**Flávio Augusto Rocha Bertholdo**

Orientador:
Prof. Arnaldo de Albuquerque Araújo

# Técnicas de Limiarização para Melhorar a Qualidade Visual de Documentos Históricos

Dissertação apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
Belo Horizonte, Setembro de 2007

*Para Gizele: seu amor me faz entender*
*que tudo que há de melhor está na vida.*

*Para Beatriz e Maria Clara: ter a presença de vocês,*
*sempre perto de mim, fortifica e alegra meu espírito.*

*Para Gledmar: sua dedicação e perseverança*
*são exemplos que busco seguir*

*Vocês são as mulheres da minha vida!*

## Agradecimentos

Ao Professor Arnaldo pelo apoio e determinação ao orientar este trabalho. Sobretudo quero agradecer a grande generosidade com que tem compartilhado seus conhecimentos comigo e com tantos outros alunos.

Agradeço ao excepcional corpo de professores do Programa de Pós-Graduação do Departamento de Ciência da Computação da UFMG. Obrigado também ao DCC/UFMG e à todos os seus funcionários por tornar viável a realização deste trabalho.

Aos colegas do Núcleo de Processamento Digital de Imagens, em especial ao Camillo, Flávio (FHC) e David Menotti pela cooperação, discussões, comentários e sugestões.

Agradeço à Deus pela perfeição do seu amor em nossas vidas.

As invenções são, sobretudo, resultado de um trabalho teimoso.

*Alberto Santos Dumont*

É preciso aprender com a prática, pois, embora você pense que sabe, só terá certeza depois que experimentar.

*Sófocles*

## Resumo

A digitalização de documentos históricos apresenta-se como forma eficaz de viabilizar o acesso público a grandes acervos e equalizar o severo compromisso entre conservação e acesso. Freqüentemente, documentos históricos apresentam alto grau de degradação, causada pela ação do tempo ou danos sofridos, resultando em imagens digitais com baixo nível de legibilidade e em alguns casos extremos, totalmente ilegíveis. Técnicas de processamento digital de imagens e análise de documentos são empregadas na solução deste problema.

A maioria das abordagens para limiarização e segmentação de documentos históricos utiliza características globais do documento para eliminar o ruído de fundo e outros problemas de degradação, buscando aprimorar a legibilidade. No entanto, a distribuição dos problemas de degradação não é uniforme e nem linear na imagem do documento. Desta forma, após a aplicação destes algoritmos, algumas regiões apresentam melhoria da legibilidade, porém, em outras, observa-se o efeito reverso.

Este trabalho apresenta uma nova abordagem para o problema de melhorar a qualidade visual e a legibilidade de imagens de documentos históricos. A solução utiliza uma abordagem híbrida, combinando características globais e locais. Pode-se dividir o processamento em quatro etapas. Primeiro, as características globais do documento são extraídas. Na segunda etapa, são identificadas as linhas que apresentam conteúdo textual. Na próxima etapa, é realizada a limiarização das linhas selecionadas, combinando características locais e globais. Finalmente, na última etapa, é realizada a binarização global do documento. Experimentos realizados com imagens de documentos históricos do acervo Dops/MG demonstram que a abordagem proposta foi eficiente em melhorar a qualidade visual e legibilidade em 80 por cento dos documentos.



### Palavras-chave:

Documentos históricos, análise de documentos, limiarização, segmentação, binarização.

## Abstract

The digitization of historical documents presents itself as an effective way of enabling public access to extensive archives and equalizing the serious commitment between preservation and access. Often historical documents present severe degradation caused by effects of time or actual damage, resulting in digital images with poor legibility, and in some extreme cases, they are totally illegible. Digital image processing techniques and document analysis are employed for the solution of this problem.

Most approaches to thresholding and segmentation of historical documents use global characteristics of documents, in order to eliminate background noise and other degradation problems aiming to improve legibility. However, the distribution of degradation problems is not homogeneous or linear on the document image. Therefore, while after the application of these algorithms some parts may present better legibility, in others the opposite effect can be observed.

This work presents a new approach to the problem of image quality improvement and image legibility of historical documents. The solution uses a hybrid approach combining global and local characteristics. The processing can be divided into four stages. First, the global characteristics of the document are extracted. Secondly, the lines that present textual content are identified and processed in the following stage. In the third stage, the thresholding of the selected lines is done combining local and global characteristics. Finally, in the last stage, global binarization of the document is done. Experiments done with images of historical documents from the Dops/MG archive showed that the proposed approach was efficient in improving the visual quality and legibility in 80 percent of the documents.



**Keywords:**

Historical documents, document analysis, thresholding, segmentation, binarization.

# Sumário

# Lista de Ilustrações

# Lista de Tabelas

# Lista de Siglas e Abreviaturas

| Sigla | Significado |
|---|---|
| APM | Arquivo Público Mineiro |
| CD | Disco Compacto, do inglês *Compact Disc* |
| CD-ROM | Disco compacto de memória apenas para leitura, do inglês: Compact Disc read-only memory |
| CPI | Comissão Parlamentar de Inquérito |
| DIA | Análise de Imagens de Documentos, do inglês *Document Image Analysis* |
| DEOPS | Departamento Estadual de Ordem Política e Social |
| DOPS | Departamento de Ordem Política e Social |
| DOPS/MG | Departamento de Ordem Política e Social do Estado de Minas Gerais |
| DPI | Pontos por polegada, do inglês *Dots per Inch* |
| IGT | Do inglês *Iterative Global Thresholding* |
| JPEG | Do inglês *Joint Photographics Experts Group* |
| OCR | Reconhecimento Óptico de Caracteres, do inglês *Optical Character Recognition* |
| PDI | Processamento digital de imagens |
| PIXEL | Elemento da figura, do inglês *PICTure ELement* |
| UFMG | Universidade Federal de Minas Gerais |

# Capítulo 1
# Introdução

Esta dissertação apresenta uma nova abordagem para o problema de melhorar a qualidade visual e a legibilidade de documentos digitalizados de acervos históricos arquivísticos. A solução utiliza uma abordagem híbrida, combinando características globais e locais. Além disso, apresenta-se um estudo comparativo de diversos algoritmos de processamento digital de imagens (PDI) e análise de documentos propostos na literatura e aplicados no processamento de documentos históricos.

Recentemente, devido à expansão dos projetos de digitalização de acervos históricos e preservação digital da memória, grandes quantidades de imagens de documentos históricos têm sido geradas por diversos projetos espalhados pelo mundo. A digitalização de documentos históricos apresenta-se como forma eficaz de viabilizar o acesso público a grandes acervos e equalizar o severo compromisso entre conservação e acesso. Porém, freqüentemente, documentos históricos apresentam alto grau de degradação, causada pelo escurecimento do papel, ressecamento da tinta, umidade, exposição à luz, armazenamento inadequado, interferência frente-verso, entre vários outros fatores, como apresentado em (BAIRD, 2004). Pode-se observar exemplos de documentos com degradação visual na Figura 1.1. Nestes documentos, observam-se problemas relativos ao ressecamento da tinta, escurecimento do papel, alto índice de ruído de fundo e baixo nível de contraste. A digitalização de documentos históricos, que apresentam esses e outros problemas, gera imagens digitais com baixo nível de legibilidade e em alguns casos extremos totalmente ilegíveis. Esta situação compromete a utilização das imagens digitais nas tarefas de indexação e pesquisa ao acervo.

Buscam-se, através do aprimoramento da qualidade visual das imagens dos documentos históricos, formas de agilizar e facilitar as tarefas de indexação e consulta de grandes acervos documentais. Observa-se que a qualidade do processo de indexação é afetado diretamente pela qualidade do documento em questão. Desta forma, a implementação de algoritmos capazes de melhorar a qualidade visual de documentos

históricos possibilita aprimoramentos na indexação, além de facilitar o acesso aos documentos, devido à melhoria da legibilidade  e da qualidade de impressão. Esta tarefa mostra-se relevante ao corroborar com a preservação da memória documental da sociedade.

Esta dissertação tem por objetivo demonstrar a viabilidade da utilização da abordagem proposta como ferramenta para aprimorar a qualidade visual e a legibilidade de imagens de documentos históricos, além de compará-la com outras abordagens propostas na literatura.

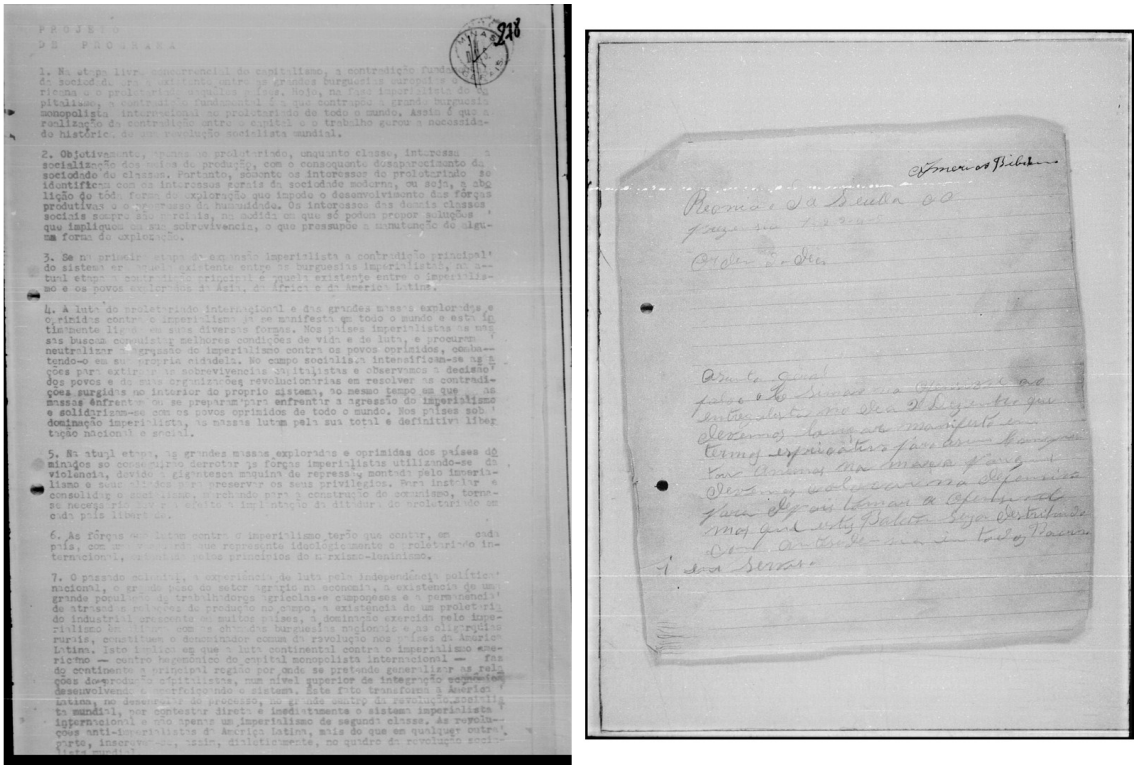


Figura 1.1 – Exemplos de documentos históricos apresentando degradação visual.

## 1.1. Motivação

Preservar a memória documental é caminho obrigatório para construir identidade cultural e garantir análise da história. Instituições destinadas a preservar arquivos públicos ou outros acervos documentais enfrentam variados problemas, em geral, provenientes do grande acúmulo de documentos e da sua fragilidade. Além destes, nos últimos anos, tem recebido especial atenção o problema de tornar mais acessíveis os

documentos aos pesquisadores e ao público em geral (ANDRADE, 2000). A fragilidade dos documentos impõe um severo compromisso entre conservação e acesso. A tecnologia digital desponta como um alternativa eficiente de disponibilizar documentos históricos para o público, sem comprometer sua integridade física (VALLE JR, 2005).

Em grandes acervos documentais, a recuperação de informação apresenta-se como peça central para garantir aos consulentes[1] o pleno acesso. Já em 1939, o arquivista Binkley defendia que:

> O objetivo da política de arquivos em um país democrático não pode ser a simples guarda de documentos. Deve ser nada menos que promover o enriquecimento da consciência histórica dos povos como um todo (BINKLEY, 1939).

Diversos trabalhos recentes descrevem a implementação de sistemas de informação multimídia destinados a realizar a recuperação de informação em acervos de documentos históricos, como os apresentados em (VALLE JR, 2002; ANDRADE, 1998). O acesso caracteriza-se como uma das dimensões mais importantes em projetos de preservação em meio digital (LIMA, 2007), pois torna-se contra-senso considerar a digitalização como meio de conservação se não for possível aos consulentes encontrarem as informações de seu interesse em meio à enormidade do acervo. A qualidade da pesquisa no acervo é influenciada diretamente pela descrição do conteúdo dos itens documentais (VALLE JR, 2002). A tarefa de descrição e indexação possui relação direta com as dimensões do acervo, podendo durar vários anos. Em geral, esta tarefa é feita de forma manual, descrevendo um documento por vez. Observa-se, então, a necessidade da criação de técnicas automáticas ou semi-automáticas para descrição e indexação de documentos históricos.

Técnicas de reconhecimento óptico de caracteres (OCR) têm sido utilizadas na captura do conteúdo de documentos históricos de caráter essencialmente textual, como apresentado em (MELLO, 1999). O sucesso desta abordagem significa grande economia de tempo, além de permitir pesquisas em todo o conteúdo textual dos documentos. Em se tratando de acervos históricos, a utilização de OCR esbarra na qualidade visual dos

1 Pessoa que consulta ou pesquisa documentos em arquivos, normalmente em uma sala de pesquisa ou leitura. É o público externo dos arquivos, também chamado usuários, pesquisadores ou leitores.

itens documentais. A ação do tempo, danos sofridos, processo de impressão empregado, entre vários outros fatores, podem degradar a qualidade visual dos documentos, afetando sensivelmente os resultados do OCR.

A Figura 1.2 apresenta um fragmento de texto de um documento. Observa-se o resultado da conversão da imagem em texto utilizando OCR, no caso, o *software Cuneiform 6.0 for Windows* (CUNEIFORM, 2006).

```
"In<Lcledo em i¿uCrito - ¿-c¿rregaão: Coai¿sio io ”e="t" Ice
taduals"

¿7em dose.".volvcn5e, como chefe do  enfartamento ãe
moino:ãndia e Sagsrior (Secretar-a de Eãucaçac àо Est: do <ie
Minas Gerais), ama intensa e dinàzica anão moraliw.dora 5e
cost ;-ice, perfeita- ente en .rocaòo coa o espfrito
revolucio¿o".
```

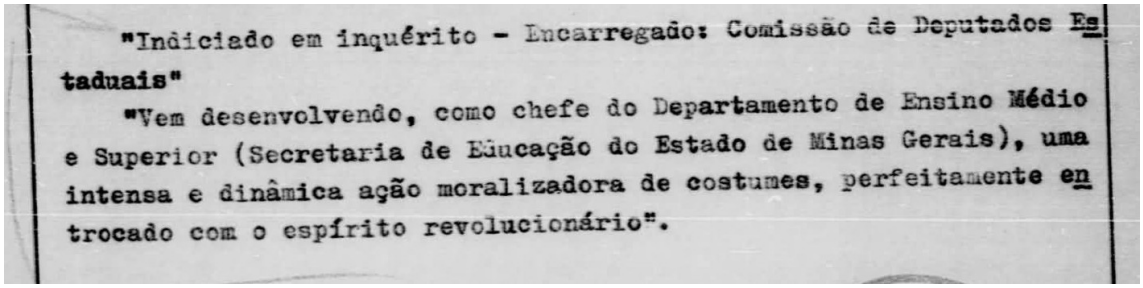

"Indiciado em inquérito - Encarregado: Comissão de Deputados Es
taduais"
"Vem desenvolvendo, como chefe do Departamento de Ensino Médio
e Superior (Secretaria de Educação do Estado de Minas Gerais), uma
intensa e dinâmica ação moralizadora de costumes, perfeitamente en
trocado com o espírito revolucionário".

Figura 1.2 – Fragmento de documento utilizado para teste de OCR.

Este simples exemplo demonstra a necessidade do desenvolvimento de técnicas para processamento de documentos históricos, capazes de aprimorar sua qualidade visual, impondo-se como condição obrigatória para o desenvolvimento de técnicas, automáticas ou semi-automáticas, de indexação de grandes acervos históricos.

## 1.2. Acervo estudado

Na realização deste trabalho, utilizou-se o acervo do Departamento de Ordem Política e Social de Minas Gerais (Dops/MG) acondicionado pelo Arquivo Público Mineiro (APM). Conhecidos genericamente como Dops, esses órgãos proliferaram por diferentes Estados Brasileiros com a finalidade de dedicar-se ao exercício das funções de “polícia política” (AQUINO, 2006). Estavam, fundamentalmente,

> à disposição dos governos quando estes decidissem vigiar e/ou aprisionar certos indivíduos, combater determinados comportamentos e estigmatizar grupos inteiros (imigrantes, dissidentes políticos, pobres das cidades) tidos sempre como "nocivos" e perigosos para a ordem pública e a segurança nacional (SOMBRA, 1996, p.41).

Nos anos 90, os arquivos públicos brasileiros iniciaram o recolhimento dos acervos das polícias políticas. Foram fundamentais para este processo o fim da ditadura militar, o início do processo de democratização, a Constituição de 1988 e a extinção dos Dops. Motta (2006) apresenta uma retrospectiva histórica da atuação da polícia política no Brasil.

Nos diversos estados da federação, o recolhimento e a abertura dos arquivos das polícias políticas ocorreram em períodos e condições diferentes. Em 1991, teve início o processo de recolhimento da documentação aos Arquivos Públicos estaduais. Naquele ano, foram liberados os arquivos do Departamento Estadual de Ordem Política e Social (Deops) de São Paulo e do Dops do Paraná (MOTTA, 2006). O Estado do Rio de Janeiro recebeu os documentos em 1992 e no Distrito Federal e Goiás, o recolhimento ocorreu em dezembro de 1995. Motta (2006) apresenta um balanço nacional do recolhimento da documentação dos arquivos dos Dops.

Em Minas Gerais, apesar da lei estadual no. 10.360/90 determinar a transferência dos acervos do Dops/MG para o APM, o recolhimento ocorreu somente em 1998, após a instalação de uma Comissão Parlamentar de Inquérito (CPI) responsável por investigar o destino dado aos arquivos do Dops/MG. A CPI foi instaurada em virtude da denúncia de uso ilegal dos registros policiais de natureza política relativos ao período militar, os quais, apesar da Lei de Anistia promulgada em 1979, continuavam presentes nos órgãos de segurança de Minas, contaminando os atestados de antecedentes criminais de cidadãos anistiados.

O acervo abrange o período compreendido entre 1927 e 1982 e pode ser considerado uma preciosa fonte de informação para a recuperação da memória nacional. Estes documentos são importantes instrumentos no estudo dos períodos de repressão que se sucederam no país bem como para resgatar os diversos aspectos da história dos movimentos políticos e sociais que marcaram a luta pela redemocratização. Sobre a potencialidade e importância dessa documentação, Motta afirma que:

> A abertura dos acervos do Dops foi conquista significativa da cidadania e passo importante no caminho de republicanizar a polícia da República. Pela primeira vez na história os cidadãos brasileiros têm o direito de consultar arquivos dos órgãos de repressão, e o significado político disso é de grande alcance (MOTTA, 2006).

Constituem o acervo do Dops correspondências policiais, prontuários de presos políticos, lista de pessoas ligadas a qualquer tipo de manifestação contrária ao regime da época, relatórios de atividades públicas realizadas em todas as regiões do Estado, como eleições, greves, apresentações teatrais, festas tradicionais, visitas de autoridades políticas ou representantes do governo, panfletos de partidos políticos e de universidades. Exemplos de documentos do acervo do Dops/MG podem ser observados na Figura 1.3.

O acervo do Dops de Minas Gerais acondicionado pelo APM está disponível em 96 rolos de microfilmes com aproximadamente 250 mil fotogramas. O acervo original (em papel) foi incinerado pela instituição que mantinha sua guarda, antes de realizar a transferência para o APM. Dessa forma, consideram-se como originais os microfilmes.

Recolhido ao APM em março de 1998, o acervo do Dops começou a ser trabalhado somente a partir de maio de 1999. A instituição envolveu quatro técnicos na elaboração de um índice onomástico[1]. Não obstante esse esforço, após doze meses de trabalho, do total de 96 rolos apenas 11 haviam sido indexados. Foi preciso criar um banco de dados, um instrumento de pesquisa que permitisse a consulta à documentação. Para tanto, foi celebrada uma cooperação entre o APM e o Departamento de História da Universidade Federal de Minas Gerais (UFMG) que em três anos de atividades resultou no arranjo de cerca de 70% do acervo (MOTTA, 2006).

A partir do projeto de digitalização e indexação, a pesquisa tornou-se mais dinâmica. Hoje, todos os fotogramas estão digitalizados e armazenados em 98 CD-ROMs[2]. Desenvolveu-se sistema informatizado para indexação e consulta às imagens digitais dos documentos. A digitalização do acervo foi realizada a partir dos

1 Compreende toda a indexação por nomes de pessoas citadas nos documentos

2 Disco compacto de memória apenas para leitura, do inglês: *Compact Disc read-only memory*

microfilmes, utilizando resolução de 300 *dpi*[1]. As imagens digitais geradas apresentam resolução adequada para o escopo deste trabalho. Porém, cabe ressaltar que optou-se pelo o armazenamento das imagens em formato *JPEG*[2]. Apesar deste formato de imagem gerar significativa economia de espaço no armazenamento, seu algoritmo de compressão implica em perdas de informação. Estudos preliminares demonstraram que estas perdas não são significativas para a realização deste trabalho. A utilização de resolução de digitalização adequada produziu imagens digitais com alto nível de detalhes, minimizando os efeitos nocivos da utilização da compressão do formato *JPEG*.

A escolha do acervo do Dops/MG como objeto de pesquisa deste trabalho deve-se a três fatores. Em primeiro, devido a sua grande importância histórica e um acentuado interesse, por parte do público, em realizar pesquisas neste acervo. Além disso, o fato de todo o acervo encontrar-se digitalizado, resultando em aproximadamente 250 mil imagens digitais. E, finalmente, por este acervo possuir atributos diferenciados: os documentos originais foram destruídos, impossibilitando qualquer iniciativa de aprimorar a qualidade visual a partir de um novo processo de digitalização, inviabilizando a aplicação das técnicas tradicionalmente utilizadas na restauração de documentos históricos em papel. O sucesso da abordagem proposta em aprimorar a qualidade visual dos documentos possibilita uma alternativa para esta documentação, onde outras são totalmente inviáveis.

1 Do inglês: *Dots Per Inch*

2 Do inglês: *Joint Photographics Experts Group*

Figura 1.3 – Exemplos de documentos do acervo do Dops.

## 1.3. Estrutura da dissertação

Nas seções anteriores, procurou-se demonstrar a importância da implementação de algoritmos para o processamento de documentos históricos, em especial, visando aprimorar a qualidade visual e legibilidade.

No Capítulo 2, apresenta-se uma revisão bibliográfica, onde são apresentadas as principais abordagens para processamento de imagens de documentos históricos, como segmentação, binarização e transformações de histograma.

O Capítulo 3 descreve em detalhes a abordagem proposta para o problema de melhorar a qualidade visual e a legibilidade de imagens de documentos históricos. Cada uma de suas etapas é apresentada e detalhada por meio de exemplos de imagens processadas, equações, gráficos e histogramas.

No Capítulo 4, estão apresentados os resultados dos experimentos utilizando imagens do acervo do Dops. Através da análise dos dados dispostos em gráficos e tabelas e exemplos de imagens processadas, pode-se constatar a viabilidade da abordagem proposta. Comparações são realizadas com os resultados apresentados por outras abordagens citadas na revisão bibliográfica.

O Capítulo 5 apresenta as conclusões e contribuições deste trabalho, além de, sugerir propostas para trabalhos futuros.

## Capítulo 2

# Trabalhos relacionados

Neste capítulo, são relacionados os principais trabalhos analisados ao longo da realização desta pesquisa. A bibliografia concentra-se em trabalhos aplicados ao processamento de documentos históricos. Diversos trabalhos têm proposto solução para o problema de aprimorar a qualidade de documentos degradados utilizando as mais diversas técnicas e abordagens. Porém, de forma muito genérica, pode-se dividí-los em dois grandes grupos:

1. Métodos de propósito geral: podem ser aplicados em qualquer tipo de imagem, pois não levam em consideração características específicas dos documentos;
2. Métodos específicos: utilizam características particulares dos documentos (como por exemplo, os *pixels*[1] do fundo são mais numerosos, problemas de iluminação, técnica de impressão utilizada), para buscar aprimoramentos nos resultados. Muitas abordagens são variações ou combinações de métodos de propósito geral.

A bibliografia selecionada apresenta exemplos de ambos os grupos. Alguns trabalhos foram selecionados por terem se consagrado como referências desta área de pesquisa, como é o caso dos métodos propostos por Otsu (1979) e Kapur (1985). De forma geral, os trabalhos concentram-se em duas grandes áreas de pesquisa, que são algoritmos de processamento digital de imagens (PDI) e Análise de Imagens de Documentos (DIA), do inglês *Document Image Analysis,* sendo que a última tem apresentado um cenário bastante efervescente no que se refere ao processamento de documentos históricos.

Vários dos algoritmos propostos utilizam a técnica de limiarização (ou binarização), que consiste em converter uma imagem em tons de cinza para uma imagem binária, ou seja, imagem apresentando apenas duas tonalidades (preto e

1 Do inglês: *PICTure ELement*

branco). Uma binarização ideal é capaz de separar perfeitamente o conteúdo textual do fundo, eliminando qualquer tipo de ruído que prejudique a legibilidade do documento. As imagens binárias são muito adequadas para o processamento posterior do documento como, por exemplo, o reconhecimento do conteúdo textual utilizando OCR. Porém, alguns trabalhos optam por manter as imagens em tons de cinza após realizarem o trabalho de restauração digital. Esta opção justifica-se por não ser necessário o processamento posterior ou por considerar imagens em tons de cinza mais adequadas ao acesso por parte do público (KAVALLIERATOU, 2006).

Os algoritmos não são apresentados em detalhes, buscando-se apenas descrevê-los de forma sucinta. O objetivo é aplicá-los às imagens do acervo, gerando, assim, referências visuais para a comparação dos mesmos.

## 2.1. Limiarização global

A limiarização ou binarização de imagens é uma grande área de pesquisa em processamento digital de imagem, que tem sido alvo dos esforços de diversos pesquisadores nos últimos 30 anos. A limiarização é uma importante abordagem para a segmentação de imagens. Gonzalez (2000) define a limiarização como uma operação que envolve testes de uma função $T$ da forma:

$$T = f(x, y, p(x, y)) \quad \textbf{(2.1-1)}$$

em que $f(x, y)$ é o nível de cinza do ponto $(x, y)$ e $p(x, y)$ denota alguma propriedade local desse ponto. Uma imagem limiarizada $g(x, y)$ é definida como:

$$g(x, y) = \begin{cases} 1 \; se \; f(x, y) > T \\ 0 \; se \; f(x, y) \leqslant T \end{cases} \quad \textbf{(2.1-2)}$$

Portanto, *pixels* rotulados como 1 (ou qualquer outro nível de cinza conveniente) correspondem ao fundo, enquanto que aqueles rotulados com 0 correspondem ao texto, tomando como referência um texto preto escrito sobre fundo branco. Quando $T$ depender apenas de $f(x, y)$, o limiar será chamado global. Se $T$ depender tanto de $f(x, y)$ quanto de $p(x, y)$, então o limiar será chamado dinâmico.

Diversos algoritmos de limiarização global já foram propostos utilizando as mais diversas abordagens. Em (SAHOO, 1988), pode-se verificar uma revisão sobre técnicas de limiarização. Lee, Chung e Park (LEE, 1990) apresentam um estudo comparativo

entre diversas técnicas de limiarização global, concentrando a avaliação em aspectos relativos à performance. Um dos mais antigos e conhecidos métodos de limiarização global é o proposto por Otsu (1979), que busca determinar um limiar de forma a maximizar a variância entre classes. A operação de limiarização consiste do particionamento dos *pixels* de uma imagem com *v* níveis de cinza em duas classes, *C0* e *C1*, que representam o objeto e o fundo, ou vice-versa, sendo que esta partição se dará no nível de cinza *t* , desta forma, tem-se:

$$C0 = \{0,1, \ldots , t\}$$
$$C1 = \{t + 1, t + 2, \ldots , v\text{-}1\}. \quad \textbf{(2.1-3)}$$

O número de pixels com nível *v* é denotado por $n_v$ e o número total de *pixels* é denotado por $N = n_0 + n_1 + \ldots + n_{V-1}$. Para simplificar, o histograma de níveis de cinza é normalizado e considerado com sendo a função de distribuição de probabilidade:

$$P_v = \frac{n_v}{N}, \quad p_v \geq 0, \quad \sum_{v=0}^{v-1} P_v = 1. \quad \textbf{(2.1-4)}$$

O método de Otsu escolhe *t* de forma a maximizar a variância inter-classes. Seja $\sigma_W^2$ a variância intraclasse, $\sigma_B^2$ a variância entre as classes e $\sigma_T^2$ a variância total. Um limiar ótimo pode ser obtido pela maximização de uma das funções critérios seguintes:

$$\lambda = \frac{\sigma_B^2}{\sigma_W^2}, \quad \textbf{(2.1-5)}$$

$$\eta = \frac{\sigma_B^2}{\sigma_T^2}, \quad \textbf{(2.1-6)}$$

$$\kappa = \frac{\sigma_T^2}{\sigma_W^2}. \quad \textbf{(2.1-7)}$$

O método de Otsu disponibiliza meios para se analisar outros aspectos além da seleção de um limiar ótimo para uma dada imagem, pode-se avaliar, por exemplo, a

separabilidade das classes *C0* e *C1* na imagem original ou a bimodalidade do histograma. Este método é considerado de uso geral e comumente utilizado como referência de comparação em trabalhos de processamento de documentos históricos. A Figura 2.1 apresenta uma imagem processada pelo algoritmo de Otsu. As Figuras 2.1.1 e 2.1.2 apresentam a ampliação dos documentos da Figura 2.1.

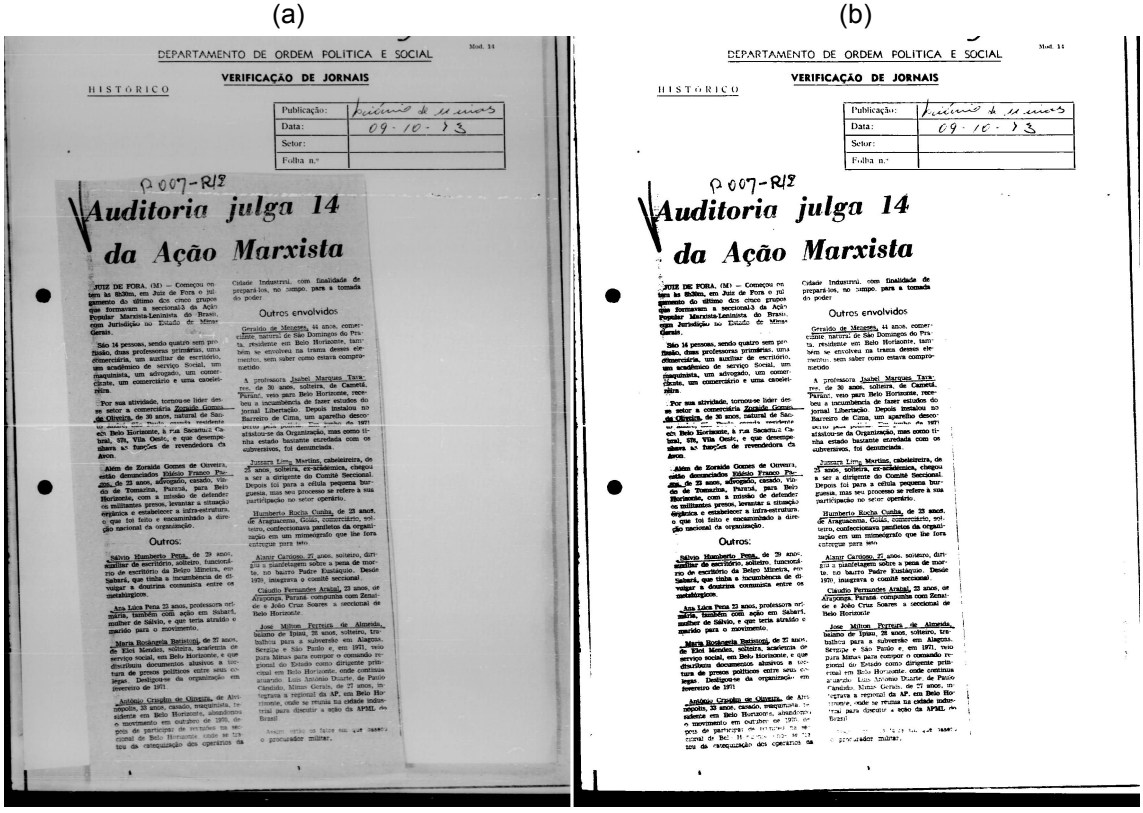
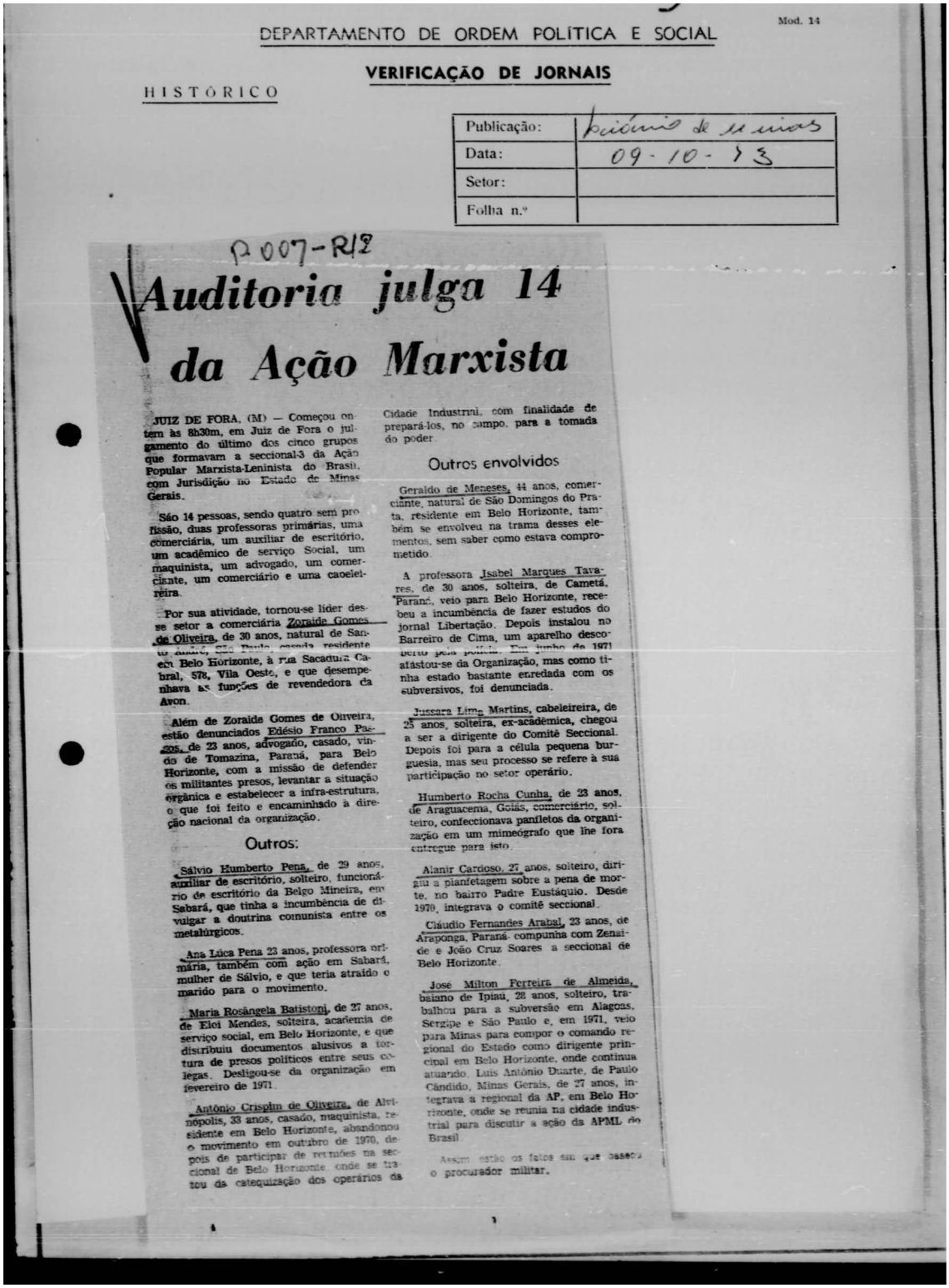



Figura 2.1 – Aplicação do algoritmo de Otsu. a) Imagem original em tons de cinza. b) Resultado da limiarização.



DEPARTAMENTO DE ORDEM POLÍTICA E SOCIAL

Mod. 14

VERIFICAÇÃO DE JORNAIS

HISTÓRICO

| | |
|---|---|
| Publicação: | Diário de Minas |
| Data: | 09-10-73 |
| Setor: | |
| Folha n.º | |

P 007-RJ

# Auditoria julga 14 da Ação Marxista

JUIZ DE FORA, (M) — Começou ontem às 8h30m, em Juiz de Fora o julgamento do último dos cinco grupos que formavam a seccional-3 da Ação Popular Marxista-Leninista do Brasil, com Jurisdição no Estado de Minas Gerais.

São 14 pessoas, sendo quatro sem profissão, duas professoras primárias, uma comerciária, um auxiliar de escritório, um acadêmico de serviço Social, um maquinista, um advogado, um comerciante, um comerciário e uma caoeleireira.

Por sua atividade, tornou-se líder desse setor a comerciária Zoraide Gomes de Oliveira, de 30 anos, natural de Santo [illegible], casada residente em Belo Horizonte, à rua Sacadura Cabral, 578, Vila Oeste, e que desempenhava as funções de revendedora da Avon.

Além de Zoraide Gomes de Oliveira, estão denunciados Edésio Franco Passos, de 23 anos, advogado, casado, vindo de Tomazina, Paraná, para Belo Horizonte, com a missão de defender os militantes presos, levantar a situação orgânica e estabelecer a infra-estrutura, o que foi feito e encaminhado à direção nacional da organização.

## Outros:

Sálvio Humberto Pena, de 29 anos, auxiliar de escritório, solteiro, funcionário de escritório da Belgo Mineira, em Sabará, que tinha a incumbência de divulgar a doutrina comunista entre os metalúrgicos.

Ana Luca Pena 23 anos, professora primária, também com ação em Sabará, mulher de Sálvio, e que teria atraído o marido para o movimento.

Maria Rosângela Batistoni, de 27 anos, de Elci Mendes, solteira, academia de serviço social, em Belo Horizonte, e que distribuiu documentos alusivos a tortura de presos políticos entre seus colegas. Desligou-se da organização em fevereiro de 1971.

Antônio Crispim de Oliveira, de Alvinópolis, 33 anos, casado, maquinista, residente em Belo Horizonte, abandonou o movimento em outubro de 1970, depois de participar de reuniões na seccional de Belo Horizonte, onde se tratou da catequização dos operários da Cidade Industrial, com finalidade de prepará-los, no campo, para a tomada do poder.

## Outros envolvidos

Geraldo de Meneses, 44 anos, comerciante, natural de São Domingos do Prata, residente em Belo Horizonte, também se envolveu na trama desses elementos, sem saber como estava comprometido.

A professora Isabel Marques Tavares, de 30 anos, solteira, de Cametá, Paraná, veio para Belo Horizonte, recebeu a incumbência de fazer estudos do jornal Libertação. Depois instalou no Barreiro de Cima, um aparelho descoberto pela polícia. Em junho de 1971 afastou-se da Organização, mas como tinha estado bastante enredada com os subversivos, foi denunciada.

Jussara Lima Martins, cabeleireira, de 25 anos, solteira, ex-acadêmica, chegou a ser a dirigente do Comitê Seccional. Depois foi para a célula pequena burguesia, mas seu processo se refere à sua participação no setor operário.

Humberto Rocha Cunha, de 23 anos, de Araguacema, Goiás, comerciário, solteiro, confeccionava panfletos da organização em um mimeógrafo que lhe fora entregue para isto.

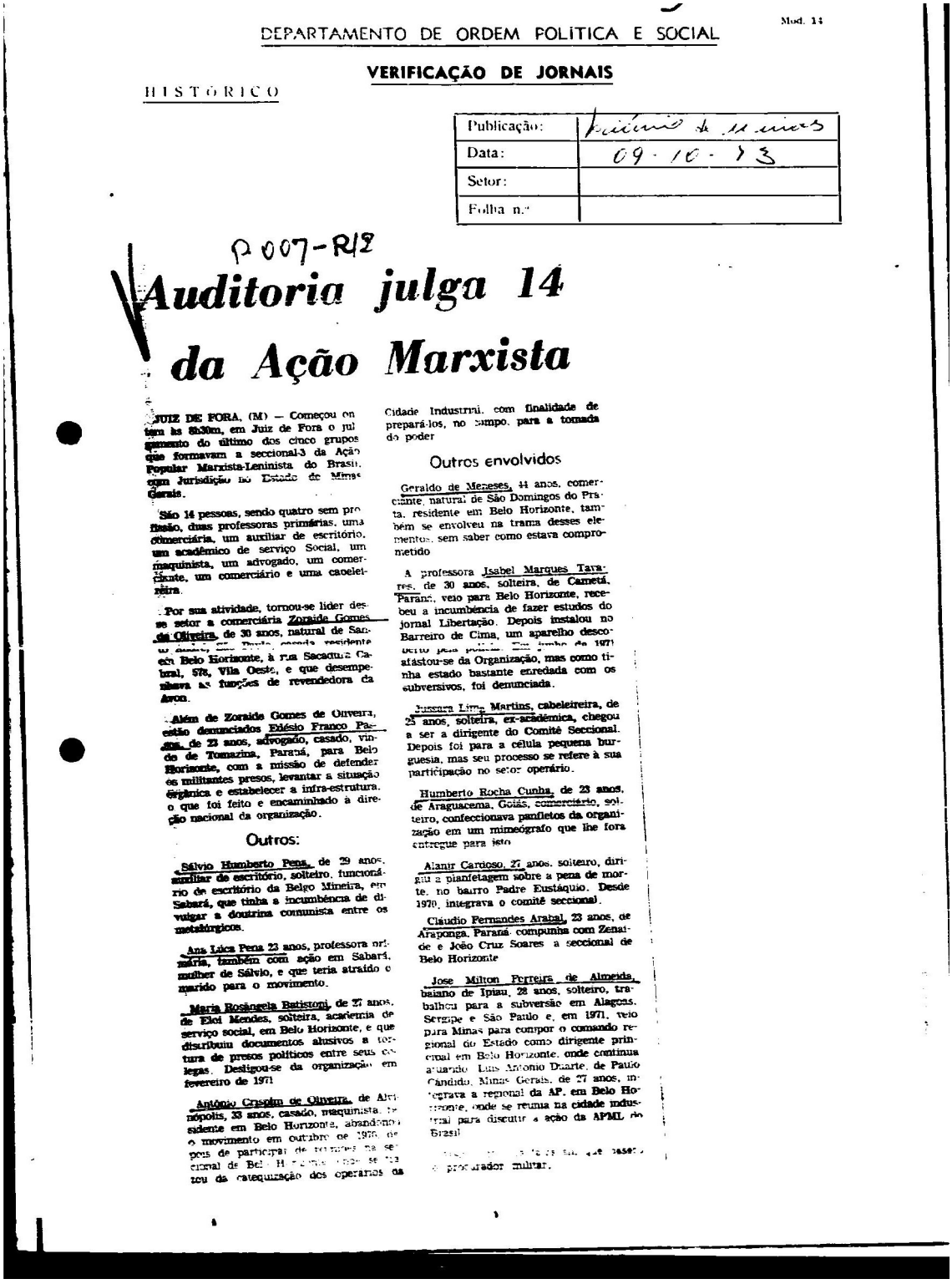
Alanir Cardoso, 27 anos, solteiro, dirigiu a pianfetagem sobre a pena de morte, no bairro Padre Eustáquio. Desde 1970, integrava o comitê seccional.

Cláudio Fernandes Arabal, 23 anos, de Araponga, Paraná, compunha com Zenaide e João Cruz Soares a seccional de Belo Horizonte.

José Milton Ferreira de Almeida, baiano de Ipiaú, 28 anos, solteiro, trabalhou para a subversão em Alagoas, Sergipe e São Paulo e, em 1971, veio para Minas para compor o comando regional do Estado como dirigente principal em Belo Horizonte, onde continua atuando. Luis Antônio Duarte, de Paulo Cândido, Minas Gerais, de 27 anos, integrava a regional da AP, em Belo Horizonte, onde se reunia na cidade industrial para discutir a ação da APML do Brasil.

Assim estão os fatos em que baseou o procurador militar.

Figura 2.1.1 – Imagem original em tons de cinza.



DEPARTAMENTO DE ORDEM POLITICA E SOCIAL

Mod. 14

**VERIFICAÇÃO DE JORNAIS**

HISTÓRICO

| | |
|---|---|
| Publicação: | Jornal do Brasil |
| Data: | 09-10-73 |
| Setor: | |
| Folha n.º | |

P 007-R/2

# Auditoria julga 14 da Ação Marxista

JUIZ DE FORA, (M) — Começou on tem às 8h30m, em Juiz de Fora o julgamento do último dos cinco grupos que formavam a seccional-3 da Ação Popular Marxista-Leninista do Brasil, com Jurisdição no Estado de Minas Gerais.

São 14 pessoas, sendo quatro sem profissão, duas professoras primárias, uma comerciária, um auxiliar de escritório, um acadêmico de serviço Social, um maquinista, um advogado, um comerciante, um comerciário e uma caoeleira.

Por sua atividade, tornou-se líder desse setor a comerciária Zoraide Gomes de Oliveira, de 30 anos, natural de San- [illegible] residente em Belo Horizonte, à rua Sacadura Cabral, 578, Vila Oeste, e que desempenhava as funções de revendedora da Avon.

Além de Zoraide Gomes de Oliveira, estão denunciados Edésio Franco Passos, de 23 anos, advogado, casado, vindo de Tomazina, Paraná, para Belo Horizonte, com a missão de defender os militantes presos, levantar a situação orgânica e estabelecer a infra-estrutura, o que foi feito e encaminhado à direção nacional da organização.

## Outros:

Sálvio Humberto Pena, de 29 anos, auxiliar de escritório, solteiro, funcionário de escritório da Belgo Mineira, em Sabará, que tinha a incumbência de divulgar a doutrina comunista entre os metalúrgicos.

Ana Lúcia Pena 23 anos, professora primária, também com ação em Sabará, mulher de Sálvio, e que teria atraído o marido para o movimento.

Maria Rosângela Batistoni, de 27 anos, de Elci Mendes, solteira, acadêmia de serviço social, em Belo Horizonte, e que distribuiu documentos alusivos a tortura de presos políticos entre seus colegas. Desligou-se da organização em fevereiro de 1971

Antônio Craspim de Oliveira, de Arinópolis, 33 anos, casado, maquinista, residente em Belo Horizonte, abandonou o movimento em outubro de 1956, depois de participar de reuniões na seccional de Bel[illegible] Horizonte, onde se tornou da catequização dos operários da Cidade Industrial, com finalidade de prepará-los, no campo, para a tomada do poder

## Outros envolvidos

Geraldo de Menezes, 44 anos, comerciante, natural de São Domingos do Prata, residente em Belo Horizonte, também se envolveu na trama desses elementos, sem saber como estava comprometido

A professora Isabel Marques Tavares, de 30 anos, solteira, de Cametá, Paraná, veio para Belo Horizonte, recebeu a incumbência de fazer estudos do jornal Libertação. Depois instalou no Barreiro de Cima, um aparelho descoberto pela polícia. Em junho de 1971 afastou-se da Organização, mas como tinha estado bastante enredada com os subversivos, foi denunciada.

Jussara Lima Martins, cabeleireira, de 23 anos, solteira, ex-acadêmica, chegou a ser a dirigente do Comitê Seccional. Depois foi para a célula pequena burguesia, mas seu processo se refere à sua participação no setor operário.

Humberto Rocha Cunha, de 23 anos, de Araguacema, Goiás, comerciário, solteiro, confeccionava panfletos da organização em um mimeógrafo que lhe fora entregue para isto

Alanir Cardoso, 27 anos, solteiro, dirigiu a panfletagem sobre a pena de morte, no bairro Padre Eustáquio. Desde 1970, integrava o comitê seccional).

Cláudio Fernandes Arabal, 23 anos, de Araponga, Paraná, compunha com Zenaide e Jcão Cruz Soares a seccional de Belo Horizonte

Jose Milton Ferreira de Almeida, baiano de Ipiau, 28 anos, solteiro, trabalhou para a subversão em Alagoas, Sergipe e São Paulo e, em 1971, veio para Minas para compor o comando regional do Estado como dirigente principal em Belo Horizonte, onde continua aguardo. Luis Antonio Duarte, de Paulo Candido, Minas Gerais, de 27 anos, integrava a regional da AP, em Belo Horizonte, onde se reunia na cidade industrial para discutir a ação da APML do Brasil

[illegible] que passou [illegible] procurador militar.

Figura 2.1.2 – Resultado da limiarização utilizando o algoritmo de Otsu.

## 2.2. Limiarização baseada em entropia

Considerando uma imagem digital com *M*×*N pixels*, define-se um conjunto de probabilidades *p* relativo aos níveis de cinza da imagem, ou seja, relativo à freqüência de ocorrência dos níveis de cinza para o conjunto total de *pixels* da imagem. O histograma *h* da imagem, normalizado pelo número total de *pixels* da imagem, representa uma boa aproximação para uma distribuição de probabilidades dos níveis de cinza em uma imagem digital. Kapur (1985) propôs um método de binarização utilizando o cálculo da entropia da imagem considerando que os *pixels* da imagem poderiam ser classificados como pertencentes ao objeto ou ao fundo. Ao definir duas classes estatisticamente independentes (objeto e fundo) como componentes de um sistema único (imagem digital), o método considera que as características de luminância dos *pixels* que compõem as classes objeto e fundo são independentes entre si.

O método da Soma de Entropias, proposto por Kapur, utiliza a propriedade de aditividade da entropia e baseia-se na maximização da medida de informação entre as duas classes. Considerando a imagem digital como um sistema que pode ser decomposto em dois subsistemas *A* e *B* estatisticamente independentes, Kapur define uma distribuição de probabilidade *A* para um objeto e uma distribuição *B* para o fundo da imagem, tal que:

$$p^A : \frac{p_1}{P_t}, \frac{p_2}{P_t}, \cdots, \frac{p_t}{P_t}, \qquad \textbf{(2.2-1)}$$
$$p^B : \frac{p_t+1}{1-P_t}, \frac{p_t+2}{1-P_t}, \cdots, \frac{p_k}{1-P_t}.$$

A entropia associada aos *pixels* pretos, *Hb*, e a entropia associada aos *pixels* brancos, *Hw*, são delimitadas pelo valor de corte *t*. O algoritmo sugere que *t* seja tal que maximize a função *H* = *Hb* + *Hw,* com *p*[*i*] de acordo com as distribuições *A* e *B*. No caso, *Hb* e *Hw* são determinadas pelas equações:

$$Hb = -\sum_{0}^{t} p[i]\log(p[i]), \qquad \textbf{(2.2-2)}$$

$$Hw = -\sum_{t+1}^{255} p[i]\log(p[i]). \qquad \textbf{(2.2-3)}$$

Para $0 \leqslant t \leqslant 255$ . Onde *p*[*i*] é a probabilidade da ocorrência dos *pixels* que apresentam o

tom de cinza *i*. Sendo calculado pela equação

$$p[i]=\frac{n_i}{N} \quad . \tag{2.2-4}$$

Onde $n_i$ é número de *pixels* com tom de cinza *i* e *N* é o número total de *pixels.*

A Figura 2.2 apresenta um documento e o resultado do processamento pelo algoritmo de Kapur. O método de Kapur preservou mais informações de contexto do documento em comparação com o método de Otsu. A Figura 2.2.1 apresenta a ampliação da imagem resultante da limiarização utilizando o algoritmo de Kapur.

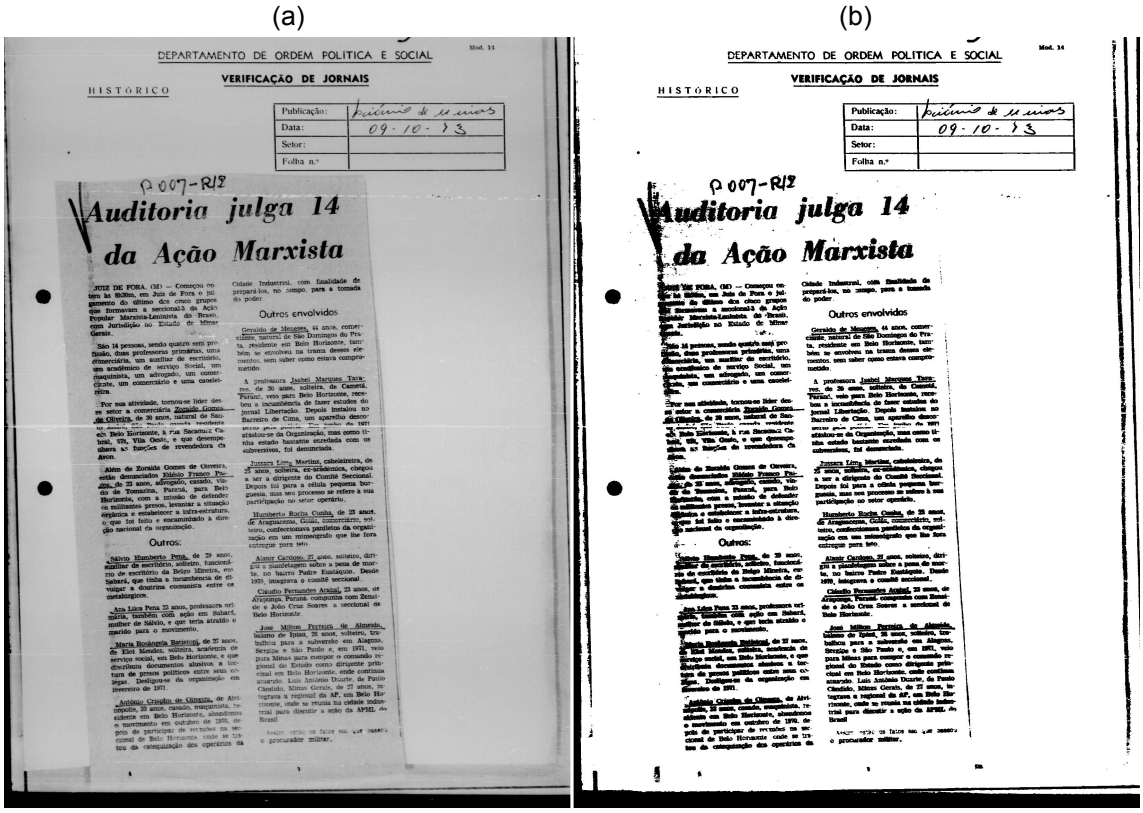
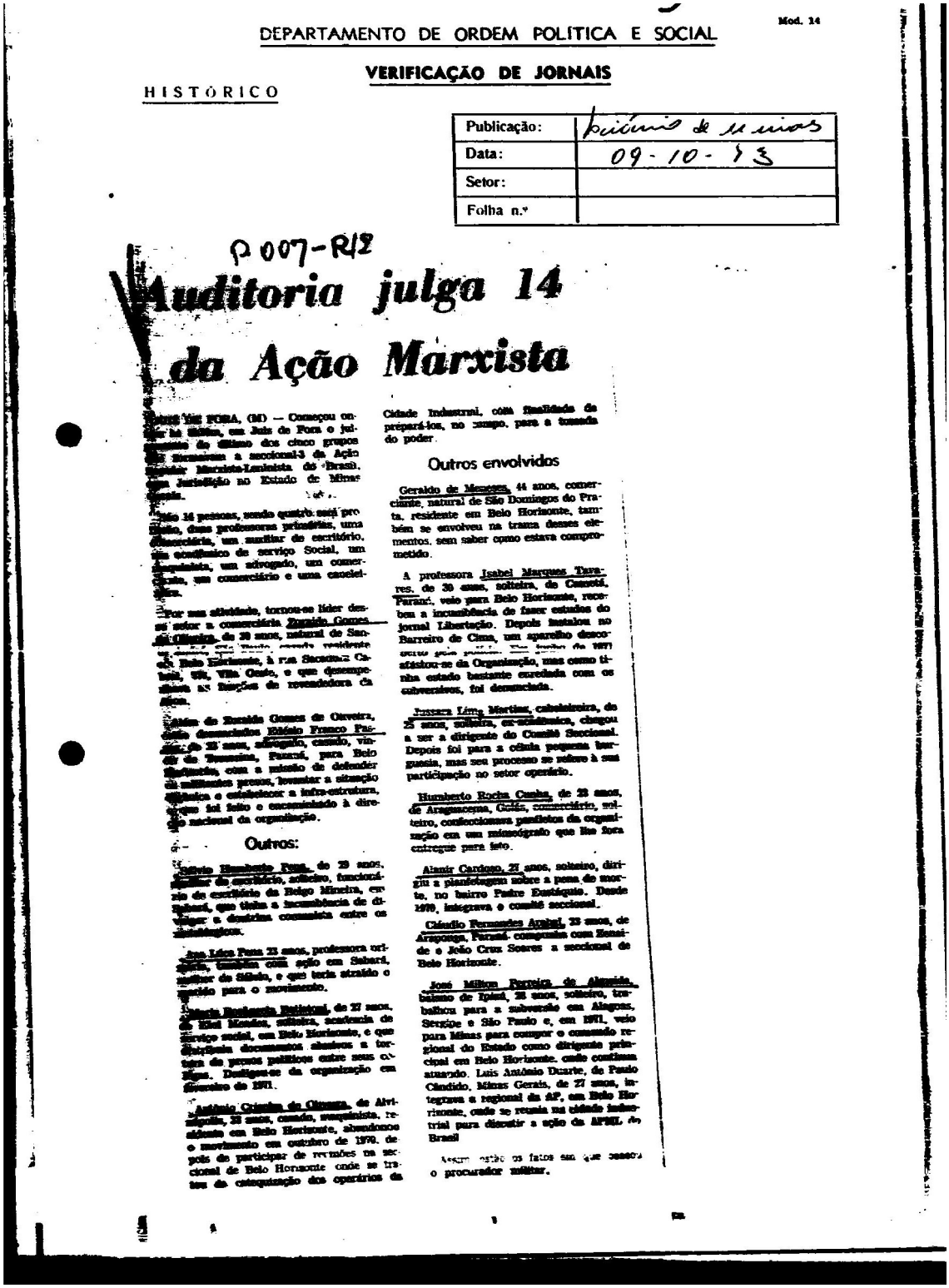



Figura 2.2 – Aplicação do algoritmo de Kapur. a) Imagem original em tons de cinza. b) Resultado da limiarização.



DEPARTAMENTO DE ORDEM POLÍTICA E SOCIAL

Mod. 14

VERIFICAÇÃO DE JORNAIS

HISTÓRICO

| | |
|---|---|
| Publicação: | Diário de Minas |
| Data: | 09-10-73 |
| Setor: | |
| Folha n.º | |

P 007-R/2

# Auditoria julga 14 da Ação Marxista

JUIZ DE FORA, (M) — Começou on-[tem] na 4ª Vara, em Juiz de Fora o julgamento de alguns dos cinco grupos que formavam a seccional-3 da Ação Libertadora Marxista-Leninista do Brasil, com jurisdição no Estado de Minas Gerais.

São 14 pessoas, sendo quatro sua prolação, duas professoras primárias, uma datilografa, um auxiliar de escritório, uma estudante de serviço Social, um [jornalista], um advogado, um comerciante, um comerciário e uma [estudante].

Por sua atividade, tornou-se líder deste setor a comerciária Zuzaida Gomes [de Oliveira], de 29 anos, natural de [Santa Rita] de Ibitipoca, residente em Belo Horizonte, à rua Sacramento Cabral, 926, Vila Cruzeiro, o que desempenhava as funções de revendedora da [área].

[Adalto] de Resende Gomes de Oliveira, [também denunciados Eldésio Franco Passos], de 23 anos, advogado, casado, vindo de Tocantins, Paraná, para Belo Horizonte, com a missão de defender os militantes presos, levantar a situação política e estabelecer a infra-estrutura, o que foi feito e encaminhado à direção nacional da organização.

## Outros:

[Celso Humberto Fonseca], de 29 anos, auxiliar de escritório, solteiro, funcionário de escritório da Belgo Mineira, em Ipatinga, que tinha a incumbência de divulgar a doutrina comunista entre os [companheiros].

[Ana Lúcia Pena] 23 anos, professora primária, também com ação em Sabará, auxiliar de Silvio, o que teria atraído o [jovem] para o movimento.

[Maria Bernardete Batistoni], de 27 anos, de Rio Monte, solteira, academia de serviço social, em Belo Horizonte, é que distribuía documentos alusivos a tortura de presos políticos entre seus colegas. Desligou-se da organização em fevereiro de 1971.

[Antônio Cristino da Cunha], de Alvinópolis, 28 anos, casado, ferroviário, residente em Belo Horizonte, abandonou o movimento em outubro de 1970, depois de participar de reuniões na seccional de Belo Horizonte onde se tratou da catequização dos operários da Cidade Industrial, com finalidade de prepará-los, no tempo, para a tomada do poder.

## Outros envolvidos

Geraldo de Menezes, 44 anos, comerciante, natural de São Domingos do Prata, residente em Belo Horizonte, também se envolveu na trama desses elementos, sem saber como estava comprometido.

A professora Isabel Marques Tavares, de 30 anos, solteira, de Cornélio, Paraná, veio para Belo Horizonte, recebeu a incumbência de fazer estudos do jornal Libertação. Depois instalou no Barreiro de Cima, um aparelho descoberto [pela polícia em 1971] afastou-se da Organização, mas como tinha estado bastante envolvida com os subversivos, foi denunciada.

Joanna Lima Martins, calculadora, de 25 anos, solteira, ex-estudante, chegou a ser a dirigente do Comitê Seccional. Depois foi para a célula pequena burguesia, mas seu processo se refere à sua participação no setor operário.

Humberto Rocha Cunha, de 28 anos, de Aragoiaçaba, Goiás, comerciário, solteiro, confeccionava panfletos da organização em um mimeógrafo que lhe fora entregue para isto.

Alamir Cardoso, 27 anos, solteiro, dirigiu a planejamento sobre a pena de morte, no bairro Padre Eustáquio. Desde 1970, integrava o comitê seccional.

Cláudio Fernandes Aranha, 23 anos, de Araponga, Paraná, congregou com Renaldo e João Cruz Soares a seccional de Belo Horizonte.

José Milton Ferreira de Almeida, baiano de Ipiaú, 28 anos, solteiro, trabalhou para a subversão em Alagoas, Sergipe e São Paulo e, em 1971, veio para Minas para compor o comando regional do Estado como dirigente principal em Belo Horizonte, onde continua atuando. Luis Antônio Duarte, de Paulo Cândido, Minas Gerais, de 27 anos, integrava a regional da AP, em Belo Horizonte, onde se reunia na cidade industrial para discutir a ação da APML do Brasil

[Acusam estão os fatos em que baseou] o procurador militar.

Figura 2.2.1 – Resultado da limiarização utilizando o algoritmo de Kapur.

## 2.3. Limiarização adaptativa

A limiarização adaptativa caracteriza-se por calcular diferentes limiares para cada *pixel* da imagem, através da análise das intensidades dos níveis de cinza dentro de janelas deslizantes. Este método também é conhecido como limiarização local ou limiar dinâmico. Os métodos mais conhecidos desta abordagem foram propostos por Niblack (1986) e Sauvola (2000).

O algoritmo proposto por Niblack calcula um limiar local utilizando o deslocamento de uma janela retangular pela imagem. O limiar $T$ para o *pixel* central da janela é calculado utilizando a média $m$ e a variância $s$ dos níveis de cinza presentes na janela:

$$T = m + k \cdot s \qquad \textbf{(2.3-1)}$$

onde $k$ é uma constante com valor -0,2, esse valor foi definido através de experimentos práticos. O valor de $k$ é utilizado para controlar quanto da borda dos objetos será considerada parte dos mesmos. Os resultados são pouco sensíveis ao tamanho da janela. Contudo, o ruído presente no fundo pode aparecer na imagem final binarizada. Em documentos com pouco conteúdo textual, uma quantidade significativa de ruído pode estar presente na imagem final. O método de limiarização adaptativa proposto por Sauvola (2000) resolve este problema adicionando a hipótese de níveis de cinza pré-determinados para os objetos e o fundo. O conteúdo textual apresenta níveis de cinza próximos de 0 e os *pixels* do fundo valores próximos a 255. Resultando na seguinte fórmula para o limiar:

$$T = m \cdot \left[ 1 + k \cdot \left( \frac{s}{R} - 1 \right) \right] \qquad \textbf{(2.3-2)}$$

onde $R$ é o intervalo dinâmico do desvio padrão fixado em 128 e a constante $k$ é fixada em 0,5. Os valores de $R$ e $k$ foram determinados por Sauvola através de experimentos práticos. Este método apresenta resultados melhores em imagens de documentos. O resultado da aplicação do método de Sauvola em uma imagem do acervo pode ser visto na Figura 2.3. A Figura 2.3.1 apresenta a ampliação da imagem resultante da limiarização utilizando o método de Sauvola.

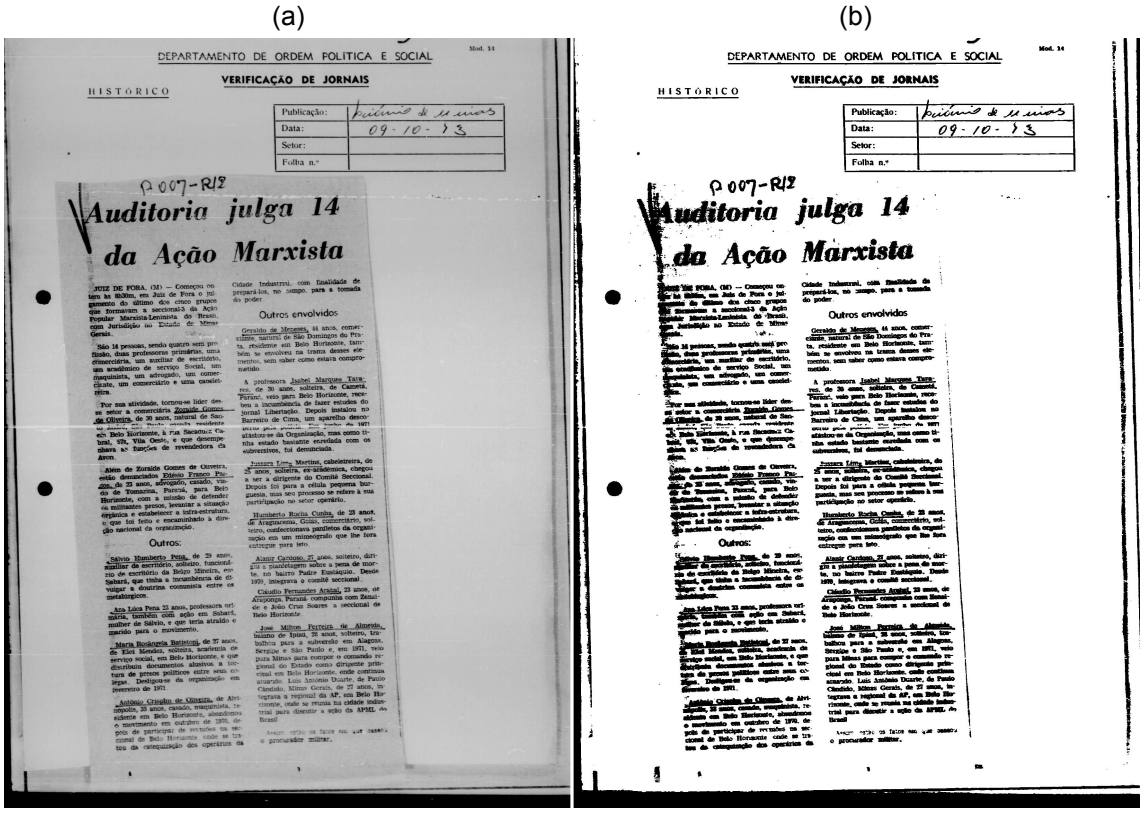
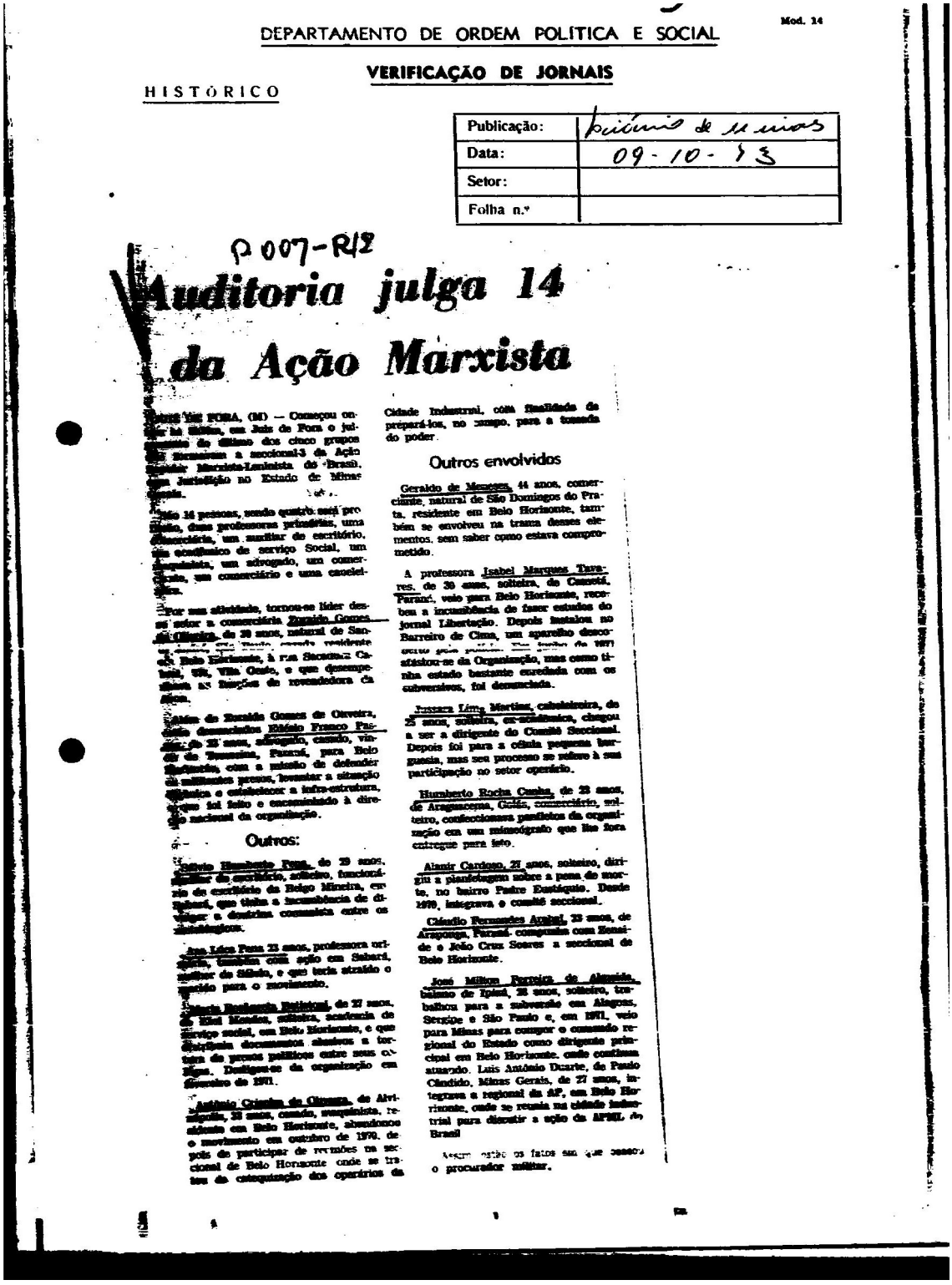



Figura 2.3 – Aplicação do algoritmo de Sauvola. a) Imagem original em tons de cinza. b) Resultado da limiarização.



DEPARTAMENTO DE ORDEM POLÍTICA E SOCIAL

Mod. 14

VERIFICAÇÃO DE JORNAIS

HISTÓRICO

| | |
|---|---|
| Publicação: | Diário de Minas |
| Data: | 09-10-73 |
| Setor: | |
| Folha n.º | |

P 007-R12

# Auditoria julga 14 da Ação Marxista

JUIZ DE FORA, (M) — Começou ontem na Auditoria, em Juiz de Fora o julgamento de alguns dos cinco grupos que formaram a seccional-3 da Ação Libertadora Marxista-Leninista do Brasil, com jurisdição no Estado de Minas Gerais.

São 14 pessoas, sendo quatro mulheres, duas professoras primárias, uma comerciária, um auxiliar de escritório, um assistente de serviço Social, um bancário, um advogado, um comerciante, um comerciário e uma enfermeira.

Por sua atividade, tornou-se líder deste setor a comerciária Zenaide Gomes de Oliveira, de 30 anos, natural de Santo [illegible], residente em Belo Horizonte, à rua Sacramento Cabral, 98, Vila Cruzeiro, e que desempenhava as funções de revendedora da célula.

Valdeir de Zenaide Gomes de Oliveira, citado denunciado Eldésio Franco Pacheco, de 25 anos, advogado, casado, vindo de Teresina, Piauí, para Belo Horizonte, com a missão de defender os militantes presos, levantar a situação política e estabelecer a infra-estrutura, o que foi feito e encaminhado à direção nacional da organização.

## Outros:

Otávio Humberto Pena, de 29 anos, auxiliar de escritório, solteiro, funcionário do escritório da Belgo Mineira, em Sabará, que tinha a incumbência de divulgar a doutrina comunista entre os trabalhadores.

Ana Lúcia Pena 23 anos, professora primária, também ação em Sabará, irmã de Otávio, e que teria atraído o marido para o movimento.

Maria Bernadete Botelho, de 27 anos, de Nova Mendes, solteira, assistente de serviço social, em Belo Horizonte, e que distribuía documentos abusivos a tortura de presos políticos entre seus colegas. Desligou-se da organização em fevereiro de 1971.

Antônio Cândido de Oliveira, de Alvinópolis, 28 anos, casado, vendedor, residente em Belo Horizonte, abandonou o movimento em outubro de 1970, depois de participar de reuniões na seccional de Belo Horizonte onde se tratou da catequização dos operários da Cidade Industrial, com finalidade de prepará-los, no tempo, para a tomada do poder.

## Outros envolvidos

Geraldo de Menezes, 44 anos, comerciante, natural de São Domingos do Prata, residente em Belo Horizonte, também se envolveu na trama desses elementos, sem saber como estava comprometido.

A professora Isabel Marques Tavares, de 30 anos, solteira, de Cascol, Paraná, veio para Belo Horizonte, recebeu a incumbência de fazer estudos do jornal Libertação. Depois instalou no Barreiro de Cima, um aparelho descoberto pelos [illegible] em 1971 afastou-se da Organização, mas como tinha estado bastante envolvida com os subversivos, foi denunciada.

Pedrara Lima Martins, cabeleireira, de 25 anos, solteira, ex-estudante, chegou a ser a dirigente do Comitê Seccional. Depois foi para a célula pequena burguesia, mas seu processo se refere à sua participação no setor operário.

Humberto Rocha Cunha, de 28 anos, de Araguacema, Goiás, comerciário, solteiro, confeccionava panfletos da organização em um mimeógrafo que lhe fora entregue para isto.

Alanir Cardoso, 27 anos, solteiro, dirigiu a planejamento sobre a zona de morte, no bairro Padre Eustáquio. Desde 1970, integrava o comitê seccional.

Cláudio Fernandes Arnaut, 23 anos, de Araponga, Paraná, companheiro com Benedito e João Cruz Soares a seccional de Belo Horizonte.

José Milton Barbosa de Almeida, baiano de Ipiaú, 28 anos, solteiro, trabalhou para a subversão em Alagoas, Sergipe e São Paulo e, em 1971, veio para Minas para compor o comando regional do Estado como dirigente principal em Belo Horizonte, onde continua atuando. Luis Antônio Duarte, de Paulo Cândido, Minas Gerais, de 27 anos, integrava a regional da AP, em Belo Horizonte, onde se reunia na cidade industrial para discutir a ação da APML do Brasil.

[illegible] os fatos em que [illegible] o procurador militar.

Figura 2.3.1 – Resultado da limiarização utilizando o algoritmo de Sauvola.

## 2.4. Processamento de histograma

Abordagens baseadas em processamento de histograma visam segmentar os elementos do documento realizando operações baseadas nas características do histograma da imagem. Os métodos desta classe utilizam técnicas de realce de contraste entre os *pixels* do texto e do fundo. Destaca-se a utilização de funções de transformação de histograma, como por exemplo a equalização. Vários métodos utilizam abordagens iterativas e geram imagens em tons de cinza, que deverão ser binarizadas posteriormente.

Em (KAVALLIERATOU, 2005b), um método especialmente desenvolvido para aprimorar a qualidade visual de documentos históricos é apresentado. O método, chamado *Iterative Global Thresholding* (IGT), consiste de procedimento iterativo baseado em equalização de histograma. O algoritmo parte da premissa que a maior parte dos *pixels* da imagem de um documento pertencem ao fundo e que raramente os *pixels* do texto (*pixels* pretos) excedem 10% do total. Desta forma, considerando-se uma imagem com contraste e brilho normais, a média das tonalidades de cinza será determinada primordialmente pelos *pixels* do fundo. Pode-se observar a aplicação do método em uma imagem do acervo na Figura 2.4. A Figura 2.4.1 apresenta a ampliação da imagem resultante da limiarização utilizando o método IGT.

O método iterativo proposto por Kavallieratou pode ser dividido em duas partes. Na primeira, o valor médio dos *pixels* é calculado e então subtraído de todos os *pixels* da imagem. Na segunda parte, é realizada a equalização do histograma, visando aumentar a escala dinâmica dos *pixels*. A cada iteração, são realizados os seguintes passos:

1. Calcular a média dos valores dos pixels (*Ti*) da imagem;
2. Subtrair *Ti* de todos os pixels da imagem;
3. Realizar a equalização do histograma;
4. Realizar os passos de 1 a 3, até a condição de parada, que será determinada à seguir;
5. Opcionalmente, pode-se binarizar a imagem resultante;

Durante a i-ésima iteração, a imagem do documento $I_i(x, y)$ será determinada pela equação:

$$I_i(x,y) = 1 - \frac{T_i - I_{i-1}(x,y)}{1 - E_i} \qquad \textbf{(2.4-1)}$$

Onde $I_{i-1}(x, y)$ é a imagem resultante da iteração anterior, $T_i$ é o limiar (média) cálculada na i-ésima iteração e $E_i$ é o menor valor de *pixel* antes da equalização do histograma. A conclusão do processamento iterativo (passo 4) é definida utilizando-se o seguinte critério:

$$\left|T_i - T_{i-1}\right| < 0.001. \qquad \textbf{(2.4-2)}$$

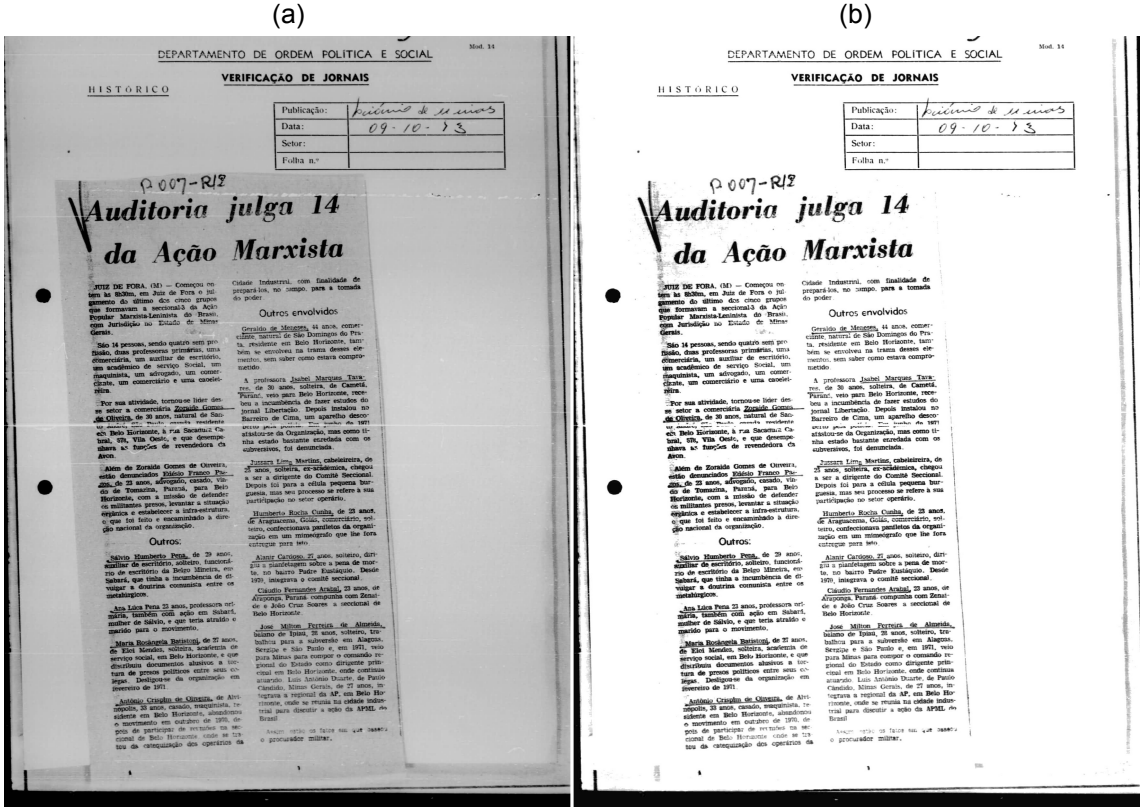
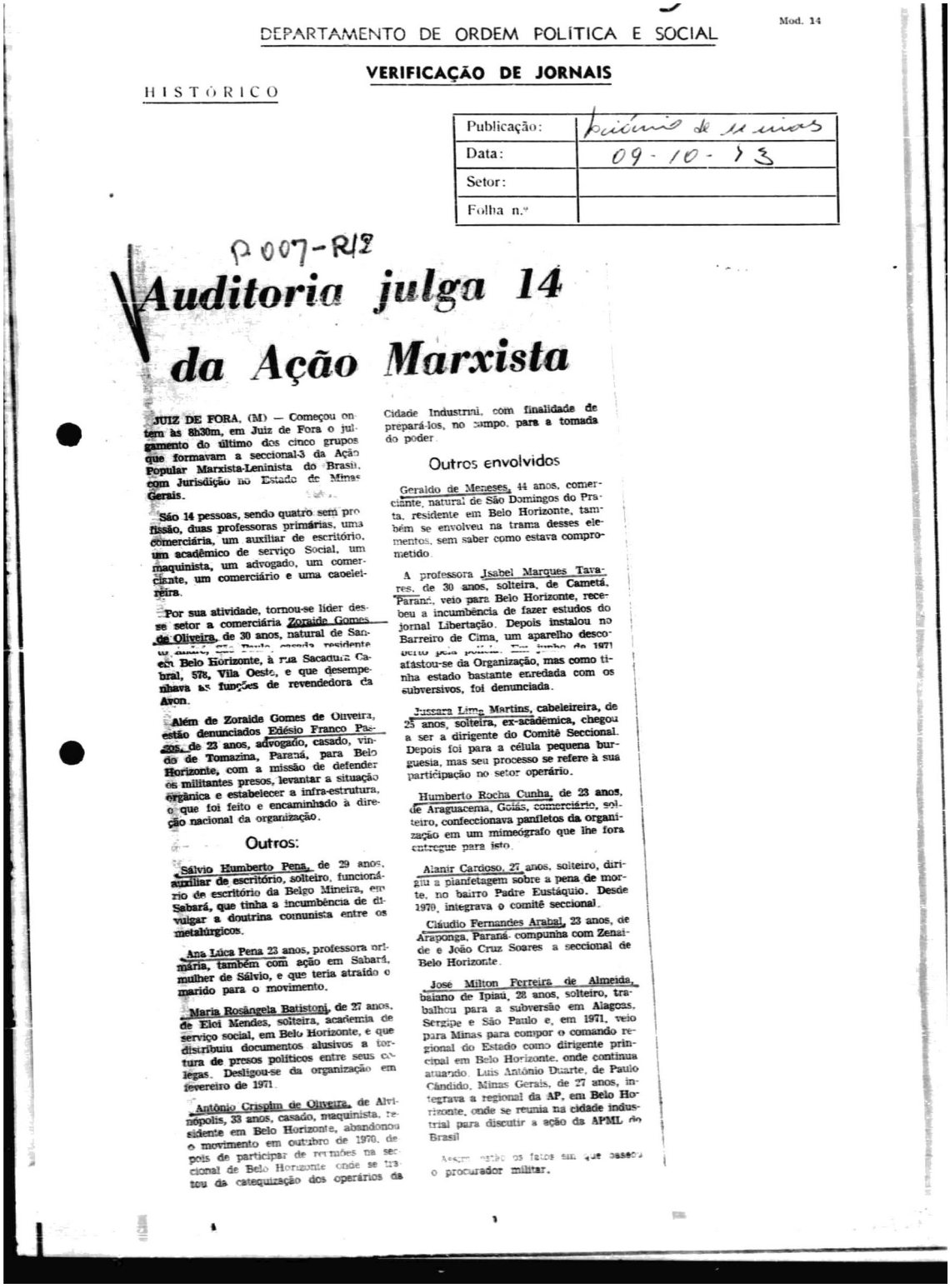



Figura 2.4 – Aplicação do algoritmo IGT. a) Imagem original em tons de cinza. b) Imagem processada em tons de cinza.

Esta abordagem funciona adequadamente em imagens de documentos históricos onde o texto é mais escuro que o fundo, ou seja, torna-se necessário um nível adequado de contraste para garantir bons resultados. Destaca-se que, neste exemplo, a imagem final (Figura 2.4) não foi binarizada (o passo 5 não foi realizado). Em algumas situações, torna-se oportuna a opção de manter a imagem final em tons de cinza, em especial, quando a imagem do documento será utilizada para leitura.

Em (KAVALLIERATOU, 2006), pode-se verificar proposta de melhoria desta abordagem. Destaca-se a combinação de técnicas globais e locais. O método de Kavallieratou apresenta implementação simples com resultados práticos significativos.



DEPARTAMENTO DE ORDEM POLÍTICA E SOCIAL

Mod. 14

**VERIFICAÇÃO DE JORNAIS**

HISTÓRICO

| | |
|---|---|
| Publicação: | Estado de Minas |
| Data: | 09-10-73 |
| Setor: | |
| Folha n.º | |

P 007-R/2

# Auditoria julga 14 da Ação Marxista

JUIZ DE FORA, (M) — Começou ontem às 8h30m, em Juiz de Fora o julgamento do último dos cinco grupos que formavam a seccional-3 da Ação Popular Marxista-Leninista do Brasil, com Jurisdição no Estado de Minas Gerais.

São 14 pessoas, sendo quatro sem profissão, duas professoras primárias, uma comerciária, um auxiliar de escritório, um acadêmico de serviço Social, um maquinista, um advogado, um comerciante, um comerciário e uma caoeleireira.

Por sua atividade, tornou-se líder desse setor a comerciária Zoraide Gomes de Oliveira, de 30 anos, natural de Santo [illegible], residente em Belo Horizonte, à rua Sacadura Cabral, 578, Vila Oeste, e que desempenhava as funções de revendedora da Avon.

Além de Zoraide Gomes de Oliveira, estão denunciados Edésio Franco Passos, de 23 anos, advogado, casado, vindo de Tomazina, Paraná, para Belo Horizonte, com a missão de defender os militantes presos, levantar a situação orgânica e estabelecer a infra-estrutura, o que foi feito e encaminhado à direção nacional da organização.

## Outros:

Sálvio Humberto Pena, de 29 anos, auxiliar de escritório, solteiro, funcionário de escritório da Belgo Mineira, em Sabará, que tinha a incumbência de divulgar a doutrina comunista entre os metalúrgicos.

Ana Lúcia Pena 23 anos, professora primária, também com ação em Sabará, mulher de Sálvio, e que teria atraído o marido para o movimento.

Maria Rosângela Batistoni, de 27 anos, de Eloi Mendes, solteira, academia de serviço social, em Belo Horizonte, e que distribuiu documentos alusivos a tortura de presos políticos entre seus colegas. Desligou-se da organização em fevereiro de 1971.

Antônio Crispim de Oliveira, de Alvinópolis, 33 anos, casado, maquinista, residente em Belo Horizonte, abandonou o movimento em outubro de 1970, depois de participar de reuniões na seccional de Belo Horizonte onde se tratou da catequização dos operários da Cidade Industrial, com finalidade de prepará-los, no campo, para a tomada do poder.

## Outros envolvidos

Geraldo de Meneses, 44 anos, comerciante, natural de São Domingos do Prata, residente em Belo Horizonte, também se envolveu na trama desses elementos, sem saber como estava comprometido.

A professora Isabel Marques Tavares, de 30 anos, solteira, de Cametá, Paraná, veio para Belo Horizonte, recebeu a incumbência de fazer estudos do jornal Libertação. Depois instalou no Barreiro de Cima, um aparelho descoberto pela polícia. Em junho de 1971 afastou-se da Organização, mas como tinha estado bastante enredada com os subversivos, foi denunciada.

Jussara Lima Martins, cabeleireira, de 25 anos, solteira, ex-acadêmica, chegou a ser a dirigente do Comitê Seccional. Depois foi para a célula pequena burguesia, mas seu processo se refere à sua participação no setor operário.

Humberto Rocha Cunha, de 23 anos, de Araguacema, Goiás, comerciário, solteiro, confeccionava panfletos da organização em um mimeógrafo que lhe fora entregue para isto.

Alanir Cardoso, 27 anos, solteiro, dirigiu a panfletagem sobre a pena de morte, no bairro Padre Eustáquio. Desde 1970, integrava o comitê seccional.

Cláudio Fernandes Arabal, 23 anos, de Araponga, Paraná, compunha com Zenaide e João Cruz Soares a seccional de Belo Horizonte.

José Milton Ferreira de Almeida, baiano de Ipiaú, 28 anos, solteiro, trabalhou para a subversão em Alagoas, Sergipe e São Paulo e, em 1971, veio para Minas para compor o comando regional do Estado como dirigente principal em Belo Horizonte, onde continua atuando. Luis Antônio Duarte, de Paulo Cândido, Minas Gerais, de 27 anos, integrava a regional da AP, em Belo Horizonte, onde se reunia na cidade industrial para discutir a ação da APML do Brasil.

Assim estão os fatos em que baseou o procurador militar.

Figura 2.4.1 – Resultado da aplicação do algoritmo IGT.

## 2.5. Outras abordagens

Devido à enorme diversidade de técnicas e abordagens propostas na literatura, torna-se extensa a tarefa de revisar a bibliografia referente à limiarização e segmentação de documentos. Esta seção cita, mesmo que de forma resumida, outros trabalhos que também apresentam resultados significativos no processamento de documentos.

Nos últimos anos, foram apresentados trabalhos propondo o desenvolvimento de algoritmos especialmente destinados ao processamento de documentos históricos. Gatos (2004) propõe um novo método que combina técnicas existentes. Um estudo comparativo de algoritmos de limiarização global em documentos degradados é apresentado por Leedham (2002). Uma interessante revisão sobre técnicas de processamento e análise de documentos manuscritos pode ser encontrada em (BRITTO JR., 2001). Diversos trabalhos concentram-se em implementar técnicas especializadas em eliminar anomalias e ruídos tipicamente presentes no fundo das imagens de documentos históricos. Este é o caso do algoritmo proposto por Shi (2005), onde uma função linear adaptativa é empregada para normalizar a intensidade de luz do fundo. Freqüentemente, encontramos documentos impressos em ambos os lados do papel e a tinta de um lado é visível do outro. Este fenômeno é conhecido como "interferência frente-verso". Silva (2006) apresenta uma nova técnica para remoção desta anomalia utilizando um algoritmo baseado em entropia. Técnicas relacionadas à segmentação combinada com limiarização têm sido empregadas em diversos trabalhos, como em (MATTANA, 1999), que apresenta um estudo utilizando imagens de cheques bancários. Droettboom (2003) propõe um método de reconstrução de caracteres baseado em grafos e aplica-o em documentos históricos degradados.

O Projeto Nabuco (NABUCO, 2006) é uma importante iniciativa nacional e relevante referência na pesquisa de técnicas computacionais aplicadas ao tratamento de imagens de documentos históricos. O projeto concentra-se no estudo de documentos de origem do século XIX, pertencentes ao acervo de Joaquim Nabuco, uma das figuras mais importantes na campanha de libertação dos escravos no Brasil (1861-1910). Vários trabalhos e algoritmos foram propostos visando realizar a limiarização dos documentos do acervo, como os apresentados em (SILVA, 2006; LINS, 1995).

## Capítulo 3

# Método Proposto

Documentos com alto grau de degradação são freqüentemente encontrados em acervos históricos. A ocorrência de anomalias nos documentos caracteriza-se como o maior empecilho para a implementação e aplicação de métodos automáticos de indexação e recuperação de informação, tais como a aplicação de *OCR*. Este trabalho, em particular, é motivado pelo problema de aprimorar a qualidade visual de documentos históricos. Especialmente, como forma de possibilitar futuros métodos de indexação automática para grandes acervos documentais.

Um algoritmo que utiliza uma abordagem híbrida, combinando características globais e locais, é proposto. O objetivo é aprimorar a qualidade visual da imagem do documento, através da correção de anomalias locais, eliminação do ruído de fundo e do aguçamento do contraste entre o texto e o fundo. O método proposto é dividido em quatro etapas, conforme apresentado na Figura 3.1.

Na primeira etapa, algumas características globais da imagem são extraídas. Utilizam-se características baseadas em medidas estatísticas, que são tomadas como parâmetros de referência nas próximas etapas. Na etapa seguinte, são calculadas medidas estatísticas para cada linha da imagem. Então, as linhas são classificadas como linhas que contêm texto ou somente fundo. A terceira etapa consiste na limiarização local apenas das linhas que contêm texto, utilizando a combinação de características locais e globais. O principal objetivo desta etapa é corrigir pequenas degradações dos caracteres do documento. Finalmente, é realizada uma limiarização global baseada no método proposto por Kavallieratou (2005). Esta etapa tem por finalidade aguçar o contraste entre texto e fundo e eliminar o ruído presente no fundo (papel). A Figura 3.2 apresenta um exemplo de imagem de documento histórico limiarizada através do método proposto.

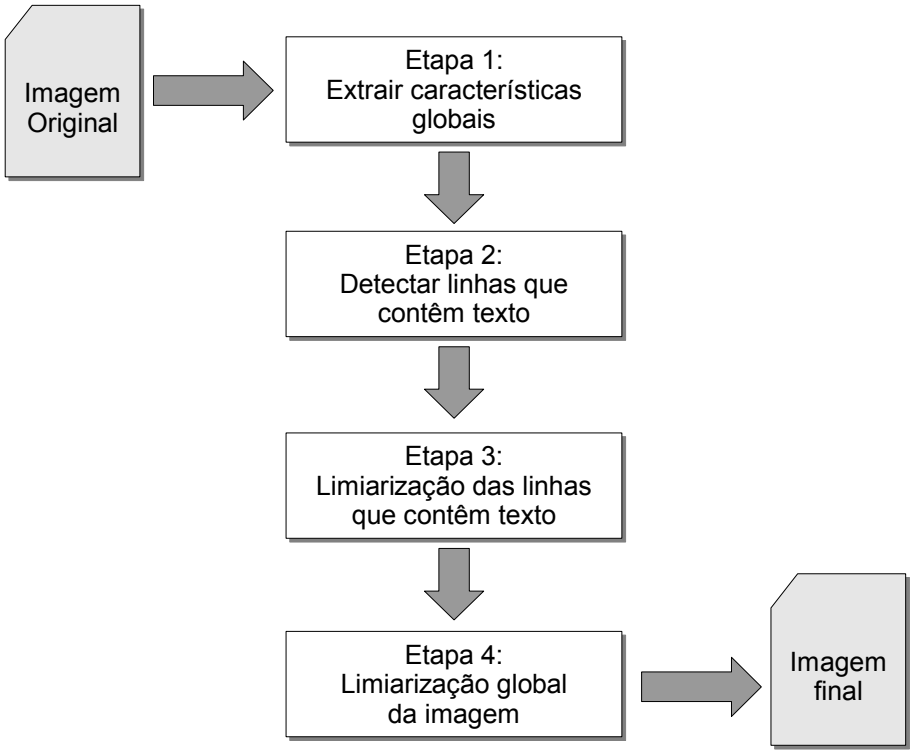
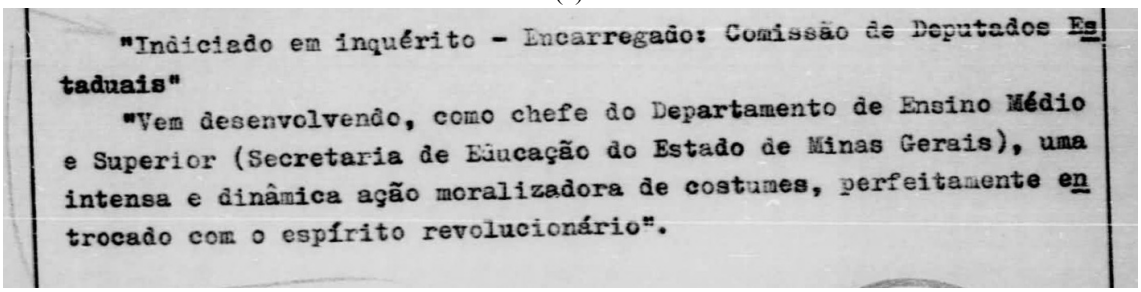


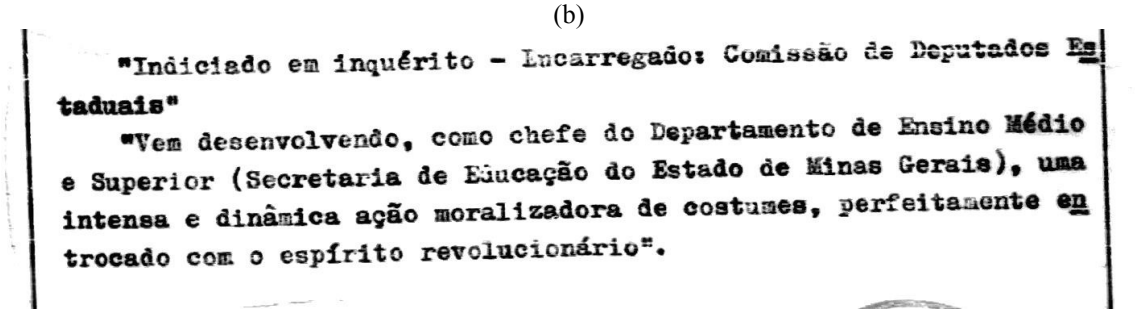

Figura 3.1 – Processo completo de limiarização de documentos históricos.

(a)


"Indiciado em inquérito - Encarregado: Comissão de Deputados Estaduais"

"Vem desenvolvendo, como chefe do Departamento de Ensino Médio e Superior (Secretaria de Educação do Estado de Minas Gerais), uma intensa e dinâmica ação moralizadora de costumes, perfeitamente entrocado com o espírito revolucionário".

(b)


"Indiciado em inquérito - Encarregado: Comissão de Deputados Estaduais"

"Vem desenvolvendo, como chefe do Departamento de Ensino Médio e Superior (Secretaria de Educação do Estado de Minas Gerais), uma intensa e dinâmica ação moralizadora de costumes, perfeitamente entrocado com o espírito revolucionário".

Figura 3.2 – Exemplo de imagem de documento histórico. a) Imagem original.
b) Imagem processada.

## 3.1. Etapa 1: Extrair características globais

A primeira etapa do método proposto utiliza diversas medidas estatísticas calculadas a partir das intensidades dos tons de cinza da imagem. Diversos trabalhos apresentados na literatura utilizam este mesmo princípio para inferir informações e/ou determinadas características da imagem. Gonzalez (2000, p.131) afirma que a média de intensidade e a variância (ou desvio padrão) são duas propriedades freqüentemente usadas devido a suas relevâncias para a aparência de uma imagem, considerando-se a média uma medida do brilho médio e a variância uma medida do contraste. Nos trabalhos de Niblack (1986) e Sauvola (2000), estas duas medidas estatísticas são utilizadas no cálculo da limiarização adaptativa. Kavallieratou (2005) afirma que em documentos com conteúdo essencialmente textual, pelo menos 10% dos *pixels* representam o texto (tonalidades escuras). Esta afirmação aponta para a utilização do percentil *P10* (10%) para categorizar documentos textuais. Porém, o estudo prático realizado, demonstrou ser mais apropriado utilizar o percentil *P5* (5%) como balizador das tonalidades dos *pixels* do texto. A mediana (ou percentil 50%) mostrou-se uma medida estatística significativa para documentos textuais, uma vez que pelo menos 50% dos *pixels* fazem parte do fundo (papel), desta forma apresentando valores superiores à mediana.

Na primeira etapa do método proposto, quatro medidas estatísticas são extraídas dos tons de cinza dos *pixels* da imagem, a saber:

- *s*: variância
- *m*: média
- *p5*: tom de cinza que abaixo do mesmo estão 5% dos pixels mais escuros da imagem.
- *p50 (mediana)*: tom de cinza que acima do qual encontram-se 50% dos pixels mais claros da imagem (fundo).

As medidas calculadas são utilizadas pelas etapas seguintes como valores de referência. Além disto, elas também são utilizadas para determinar se é adequada a aplicação do algoritmo em determinada imagem. Experimentos práticos demonstram que não é adequada a utilização do algoritmo em imagens muito esmaecidas (contraste demasiadamente baixo entre texto e fundo). Neste casos, a aplicação do algoritmo pode provocar deterioração da qualidade visual da imagem, podendo gerar artefatos

indesejados ao redor dos caracteres, intensificação de ruído de fundo e até mesmo tornando ainda mais ilegível certos caracteres muito esmaecidos. A imagem deve satisfazer a seguinte regra prática para a aplicação do algoritmo:

$$p50 - p5 \geqslant m - 2s \tag{3.1}$$

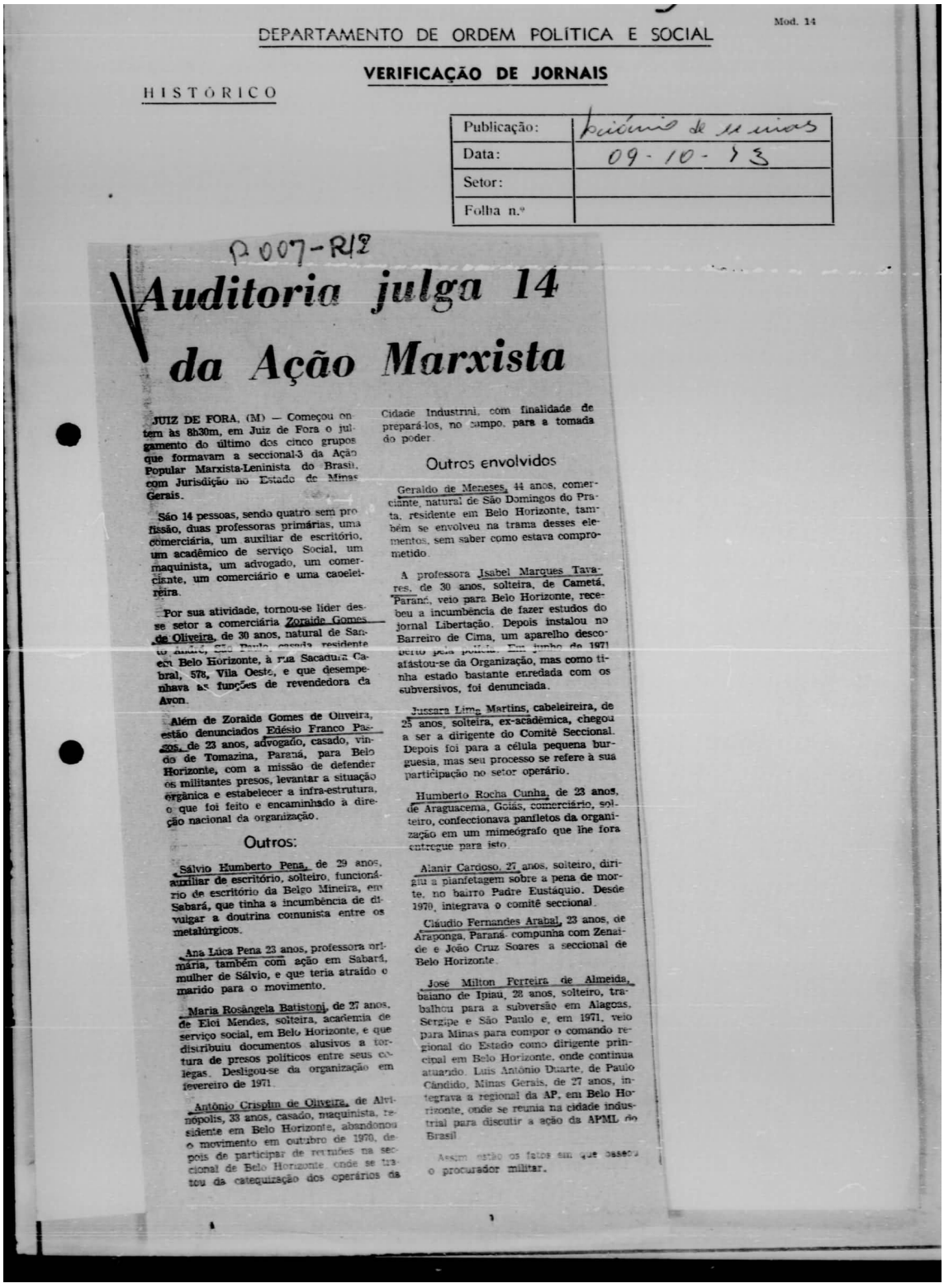
Na literatura, não foram encontradas referências sobre a definição de critérios de viabilidade para aplicação de técnicas de limiarização. Em geral, esta etapa fica atribuída a operadores humanos, como apresentado em (LINS, 1995). Porém, em se tratando de grandes acervos documentais, a definição de métodos automáticos torna-se de extrema importância.

O acervo do Dops/MG, além de muito extenso, possui itens documentais bastante heterogêneos quanto ao quesito qualidade visual. Em um mesmo rolo de microfilme, é comum haver documentos com qualidade visual satisfatória e outros totalmente degradados. A separação destes dois grupos é essencial para realizar de forma automática o tratamento das imagens do acervo. O critério proposto apresentou resultados satisfatórios na classificação das imagens do acervo. As Figuras 3.3 e 3.4 apresentam dois exemplos de imagens, um de cada grupo. Na Tabela 3.1, as medidas estatísticas globais de cada uma das imagens são apresentadas. De acordo com o critério estabelecido, o documento apresentado na Figura 3.3 satisfaz a Equação 3.1, enquanto o da Figura 3.2 não satisfaz.

A Figura 3.5 apresenta o resultado da aplicação do algoritmo proposto no documento apresentado na Figura 3.4. A utilização da Inequação 3.1 indica não ser adequada a aplicação do algoritmo neste documento. Porém, aplicou-se mesmo assim apenas para demonstrar o efeito negativo em determinadas imagens. Na Figura 3.5, observa-se que a legibilidade ficou ainda mais comprometida. A utilização da regra estabelecida pela Equação 3.1 mostrou-se efetiva na eliminação desse efeito indesejado do algoritmo, conforme atestam os experimentos apresentados no próximo capítulo.

Tabela 3.1 – Medidas estatísticas globais das imagens das Figuras 3.3 e 3.4.

| Imagem | m | s | p5 | p50 | p50 - p5 | m - 2s |
|---|---|---|---|---|---|---|
| Figura 3.3 | 165 | 49 | 26 | 180 | 154 | 67 |
| Figura 3.4 | 190 | 24 | 157 | 188 | 31 | 142 |



DEPARTAMENTO DE ORDEM POLÍTICA E SOCIAL

Mod. 14

**VERIFICAÇÃO DE JORNAIS**

HISTÓRICO

| | |
|---|---|
| Publicação: | Estado de Minas |
| Data: | 09-10-73 |
| Setor: | |
| Folha n.º | |

P 007-R/2

# Auditoria julga 14 da Ação Marxista

JUIZ DE FORA, (M) — Começou ontem às 8h30m, em Juiz de Fora o julgamento do último dos cinco grupos que formavam a seccional-3 da Ação Popular Marxista-Leninista do Brasil, com Jurisdição no Estado de Minas Gerais.

São 14 pessoas, sendo quatro sem profissão, duas professoras primárias, uma comerciária, um auxiliar de escritório, um acadêmico de serviço Social, um maquinista, um advogado, um comerciante, um comerciário e uma caoeleireira.

Por sua atividade, tornou-se líder desse setor a comerciária Zoraide Gomes de Oliveira, de 30 anos, natural de Santo André, São Paulo, casada residente em Belo Horizonte, à rua Sacadura Cabral, 578, Vila Oeste, e que desempenhava as funções de revendedora da Avon.

Além de Zoraide Gomes de Oliveira, estão denunciados Edésio Franco Passos, de 23 anos, advogado, casado, vindo de Tomazina, Paraná, para Belo Horizonte, com a missão de defender os militantes presos, levantar a situação orgânica e estabelecer a infra-estrutura, o que foi feito e encaminhado à direção nacional da organização.

## Outros:

Sálvio Humberto Pena, de 29 anos, auxiliar de escritório, solteiro, funcionário de escritório da Belgo Mineira, em Sabará, que tinha a incumbência de divulgar a doutrina comunista entre os metalúrgicos.

Ana Luca Pena 23 anos, professora primária, também com ação em Sabará, mulher de Sálvio, e que teria atraído o marido para o movimento.

Maria Rosângela Batistoni, de 27 anos, de Elci Mendes, solteira, acadêmia de serviço social, em Belo Horizonte, e que distribuiu documentos alusivos a tortura de presos políticos entre seus colegas. Desligou-se da organização em fevereiro de 1971.

Antônio Crispim de Oliveira, de Alvinópolis, 33 anos, casado, maquinista, residente em Belo Horizonte, abandonou o movimento em outubro de 1970, depois de participar de reuniões na seccional de Belo Horizonte, onde se tratou da catequização dos operários da Cidade Industrial, com finalidade de prepará-los, no tempo, para a tomada do poder.

## Outros envolvidos

Geraldo de Meneses, 44 anos, comerciante, natural de São Domingos do Prata, residente em Belo Horizonte, também se envolveu na trama desses elementos, sem saber como estava comprometido.

A professora Isabel Marques Tavares, de 30 anos, solteira, de Cametá, Paraná, veio para Belo Horizonte, recebeu a incumbência de fazer estudos do jornal Libertação. Depois instalou no Barreiro de Cima, um aparelho descoberto pela polícia. Em junho de 1971 afastou-se da Organização, mas como tinha estado bastante enredada com os subversivos, foi denunciada.

Jussara Lima Martins, cabeleireira, de 25 anos, solteira, ex-acadêmica, chegou a ser a dirigente do Comitê Seccional. Depois foi para a célula pequena burguesia, mas seu processo se refere à sua participação no setor operário.

Humberto Rocha Cunha, de 23 anos, de Araguacema, Goiás, comerciário, solteiro, confeccionava panfletos da organização em um mimeógrafo que lhe fora entregue para isto.

Alanir Cardoso, 27 anos, solteiro, dirigiu a panfetagem sobre a pena de morte, no bairro Padre Eustáquio. Desde 1970, integrava o comitê seccional.

Cláudio Fernandes Arabal, 23 anos, de Araponga, Paraná, compunha com Zenaide e João Cruz Soares a seccional de Belo Horizonte.

José Milton Ferreira de Almeida, baiano de Ipiaú, 28 anos, solteiro, trabalhou para a subversão em Alagoas, Sergipe e São Paulo e, em 1971, veio para Minas para compor o comando regional do Estado como dirigente principal em Belo Horizonte, onde continua atuando. Luis Antônio Duarte, de Paulo Cândido, Minas Gerais, de 27 anos, integrava a regional da AP, em Belo Horizonte, onde se reunia na cidade industrial para discutir a ação da APML do Brasil.

Assim estão os fatos em que baseou o procurador militar.

Figura 3.3 – Imagem de documento que satisfaz a Equação 3.1.

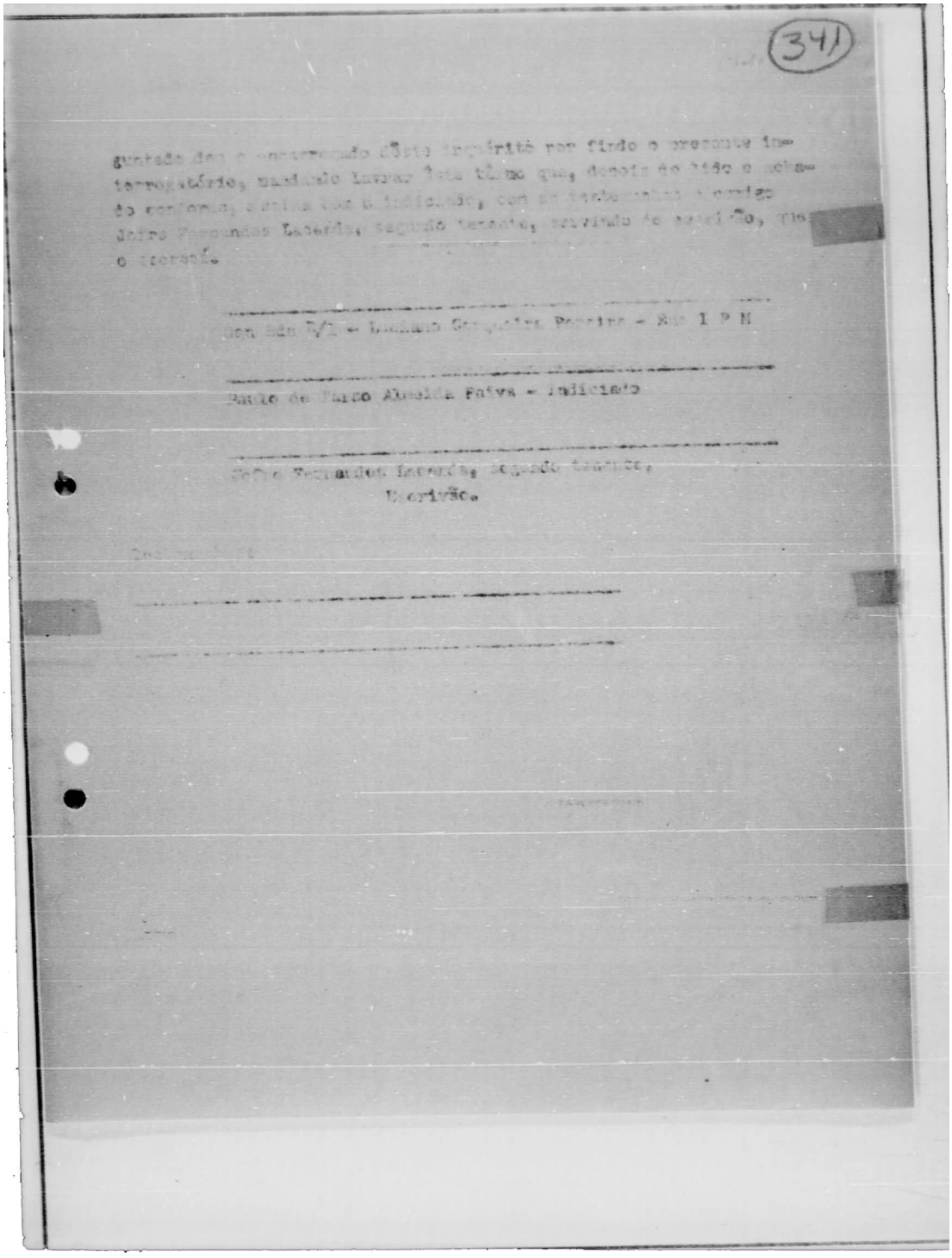



Figura 3.4 – Imagem de documento que não satisfaz a Equação 3.1.

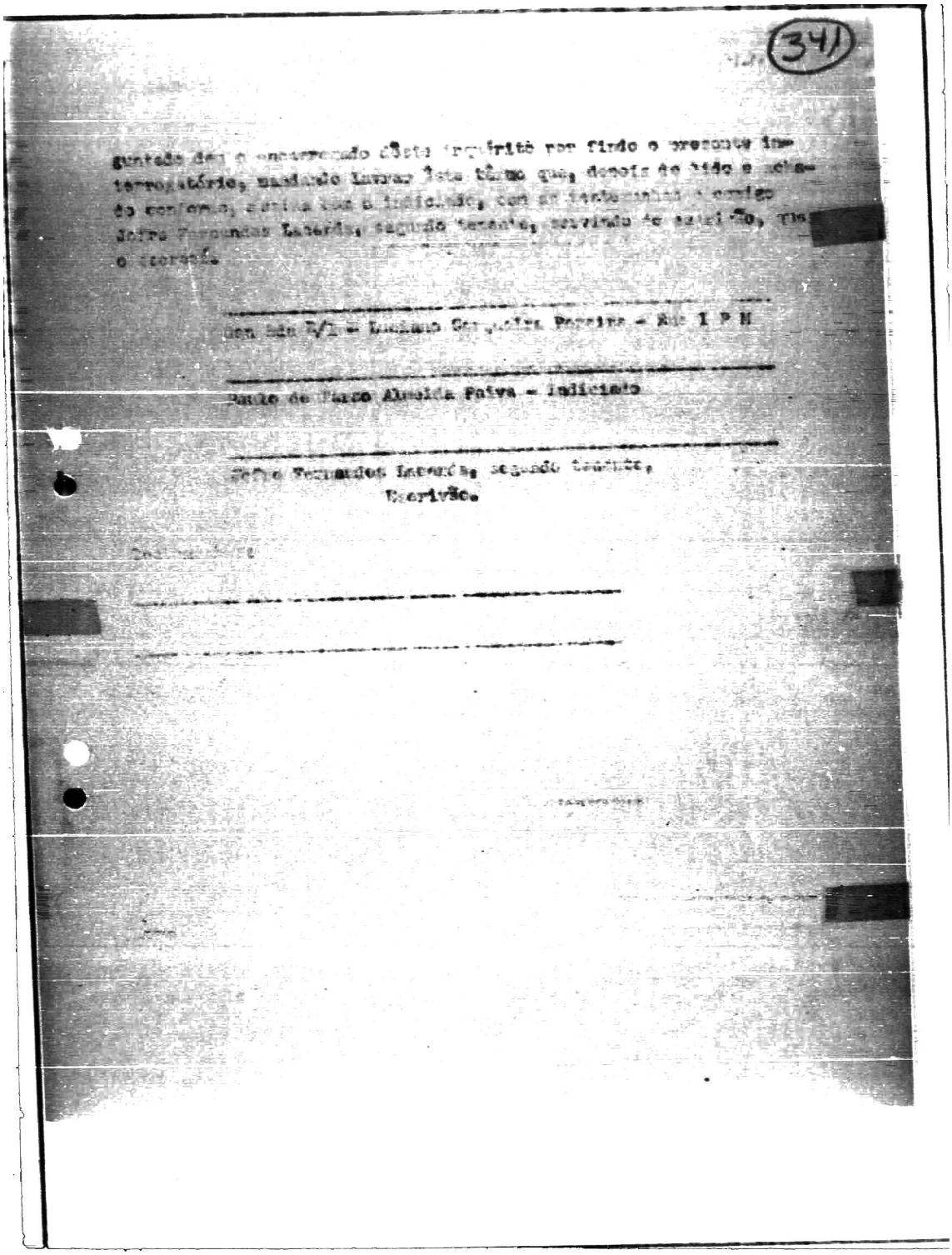


Figura 3.5 – Resultado da aplicação do algoritmo no documento apresentado na Figura 3.4.

## 3.2. Etapa 2: Detectar linhas que contêm texto

A segunda etapa consiste em classificar cada linha horizontal do documento (com altura de 1 *pixel*) como contendo ou não texto. Um método bastante simples foi desenvolvido para realizar esta tarefa. O método proposto é resultado de observação e exaustivos testes nos documentos do acervo, além de, basear-se na premissa que linhas que contêm texto possuem alta freqüência espacial, devido à ocorrência dos caracteres. Esta identificação é realizada através da extração e análise de duas características estatísticas de cada linha horizontal da imagem: a média e a moda. Uma varredura da linha é realizada e as duas medidas são calculadas. A linha é identificada como contendo texto se a diferença entre a moda e a média for superior a uma constante *C1*, conforme apresentado na Inequação 3.2:

$$moda - média > C1. \tag{3.2}$$

Pode-se explicar este resultado pelo fato da moda ser determinada predominantemente pelos *pixels* do fundo, que ocorrem em maior quantidade (mesmo nas linhas de texto) como pode ser observado nos histogramas da Figura 3.6. A maior ocorrência de texto em uma determinada linha provoca a diminuição do valor da média, deslocando-a da moda. Isto ocorre devido ao texto ser mais escuro que o fundo, desta forma os *pixels* do texto apresentam valores menores do que os apresentados pelo fundo. Na Figura 3.6b, o texto é representado pela cauda do histograma que se estende pela região dos tons de cinza mais escuros. Comparando os histogramas apresentados na Figura 3.6, pode-se verificar que em ambos ocorre uma grande concentração de *pixels* na região de tonalidades mais claras, representando o fundo do documento. Em documentos textuais, a moda será definida por este conjunto de *pixels* do fundo. Ressalta-se o fato de todas estas observações serem válidas apenas para documentos textuais, onde o fundo é representado por tonalidades claras e o texto por tonalidades escuras.

Resultados práticos demonstram que, ao atribuir-se o valor 20 a *C1,* as linhas que contêm texto são identificadas satisfatoriamente. Um exemplo da identificação de linhas que contêm texto pode ser visto na Figura 3.7. Com a finalidade de destacar o resultado desta etapa do método, o fragmento de documento apresentado na Figura 3.2a foi

processado. Utilizou-se a seguinte definição: para as linhas que satisfazem a Equação 3.2, ou seja aquelas que contêm texto, os *pixels* foram subtraídos de uma constante *Cb*, definida através de experimentos práticos, tornando-se mais escuros. Enquanto para as linhas que não satisfazem a Equação 3.2, classificadas como não contendo texto, os *pixels* foram convertidos para a cor branca. Cabe destacar que este procedimento foi utilizado apenas para apresentar o resultado parcial da detecção das linhas que contêm texto. No método proposto, as linhas são apenas classificadas nesta etapa sem realizar nenhuma alteração nos tons de cinza dos *pixels*. O fragmento de documento da Figura 3.2a, utilizado neste exemplo de detecção de linhas de texto, apresenta as linhas de texto levemente inclinadas. Pode-se observar na Figura 3.7 que esta inclinação não prejudicou significativamente a detecção das linhas de texto.

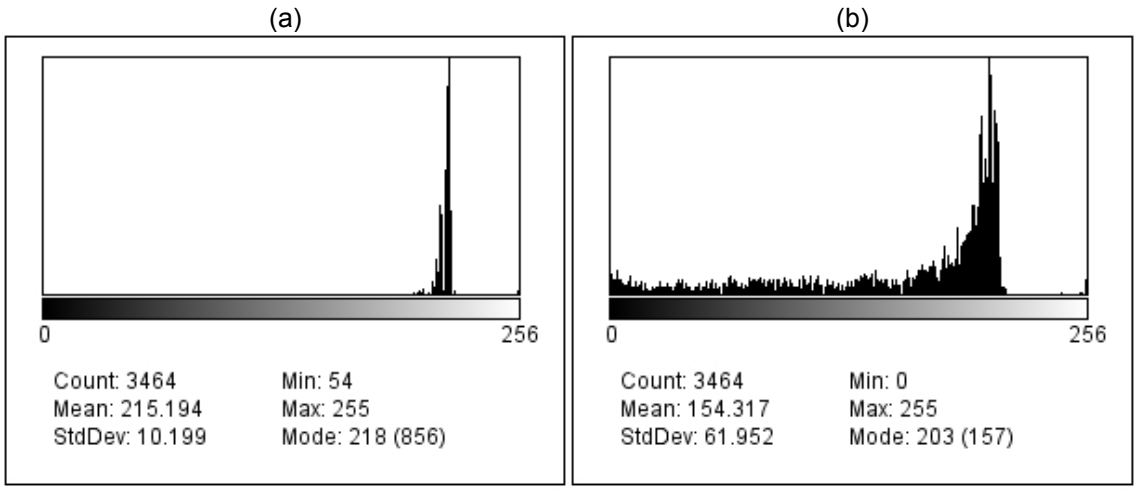



Figura 3.6 – Exemplos de histogramas de linhas horizontais do documento da Figura 3.2a.
a) Histograma de linha que não contem texto. b) Histograma de linha que contem texto.

O método proposto pode sofrer uma pequena variação para diferenciar linhas que contêm muito ou pouco texto. Pode-se obter este resultado utilizando duas constantes *C11* e *C12* , em vez de apenas uma, na Equação 3.2. Se a diferença entre a moda e média for inferior a *C11* a linha será classificada como não contendo texto (apenas fundo), se a diferença ficar no intervalo entre *C11* e *C12* será classificada como contendo pouco texto e finalmente se a diferença for maior que *C12* a linha será classificada como contendo muito texto. Testes práticos demonstraram que ao atribuir os valores 10 e 20 para *C11* e *C12,* respectivamente, a detecção é realizada de maneira satisfatória. Porém, este método de classificação utilizando duas constantes não foi

utilizado neste trabalho por não ser relevante para as próximas etapas. Optou-se apenas em classificar as linhas como contendo ou não texto.

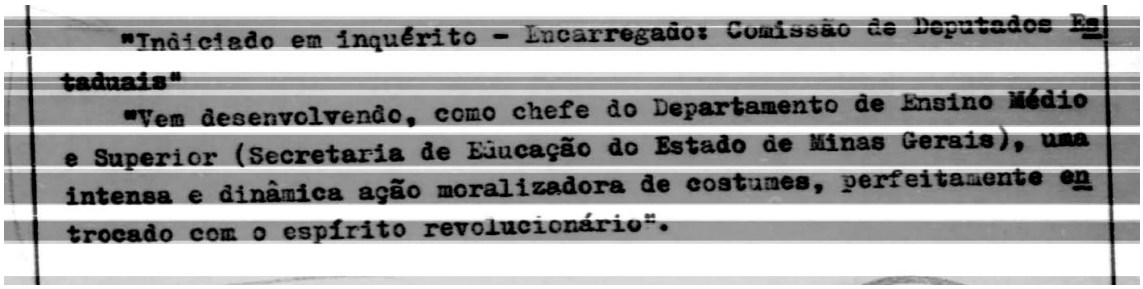



Figura 3.7 – Detecção das linhas que contêm texto.

Em geral, trabalhos que utilizam a limiarização adaptativa, ou seja, cálculo local do limiar para cada *pixel*, apresentam o inconveniente de segmentar determinadas áreas escuras do fundo como texto. Pequenos artefatos ou mesmo manchas escuras podem ser gerados na imagem binarizada. Este problema ocorre mesmo em áreas do documento, onde não se verifica a ocorrência de nenhum conteúdo textual. Em geral, estas anomalias estão associadas a problemas de iluminação ou variações de tonalidade do papel (devido às manchas ou ação do tempo). A inserção de uma etapa de detecção de linhas que contêm texto visa reduzir a área de aplicação da limiarização local. E por conseguinte, minimizar os efeitos nocivos da limiarização local sobre os *pixels* do fundo. Após identificadas as linhas que contêm texto, o método de limiarização local será aplicado apenas nelas. Além disto, pode-se implementar métodos de limiarização local que se beneficiem do conhecimento prévio de serem aplicados apenas em regiões onde ocorre conteúdo textual. O método proposto, neste trabalho, para detectar linhas que apresentam conteúdo textual possui diversas limitações, podendo apresentar pequenos erros na detecção em determinados casos. Porém, devido a simplicidade e baixo custo de tempo, mostrou-se bastante adequado para o escopo deste trabalho. A implementação desta etapa, também propicia uma redução do tempo de execução da limiarização local (próxima etapa), ao reduzir consideravelmente a quantidade de *pixels* a ser processada.

## 3.3. Etapa 3: Limiarização das linhas que contêm texto

Nesta etapa, as linhas horizontais da imagem que contêm texto, detectadas na etapa anterior, são limiarizadas através de um método que combina características locais e globais. O desenvolvimento do método baseou-se na limiarização adaptativa proposta por Niblack (1986) e no realce de bordas através da detecção de rampas proposto por Petrou (1991). Os trabalhos citados foram utilizados como fonte de idéias, porém, a implementação do método proposto não mantém relação estreita com as implementações destes trabalhos.

Inicialmente, o limiar $T$ da linha é calculado utilizando a média $m_l$ e a variância $s_l$ dos níveis de cinza presentes na linha e a média $m$ dos níveis de cinza da imagem calculada na primeira etapa do método:

$$T = m_l - \frac{m}{m_l} \cdot s_l . \tag{3.3}$$

Inicialmente, todos os *pixels* da linha que possuem valores superiores a $T$ têm seus valores adicionados à variância da linha. Esta medida suaviza o ruído de fundo do papel ao deslocar os *pixels* em direção à região de tonalidades mais claras.

O próximo passo consiste em detectar degraus. Uma varredura da esquerda para a direita é realizada nos *pixels* da linha a procura de variações decrescentes nos tons de cinza (rampa de descida) ou variações crescentes (rampa de subida). Considera-se como rampa qualquer variação absoluta superior à metade da variância *sl* da linha. A Figura 3.8 apresenta um esquema da detecção de rampas. O intervalo entre uma rampa de descida e uma rampa de subida consecutivas é considerado como um possível caracter ou parte dele. Para ser considerada como parte do texto, a área compreendida entre o final da rampa de descida e o início da rampa de subida deverá satisfazer as seguintes condições:

1. A profundidade da rampa de descida deve ser superior ou igual à variância *sl* da linha.
2. A média dos tons de cinza dos *pixels* da área deve ser inferior a *T*.

Caso as condições anteriores sejam satisfeitas, a área tem seus valores subtraídos da variância *sl* da linha. Optou-se por subtrair a variância da linha para suavizar o efeito

deste procedimento. Testes práticos demonstraram que a intensificação acentuada do contraste nesta etapa prejudica consideravelmente os resultados da etapa seguinte (limiarização global). Destaca-se o problema causado pela perda de detalhes em caracteres e outros elementos dos documentos, como por exemplo carimbos e anotações manuscritas.

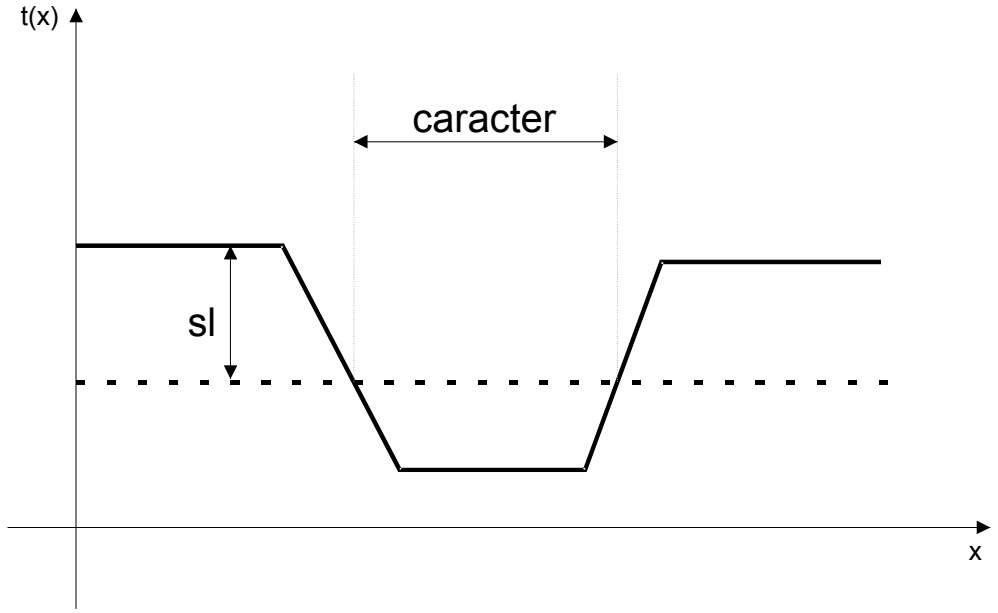



Figura 3.8 – Esquema de detecção de rampas.

A aplicação deste método corrige pequenas imperfeições em caracteres e equaliza pequenas áreas claras do documento. Nos documentos do acervo estudado, pequenas manchas ou linhas claras são comumente encontradas. Esta etapa do algoritmo corrige ou ameniza essas anomalias. Em determinados casos, as anomalias são tão graves que torna-se inviável sua correção sem causar danos maiores nas suas adjacências. A Figura 3.9 apresenta um exemplo da aplicação do método em um fragmento de texto. Os histogramas das imagens apresentadas nas Figuras 3.9a e 3.9d podem ser vistos na Figura 3.10. Observa-se no histograma da Figura 3.10b que os *pixels* foram sensivelmente deslocados para as extremidades do espectro.

Segue um resumo dos passos executados nesta etapa:

1. Calcular o limiar $T$ de acordo com a Equação 3.3;
2. Adicionar $s_l$ a todos os pixels que possuem valor superior a $T$;
3. Detectar a ocorrência de possíveis caracteres através de identificação de rampas descendentes e ascendentes;

4. Caso seja identificada como conteúdo textual, a área compreendida entre as rampas descendente e ascendente devem ter seus valores subtraídos de *sl*.

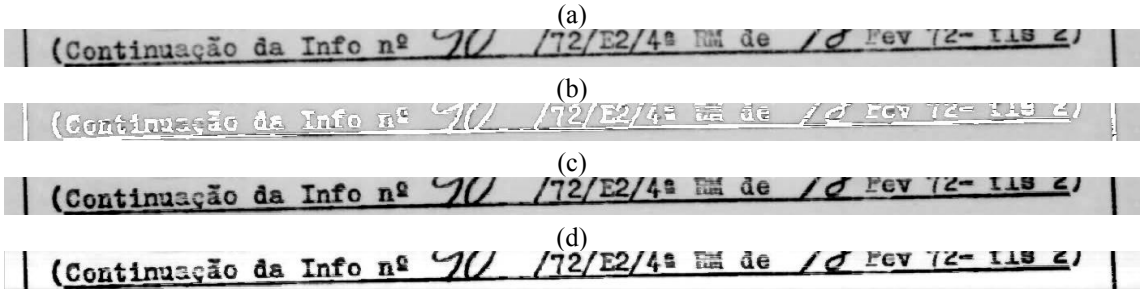



Figura 3.9 – Exemplo da aplicação da Etapa 3. a) Imagem original. b) Segmentação de caracteres pelo método de detecção de rampas. c) Escurecimento dos caracteres sem remoção do fundo. d) Resultado final.

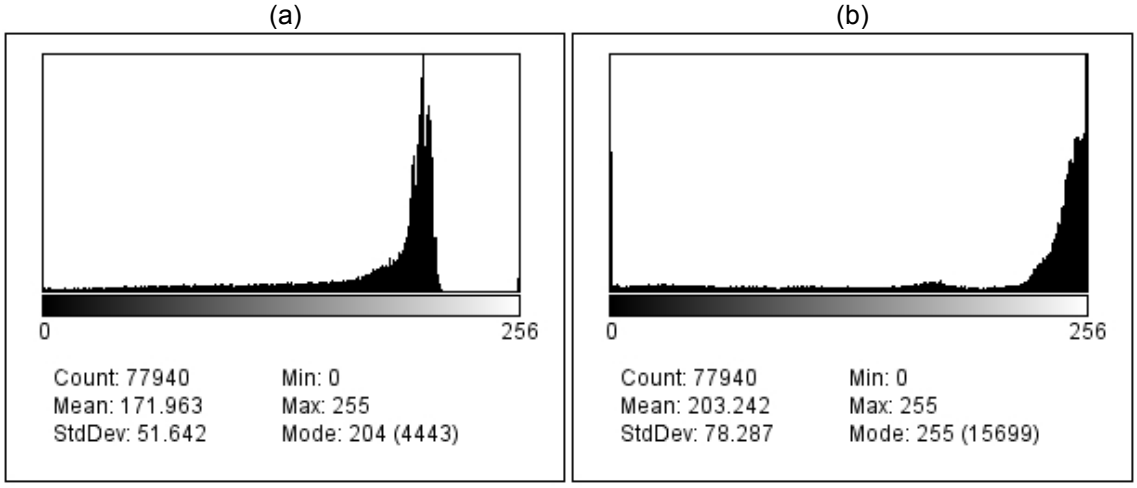



Figura 3.10 – Histogramas dos fragmentos de documentos apresentados na Figura 3.9. a)Histograma da Figura 3.9a (imagem original). b) Histograma da Figura 3.9d (imagem processada).

## 3.4. Etapa 4: Limiarização global da imagem

A quarta e última etapa do método consiste numa adaptação do método proposto por Kavalliertou (2005) e detalhado na Seção 2.4. Utilizou-se o método *Iterative Global Thresholding* (IGT) a fim de se realizar a equalização global do documento. Trata-se de um procedimento iterativo baseado em equalização de histograma. Como entrada, assume-se a imagem em tons de cinza gerada pela etapa anterior. A imagem é descrita

pela equação:

$$I(x,y)=t, t\in[0,1]. \quad \textbf{(3.4)}$$

A cada iteração, são realizados os seguintes passos:

1. Calcular a média dos valores dos pixels (*Ti*) da imagem;
2. Adicionar *1 - Ti* a todos os pixels da imagem;
3. Realizar a equalização do histograma;
4. Realizar os passos de 1 a 3, até a condição de parada definida pela Equação 2.4-2;
5. Opcionalmente, pode-se binarizar a imagem resultante.

O algoritmo IGT foi apresentado em detalhes na Seção 2.4, por este motivo não será repetida a explicação sobre seu funcionamento. Apresenta-se a seguir as adaptações realizadas e as justificativas pela escolha deste método.

No segundo passo, optou-se por somar a diferença entre a tonalidade de valor 1 (cor branca) e a média a todos os *pixels* da imagem. Desta forma, após este passo, toda a imagem estará mais clara. No algoritmo IGT, utiliza-se a subtração de *Ti* dos *pixels* da imagem. Porém, nos experimentos verificou-se que deslocar os *pixels* em direção ao branco propicia a obtenção de melhores resultados. Durante a i-ésima iteração, a imagem do documento $I_i(x, y)$ será determinada pela equação:

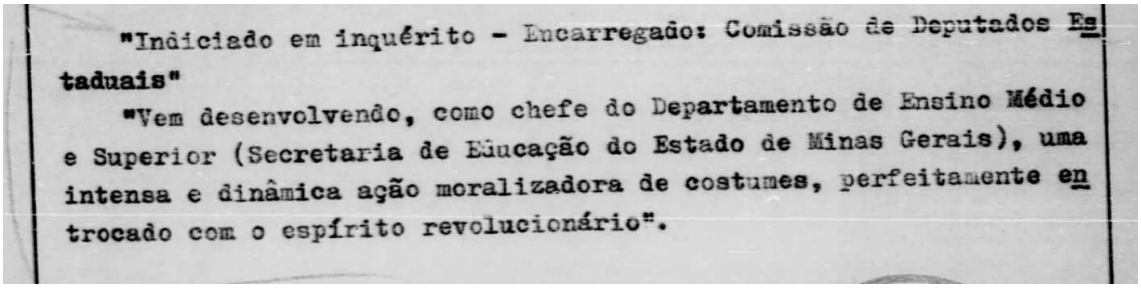
$$I_i(x,y)=I_{i-1}(x,y)+1-T_i \quad \textbf{(3.5)}$$

Para a realização da equalização do histograma (passo 3), foi utilizado um algoritmo modificado conforme apresentado em (KIRK, 2007). O algoritmo utiliza a raiz quadrada dos valores do histograma. Este algoritmo possibilitou melhoria nos resultados se comparado à utilização de equalização de histograma convencional conforme definida em (GONZALEZ, 2000). A Figura 3.11 apresenta exemplos da equalização de histograma no fragmento de documento apresentado na Figura 3.2a.

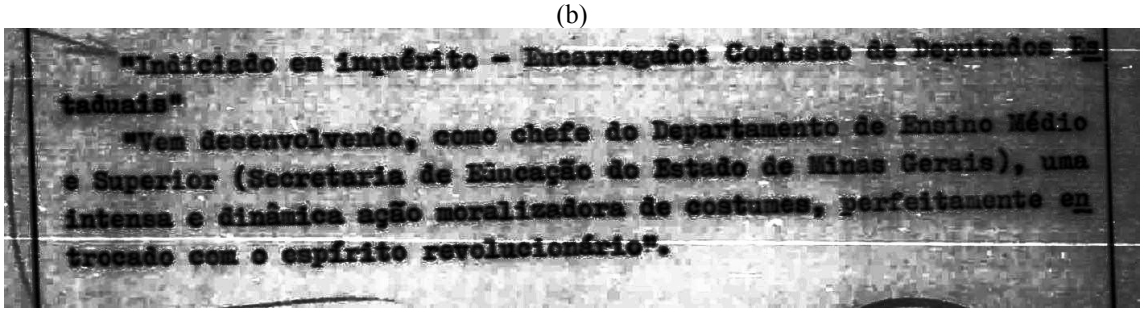
Optou-se pela utilização do algoritmo IGT modificado devido a sua simplicidade e pelos ótimos resultados obtidos nos documentos do acervo. No próximo capítulo, será apresentada em detalhes análise comparativa com outros algoritmos. A Figura 3.12

apresenta exemplo de documento processado pelo método além de imagens de etapas intermediárias.

(a)




(b)




(c)

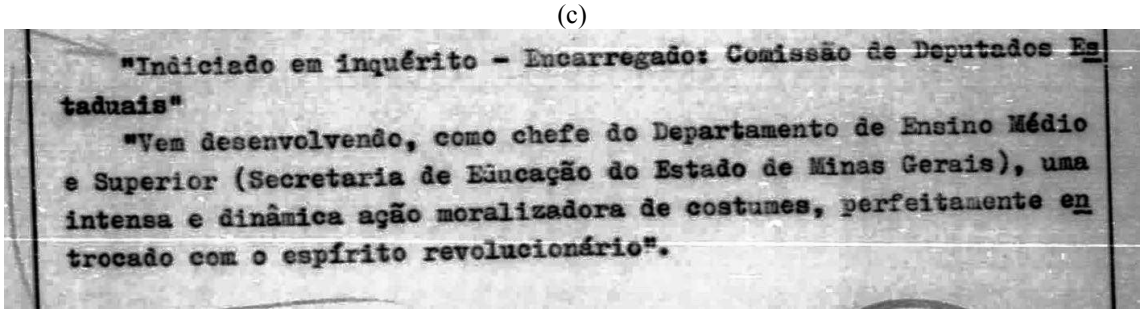



Figura 3.11 – Exemplos da equalização de histograma. a) Imagem original. b) Equalização de histograma convencional. c) Equalização utilizando o método proposto por Kirk.

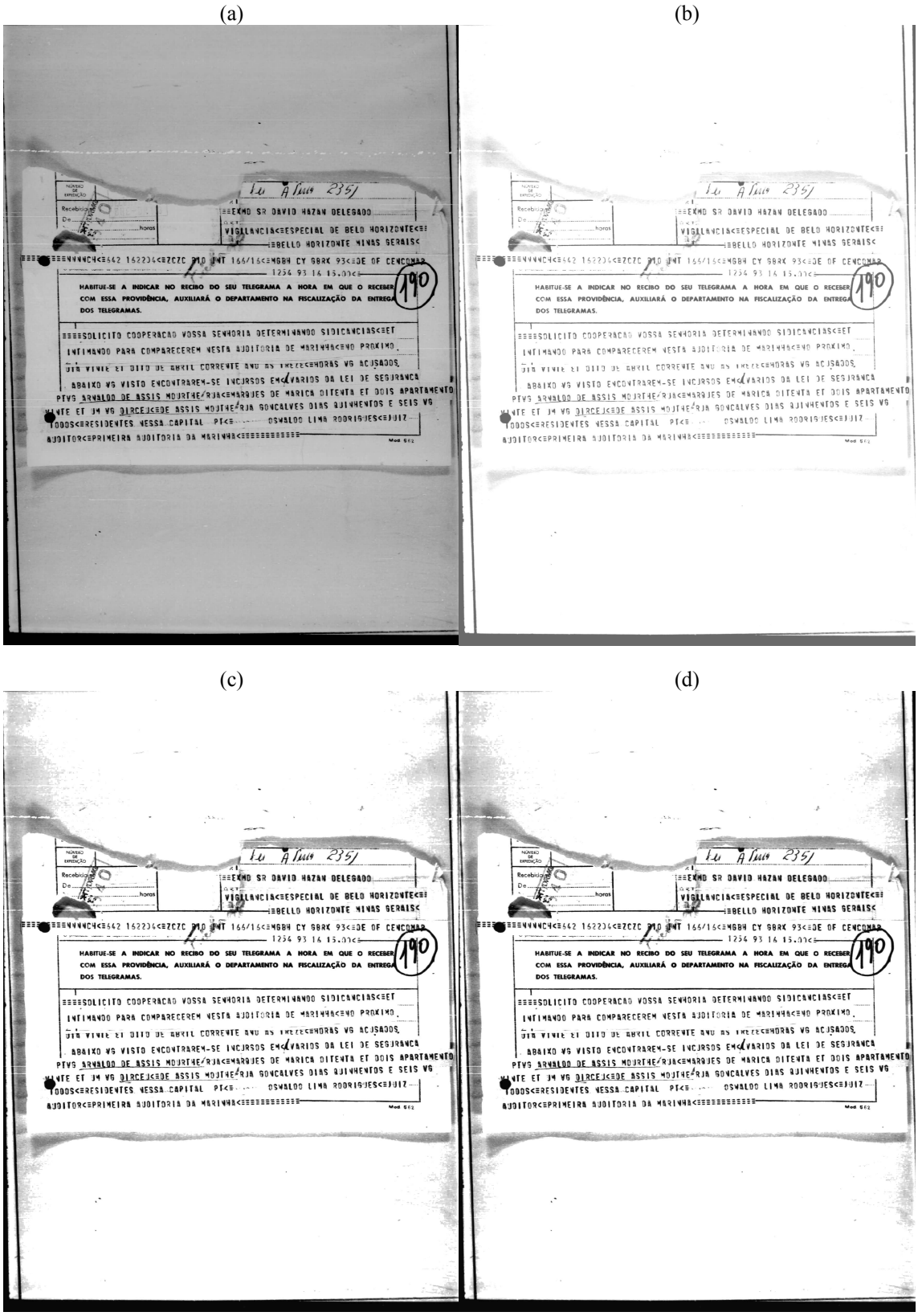



Figura 3.12 – Exemplo da aplicação do método IGT modificado. a) Imagem original. b) Imagem após o passo 2 da primeira iteração. c) Imagem após a equalização de histograma da primeira iteração. d) Imagem processada após três iterações.

## Capítulo 4

# Experimentos e Resultados

O texto seguinte tem por objetivo apresentar os resultados obtidos na aplicação do método proposto nos documentos do acervo do Dops/MG. As etapas dos método foram apresentadas nas Seções 3.1 à 3.4. Em um primeiro momento, são exibidos testes realizados com o objetivo de observar o comportamento do método proposto no tratamento de diversos problemas apresentados pelas imagens dos documentos históricos. O método foi experimentado em uma base de 325 imagens de documentos do acervo Dops/MG. Estes documentos incluem recortes de jornais, manuscritos, documentos datilografados e correspondências.

Em seguida, experimentos foram conduzidos visando apresentar uma análise comparativa com outros trabalhos encontrados na literatura. Os resultados do método proposto são comparados com os resultados dos trabalhos apresentados no Capítulo 2. Utilizam-se técnicas de OCR para determinar métricas de comparação entre os métodos.

Toda a implementação foi realizada através do *ImageJ 1.38x* (IMAGEJ, 2006), utilizando a linguagem de programação *Java 1.6.0*_02 (JAVA, 2007). Os testes utilizaram um *notebook HP Compaq nx6105* equipado com *AMD Turion 64 Mobile* e 1,5 GB de mémoria *RAM*. Os algoritmos propostos foram implementados na forma de *plugins* do *ImageJ*. Em todos os testes de conversão de imagem em texto através de OCR, utilizou-se o *software Tesseract-ocr* (TESSERACT, 2007).

## 4.1. Resultados experimentais no acervo do Dops/MG

Um total de 325 imagens provenientes de cinco bases (Tabela 4.1) foi utilizado para testes do método proposto. Todas as imagens são de documentos do acervo do Departamento de Ordem Política e Social (Dops/MG) acondicionado pelo Arquivo Público Mineiro (APM). As imagens possuem resolução aproximada de 1200x1500 *pixels* e foram armazenadas em formato *JPEG* com compressão. Para a realização deste

trabalho, o APM disponibilizou as imagens digitais referentes ao rolo de micofilme nº 87. Os documentos de cada rolo são organizados em pastas, preservando a organização original dos mesmos. O Rolo nº 87 contem 2432 fotogramas, organizados em 20 pastas. Neste estudo, optou-se por utilizar o conteúdo de apenas três pastas devido às restrições referentes ao uso e difusão de informações impostas aos documentos do acervo do Dops por força de lei. Além destes, foi utilizado um conjunto de 44 imagens de exemplos de documentos do acervo fornecido pelo APM, caracterizando-se como amostras dos itens documentais presentes no acervo. Neste conjunto, encontram-se correspondências, prontuários, listas de pessoas, declarações, anotações manuscritas, recortes de jornais e correspondências. O último conjunto documental utilizado é composto por 14 amostras de documentos datilografados com alto índice de degradação visual pertencentes ao Rolo nº 02. As imagens resultantes da digitalização dos documentos do Rolo nº 02 possuem resolução de 1600x2144 *pixels*.

Tabela 4.1 – Sumário das bases de imagens utilizadas nos testes.

| **Bases** | **Nº imagens** |
|---|---|
| A – Imagens diversas fornecidas pelo APM como amostras dos documentos do acervo Dops/MG. | 44 |
| B – Recortes de Jornais de Junho/Julho 1970. Rolo 087 Pasta 5318. | 123 |
| C – Transcrições datilografadas de entrevistas . Rolo 087 Pasta 5333. | 79 |
| D – Revista Policial Mineira. Rolo 087 Pasta 5334. | 65 |
| E – Amostras de documentos datilografados com alto índice de degradação visual. Rolo 002. | 14 |

A Figura 4.1a mostra um documento da base de imagens D e a Figura 4.1b apresenta o mesmo documento após o processamento pelo método proposto. A Figura 4.2 apresenta os histogramas das Figuras 4.1a e 4.1b. No histograma da Figura 4.2b pode-se observar uma grande alteração da distribuição dos *pixels*. Esta mudança é provocada principalmente pela etapa 4 do método, necessitando de apenas duas iterações para obter o resultado final. A imagem resultante continua em tons de cinza, porém a maior parte dos *pixels* da imagem foi deslocada para as extremidades do histograma.

(a) (b)

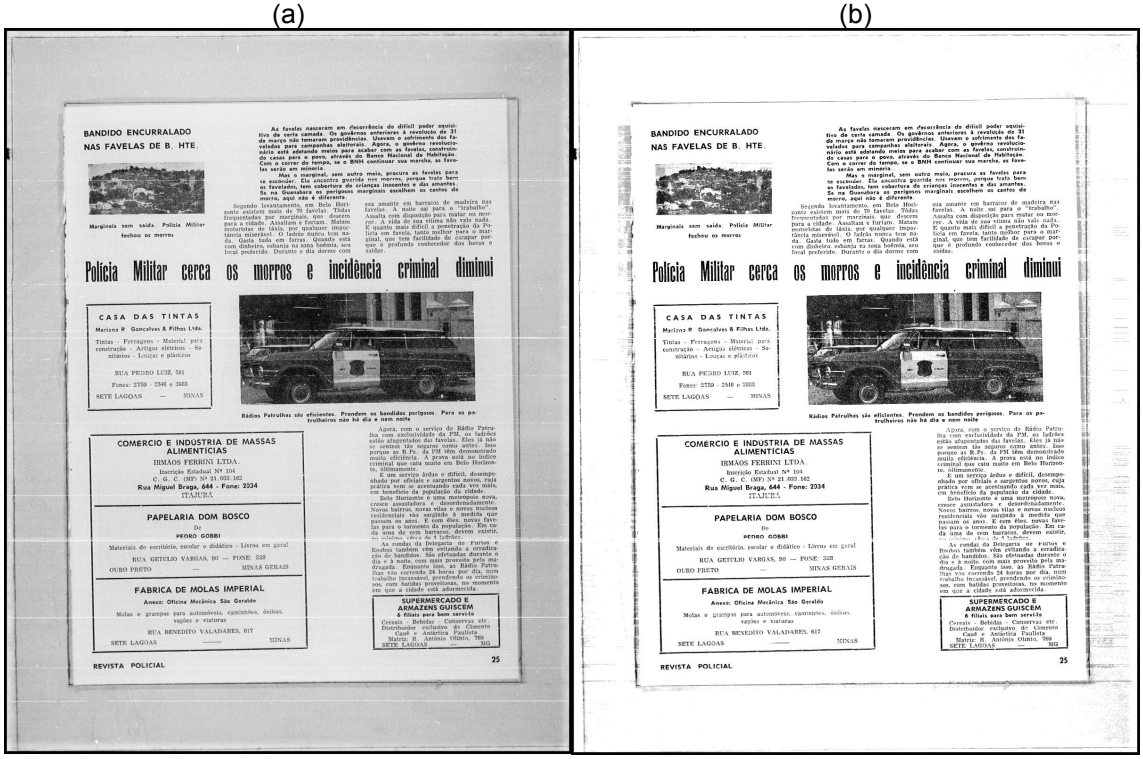



Figura 4.1 – Exemplo do resultado do processamento de documento da base D. a) Imagem original. b) Imagem processada.

(a) (b)

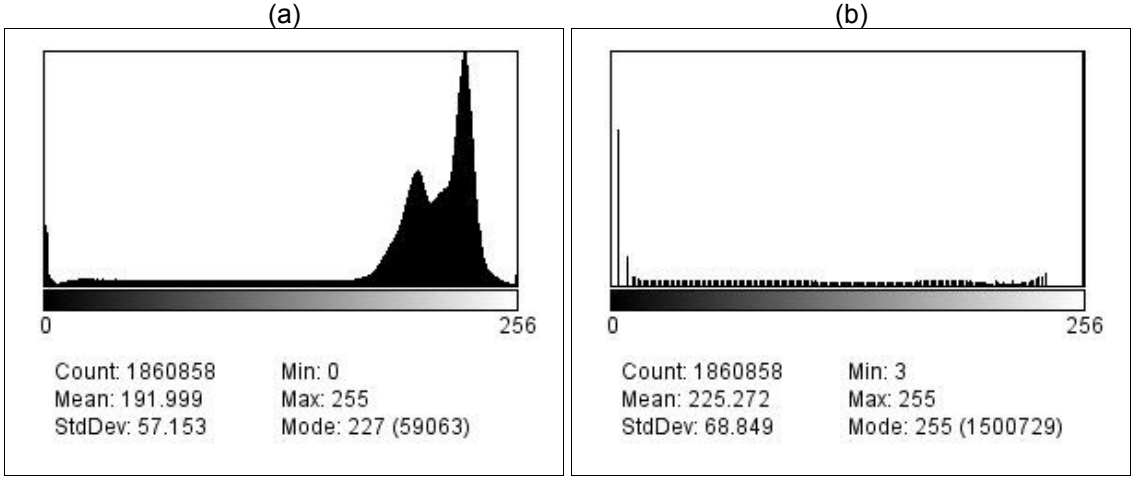



Figura 4.2 – Histogramas dos documentos apresentados na Figura 4.1. a) Histograma da Figura 4.1a (imagem original). b) Histograma da Figura 4.2b (imagem processada).

A Figura 4.3 mostra detalhes do documento apresentado na Figura 4.1. Através da comparação entre as Figuras 4.3a e 4.3b, verifica-se que o ruído de fundo foi totalmente eliminado, além de, considerável realce do contraste. Observa-se, também, que a figura

localizada abaixo do título foi preservada.

(a) (b)

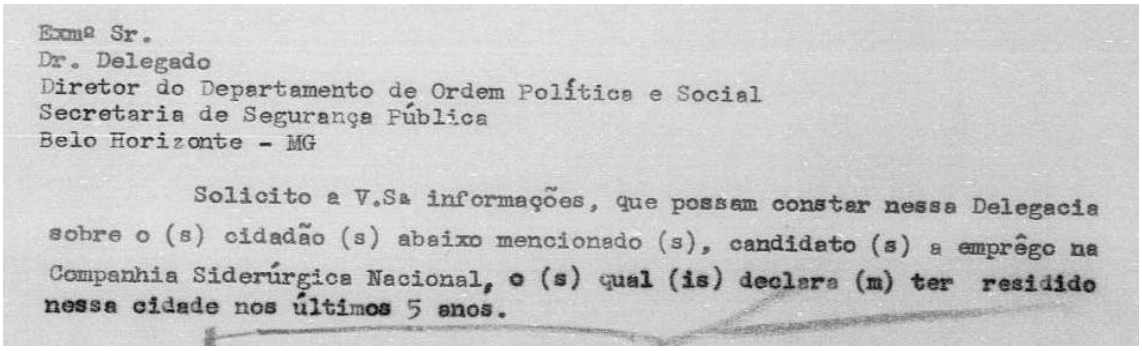

Figura 4.3 – Detalhe do documento apresentado na Figura 4.1. a) Imagem original. b) Imagem processada.

Documentos datilografados são freqüentemente encontrados no acervo do Dops/MG. Neste tipo documental, observa-se uma anomalia bastante típica, grande variação de tonalidade entre caracteres de uma mesma linha ou parágrafo. Esta anomalia é um empecilho para técnicas de limiarização global. A etapa 3 do método proposto busca minimizar este problema. A Figura 4.4 apresenta um fragmento de documento, onde verifica-se a anomalia nas tonalidades dos caracteres datilografados. Observa-se, na Figura 4.4a, que os caracteres das linhas inferiores apresentam tonalidade mais escura que os caracteres das linhas superiores. Na Figura 4.4b, pode-se observar o realce do contraste dos caracteres com o fundo após a execução da etapa 3 do método proposto. A imagem processada é apresentada na Figura 4.4c.

A Figura 4.5 apresenta um experimento com um fragmento da Revista Policial Mineira (base D). Na Figura 4.5b, observa-se que a fotografia presente não sofreu alterações significativas. Este resultado é bastante apropriado para algoritmos de limiarização de documentos. No acervo do Dops/MG, encontra-se um grande número de

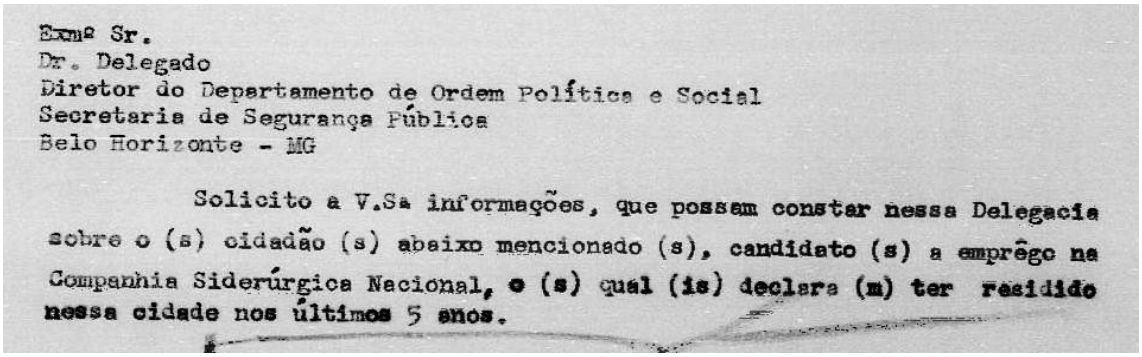
documentos com fotografias, sejam eles: recortes de jornais ou revistas, prontuários ou mesmo fotografias com anotações manuscritas. Este comportamento do método proposto mostrou-se adequado ao permitir a limiarização do conteúdo textual sem deteriorar as fotografias contidas no documento.

(a)


Exmº Sr.
Dr. Delegado
Diretor do Departamento de Ordem Política e Social
Secretaria de Segurança Pública
Belo Horizonte - MG

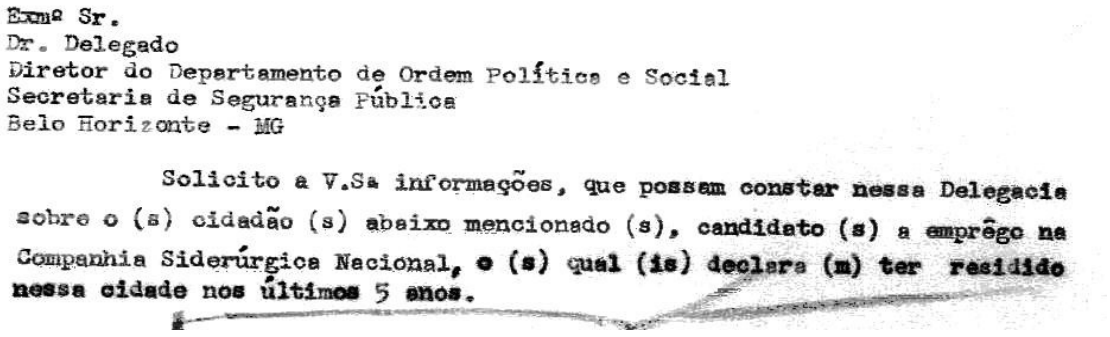
Solicito a V.Sa informações, que possam constar nessa Delegacia sobre o (s) cidadão (s) abaixo mencionado (s), candidato (s) a emprêgo na Companhia Siderúrgica Nacional, o (s) qual (is) declara (m) ter residido nessa cidade nos últimos 5 anos.

(b)


Exmº Sr.
Dr. Delegado
Diretor do Departamento de Ordem Política e Social
Secretaria de Segurança Pública
Belo Horizonte - MG

Solicito a V.Sa informações, que possam constar nessa Delegacia sobre o (s) cidadão (s) abaixo mencionado (s), candidato (s) a emprêgo na Companhia Siderúrgica Nacional, o (s) qual (is) declara (m) ter residido nessa cidade nos últimos 5 anos.

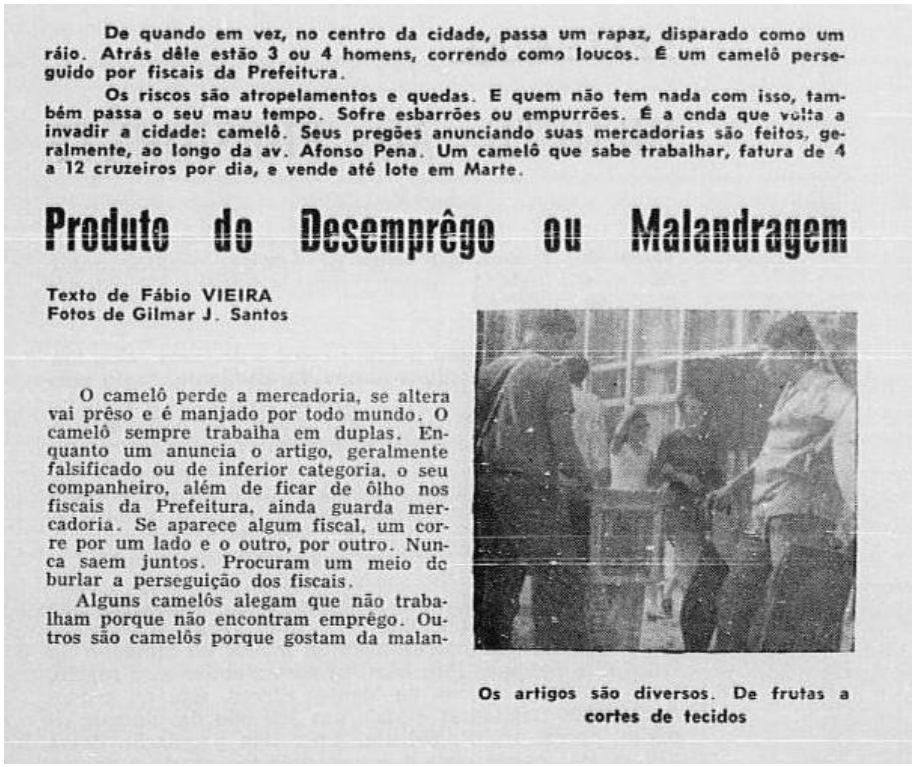
(c)


Exmº Sr.
Dr. Delegado
Diretor do Departamento de Ordem Política e Social
Secretaria de Segurança Pública
Belo Horizonte - MG

Solicito a V.Sa informações, que possam constar nessa Delegacia sobre o (s) cidadão (s) abaixo mencionado (s), candidato (s) a emprêgo na Companhia Siderúrgica Nacional, o (s) qual (is) declara (m) ter residido nessa cidade nos últimos 5 anos.

Figura 4.4 – Exemplo de anomalias nos caracteres datilografados. a) Imagem original. b)Imagem após a etapa 3. c) Imagem processada.

(a)



**De quando em vez, no centro da cidade, passa um rapaz, disparado como um ráio. Atrás dêle estão 3 ou 4 homens, correndo como loucos. É um camelô perseguido por fiscais da Prefeitura.**

**Os riscos são atropelamentos e quedas. E quem não tem nada com isso, também passa o seu mau tempo. Sofre esbarrões ou empurrões. É a onda que volta a invadir a cidade: camelô. Seus pregões anunciando suas mercadorias são feitos, geralmente, ao longo da av. Afonso Pena. Um camelô que sabe trabalhar, fatura de 4 a 12 cruzeiros por dia, e vende até lote em Marte.**

# Produto do Desemprêgo ou Malandragem

**Texto de Fábio VIEIRA**
**Fotos de Gilmar J. Santos**

O camelô perde a mercadoria, se altera vai prêso e é manjado por todo mundo. O camelô sempre trabalha em duplas. Enquanto um anuncia o artigo, geralmente falsificado ou de inferior categoria, o seu companheiro, além de ficar de ôlho nos fiscais da Prefeitura, ainda guarda mercadoria. Se aparece algum fiscal, um corre por um lado e o outro, por outro. Nunca saem juntos. Procuram um meio de burlar a perseguição dos fiscais.

Alguns camelôs alegam que não trabalham porque não encontram emprêgo. Outros são camelôs porque gostam da malan-

**Os artigos são diversos. De frutas a cortes de tecidos**

(b)

**De quando em vez, no centro da cidade, passa um rapaz, disparado como um ráio. Atrás dêle estão 3 ou 4 homens, correndo como loucos. É um camelô perseguido por fiscais da Prefeitura.**

**Os riscos são atropelamentos e quedas. E quem não tem nada com isso, também passa o seu mau tempo. Sofre esbarrões ou empurrões. É a onda que volta a invadir a cidade: camelô. Seus pregões anunciando suas mercadorias são feitos, geralmente, ao longo da av. Afonso Pena. Um camelô que sabe trabalhar, fatura de 4 a 12 cruzeiros por dia, e vende até lote em Marte.**

# Produto do Desemprêgo ou Malandragem

**Texto de Fábio VIEIRA**
**Fotos de Gilmar J. Santos**

O camelô perde a mercadoria, se altera vai prêso e é manjado por todo mundo. O camelô sempre trabalha em duplas. Enquanto um anuncia o artigo, geralmente falsificado ou de inferior categoria, o seu companheiro, além de ficar de ôlho nos fiscais da Prefeitura, ainda guarda mercadoria. Se aparece algum fiscal, um corre por um lado e o outro, por outro. Nunca saem juntos. Procuram um meio de burlar a perseguição dos fiscais.

Alguns camelôs alegam que não trabalham porque não encontram emprêgo. Outros são camelôs porque gostam da malan-

**Os artigos são diversos. De frutas a cortes de tecidos**

Figura 4.5 – Exemplo de documento contendo fotografia. a) Imagem original. b) Imagem processada.

Através de testes utilizando o critério de viabilidade proposto neste trabalho, verificou-se que aproximadamente 20% das imagens apresentam condições que impedem o tratamento através do algoritmo apresentado no Capítulo 3. Estas imagens não satisfazem o critério definido pela Inequação 3.1, conforme definido na Seção 2.1. Porém, destaca-se o fato de que em alguns rolos quase a totalidade das imagens não satisfaz o critério para processamento, como por exemplo na base de imagens E. A Tabela 4.2 apresenta os resultados dos testes de viabilidade para cada uma das bases de imagens estudadas. Observou-se que entre os documentos datilografados o percentual de imagens rejeitadas é maior. Os menores índices de rejeição são encontrados nas imagens de recortes de jornais e revistas.

Tabela 4.2 – Teste de viabilidade para aplicação do método utilizando a Equação 3.1.

| **Bases** | **Nº imagens** | **Imagens rejeitadas** | **% rejeitadas** |
|---|---|---|---|
| A – Imagens diversas | 44 | 13 | 29,6 |
| B – Recortes de Jornais | 123 | 15 | 12,2 |
| C – Transcrições datilografadas de entrevistas. | 79 | 23 | 29,1 |
| D – Revista Policial Mineira. | 65 | 6 | 9,2 |
| E – Documentos com alto índice de degradação visual. | 14 | 14 | 100,0 |

Nesta seção, foram apresentados resultados de três dos experimentos realizados. Todos os 254 documentos classificados de forma positiva na análise de viabilidade foram processados utilizando o método proposto. Os resultados são similares aos dos experimentos apresentados. Destaca-se os resultados obtidos na eliminação do ruído de fundo, no realce do contrate e na suavização de pequenas anomalias nos caracteres. O método proposto mostrou-se capaz de manter imagens presentes no corpo do documento, sem realizar alterações que comprometam seu aspecto visual.

A utilização do método proposto mostrou-se adequada em relação aos documentos do acervo do Dops/MG. Um conjunto significativo de documentos, mais de 30% do acervo, encontra-se em alto grau de degradação e microfilmados com qualidade muita baixa. As imagens resultantes da digitalização desse conjunto são praticamente ilegíveis. A aplicação do método proposto nessas imagens resultou em significativa melhoria da

qualidade visual. Destaca-se a melhoria obtida na impressão dos documentos, em diversos casos, a melhoria da impressão foi mais significativa do que a obtida na visualização do documento na tela do computador.

## 4.2. Comparação com outras abordagens

Esta seção apresenta um estudo comparativo do método proposto com outras abordagens tipicamente utilizadas na limiarização de documentos. Para comparar os resultados, utilizou-se os métodos descritos em (OTSU, 1979; KAPUR, 1985; SAUVOLA, 2000 e KAVALLIERATOU, 2005). Kavallieratou (2005b, p.686) afirma que devido ao fato da área de pesquisa de processamento de documentos históricos ser relativamente nova, ainda não estão disponíveis bases para validação dos resultados. Na literatura, adota-se como métrica de validação a comparação com outros métodos já consagrados.

As Figuras 4.6.1 e 4.6.2 apresentam os resultados do processamento da capa da Revista Policial Mineira. Observa-se que os algoritmos de Otsu, Kapur e Sauvola deterioraram significativamente a fotografia central, este resultado é provocado pelo fato dos algoritmos em questão gerarem imagens binárias (apenas dois tons de cinza). O algoritmo de Kavallieratou conseguiu manter a fotografia central sem deterioração. A palavra “Policial” no título da página foi totalmente suprimida da imagem processada pelos algoritmos de Kapur, Kavallieratou e Sauvola, enquanto o algoritmo de Otsu tornou-a praticamente ilegível. O método proposto foi capaz de manter a fotografia central com qualidade adequada e limiarizar corretamente a palavra “Policial” no título da página. Este experimento demonstra o equilíbrio do método ao processar documentos que contêm texto e imagens.

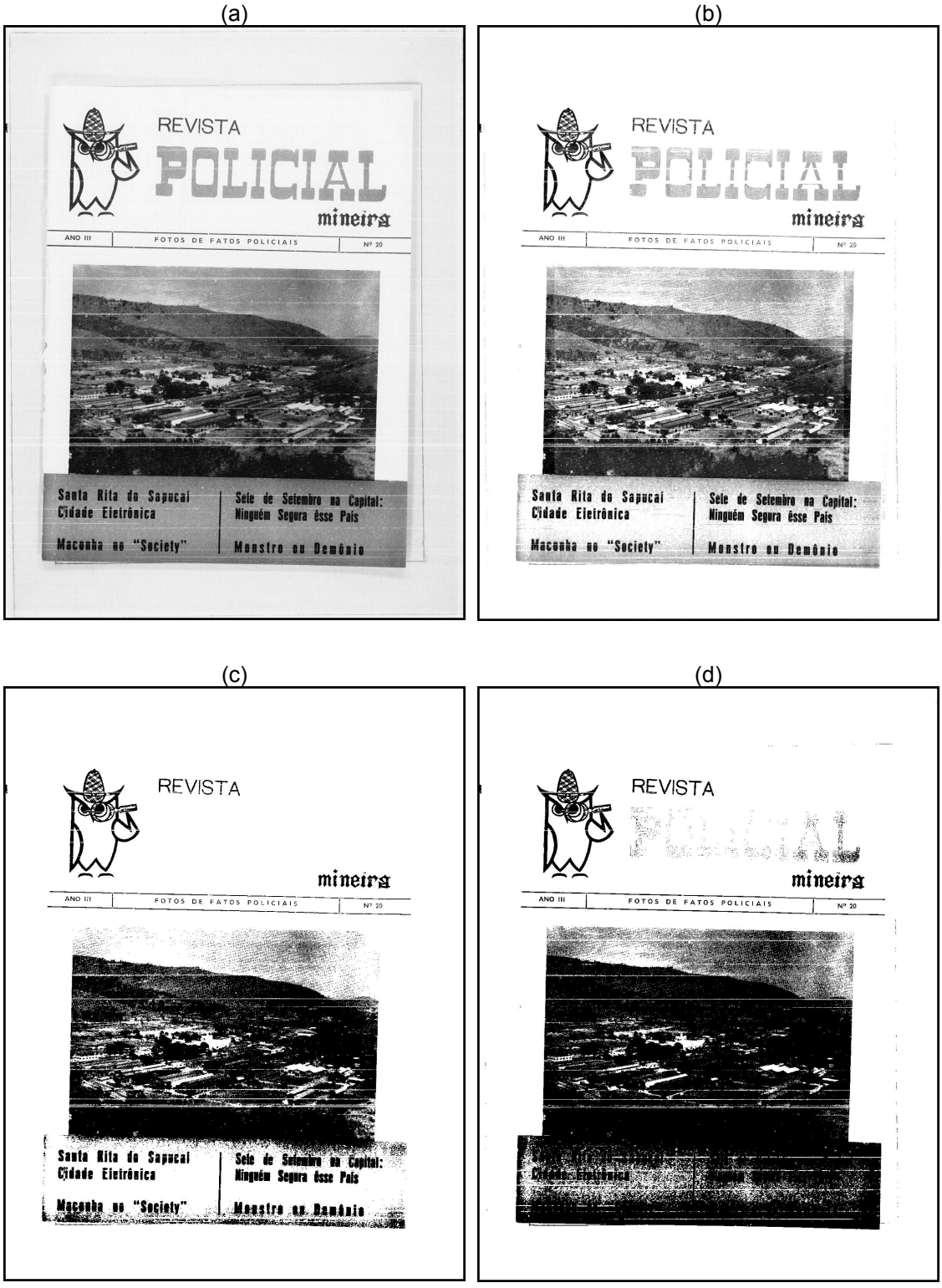



Figura 4.6.1 – Exemplo de documento contendo fotografia. a) Imagem original.
b) Método proposto. c) Kapur. d) Otsu.

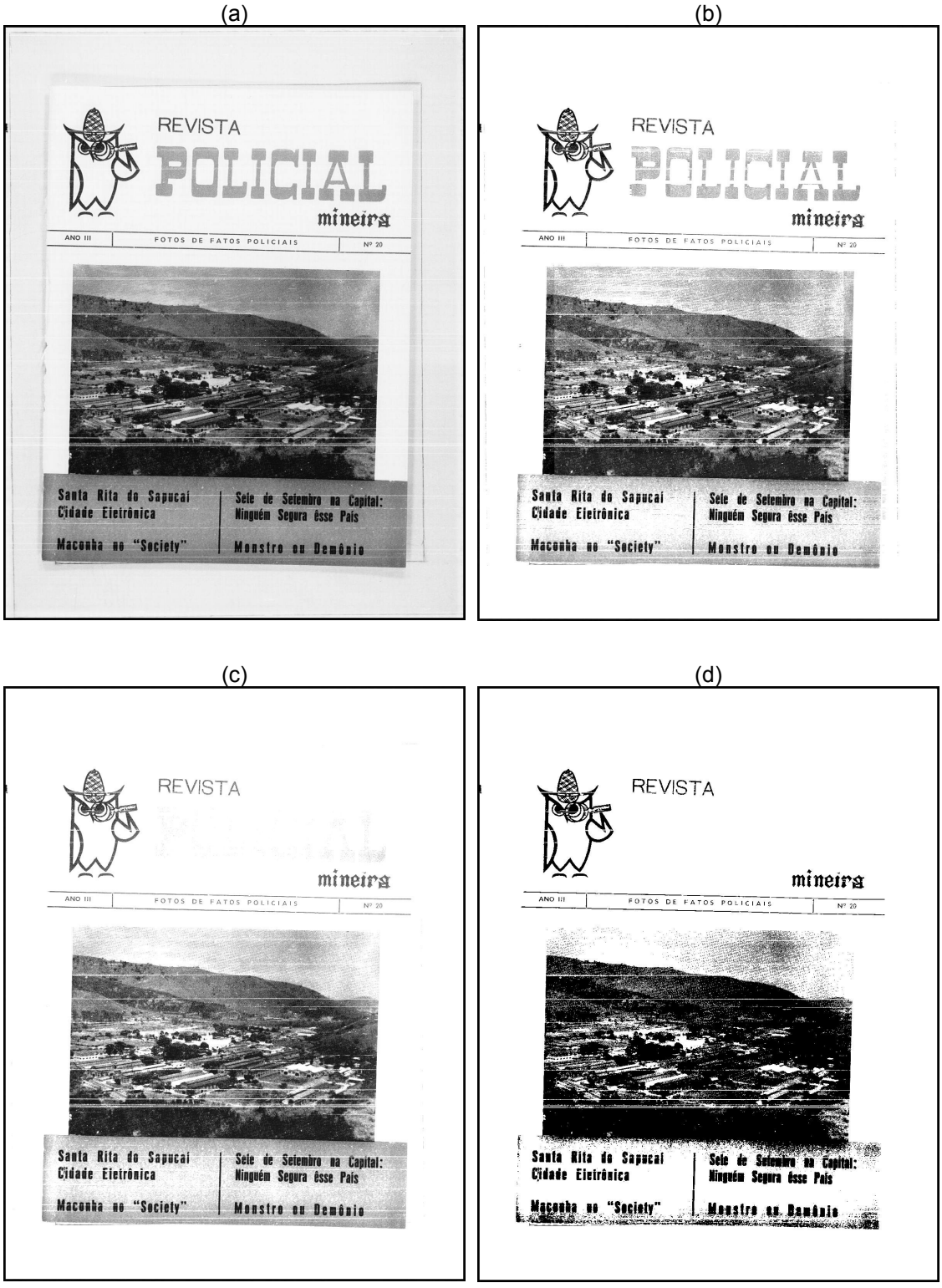



Figura 4.6.2 – Exemplo de documento contendo fotografia. a) Imagem original.
b) Método proposto. c) Kavallieratou. d) Sauvola.

Na Figura 4.7, observam-se os resultados da limiarização de um documento contendo apenas texto, no caso, um recorte de jornal. O resultado do método proposto é similar aos dos métodos testados. Destaca-se que o método proposto gerou alguns artefatos indesejáveis (lado esquerdo do recorte). Porém, este comportamento não prejudicou significativamente a limiarização do conteúdo textual.

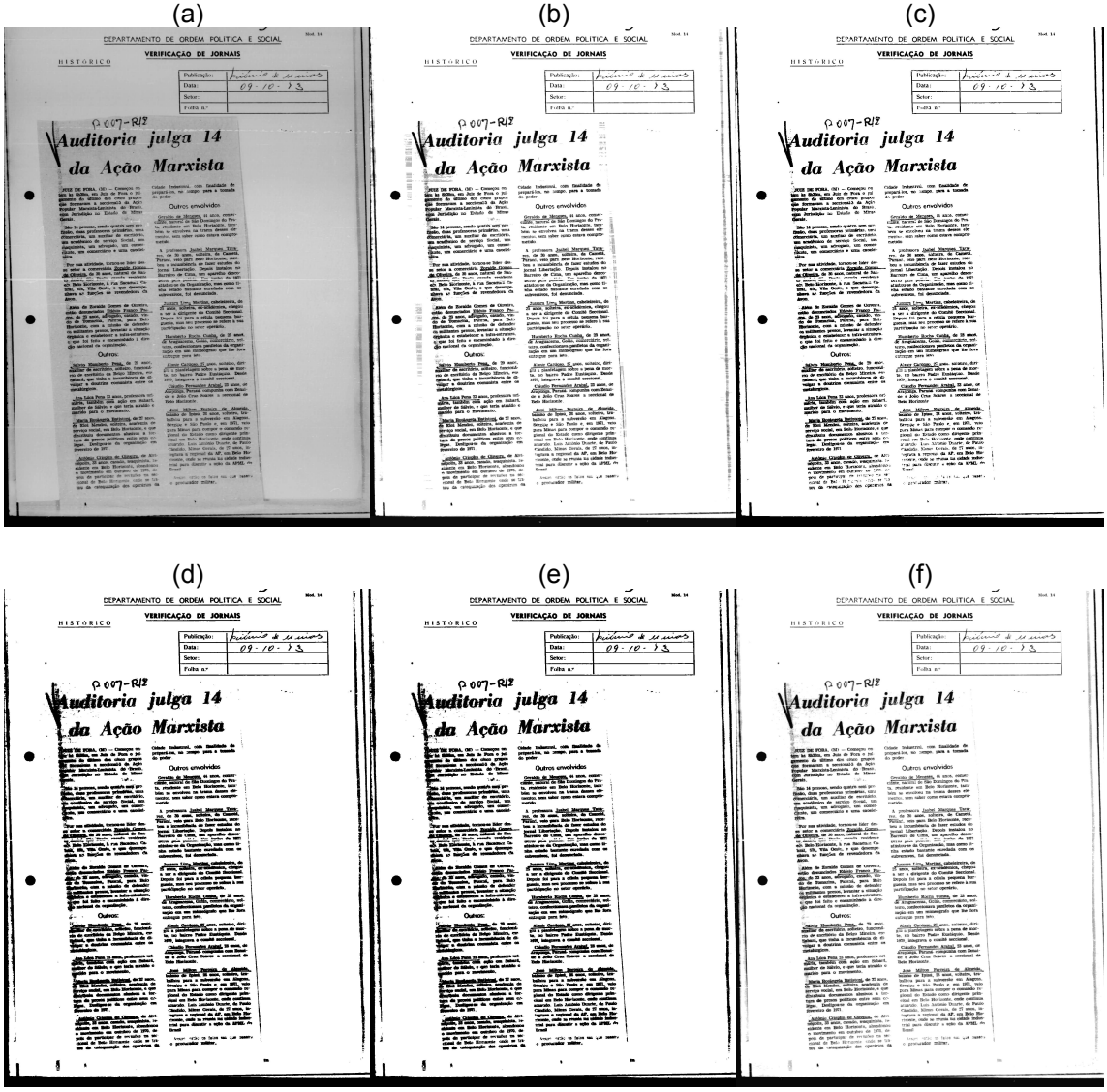
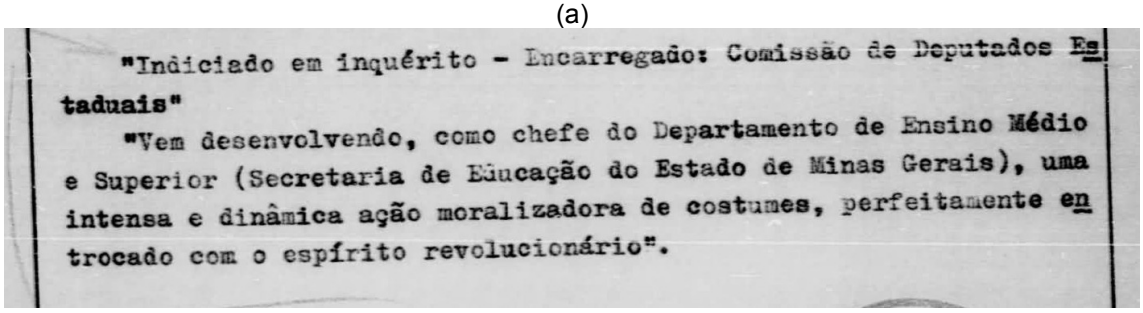


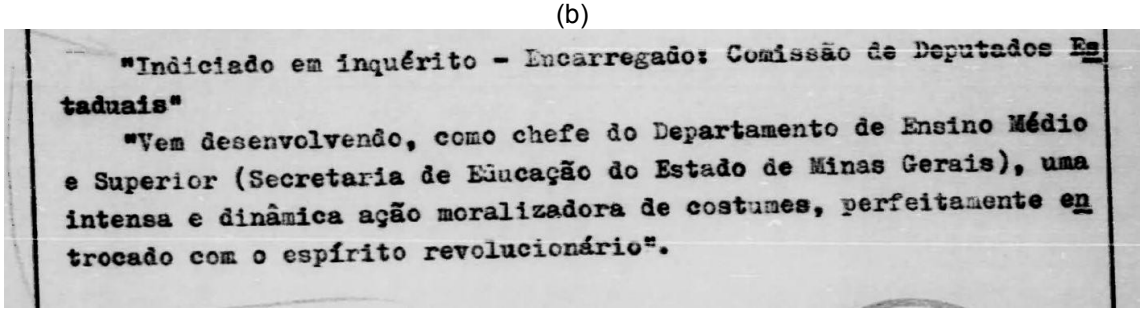

Figura 4.7 – Exemplo de documento contendo recorte de jornal. a) Imagem original. b)Método proposto. c) Otus. d) Tapuru. e) Sauvola. f) Kavallieratou.

Nas Figura 4.8.1 e 4.8,2, observam-se os resultados da limiarização do fragmento de texto apresentado na Figura 1.2. O fragmento foi limiarizado utilizando o método proposto, e os métodos de Kapur, Otsu, Kavallieratou e Sauvola.

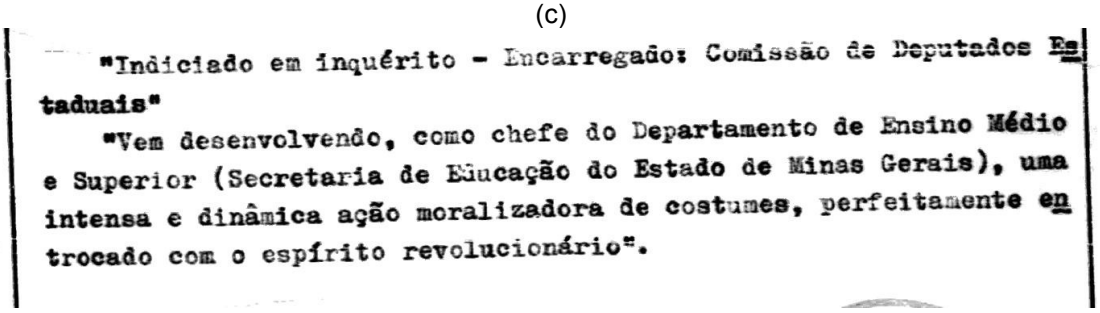
(a)



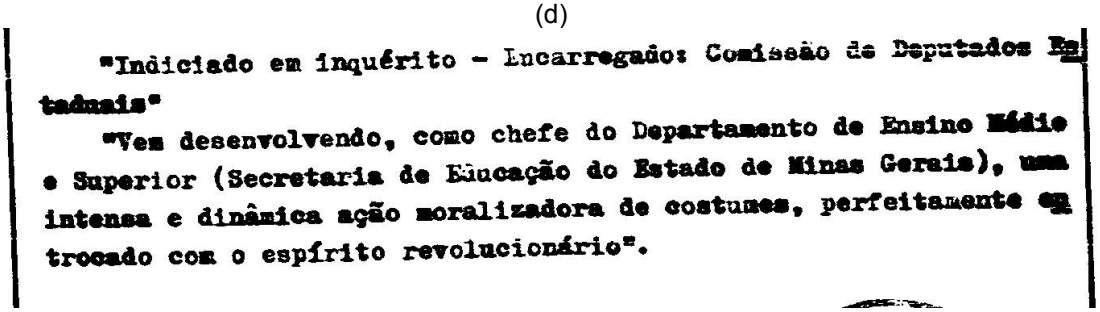

(b)




(c)




(d)




(e)

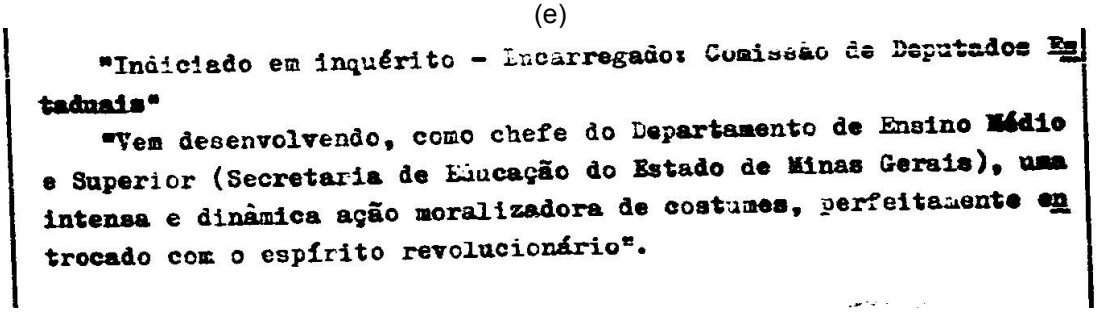
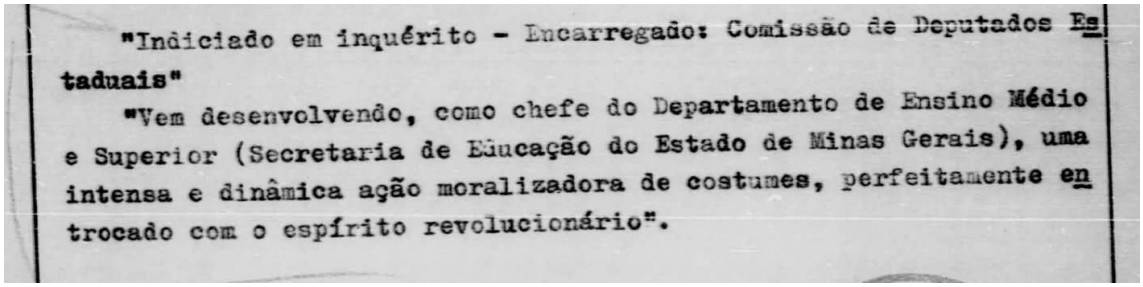
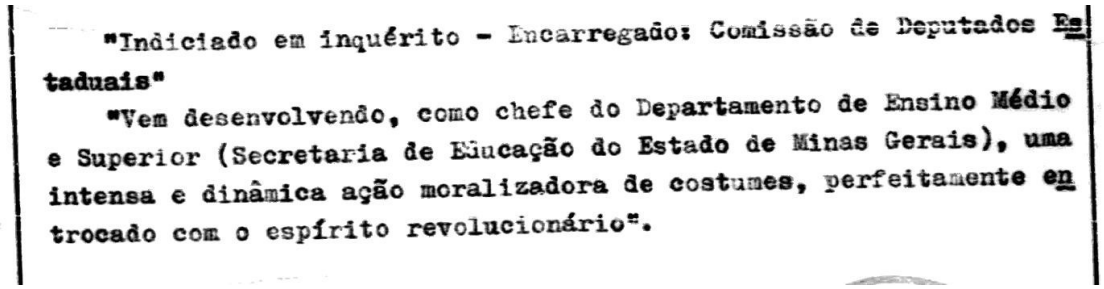



Figura 4.8.1 – Limiarização de fragmento de documento. a) Imagem original. b) Etapa 3. c)Método proposto. d) Kapur. e) Otsu.

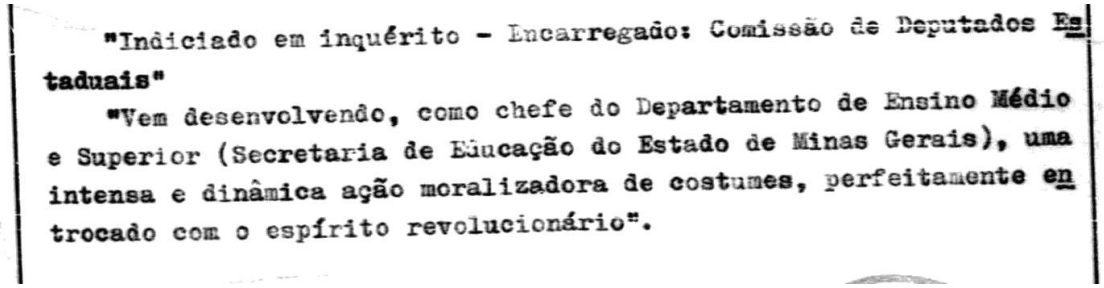
(a)




(b)




(c)




(d)

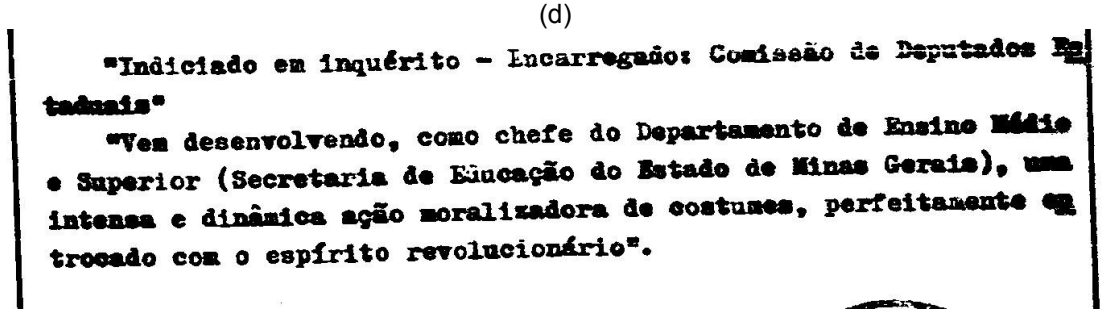



Figura 4.8.2 – Limiarização de fragmento de documento. a) Imagem original.
b) Método proposto. c) Kavallieratou. d) Sauvola.

As imagens limiarizadas das Figuras 4.8.1 e 4.8.2 foram submetidas ao reconhecimento óptico de caracteres (OCR) e os resultados são apresentados a seguir.

Resultado da aplicação de OCR na imagem da Figura 4.8.1a (Imagem original):

```
\ "lnóicißdo em inquérito - lncărrogædoz Comíxèàù às Deputado: Eá
talas!]"
'Vem desenvolvendo, como chefe do Dcpsrtanonto de Bnsiuo lldío
o Suporíor (Secroturin do Blmcação do Bstado do limas Gerais), II
intensa c dinâmica ação moralizadora. do costumou, pcrfeitamento •g_
trocado com o espírito revO1ucíOmh'ío‘‘.
```

Resultado da aplicação de OCR na imagem da Figura 4.8.1b (Etapa 3):

```
••‘Iuã1c1Ado em inquérito - Lucărregaãox Comiă-são às æputădos
t•dnd.••
Wen desenvolvendo, como chefe do Dcpsrtamento de Eusino lldiø
— • Supox-1Or (Secretsria de Eãucação do Bstudo do unas Gerais), na
— intensa e dinâmica ação moralizadora de costumes, pcrfe1tu:xent•
•g
d troudo com o espírito revo1uciOnš.:~ío=.
```

Resultado da aplicação de OCR na imagem da Figura 4.8.1c (Método proposto):

```
"ïučíciaßo em inquérito - Imcărrcgaão: Cumíùsão às æpctadce Be
tenhais"
Wen desenvolvendo, como chefa do Dcpartanento de Eusino lčdio
— • Suporior (Secrotańa de Eûucação do Bstado do unas Gerais), un
— intensa e dinâmica ação moralizadora do costumes, pcrfeitasxento
og
  trocado com o espírito rcvO1ucionário=.
```

Resultado da aplicação de OCR na imagem da Figura 4.8.1d (Kapur):

```
"ïuãiciado em inquérito - Lucarngaûos Conícssäo às D"•p:'.‘::Åos
tdnt.iI'
'Y•• desenvolvendo, como chefe do Dopgrtamntø do Enainø Hl1•
• Suyorior (Secr•t•r1• ă• Eàucaçíø do lotado d• lusa G•1··1•), 1
— :I.nt•u•• e dixńmíoa gção noralizuăorn do costuma, p•r!eitn.m•ut•
Q
  trocado com o espírito ravoluciøuárioä
```

Resultado da aplicação de OCR na imagem da Figura 4.8.1e (Otsu):

```
\ 'ïndiciedo em inquérito - lncărregsdoz Cumíxèào ćõ Deputado: ğl
talas!)"
'Vem desenvolvendo, como chefe do Dopsrtsnonto do Bnsino lldío
• Superíor (Secreturís do Bãuœção do Bstado do limas Gerais), uß
intensa e dinâmica ação moralizadora do costuma, pcrfeítsuento •g
trocado com o espírito revo1ucíomh'ío‘‘.
```

Resultado da aplicação de OCR na imagem da Figura 4.8.2c (Kavallieratou):

```
\ 'ïnúiciedo em inquérito  încărregaãøs Cumíùèào às Deputado: ğj
tnduníi"
'Vem ãesenvclvenõo, como chefe do Departsnento de Bxninø ßdíø
e Superíør (Secretarís de Bãuœção do Bstado do uma Gerais), nu
intensa e dinâmica ação moralizadora de coetumoa, perfe1taăxcut• øg
trocado com O espírito revO1uciOu.ă.río‘.
```

Resultado da aplicação de OCR na imagem da Figura 4.8.2d (Sauvola):

```
'ïndíciado em inquérito - 1ncarr•gn.ã0: Gonissão às Ikpztaáot
tolhi.!"
'Vu desenvolvendo, como chato do Doportsnønto •1• Ensim ßlíø
Z • Snporior (Socr•t•r1• d• Eãuooçíø do lotado d• num Gonul, 1
— i.nt•n•• e dinâmica oçío noralizoãcro do costuma, p•1·feita.m•ut•
Q
  trocado coa o espírito revolucionário".
```

Observa-se que nenhum dos métodos foi capaz de produzir resultados significativos no reconhecimento óptico de caracteres. Os resultados são similares, apenas apresentando pequenas variações em certas regiões da imagem. Os testes realizados não obtiveram sucesso na conversão da imagem em texto. Acredita-se que seja necessário criar métodos de OCR específicos para o acervo em questão, a fim de se aprimorar os resultados do reconhecimento óptico de caracteres. Apesar do método proposto apresentar bons resultados no aprimoramento da qualidade visual dos documentos, o mesmo não se mostrou viável como uma etapa de pré-processamento para o reconhecimento óptico de caracteres.

## 4.3. Considerações finais

Devido à simplicidade de todas as etapas do método, o custo computacional é baixo em comparação com outros algoritmos para processamento de documentos textuais. Experimentos práticos demonstraram que o método proposto apresentou performance equivalente aos outros métodos testados. O estudo do comportamento assintótico do método proposto demonstra que o mesmo apresenta complexidade linear *O(n)*, onde *n* é número de *pixels* da imagem.

O método proposto será integrado ao sistema de acesso ao acervo do Dops disponível no APM. Este sistema permite ao público realizar consultas aos documentos do Dops/MG. Após a integração do método proposto ao sistema, os usuários poderão

realizar a limiarização das imagens dos documentos sempre que julgarem necessário. O método proposto permite o processamento automático das imagens do acervo, podendo processá-las em lote. Porém, acredita-se que, em um primeiro momento, seja mais adequado deixar a decisão da aplicação do método sob a responsabilidade dos usuários. Quando o usuário solicitar, a imagem do documento será processada. Uma nova imagem será gerada, apenas para exibição, enquanto a imagem original será preservada integralmente. Atualmente, o sistema de acesso ao acervo Dops/MG utiliza esta estratégia nos ajustes de brilho e contraste das imagens do acervo, deixando que o usuário decida pela aplicação.

# Capítulo 5
# Conclusão

Esta dissertação apresentou as etapas necessárias para realizar a limiarização de documentos históricos. O método apresentado mostrou-se eficaz na tarefa de aprimorar a qualidade visual de documentos históricos. Os resultados obtidos são similares ou superiores aos encontrados na literatura. A definição de um critério de viabilidade de aplicação do método mostrou-se relevante e funcional. Principalmente, ao viabilizar o tratamento de grandes acervos documentais de forma automática. A utilização de uma abordagem híbrida, combinando limiarização global e local (adaptativa), apresentou resultados significativos na comparação com outras abordagens.

Como pode ser observado, o método é capaz de eliminar problemas como ruído de fundo, baixo nível de contraste e amenizar anomalias nos caracteres. A aplicação do método possibilitou melhoria na legibilidade e na qualidade visual dos documentos do acervo do Dops/MG. Os testes demonstraram a viabilidade da aplicação do método nas imagens dos documentos do acervo.

Este trabalho atesta, como vários outros, a viabilidade do desenvolvimento de algoritmos para processamento de documentos históricos. Em particular, a melhoria da qualidade visual do acervo permite a obtenção de melhores resultados na indexação manual, além de, facilitar o acesso aos documentos, devido à melhoria da legibilidade e melhor qualidade de impressão.

## 5.1. Contribuições

Com a expansão dos projetos de digitalização de acervos de documentos históricos, as coleções de imagens digitais crescem rapidamente. Porém, muitas vezes a qualidade das imagens geradas não é satisfatória. Aprimorar, de forma automática, a qualidade visual destas imagens é uma tarefa que ainda apresenta grandes desafios.

Este trabalho contribui da seguinte maneira:

1. discutindo questões relacionadas ao acesso a grandes coleções de imagens digitais de acervos históricos;
2. propondo uma solução para o problema de aprimorar a qualidade visual de imagens de documentos históricos. A solução utiliza uma abordagem híbrida e sem intervenção humana. Isto a diferencia da maioria das estratégias apresentadas na literatura;
3. propondo um método simples, baseado em medidas estatísticas, para avaliar a viabilidade da aplicação da limiarização. Esta análise preliminar da imagem evita que a aplicação indevida do método prejudique ainda mais a qualidade visual de determinados documentos;
4. propondo um novo método de detecção de linhas horizontais que apresentam conteúdo textual, permitindo aprimorar e otimizar a etapa de limiarização adaptativa do método;
5. realizando experimentos do método em bases de imagens do Dops/MG e demonstrando sua viabilidade.

## 5.2. Trabalhos futuros

Os resultados já alcançados são satisfatórios, motivando novas pesquisas. Em especial a comparação da resposta de ferramentas de OCR em documentos originais e os obtidos após a aplicação do método proposto. A implementação de uma ferramenta de OCR especializada no tratamento de documentos históricos seria necessária.

O estudo e definição de métricas para avaliar os resultados de algoritmos de processamento de documentos ainda demandam esforços. Destaca-se a falta de critérios objetivos e quantificáveis, tornando-se um grande empecilho na comparação dos resultados dos métodos.

O estudo de melhorias no método proposto, ou mesmo a inserção de novas etapas, pode aprimorar os resultados. Sugere-se a criação de novas etapas para tratar problemas específicos, como por exemplo o fenômeno conhecido como "interferência frente-verso", ocorrendo em documentos escritos em ambos os lados e fazendo a tinta de uma lado ser visível do outro.

Diversos outros acervos do APM podem ter a qualidade visual dos documentos

aprimorada através de técnicas de processamento digital de imagens. Novos estudos podem ser realizados para determinar a viabilidade da aplicação de técnicas existentes ou a implementação de abordagens específicas. O APM guarda documentos de variadas fases históricas de Minas Gerais, a grande diversidade deste acervo propicia a ocorrência de grande variedade de problemas que podem tornar-se foco de novos estudos.

A utilização do método apresentado neste trabalho combinado com modernos sistemas de informação, em especial ferramentas de recuperação de informação, pode garantir excelentes níveis de qualidade de acesso à acervos documentais. Esta abordagem propicia melhores resultados nas pesquisas realizadas pelos consulentes, além de, reduzir tempo e esforços de todas as partes envolvidas. Finalmente, acredita-se que o desenvolvimento de tecnologias de extração de conteúdo e recuperação de informação sejam peças-chave para viabilizar o pleno acesso aos acervos documentais, garantindo qualidade e eficiência.

# Referências Bibliográficas


[ANDRADE, 1998] ANDRADE, Nélson S. de; ARAÚJO, A. A.; MELO, C. H. de. *A multimedia information system for governamental historical documents*. Proceedings (CD-ROM) of the Museums and the Web: An International Conference, Toronto, Canada, 1998.

[ANDRADE, 2000] ANDRADE, Nelson Spangler de; ARAÚJO, Arnaldo de Albuquerque. *Multimídia para Acesso a Acervos Históricos*. Ip Informática Pública, Belo Horizonte, v. 2, n. 1, p. 49-66, 2000.

[AQUINO, 2006] AQUINO, Maria Aparecida de. *As Vísceras expostas do autoritarismo*. Revista do Arquivo Público Mineiro – RAPM, Belo Horizonte, MG, Brasil, v. XLII, n. 1, p. 20-39, 2006.

[BAIRD, 2004] BAIRD, Henry S. *Difficult and Urgent Open Problems in Document Image Analysis for Libraries*, DIAL'04, p. 25-32, 2004.

[BINKLEY, 1939] BINKLEY, Robert C. *Strategic Objectives in Archive Politics*. American Archivist, p. 162-168, julho 1939. apud [CONWAY, 1997]

[BRITTO JR, 2001] BRITTO JR, Alceu de Souza et al. *Técnicas em Processamento e Análise de Documentos Manuscritos*. Rita n. 2, p. 47-68, 2001.

[CONWAY, 1997] CONWAY, Paul. *Preservação no universo digital*. Coord. Ingrid Beck, Trad. Olga Marder. Rio de Janeiro, Arquivo Nacional, 1997. 24p. (Tradução de Preservation in the digital world).
Disponível em: <http://www.clir.org/pubs/reports/conway2/index.html> Acesso em: 15 fev. 2007.

[CUNEIFORM, 2006] Cuneiform 6.0 for Windows. Disponivel em: <http://www.ocr.com>. Acesso em: 10 dez. 2006.

[DROETTBOOM, 2003] DROETTBOOM, M. *Correcting broken characters in the recognition of historical printed documents*. Joint Conference on Digital Libraries, p. 364-366, 2003.

[GATOS, 2004] GATOS, B.; PRATIKAKIS, I.; PERANTONIS, S.J. *An adaptive binarization technique for low quality historical documents*. IAPR Workshop on Document Analysis systems, LNCS 3163, p. 102-113, 2004.

[GONZALEZ, 2000] GONZALEZ, Rafael C.; WOODS, Richard E. *Processamento de Imagens Digitais*. Edgard Blücher Ltda, 2000.

[IMAGEJ, 2006] ImageJ 1.38x. Disponivel em: <http://rsb.info.nih.gov/ij/>. Acesso em: 02 out. 2006.

[JAVA, 2007] Java. Disponivel em: <http://java.sun.com/>. Acesso em: 08 ago. 2007.

[KAPUR, 1985] KAPUR, J.N.; SAHOO, P.K.; WONG, K.C. *A New Method for Gray-Level Picture Thresholding using the Entropy of the Histogram*, Computer Vision, Graphics and Image Processing, p. 29, 1985.

[KAVALLIERATOU, 2005] KAVALLIERATOU, Ergina. *A Binarization Algorithm Specialized on Document Images and Photos*, Eighth International Conference on Document Analysis and Recognition (ICDAR'05), p. 463-467, 2005.

[KAVALLIERATOU, 2005b] KAVALLIERATOU, E.; ANTONOPOULOU, H. *Cleaning and Enhancing Historical Document Images*, AICVS - Advanced Concepts for Intelligent Vision Systems, LNCS 3708, p. 681-688, 2005.

[KAVALLIERATOU, 2006] KAVALLIERATOU, Ergina; STAMATATOS, *Efstathios. Improving the Quality of Degraded Document Images*, Second International Conference on Document Image Analysis for Libraries (DIAL'06), p. 340-349, 2006.

[KIRK, 2007] KIRK, Richard. *Enhance contrast,* ImageJ Documentation Portal. Disponivel em: <http://imagejdocu.tudor.lu/imagej-documentation-wiki/gui-commands/enhance-contrast.>. Acesso em: 08 jun. 2007.

[LEE, 1990] LEE, S.U.; CHUNG, S.Y.; PARK, R.H. *A comparative performance study of several global thresholding techniques for segmentation*. Comput. Vision, Graphics, Image Proc., v. 52, n. 2, p. 171-190. apud [GONZALEZ, 2000, p.340]

[LEEDHAM, 2002] LEEDHAM, G et al. *Separating Text and Background in Degraded Document Images*, Proceedings Eighth International Workshop on Frontiers of Handwriting Recognition, p. 244-249, 2002.

[LIMA, 2007] LIMA, Clarissa Costa. *Preservação Digital: A Experiência da Pesquisa Guignard*. 100 f. Dissertação Mestrado em Artes Visuais, Universidade Federal de Minas Gerais, Belo Horizonte, 2007.

[LINS, 1995] LINS, R.D. et al. *An Environment for Processing Images of Historical Documents*. Microprocessing & Microprogramming, p. 111-121, 1995.

[MATTANA, 1999] MATTANA, Marco F., FACON, Jacques, BRITTO JR, Alceu Souza, *Evaluation by recognition of thresholding-based segmentation techniques on Brazilian bank checks*, Proceedings of SPIE, v. 3572, 3rd Iberoamerican Optics Meeting and 6th Latin American Meeting on Optics, Lasers, and Their Applications, p. 344-348, 1999.

[MELLO, 1999] MELLO, C.A.B; LINS, R.D. *A Comparative Study on OCR Tools*, Vision Interface 99, 1999.

[MOTTA, 2006] MOTTA, Rodrigo Patto Sá. *Ofício das sombras*. Revista do Arquivo Público Mineiro – RAPM, Belo Horizonte, MG, Brasil, v. XLII, n. 1, p. 52-67, 2006.

[NABUCO, 2006] Projeto Nabuco. Disponivel em: <http://www.di.ufpe.br/~nabuco>. Acesso em: 02 dez. 2006.

[NIBLACK, 1986] NIBLACK, W. *An Introduction to Digital Image Processing*. Englewood Cliffs, N. J., Prentice Hall, 1986. p. 115-116.

[OTSU, 1979] OTSU, N. *A Threshold Selection Method from Gray-level Histograms*, IEEE Transactions on Systems, Man and Cybernetics, v. SMC 9, n. 1, p. 62-66, 1979.

[PETROU, 1991] PETROU, M; KITTLER, J. *Optimal edge detector for ramp edges*, IEEE Trans. Pattern Analysis and Machine Intelligence, v. 13, p. 483–491, 1991.

[SAHOO, 1998] SAHOO, P.K. et al. *A Survey of Thresholding Techniques*. Computer Vision, Graphics and Image Processing., v. 4, p. 233-260, 1998. apud [GONZALEZ, 2000, p.340]

[SAUVOLA, 2000] SAUVOLA, J.; PIETIKAINEN, M. *Adaptive Document Image Binarization*. Pattern Recognition 33, p. 225-236, 2000.

[SHI, 2005] SHI, Z.; GOVINDARAJU, V. *Historical Document Image Segmentation Using Background Light Intensity Normalization*, SPIE Document Recognition and Retrieval XII, p.16-20, 2005.

[SILVA, 2006] SILVA, J. M. M. Da; LINS, R.D.; ROCHA JR, V.C. da. *Binarizing and Filtering Historical Documents with Back-to-Front Interference*, ACM-Document Engineering-2006, Dijon, França, 2006.

[SOMBRA, 1996] SOMBRA, Luiz Henrique. *Departamento Federal de Segurança Pública: ruptura ou permanênia?* In: Dops a lógica da desconfiança. Rio de Janeiro: Secretaria de Estado da Justiça, Arquivo Público do Estado, p. 37-41, 1996.

[TESSERACT, 2007] Tesseract OCR versão 2.0 Windows. Disponivel em: <http://code.google.com/p/tesseract-ocr/>. Acesso em: 10 nov. 2007.

[VALLE JR, 2002] VALLE JR., E.A; ARAÚJO, A. de A. *Preserving Historical Collections Using Multimedia Information Systems*, Proceedings of the VIII Brazilian Symposium on Multimedia and Hypermedia Systems - SBMIDIA, Thesis and Dissertation Workshop, Fortaleza, Ceará, Brasil, p. 317-324, 2002.

[VALLE JR, 2005] VALLE JR., E.A.; ARAÚJO, A. de A. *Digitalização de Acervos, Desafio para o Futuro*. Revista do Arquivo Público Mineiro - RAPM, Belo Horizonte, MG, Brasil, v. XLI, p. 128-143, 2005.