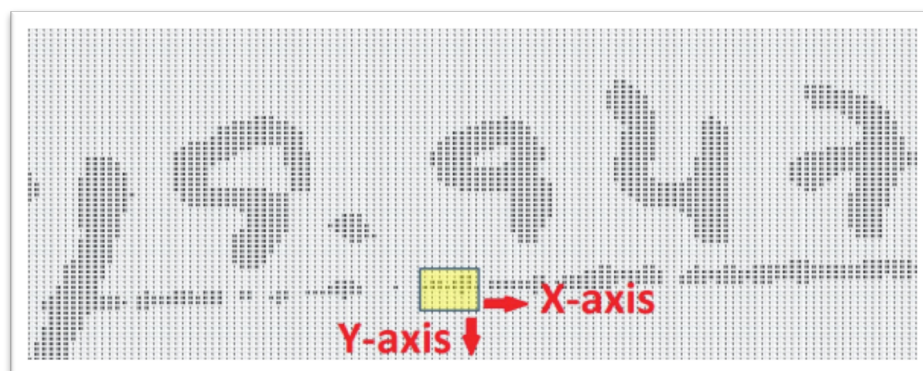


Figura 25 - Exemplo de imagem durante a execução do algoritmo.

O resultado desta etapa, para a mesma imagem do campo do número de registro (nº 19.947 neste exemplo), pode ser observado na Figura 26, comparando a imagem antes (primeira coluna) e depois (segunda coluna) de ser processada pelo algoritmo proposto, nesta podemos observar a linha horizontal removida.

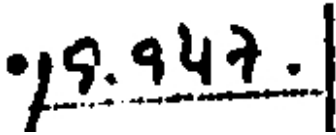
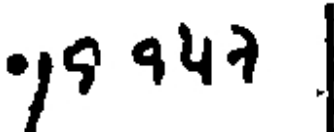
Imagem Original	Imagem Processada
	

Figura 26 - Número de Registro antes e depois da remoção da linha horizontal.

4.2. Remoção de Ruídos

O próximo passo na etapa de pré-processamento é a remoção de ruídos dilatados. Os ruídos aqui trabalhados podem ser gerados por manchas, mofo, envelhecimento do papel, poeira, etc; quer sejam ruídos físicos ou introduzidas durante a digitalização, como poeira na lente da câmera ou no livro durante a aquisição da imagem; ou até mesmo introduzidos durante o processo de binarização (LINS, 2009).

4.2.1 Características

Apesar de toda a imagem estar coberta pelo ruído de sal-e-pimenta, tratado na última fase de pré-processamento realizado pela Plataforma HistDoc (SILVA *et al.*, 2010), o ruído tratado nesta seção trata de ruído de maior granularidade, com dimensões maiores que os de sal-e-pimenta, apresentando características peculiares que dificultam o seu tratamento, como a não homogeneidade na sua distribuição.

Esse tipo de ruído também é conhecido como *Stroke-like Pattern Noise* (SPN) (AGRAVAL e DOERMANN, 2011) e é adicionalmente caracterizado por fazer parte do documento físico ou durante sua produção, escaneamento, transmissão, armazenamento ou conversão de um formato para outro; como está descrito em (AGRAVAL e DOERMANN, 2011). Esse ruído também pode ser chamado de *clutter noise*, conforme descrito na referência (AGRAWAL e DOERMANN, 2009).

A filtragem desse tipo de ruído, gerado pelos fatores apresentados, melhora a qualidade da imagem e aumenta a taxa de acerto no reconhecimento. A Figura 27 apresenta um exemplo de um campo de informação (número do registro) do documento em escala de cinza (original) e os ruídos gerados após a fase de binarização.

Como se pode observar na Figura 27, o tamanho dos "grãos" de ruído são muito mais grossos do que o do ruído de sal-e-pimenta.

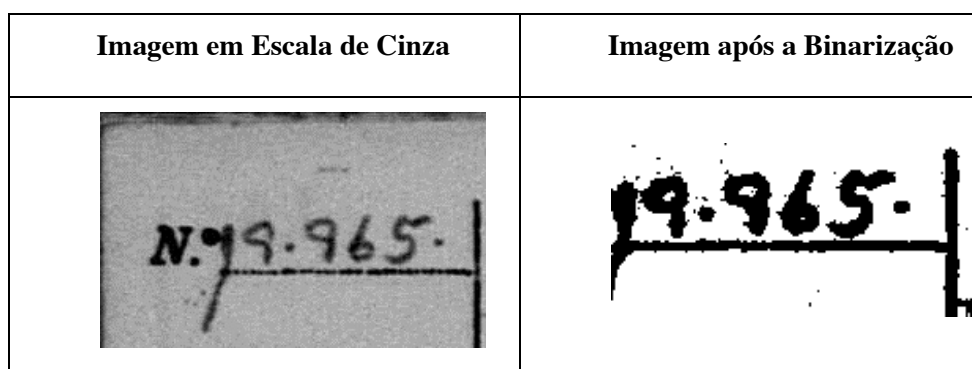


Figura 27- Ruído apresentado na imagem original e binarizada.

4.2.2 Algoritmo

Para remover esse tipo de ruído, foi desenvolvido um algoritmo que percorre toda a imagem à procura de regiões que apresentem características similares a esses “grãos”.

O algoritmo proposto percorre toda a imagem, da esquerda para a direita e de cima para baixo, utilizando uma máscara de 4x4 pixels. Os pixels sob a máscara são apagados (atribuído valor 1) quando se percebe as seguintes características no bloco analisado:

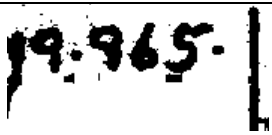
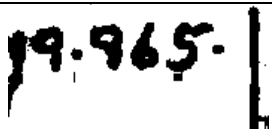


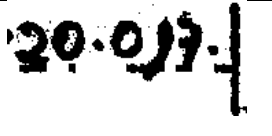
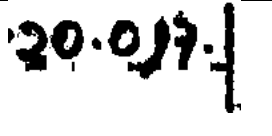
- (i) Tanto a primeira quanto a última linhas do bloco forem completamente brancas;
- (ii) Quando mais de 50% do número de pixels (mais de 8 pixels) sob a máscara for branca; e
- (iii) Quando a primeira coluna da máscara tiver pelo menos um pixel preto. Pelo mesmo motivo explicado na seção 4.2.1.

Após a aplicação do algoritmo de remoção de ruído para os “grãos-grossos”, a imagem é re-processada com o algoritmo para remover linhas horizontais novamente, para melhorar os resultados, o qual retirou partes de linhas que não haviam sido tratadas da primeira vez pela quantidade de ruídos, que fazia com que a linha se confundisse com uma parte do número.

Na Tabela 2 é possível observar exemplos de resultados alcançados nos números de registros antes e depois da aplicação de toda a etapa para retirada de ruído; a primeira coluna apresenta a imagem antes da etapa de remoção de ruídos e a segunda coluna apresenta as imagens após a aplicação do algoritmo.

O algoritmo desenvolvido para realizar a implementação de remoção de ruídos grossos, pode ser encontrado no Apêndice E desta dissertação.

Tabela 2 - Exemplos de números de Registros antes e depois de aplicar o algoritmo de redução de ruído.

Antes do Algoritmo	Depois do Algoritmo
	
	
	

4.3. Remoção de Linhas Verticais

A última etapa do pré-processamento antes do reconhecimento dos manuscritos visa eliminar a linha vertical presente nas imagens do número do registro.

O algoritmo desenvolvido para este fim é baseado na avaliação automática da distribuição dos pixels pretos ao longo da figura inteira, coluna a coluna, utilizando a técnica de *project profile* (HA *et al.*, 1995). Foi possível observar que as linhas verticais apresentam uma coluna negra que concentra um maior número de pixels pretos do que as áreas que correspondem aos manuscritos.

Neste contexto, foram analisados, para cada imagem, o valor máximo de *pixels* pretos por coluna, doravante utilizada a variável “*maxColumnSum*” para representar esse valor; e também foi analisado o valor médio de *pixels* pretos por coluna (variável “*meanColumnSum*”); e a partir desses valores, foi gerado um limiar específico para exclusão (variável “*thresholdDel*”) diferente para cada imagem processada. O limiar variável é importante para diferenciar as imagens mais claras ou que apresentam linhas mais

finas, daquelas que possuem caracteres ou linhas mais dilatados. Assim, cada coluna que possuir linhas com número de pixels maior que o limiar será apagada. Além disso, nos algoritmos também foi considerado que as colunas imediatamente anteriores e posteriores à apagada, e que possuísem o número de pixels maior que o limiar, também seriam excluídas. A Equação 1 abaixo, representa o cálculo do valor do limiar de exclusão (thresholdDel).


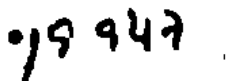
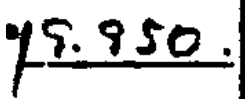
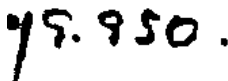
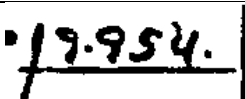
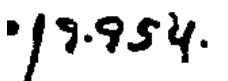
$$\text{thresholdDel} = \text{maxColumnSum} - 2 * \text{meanColumnSum}$$

Equação 1 - Cálculo do Limiar de apagamento das linhas verticais.

Os resultados deste algoritmo podem ser observados na imagem da Tabela 3, na qual podem ser analisadas as imagens antes e depois de toda a etapa de pré-processamento (remoção de linhas verticais e horizontais e redução de ruídos), o último passo antes da etapa de reconhecimento.

O algoritmo desenvolvido para a realização da remoção das linhas verticais está disponível no Apêndice F desta dissertação.

Tabela 3 - Exemplos de numeros de Registro antes e depois da aplicação de todo o estágio de pré-processamento.

Imagem Original	Imagem depois de todo o estágio de pré-processamento.
	
	
	

4.4. Reconhecimento e Classificação

Apesar de não ser o foco desta dissertação, fazem-se necessário apresentar aqui as técnicas e resultados obtidos para reconhecimento e classificação das informações manuscritas que fazem parte dos registros de óbito preprocessados pela plataforma *Thanatos* (ALMEIDA *et al.*, 2011).

4.4.1 Reconhecimento

Reconhecer símbolos manuscritos é muito mais difícil quando eles estão conectados. As alternativas encontradas para aumentar a taxa de reconhecimento global foi, tanto de dividir a palavra em símbolos ou de tentar reconhecer a palavra inteira (ou a parte conectada dela). O classificador desenvolvido para a Plataforma *Thanatos* assume que os símbolos numéricos formam um bloco contíguo e sem necessidade de segmentação futura. No caso de campos não-numéricos, há a necessidade de aplicar as técnicas de pré-processamento descritas acima, em que consiste em dividir a imagem em duas partes, tal como apresentado na Figura 28.



Figura 28 - Reconhecimento de Campo não-numérico.

A maioria dos estudos sobre reconhecimento de padrões tem como ponto central a escolha de um conjunto de características capazes de representar e discriminar as diferentes formas de classificação. Encontrar tais conjuntos de características está longe de ser uma

tarefa fácil. A literatura técnica apresenta várias técnicas para tal propósito (LIU *et al.*, 2003), (LIU *et al.*, 1997), (OLIVEIRA *et al.*, 2002), (HU, 1962), que podem ser aplicadas ao reconhecimento de texto manuscrito, as quais podem ser resumidas em três diferentes classes: Primitivas baseadas em transformada global e séries de expansão, como Fourier, Walsh, Harr, etc. (OLIVEIRA, 2007). As quais não são sensíveis a algumas transformadas globais, tais como rotação e translação. Essas técnicas necessitam de um alto poder computacional e consomem muito tempo de processamento.

Primitivas com base na distribuição estatística dos pontos. Elas incluem momentos, n-tuplas, cruzamentos e distâncias. Elas suportam distorção de forma e leva em consideração variações na forma de escrita, em alguns casos. Eles têm baixa complexidade de implementação;

Primitivas geométricas e perceptuais. Estas são as primitivas mais amplamente usadas para representar propriedades globais e locais dos caracteres. Nesta classe é possível encontrar: inflexões ascendentes e descendentes, *loops*, linha de intersecção do segmento, os pontos de término, propriedades angulares, as relações entre as inflexões, etc. Essas primitivas têm uma alta tolerância a distorções, a variação de estilo, translação e rotação.

O presente estudo segue a abordagem apresentada em (SILVA e LINS, 2011) e faz uso de um conjunto de características geométricas e de percepção extraídas no "zoneamento" da imagem. Esta técnica conta o número de loops, concavidades, inflexões horizontais e verticais, etc. O "zoneamento" pode ser visto como a divisão de um padrão complexo em vários outros mais simples. No caso de textos degradados, a preocupação deste trabalho, torna-se uma importante base de discriminação entre classes, pelo fato de a informação "real" é limitada apenas a algumas classes.

Alguns pesquisadores propõem apenas o zoneamento "empírico" (SUEN *et al.*, 1994), (LI *et al.*, 1995), (FREITAS *et al.*, 2007), no qual cada caracter é representado por um retângulo

Z, que pode assumir vários formatos diferentes, tais como os apresentados na Figura 29. Outros pesquisadores propõem métodos de zoneamento automático (RADTKE *et al.*, 2003).

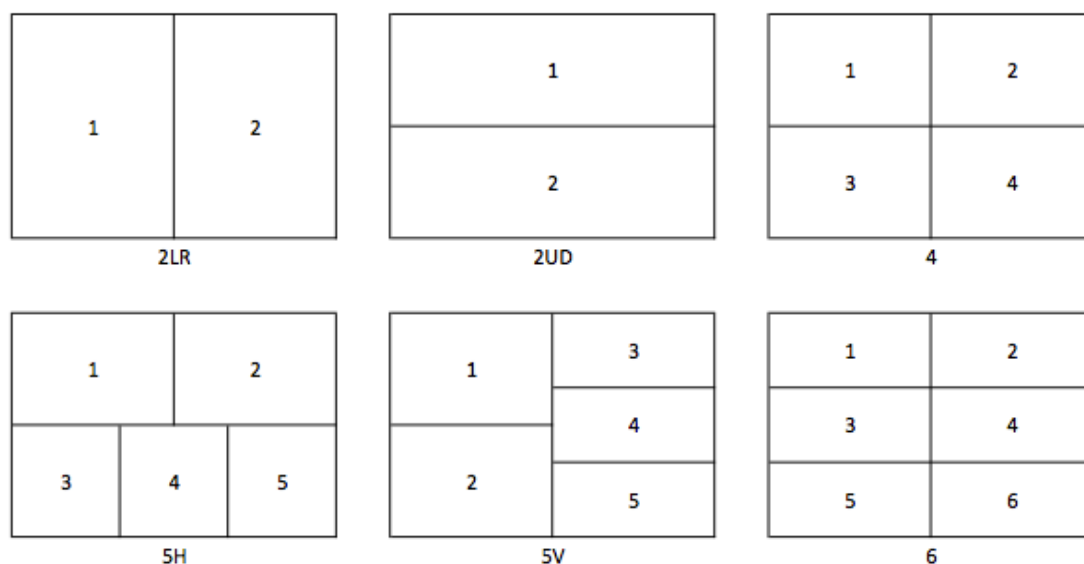


Figura 29 - Zoneamento automático de caracteres

A plataforma *Thanatos* adotou uma abordagem muito pragmática para o reconhecimento e usa as diferentes técnicas para analisar qual delas forneceu o melhor resultado para cada campo de informação.

4.4.2 Classificador de Campo Numérico

O classificador numérico da Plataforma *Thanatos* fez uso de 10 classes de dígitos numéricos, cada um deles com 2.500 imagens por classe, 2.000 delas foram obtidas da base de dados do NIST (NIST, 1968), e as outras 500 foram extraídas do acervo Mórmon-TJPE. O reconhecimento usou três subconjuntos para treinamento, teste e validação que correspondem a: 50%, e 25% a 25%, do total dos dados, respectivamente.

Dois classificadores foram testados, sendo um o MLP (Multi-layer perceptron) e o outro o RBF (radial basis functions). Esse último apresentou um desempenho ligeiramente melhor, como pode ser observado nos resultados apresentados na Tabela 4 para reconhecimento de caracteres isolados e na

Tabela 5 para o reconhecimento completo do campo.

Tabela 4 - Taxa de acerto no reconhecimento dos dígitos no campo "número de registro"

Digito	MLP	RBF
0	100%	100%
1	100%	100%
2	98%	99%
3	100%	100%
4	95%	98%
5	96%	96%
6	98%	98%
7	100%	100%
8	100%	100%
9	96%	95%

Tabela 5 - Taxa de Acerto no reconhecimento dos campos "número de registro" e "dia".

Campo	MLP	RBF
Numero do Registro	92%	95%
Dia	93%	96%

O diagrama para a Plataforma *Thanatos* (ALMEIDA *et al.*, 2011), mostrado na Figura 20, apresenta um bloco para "análise do contexto". O número de registro segue uma ordem sequencial e esta informação é levada em consideração para aumentar a taxa de reconhecimento das informações neste campo. O campo "dia" é o mesmo do Cadastro anterior ou é cíclico ascendente, começando no dia 1 e terminando em 28 ou 29 ou 30 ou 31 e retornando para o dia 1 (1, 2...28/29/30/31/1...). O uso de tais

informações permitiu uma taxa de acerto de 100% de reconhecimento dos dados para esses campos, para ambos os classificadores.

4.4.3 Classificador de Campo Não-Numérico

A fase de reconhecimento de campos com caracteres não-numéricos fez uso de duas estratégias que leva em conta a variação da informação no campo. A primeira abordagem é utilizada para campos que a variação máxima é de quatro palavras, nas quais a extração da geometria e perspectiva são utilizadas. Este é o caso para o campo "estado civil", que apresenta apenas quatro opções por gênero: "solteiro", "casado", "viúvo" e "divorciado"; o último encontrado apenas nos registros mais recentes.

No caso dos campos que possuem uma ampla gama de possibilidades, como: "Estado", "Cidade", números por escrito, nome do mês, etc; o uso de características geométricas e perspectiva rendeu resultados insatisfatórios e foi substituído pelo mecanismo de zoneamento. O resultado obtido para 300 registros de óbitos é apresentados na Tabela 6.

Tabela 6 - Taxa de Reconhecimento para campos não-numéricos.

Campo	Taxa de Acerto no Reconhecimento
Nome do Cartório	98%
Cidade do Cartório	71%
Estado do Cartório	98%
Local da Morte	31%
Números por extenso: (Horário do óbito, data da morte e data de nascimento)	69%
Cor da Pele	100%
Estado Civil	100%

É importante ressaltar que a análise automática do contexto da plataforma *Thanatos* também foi usada aqui para alcançar tão bons resultados. Por exemplo, no caso do “nome do escrivão”, cada tabelionato tem, no máximo, quatro oficiais em tal posição que permanecem ativos por longos períodos de tempo (no Brasil esse serviço é uma concessão do Estado e que se mantém ao longo da vida).

Pode-se observar que, no caso do campo "local da morte" a taxa de reconhecimento foi baixa. A análise contextual não foi implementada. A adição de um dicionário de todas as cidades do Estado de Pernambuco, juntamente com a informação de jurisdição (óbitos devem ser registrados no cartório local) ainda estão para ser implementadas. Os outros campos não-numéricos foram treinados usando as informações no banco de dados de palavras usadas em cheques bancários no Brasil, consistindo apenas na base segmentada da empresa AiLider. Da mesma forma, pode-se implementar um dicionário de médicos que trabalhavam em uma determinada região para um melhor reconhecimento dos dados.

4.5. Trabalhos Relacionados

Nesta seção são apresentados alguns artigos publicados na área de pré-processamento de imagens, mais especificamente em remoção de ruídos e linhas. Também é realizada uma análise qualitativa dos trabalhos descritos, comparativamente com as técnicas utilizadas pela Plataforma *Thanatos* e apresentadas nessa dissertação.

O artigo de AGRAWAL e DOERMANN (2011) trata de um tipo de ruído, que possui características semelhantes aos estudados nessa dissertação, por estarem espalhados por todo o documento e por possuírem dimensões maiores que o ruído de sal-e-pimenta. Nesse artigo, os autores, primeiramente, extraíram características do documento, como: área, perímetro, tamanho dos eixos e excentricidade (afastamento do eixo central) dos proeminentes componentes de texto (PTC). Após essa etapa, utilizaram redes neurais do tipo *Radio*

Function Base (RFBs) para classificar os PTCs e diferenciá-los dos não-PTCs e ruídos, inclusive dos pequenos componentes de texto, que poderiam se confundir com ruídos e serem excluídos.

A medida utilizada nesse artigo foi o cálculo da precisão do algoritmo, a qual relacionou o número de pixels de ruído removidos com o número de pixels removidos. A precisão da técnica atingiu os 86%, podendo ser comparada com a da Plataforma *Thanatos* (ALMEIDA *et al.*, 2011), que tratou ruídos com características semelhantes e não apresentou perda de informação com o apagamento de pixels de texto. Contudo a Plataforma *Thanatos* não foi testada na base de dados do artigo supra-citado, que consiste em textos manuscritos em Árabe.

Esse tipo de ruído irregular que é tratado no trabalho, chamado de *clutter noise* (AGRAWAL e DOERMANN, 2009) ou também *Stroke-like Pattern Noise* (AGRAWAL e DOERMANN, 2011), ele tem sido tipicamente classificado com regras simples para sua classificação. (OZAWA e NAKAGAWA, 1993), (WANGAND e TAN, 2001), (NEGISHI *et al.*, 1999) utilizam níveis de cinza para distinguir o primeiro plano do plano de fundo da imagem. (FAN *et al.*, 2011) considerou o tamanho, posição e vizinhança do ruído para detectar e remover esses ruídos. Em (LIANG *et al.*, 2011) se analisa a periodicidade e regularidade do ruído para removê-lo. Entretanto, não existem muitos trabalhos relacionados na remoção desses ruídos irregulares em imagens binarizadas de documentos (AGRAWAL e DOERMANN, 2009). A plataforma *Thanatos* estende as pesquisas para esse tipo de ruído, identificando-o de forma mais simples, facilitando o seu processamento.

O artigo intitulado *A Model-based Line Detection Algorithm in Documents* (ZHENG *et al.*, 2003) tem como objetivo a detecção e remoção de linhas em documentos manuscritos. O escopo desse artigo abrange a detecção de linhas paralelas em documentos, com espaçamento fixos entre elas (denominados *gaps*) e descontínuas na sua extensão. Além disso, possui

restrições na sua execução, como a consideração de que todas as linhas começam na borda da esquerda e terminam na borda da direita (ZHENG *et al.*, 2003). Um exemplo de imagem processada por este algoritmo pode ser observada na Figura 30.

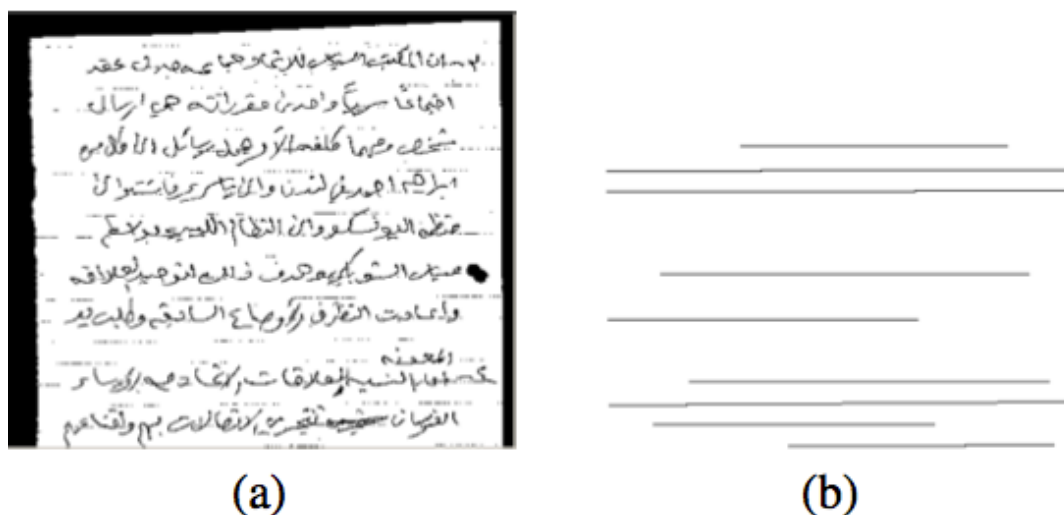


Figura 30 - Resultado da detecção de linhas fortemente descontínuas.
(a) Imagem de um documento manuscrito em árabe; (b) resultado da detecção das linhas.

Concluí-se que esse artigo pode ser comparado com o escopo utilizado na Plataforma *Thanatos* (ALMEIDA *et al.*, 2011) por se tratar de imagens de documentos manuscritos e com linhas descontínuas. Entretanto, além de abranger esses tipos de documentos, a Plataforma proposta nessa dissertação não apresenta restrições para início e fim das linhas; nem exige paralelismo e equidistância entre as linhas, possibilitando o processamento de um escopo mais variado de documentos e formulários; além de ser sensível às partes dos textos que cortam as linhas, preservando-as.

Capítulo 5.

C

onclusões e Trabalhos Futuros

A necessidade crescente de preservar informações históricas e com valor legal, presentes em documentos não digitais, bem como extrair dados desses documentos, com o objetivo de disponibilizá-los para consulta dessas informações de forma célere, bem como a realização de análise estatísticas, demográfica ou ainda estudos genealógicos, motivaram o desenvolvimento deste trabalho.

Este capítulo apresenta as principais contribuições e os aspectos mais relevantes desta dissertação. As limitações dos modelos desenvolvidos e os trabalhos futuros a serem realizados dentro desta linha de pesquisa também são destacados.

5.1. Contribuições

Thanatos é uma plataforma desenvolvida para extrair informações das imagens digitalizadas dos registros de óbito do Estado de Pernambuco. Esta dissertação está inserida no contexto dessa plataforma e foca no pré-processamento dessas imagens históricas, com o objetivo de melhorar a qualidade das imagens adquiridas, permitindo melhores resultados para as etapas posteriores de extração de características e reconhecimento automático dos manuscritos.

Mais especificamente, nesta dissertação foi apresentada uma etapa inicial para retirada de borda, alinhamento, remoção de interferência frente-e-verso e de ruído de sal-e-pimenta,

usando a Plataforma HistDoc (SILVA *et al.*, 2010) na versão 2.0 (LINS *et al.*, 2011). Após essa etapa, foi apresentada a Plataforma *Thanatos* (ALMEIDA *et al.*, 2011), proposta nessa dissertação, na qual são realizadas as etapas de segmentação dos registros e uniformização das imagens para padronizá-las devido as variações existentes em resolução e posicionamento relativos dessas imagens; extração dos campos dos registros; retirada de ruídos gerados a partir de processos de aquisição das imagens e binarização, os quais são mais espessos do que o ruído de sal-e-pimenta; e remoção de linhas.

Os processamentos relacionados à remoção de ruídos e de linhas verticais e horizontais, compõem uma importante contribuição desta dissertação, pois apresenta uma nova técnica para tratamento dessas características, identificando ruídos e linhas independentemente de sua forma ou posição, sendo um diferencial do algoritmo. Os resultados obtidos foram apresentados no *Historical Documents Imaging and Processing Workshop* (HIP), evento ligado a *International Conference on Document Analysis and Recognition* (ICDAR), realizada em setembro de 2011, na cidade de Pequim na China, na qual foi realizada apresentação oral do artigo denominado *Thanatos – Automatically Retrieving Information of Death Certificates in Brazil* (ALMEIDA *et al.*, 2011), cuja cópia encontra-se no Anexo A desta dissertação.

5.2. Trabalhos Futuros

Nesta dissertação, os estudos desenvolvidos possibilitam o reconhecimento de alguns campos manuscritos dos registros de óbito, como: número do registro, data do registro e do óbito, nome do cartório, cidade, estado, cor da pele e estado civil. Dessa forma, se propõe como trabalho futuro, o reconhecimento de outros campos do registro, o desenvolvimento de uma interface para consulta dos documentos e das informações extraídas, bem como a análise estatística de dados demográficos da população.

Outra linha de trabalho futuro proposta é estender a aplicação para outros tipos de registros civis, como de casamento e nascimento, customizando os parâmetros utilizados nos algoritmos e verificando os aspectos convergentes para se automatizar ainda mais o processo.

Também podem ser iniciados trabalhos para pré-processamento e reconhecimento dos documentos totalmente manuscritos, mais antigos, datados no início do século XX.

Neste trabalho, foi realizada uma análise qualitativa de inspeção visual dos resultados obtidos na etapa de pré-processamento e seu impacto quantitativo no reconhecimento automático dos manuscritos, bem como sua comparação com trabalhos correlatos. Assim, fica também como sugestão de trabalho futuro, a realização de uma análise quantitativa dos resultados obtidos, medindo o percentual de acerto da técnica utilizada, bem como a sua comparação com outras técnicas que se propõem a pré-processamentos semelhantes de retirada de ruídos e linhas horizontais e verticais.

Referências Bibliográficas

(AGRAWAL e DOERMANN, 2009)

AGRAWAL, M.; DOERMANN, D. Clutter noise removal in binary document images. Proceedings of 10th Int'l Conference on Document Analysis and Recognition (ICDAR 2009). New Jersey: IEEE Computer Society Conference Publishing Services (CPS). 2009. p. 556-560.

(AGRAWAL e DOERMANN, 2011)

AGRAWAL, M.; DOERMANN, D. Stroke-like Pattern Noise Removal in Binary Document Images. Proceedings of 11th International Conference on Document Analysis and Recognition. New Jersey: IEEE Computer Society Conference Publishing Services. 2011. p. 17-21.

(ALMEIDA et al., 2011)

ALMEIDA, A. B. S.; LINS, R. D.; SILVA, G. F. P. *Thanatos* - Automatically Retrieving Information from Death Certificates in Brazil. Proceedings of the 2011 Workshop on Historical Documents Imaging and Processing. New York: ACM Press. 2011. p. 146-153.

(AVILA e LINS, 2004)

AVILA, B. T.; LINS, R. D. A new algorithm for removing noisy borders from monochromatic documents. ACM Symposium on Applied computing (SAC). New York: ACM Press. 2004. p. 1219-1225.

(AVILA e LINS, 2005)

ÁVILA B. T.; LINS, R. D. A Fast Orientation and Skew Detection Algorithm for Monochromatic Document Images. ACM International Conference on Document Engineering, 2005. ACM Press. 2005. p.118 - 126

(CASTLEMAN,1996)

CASTLEMAN, Kenneth. R. Digital Imaging Processing. 1 ed. New Jersey: Prentice Hall, Inc.1996.

(FAN et al., 2001)

FAN, K. C.; WANG, Y. K.; LAY, T. R. Marginal noise removal of document images. Proceedings of 6th Int'l Conference Document Analysis and Recognition (ICDAR'01). New Jersey: IEEE Computer Society Conference Publishing Services (CPS). 2001. p. 317-321.

(FORMIGA e LINS, 2009)

FORMIGA, A. A.; LINS, R. D. Efficient Removal of Noisy Borders of Monochromatic Documents. In: International Conference on Image Analysis and Recognition 2009. Springer Verlag, LNCS v.5627. 2009. p.158 – 167.

(FREITAS et al., 2007)

FREITAS, C. O. A.; OLIVEIRA, L. S.; AIRES, S. B. K.; BORTOLOZZI, F. Zoning and metaclasses for character recognition. In: ACM– SAC 2007. 2007. P. 632-636.

(GENEALOGICAL, 2008)

GENEALOGICAL SOCIETY OF UTAH. Genealogical Society of Utah. 2008. Disponível em: <<http://www.gensocietyofutah.org/default.asp>>. Acesso em: novembro de 2010.

(GONZALEZ e WOODS, 2010)

GONZALEZ, R. C.; WOODS, R. E. Processamento Digital de Imagens. 3 edição. ed. São Paulo: Pearson Prentice Hall, 2010.

(HA et al., 1995)

HA, J.; HARALICK, R.M.; PHILLIPS, I.T. Recursive X-Y cut using bounding boxes of connected components. In: Document Analysis and Recognition. Proceedings of the Third International Conference on Document Analysis and Recognition. 1995. P. 952-955.

(HU, 1962)

HU, M. Visual pattern recognition by moment invariants. In: IEEE Transactions on Information Theory, 1962. 8(2):179-187.

(KASTURI et al., 2002)

KASTURI, R.; O'GORMAN, L.; GOVINDARAJU, V. Document image analysis: A primer. SADHANA - Academy Proceedings in Engineering Sciences, p. 3-22, 2002.

(LI et al., 1995)

LI, Z. C.; SUEN, C. Y.; GUO, J. A Regional Decomposition Method for Recognizing Handprinted Characters. In: IEEE Transactions on Systems, Man, and Cybernetics, N. 25. 1995. p. 998-1010.

(LIANG et al., 1994)

LIANG, S.; AHMADI, M.; SHIRIDHAR, M. A morphological approach to text string extraction from regular periodic over-lapping text/background images. Proceedings of IEEE Int'l Conference Image Processing (ICIP-94). New Jersey: IEEE Computer Society Conference Publishing Services (CPS). 1994. p. 144-148.

(LINS, 2009)

LINS, R. D. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. ICIAR 2009, LNCS v.5627, p.844 - 854 Springer Verlag, 2009.

(LINS et al., 2010)

LINS, R. D.; OLIVEIRA, D. M.; TORREAO, G.; FAN, J.; THIELO, M. Correcting Book Binding Distortion in Scanned Documents. ICIAR 2010. Springer Verlag, v.LNCS. 2010.

(LINS et al., 2011)

LINS, R. D.; SILVA, G. F. P.; FORMIGA, A. A. HistDoc v. 2.0 - Enhancing a Platform to Process Historical Documents. In: Workshop on Historical Documents Imaging and Processing: beijing, China: 2011. Proceedings of the 2011 Workshop on Historical Documents Imaging and Processing. New York: ACM Press, 2011. P. 169-176.

(LIU et al., 2003)

LIU, C.; NAKASHIMA, K.; SAKO, H.; FUJISAWA, H. Handwritten digit recognition: benchmarking of state-of- the-art techniques. Pattern Recognition, 2003. 36(10):2271-2285.

(LIU et al., 1997)

LIU, C.; LIU, Y.; DAI, R. Preprocessing and statistical/ structural feature extraction for handwritten numeral recognition. In: Progress of Handwriting Recognition. A.C. Downton and S. Impedovo eds., World Scientific, 1997.

(MATTOS et al., 2008)

MATTOS , G. G.; FORMIGA, A. A.; LINS, R. D.; MARTINS, F. M. J. BigBatch: a document processing platform for clusters and grids. Proceedings of ACM-SAC 2008. [S.l.]: ACM Press. 2008. p. 434-441.

(NEGISHI, et al., 1999)

NEGISHI, H.; KATO, J.; HASE, H.; WATANABE, T.. Character extraction from noisy background for an automatic reference system. Proceedings of 5th Int'l Conf. Document Analysis and Recognition (ICDAR'99). New Jersey: [s.n.]. 1999. p. 143-146.

(NIST, 1968)

NIST Scientific and Tech. Databases. 1968. Disponível em: <http://www.nist.gov/srd/index.cfm>. Acessado em: março de 2011.

(OLIVEIRA, 2007)

OLIVEIRA, HELIO M. 2007. Análise de Fourier e Wavelets. 1 ed. Recife: Editora da Universidade Federal de Pernambuco.

(OLIVEIRA et al., 2002)

OLIVEIRA, L.; SABOURIN, R.; BORTOLOZZI, F.; SUEN, C. Automatic recognition of handwritten numerical strings: A recognition and verification strategy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2002. 24(11):1438-1454.

(OZAWA e NAKAGAWA, 1993)

OZAWA, H.; NAKAGAWA, T. A character image enhancement method from characters with various background images. *Proceedings of 2nd International Conference Document Analysis and Recognition (ICDAR'93)*. New Jersey: IEEE Computer Society Conference Publishing Services (CPS). 1993. p. 58-61.

(PRESIDENCIA, 1968)

PRESIDÊNCIA DA REPÚBLICA. Presidência da República - Casa Civil. lei nº 5.433 de 8 de maio de 1968, 8 maio 1968. Disponível em: <<http://www.planalto.gov.br/ccivil/leis/15433.htm>>. Acesso em: 22/12/2010.

(PRESIDENCIA, 2006)

PRESIDENCIA DA REPÚBLICA. Presidencia da República - Casa Civil. Decreto nº 1.799 de 30 de janeiro de 1996, 30 jan. 2006. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/Antigos/D1799.htm>. Acesso em: 22/12/2010.

(RADTKE et al., 2003)

RADTKE, P. V. W.; OLIVEIRA, L.S.; SABOURIN, R.; WONG, T. Intelligent Zoning Design Using Multi-Objective Evolutionary Algorithms. *Proceedings of 7th Int'l Conference on Document Analysis and Recognition (ICDAR 2003)*. New Jersey: IEEE Computer Society Conference Publishing Services (CPS). 2003. p.824-828.

(SHARMA, 2001)

SHARMA, G. Show-through cancellation in scans of duplex printed documents. *IEEE Transaction Image Processing*. 2001. p. 736-754.

(SILVA et al., 2010)

SILVA, F. P.; LINS, R. D.; SILVA, J. M. M. HistDoc - A Toolbox for Processing Images of Historical Documents. *ICIAR 2010*. Springer Verlag. 2010. p. 1-11.

(SILVA et al., 2010)

SILVA, F. P.; LINS, R. D.; SILVA, J. M. M.; BANERGEE, S.; KUCHIBHOTLA, A.; THIELO, M. Enhancing the Filtering-Out of the Back-to-Front Interference in Color Documents with a Neural Classifier. *ICPR 2010*. IEEE Press. 2010. P. 2415-2419.

(SILVA e LINS, 2011)

SILVA, F. P.; LINS, R. D. An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. In: Proceedings of 11th Int'l Conference on Document Analysis and Recognition (ICDAR 2011). ICDAR 2011, Beijing. IEEE Computer Society Conference Publishing Services (CPS). 2011.

(SUEN et al., 1994)

SUEN, C. Y.; GUO, J.; LI, Z. C. Analysis and Recognition of Alphanumeric Handprints by parts. In: IEEE Transactions on Systems, Man, and Cybernetics. N. 24. 1994. p. 614-631.

(WANGAND e TAN, 2001)

WANGAND, Q.; TAN, C. L. Matching of double-sided document images to remove interference. Proc. Comp. Vision and Patt. Recognition (CVPR'01). New Jersey: IEEE Computer Society Conference Publishing Services (CPS). 2001. p. 1084-1089.

(ZHENG et al., 2003)

ZHENG, Y.; LI, H.; DOERMANN, D. A model-based line detection algorithm in documents. Proceedings of 7th Int'l Conference on Document Analysis and Recognition (ICDAR 2003). New Jersey: IEEE Computer Society Conference Publishing Services (CPS). 2003. p. 44-48.

Anexo A – Artigo Publicado

Título do Artigo: *Thanatos*– Automatically Retrieving Information from Death Certificates in Brazil.

Conferência: Workshop on Historical Document Imaging and Processing.

Local da Conferência: Beijing – China.

Data da Conferência: 16 e 17 de setembro de 2011.

Thanatos

Automatically Retrieving Information from Death Certificates in Brazil

Alessandra B. S. Almeida
PPGEE - U.F.PE.
Recife-Pernambuco-Brazil
+55 81 9422-4537
alessandrabsa@gmail.com

Rafael Dueire Lins
U.F.PE.
Recife-Pernambuco-Brazil
+55 81 8896-0698
rdl.ufpe@gmail.com

Gabriel de F. Pereira e Silva
PPGEE - U.F.PE.
Recife-Pernambuco-Brazil
+55 81 8803-8715
gfps.cin@gmail.com

ABSTRACT

Death certificates provide important data such as *causa mortis*, age of death, birth and death places, parental information, etc. Such information may be used to analyze not only what caused the death of the person, but also a large number of demographic information such as internal migration, the relation of death cause with marital status, sex, profession, etc. Thanatos is a platform designed to extract information from the Death Certificate Records in Pernambuco (Brazil), a collection of “books” kept by the local authorities from the 16th century onwards. The current phase of the Thanatos project focus on the books from the 19th century.

Categories and Subject Descriptors

I.4.9 [Image Processing and Computer Vision]: Applications.

General Terms

Algorithms, Document image analysis.

Keywords

Document processing, death certificates, image processing, historical documents.

1. INTRODUCTION

The Mormon Church teaches that their members are responsible to be baptized for dead loved ancestors. If a person dies having never been baptized in this life, a Mormon relative can be baptized in his place. Then the dead person may have a chance after death to believe the gospel, repent, and be saved. Joseph Smith, the founder of Mormonism, taught that seeking the dead in this manner is a Mormon’s greatest responsibility [1]. Such duty motivates the Mormon Church to have genealogical records all over the world. In the state of Pernambuco (Brazil), the Mormon Church celebrated an agreement with the Judiciary Power of the State (Tribunal de Justiça de Pernambuco - TJPE) to digitize all

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HIP’2011, Sep. 16-17, 2011, Beijing, China.
Copyright 2011 ACM 1-58113-000-0/00/0010...\$10.00.

“books” of Civil Records kept by the local authorities. The complete Mormon-TJPE file encompasses over one million records and includes birth, wedding, and death certificates from all over the state of Pernambuco. The oldest records are from the 16th century just after Trento Council (1543 to 1563), when the Catholic Church that it was mandatory for people to have birth, christening, wedding, and death certificates. Figure 1 presents the platform used in the digitization of such records. It is a camera-based platform with controlled illumination. The height of the camera is adjustable to allow for larger volumes. The images were made available in grey-scale (8-bits) with an (approximate) resolution of 200 dpi. They were obtained between 1998 and 2000.



Figure 1 – Digitalization platform used by the Mormon Church to acquire the images of the death certificates in Pernambuco (Brazil).

Figure 2 presents an example of a document from the Mormon-TJPE file. It is a civil wedding certificate dated from 1948 that took place at the city of Recife.

The information contained in Death Certificates goes far beyond genealogical data. It may be used to analyze not only the *causa mortis* of an individual, what caused the death of the person, but also a large number of demographic information such as inter and extra-regional migration, the relation of death cause with the person’s age, marital status, sex, profession, the way diseases were transmitted, the quality health assistance, sanitary

conditions, infant mortality, etc. The correlation of such information may provide an invaluable testimony of a society in many different aspects.

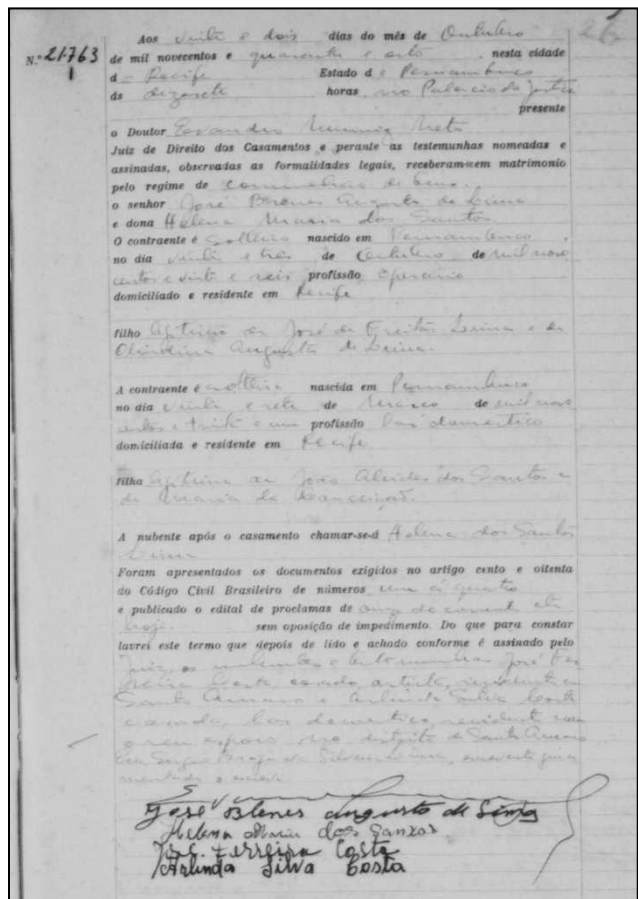


Figure 2 – Wedding certificate from the Mormon-TJPE file.

In Greek mythology, *Thanatos* (in Greek, *Θάνατος*—"Death") was the personification of death [2]. This paper presents the *Thanatos* platform, designed to extract data from the Death Certificate Records in Brazil. The current phase of the *Thanatos* project focuses on the books from the 19th century. The platform pre-processes the image to perform image binarization, border removal, skew correction, and content extraction.

2. The *Thanatos* Platform

The last two decades has witnessed an important growth in all fields of document engineering. Several new techniques, algorithms, tools, and platforms have been developed for document acquisition, processing, enhancement, and content extraction. New challenges appear every day in this area as the digitalization pace moves faster.

The analysis of the documents in the Mormon-TJPE bequest shows that several of them present several kinds of noise, which according to the classification in reference [3], are either physical (back-to-front interference, paper aging, faded ink, stains, folding marks, torn-off regions, etc.) or digitalization (borders, skew, salt-and-pepper, etc.). In the specific case of the death certificates from the file the kinds of noises found were similar, as one may

observe in the image shown in Figure 3, which presents an example of a two-page image with four death certificates of such a book from the 19th century. The book of records had little change over the centuries. The predominant ones from late 19th and the 20th centuries up to the late 1990s, when notaries were informatized, were pre-printed with fields to be filled in (that is also the case of the wedding certificate in Figure 2).

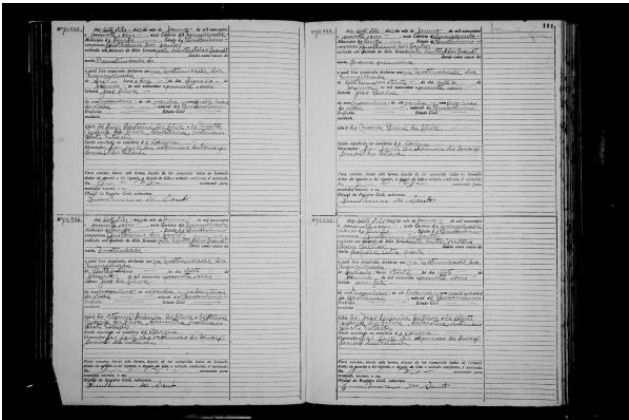


Figure 3 – Two-page image of a “book” showing four Death Certificates acquired by the platform of Figure 1

Although hard binding warp was observed in many of the images of documents the authors opted not to address such problems for two reasons:

- The margin between the binding and the start of textual information is wide enough not to warp significantly textual areas.
- In most of cases, only certificate number was affected.

2.1 Preprocessing

As already mentioned, the images from the Mormon-TJPE file were acquired with a digitalization platform such as the one presented in Figure 1. One may observe that there is a fixed basis in Black background plane where lays the document to be photographed. The camera is held at a column perpendicular to the basis in such a way as to move up and down as parallel as possible to the background plan. Before the start of the image acquisition process there is a camera calibration phase which performs diaphragm aperture control to check the intensity of the lightning that reaches the sensor, gray scale checking, focus adjustment, heating of illumination lamps, etc.

In spite of all the care taken in image acquisition and initial calibration, there are variations from one set of images to another that yield significant distortions that complicate image processing, segmentation, information extraction and automatic recognition. As one may observe in Figure 3, the image shows the presence of black borders (that correspond to the mechanical support where the book rested), a small skew angle (generally of less than 3 degrees), and some salt-and-pepper noise possibly due to dust and small stains in the paper.

The closer analysis of the documents of the death certificates showed that many of the documents exhibit a weak back-to-front interference (also known as bleeding [4] or show-through [5]). The new version [22] of the HistDoc platform [6] was used to

remove such noise by applying the algorithm described in reference [7], which assesses the intensity of the interference for tuning the global threshold algorithm. After the removal of the back-to-front interference the image is binarized. Besides the problems already listed neither the illumination of the document is completely even nor is the resolution uniform as there is a variation of the height of the camera. Both factors cause difficulties for document processing and information extraction. A general scheme of the *Thanatos* platform and its integration with the HistDoc platform is shown in Figure 4.

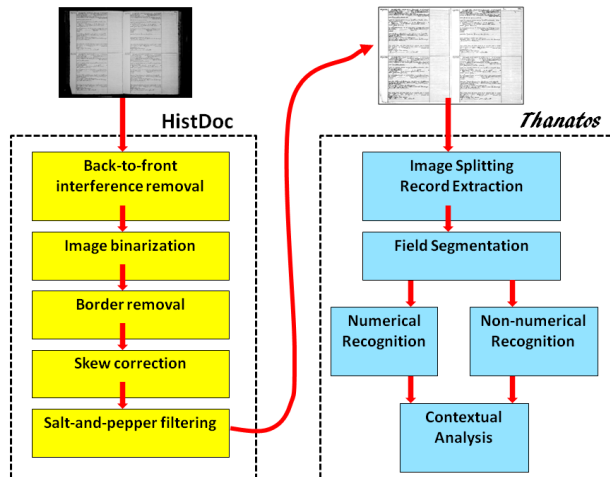


Figure 4 – Block diagram of the *Thanatos* platform and its integration with HistDoc.

The removal of the black surrounding border is performed by using the algorithm presented in reference [8]. Skew correction is performed by the algorithm described in reference [9], and finally salt-and-pepper filtering is applied. The new version of HistDoc works similarly to BigBatch [10] in either operator-driven mode, or stand-alone batch mode that also makes possible working in clusters and grids. Figure 5 shows the document in Figure 3 after being processed with the filters of HistDoc [6], [22] and BigBatch, followed by cropping.

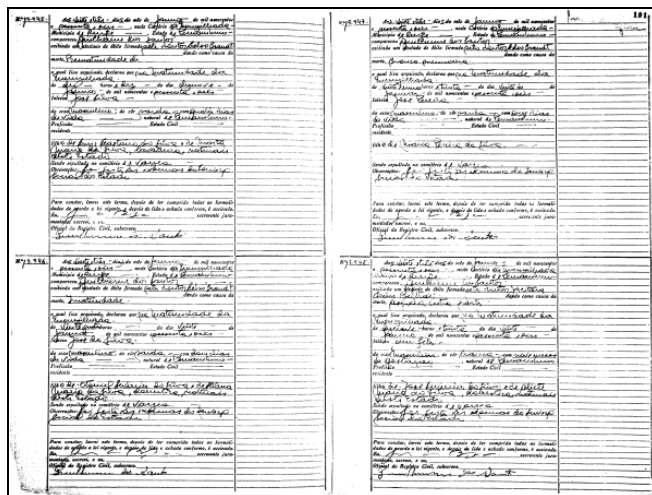


Figure 5 – Image after processed by HistDoc and BigBatch.

The next step in the document processing is performed by the core of the *Thanatos* platform that splits the image either in two (for the first page of the book) or in four (for the remaining pages) death certificates. This operation is performed by automatically seeking the central line of the image, which corresponds to the volume spine, yielding the left and right page images with two certificates each. Then, a search is performed horizontally to find the central line of the document. Due to the image acquisition process, such line does not correspond to the median line in the image. This task is performed by getting a central area from the page and performing a horizontal projection profile in it. The desired line is slightly thicker than the other blank lines. This method works satisfactorily and yields the upper and lower images, an example of which is shown in Figure 6.

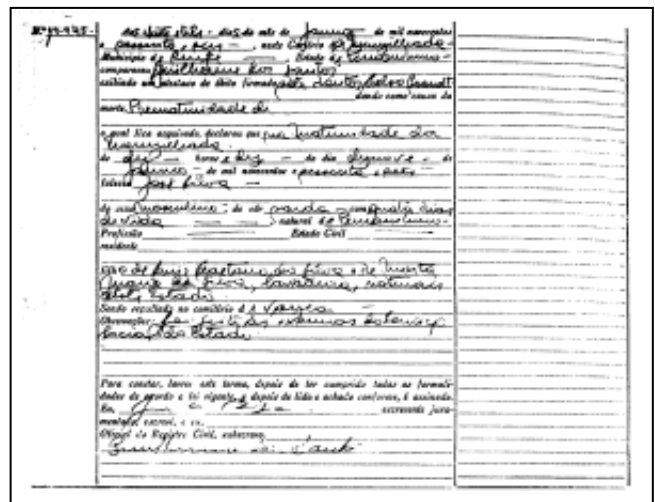


Figure 6 – Death Register after segmentation by the *Thanatos* platform.

The direct segmentation of the certificate image such as the one in Figure 6, has been unsuccessful in automatic content extraction due to book binding warp that increases the complexity of such a task. Image uniformization has to be performed because the four certificates on an image (top/bottom-left, top/bottom-right) have different “shapes”. Image pre-processing is performed to automatically search for “field” delimiters, as shown in Figure 7.

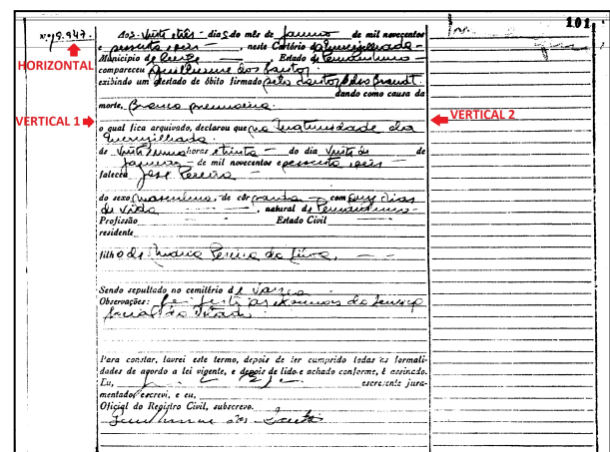


Figure 7 – Finding reference points for content segmentation.

Several problems were found in the development of the standardization process. As one may observe in Figure 7, the vertical lines have several missing points due to bad quality typographical printing and also during image binarization. Book binding warp in some cases also caused skew in the horizontal lines and this made boundary line detection difficult. Besides those factors there is a additional problem: certificate books also varied in size. Some of them were smaller than the 1,900 by 1,470 pixels as the horizontal and vertical dimensions adopted as the minimum for a register. In such cases, white pixel padding was performed at the bottom and right-hand side margins of the images. After the register standardization process all death registers had the same dimensions (1,900 x 1,470 pixels) and the same relative positions between different images for the same information fields, allowing segmentation to take place.

2.2 Segmentation

The aim of this phase is to automatically extract the information from each field in the death certificate form. Fields were hand-written and often they overlapped the fields. The ten first fields of the death certificate are described below. (The underlined text corresponds to the information the notary wrote in the field of the form):

- **N°** (Register number) – placed at the top of the left margin of the register. It conveys numerical information only. Example: N° 19.945.
- **Data** (Date) – the date is written in words and the information is filled in three fields for *day*, *month*, and *year* in this sequence. Example: Aos vinte e três dias do mês de janeiro de mil novecentos e sessenta e seis (At the twenty three days of the month of January of one thousand nine hundred and sixty six).
- **Nome do cartório** (Notary name) – this field holds the name of the place where the notary office was found. Example: neste cartório da Encruzilhada (at this notary office at Encruzilhada).
- **Município do Cartório** (City of the notary office) – Example: município de Recife (at the city of Recife).
- **Estado do Cartório** (State of the notary office) - Example: Estado de Pernambuco (State of Pernambuco).
- **Nome do Declarante** (Name of declarer) – Name of who attended the office to inform the death. Example: compareceu Guilherme dos Santos (attended Guilherme dos Santos).
- **Nome do Médico** (Name of the Medical Doctor) – Name of the M.D. who checked the death. Example: exibindo um atestado de óbito firmado pelo doutor José Ricardo (showing a death declaration signed by doctor José Ricardo).
- **Causa mortis** – Specifies the reason of the death in the declaration from the M.D. Example: dando como causa da morte edema pulmonar, o qual fica arquivado (that states as *causa mortis* lung edema, which is archived).

For better understanding of the correspondence of the information above the fields, they were translated on a word-to-word basis between (Brazilian) Portuguese and English.

The automatic information extraction was performed by masks for each of the fields of the death certificate. Each field mask has variable size and is placed onto a matrix of masks (MM) as a two-dimensional array, with the following lay-out:

$$MM = \begin{bmatrix} X0_1 Y0_1 W_1 H_1 \\ X0_2 Y0_2 W_2 H_2 \\ \vdots \vdots \vdots \vdots \\ X0_n Y0_n W_n H_n \end{bmatrix}$$

In which ‘n’ is the n^{th} element from MM matrix ‘X0’ is the initial coordinate of the X axis, ‘Y0’ is the initial coordinate of the Y axis, ‘W’ is the width of the n^{th} field and ‘H’ is the height of the n^{th} field.

The creation of the MM matrix allows cropping of each image field. As a result of the application of the matrix of masks to a (standardized) death register one obtains the image for the fields of the death certificate as shown in Table 1 (the images correspond to the item by item examples provided above).

Mask n°	Field Name	Image
01	Registry Number	
02	Date	
03	Month	
04	Year	
05	Name/Place of notary	
06	City	
07	State	
08	Name of declarer	
09	Name of the M.D. who certified the death.	
10	Cause of death	

Table 1 – Segmented fields of death records.

The direct recognition of the handwritten fields above has shown to provide very bad results. Several problems were identified in the images that had deleterious effects in the accuracy of character recognition. They were: the horizontal lines of the gaps to be filled in, the vertical lines that set the margins of the record, remaining stains and physical noise within the image, and too

narrow characters. To solve such problems, each of the records has to go through a new preprocessing stage. These problems can be observed in the column “image” in Table 1.

2.3 Recognition Preprocessing

The first preprocessing step for the segmented images is to remove the printed horizontal lines. To solve this problem, an algorithm was developed to sweep the field image from left to right and from top to bottom, scanning it in blocks of 8x8 pixels to identify regions which resemble a horizontal line, even if it has “holes” in it as found in several images. If the first and last horizontal lines of the block are all white the block under the mask is replaced by white pixels. If most pixels under the mask are white and the first column has at least one black pixel the area must remain unchanged to avoid erasing areas that are the starting points of character strokes, as shown Figure 8.

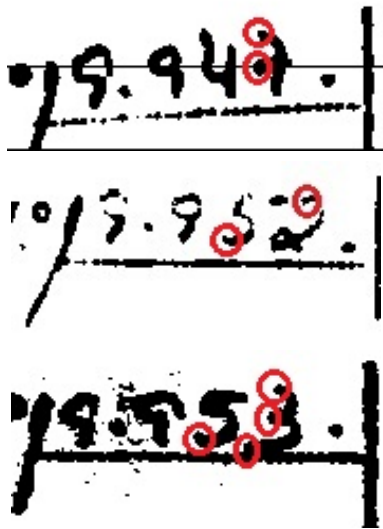


Figure 8 - Areas of the image that may be erased if the first column of the block is not checked.

Figure 9 illustrates the algorithm for removing horizontal lines, displaying the 8x8 block of pixels, comparing their size and the image of a block being processed and the example of a block that will be erased with this technique.

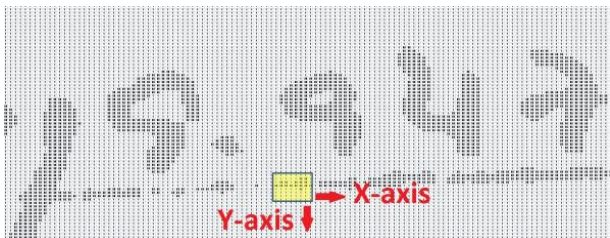


Figure 9 – Example image under the line removal algorithm.

The result of the same image of the Register Number field (Nº 19.947) before and after being processed by the algorithm proposed here can be seen in Figure 10.



Figure 10 – Register number before and after horizontal line removal.

The second step of the recognition preprocessing is noise reduction. Such noise is either physical (stains, mould, paper aging, dust, etc) or introduced during digitalization (dust on the camera lens or on the book during image acquisition), or even introduced during binarization. Although the whole image had the salt-and-pepper noise filtered out in the last phase of preprocessing performed by HistDoc, field-specific filtering enhances the quality of the image and increases the correct recognition rate. Figure 11 presents an example of such a field image in the original gray-scale document and after binarization.

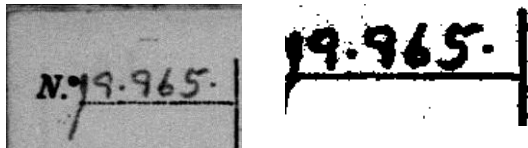


Figure 11 - Noise present in the original and binarized images.

As one may observe in the image of Figure 11 the size of the noise “grains” is much coarser than the salt-and-pepper one. To remove such noise the whole image is scanned from left to right and from top to bottom, with a 4x4 pixel mask. The pixels under the mask are painted white if either the first and last lines are white, if over 50% of the number of pixels under the mask is white, or if the first column of the mask has at least one black pixel. After the application of coarse-grain noise removal algorithm, the image is re-processed with the algorithm to remove horizontal lines to improve the results.

Before Algorithm	After Algorithm

Table 2 – Examples of Registry number before and after applying the noise reduction algorithm.

The third and last step of the recognition preprocessing aims to remove the vertical line in the images of the record number. The algorithm developed for this purpose is based on the automatic evaluation of the distribution of black pixels along the projection profile of the whole figure. It was possible to

observe that vertical lines have a greater number of black pixels per column than the areas that correspond to characters. In this context, we use, for each image, the maximum value of black pixels per column ($maxColumnSum$) and the mean value ($meanColumnSum$), per column, for the entire image in both cases to generate a specific threshold deletion ($thresholdDel$) for each image processed. The variable threshold is important to differentiate the clearer images (or with fine lines), from the ones with more dilated characters. Additionally, the columns immediately anterior and posterior to the columns with the number of pixels greater than the threshold value are also deleted. The equation below represents the calculation of the deletion threshold value.

$$thresholdDel = maxColumnSum - 2 * meanColumnSum$$

The results for this algorithm are shown in the images of Table 2, where one may observe images before and after preprocessing process (vertical and horizontal line removal and noise reduction), the last step before the information recognition step.

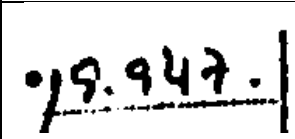
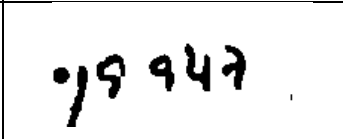
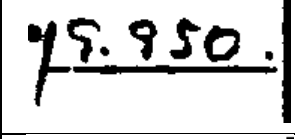
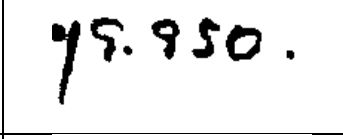
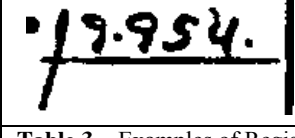
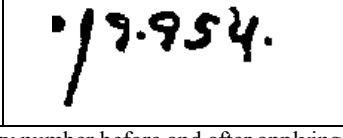
Original Image	Image after the recognition preprocessing stage
	
	
	

Table 3 – Examples of Registry number before and after applying the entire Recognition Preprocessing Stage.

2.4 Recognition and Classification

The first strategy used for information recognition was to transcribe the fields using the commercial OCR tool ABBYY FineReader 10 Professional Editor [11]. The results obtained were zero correct recognition for all fields, including even the numerical ones. Such disappointing results forced the authors to develop a recognition tool for the *Thanatos* platform.

Recognizing handwritten symbols is much harder when they are connected. The alternatives one finds to increase the global recognition rate is either to split symbols or to try to recognize the whole word (or the connected part of it). The classifier developed for the *Thanatos* platform assumes that numerical symbols form a contiguous block and no further segmentation is needed. In the case of non-numerical fields, there is the need of applying the preprocessing techniques described above, in which the image is split into two parts as presented in Figure 12.

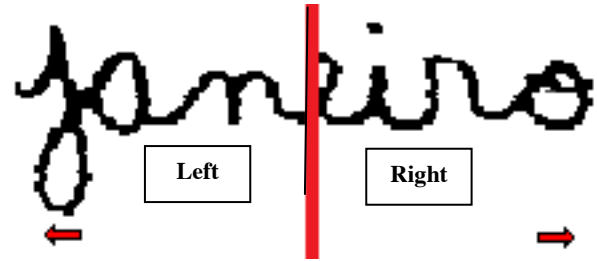


Figure 12 – Non-numerical field recognition

The majority of studies using pattern recognition have as central theme the selection of a set of features capable of representing and discriminating between the different shapes to be classified. To find such set of features is far from being an easy task. The technical literature presents several techniques for such a purpose [12][13][14][15] which when applied to handwritten text recognition may be summarized into three different classes:

- Primitives based on global transforms and expansion series, such as Fourier, Walsh, Harr, etc provide invariants to some global transformations such as rotation and translation. These techniques require greater processing power and are time consuming.
- Primitives based on the statistical distribution of the points. They include moments, n-tuples, crossing and distances. They allow for shape distortion and take into account hand written style variation, in some cases. They have low implementation complexity.
- Geometrical and perceptual primitives. These are the primitives more widely used to represent global and local properties of characters. In this class one finds: ascending and descending strokes, loops, line-segment intersection, ending points, angular properties, relations between strokes, etc. These primitives have a high tolerance to distortions, style variation, translation and rotation.

The current study follows the approach in reference [16] and makes use of a set of geometrical and perceptual features extracted from “zoning” the image. This technique counts the number of loops, concavities, horizontal and vertical strokes, etc. “Zoning” may be seen as splitting a complex pattern in several simpler ones. In the case of degraded texts, the concern of this paper, this becomes an important discrimination basis amongst classes, as the “real” information is limited only to some classes. Some researchers propose only the “empirical” zoning [17][18][19], in which each character is represented by a rectangle Z, that may assume several different formats, such as the ones presented in Figure 13. Other researchers propose methods of automatic zoning [20].

This work adopted a very pragmatic approach to recognition and used the different techniques to analyze which one provided the best result for each information field.

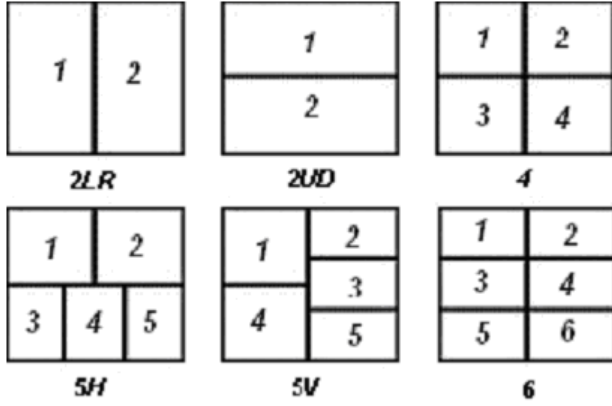


Figure 13 – Automatic character zoning

A. Numerical Field Classifier

The *Thanatos* numerical field classifier made use of 10 classes of numerical digits, each of them with 2,500 images per class, 2,000 of them were obtained from the NIST database [21], and the other 500 were extracted from the documents from the The Church-TJPE files. The recognition used three subsets for testing, training and validation that correspond to: 50%, 25% and 25%, of the total data respectively. Two classifiers were tested: a MLP (Multi-layer perceptron) and a RBF (Radial Basis Function). The latter presented a slightly better performance, as may be observed in the results presented in Table 4 for isolated character recognition and in Table 5 for the complete recognition of the field.

Digit	MLP	RBF
0	100%	100%
1	100%	100%
2	98%	99%
3	100%	100%
4	95%	98%
5	96%	96%
6	98%	98%
7	100%	100%
8	100%	100%
9	96%	95%

Table 4 – Correct recognition rates for digits in field Register Number.

Field	MLP	RBF
Numerical Register	92%	95%
Day	93%	96%

Table 5 – Correct recognition rates for fields Register Number and Day of a month.

The diagram for the *Thanatos* platform shown in Figure 4 includes a block for context analysis. The Register Number follows a sequential order and this information is taken into account to increase the correct recognition rate of the information in this field. The “Day” field is either the same as the previous Register or is (28/29/30/31 cyclically) incremented. The use of such information allowed a 100% correct data recognition for these fields, for both classifiers.

B. Non-Numerical Field Classifier

Character recognition for non-numerical fields makes use of two strategies that take into account the variance of the information in the field. The first approach is used for fields where the maximum variance is four words in which the extraction of geometrical and perceptive information is used. This is the case for the field “estado civil” (marital status) that presents only four options per gender: “solteiro” (single), “casado” (married), “viúvo” (widower), and “separado” (divorced); the latter one found only in the most recent registers. In the case of other fields that have a wider range of possibilities, such as “Estado” (State), “Cidade” (city), numbers in writing, name of the month, etc. the use of geometrical and perceptive features yielded unsatisfactory results and was replaced by the zoning mechanism.

The results obtained for 300 death certificates are presented in Table 6.

Field	Correct recognition rate
Name of Notary	98%
City of the Notary	71%
State of the Notary	98%
Place of death	31%
Numbers in writing: (Time of obit, date of death, date of birth)	69%
Color of skin	100%
Marital status	100%

Table 6 – Recognition rate for non-numerical fields

It is important to stress that the automatic contextual analysis of the *Thanatos* platform was also used here to Grant such good results. For instance, in the case of the Name of Notary, each office has at maximum four officers in such position who remain active for long periods of time (in Brazil that service is a concession of the State and that is a life-long position). One may observe that in the case of the field “Place of death” the recognition rate was low. The contextual analysis was not implemented. The addition of a dictionary of all the cities of the State of Pernambuco, together with the information of jurisdiction (the obits within a region must be registered in the closest possible notary office) are yet to be implemented. The other non-numerical fields were trained using the information in the database of words used in bank cheques in Brazil. Similarly, one may implement a dictionary of medical doctors that worked at a certain region for better data recognition.

3. Conclusions and lines for further work

Death certificates provide important anthropological, sociological, and medical information of populations.

This paper presents the *Thanatos* platform, a platform designed to extract information from death certificates from the Mormon-TJPE files from Pernambuco, Brazil. The platform introduced several new strategies for content extraction and recognition that was able to retrieve information at a very high accuracy rate. It is important to stress that even if the correct recognition rate is lower than 100% the information may be useful for demographic studies.

The current version of the platform focused on the books from late 19th century to close to the year 2000, when notaries used pre-printed books to fill information in the fields. Notary books from the earlier periods have a much higher degree of difficulty for processing and information extraction as there are no patterns to guide the register and field extraction. Addressing such documents is left for further work.

4. Acknowledgements

The authors are grateful to the The Church of Jesus Christ of Latter-day Saints (Family Search International) for the initiative of digitizing the death certificate records of Pernambuco (Brazil) and to Tribunal de Justiça de Pernambuco (TJPE) to allow the use of such data for research purposes.

Research presented here is partly sponsored by CNPq- Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico, Brazilian Government.

5. REFERENCES

- [1] Marvelous Work and a Wonder, p. 189.
- [2] <http://en.wikipedia.org/wiki/Thanatos>, visited on 18/06/2011.
- [3] R. D. Lins. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. ICIAR 2009, LNCS v.5627, p.844 - 854 Springer Verlag, 2009.
- [4] R. Kasturi, L. O’Gorman and V. Govindaraju, "Document image analysis: A primer", Sadhana, (27):3-22, 2002.
- [5] G.Sharma, "Show-through cancellation in scans of duplex printed documents", IEEE Trans. Image Processing, v10(5):736-754, 2001.
- [6] G. F. P e Silva, R. D. Lins, J. M. M. da Silva. HistDoc - A Toolbox for Processing Images of Historical Documents, ICIAR 2010, LNCS v.6112, p.1 – 11. Springer Verlag, 2010.
- [7] G. F. P e Silva, R. D. Lins, J. M. M. da Silva, S. Banergee, A. Kuchibhotla, M. Thielo. Enhancing the Filtering-Out of the Back-to-Front Interference in Color Documents with a Neural Classifier. ICPR 2010. pp: 2415-2419. IEEE Press.
- [8] A. de A. Formiga and R. D. Lins. Efficient Removal of Noisy Borders of Monochromatic Documents. International Conference on Image Analysis and Recognition, 2009, LNCS v.5627. p.158 – 167, Springer Verlag, 2009.
- [9] B. T. Ávila and R. D. Lins. A Fast Orientation and Skew Detection Algorithm for Monochromatic Document Images. ACM International Conference on Document Engineering, 2005. ACM Press, 2005. p.118 - 126
- [10] G. G. de Mattos, A. de A. Formiga, R. D. Lins and F. M. J. Martins. BigBatch: a document processing platform for clusters and grids. Proceedings of ACM-SAC 2008. v.1. p.434 – 441, ACM Press, 2008.
- [11] ABBYY FineReader 10 Professional Editor, <http://finereader.abbyy.com/>.
- [12] C. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art techniques", Pattern Recognition, 36(10):2271-2285, 2003.
- [13] C. Liu, Y. Liu, and R. Dai, "Preprocessing and statistical/structural feature extraction for handwritten numeral recognition", Progress of Handwriting Recognition, A.C. Downton and S. Impedovo eds., World Scientific, 1997.
- [14] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, "Automatic recognition of handwritten numerical strings: A recognition and verification strategy", IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(11):1438-1454, 2002.
- [15] M. Hu, "Visual pattern recognition by moment invariants", IEEE Transactions on Information Theory, 8(2):179-187, 1962.
- [16] G. F. P. e Silva and R. D. Lins. An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. ICDAR 2011, Beijing, September, IEEE Press, 2011.
- [17] C.Y. Suen, J. Guo, Z.C Li, Analysis and Recognition of Alphanumeric Handprints by parts, IEEE Transactions on Systems, Man, and Cybernetics, N. 24, p. 614-631, 1994.
- [18] Z.C. Li, C.Y. Suen, J. Guo, A Regional Decomposition Method for Recognizing Handprinted Characters, IEEE Transactions on Systems, Man, and Cybernetics, N. 25, p. 998-1010, 1995.
- [19] C. O. A. Freitas, L.S. Oliveira, S.B.K. Aires, F. Bortolozzi, Zoning and metaclasses for character recognition. ACM-SAC 2007. P. 632-636, 2007.
- [20] P. V. W. Radtke, L.S. Oliveira, R. Sabourin, T. Wong, Intelligent Zoning Design Using Multi-Objective Evolutionary Algorithms, ICDAR 2003, p.824-828, 2003.
- [21] NIST Scientific and Tech. Databases <http://www.nist.gov/data/>.
- [22] R. D. Lins, G.F.P e Silva, A. de A. Formiga, HistDoc v. 2.0 - Enhancing a Platform to Process Historical Documents. HIP 2011, ACM Press, 2011.

Apêndice A

Código da Segmentação dos Registros

%FINAL07 - SEGMENTAÇÃO DOS 4 REGISTROS (AUTOMATICO)

```
files = dir('<dir>');
```

```
tic
```

```
for f=3:size(files,1)
```

```
%for f=4:4
```

```
    reg = imread(['<dir>',files(f).name]);
```

```
    largura = size(reg,2);
```

```
    altura = size (reg,1);
```

```
%divide a imagem em 2 regioes para identificar o meio vertical
```

```
rect1 = [1850 1 300 altura]; %Xinicial do crop esta relacionado com i1
```

```
reg1 = imcrop (reg,rect1);
```

```
%    imshow(reg1)
```

```
%SOMA OS PIXELS DAS COLUNAS (DISTRIBUIÇÃO HORIZONTAL - barras verticais)
```

```
%i1 e i2 são a posição do MAX(linha)
```

```
colsum1 = sum(~reg1,1);
```

```
[maxCol1,i1] = max(colsum1); %maxCol1=valor maximo do vetor soma E i1= posição desse valor MAXimo
```

```
f
```

```
i1;
```

```
maxCol1;
```

```
%CORTE VERTICAL
```

```
%Lado Esquerdo
```

```
rectE = [1 1 1850+i1 altura];
```

```
regE = imcrop (reg,rectE);
```

```
imwrite(regE, ['<dir>',files(f).name(1:end-4)], '_E.tif');
```

```
%Corte horizontal Direita (divide a imagem em 2 regioes - Superior=1 e Inferior=2)
```

```
larguraE = size(regE,2);
```

```
alturaE = size (regE,1);
```

```
rectE = [1 1350 larguraE 300]; %imagem temporaria para idenficar o meio
```

```
regEt = imcrop (regE,rectE);
```

```
colsumE = sum(~regEt,2);
```

```
[maxColE,iE] = max(colsumE); %maxCol1=valor maximo do vetor soma E i1= posição desse valor MAXimo
```

```
rectE1 = [1 1 larguraE 1350+iE]; %Lado Esquerdo1 (superior)
```

```
regE1 = imcrop (regE,rectE1);
```

```
imwrite(regE1, ['<dir>',files(f).name(1:end-4)], '_E1.tif');
```

```
FE = alturaE -(iE +1350);
```

```
rectE2 = [1 1350+iE larguraE FE]; %Lado Esquerdo2 (inferior)
```

```

regE2 = imcrop (regE,rectE2);
imwrite(regE2, ['<dir>',files(f).name(1:end-4)],'_E2.tif');

%-----
%Lado Direito
F = largura -(i1 +1850);
rectD = [1851+i1 1 F altura];
regD = imcrop (reg,rectD);
imwrite(regD, ['<dir>',files(f).name(1:end-4)],'_D.tif');

%Corte horizontal Esquerda
larguraD = size(regD,2);
alturaD = size (regD,1);

rectD = [1 1350 larguraD 300]; %imagem temporaria para idenficar o meio
regDt = imcrop (regD,rectD);

colsumD = sum(~regDt,2);
[maxColD,iD] = max(colsumD); %maxCol1=valor maximo do vetor soma E i1= posição desse valor
MAximo

rectD1 = [1 1 larguraD 1350+iD]; %Lado Direito1 (superior)
regD1 = imcrop (regD,rectD1);
imwrite(regD1, ['<dir>',files(f).name(1:end-4)],'_D1.tif');

FD = alturaD -(iD +1350);
rectD2 = [1 1350+iD larguraD FD]; %Lado Direito2 (inferior)
regD2 = imcrop (regD,rectD2);
imwrite(regD2, ['<dir>',files(f).name(1:end-4)],'_D2.tif');

end

toc

```


Apêndice B

Código da Uniformização das Imagens dos Registros

%HIP 00 - Uniformizar as Imagens de Registro

%LER A IMAGEM DO NÚMERO DO REGISTRO BINARIZADA (.tif)

files = dir('<dir>');

tic

for f=3:size(files,1)

reg = imread(['<dir>',files(f).name]);

files(f).name;

%imshow(reg)

largura = size(reg,2);

altura = size (reg,1);

%%

%Uniformizar a LARGURA tendo como referencia a linha pós numero

%%

%divide a imagem em 2 regioes para identificar as linhas verticais

rect1 = [50 1 450 altura]; %Xinicial do crop esta relacionado com i1

rect2 = [500 1 1430 altura];

reg1 = imcrop (reg,rect1);

reg2 = imcrop (reg,rect2);

%SOMA OS PIXELS DAS COLUNAS (DISTRIBUIÇÃO HORIZONTAL - barras verticais)

%i1 e i2 são a posição do MAX(linha)

colsum1 = sum(~reg1,1);

[maxCol1,i1] = max(colsum1); %maxCol1=valor maximo do vetor soma E i1=posição desse valor Maximo

colsum2 = sum(~reg2,1);

[maxCol2,i2] = max(colsum2);

f;

i1;

maxCol1;

%Formação da imagem Uniformizada

xmin = i1-120;

size(reg,2);

larguraF = 1899; %largura final da imagem uniformizada (é 1900 o total, mas aqui é 1(X0) +larguraF(1899)

%Se ponto inicial(Xmin)+largura final(larguraF), ultrapassar o tamanho

%total da imagem(1900), entao deve-se fazer um preenchimento da imagem PARA ALCANÇAR a larguraF.


```

%IF PARA LARGURA TOTAL
if larguraF+xmin > size(reg,2)
    %f
    %info = 'ultrapassou o maximo'
    larguraO = size(reg,2)-xmin; %larguraO= largura original da imagem sem enchimento
    x = larguraF - larguraO; % diferença entre largura desejada e original = tamanho do enchimento
    matrizX = ones(size(reg),x); %cria uma matriz do tamanho do enchimento necessário
    rect3 = [xmin 1 larguraO altura];
    reg3 = imcrop (reg,rect3);
    regT = [reg3 matrizX]; % Agrupa as 2 matrizes - reg original + enchimento = regT (registro
Temporario)
    xmin=1; %como a imagem foi refeita (crop a partir de Xmin) então o inicio agora passa a ser
em 1;
    %imshow(regT);
    size(regT);

else
    regT=reg;
end
size(regT);
%Ajusta totalmente a LARGURA
rectH = [xmin 1 larguraF altura];
regT = imcrop(regT,rectH);
size(regT);
%imshow(regT)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Uniformizar a ALTURA tendo como referencia a linha do numero do
%registro
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
largura2 = size(regT,2);
%fazer crop na regioao q contem o numero do registro para identificacao
%da linha
rectV = [1 10 170 150]; %largura=1 a i1
regV = imcrop (regT,rectV);
%imshow(regV);
%identificar o numero maximo de pixels nessa regioao (linha do numero)
linesumV = sum(~regV,2);
[maxLineV,iV] = max(linesumV); %maxLineV=numero de pixels maximo e iV=posição da linha do numero
f;
iV;

%PREENCHIMENTO - IF PARA ALTURA ACIMA DA LINHA DE NUMERO
if (iV+10) < 50
    %f;
    %info0 = 'altura acima da linha menor q 50'
    V = 40-iV; %50-(iV+10)
    matrizV = ones(V, largura2); %cria uma matriz do tamanho do enchimento necessário
    regT = [matrizV; regT]; % Agrupa as 2 matrizes - imagem original + enchimento = regV (registro
Temporario Vertical)
    iV=40;
else
end

%PREENCHIMENTO - IF PARA ALTURA ABAIXO DA LINHA DE NUMERO

```

```

alturaV = size(regT,1)-(iV+10); %alturaV = altura de iV até o final

if alturaV < 1420          %se essa altura for menos q 1420, precisa de preenchimento
    %f
    %info1 = 'altura total menor q a minima (1470)'
    y = 1420 - alturaV;    % diferença entre largura desejada e original = tamanho do enchimento
    matrizY = ones(y, largura2); %cria uma matriz do tamanho do enchimento necessário
    regT = [regT; matrizY]; % Agrupa as 2 matrizes - reg original + enchimento = regT (registro
Temporario)
    else
    end

    %PREENCHIMENTO TOTAL
    iT= abs(iV-40);
    if iT==0 %teste para evitar iT=0 que tinha como consequencia alturas finasis=1469;
        iT=1;
    else
    end

    %Ajusta totalmente a IMAGEM FINAL (uniformizada)
    rectF = [1 iT larguraF 1469];
    regF = imcrop(regT,rectF);

    imwrite(regF, ['<dir>',files(f).name(1:end-5)], '_E2unif.tif');

    %size(regF)
    if size(regF,1) ~=1470
        f
        size(regF)
    else
    end
end
%FINAL09 - TESTA A UNIFORMIZAÇÃO

files = dir('<dir>');

tic

%for f=3:size(files,1)
for f=1:40
    reg = imread(['<dir>',files(f).name]);
    f
    largura = size(reg,2);
    altura = size (reg,1)

    if largura ~= 1900
        f
        err = 'largura errada'
        largura
    else
    end

    if altura ~= 1470
        f
        err2 = 'altura errada'
        altura
    else

```

Apêndice C

Código para geração das Máscaras e Extração dos Campos dos Registros

% FINAL 12 - OBJETIVO: Definir área de crop após aplicação da máscara,
% automaticamente. Gerar MM (Matriz de Mascaras) e realizar o Crop.

```
clear all
tic
%LER A IMAGEM DO NÚMERO DO REGISTRO BINARIZADA (.tif)
files = dir('<dir>');

SF = size(files,1)-2 ; %numero de imagens
MM = ones(SF,4);

for f=3:size(files,1)
%for f=4:4
    reg = imread(['<dir>',files(f).name]);
    pos=f-2;

    %definição da LARGURA da Máscara
    largura = size(reg,2);
    colsum = sum(~reg,1);
    [maxCol,i1] = max(colsum);

    %Posição inicial da mascara (eixo dos X)
    MM(pos,1)=i1;
    MM(pos,4)=maxCol; % (pos,4) o número de pixels da coluna = altura (por isso inverte com (pos,3))

    %definição da ALTURA da Máscara
    altura = size (reg,1);
    linesum = sum(~reg,2);
    [maxLine,i2] = max(linesum);

    MM(pos,2)=i2;
    MM(pos,3)=maxLine;

end
MM %matriz de mascaras (MM) definida!

%realizar o crop a partir de MM

files2 = dir('<dir>');

%for i=3:size(files2,1)
for i=6:402
    ima = imread(['<dir>',files2(i).name]);

    %Habilitar de acordo com o campo que será segmentado.
    %'leu ima'
```

```

% %Crop regioao 1
% rectm1 = [MM(1,1) MM(1,2) MM(1,3) MM(1,4)];
% regm1 = imcrop (ima,rectm1);
% imwrite(regm1, ['<dir>',files2(i).name(1:end-)],'_crop_m01.tif');
%
% %Crop regioao 2
% rectm2 = [MM(2,1) MM(2,2) MM(2,3) MM(2,4)];
% regm2 = imcrop (ima,rectm2);
% imwrite(regm2, ['<dir>',files2(i).name(1:end-)],'_crop_m02.tif');
%
% %Crop regioao 3
% rectm3 = [MM(3,1) MM(3,2) MM(3,3) MM(3,4)];
% regm3 = imcrop (ima,rectm3);
% imwrite(regm3, ['<dir>',files2(i).name(1:end-)],'_crop_m03.tif');
%
% %Crop regioao 4
% rectm4 = [MM(4,1) MM(4,2) MM(4,3) MM(4,4)];
% regm4 = imcrop (ima,rectm4);
% imwrite(regm4, ['<dir>',files2(i).name(1:end-)],'_crop_m04.tif');
%
% %Crop regioao 5
% rectm5 = [MM(5,1) MM(5,2) MM(5,3) MM(5,4)];
% regm5 = imcrop (ima,rectm5);
% imwrite(regm5, ['<dir>',files2(i).name(1:end-)],'_crop_m05.tif');
%
% %Crop regioao 6
% rectm6 = [MM(6,1) MM(6,2) MM(6,3) MM(6,4)];
% regm6 = imcrop (ima,rectm6);
% imwrite(regm6, ['<dir>',files2(i).name(1:end-)],'_crop_m06.tif');
%
% %Crop regioao 7
% rectm7 = [MM(7,1) MM(7,2) MM(7,3) MM(7,4)];
% regm7 = imcrop (ima,rectm7);
% imwrite(regm7, ['<dir>',files2(i).name(1:end-)],'_crop_m07.tif');
%
% %Crop regioao 8
% rectm8 = [MM(8,1) MM(8,2) MM(8,3) MM(8,4)];
% regm8 = imcrop (ima,rectm8);
% imwrite(regm8, ['<dir>',files2(i).name(1:end-)],'_crop_m08.tif');
%
% %Crop regioao 9
% rectm9 = [MM(9,1) MM(9,2) MM(9,3) MM(9,4)];
% regm9 = imcrop (ima,rectm9);
% imwrite(regm9, ['<dir>',files2(i).name(1:end-)],'_crop_m09.tif');
%
% %Crop regioao 10
rectm10 = [MM(10,1) MM(10,2) MM(10,3) MM(10,4)];
regm10 = imcrop (ima,rectm10);
imwrite(regm10, ['<dir>',files2(i).name(1:end-8)],'_crop_m10.tif');

```

end

toc

% FINAL 11 - OBJETIVO: GERAR MM, Definir área de crop, pós aplicação da máscara, automaticamente.

```

clear all
tic
%LER A IMAGEM DO NÚMERO DO REGISTRO BINARIZADA (.tif)
files = dir('<dir>');

SF = size(files,1)-2 %numero de imagens
MM = ones(SF,4)

for f=3:size(files,1)
%for f=4:4
    reg = imread(['<dir>',files(f).name]);
    pos=f-2

    %definição da LARGURA da Máscara
    largura = size(reg,2);
    colsum = sum(~reg,1);
    [maxCol,i1] = max(colsum)

    %Posição inicial da mascara (eixo dos X)
    MM(pos,1)=i1
    MM(pos,3)=maxCol

    %definição da ALTURA da Máscara
    altura = size (reg,1);
    linesum = sum(~reg,2);
    [maxLine,i2] = max(linesum)

    MM(pos,2)=i2
    MM(pos,4)=maxLine

end
MM
toc

```

%FINAL15 - Segmentar Campo do Numero do Registro com Mascara

%criado teste para verificar se a imagem do registro está binarizada, %senão deve binarizá-la em tempo de execução; para evitar erros de %classes diferentes durante a execução da função PLUS.

%Incluir laço para varrer o diretório de mascaras automaticamente tb.

%Limpa todas as Variáveis%
clear all;

```

tic
%LER A IMAGEM DO NÚMERO DO REGISTRO BINARIZADA e uniformizada(.tif)
%LER a mascara para o numero do registro
files = dir('<dir>');
masks = dir('<dir>');

for m=8:size(masks,1)
    mask = imread(['<dir>',masks(m).name]);
    m
    for f=3:size(files,1)
        %for f=3:3

```

```

f;
reg = imread(['<dir>',files(f).name]);
if size(reg,1)~=1470
    f
else
    f
    size(reg);
    size(mask);
    %Filename
    %N = file(f).name(1:end)

    %teste para verificar se todas as imagens estao binarizadas, pois se
    %não, não é possível executar o PLUS(+) com imagens de classes diferentes

    flagREG=isbw(reg);
    if flagREG==1
        reg1=reg;
    else
        level = graythresh(reg);
        level1 = level +0.05;
        reg1 = im2bw(reg,level1);
    end

    A = reg1 + mask;

    %imshow(A)
    if m==3
        imwrite(A, ['<dir>',files(f).name(1:end-4)], '_m01.tif');
    elseif m==4
        imwrite(A, ['<dir>',files(f).name(1:end-4)], '_m02.tif');
        ...

    elseif m==12
        imwrite(A, ['<dir>',files(f).name(1:end-4)], '_m10.tif');

    end
end
end
end

end

toc

```

Apêndice D

Código para Remoção de Linhas Horizontais

%HIP 01 - FINAL17 - RETIRA LINHAS horizontais

%LER A IMAGEM DO NÚMERO DO REGISTRO BINARIZADA (.tif)

```
files = dir(<dir>');
```

tic

```
for f=3:size(files,1)
```

```
bw = imread([<dir>',files(f).name]);
```

```
flag = isbw(bw)
```

```
subplot(1,2,1)
```

```
imshow(bw);
```

if flag==0

'gray'

```
level = graythresh(bw);
```

```
bw = im2bw(bw,level);
```

'ok, now is bw'

else

end

```
%retirar as linhas HORIZONTAIS
```

```
x=1;
```

```
for y=40:[size(bw,1)-8]
```

```
x=1;
```

```
for i=1:(size(bw,2)-8)
```

i

 y

X

[illegible]

'chegou ao final da linha x='

X;

break

else

```
%formação do bloco 8x8 para varredura
```

```
mat=bw(y:y+7 , x:x+7);
```

%mat

'gerou mat'

```
soma = sum (mat(:));
```

```
if soma==64
```

```
x=x+8;
```

'soma = 64'

else

```
if soma >=20
```

'mais de 32 pixels brancos o bloco 8x8 - predominantemente branco'

```
if sum(mat(1,:))==8
```

```

    '1ª linha é toda branca'
    if sum(mat(8,:))==8
        '8ª linha é toda branca'
        if sum(mat(:,1))<8
            '1ª coluna tem pelo menos 1 pixel preto'
            bw(y:y+7 , x:x+7) = 1;
            bw(y:y+7 , x:x+7)
            x=x+8;
            'apagou mat'
        else
            x=x+1;
        end
    else
        x=x+1;
        sum(mat(8,:));
    end
end
else
    x=x+1;
    sum(mat(1,:));
end
end
else
    x=x+1;
end
end
end
y;
x;

end
end
imwrite(bw, [<dir>', [files(f).name(1:end-4)], '_L.tif']);
subplot(1,2,2)
imshow(bw);
end
toc
t=toc

```


Apêndice E

Código para Remoção de Ruídos

```
% HIP02 - retirar ruido
%aplicando bloco de varredura de dimensões 4x4
% Início da varredura = variável 'y'

%LER A IMAGEM DO NÚMERO DO REGISTRO BINARIZADA (.tif)
files = dir('<dir>');

tic

for f=3:size(files,1)
    bw = imread(['<dir>',files(f).name]);

    flag = isbw(bw)
    subplot(1,2,1)
    imshow(bw);

    if flag==0
        'image gray'
        level = graythresh(bw);
        bw = im2bw(bw,level);
        'ok, now is bw'
    else
        end

    %retirar RUIDOS
    x=1;
    for y=1:(size(bw,1)-8)        %a varredura é realizada na imagem toda, a partir de y=1.

        x=1;
        for i=1:(size(bw,2)-8)
            i
            y
            x

        if x>=(size(bw,2)-8)      %se chegar ao final da imagem (pelo eixo dos x) passa para o proximo Y.
            'chegou ao final da linha x='
            x;
            break
        else
            %formação do bloco 4x4 para varredura
            mat=bw(y:y+3 , x:x+3);
            %mat
            'gerou mat'
            soma = sum (mat(:));
            if soma==16
                x=x+4;
                'soma = 16 (bloco totalmente branco)'
            else
                if soma >=11      %apaga até 5 pixels pretos
                    'mais de 11 pixels brancos o bloco 4x4 - predominantemente branco'
```

```

if sum(mat(1,:))==4
    '1ª linha é toda branca'
if sum(mat(4,:))==4
    '4ª linha é toda branca'
if sum(mat(:,1))<4
    '1ª coluna tem pelo menos 1 pixel preto'
    bw(y:y+3 , x:x+3) = 1;
    bw(y:y+3 , x:x+3); %soh para debugar
    x=x+4;
    'apagou bloco (mat)'
else
    x=x+1;
end
else
    x=x+1;
    sum(mat(4,:));
end
end
else
    x=x+1;
    sum(mat(1,:)); %soh para debugar
end
end
else
    x=x+1;
end
end
end
end
end
imwrite(bw, ['<dir>',files(f).name(1:end-4)], '_LR.tif');
subplot(1,2,2)
imshow(bw);
end

```

Apêndice F

Código para Remoção de Linhas Verticais

```
%FINAL18 - RETIRA COLUNAS
```

```
%LER A IMAGEM DO NÚMERO DO REGISTRO BINARIZADA (.tif)
files = dir('<dir>');
tic
```

```
for f=3:size(files,1)
%for f=4:4
    reg = imread(['<dir>',files(f).name]);
```

```
%SOMA OS PIXELS DAS COLUNAS (DISTRIBUIÇÃO HORIZONTAL - barras verticais)
colsum = sum(~reg,1);
maxCol = max(colsum)
meanCol = mean(colsum)
limiarCol = maxCol - (2*meanCol)
```

```
%armazena os valores max e medios dos pixels pretos das colunas
matMaxMeanCol (1, f-2) = maxCol;
matMaxMeanCol (2, f-2) = meanCol;
```

```
%RETIRA DA IMAGEM COLUNAS VERTICAIS - Matriz Bidimensional
init = (size(reg,2)/2)+0.5
for z=init:size(reg,2)-1
    A = colsum(z); % A recebe o somatorio dos pixels de uma coluna
    %A = A*0,7
    if A > limiarCol
        reg(:,z)= 1;
        reg(:,z-1)= 1;
        reg(:,z+1)= 1;
        %reg2(:,z)=255;
    else
        end
    end
end
```

```
imwrite(reg, ['<dir>',files(f).name(1:25)], '_lr2c.tif');
f
```

```
end
```

```
toc
t=toc
```

Apêndice G

Código Reverso para Remoção de Linhas Horizontais

%FINAL19 - RETIRA LINHAS HORIZONTAIS - REVERSO

%LER A IMAGEM DO NÚMERO DO REGISTRO BINARIZADA (.tif)

```
files = dir('<dir>');
tic
```

```
for f=3:size(files,1)
```

```
%for f=13:13
```

```
    bw = imread(['<dir>',files(f).name]);
```

```
    flag = isbw(bw)
```

```
    subplot(1,2,1)
```

```
    imshow(bw);
```

```
    if flag==0
```

```
        'gray'
```

```
        level = graythresh(bw);
```

```
        bw = im2bw(bw,level);
```

```
        'ok, now is bw'
```

```
    else
```

```
    end
```

%retirar as linhas HORIZONTAIS - REVERSO

```
x=1;
```

```
for y=[size(bw,1)-8]:-1:40      %começa NA ÚLTIMA LINHA e vai até 40 (q é um pouco antes da
identificação da linha, q é em y=50).
```

```
    x=size(bw,2)-8;
```

```
    for i=1:(size(bw,2)-8)
```

```
        i
```

```
        y
```

```
        x
```

```
    if x<=1
```

%se chegar ao INICIO da imagem (pelo eixo dos x) passa para o Y ANTERIOR.

```
        'chegou ao inicio da linha x='
```

```
        x;
```

```
        break
```

```
    else
```

```
        %formação do bloco 8x8 para varredura
```

```
        mat=bw(y:y+7 , x:x+7);
```

```
        %mat
```

```
        'gerou bloco 8x8 (mat)'
```

```
        soma = sum (mat(:));
```

```
        if soma==64
```

```
            x=x-8;
```

```
            'soma = 64'
```

```
        else
```

```

if soma >=20
    'mais de 32 pixels brancos o bloco 8x8 - predominantemente branco'
    if sum(mat(1,:))==8

        '1ª linha é toda branca'
        if sum(mat(8,:))==8
            '8ª linha é toda branca'
            if sum(mat(:,8))<8
                '8ª coluna tem pelo menos 1 pixel preto'
                bw(y:y+7 , x:x+7) = 1;
                bw(y:y+7 , x:x+7)
                x=x-8;
                'apagou mat'
            else
                x=x-1;
            end
        else
            x=x-1;
            sum(mat(8,:));
        end
    else
        x=x-1;
        sum(mat(1,:));
    end
else
    x=x-1;
end
end
end
y;
x;
%break
end
end
imwrite(bw, ['<dir>',files(f).name(1:25)], '_LRLlr.tif');
subplot(1,2,2)
imshow(bw);
end

toc
t=toc

```

Apêndice H

Código para Segmentação dos Dígitos Numéricos

```
%FINAL03_SEGMENTAÇÃO
%Teste para segmentação de cada dígito do Número do Registro

%LER A IMAGEM DO NÚMERO DO REGISTRO BINARIZADA (.tif)
files = dir('<dir>');
files2 = dir('<dir>');
tic

for f=3:size(files,1)
%for f=7:7
    reg = imread(['<dir>',files(f).name]);

    %SOMA OS PIXELS DAS COLUNAS (DISTRIBUIÇÃO HORIZONTAL - barras verticais)
    colsum = sum(~reg,1);
    bar(colsum);

    %GERAÇÃO DE MATRIZES E VARIÁVEIS DE APOIO
    colsum2 = size(colsum,2)/2; %Calcula o tamanho da matriz de medias = metade da largura da imagem
    (230/2=115)
    media=zeros(1, colsum2); %MATRIZ DO VALOR MEDIO DOS PIXELS DE 2x2 COLUNAS
    ponteiro=zeros(1, 14); %MATRIZ DOS VALORES DOS PONTEIROS QUE MARCAM O INICIO E FIM
    DE CADA DÍGITO
    i = 1; %Indice da matriz de ponteiros
    m1 = 1; %Indice da matriz de VALOR MÉDIO
    S = 0; %CONTADOR DOS NIVEIS DE SUBIDA
    D = 0; %CONTADOR DOS NIVEIS DE DESCIDA

    %GERA MATRIZ COM O VALOR MEDIO DOS PIXELS DE 2x2 COLUNAS
    for m=1:2:size(colsum,2)
        A1=colsum(m);
        A2=colsum(m+1);
        A=(A1+A2)/2;
        media(1, m1)=A;
        m1 = m1+1;
    end

    %COMPARA OS VALORES DAS MEDIAS PARA IDENTIFICAR SUBIDAS E DESCIDAS E
    %GERA MATRIZ DE PONTEIROS

    for p=1:(size(media,2)-1) %media tem tamanho da figura /2 - 230/2 = 115
        if i >= 14
            break
        else
            B1=media(p);
            B2=media(p+1);
            if B2 > B1 %representa uma subida
                S = S+1;
```

```

D = 0;
if S == 3
    if mod(i,2) == 1 %SE i FOR IMPAR, SALVA DIRETO. SE FOR PAR, INCREMENTA 1
        POSIÇÃO NA MATRIZ DE PONTEIROS
        ponteiro(1,i) = (p-2)*2; %armazena a referencia do inicio da subida(p-2) e *2 para fazer
referencia a imagem original
        i = i+1;
    else
        ponteiro(1,i+1) = (p-2)*2; %armazena a referencia do inicio da subida(p-2) e *2 para fazer
referencia a imagem original
        i = i+1;
    end
else
    end
elseif B2 < B1 %representa uma descida
    D = D+1;
    S = 0;
    if D == 3
        if mod(i,2) ~= 1 %SE i FOR PAR, SALVA DIRETO. SE FOR PAR, INCREMENTA 1 POSIÇÃO
        NA MATRIZ DE PONTEIROS (PQ AULTIMA JÁ FOI DESCIDA)
        ponteiro(1,i) = (p*2)+4; %armazena a referencia do fim da descida(p) e *2 para fazer referencia a
imagem original + 4pixels para folga
        i = i+1;
    else
        ponteiro(1,i+1) = (p*2)+4; %armazena a referencia do fim da descida(p) e *2 para fazer
referencia a imagem original + 4pixels para folga
        i = i+1;
    end
    end
else
    end
end

end
end
ponteiro %ponteiro tem tamanho 14, para comportar as subidas e descidas de cada um dos 7 digitos.

```

```

for elimina = 1:size(ponteiro,2)
    E = ponteiro(elimina);
    if E == 0
        if elimina == 1
            sprintf('problema - borda a esquerda')
        else
            if mod(elimina,2) == 1 %se for impar, representa ponteiro de subida soma 2 em relação a posição
anterior
                ponteiro(elimina) = ponteiro(elimina-1)+2; %Substitui aquela posição do vetor pela anterior +2
pixels
            else %se for par, representa ponteiro de descida subtrai 2 em relação a posição posterior
                if elimina == 14
                    sprintf('problema - borda a direita')
                    pont = ponteiro(elimina-1)+20;
                    if pont <=230
                        ponteiro(elimina) = pont; %Substitui aquela posição do vetor pela anterior +20
                    else
                        ponteiro(elimina) = 230;
                    end
                end
            end
        end
    end
end

```

```

else
    if ponteiro(elimina+1) == 0 %se a proxima posição tb for zero, então será somada
        pont2 = ponteiro(elimina-1)+20;
        if pont2 <=230
            ponteiro(elimina) = pont2; %Substitui aquela posição do vetor pela anterior +20
        else
            ponteiro(elimina) = 230;
        end
    else
        ponteiro(elimina) = ponteiro(elimina+1)-2; %Substitui aquela posição do vetor pela posterior -
2
    end
end
end
end
end
else
end
end
sprintf('ponteiro após indentificação de zeros')
ponteiro

%UTILIZAR MATRIZ DE PONTEIRO PARA SEGMENTAR A IMAGEM EM ESCALA DE CINZA
%- SEM BORDA

nameNR = [files(f).name(1:end-11), 'NR_NROKgray.tif']; NRi = imread(['<dir>',nameNR]);

for segmenta = 1:2:(size(ponteiro,2)-1)
    pi = ponteiro(segmenta);
    pf = ponteiro(segmenta+1);
    p = pf-pi;
    if ponteiro(elimina) >= 228 %verifica se chegou ao fim das colunas para evitar travamento no momento da
segmentação (imcrop)
        break
    else
        NR = imcrop(NRi, [pi 1 p 150]); %A altura ficou fixa e igual ao tamanho da imagem (150pixels)
        if segmenta == 1
            dig1 = [nameNR(1:end-13),'OK_d1.tif'];
            imwrite(NR, ['<dir>',dig1]);
        elseif segmenta == 3
            dig2 = [nameNR(1:end-13),'OK_d2.tif'];
            imwrite(NR, ['<dir>',dig2]);
        elseif segmenta == 5
            dig3 = [nameNR(1:end-13),'OK_d3.tif'];
            imwrite(NR, ['<dir>',dig3]);
        elseif segmenta == 7
            dig4 = [nameNR(1:end-13),'OK_d4.tif'];
            imwrite(NR, ['<dir>',dig4]);
        elseif segmenta == 9
            dig5 = [nameNR(1:end-13),'OK_d5.tif'];
            imwrite(NR, ['<dir>',dig5]);
        elseif segmenta == 11
            dig6 = [nameNR(1:end-13),'OK_d6.tif'];
            imwrite(NR, ['<dir>',dig6]);
        else segmenta == 13
            dig7 = [nameNR(1:end-13),'OK_d7.tif'];
            imwrite(NR, ['<dir>',dig7]);
        end
    end
end
end

```