Regiane Kowalek Hanusiak

Verificação da Autoria de Manuscritos com Base em Atributos Genéticos e Genéricos da Escrita

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Curitiba
Maio / 2010

Regiane Kowalek Hanusiak

Verificação da Autoria de Manuscritos com Base em Atributos Genéticos e Genéricos da Escrita

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Área de Concentração: *Metodologias e Técnicas de Computação*

Orientador: Prof. Dr. Edson J. Rodrigues Justino
Co-orientador: Luiz Eduardo Soares de Oliveira

Curitiba
Maio / 2010

Hanusiak, Regiane Kowalek

Verificação da Autoria de Manuscritos com Base em Atributos Genéticos e Genéricos da Escrita. Curitiba, 2010, 97p.

Dissertação de Mestrado – Pontifícia Universidade Católica do Paraná - Programa de Pós-Graduação em Informática Aplicada.

1. Grafoscopia 2. Atributos Genéricos e Genéticos 3. Verificação de Manuscritos 4. *Support Vector Machine* 5. Características de Textura I.Pontifícia Universidade Católica do Paraná. Centro de Ciências Exatas e Tecnologia. II.Programa de Pós-Graduação em Informática Aplicada.

Dedico este trabalho em memória de meu pai Eugênio Hanusiak, meu exemplo de valores pessoais, à minha mãe Regina Hanusiak e ao meu namorado pelo carinho e compreensão.

Agradecimentos

A Deus que me ajudou me deu forças para superar os problemas durante todo o desenvolvimento deste trabalho.

À minha família que sempre me apoiou, principalmente meu pai, o qual me ensinou enquanto em vida, virtudes como a paciência, a dedicação e a busca de realização de meus sonhos.

Ao meu namorado, que me proporcionou muita alegria, carinho e compreensão.

Ao Professor Orientador Dr. Edson Justino por toda amizade e compreensão nos momentos difíceis e também por todo conhecimento, contribuições, incentivos e ajuda no trabalho. Ao Professor Co-Orientador Luiz Eduardo Soares de Oliveira, pelas suas ajudas com contribuições científicas.

À Pontifícia Universidade Católica do Paraná, em especial ao Programa de Pós-Graduação em Informática Aplicada (PPGIA), pelo apoio estrutural e organizacional que permitiu a realização deste trabalho.

A todos que direta ou indiretamente contribuíram para a realização deste trabalho.

SUMÁRIO

LISTA DE FIGURAS.....X
LISTA DE TABELAS..... XII
LISTA DE ABREVIATURAS E SIGLAS.....XIV
LISTA DE SÍMBOLOS..... XV
RESUMO.....XVII
ABSTRACT.....XIX
1. INTRODUÇÃO.....1
1.1 DESAFIOS..... 4
1.2 MOTIVAÇÃO..... 4
1.3 OBJETIVOS..... 5
1.4 CONTRIBUIÇÕES..... 6
1.5 ORGANIZAÇÃO DA DISSERTAÇÃO..... 6
2. ESTADO DA ARTE.....7
2.1 AVANÇOS NA IDENTIFICAÇÃO E NA VERIFICAÇÃO DE AUTORES..... 7
2.2 ABORDAGENS ATUAIS..... 8
2.3 VISÃO CRÍTICA..... 19
3 FUNDAMENTAÇÃO TEÓRICA..... 21
3.1 UM HISTÓRICO SOBRE A VERIFICAÇÃO DA AUTORIA DE MANUSCRITOS..... 21
3.1.1 Ciência Forense..... 21
3.1.2 Documentos Manuscritos.....22
3.2 A VERIFICAÇÃO DE AUTORIA DE MANUSCRITOS PARA FINS FORENSES..... 25
3.2.1 Dificuldades a serem enfrentadas..... 26
3.2.2 Grafoscopia..... 29
3.2.3 Atributos genéricos e genéticos da Grafoscopia.....33

3.3 RECONHECIMENTO DE PADRÕES..... 36
3.3.1 Tipos de Abordagens..... 37
3.3.2 Extração de Características..... 39
3.3.2.1 Características relacionadas à textura do traçado..... 39
3.3.2.1.1 Filtros de Gabor..... 40
3.3.2.1.2 Matriz de co-ocorrência (GLCMs) e Descritores de Haralick..... 41
3.3.2.2 Inclinação Axial..... 45
3.3.3 Treinamento de Modelos..... 45
3.3.4 Classificação..... 46
3.3.4.1 K-NN (*k-nearest neighbors*)..... 47
3.3.4.2 SMV (*Support Vector Machine*)..... 48
3.3.4.3 WED (*Weighted Euclidian Distance*)..... 50
3.3.4.4 Fusão de Resultados..... 51
3.4 COMENTÁRIOS FINAIS..... 51
4. MÉTODO PROPOSTO..... 53
4.1 INTRODUÇÃO..... 53
4.2 AQUISIÇÃO E PREPARAÇÃO DA BASE..... 57
4.3 PRÉ-TRATAMENTO..... 59
4.3.1 Segmentação do manuscrito com base nos atributos genéticos e genéricos..... 59
4.3.2 Processamento para 16, 8, 4 e 2 níveis de cinza..... 66
4.3.3 Binarização..... 67
4.3.4 Detecção das bordas da escrita por Dilatação e Erosão..... 68
4.3.5 Divisão do manuscrito em fragmentos..... 71
4.4 EXTRAÇÃO DE CARACTERÍSTICAS..... 73
4.4.1 GLMCs e Descritores de Haralick..... 74
4.4.2 Inclinação Axial..... 79
4.5 COMPARAÇÃO..... 81
4.6 DECISÃO..... 83

4.7 COMENTÁRIOS FINAIS.....84
5. EXPERIMENTOS E ANÁLISE DE ERROS..... 85
5.1 BASE DE MANUSCRITOS..... 85
5.1.1 Treinamento..... 86
5.1.2 Testes..... 87
5.2 EXPERIMENTOS COM DESCRITORES DE HARALICK..... 87
5.3 EXPERIMENTOS COM INCLINAÇÃO AXIAL..... 95
5.4 COMENTÁRIOS FINAIS.....95
6. CONCLUSÃO.....96
REFERÊNCIAS BIBLIOGRÁFICAS.....98

Lista de Figuras

Figura 1	Exemplo de um manuscrito, obtido por digitalização, da base de cartas PUC-PR.	23
Figura 2	Imagem de um manuscrito que possui grande espaçamento lateral de margem e entre parágrafos.	28
Figura 3	Exemplo de fragmento de um manuscrito digitalizado que possui grande espaçamento entre palavras.	29
Figura 4	(a) Texto com conteúdo conexo; (b) Texto com alinhamento irregular; (c) Texto ilegível.	38
Figura 5	Exemplo de posições de *pixels*.	43
Figura 6	Exemplo de escita com inclinação axial: (a) à direita; (b) à esquerda; (c) nula.	45
Figura 7	Esquema do processo de decisão na verificação de manuscritos baseado na visão pericial [BARANOSKI, 2005]	54
Figura 8	Comparação das etapas no processo de verificação de manuscritos: (a) processo de análise e decisão pericial e (b) no método computacional estabelecido neste trabalho.	55
Figura 9	Exemplo de um manuscrito digitalizado da base PUC-PR.	58
Figura 10	Exemplos de trechos selecionados pelo algoritmo *fill area.*	60
Figura 11	Exemplo de trecho que ao ser recortado pelo *bounding box*, trouxe consigo parte da escrita não selecionada pelo *fill area* e que é eliminada pelo algoritmo.	61
Figura 12	Um exemplo do processo de segmentação; (a) Exemplo de uma carta da base PUCPR; (b) Exemplo da carta segmentada; (c) melhor vizualização do texto segmentado.	63
Figura 13	Fragmento de um manuscrito da base original PUC-PR; (a) Fragmento selecionado em 256 níveis de cinza; (b) O mesmo fragmento após ser binarizado pelo método Otsu.	67
Figura 14	Fragmento de um manuscrito da base compactada PUC-PR; (a)	68

	Fragmento selecionado em 256 níveis de cinza; (b) O mesmo fragmento após ser binarizado pelo método Otsu.	
Figura 15	Fragmento de um manuscrito da base original PUC-PR; (a) Fragmento binarizado; (b) fragmento erodido; (c) fragmento dilatado; (d) fragmento com bordas extraídas.	69
Figura 16	Fragmento de um manuscrito da base compactada PUC-PR; (a) Fragmento binarizado; (b) fragmento erodido; (c) fragmento dilatado; (d) fragmento com bordas extraídas.	70
Figura 17	Exemplo de manuscrito segmentado da base original PUC-PR.	72
Figura 18	Exemplo de manuscrito segmentado da base compactada PUC-PR.	73
Figura 19	Exemplo de um elemento estruturante com comprimento k=5 e L = 17 direções.	81
Figura 20	Diagrama do processo de cálculo da distância Euclidiana.	82
Figura 21	Curvas ROC geradas com o uso do *kernel* linear (a) e gaussiano (b) para os descritores de Haralick.	89
Figura 22	Curvas ROC comparando os descritores para dois níveis de cinza.	90
Figura 23	Curvas ROC geradas com o uso do *kernel* gaussiano para os descritores de Haralick	93

Lista de Tabelas

Tabela 1	Resumo do Estado da Arte	18
Tabela 2.1	Atributos genéricos e genéticos.	34
Tabela 2.2	Exemplos de atributos genéricos e genéticos.	34
Tabela 3	Descritores de Haralick utilizados neste trabalho.	44
Tabela 4	*Kernels* do SVM.	49
Tabela 5	Exemplos dos atributos genéticos e genéricos observáveis após o processo de segmentação.	66
Tabela 6	Descritores de Haralick utilizados neste trabalho.	74
Tabela 7	Resultados obtidos com os descritores individualmente, com as imagens dos fragmentos em 16 níveis de cinza.	88
Tabela 8	Resultados obtidos com a variação no número de escritores na geração do modelo para dois níveis de cinza, com regra de fusão pela soma.	89
Tabela 9	Resultados obtidos com a variação nos níveis de cinza dos fragmentos manuscritos para 50 escritores de treinamento e 115 para testes, com regra de fusão pela soma.	90
Tabela 10	Resultados obtidos com os diferentes tipos de votos para o descritor entropia, com fragmentos em 2 níveis de cinza.	91
Tabela 11	Resultados obtidos com variando a quantidade de escritores para teste, sendo 75 escritores fixos para treinamento.	91
Tabela 12	Resultados obtidos com variando a quantidade de escritores para teste, sendo 50 escritores fixos para treinamento.	92
Tabela 13	Resultados obtidos com a combinação de classificadores.	92
Tabela 14	Resultados obtidos pela fusão dos resultados obtidos pelos 6 diferentes descritores.	94
Tabela 15	Resultados obtidos das diferentes bases, utilizando a característica de entropia.	94
Tabela 16	Resultados obtidos das base original PUC-PR, utilizando a característica de inclinação axial.	95

Tabela 17 Resultados obtidos das base compactada PUC-PR, utilizando a característica de inclinação axial. 95

Lista de Abreviaturas e Siglas

DPI	*Dot per inch*
FDA	*Fisher Discriminant Analysis*
GLCMs	*Gray-Level Co-occurrence Matrix*
K-NN	*K- Nearest Neighbors*
MMH	Hiperplano de Margem Máxima
OCR	*Optical Character Recognition*
PUC-PR	Pontifícia Universidade Católica do Paraná
RBF	Redes com Funções de Base Radial
ROC	*Receiver Operating Characteristic*
ROI	Regiões de Interesse de um Manuscrito
SRM	Minimização de Risco Estrutural
SVM	*Support Vector Machine*
WD	*Writer- Dependent*
WED	*Weighted Euclician Distance*
WI	*Writer–Independent*

Lista de Símbolos

Filtros de Gabor

$g(x, y)$	Função Gaussiana 2-D
h_e e h_o	Par de filtros de Gabor
f	Frequência
θ	Função radial
$N \times N$	Tamanho da imagem

K-NN

X	Valores das características de um autor X
Y	Valores das características de um autor Y
K	Valor dos vizinhos mais próximos
$d(x, y)$	Distância Euclidiana

SVM

$K(x_i, x)$	Função do *Kernel*
$\vec{p}$	Vetor de pesos
b	Limiar
$\vec{x}$	Padrões de entrada
$f(\vec{x})$	Função de decisão do SVM
S_l	Conjunto de Treinamento
w	Grupo ou classe
ξ	Magnitude do erro de classificação
C	Penalidade de erro no SVM
V_m	Voto Majoritário

WED

f	Característica de entrada
k	Distância mínima
N	Número total de características
d	Distância Euclidiana
M_{ki}	Amostras de manuscritos de autoria desconhecida (referência)
M_o	Amostra do manuscrito de autoria desconhecida (questionada)
L	Características grafoscópicas
D_i	Decisão
R_i	Laudo pericial resultante

GLCMs

d	Distância entre os *pixels*
θ	Ângulo entre os *pixels*
P(i,j)	Valor do pixel da matriz de co-ocorrência
M	Matriz de co-ocorrência
$p(i,j)$	Valor normalizado da célula da matriz de co-ocorrência
σ	Desvio padrão
μ	Média

Inclinação Axial

K	Comprimento do elemento estruturante
L	Direções
θ	Ângulo

Resumo

A verificação de autoria trata-se de uma atividade relacionada às Ciências Forenses, utilizada para auxiliar na identificação ou constatação de fraudes de documentos. Os textos manuscritos estáticos ou *offline*, as assinaturas e rubricas possuem muitas variabilidades na escrita de um mesmo autor e também semelhanças entre autores diferentes. As variabilidades da escrita de um autor são causadas por vários motivos, como fatores psicológicos, ou fatores como sexo, idade, diferenças de região e cultura, e estas dificultam a distinção de autoria e não-autoria na verificação.

Atualmente, há diversos estudos em que a computação auxilia com técnicas de verificação de autoria de manuscritos em busca de resultados com índices de eficácia necessários em aplicações de técnicas para usuários finais. Estas técnicas devem seguir regras padronizadas para que sejam confiáveis em sua prática.

As abordagens desenvolvidas geralmente possuem as etapas de: aquisição de dados que envolvem a colheita e digitalização de manuscritos; pré-processamento, em que as imagens são preparadas ou alteradas para a etapa de extração de características; a extração de características, relacionadas à grafoscopia para caracterizar a individualidade da escrita; treinamento, produção de um modelo e decisão a partir de um classificador. Particularmente, cada uma das etapas das abordagens possui complexidades que acabam dificultando o desenvolvimento de aplicações computacionais semi-automáticas e automáticas.

Este trabalho apresenta um modelo que envolve uma abordagem global de classificação na etapa de reconhecimento, a partir da análise dos atributos genéticos e genéricos da escrita. A abordagem adotada utiliza características globais em relação à textura do traçado da escrita e a inclinação angular da escrita, com base em imagens de texto segmentado. Na etapa de classificação é utilizado o classificador SVM. O modelo proposto utiliza a base de cartas PUC-PR e contribuiu com a melhoria dos resultados obtidos pelos trabalhos já realizados para procedimentos automáticos e semi-automáticos de verificação, e obteve uma taxa de acerto em torno de 95%, contribuindo também na

redução das dificuldades encontradas em métodos de verificação de autoria de manuscritos questionados.

Palavras-chave: 1. Grafoscopia 2. Atributos Genéricos e Genéticos 3. Verificação de Manuscritos 4. *Support Vector* Machine 5. Características de Textura

Abstract

The authory verification is an activity related to Forensic science, used to support document identification or fraud. The static or offline handwriting, the signatures have many writing variation of the same author and thus similarities between different ones. The writing variation of an author are caused by many motives, like psychological facts, gender, age, religion and cultural differences, and those make the analysis difficult.

Nowadays, there are many studies where computation helps with handwriting verification techniques in search results with efficacy rates required in technical applications to end users. These techniques must follow standard rules to be trustful in their practice.

Generally there are the approaches stages on the developed topics: data acquisition that involves handwriting collection and digitalization; pre-processment, in which the images are prepared or modified for the next stage, features extraction; the features extraction, related to graphology to show the writing individuality; training, model production and decision attached to a classifier. Particularly, each of the stages detects complexity that difficults the development of automatic and semiautomatic computing applications

This project presents a model which involves a global approach to classification in the recognizement stage, from the analysis of genetic and generic attributes of the writing. The approach uses global features in relation to texture and inclination angle of the writing.

The proposed model uses the base handwriting and has contributed to the improved performance of the work already carried out procedures for automatic and semi-automatic authentication and obtained the accuracy around 95%, also reducing difficulties encountered in methods of verification of authorship of manuscripts questioned.

Key-words: *1. Grafoscopy 2. Genetic and Generic Attributes 3. Handwriting Verification 4. Support Vector Machine 5. Texture Features*

Capítulo 1

Introdução

A análise e autenticação de manuscritos são alguns dos desafios da área forense e são atividades relacionadas à análise de documentos questionados, aplicadas em questões judiciais em que há suspeita de fraudes e falsificações de documentos.

A documentoscopia é uma das ciências relacionada ao estudo ou análise de documentos, que considera características como a origem e o tipo do papel, a data de elaboração, elementos tipografados e carimbos, que devem ser considerados na análise pericial e na criação de um laudo. Já a grafoscopia é um campo da Ciência Forense e subárea da documentoscopia que trata somente de aspectos da escrita e sua autoria, como a autenticação de documentos e verificação de autoria de manuscritos.

Os estudos realizados neste campo colaboram para a descoberta de ferramentas para identificação e verificação de autoria de documentos questionados na área judicial, uma vez que auxilia nos resultados para solucionar crimes e identificar suspeitos.

No processo de análise de documentos questionados, podem surgir muitos desacordos, isto devido à falta de normatização de técnicas com base em estudos científicos comprovados que determinam a individualidade da escrita. Assim, o trabalho de análise de documentos questionados realizado pelos peritos, que envolve a aplicação técnicas grafométricas baseadas em atributos grafoscópicos, se torna subjetivo, já que os peritos acabam produzindo laudos de diferentes conteúdos, de acordo com suas respectivas análises.

Outra dificuldade apresentada no procedimento de análise pericial que colabora para a subjetividade é a ausência de métodos computacionais automáticos e semi-automáticos eficientes que sejam seguros e corretos para auxiliar na análise e na

identificação e verificação da autoria do documento questionado, e assim produzir laudos que sejam aceitos juridicamente. Sem o auxilio de soluções computacionais, o trabalho manual executado pelo perito, muitas vezes é exaustivo quando existe uma grande quantidade de documentos.

No contexto da grafoscopia, dois objetos de análise se apresentam: os manuscritos e as assinaturas. Mesmo possuindo características distintas, ambos mantêm uma estreita relação entre si, possuindo a mesma raiz ou origem no processo de aprendizado do escritor. Isto é, carregam consigo as experiências adquiridas pelo escritor, durante o seu processo de aprendizado e posteriormente, através do aperfeiçoamento do estilo pessoal de escrita [BARANOSKI, 2005].

A escrita está sujeita às inúmeras mudanças decorrentes de causas variadas. Por exemplo, a escrita de uma pessoa pode sofrer alterações com o passar do tempo de acordo com fatores psicológicos do autor e/ou utilização de tipos de canetas e texturas de papel. A variabilidade da escrita também pode ser influenciada por fatores como alfabeto, sexo, idade, etnia da população analisada. As variabilidades conhecidas como intrapessoais, ou são as que procedem da instabilidade que existe entre as escritas do mesmo autor. Existem também as similaridades interpessoais, que representam as semelhanças como a forma e o estilo, que ocorrem na escrita de autores diferentes. As variabilidades intrapessoais e as similaridades interpessoais tornam o processo de verificação complexo, causando dificuldades de distinção da autoria e não-autoria.

Muitas pesquisas estão sendo ampliadas no campo de análise para autenticação de documentos com o uso da grafoscopia, por meio de abordagens computacionais automáticas ou semi-automáticas, podendo ser globais e/ou locais, *offline* (também chamada de estática) ou *online* (também chamada de dinâmica), que sejam seguras com possibilidade de serem desenvolvidas, superem as variabilidades da escrita e que assim forneçam informações comprovadas cientificamente dentro de normas estabelecidas.

A verificação computacional de autoria de manuscritos pode estar relacionada a dois tipos de abordagens, quando se refere ao tipo de aquisição de dados: abordagem *off-line* ou estática em que o manuscrito em folha de papel é digitalizado com uma câmera; e a abordagem *online* ou dinâmica em que um dispositivo especial captura a escrita conforme ela vai sendo feita. Em relação à automatização do processo, quando uma abordagem computacional auxiliar automaticamente em algumas etapas do processo de

verificação de autoria de manuscritos, esta é semi-automática. Quando a abordagem computacional realizar todas as etapas da verificação de autoria de manuscritos, desde a aquisição de dados, até a decisão esta é considerada automática.

Os atributos grafoscópicos se apresentam em grande número [GOBINEAU, 1954] [HUBER & HEADRICK, 1999] [MORRIS, 2000]. Para a grafoscopia, os atributos grafoscópicos se dividem em dois elementos relevantes de análise para identificação de manuscritos: atributos genéricos e atributos genéticos. Os primeiros abordam critérios mais globais de análise, tais como a altura, comprimento e forma. Os segundos abordam elementos dinâmicos do traçado, tais como inclinação axial, pontos de ataque e remates [GOBINEAU, 1954].

A verificação automática de documentos manuscritos busca determinar se o manuscrito é do próprio punho do autor ou não. A metodologia deste trabalho propõe uma técnica *offline* de verificação de autoria de documentos manuscritos, baseada na abordagem de análise utilizada pela perícia grafoscópica, através de uma abordagem global. A abordagem global também conhecida como abordagem independente do escritor (*Writer–Independent*), em que um modelo geral é utilizado, não necessitando de novo treinamento se novos autores sejam analisados. A base utilizada na abordagem, trata-se da base PUC-PR, sendo exemplares coletados de origem natural e isenta de qualquer tipo de falsificação. A utilização desta base tem propósitos comparativos em relação a trabalhos realizados anteriormente.

Em termos de classificação, a técnica de verificação de autoria deste trabalho será baseada em na abordagem global, em que existem duas classes: a de associação w_1 é composta por exemplares genuínos (autoria), enquanto a classe de dissociação w_2 é composta por exemplares de autores distintos (não-autoria). A dissimilaridade de verificação é aplicada em cima da extração dos atributos grafoscópicos (atributos genéricos e genéticos do escritor). Os atributos são extraídos a partir da produção da textura através do texto escrito do autor, que é obtida por meio de um pré-processo de segmentação do texto e compactação, em que é gerada uma textura.

1.1 Desafios

A partir do estudo realizado, uma abordagem de verificação de autoria de manuscritos é proposta e também associada a outras, em busca de melhores resultados para obter mais confiança aos métodos utilizados computacionalmente e assim aplicá-los ao trabalho de verificação e identificação de documentos, reduzindo as dificuldades encontradas na área de análise pericial de manuscritos para questões judiciais, como a subjetividade que ocasiona desacordos.

Este trabalho teve como desafio pesquisar e propor uma técnica computacional semi-automática para auxiliar o trabalho de verificação da autoria da escrita manuscrita em documentos. A técnica proposta está associada primeiramente à etapa de preparação das imagens digitalizadas, extração e compactação do conteúdo da escrita para se obter uma textura, extração de características globais a partir da textura formada e a classificação.

O principal desafio é propor um método com utilização de abordagem global que auxiliem o processo de verificação de manuscrito e superem as complexidades que a escrita traz, como por exemplo, as variabilidades.

1.2 Motivação

A partir dos problemas encontrados no processo de análise pericial, a motivação deste trabalho é criar soluções computacionais que sejam úteis na atividade de verificação de autoria. Outra motivação é fazer com que estas colaborarem para minimizar a subjetividade da análise, desde que as abordagens propostas realizem uma comparação confiável entre o documento questionado e o documento conhecido.

A solução apresentada foi elaborada de tal forma que forneça respostas confiáveis para a realização de laudos aceitáveis, devendo ser aprovada por um rigoroso processo de avaliação que envolve resultados estatísticos [JUSTINO, 2002].

A abordagem computacional criada neste trabalho é baseada nos princípios da grafoscopia, ou seja, nas técnicas que a perícia utiliza no processo de identificação da autoria de manuscritos. A análise bibliográfica mostra que muitos estudos vêm sido realizados envolvendo a textura obtida através da compactação da escrita do autor. Assim

sendo, a abordagem adotada neste trabalho baseou-se nestes estudos para buscar uma metodologia que possa contribuir com os trabalhos já realizados.

1.3 Objetivos

Este trabalho tem como objetivo principal apresentar um modelo próximo da realidade, com base na visão pericial, contendo as etapas de aquisição de base, pré-tratamento de imagens, extração de características, comparação e decisão. Assim, apresenta-se uma abordagem computacional para auxiliar a verificação de autoria de documentos manuscritos para extração de características globais a partir de atributos grafoscópicos. A classificação da abordagem é baseada na abordagem independente do escritor, WI (*Writer–Independent*), em que existem as classes de associação e dissociação.

Abaixo são listados outros objetivos secundários de acordo com o escopo da abordagem:

- Propor uma técnica para implementação dos métodos de segmentação e compactação do texto, capaz de eliminar os efeitos indesejáveis, produzidos pela segmentação em linhas e palavras;
- Estudar as relações entre as características da grafoscopia e as possíveis aplicações computacionais relacionadas à extração de características globais da escrita manuscrita;
- Realizar estudos comparativos em relação as técnicas já implementadas em realação à textura e inclinação axial;
- Estudar o desempenho das características utilizadas no contexto individual (similaridades interpessoais) e em grupo (variações intrapessoais);
- Estudar a fusão dos resultados dos classificadores, tendo em vista a combinação das características;
- Através da técnica de segmentação e compactação do texto, criar uma base nova com redução dos espaços em branco, para ser trabalhada por outras características globais;
- Realizar a extração de características em relação à textura da escrita, e re-implementar a extração de características de inclinação axial comparando

resultados da base com texto compactado, em relação à base original, em busca de resultados mais precisos.

1.4 Contribuições

Este trabalho apresenta as seguintes contribuições:

- Um método computacional que auxilie a análise pericial, tanto nas etapas de extração de características como no processo da verificação de autoria;
- Criação de uma nova base, a partir da base original de manuscritos digitalizados (PUC-PR), com imagens contendo a concentração do conteúdo da escrita reduzindo os espaços em branco da imagem. Esta base poderá ser utilizada em trabalhos futuros;
- Um estudo que relaciona as características grafoscópicas aos modelos computacionais baseados em uma abordagem global de segmentação de conteúdo, e nesse caso a textura;
- Resultados a partir da classificação baseada na abordagem independente do escritor, em que existem duas classes, de autoria e não-autoria;

1.5 Organização da Dissertação

Este documento de dissertação é apresentado em seis capítulos sendo eles: introdução, em que é descrito um breve resumo sobre a contextualização da verificação semi-automática de manuscritos, os principais problemas encontrados e objetivos gerais e específicos do trabalho. O segundo capítulo aborda a fundamentação teórica dos campos em que o trabalho foi baseado, inclusive elementos do processo computacional de verificação de manuscritos. Estes elementos são utilizados por abordagens recentes, sendo estas apresentadas na revisão de trabalhos no terceiro capítulo, que serviram como base para o desenvolvimento deste trabalho. O quarto capítulo apresenta as etapas do método proposto e implementa a abordagem do trabalho. Os resultados dos experimentos realizados são apresentados no quinto capítulo e, por último, o sexto capítulo apresenta as conclusões do trabalho e propostas que poderão ser realizadas futuramente.

Capítulo 2

Estado da Arte

Neste capítulo serão abordados métodos recentes de verificação de autoria, que incorporam em sua abordagem, técnicas de extração de características de textura, que representam o foco deste trabalho. Também são citados métodos que envolvem a extração de característica a partir do ângulo da escrita.

Atualmente, há uma quantidade importante de diferentes abordagens para a identificação e verificação de autoria. Estas abordagens adotam diferentes formas de extração de características e classificadores. A maioria dos trabalhos desenvolvidos recentemente propõe atividades futuras para dar continuidade e realizar melhorias para obter melhores resultados, a partir dos métodos já propostos.

Entre as abordagens, destacam-se as globais, que basicamente, são descritas em seguida por estarem relacionadas com as características de textura de inclinação do manuscrito, e são utilizadas como referência.

2.1 Avanços na identificação e verificação de autoria

Alguns sistemas atuais na área forense são destacados: Fish/BKA, Script/TNO e Cedar/FOX possuem grande confiança na perícia humana, que oferecem medidas manuais como características gerais geométricas da escrita, espaçamento de linhas e entre palavras e o ângulo total de inclinação do manuscrito. Entretanto os métodos manuais são muito custosos e podem ser subjetivos [SCHOMAKER, 2007]. É mais apropriado utilizar sistemas automáticos, em que é possível selecionar regiões de interesse (ROI) de um manuscrito, para as características serem automaticamente computadas e comparadas

entra amostras. O sistema também pode, através dos Modelos Escondidos de Markov, segmentar um manuscrito em linhas para comparação entre amostras. Nos sistemas semi-automáticos, os caracteres são selecionados e suas respectivas características são computadas e designadas ao sistema OCR. São utilizadas as características de Gabor, envolvendo histogramas de ângulos de inclinação e curvatura de textura do manuscrito. Para elementos da forma da grafia, são utilizadas características do caractere ou da forma do caractere, imagens de tamanho normalizado e componentes conectados. Os descritores de Fourier estão relacionados com as características de colocações (*layout*) e deposição de tinta no manuscrito. Schomaker [SCHOMAKER, 2007] afirma que embora o desempenho das características de colocação (*layout*) seja menor do que as de textura e relacionadas à forma da escrita, se as características de colocação forem combinadas com outras características, acabam impulsionando o desempenho no sistema total.

Para a comparação de métodos, devem ser analisados, fatores como o número de parâmetros, esforço gasto para o treinamento do sistema e a quantidade de texto. Os métodos mais apropriados são os que não necessitam de treinamento em nível de escritores individuais e utilizam somente poucas linhas de texto.

Atualmente as tecnologias de caixa-preta utilizam características globais de textura da imagem e distâncias para cálculo de similaridade. A desvantagem é que apesar do avanço dos sistemas de caixa-preta, eles ainda não respondem a algumas dúvidas. A vantagem é que a combinação de abordagens, futuramente, pode ser auxiliar em relatórios verbais com a descrição dos resultados. Outra vantagem é que desde que o sistema judicial tem utilizado o raciocínio Bayesiano, será possível no futuro criar relatórios que sejam legíveis ao ser humano, inclusive relações de probabilidade para as decisões obtidas pelo sistema.

2.2 Abordagens atuais

Franke [FRANKE, 2002] apresenta uma abordagem global para auxiliar computacionalmente o reconhecimento de tinta utilizado em um manuscrito, a partir da análise da escrita, sendo utilizados 62 tipos de canetas e refis para produzir 737 amostras, em que 368 amostras foram utilizadas para testes.

Segundo Franke [FRANKE, 2002], características como o ritmo da pressão da escrita e a estrutura do curso são relevantes para diferenciar manuscritos verdadeiros de falsos. Os métodos atuais relacionados ao domínio de textura, não consideram materiais como o papel e a tinta. A desvantagem dos modelos desenvolvidos até agora, era a seleção manual do modelo de distribuição de tinta de acordo com o tipo de caneta correspondente. Sendo assim, um modelo de distribuição de tinta foi proposto por Franke [FRANKE, 2002] para superar as fraquezas da análise de variações de largura do curso, podendo ser adaptado às propriedades específicas de lápis, canetas esferográficas e canetas-tinteiro.

Na abordagem, três classes fundamentais de tinta foram selecionadas: sólida (grafite), viscosa (esferográfica) e fluída (gel). Para a produção das amostras de testes foram utilizadas 8 lápis/lapiseiras de tinta sólida 28 canetas esferográficas de tinta viscosa e 26 canetas de tinta fluída. Uma sentença foi escrita com cada uma das canetas em 5 folhas do mesmo papel.

O primeiro procedimento realizado foi a extração das características, em que a computação da matriz de co-ocorrência é feita, e a partir da mesma, são calculadas quatorze características: energia, correlação, momento inverso da diferença, entropia, diferença de soma e contraste, diferença e soma da energia, diferença e soma da média, diferença e soma das divergências, diferença e soma das entropias. Na segunda etapa Franke [FRANKE, 2002] utiliza da técnica FDA (Análise Discriminante de Fisher) e a busca exaustiva pela característica de maior valor, através do classificador de um vizinho mais próximo (KNN – 1NN) para avaliar se as 14 características extraídas de acordo com a matriz de co-ocorrência permitem a identificação de tintas em geral. Com esta técnica foi possível obter uma boa separação dos três tipos de classes de tintas. Para a última etapa, a classificação, foi utilizado um SVM com *kernel* polinomial, que através de seu algoritmo de treinamento, busca o hiperplano ótimo de separação.

O resultado da taxa de acerto do reconhecimento do tipo de tinta foi de 99.7% para imagens digitalizadas em 600 DPI e 98.4% para imagens em 300 d.p.i. O método apresenta resultados, indicando que a combinação de características pode influenciar na taxa de erro. A vantagem desta abordagem é que apesar de não estar ligada diretamente à identificação de autoria, ela pode ser colaborativa ao reconhecimento de autores, pois de acordo com o resultado obtido, mostra que as características de textura têm grande

importância. Outros estudos podem ser realizados também, como a aplicação de modelos adaptados da distribuição da tinta para análise de manuscritos pseudo-dinâmicos, descobrindo falsificações [FRANKE, 2002].

A abordagem de Said [SAID, 1998] propõe uma abordagem baseada na análise na textura em que cada manuscrito do autor é considerado como uma textura diferente. A abordagem apresenta um algoritmo automático de identificação *offline* de autor, a partir das imagens digitalizadas dos manuscritos, em 150 d.p.i, de textos independentes e inclinados não-uniformemente. A maioria das técnicas desenvolvidas utiliza textos com mesmo conteúdo, como as técnicas para verificação de assinaturas, em que o autor escreve o mesmo texto fixo, porém, estas técnicas de texto-dependente do autor podem ser propensas a falsificação da identificação. A utilização de texto-dependente possui outra desvantagem: não é aplicável na identificação de autores de manuscritos arquivados, e também da identificação de suspeitos de crimes na área de ciências forenses. Uma das vantagens, é que o método proposto por Said funciona para casos em que existem ruídos. Foram utilizadas 150 originais de teste de 10 autores. Os procedimentos da abordagem são: digitalização das imagens, normalização, extração de características e identificação.

As palavras e caracteres podem não ficar alinhados em posição correta após a digitalização. Essas inclinações podem afetar consequentemente, a identificação do autor, nas etapas finais. Os métodos em geral são limitados para imagens de textos com linhas inclinadas uniformemente. Esta é a vantagem do método de Said, que introduz uma etapa de pré-processamento em que se detecta o ângulo de inclinação das palavras individuais para realizar a normalização, utilizando linhas adequadas e componentes conectados, sendo assim, um método aplicável também para linhas inclinadas não uniformemente. Para estimar a inclinação, é realizada a projeção do perfil vertical/horizontal, em que é criado um histograma, relativo ao número de *pixels* pretos ao longo das linhas. Segundo o autor, a textura é afetada por diferentes espaçamentos das palavras, pela variação do espaço da linha, etc. A normalização do texto minimiza a influência destes fatores. No método, a partir de uma imagem binária do original, as palavras são normalizadas e os gráficos e pinturas são removidos.

Para o procedimento de extração de características, a abordagem utiliza a técnica multi-canal dos Filtros de Gabor, uma das mais populares e mais reconhecidas, gerando

16 imagens de saída (quatro imagens para cada uma das quatro freqüências: 4, 8, 16, 32), das quais as características de média e desvio padrão são extraídas. Depois, 32 características são calculadas a partir dessas imagens de saída. Foram realizados testes usando estas características separadamente e em conjunto. Outra técnica foi utilizada, as matrizes de co-ocorrência (GLCMs), com cinco distâncias (d = 1, 2, 3, 4 e 5) e quatro direções ($\theta = 0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$. A partir da matriz de co-ocorrência foram extraídas características de energia, entropia e contraste e correlação. Said [SAID, 1998] afirma que a técnica de matriz de co-ocorrência é cara de ser computada.

Na etapa de identificação de autor, a abordagem emprega o uso de dois classificadores: Peso da Distância Euclidiana (WED) e KNN (*k-nearest neighbors*). Existem outros classificadores mais sofisticados como as redes neurais, contudo, Said optou-se pela escolha dos dois classificadores citados, devido sua simplicidade computacional.

Para a classificação 25 segmentos (128 x 128) não sobrepostos, de cada autor, foram extraídos para serem utilizados em dois conjuntos. O conjunto A, com 10 imagens por autor para treinamento e 15 para testes, e o conjunto B, com 15 imagens para treinamento e 10 para testes. As imagens de testes não aparecem no conjunto de treinamento e foram realizados testes combinando diferentes características sob ambos classificadores [SAID, 1998].

O principal resultado foi 96% de exatidão da identificação. Para obter estar taxa, foram utilizadas todas as características extraídas sob os filtros de Gabor, e a classificação, realizada pelo classificador WED (particularmente para o conjunto B). Em relação às características da matriz de co-ocorrência, 96% de exatidão foi obtido através do uso de características envolvendo as distâncias d=1 e d=2 e para a classificação, o KNN. Outros resultados são apresentados no artigo, em que os dois classificadores apresentaram resultado bom, mas em alguns casos o classificador KNN teve baixo desempenho em relação ao classificador WED. Os resultados também indicam que quando utiliza-se a matriz de co-ocorrência e classificador WED, os resultados são menos exatos do que os obtidos com o método de filtros de Gabor [SAID, 1998].

A combinação de características e o uso de dois classificadores mostra que é possível realizar uma melhor análise a partir dos resultados gerados. Além da abordagem global, uma abordagem local foi considerada para buscar características específicas do

autor, para melhorar a exatidão do reconhecimento. Isso indica que a integração desses dois tipos de abordagens em um sistema único, pode melhorar a identificação.

Busch [BUSCH, 2005] apresenta um método para a identificação de textos impressos em diferentes idiomas. O texto impresso é inicialmente submetido a um processo de correção da inclinação das linhas, a fim de eliminar imperfeições do processo de digitalização do documento. Em seguida, é feita a segmentação das linhas, com o objetivo de reduzir os espaços em branco existente entre as mesmas e entre palavras. Quando a linha apresenta pouco texto, assim como em Said [SAID, 1998], o restante da mesma é incrementado com conteúdo redundante obtido na mesma linha. Com base na textura gerada, várias abordagens para a extração de características são utilizadas, entre elas, descritores de textura, filtros de Gabor e Wavelets. Os índices de acerto ficaram entre 91% e 97% sendo que os melhores resultados foram alcançados com Wavelets e filtros de Gabor.

Al-Dmour e Zitar [AL-DMOUR, 2007] apresentam um método para a identificação de manuscritos arábicos. A textura é obtida seguindo um protocolo similar aos vistos anteriormente, isto é, corrigindo a inclinação das linhas de texto e retirando espaços entre linhas e entre palavras. Na extração das características foram utilizadas as matrizes de co-ocorrência, filtros de Gabor e as transformas de Fourier. Os resultados obtidos ficam em torno dos 90% de acerto. O melhor desempenho, nesse caso, ficou com as transformadas de Fourier.

Shen [SHEN, 2002] propôs uma abordagem global de identificação de autoria, que envolve um pré-processamento para normalizar a imagem. Após usa a técnica de Filtros de Gabor para extração de características e o classificador K-NN para identificar. Foram utilizados 110 espécimes de 50 autores, sendo o método independente do conteúdo do texto. A porcentagem de acerto foi de 97.6%.

Nas pesquisas recentes, as técnicas de filtros espaciais multi-canais obtiveram muito sucesso em suas aplicações. Uma vantagem destes filtros é que eles provem resoluções ótimas simultâneas em ambos os espaços e ambos os domínios de frequência [SHEN, 2002]. Por este motivo a transformada de Gabor 2-D foi selecionada por Shen para extrair características.

A abordagem possui dois passos na etapa de pré-processamento, realizada para minimizar a influência de vários fatores e obter uma imagem uniforme. O primeiro passo

do pré-processamento é a redução de ruídos e do fundo da imagem, mantendo apenas o texto escrito; o segundo pré-processamento consiste em uma transformação em escala de cinza e após, uma normalização devido à inclinação das linhas, com objetivo de remover os espaços entre as linhas e palavras. Esta normalização para retirar espaços em branco é realizada para que na próxima etapa, de segmentação, os fragmentos obtidos tragam o texto compactado em forma de textura.

A idéia futura proposta por Shen é que a abordagem seja aplicada para um número maior de autores, mas que se utilize outro classificador como o SVM, que é mais eficaz classificação de grandes quantidades de espécimes.

As técnicas online ainda têm sido estudadas e as técnicas offline possuem muitos problemas porque as informações dos manuscritos só podem ser extraídas de imagens.

Em sua abordagem Zhenyu He [HE, 2005] utiliza como base, 20 manuscritos chineses de 10 pessoas, em que foram criados blocos de 512 x 512 *pixels*, com 64 caracteres cada, sendo estes caracteres de tamanho 64 x 64 *pixels*.

A abordagem possui uma fase de pré-processamento que primeiramente faz a remoção de ruídos, após localiza cada linha de texto e separa cada caractere usando projeção; por último normaliza cada caractere para o mesmo tamanho. Com este pré-processamento obtêm-se blocos de texto (de tamanho 512 x 512 *pixels*) que são as imagens texturizadas.

Zhenyu He [HE, 2005] afirma que a Transformada de Gabor 2-D é uma boa técnica para o método *offline* de reconhecimento de manuscritos, porém possui algumas desvantagens como o alto custo computacional, porque o filtro tem de realizar uma operação de duas entradas e uma saída de toda imagem para cada orientação e frequência. A abordagem utiliza a Densidade Gaussiana Generalizada (GGD) para extração de características que atingiu resultados melhores e reduz muito o tempo decorrido.

Para a classificação foi utilizada a distância euclidiana com o classificador KLD (*Kullback-Leibler Distance*). A porcentagem de precisão foi de 70% para as Transformadas de Gabor e de 80% para o GGD.

Brink [BRINK, 2007] descreve um método *offline* para verificação e identificação de autoria, que codifica características do autor como uma mistura de estilos típicos do manuscrito, em que estes autores são conhecidos como “classe de autores”. A vantagem do método em relação aos demais métodos automáticos existentes é apresentar relatórios

compreensíveis, ou seja, apresentar transparência. Os sistemas de verificação e identificação, apesar de apresentar ótimas implementações e desempenhos, não convencem os peritos, devido as saídas do sistema serem difíceis de interpretar, tornando os sistemas vistos como caixa-preta e seus funcionamentos não serem claros. Cada um dos autores da classe é representado por um original selecionado dos dados de entrada, sendo possível desta forma, visualizar o manuscrito, tornando os vetores de características compreensíveis. O método tem como principal efeito, a redução da dimensionalidade do vetor de características computado [BRINK, 2007].

Duas bases foram utilizadas separadamente nas experiências. A base de Firemaker envolve 252 estudantes que escreveram quatro páginas cada um, de texto cursivo conectado e ilimitado, resultando em 1008 páginas, porém só foram utilizadas as páginas 1 e 4. A base de manuscritos NFI é nova e heterogênea, coletadas pelo NFI (Instituto Forense Nacional Holandês). O conjunto consiste em 3501 documentos digitalizados que foram escritos por 1311 suspeitos criminais. Foram produzidos em média, 2 (dois) manuscritos por autor, de um texto padrão ditado, em que uma parte do texto é cursivo conectado e a outra, com letras maiúsculas. Essa base de dados é considerada como "suja", porque os textos não são lineares e o nível de instrução é baixo. Além disso, algumas páginas contêm rasuras.

Para cada original de entrada, foi computado um vetor de características básicas, através da implementação da característica da dobradiça [BRINK, 2007], que captura a orientação e a curvatura do traço da tinta. Para criar os perfis das classes, os autores podem ser selecionados manualmente para representar, por exemplo, estilos de países, sexo, idades, etc. Na abordagem de Brink [BRINK, 2007], as amostras foram selecionadas aleatoriamente, 50 da base Firemaker e 25 da base NFI e para cada amostra foi computado um perfil de classe, executando-se após, a verificação ou a identificação do autor a partir da base de treinamento. As amostras que possuíram o maior desempenho a partir da base de treinamento foram designadas para o conjunto final de autores de classe. O perfil de classe de um original é indicado pela distância de seu vetor da característica básica e os vetores das características básicas dos originais de cada um dos autores da classe. Assim, um novo vetor de características é usado para discriminar autores, mostrando que os perfis das classes podem ser vistos como uma forma de redução de dimensionalidade [BRINK, 2007].

A etapa de verificação implica o teste de verificar a dissimilaridade, computada pela distância Euclidiana, entre duas classes de perfis estava abaixo de um limiar θ. Caso positivo, a decisão era verdadeira, ou seja, mesmo escritor, se não, escritor diferente. As distribuições de probabilidade das distâncias em ambas as classes foram criadas usando a janelas de Parzen com *kernel* Gaussiano. A identificação do autor foi realizada através da classificação dos *s* vizinhos mais próximos, com *s* = 1, 10 e 100.

A base foi dividida, em que 25% foi separada para treinamento e 75% para testes. Na verificação uma parte do treinamento foi usada para selecionar os autores das classes, computar os perfis e determinar um limiar. Após os autores e o limiar foram usados para executar a verificação. Com os resultados Brink [BRINK, 2007] conclui que o desempenho quando se utiliza os perfis é similar ao desempenho usando a característica da dobradiça, diretamente e que o número de autores não possui muita influência.

Na identificação, uma parte do treinamento também foi usada para selecionar autores e computar perfis e após os autores participaram da execução da identificação do autor. De acordo com os resultados, o método da abordagem não trabalha tão bem para a identificação como para a verificação [BRINK, 2007].

Na identificação, utilizando o vizinho mais próximo, com n=1, foi obtido uma média de 67.3% de resultado verdadeiro-positivo para a base Firemaker, e uma média de 53.5% de resultados de verdadeiro-positivo para a base NFI.

Já na verificação, a média obtida de resultado verdadeiro-positivo para a base Firemaker foi de 96.1% e para a base NFI, 79%. Na verificação, conforme cresce o número de autores (2, 4, 5, 50) o resultado decai, cerca de 8% para a base Firemaker e 4% para base NFI.

A idéia de Brink [BRINK, 2007] é melhorar o método, como por exemplo, através de uma melhoria no pré-processamento dos documentos da base NFI; construir um perfil de classe baseando-se em uma ou mais características básicas; selecionar autores de classe utilizando uma amostragem aleatória ou manual, pelos peritos (para representar grupos de idade, sexo, nacionalidade, etc); e realizar a avaliação através de outras medidas de distâncias.

Na abordagem de Imdad [IMDAD, 2007], é mostrado que as características Dirigidas de Hermite são muito úteis, pois extraem uma grande quantidade de informações em múltiplas escalas, particularmente para dados caracterizados por

características orientadas, curvas e segmentos. O método proposto é baseado nos princípios do Sistema Visual Humano, independente do texto escrito.

As amostras digitalizadas são de coleções de manuscritos antigos, de 30 autores diversificados da base IAM. A base de testes consiste em algumas linhas de manuscritos (na maioria cinco). Apesar dos autores utilizarem um classificador SVM para a classificação, a técnica proposta na abordagem é independente de todo o esquema específico da classificação. Apesar de existirem alternativas além da transformada de Hermite, como a transformada Wavelet e a transformada de Gabor, a transformada de Hermite possui alguns benefícios adicionais em relação às outras duas. Similarmente à transformada de Gabor, a transformada de Hermite também fornece uma resposta semelhante à visão humana, mas em contraste à Gabor, a resposta de Hermite é um modelo mais exato da visão humana [IMDAD, 2007]. Assim como a transformada Wavelet, em que as propriedades da imagem podem ser observadas em múltiplas escalas, a transformada de Hermite também possui estas propriedades, como as características de Hermite que usam filtros escalados para capturar detalhes das imagens em muitos níveis, mas tem uma vantagem em relação à Wavelets, de possuir a propriedade de ortogonalidade [IMDAD, 2007].

Para resolver problemas da identificação do autor e caracterizar a impressão da textura de forma global, os autores introduziram os filtros de Hermite. O autor mostra que os filtros de Krawtchouk de tamanho N são aproximados aos filtros de Hermite de expansão N/2 [IMDAD, 2007].

O primeiro estágio do método é o treinamento do classificador (SVM), em que as imagens são limiarizadas pelo método Otsu e após são extraídos os coeficientes de Hermite (para 4 escalas e 6 orientações). Durante os testes foi descoberto um classificador linear para ser altamente eficaz na detecção do autor, ainda que a exatidão possa ser estendida, utilizando um *kernel* adaptado para o SVM. Uma vantagem de utilizar o SVM é que pode ser utilizado um *kernel* não-linear que possa fornecer a separação de características selecionadas [IMDAD, 2007].

Assim [IMDAD, 2007] conclui em sua abordagem que com um número pequeno de autores, a exatidão foi de quase 100%, porém, se aumentar o número de autores, sem aumentar o número de características ou sem aumentar o processo de treinamento, os resultados decaem, mas não deixam de ser robustos. Imagens de alta resolução (maiores

que 256 x 256 utilizadas na abordagem) e um vetor com maior quantidade de características podem melhorar a exatidão [IMDAD, 2007]. No reconhecimento de autor, para 150 imagens de treinamento por autor, e 150 imagens para testes, com uma quantidade total de 30 autores, a taxa de exatidão é de 83%.

O algoritmo foi testado nas texturas de Brodatz (111 texturas) para verificar a robustez das características de Hermite. Os procedimentos foram semelhantes aos que foram feitos para a identificação do autor, e foi possível chegar a uma exatidão de 90%.

As características Dirigidas de Hermite, na maioria dos casos, fornecem resultados melhores que as de Gabor e Wavelet, e podem ser utilizadas para a maioria dos trabalhos de reconhecimento e de classificação na análise do original [IMDAD, 2007]. Esta mesma abordagem foi testada no reconhecimento de assinaturas e a exatidão foi comparável ao nível de exatidão do reconhecimento de autor por texto.

Assim como as abordagens de Brink [BRINK, 2007] e Imdad [IMDAD, 2007], uma técnica que envolve a angulação da escrita foi implementada por Baranoski [BARANOSKI, 2005] em que a inclinação axial é extraída a partir da borda da escrita. A técnica aplica os pré-tratamentos de binarização primeiramente, e após, dilatação e erosão para extração da borda. Sendo k o elemento estruturante e L a posição, o algoritmo utiliza a borda-direcional com $k=5$ e $L=17$, os quais apresentam resultados satisfatórios. Foram utilizados os manuscritos da base PUC-PR, que contém 945 cartas digitalizadas de 315 autores, sendo 3 manuscritos por autor. Foi utilizado o classificador SVM e o resultado foi de 90% de acerto.

A tabela a seguir (Tabela 1) apresenta um resumo dos métodos discutidos neste capítulo, com as principais informações como tipo de abordagem, base, extração de características, classificadores e resultados, das abordagens dos autores.

Referência	Tipo de Abordagem	Base	Características	Classificador	*Resultados*
[FRANKE et al., 2002]	Global	62 tipos de canetas e refis para produzir 737 amostras, destas 368 usadas para testes.	Matriz de co-ocorrência: Energia, Entropia, Correlação, Momento da diferença inversa, diferença e soma de contraste, diferença e soma de energia, diferença e soma da média, diferença e soma das divergências, diferença e soma das entropias.	K-NN SVM	*99.7% para 600 dpi* *98.4% para 300 dpi*
[BUSCH et al., 2005]	Global	10 manuscritos latinos, 4 chineses, 4 japoneses, 4 persas 3 hebreus, 3 gregos, 3 cirilicos, 3 sânscritos No total, 100 para treinamento e teste de cada tipo.	Matriz de co-ocorrência, filtros de Gabor, Wavelets	GMM Modelos de misturas gaussianas	*91 a 97% para Wavelets e filtros de Gabor*
[AL-DMOUR et al., 2007]	Global	20 autores escreveram duas cartas cada um, a primeira para treinamento e a segunda para teste.	Matriz de co-ocorrência, filtros de Gabor, as transformas de Fourier	SVM K-NN Distancia Euclidiana Análise Discriminante Linear	*Em torno de 90% para Análise Discriminante Linear com seleção de características.*
[SAID et al., 1998]	Global	25 segmentos por autor; conjunto A: 10 imagens por autor p/ treinamento e 15 para testes conjunto B:15 imagens para treinamento e 10 para testes	Pré-tratamento: inclinação angular Textura: Filtros de Gabor e matrizes de co-ocorrência (GLCMs): energia, entropia e contraste e correlação	Peso da distância Euclidiana (WED) e K-NN	*96% (conjunto B + WED + Filtros de Gabor)* *96% (GLCMs , com d=1 e d=2 + KNN)*
[SHEN et al., 2002]	Global	110 espécimes de 50 autores	Transformada de Gabor 2-D	K-NN	*97.6%*
[HE et al., 2005]	Global	20 manuscritos chineses de 10 autores.	Densidade Gaussiana Generalizada.	KLD	*80%*
[BRINK et al., 2007]	Global	50 amostras da base Firemaker 25 amostras da	Característica da dobradiça: orientação e curvatura do traço da escrita.	Distância Euclidiana e K-NN com k=1, k=10 e k=100	*Identificação com k=1: base Firemaker 67.3% e base NFI 53.5%*

		base NFI			*Verificação: Firemaker 96.1% e NFI 79%*
[IMDAD* et al*., 2007]	Global / Local	150 imagens de treinamento por autor, e 150 imagens para testes, sendo 30 autores diversificados da base IAM. Texturas de Brodatz (111 texturas)	Características de Hermite: relação com características orientadas, curvas e segmentos.	SVM	*83% de exatidão para base IAM.* *90% de exatidão sob as texturas de Brodatz*
[BARANOSKI et al., 2005]	Global	3 cartas por autor (315 autores) em um total de 945 cartas digitalizadas	Inclinação Axial	SVM	*90% de exatidão*

Tabela 1. Resumo do Estado da Arte

2.3 Visão Crítica

A visão crítica em relação ao estado da arte elaborado visa contribuir para a elaboração de uma abordagem mais consistente e que possua uma conotação prática. Porém a comparação dos resultados obtidos é de difícil análise, devido às abordagens descritas possuírem diferentes bases, características e classificadores selecionados e utilizados.

De acordo com as abordagens analisadas, pode-se observar que cada vez mais as características em relação à textura em imagens vêm sendo aplicadas nos métodos, incluindo a utilização da técnica da matriz de co-ocorrência que é a base para extração de características. Os resultados de classificação para as abordagens relacionadas à textura são promissores.

Por estes motivos, a análise da textura do traçado por meio da técnica da matriz de co-ocorrência e extração de características com base nesta, foi selecionada para ser desenvolvida no método proposto neste trabalho, que apresenta um método global de análise dos atributos genéticos e genéricos da escrita com base em imagens de textura utilizando uma abordagem de segmentação não-contextual. Isso propicia a análise de

manuscritos em que o teor do texto não se apresenta legível e, portanto, de difícil segmentação. O processo proposto não exige correções na inclinação das linhas de texto e dispensa os processos de correção dos alinhamentos entre palavras de uma mesma linha. Das imagens de textura geradas, são calculadas as matrizes de co-ocorrência, as quais são utilizadas em conjunto com os descritores de textura de Haralick [HARALICK, 1973]. Os descritores de Haralick foram selecionados devido às propriedades que os mesmos possuem no trato de textura onde a irregularidade dos padrões é predominante. Outra idéia realizada pela abordagem foi a aplicação da extração de característica de acordo com a metodologia proposta por [BARANOSKI, 2005], buscando melhores resultados.

Capítulo 3

Fundamentação Teórica

Este capítulo aborda os principais conceitos da ciência e campo em que o trabalho baseou-se. Também apresenta uma base teórica com os conceitos do objeto manuscrito, atributos grafoscópicos, reconhecimento de padrões, tipos de abordagens, extração de características, treinamento, classificadores, que se trata de elementos de um processo de verificação de autoria de documentos manuscritos.

3.1 Um histórico sobre a verificação da autoria de manuscritos

3.1.1 Ciência Forense

A Ciência Forense é um campo que possui uma grande área de atuação e inclui conhecimentos provenientes de diferentes áreas como a física, biometria, biologia, psicologia, geologia, química, matemática, engenharias, documentoscopia, grafoscopia, medicina, odontologia, balística, engenharias, informática, administração, toxologia, contabilidade, fonética. Estas áreas oferecem métodos e técnicas para que a ciência forense utilize em seu trabalho de perícia técnico-científica, permitindo analisar a ocorrência de fatos através de evidências ou provas descobertas. Os fatos colaboram para a tomada de decisão do juiz em um processo judicial.

Uma das áreas envolvidas na Ciência Forense é a documentoscopia, uma ciência que é responsável pelo estudo e análise de documentos, com o objetivo de identificar falsificações, irregularidades ou adulterações em documentos, sendo que o resultado da análise pode ser considerado como uma prova em processos criminalísticos.

Entende-se como documento qualquer objeto ou fato que serve como prova, confirmação ou testemunho. A classificação do objeto ou fato pode estar associada, entre

outras, ao material de suporte onde o mesmo foi aposto. Assim sendo, o registro dos fatos pode estar presente em papéis, fitas de áudio, fitas de vídeo, pinturas, quadros, fotos, discos magnéticos, discos óticos, entre outros, podendo ser também encontrado em um pequeno fragmento dos mesmos. Em aplicações forenses, a documentoscopia é normalmente utilizada para determinar os fatos relacionados a uma prova específica, anexa aos autos do processo.

A grafoscopia é um ramo da documentoscopia, que trata apenas aspectos de escrita e sua autoria. A escrita pode estar relacionada a vários fatores, como por exemplo, à autenticidade de autoria e determinação da contemporaneidade do manuscrito. O conceito do documento manuscrito é contextualizado a seguir.

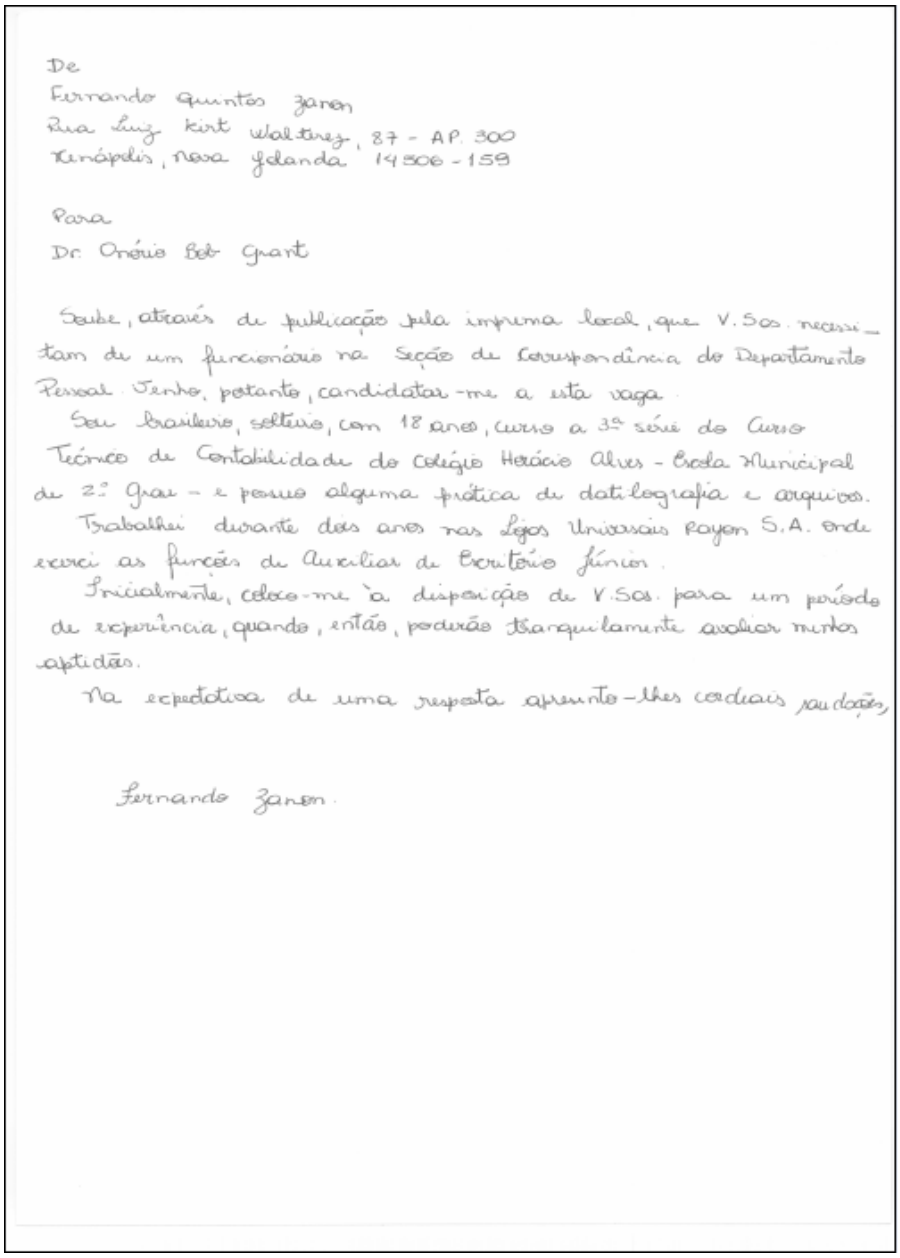
O ramo da grafoscopia desempenha um estudo de características individuais da grafia do ser humano com o objetivo de fornecer uma base para a verificação de autenticidade para as perícias.

Diversos métodos são propostos para a verificação e identificação da autoria, envolvendo a grafoscopia, porém cada uma com sua particularidade, ou seja, características que as diferenciam (como alto/baixo custo computacional; quantidade de autores para a base, formas de extração de características, classificações) e desta forma fornecem um estudo para que seja possível criar melhorias em relação a essas abordagens, ou então propor novas abordagens.

A descrição da grafoscopia associada aos manuscritos será apresentada em detalhes pois os conceitos constituem o desenvolvimento da abordagem deste trabalho.

3.1.2 Documentos Manuscritos

O manuscrito é uma habilidade adquirida, a qual é uma tarefa complexa de percepção motora às vezes relacionada à atividade neuromuscular (Figura 1). A habilidade de uma pessoa em manipular um objeto de escrita é precisamente coordenada pelo sistema nervoso que controla os movimentos do braço, da mão e dos dedos. A precisa ordem e tempo dos movimentos determinam a estrutura e o padrão que é reproduzido pelo objeto de escrita [HUBER & HEADRICK, 1999].



De
Fernando Quintos Zanon
Rua Luiz Kirt Walterez, 87 - AP. 300
Xinápolis, Nova Yolanda 14506-159

Para
Dr. Onório Bob Grant

Soube, através da publicação pela imprensa local, que V.Sas. necessitam de um funcionário na Seção de Correspondência do Departamento Pessoal. Tenho, portanto, candidatar-me a esta vaga.

Sou brasileiro, solteiro, com 18 anos, curso a 3ª série do Curso Técnico de Contabilidade do Colégio Heráclio Alves - Escola Municipal de 2º Grau - e possuo alguma prática de datilografia e arquivos.

Trabalhei durante dois anos nas Lojas Universais Rayon S.A. onde exerci as funções de Auxiliar de Escritório Júnior.

Inicialmente, coloco-me à disposição de V.Sas. para um período de experiência, quando, então, poderão tranquilamente avaliar minhas aptidões.

Na expectativa de uma resposta aguardo-lhes cordiais saudações,

Fernando Zanon.

Figura 1. Exemplo de um manuscrito, obtido por digitalização, da base de cartas PUC-PR.

No texto de um manuscrito de um autor, existem características a partir da forma como é criado, por uma sequência estrutural e de movimentos coordenados, sendo que estes movimentos acontecem em determinado tempo e sequência. A partir destes movimentos é possível obter um padrão, com características individuais de cada escritor.

Na área forense, um manuscrito genuíno ou *standart* é definido como um exemplar conhecido que pode ser comparado a um exemplar desconhecido. O exemplar genuíno possui normalmente uma quantidade suficiente de texto escrito para identificar características da individualidade do autor.

Na análise pericial, o perito deve confrontar a escrita dos exemplares genuínos com a escrita do documento questionado e com isto, produzir um laudo técnico, no qual o parecer técnico demonstre sua autenticidade ou discordâncias [JUSTINO, 2003].

Para a comparação da escrita manuscrita, o ideal é obter um exemplar original sob as mesmas condições com o qual o documento questionado foi produzido, possuindo a mesma quantidade de palavras números e símbolos, utilizando os mesmos recursos (papel e caneta), inclusive, deve ser produzido sem que o autor conheça o propósito do seu uso. Estes requisitos são relevantes e devem ser atendidos sempre que possível, porém, nem todos conseguem ser atendidos.

Para modelos de manuscritos, dois tipos são utilizados: os colhidos e coletados. Exemplares coletados são os documentos de escrita bem simples que foram indiscutivelmente preparados pelo escritor quando o mesmo não tinha razões para pensar que poderiam ser usados em uma demanda judicial, sendo, portanto, livres da tentativa de disfarce. Exemplares colhidos são aqueles nos quais o indivíduo é intimado a reproduzir um material escrito específico.

Um exemplar coletado tem como desvantagem a dificuldade de encontrar espécimes que reproduzem o formato e o texto do documento questionado, porém possui a vantagem de eliminar a possibilidade de disfarce.

Um exemplar colhido possui uma vantagem em relação ao coletado, por ser produzido de acordo com as orientações do perito com o mesmo formato e conteúdo do documento questionado. Porem possui a desvantagem do autor conhecer a finalidade do documento a ser produzido, que pode ser usado contra seus interesses.

Apesar de não possuir o mesmo conteúdo dos documentos questionado, diversos modelos de coleta são usados em vários países e possuem muitas associações de palavras, letras e símbolos encontradas em cartas comuns, sendo adaptados aos padrões de grafia do idioma usado. Na língua inglesa são citados como exemplos os modelos: “Carta da Classe 16”, “Carta do Egito”, “Carta de Londres” [BARANOSKI, 2005].

3.2 A verificação de autoria de manuscritos para fins forenses

A escrita se desenvolve em um indivíduo a partir da cultura e local que sofre mudanças de acordo com os sistemas e características de uma nação. Com a prática e fato que, com prática e habilidade, a execução da escrita torna-se mais automático, levando o processo da escrita ser menos sujeita ao controle consciente. [HUBER & HEADRICK, 1999].

Segundo Huber [HUBER & HEADRICK, 1999], a verificação de autoria de um manuscrito é baseada em duas premissas: o hábito e individualidade ou heterogeneidade da escrita.

As pessoas desenvolvem a escrita a partir de hábitos, como a formação de letras, palavras, sentenças, e a ação de colocar em prática estes hábitos dependem do processo de pensamento do indivíduo. O hábito não é instintivo, tampouco hereditário, mas sim um processo complexo de aprendizado que é gradualmente desenvolvido.

Na comparação de manuscritos para verificação de autoria, letras, combinações de letras, palavras, ou sentenças devem ser consideradas de acordo com o grau em que constituem um hábito coletivo, de modo que o todo representa mais do que suas partes. Assim, a influência de letras conjuntas sobre outra irá variar segundo o papel que estas letras representam nas palavras ou frases as quais se tornaram um hábito de escrita, mais que como letras isoladas. A variação no formato e movimento pode ser esperada como uma alteração relacionada a este fator.

Para a premissa da individualidade ou heterogeneidade, o manuscrito é único e individual, e todo perito deve se basear nisto. Anos atrás, este argumento era difundido, porém simplesmente, baseado na crença “a natureza nunca nos oferece seu trabalho em cópias”. Além disto, as pessoas eram comparadas a folhas ou pedras, não foi encontrado nem dois de cada espécie exatamente iguais. Isaac D’Israeli é citado, há mais de um século e meio, “Para cada indivíduo, a natureza deu um diferente tipo de escrita, como se tivesse dado forma, voz e “gestos característicos”. Estes gestos são definidos, no caso da escrita, por “gênese da escrita”, a qual é composta de atributos genéticos como o espaçamento, o calibre, momentos, proporcionalidades, entre outros, detalhados neste trabalho posteriormente.

É axiomático que cada dois itens na natureza devem ser distintos, desde que a escala de julgamento tenha um nível suficiente de precisão. Mas enquanto não existe nada como identidade verdadeira, a questão real para peritos em manuscrito é se, na diferenciação da escrita, o julgamento do perito e suas ferramentas são capazes ou não de detectar tal precisão a fim de que consiga fazer a distinção necessária. Não é suficiente, e pouco científico, discutir isto porque alguns manuscritos são obviamente diferentes, de forma alguma dois textos de escritores diferentes podem ser tão coincidentemente similares que sejam equivocadamente avaliadas como produção de uma mesma pessoa.

3.2.1 Dificuldades a serem enfrentadas

Atualmente, várias abordagens foram e estão sendo estudadas para auxiliar a verificação e identificação de autoria para fins forenses, baseada em atributos grafoscópicos, em busca de resultados cada vez melhores [SAID, 1998],[FRANKE, 2002],[BUSCH, 2005],[AL-DMOUR, 2007],[SHEN, 2002],[HE, 2005],[BARANOSKI, 2005],[BRINK, 2007],[IMDAD, 2007]. Porém, vários problemas são encontrados no desenvolvimento das técnicas, alguns deles, relevantes, são citados a seguir.

Em relação à escrita de manuscritos, muitos acidentes podem aparecer. Estes acidentes são divergências isoladas, breves, ou temporárias da prática normal da escrita. Há ocorrências na escrita que tem pequena ou nenhuma explicação plausível. Estas são formas incomuns, formato ou movimento, quebras na linha, até a duplicação de letras ou parte de letras. Elas são mais freqüentemente menores em natureza, sem freqüência e de importância insuficiente para que o escritor dê atenção ou se importe em corrigir.

Raramente acidentes são percebidos ou observados nos padrões da escrita. É mais uma designação ou rótulo dado a um elemento de um texto questionado que destoa significativamente da escrita normal e natural observada em um padrão, e pelo qual não há explicação aceitável. É simplesmente uma qualificação para assinalar que é diferente [HUBER & HEADRICK, 1999].

Várias condições e circunstâncias contribuem para a natureza da escrita comum e a qualidade do desempenho da escrita. Alguns destes fatores são variáveis e além de nosso controle voluntário. Eles pertencem à natureza do escritor e incluem aqueles fatores que são físicos e outros são mentais. Idade e a falta de firmeza são influências externas

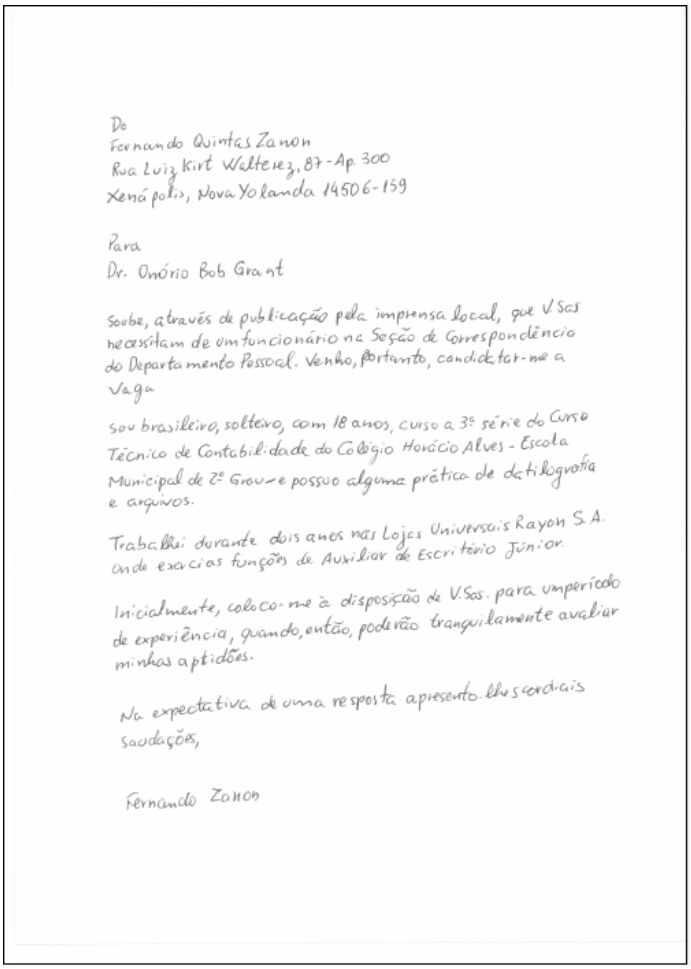
invariáveis. Sinistralidade (canhoto) é outro fator freqüentemente encontrado. Aderência a um sistema próprio de escrita ou habilidade com instrumentos de escrita são dois outros. Outras variantes são intrínsecas e de algum modo circunstanciais, ou seja, são fatores sobre os quais se pode exercer algum controle, quando desejado, como, por exemplo, a imitação da ascendência e outras práticas, ou a marca do instrumento de escrita. Juntamente com estes fatores intrínsecos está um conjunto de condições temporárias, derivadas de alucinógenos, álcool, hipnose, estresse e fadiga, que exercem influência sobre a escrita, a despeito da forma que de outro modo toma, sendo estas consideradas como influências as quais se submete voluntariamente [HUBER & HEADRICK, 1999]. Para este trabalho, os acidentes da escrita em um documento não são considerados, pois os textos utilizados da base PUC-PR são de origem natural, isentos de qualquer tipo de falsificação ou dissimulação, tais como tentativas de disfarce e outros.

Já para o documento manuscrito, muitas circunstâncias ou condições que afetam as conclusões que podem ser tiradas, quando há um exame de manuscrito. Estas incluem:

- Insuficiência qualitativa (falta de significado) de hábitos presentes no material questionado, ex., o predomínio de letras que apresentem menos oportunidade de individualização;
- Grande variação nos padrões de um texto para outro;
- Insuficiência quantitativa de hábitos que o material questionado contém;
- Escrita precária e degeneração no formato das letras;
- Falta de credibilidade da reprodução, como registro dos hábitos de escrita, e do caráter do documento original (ex., consistência da tinta e do papel, a seqüência de paradas), quando o exame dos originais não é possível;
- Distorção deliberada ou disfarce do documento questionado ou dos padrões de escrita;
- Condição anômala do escritor ou das circunstâncias para escrita do documento questionado.

Um problema frequente encontrado nos manuscritos é os espaços em branco de um documento manuscrito, que dificultam os resultados para a verificação e identificação da autoria, quando a técnica se trata de técnicas de extração de características baseado somente no conteúdo da escrita.

Um exemplo da causa de espaços em branco em um documento é as expansões laterais. Huber [HUBER & HEADRICK, 1999] descreve que nos manuscritos, a expansão lateral é determinada pela dimensão horizontal de um grupo sucessivo de letras e palavras, ou seja, é um produto da formação da letra, tamanho das letras, e espaçamento entre letras e palavras que varia de contraído a expandido. Apesar do formato e tamanho das letras contribuírem para a expansão lateral, o contribuinte principal tende a ser o espaçamento entre letras e palavras. O espaçamento é um aspecto da escrita que é freqüentemente distintivo e bastante consistente em alguns indivíduos (Figura 2 e Figura 3).



De
Fernando Quintas Zanon
Rua Luiz Kirt Walterez, 87 - Ap. 300
Xenápolis, Nova Yolanda 14506-159

Para
Dr. Onório Bob Grant

Soube, através de publicação pela imprensa local, que V.Sas necessitam de um funcionário na Seção de Correspondência do Departamento Pessoal. Venho, portanto, candidatar-me a vaga

sou brasileiro, solteiro, com 18 anos, curso a 3ª série do Curso Técnico de Contabilidade do Colégio Horácio Alves - Escola Municipal de 2º Grau - e possuo alguma prática de datilografia e arquivos.

Trabalhei durante dois anos nas Lojas Universais Rayon S.A. onde exercias funções de Auxiliar de Escritório Júnior.

Inicialmente, coloco-me à disposição de V.Sas. para um período de experiência, quando, então, poderão tranquilamente avaliar minhas aptidões.

Na expectativa de uma resposta apresento-lhes cordiais saudações,

Fernando Zanon

Figura 2. Imagem de um manuscrito que possui grande espaçamento lateral de margem e entre parágrafos.

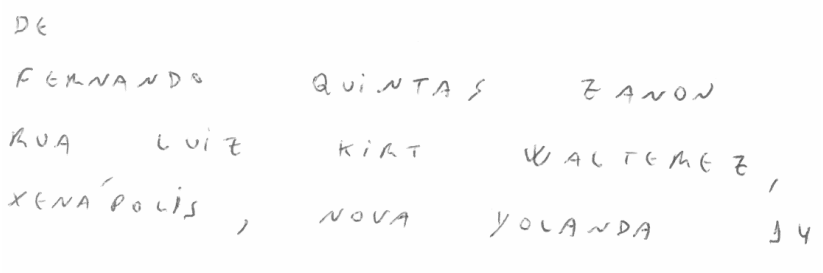


Figura 3. Exemplo de fragmento de um manuscrito digitalizado que possui grande espaçamento entre palavras.

Conseqüentemente, a expansão lateral não é uma característica da escrita, por si, e pode dificultar o julgamento com propósitos comparativos, exceto em casos especiais. Com o objetivo de extrair características da textura do traçado, a partir das imagens originais digitalizadas dos manuscritos da base PUC-PR, outras imagens foram criadas através de um algoritmo que extrai apenas o conteúdo da escrita, obtendo-se a escrita concentrada. Este algoritmo aplicado na base original PUC-PR foi realizado para reduzir os espaços em branco do manuscrito. Com a remoção da parte sem conteúdo e compactação da escrita para a formação da textura, os resultados na verificação da autoria foram melhores.

3.2.2 Grafoscopia

A Grafoscopia tradicional foi concebida com o objetivo de esclarecer questões criminais. Tratando-se de um campo da criminalística, ela tem sido conceituada como a área cuja finalidade é a verificação da autenticidade da autoria de um documento a partir de características gráficas utilizadas na elaboração de um documento [JUSTINO, 2001].

Como a escrita está sujeita à inúmeras mudanças, decorrentes de causas variadas, ela exige conveniente interpretação técnica para o completo êxito dos exames grafoscópicos periciais [JUSTINO, 2001]. Para a correta análise do perito grafotécnico, tanto para a identificação quanto para a autenticação de autoria, existe a necessidade de entender os princípios básicos do processo de aprendizado da escrita.

Nos primeiros anos do processo de aprendizado da escrita o indivíduo não possui estilo ou escrita própria, mas sim, apenas uma reprodução do modelo treinado. Com o passar do tempo, após o modelo memorizado, o indivíduo passa a introduzir

variabilidades ou desvios do modelo inicial, sendo esse o processo de desenvolvimento da sua própria escrita ou estilo [JUSTINO, 2001].

Os desvios do modelo aprendido são alguns elementos que o autor introduz em sua escrita, tais como embelezamento, escrita mais veloz e pequenos cortes; a imagem mental e a habilidade de lembrar o modelo inicial são gradativamente substituídos pelo modelo pessoal [JUSTINO, 2001].

Outro aspecto importante que também está presente na escrita do autor são as classes de características: semelhanças de grafia apresentadas por indivíduos ou grupos de indivíduos que foram ensinados através de sistemas de aprendizado iguais ou semelhantes. Estas classes podem ajudar na redução da procura, num universo finito de autores, quando se compara um autor questionado com os padrões de vários autores diferentes [JUSTINO, 2001].

A grafoscopia busca a padronização e auxiliar os procedimentos da perícia para a identificação e verificação da autoria de textos manuscritos, através de técnicas grafométricas que envolvem um conjunto de atributos grafoscópicos. Neste conjunto duas classes se destacam: atributos genéricos e genéticos [GOBINEAU, 1954].

As características individuais da escrita de um autor são definidas como determinados elementos que servem para diferenciação e verificação ou identificação de autores entre membros de algum ou todos os grupos.

Na análise grafotécnica da escrita, são destacados importantes elementos da grafia, conforme [BARANOSKI, 2005] descreve: o espaço bidimensional onde a escrita é feita, nomeado campo gráfico; movimento gráfico que indica o movimento realizado pelos dedos do escritor, formando um traço gráfico; um traço é um trajeto que realizado pela escrita, em um único gesto; traço descendente, fundamental, pleno ou grosso; traço ascendente ou perfil sendo um traço fino; elementos em formas de círculo, como por exemplo, as letras “a, o, q, q” denominadas ovais; os traços plenos (movimento de descanso) que são as hastes, encontradas em letras como “l, t, b, f”e também nos traços verticais das letras “m, n”; laçadas inferiores, conhecidos como traços plenos (descendentes) de letras como “g, j, y, f”; bucles, que representa os traços ascendentes (perfis) das hastes das laçadas inferiores e todo movimento que ascende cruzando a haste e unindo-se a ela formando círculo; partes essencias que correspondem ao esqueleto da letra e parte secundária ou acessória que corresponde ao revestimento ornamental.

Também são determinadas três zonas nas letras: zona inicial, que indica a área em que existe o ponto no qual se inicia a letra; zona final, que indica a área em que está o ponto que termina a letra; zona superior que corresponde a área em que está o ponto mais alto ocupado pelas hastes, pelos pontos e acentos, pelas barras da letra “t”e parte das letras maiúsculas; zona média que é a área central ocupada por todas as vogais minúsculas “a, e, i, o, u” e pelas letras “m, n, r” e zona inferior que representa a zona baixa da escrita a partir da base de todos os ovais descendentes, das letras maiúsculas ou de todas as letras.
Normalmente, características individuais são percebidas como aspectos ou particularidades da escrita que são peculiares a um escritor específico. Neste este sentido, eles possuem um caractere que é freqüentemente encontrado. Existe, contudo, um grande número de elementos da escrita comumente encontrados que podem ser descritos como desenhos, invenções e desenvolvimento do escritor que, quando considerados em combinação como um grupo, da a escrita exclusividade. Sendo assim, é a composição da combinação a responsável pela individualidade da escrita [HUBER & HEADRICK, 1999].

Individualidade é provavelmente mais freqüentemente exibida por escritores na execução de textos mais complexos. Algumas letras do nosso alfabeto exigem movimentos complexos os quais muitos escritores acham difícil de executar. Como resultado a cópia, ou modelos de texto, são normalmente um pouco alterados pelo indivíduo para uma estrutura ou formato mais conveniente para reproduzir. Estas modificações, as vezes sutis, as vezes profundas, são as características individuais de cada escritor [HUBER & HEADRICK, 1999].

Justino [JUSTINO, 2002] realizou um estudo sobre as características particulares do autor que são utilizadas na análise pela grafoscopia e podem ser indicados como elementos discriminantes em um manuscrito:

- Forma caligráfica: é a representação pictórica da escrita, podendo ser cursiva, de caixa alta ou tipográfica, ou mista;
- Nível de habilidade: autores com alto nível de habilidade produzem textos rítmicos bem traçados, artisticamente embelezados; autores com baixo nível de habilidade produzem textos com escrita vacilante, traçada lentamente;

- Inclinação axial: é o ângulo de inclinação da escrita em relação ao eixo vertical de um sistema de eixos cartesianos, em que o eixo horizontal é representado por uma linha de base imaginária; a inclinação pode ocorrer à esquerda, à direita ou ser nula. Alguns autores possuem inclinação mista;
- Movimento: é a direção do movimento dos instrumentos de escritura, como lápis ou a caneta;
- Proporções: referem-se às simetrias das letras individualmente;
- Relações de altura: é a comparação ou correlação da altura de uma letra ou segmento de letra em relação à outra letra, normalmente dentro da mesma palavra;
- Mínimos gráficos: são pequenas proporções de escritura como pontos finais, vírgulas, acentos gráficos e cedilhas.
- Corte da letra “t”: o corte da letra pode estar alinhado na horizontal, apresentar inclinações, apresentar elevação do traço à direita ou à esquerda, ou estar conectado a um golpe de saída de uma letra terminal de uma palavra;
- Laçadas: é um traçado que apresenta um movimento de retorno para o ponto de partida e ocorre geralmente em letras cursivas possuindo elementos ascendentes e descendentes, formas pontiaguda ou arredondadas, simétricas ou assimétricas;
- Pressão: representa a variabilidade da largura do traçado e acúmulo do material em uma determinada região do traço;
- Alinhamento em relação à linha de base: capacidade do autor de produzir linhas de textos alinhadas com uma linha guia horizontal imaginária em papel não pautado ou linha com papel pautado;
- Embelezamento: localiza-se usualmente no começo de uma letra podendo estar presente ao longo do manuscrito;
- Retraço: é o processo no qual o objeto da escrita repinta uma porção escrita da linha, normalmente em direção oposta, com um movimento descendente seguido por um movimento ascendente sobre a linha existente;

- Erros de ortografia e espaçamento: a ortografia incorreta das palavras pode ser um indicativo de uma característica individual do autor, o espaçamento é formado pela interrupção do curso da escrita entre combinação de letras especificas;
- Formato: são elementos gráficos e abreviações em um documento;
- Entradas e golpes de saída do traçado: podem ser movimentos habituais e podem repetir-se em formações de letras semelhantes como nas letras “U, V, M, N”.

Porque a escrita deve ser lida, alguma conformidade com o desenho do caderno deve ser estabelecida a qual limita a extensão da divergência aceitável das formas. O resultado é que traços individuais similares podem ser encontrados na escrita de outras pessoas, embora um grupo em particular de tais traços não é provável de serem duplicados. É este grupo ou a combinação de tais traços, mais do que um elemento específico, que serve, normalmente, para distinguir um escritor de outro.

Elementos de execução consistem em abreviações, alinhamento, começos e terminações, diacríticas e pontuação, embelezamento, continuidade, qualidade ou fluência (velocidade), controle da pena (que inclui o modo de segurar, a posição e a pressão) movimento da escrita (incluindo o ângulo), e legibilidade ou qualidade da escrita (a qual inclui formato da letra ou formato das letras para qualquer texto dado); consistência ou variação natural e persistência; expansão lateral e proporção das palavras.

3.2.3 Atributos genéricos e genéticos da Grafoscopia

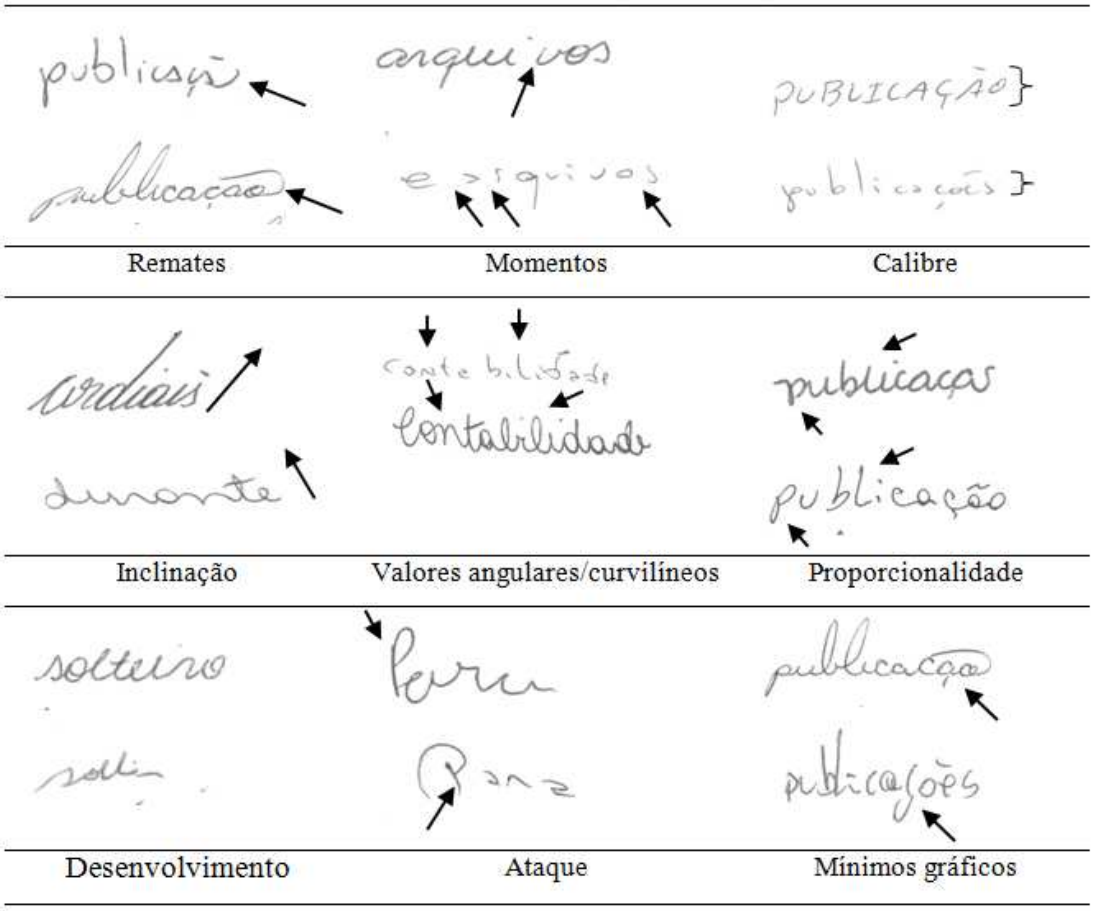
Quando os autores tratam das características grafoscópicas, buscam agrupá-las em uma classificação, e nesse contexto duas classes se destacam, os chamados atributos genéticos e os atributos genéricos. Os atributos genéricos possuem elementos relacionados com a forma geral do texto [GOMIDE & GOMIDE, 1995]. Os atributos genéticos traduzem a gênese da escrita do autor, apresentando aspectos do traçado elaborados de forma inconsciente. Os mesmos são de difícil ocultação pelo escritor e de difícil reprodução por terceiros. Cada classe de atributos pode então ser enquadrada em um exame pericial, como convergência ou divergência mínima, média ou máxima.

Na Tabela 2.1 é possível ver ambas as classes e seus atributos relacionados. Na Tabela 2.2 são apresentados exemplos de atributos genéricos e genéticos.

Tabela 2.1 Atributos genéricos e genéticos.

Atributos Genéticos	**Atributos Genéricos**
Pressão	Calibre
Progressão	Espaçamento
Ataques	Comportamento pauta
Remates	Comportamento base
Desenvolvimento	Valores angulares
Inclinação Axial	Valores curvilíneos
Mínimos Gráficos	Momentos
	Proporcionalidade

Tabela 2.1 Exemplos de atributos genéricos e genéticos.



- O calibre determina a escala geral da escrita, tal como o tamanho das letras;
- O espaçamento determina o espaço médio entre palavras;
- O comportamento pauta determina a distância média do texto escrito, em relação à linha base ou pauta;
- O comportamento base determina o ângulo médio de inclinação das linhas de texto, para cima ou para baixo, de uma linha de base imaginária. O comportamento base é enfatizado quando o texto é redigido em papel sem pauta;
- Os valores angulares e curvilíneos estabelecem o comportamento angular nos pontos de mudança de direção do traçado, tais como nas letras “L”, “M”, “G”, entre outras. Nesses casos, o traçado pode apresentar uma curvatura suave ou uma mudança brusca de direção;
- Os momentos determinam os pontos de interrupção do traçado, no decorrer da escrita de uma palavra. Estes estão presentes, usualmente, na conexão entre letras;
- A proporcionalidade determina o grau de variabilidade entre ascendente, descendentes e corpo das letras;
- A pressão determina as variabilidades da força que o objeto de escrita exerce sobre o papel, durante a evolução do traçado;
- A progressão determina a velocidade imposta pelo escritor, ao objeto de escrita, durante o transcorres da escrita;
- O ataque determina a forma do traçado no início de uma palavra;
- O remate determina a forma do traço na finalização de uma palavra;
- O desenvolvimento determina o grau de destreza do escritor na execução do traçado;
- A inclinação axial determina o ângulo médio de inclinação da escrita;
- Os mínimos gráficos determinam as formas e angulações das acentuações e pontuações presentes no texto manuscrito.

Do ponto de vista de um perito, o resultado da análise é decorrência de evidenciação de um conjunto mínimo de atributos grafoscópicos e dos respectivos graus de convergência, encontrados nos manuscritos de referência (de autoria conhecida) e no questionado (de autoria desconhecida). No que se refere ao grau de confiabilidade dos

atirbutos, tanto os genéticos como os genéricos são importantes na tomada de decisão durante o exame pericial.

A análise dos atributos genéticos e genéricos é feita tendo como base a dinâmica do traçado da escrita. Esta dinâmica representa um conjunto de fenômenos gráficos, usualmente produzidos de forma inconsciente pelo escritor, muitas vezes denominados de "gestos característicos" [HUBER & HEADRICK, 1999]. Os atributos genéticos e genéricos podem ser considerados complementares.

Do ponto de vista computacional, o processo de segmentação e de extração das classes de características genéticas e genéricas é de elevada complexidade, tendo em vista restrições impostas pela segmentação do texto, eliminando espaços em branco e, por conseguinte, da extração das características. No entanto, a adoção de um método que aproveite as propriedades inerentes das duas classes de características, tanto no processo de segmentação como no de extração das características, podem favorecer a criação de um modelo computacional robusto e que atenda, de forma satisfatória, às expectativas requeridas em uma análise pericial.

3.3 Reconhecimento de Padrões

Um padrão é uma descrição de um objeto que pode ser um conjunto de medidas ou observações normalmente representadas através de um vetor ou notação de matriz. A verificação de autoria em manuscritos pode ser enquadrado neste universo, onde o manuscrito é um exemplo de padrão que pode ser representado por uma matriz de *pixels*.

O conceito de reconhecimento de padrões envolve a categorização de dados de entrada, dentro de classes identificáveis por meio de extração de características significantes ou atributos com detalhes relevantes. Um sistema de reconhecimento de padrões engloba etapas de: aquisição do sinal por meio de um sensor que elimina detalhes de pouca importância no reconhecimento de padrões; pré-processamento; extração de características sendo estas medidas de um padrão que podem contribuir para a etapa de classificação que por sua vez, associa dados de entrada de uma ou mais classes pré-definidas. Na etapa de verificação de autoria de manuscritos, através das características extraídas, é estabelecer uma regra de decisão através da comparação realizada com o

modelo de referência que descreve uma representação análoga obtido em fase anterior denominada treinamento. Cada etapa será descrita especificamente a seguir.

Da escolha do tipo de representação (os tipos primitivas) constitui uma etapa essencial na elaboração de um método de verificação. As dificuldades surgem principalmente da maneira com a qual são tratadas as entidades naturais usadas para obter a descrição matemática, induzida por um método teórico formal. Essa indução possui dois reflexos, sendo dois métodos: estruturais e estatísticos.

Os métodos estruturais buscam descrever informações geométricas de maneira estrutural, representando formas complexas a partir de componentes elementares, chamadas primitivas. Existem dois tipos de métodos estruturais: os propriamente ditos, nos quais a estrutura utilizada é um grafo que permite representar formas, as primitivas e as relações entre elas; e os métodos sintáticos, nos quais a estrutura é utilizada para codificar a forma em uma lista, utilizando um alfabeto cujos termos representam elementos da forma a descrever [JUSTINO, 2001].

Os métodos estatísticos consistem em efetuar as medições do espaço métrico através da estatística, sendo que o aprendizado é executado através da separação de um conjunto de características comuns. São importantes nos sistemas cujas classes possuem uma elevada instabilidade entre vários espécimes: os paramétricos que trabalham com hipóteses de que as classes em questão possuem uma distribuição de probabilidade com comportamento determinado, e os não paramétricos, que assumem que leis de formação da probabilidade de uma classe são desconhecidas.

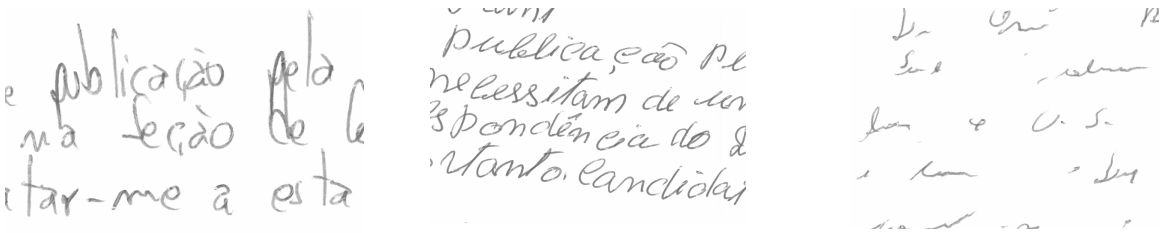
3.3.1 Tipos de Abordagens

As abordagens relacionadas à verificação automática ou semi-automática de manuscritos estão diretamente relacionadas com o método de aquisição de dados. Se o processo de aquisição e verificação ocorre ao mesmo tempo em que o autor escreve, o método é dito *on-line* ou dinâmico, neste caso havendo a necessidade de um dispositivo de acesso especial quando o manuscrito é produzido. O método *off-line* ou estático caracteriza-se pela aquisição da informação, provavelmente de uma folha de papel, feita por um digitalizador ou câmera para posterior análise da imagem.

Em relação ao processo de extração de características da escrita, é possível encontrar dois tipos de abordagens computacionais básicas: as locais e as globais [SCHOMAKER, 2007].

As locais geralmente fazem uso de processos de segmentação contextual [SRIHARI, 2002][BUSCH, 2005][AL-DMOUR, 2007], segmentando o texto em linhas, palavras, letras e segmentos de letras ou traços. O produto gerado pelo processo de segmentação é posteriormente submetido a diferentes processos de extração de características, as quais podem ou não estar associadas às classes de atributos grafoscópicos vistos anteriormente [CHA, 2001][PERVOUCHINE, 2007].

As abordagens que fazem uso de processo de segmentação contextual apresentam elevada complexidade e, podem não atender a todos os escritores. Existem casos em que o conteúdo de linhas adjacentes se sobrepõe (comportamento pauta - Figura 4a), a inclinação das mesmas é elevada e ocorre de forma irregular (comportamento base - Figura 4b), e o texto é ilegível, impedido a identificação dos pontos de conectividade entre letras (momentos) e os espaçamentos entre palavras (espaçamentos – Figura 4c). Em decorrência disso, a solução usualmente adotada nesses casos é a segmentação manual ou semi-automática. As abordagens globais utilizam segmentações não-contextuais, isto é, sem levar em consideração o teor do texto manuscrito, em que fragmentos ou subáreas da imagem do manuscrito são submetidos ao processo de extração de característica [SAID, 1999][SRIHARI, 2002]. Uma terceira abordagem, denominada mista, também pode ser encontrada sendo esta uma combinação das duas anteriores [BULACU, 1993].



(a) (b) (c)

Figura 4. (a) Texto com conteúdo conexo; (b) Texto com alinhamento irregular; (c) Texto ilegível.

3.3.2 Extração de Características

Os tipos de características extraídas a partir da escrita do autor estão diretamente ligadas aos tipos de abordagens: globais e locais. Características globais são características extraídas de segmentos do manuscrito, como por exemplo, parágrafos, pedaços da imagem. Já as características locais são extraídas a partir de letras e palavras segmentadas do documento manuscrito.

Este trabalho envolve um estudo relacionado à obtenção de características globais a partir da textura do traçado da escrita. Dentre as abordagens globais, a produção de texturas a partir do texto manuscrito, vem se mostrando uma alternativa promissora na extração de características.

A remoção da predominância dos espaços em branco existentes entre linhas e palavras é o procedimento mais utilizado nesses casos. Em outras palavras, técnicas de segmentação contextuais são utilizadas para produzir imagens que serão analisadas de maneira global. Com a compactação do texto e eliminação dos espaços em branco as características de textura trouxeram melhores resultados.

Neste trabalho, também foi aplicada a extração da característica de inclinação axial proposta na técnica da Baranoski [BARANOSKI, 2005]. A aplicação foi realizada tanto para base PUC-PR original de manuscritos quanto para a base com a escrita compactada após a segmentação para remoção dos espaços em branco, para comparação dos resultados.

3.3.2.1 Características relacionadas à Textura do traçado

A textura de uma imagem pode ser definida como um conjunto de variações de intensidade que de certa forma se repetem formando um padrão. Os padrões são o resultado das propriedades físicas como o contraste, a direcionalidade, a regularidade, linearidade.

Há três tipos de abordagens para descrever a textura: estatística, estrutural e espectral.

A técnica estatística caracteriza a textura utilizando as propriedades estatísticas dos níveis de cinza dos *pixels* abrangendo a área da imagem; as estatísticas de primeira

ordem são usadas para calcular a probabilidade de observar um valor de cinza em um local da imagem escolhido aleatoriamente. Pode-se calcular as estatísticas de primeira ordem usando um histograma das intensidades dos *pixels*. Uma abordagem freqüentemente usada para analisar a textura é baseada nas propriedades estatísticas da intensidade do histograma. Estatísticas de segunda ordem são definidas como a probabilidade de observar um par de valores de cinza nos extremos de um local dipolo na imagem, em uma localização e orientação aleatória. Estas são as propriedades de um par de *pixels*. Normalmente, estas propriedades são calculadas utilizando os níveis de cinza da matriz de co-ocorrência da imagem.

Uma técnica estrutural caracteriza a textura como sendo um composto de simples estruturas primitivas chamadas "*texels*" ou elementos da textura. Estes elementos são organizados regularmente sobre uma área, de acordo com algumas regras de combinação da área.

As técnicas espectrais são baseadas nos espectros de Fourier e descreve a periodicidade global dos níveis de cinza na área da imagem identificando os picos de alta energia no espectro de Fourier.

Com base no estudo feito para o estado da arte deste trabalho, três técnicas bastante utilizadas em relação à textura serão descritas a seguir. A matriz de co-ocorrência é uma técnica relacionada aos níveis de cinza de uma imagem, e esta foi selecionada para ser aplicada na textura do traçado da escrita para obtenção das características utilizadas na abordagem que se referem às fórmulas dos descritores de Haralick [HARALICK, 1973], devido sua fácil forma de implementação e baixo custo computacional.

3.3.2.1.1 Filtros de Gabor

A técnica multi-canal de Gabor é inspirada nas descobertas psicofísicas em que o processamento da informação pictorial no córtex da visão humana envolve um conjunto de mecanismos paralelos e quase independentes ou canais corticais que podem ser modelados por filtros *bandpass* (filtros passa-faixas).

Resumidamente, cada canal cortical é modelado por um par de filtros de Gabor $h_e(x, y; f, \theta)$ e $h_o(x, y; f, \theta)$. Os dois filtros de Gabor são de simetrias opostas e são dados por:

$$h_e(x, y; f, \theta) = g(x, y)\cos(2\pi f(x\cos\theta + y\sin\theta)) \quad (1)$$

$$h_o(x, y; f, \theta) = g(x, y)\sin(2\pi f(x\cos\theta + y\sin\theta)) \quad (2)$$

em que *g(x, y)* é uma função Gaussiana 2-D e f e θ são a frequência e a orientação radial que define o local do canal e a frequencia no plano. As frequências mais comuns utilizadas são de frequência 2. Para qualquer imagem de tamanho *N x N* os componentes de frequência importantes podem ser encontrados por $f \leq \frac{N}{4}$ ciclos/grau. Por estas razões foram utilizadas frequências de 4, 8, 16 e 32 ciclos na abordagem de Said [SAID, 1998]. Para cada frequência central *f*, a filtragem é realizada em θ=0°, 45°, 90° e 135°. Com isto se obtém 16 imagens de saída, 4 para cada frequência, a partir das quais as características são extraídas. Estas características são a média e o desvio-padrão de cada imagem de saída. Assim são calculadas 32 características para cada imagem de entrada.

Os Filtros de Gabor é uma das técnicas mais populares e reconhecidas [SAID, 1998], porém possui como desvantagem o alto custo computacional. Esta técnica foi aplicada em recentes abordagens [BUSCH, 2005] [SAID, 1998][AL-DMOUR, 2007].

3.3.2.1.2 Matriz de co-ocorrência (GLCMs) e Descritores de Haralick

As matrizes de co-ocorrência de níveis de cinza (GLCMs – *Gray-Level Co-occurrence matriz*) são utilizadas para representar uma estatística de um conjunto de pares de *pixels* de uma imagem e tem sido utilizada por muitos anos como um meio de caracterizar a textura e classificar imagens [BUSCH, 2005].

As matrizes de co-ocorrência são poderosas ferramentas na classificação de imagens. A primitiva mais simples que pode ser definida em uma imagem digital em níveis de cinza é um *pixel*, que tem como propriedade seu nível de cinza. Conseqüentemente, a distribuição dos níveis de cinza dos *pixels* pode ser descrita por estatísticas de primeira ordem, como média, variância, desvio padrão, inclinação *skewness* ou estatísticas de segunda ordem como a probabilidade de dois *pixels* terem um determinado nível de cinza ocorrendo com um relacionamento espacial particular. Essa

informação pode ser resumida em matrizes de co-ocorrência bidimensionais, calculadas para diferentes distâncias e orientações [ROCHA & LEITE, 2007].

A técnica para cálculo da matriz de co-ocorrência vem sido utilizada em abordagens recentes [FRANKE, 2002], [SAID, 1998], [AL-DMOUR, 2007], [BUSCH, 2005] e a partir das matrizes são utilizadas características associadas a elas, para extração. Devido a estes estudos, a abordagem deste trabalho, envolve a utilização das matrizes na análise do traçado da escrita, inclusive as características referentes aos descritores de Haralick.

A matriz de co-ocorrência é uma tabulação de quantas combinações diferentes de valores de intensidade dos *pixels* (níveis de cinza) ocorrem em uma imagem. A idéia principal da matriz de co-ocorrência é descrever a textura através de um conjunto de características para as ocorrências de cada nível de cinza nos *pixels* da imagem considerando múltiplas direções [ROCHA & LEITE, 2007].

Em 1979, Haralick propôs uma metodologia para descrição de texturas com base em estatística de segunda ordem, em que são definidas características provenientes do calculo de matrizes, denominadas matrizes de co-ocorrência, que consistem em uma tabulação da contagem de quantas combinações diferentes de valores de intensidade dos *pixels* ou níveis de cinza ocorrem em uma imagem, em uma determinada direção (Figura 5) [ALVES, 2006]. Para cada *pixel* P(i,j) processado a partir de uma imagem, há uma janela em torno deste *pixel*, com distância d=1 nas quatro direções angulares.

Na prática, as correlações mais relevantes ocorrem em distâncias curtas e, assim, os valores de *d* são normalmente mantidas pequenas, e expressa na forma (d,θ), com *d* representando a distância linear em *pixels*, e θ o ângulo entre eles [BUSCH, 2005].

Tipicamente, θ é limitado aos valores {0^o, 45^o, 90^o, 135^o}, e *d* é limitado a um pequeno intervalo de valores. As distâncias são escolhidas de acordo com a granularidade das imagens manipuladas. Nesta abordagem foram utilizadas distâncias *d* = 1, 2, 3, 4 e 5. Para cada direção e cada distância gera-se uma matriz de co-ocorrência.

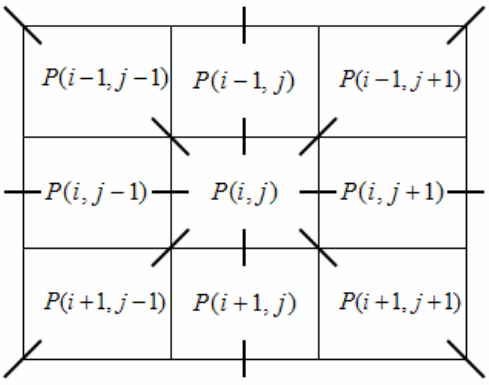


Figura 5. Exemplo de posições de *pixels.*

As matrizes de co-ocorrência formam a base para elaboração de diversas medidas estatísticas conhecidas como descritores de Haralick, como o segundo momento angular, contraste, entropia, momento da diferença inversa, variância, dissimilaridade, media [ALVES, 2006].

Dentre os 14 descritores de Haralick [HARALICK, 1973], que se tratam de características da textura, foram selecionadas seis (Tabela 3), que se tratam das mais utilizadas, de acordo com as abordagens recentes [BUSCH, 2005][AL-DMOUR, 2007][SAID, 1998]. As demais equações de Haralick são derivações das principais.

Tabela 3. Descritores de Haralick utilizados neste trabalho.

Descritor	Equação	
2º Momento Angular	$\sum_{i=0}^{n}\sum_{j=0}^{m}(p(i,j))^2$	(3)
Entropia	$-\sum_{i=0}^{n}\sum_{j=0}^{m}p(i,j).\log(p(i,j))$	(4)
Homogeneidade	$\sum_{i=0}^{n}\sum_{j=0}^{m}p(i,j)/(1+\lvert i-j\rvert)$	(5)
Dissimilaridade	$\sum_{i=0}^{n}\sum_{j=0}^{m}p(i,j)/\lvert i-j\rvert)$	(6)
Variância Inversa	$\sum_{i=0}^{n}\sum_{j=0}^{m}p(i,j)/(i-j)^2) \quad i \neq j$	(7)
Energia	$\sqrt{\sum_{i=0}^{n}\sum_{j=0}^{m}(p(i,j))^2}$	(8)

- **Segundo momento angular**: avalia a uniformidade textural em uma imagem;
- **Entropia ou Suavidade:** fornece o grau de dispersão de níveis de cinza de uma imagem;
- **Homogeneidade ou momento da diferença inversa**: refere-se à distribuição dos *pixels;*
- **Dissimilaridade**: mede o desvio dos valores da combinação de pares de *pixels* diagonais, em que apenas a contribuição do desvio é considerada;
- **Variância inversa**: corresponde ao inverso de contraste;
- **Energia:** é obtida pela raiz do segundo momento angular;

A análise da imagem é feita sob um conjunto de matrizes de co-ocorrência para caracterizar a textura à qual elas se referem, utilizando-se um ou mais descritores. A seleção dos descritores ou características a serem adotados, baseia-se em testes empíricos

sobre um domínio específico, como nesta abordagem, a textura do traçado da escrita, para verificar quais trazem resultados mais satisfatórios.

A matriz de co-ocorrência foi selecionada para ser aplicada de acordo com a abordagem deste trabalho, devido sua simplicidade de ser implementada e o baixo custo computacional. Trata-se também de uma das técnicas mais aplicadas para extração de características de textura, como foi utilizada nas abordagens atuais [FRANKE, 2002], [BUSCH, 2005], [AL-DMOUR, 2007], [SAID, 1998].

3.3.2.2 Inclinação Axial

A inclinação axial é uma característica grafocinética que descreve o aspecto dinâmico do traçado e o ângulo de inclinação da escrita em relação ao eixo vertical de um sistema de eixos cartesianos, onde o eixo horizontal é representado por uma linha base imaginária. A inclinação pode ocorrer à direita, à esquerda ou ser nula (alinhada ao eixo vertical), podendo ainda ocorrer para alguns autores, um misto de inclinações em sua escrita (Figura 6) [BARANOSKI, 2005].

(a) (b) (c)

Figura 6. Exemplo de escita com inclinação axial: (a) à direita; (b) à esquerda; (c) nula

Esta característica foi implementada na abordagem de Baranoski [BARANOSKI, 2005] utilizando a base original PUC-PR. Por este motivo, ela foi selecionada e desenvolvida na abordagem deste trabalho para ser aplicada na base compactada e assim poder comparar as diferenças de resultados dessa característica em relação às duas bases.

3.3.3 Treinamento de Modelos

Os métodos computacionais destinados à identificação e verificação da autoria de manuscritos baseiam-se, usualmente, em duas abordagens, a abordagem independente do escritor ou WI (*Writer–Independent*) e a abordagem dependente do escritor ou WD (*Writer-Dependent*) [BERTOLINI, 2008]. A abordagem WD utiliza um modelo por

escritor, enquanto que a outra faz uso de um modelo geral para todos os escritores, ou seja, o autor é obrigado a escrever o mesmo texto fixo em todos os espécimes. A abordagem WD possui a vantagem de modelar adequadamente os atributos genéticos e genéricos do escritor e como desvantagem, exigir um conjunto elevado de exemplares de manuscritos, na geração de um modelo pessoal robusto.

A abordagem WI, em que o autor não escreve o mesmo texto, possui as vantagens de generalização e a necessidade de um número reduzido de exemplares de manuscritos para cada escritor e de não necessitar de um novo treinamento do modelo, diante da análise de manuscritos de outros autores, que não tenha participado da geração do modelo inicial.

No treinamento do modelo WI, a classe *w1* representa a classe de manuscritos do mesmo escritor. A classe *w2* representa o conjunto de manuscritos de escritores distintos. Na verificação, o modelo gerado é então utilizado para a comparação com os manuscritos de escritores desconhecidos. Os vetores de características são extraídos das imagens de referência e das imagens questionadas. O vetor de dissimilaridade, que nada mais é do que o módulo das diferenças entre os componentes dos vetores é calculado e usado para treinar o classificador. A idéia por traz disso é que se as amostras de referência e questionada pertencerem aos mesmos autores, o vetor de dissimilaridade terá componentes próximo a zero, caso contrário, os componentes serão bem maiores que zero.

3.3.4 Classificação

Existem diversos classificadores que são utilizados para auxiliar estudos na área reconhecimento de padrões, e para a autenticação de manuscritos, três destes serão conceituados, que foram mais utilizados por outros estudos científicos, de acordo com o estado da arte demonstrado neste trabalho. O SVM *(Support Vector Machine)* foi o classificador selecionado para auxiliar a classificação na abordagem deste trabalho, pois é adequado para um grande número de manuscritos, segundo observações de Shen [SHEN, 2002].

3.3.4.1 K-NN (*k-nearest neighbors)*

O método K-NN (*k-nearest neighbors*) k-vizinhos mais próximos, é uma abordagem importante para a classificação não-paramétrica, sendo considerado fácil e eficiente. Os "vizinhos mais próximos são determinados, considerando o vetor e a distância mensurada." Usualmente é utilizado a distância euclidiana para medir os k-vizinhos mais próximos. [JIANGSHENG, 2002]

O K-NN é, em suma, um classificador, que recebe como parâmetro o valor de K, sendo que a partir deste valor, o classificador busca os K elementos (ou vizinhos mais próximos) do grupo de treinamento que possuem a menor distância em relação ao elemento que está sendo identificado. Após, o método analisa quais são as classes respectivas aos elementos vizinhos identificados e determina qual é a classe que possui maior freqüência. O método obtém a métricas utilizando o cálculo da distância Euclidiana.

Sejam dois pontos:

$$X = (x_{1,} x_{2,...,} x_n) \text{ e } Y = (y_{1,} y_{2,...,} y_n)$$

A distância Euclidiana entre os pontos X e Y é dada por:

$$\mathrm{d(x,y)} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} \qquad (9)$$

Em termos de desvantagem, o K-NN é um método computacional que se torna exaustivo quando há uma grande quantidade de dados e é intolerante a ruídos, porém é um método vantajoso por ser flexível e possibilitar alterações, e por este motivo, é um classificador muito utilizado nos estudos de áreas de classificação.

Este classificador apresentou bons resultados em relação à utilização da matriz de co-ocorrência nas abordagens de Said e Al-Dmour [SAID, 1998] [AL-DMOUR, 2007] e em relação à características da dobradiça em [BRINK, 2007]. Também foi utilizado na abordagem de [SHEN, 2002], aplicado após o uso dos Filtros de Gabor, porém, para poucos espécimes.

3.3.4.2 SVM (*Support Vector Machine*)

O classificador denominado SVM (*Support Vector Machine*) se destaca por atuar em aplicações do campo de reconhecimento de faces, classificação de impressões digitais, e em relação aos manuscritos, é aplicado na verificação de assinaturas, reconhecimento da cadeia de dígitos, identificação e verificação de autoria. Segundo Shen [SHEN, 2002], o SVM é mais eficaz classificação de grandes quantidades de espécimes. Devido aos envolvimentos nestes campos, o SVM foi o classificador selecionado para uso na abordagem deste trabalho.

O objetivo do SVM é separar da melhor forma possível os pontos dos dados de classes com superfícies que maximizem a margem entre elas. O SVM identifica os pontos mais próximos do limite entre as duas classes que auxiliarão a definir a forma da melhor superfície.

O SVM possui como vantagem a robustez diante de dados de grande dimensão e a boa capacidade de generalização, pois procura minimizar o risco estrutural (SMR).

No SVM, os padrões de entrada são transformados em um vetor de características de alta dimensionalidade. O mapeamento das características no espaço pode ser linear ou não, dependendo da função *kernel*, que é responsável pelo preenchimento da superfície de hiperplano por dados. A função *kernel* pode ser linear, polinomial ou RBF.

A literatura apresenta várias possibilidades de *kernels* para o SVM em aplicações envolvendo o reconhecimento de padrões [VAPNIK, 1998][BURGES, 1998]. Nesse estudo foram testados os principais *kernels*, mas os melhores desempenhos foram obtidos utilizando o *kernel* linear e gaussiano, cujos resultados são apresentados no capítulo 5.

Durante o processo de treinamento o SVM seleciona vetores de suporte ao longo da superfície da função *kernel* (Tabela 4) determinada, permitindo classificar uma faixa de problemas maiores.

Kernel	*Expressão*
Linear	$K(x_i, x) = x.x_i$ (10)
Polinomial de grau *d*	$K(x_i, x) = (1 + x.x_i)^d$ (11)
Gaussiano RBF	$K(x_i, x) = \exp(-\|x - x_i\|^2)$ (12)

Tabela 4. *Kernels* do SVM

Após o espaço adequado de características ser definido, o SVM seleciona o hiperplano particular, denominado hiperplano de margem máxima (MMH), o qual corresponde a maior distância de seus padrões no conjunto de treinamento.

A função de decisão do SVM, $f(\vec{x})$, treinada linearmente, é descrita pelo vetor de pesos $\vec{p}$, um limiar b e padrões de entrada $\vec{x}$:

$$f(\vec{x}) = sign(\vec{p}.\vec{x} + b) \qquad (13)$$

Dado um conjunto de treinamento S_l composto por duas classes separadas $w_1(y_1 = +1)$ e $w_2(y_2 = +1)$, o SVM encontra o hiperplano com a máxima distância Euclidiana. De acordo com os princípios do SRM, haverá apenas um hiperplano ótimo com margem máxima δ, definida como a soma das distâncias do hiperplano para os pontos mais próximos das classes. Esse limiar do classificador do linear é o hiperplano ótimo separador, conforme demonstrado na fórmula:

$$S_l = ((\vec{x}_1, y_1), ..., (\vec{x}_l, y_l)), \vec{x}_1 \in \Re^n, y_1 \in \{-1, +1\} \quad (14)$$

No caso de conjuntos de treinamentos não separáveis, o *i-ésimo* ponto de dados tem uma variável ξ inativa, a qual representa a magnitude do erro de classificação. Uma função de penalidade $f(\xi)$ representa a soma dos erros de má classificação [BARANOSKI, 2005]:

$$f(\xi) = \sum_{i=1}^{l} \xi \qquad (15)$$

A solução do SVM pode ser encontrada se mantiver o limite superior na dimensão VC (número de pontos máximos que pode ser separado para um conjunto de dados), e por minimizar o limite superior de risco empírico, isto é, o número de erros de treinamento, com a seguinte minimização:

$$\min_{w,b,\xi} = \frac{1}{2} \vec{p}.\vec{p} + C\sum_{i=1}^{l} \xi \qquad (16)$$

sendo que C > 0 determina o compromisso entre o erro empírico e o termo de complexidade. O parâmetro de C é escolhido livremente. Um grande valor para C corresponde à associação de uma penalidade mais alta de erros [JUSTINO, 2003].

Este classificador foi utilizado na abordagem de Imdad [IMDAD, 2007], que utiliza as características de Hermite, e na abordagem de Al-Dmour e Zitar, apresentando vantagem nos resultados em relação aos classificadores K-NN e WED. Também foi utilizado na abordagem de Franke [FRANKE, 2002], que envolve as matrizes de co-ocorrência.

3.3.4.3 WED *(Weighted Euclidian Distance)*

A WED se destaca por possuir a vantagem ser um classificador computacionalmente simples [SAID, 1998]. Características representativas são extraídas de textos manuscritos treinados para cada autor. Uma operação similar de extração de característica é realizada em seguida para um bloco de entrada de texto manuscrito de um autor desconhecido, e essas características são comparadas com características representativas de um grupo de escritores conhecidos. O autor do manuscrito é identificado como escritor *k* pelo classificador WED se a função que se segue é uma distância mínima de *k*.

$$d(k) = \sum_{i=1}^{N} \frac{\left(f_n - f_n^k\right)^2}{\left(v_n^k\right)^2} \qquad (17)$$

em que f_n é a n^a característica de entrada do documento, $f_n(k)$ é a média da amostra e $v_n(k)$ é o desvio padrão da amostra da n^a característica do autor *k*, e *N* é o número total de características extraída de um único autor.

O classificador WED apresentou bons resultados na abordagem de [SAID, 1998], após utilização da técnica de Filtros de Gabor e na abordagem de [BRINK, 2007].

3.3.4.4 Fusão de resultados

Para determinar estatisticamente os resultados da classificação realizada pelo SVM, decisões por fusão de resultados podem ser aplicadas no arquivo de saída do SVM com intenção de buscar maiores taxas de acerto. Um dos votos realizados a partir do arquivo original de saída do classificador SVM é o voto majoritário em que a minoria é desconsiderada e a maioria dos votos é levada em consideração. Por exemplo, na abordagem deste trabalho, são analisados grupos de 5 saídas do SVM. Se a maioria forem resultados positivos, significa que o classificador classificou como autor. Caso contrário, sendo a maioria negativo, trata-se de não-autor.

Existem também as fusões realizadas a partir dos resultados de saída do SVM. A fusão por valor máximo, em que é considerado o maior valor de um grupo de resultados. Se este valor for positivo, trata-se de autor, se negativo, é considerada não-autoria. Para a fusão por valor mínimo, o menor valor é selecionado, aplicando-se a mesma regra de decisão da máxima. No caso da fusão por média, os valores de um grupo são colocados em ordem crescente, e é selecionado o valor que se encontra na posição mediana para ser avaliado. A mesma regra é aplicada para decisão de autoria, assim como na fusão por valor máximo e mínimo. Na fusão por soma, todos os valores do grupo são somados e o resultado indica a autoria ou não-autoria.

3.4 Comentários Finais

Neste capítulo foram apresentados os conceitos dos elementos que fazem parte do processo para a verificação da autoria, fundamentais para definição da abordagem deste trabalho. Esta fundamentação teórica descreveu as abordagens e métodos (extração de características, treinamento, classificadores) que fazem parte do processo deste trabalho

que propõe a verificação de autoria de manuscritos dentro das regras de reconhecimento de padrões. No próximo capítulo será apresentada a metodologia da abordagem deste trabalho.

Capítulo 4

Método Proposto

Neste capítulo é apresentado o método proposto para a verificação de autoria de manuscritos. São detalhadas as etapas da base, pré-tratamento dos manuscritos, extração de características, comparação e decisão.

4.1 Introdução

Os peritos grafotécnicos classificam os textos manuscritos em relação à autoria como: associação w_1 que indica que a grafia presente no manuscrito foi elaborada, de próprio punho, pelo autor avaliado; ou dissociação w_2, que indica que o manuscrito não foi produzido de próprio punho, pelo autor avaliado.

Baranoski [BARANOSKI, 2005] cita que com base no modelo da visão pericial, um modelo computacional pode ser estruturado. Matematicamente, durante a prova pericial, o perito utiliza um conjunto n de amostras de manuscritos de autoria desconhecida (referência) $M_{ki}(i = 1,2,3,...,n)$, em comparação com a amostra do manuscrito de autoria desconhecida (questionada) M_Q. O perito observa as diferenças $D_i(i = 1,2,3,...,n)$ entre as L características grafoscópicas do conjunto de amostras de referência $fv_{ki,j}(i = 1,2,3,...,n)(j = 1,2,3,...,L)$ e da questionada $fv_{Qj}(j = 1,2,3,...,L)$. Após este procedimento, toma a decisão $R_i(i = 1,2,3,...,n)$. O laudo pericial resultante D depende da soma dos resultados obtidos das comparações individuais dos pares (referência / questionada) (Figura 7).

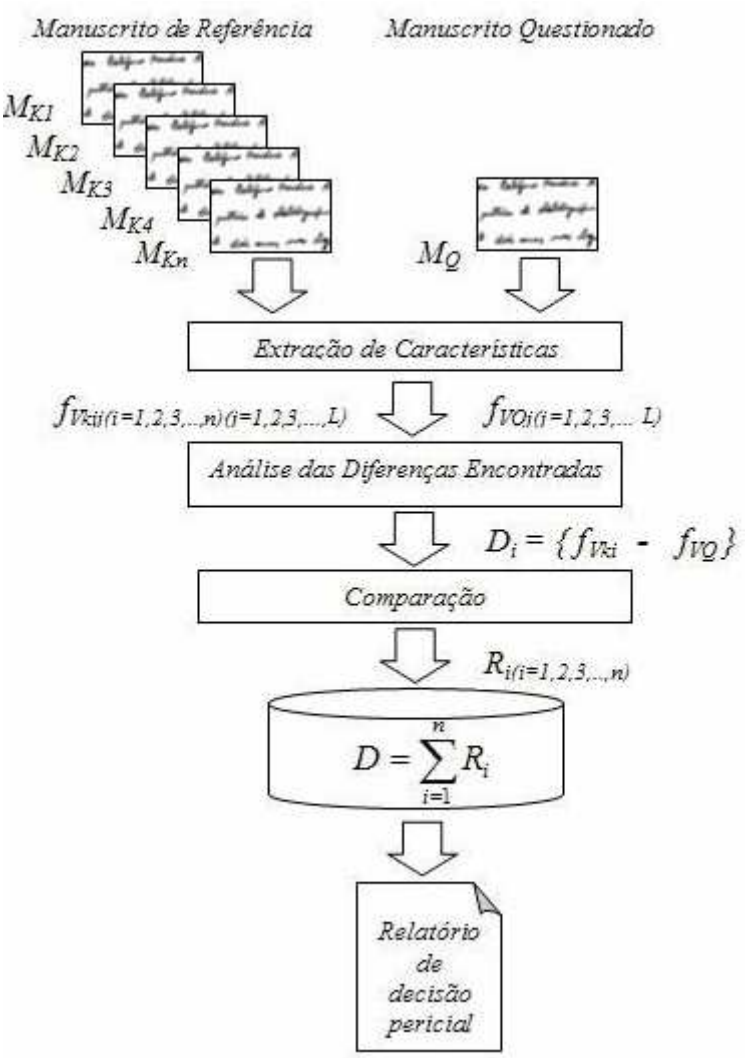
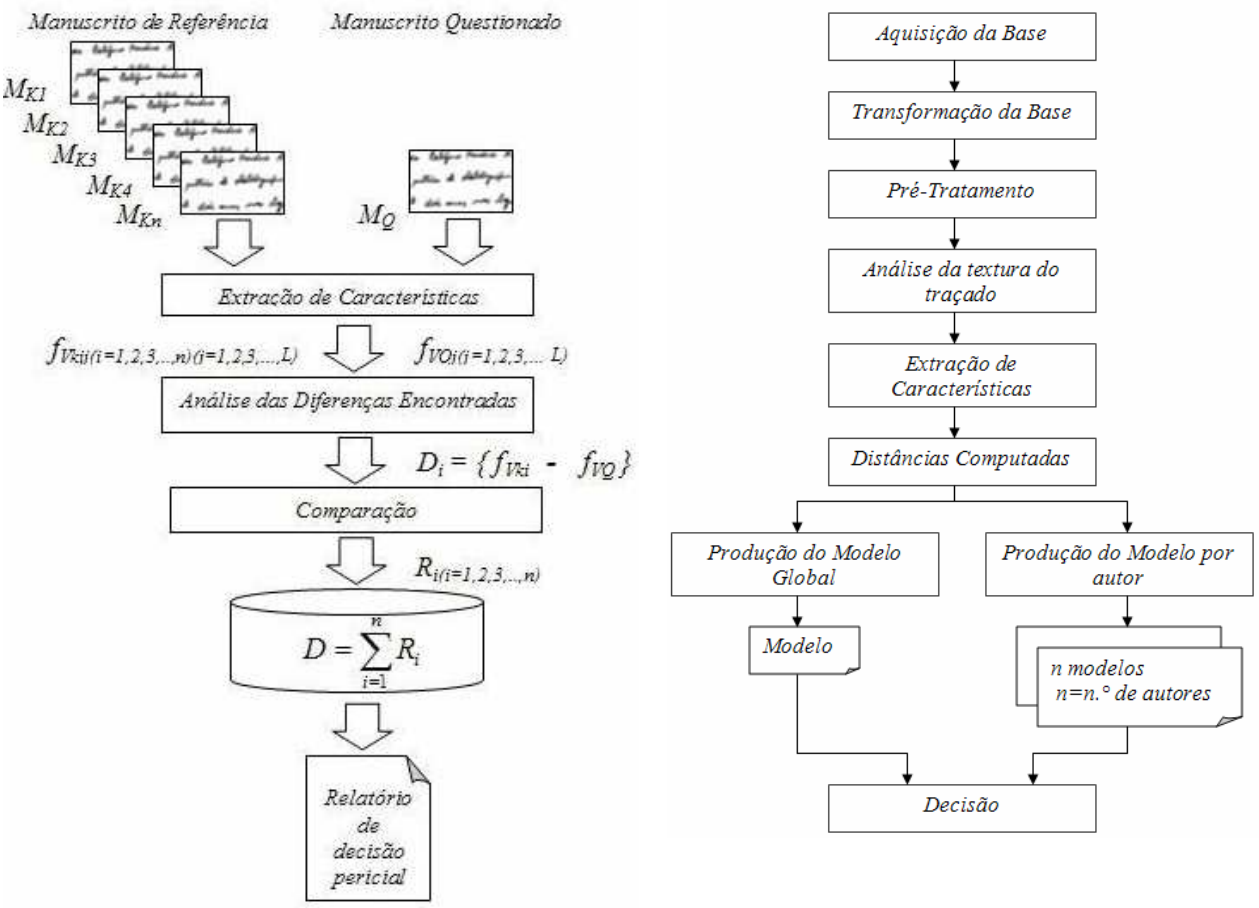


Figura 7. Esquema do processo de decisão na verificação de manuscritos baseado na visão pericial

Este trabalho tem como objetivo estabelecer um processo que utilize os recursos da grafoscopia buscando criar um método computacional que seja semelhante à visão que o perito tem no processo de análise de um documento manuscrito.

Para criar um método de verificação de autoria semi-automática de manuscritos, a abordagem deste trabalho foi elaborada contendo as seguintes etapas: transformação da base de dados, em que será adquirida nova base com a escrita compactada; pré-processamento das imagens; extração de características; treinamento, para obtenção de um modelo; classificação; realização de novos experimentos em busca de resultados mais satisfatórios.

A implementação de um método automático do trabalho proposto para verificação de autoria de manuscritos requer alguns procedimentos adicionais. Abaixo é demonstrado o processo pericial, comparado ao processo automático de autenticação de manuscritos (Figura 8).



(a) (b)

Figura 8. Comparação das etapas no processo de verificação de manuscritos: (a) processo de análise e decisão pericial e (b) no método computacional estabelecido neste trabalho.

Os sistemas automáticos de verificação da autoria de textos manuscritos baseiam-se usualmente em duas abordagens de modelos para classificação: o modelo dependente do escritor e o modelo independente do escritor, a pessoal e o global [JUSTINO, 2003].

O modelo dependente do escritor é baseado no conceito de policotomia, em que o problema é classificado em n-classes e cada autor representa uma classe. Este modelo tem como vantagem a descrição adequada das variabilidades intrapessoais do autor, porém possui como desvantagem a geração de um modelo novo a cada autor que seja incluído. Outra desvantagem é a necessidade de um conjunto elevado de genuínos para a geração do modelo, pois para cada autor será gerado um modelo específico com descrição das características do mesmo.

O modelo independente do escritor é baseado no conceito da dicotomia, em que o modelo é dividido em duas classes, a autoria e não-autoria. A geração do modelo global ocorre com um conjunto de autores escolhidos aleatoriamente, combinando-se espécimes de um mesmo autor e de autores diferentes. No treinamento do modelo global, a classe w_1 representa a classe de espécimes genuínos dos autores usados para o treinamento. A classe w_2 representa o conjunto de espécimes pertencentes a autores distintos. Na verificação o modelo gerado é então utilizado para a comparação com o espécime desconhecido. Este modelo possui como desvantagem da generalização. Porém, foi selecionado para ser aplicado neste trabalho, devido suas vantagens, como a utilização de um número reduzido de exemplares de cada autor e a desnecessidade de um novo treinamento do modelo quando um autor é incluído, o que torna o processo simples e viável ao aplicá-lo em situações reais.

Para o desenvolvimento do sistema de verificação de manuscritos, tornam-se indispensáveis as etapas fundamentais, como: aquisição de dados, em que a imagem do manuscrito é produzida a partir de um *scanner*; pré-processamento, em que há a preparação das imagens para a extração de características, sendo possível aplicar mais que um pré-tratamento, dependendo do caso; segmentação: em que ocorre uma divisão da imagem do manuscrito em diversos fragmentos usados na fase de extração de características; cálculo da distância entre as características, em que são calculadas as diferenças entre os vetores de características extraídas, usadas na produção do modelo e no processo de decisão; produção de um modelo; conjunto de referências de manuscritos gerado para se realizar o processo comparativo; e o processo de decisão ou classificação, em que ocorre a avaliação da saída do modelo produzido, verificando se o manuscrito caracteriza associação (autoria) ou dissociação (não-autoria);

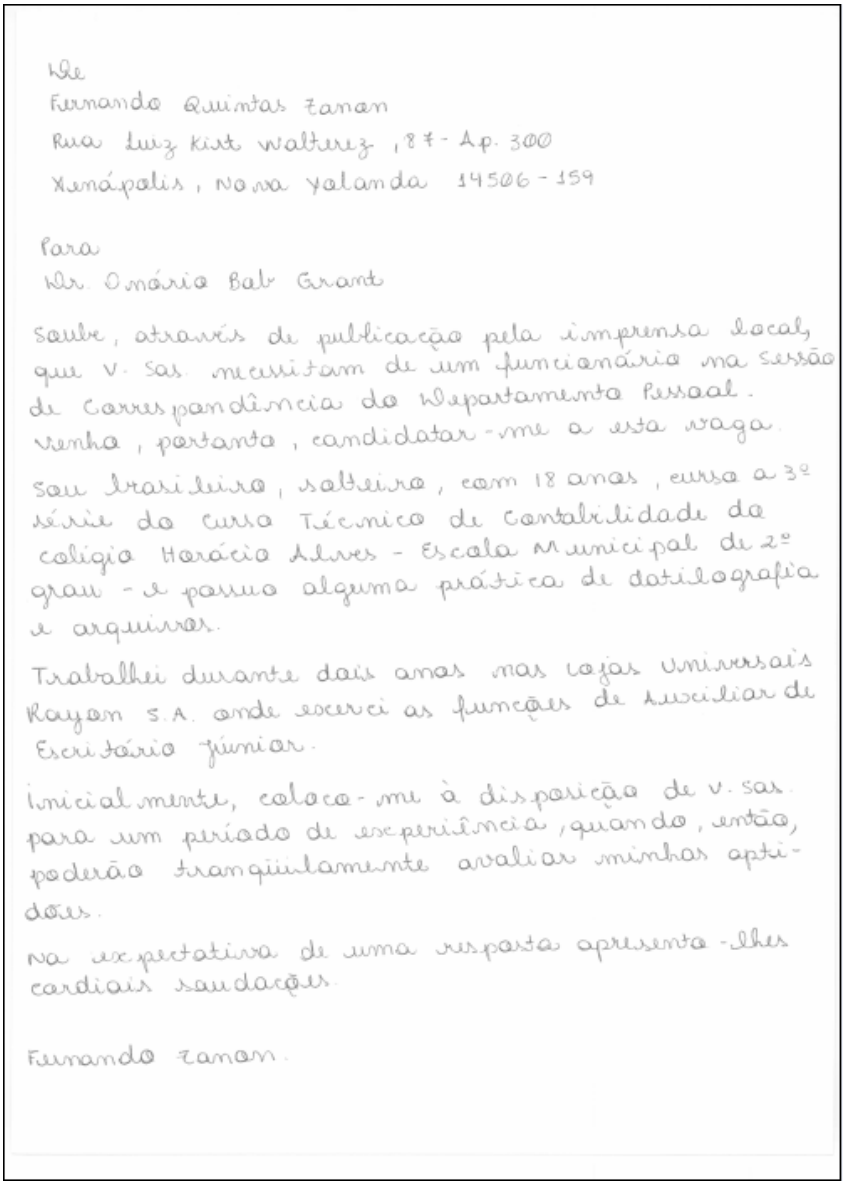
A seguir, a descrição do processo realizado na abordagem deste trabalho é descrito.

4.2 Aquisição e Preparação da Base

Os resultados obtidos a partir do aprendizado e os testes de um método desenvolvido para autenticação de manuscritos dependem da composição de dados o qual está utilizando. A base deve ser formada por uma quantidade suficiente de autores para que seja possível obter uma validação estatística. Outro fator importante é a quantidade de espécimes por autor, para que possa representar de maneira satisfatória as variações intrapessoais de cada autor.

A base utilizada nesta abordagem é a Base de Dados PUC-PR, com um total de 315 autores, sendo três amostras redigidas por autor, de mesmo teor, totalizando em uma base de 945 imagens. A base encontra-se sob os cuidados do Programa de Pós-Graduação em Informática. Esta base foi selecionada para utilização no trabalho, devido a eliminação de suspeitas de falsificações, por ser de origem natural e também com o objetivo de comparações, por já ter sido utilizada em outros trabalhos como na abordagem de Baranoski [BARANOSKI, 2005].

Esta base foi obtida pela PUC-PR devido à ausência de um modelo para a escrita latina, mais especificamente a de língua portuguesa. O conteúdo da carta foi elaborado de forma a contemplar o conjunto de letras do alfabeto da Língua Portuguesa tanto minúscula quando minúscula. O modelo de manuscrito possui todas as particularidades da escrita da Língua Portuguesa, como símbolos da acentuação (tais como o til, cedilha, acento circunflexo, acento grave, acento agudo e pingo nos i´s) e mínimos gráficos, contendo um léxico de 124 palavras (Figura 9).



De
Fernando Quintas Zanan
Rua Luiz Kist Walteriz, 87 - Ap. 300
Xenápolis, Nova Yolanda 14506-159

Para
Dr. Omário Bab Grant

Soube, através de publicação pela imprensa local, que V. Sas. necessitam de um funcionário na sessão de Correspondência do Departamento Pessoal. Venho, portanto, candidatar-me a esta vaga.

Sou brasileiro, solteiro, com 18 anos, curso a 3ª série do curso Técnico de Contabilidade do colégio Horácio Alves - Escola Municipal de 2º grau - e possuo alguma prática de datilografia e arquivar.

Trabalhei durante dois anos nas Lojas Universais Rayon S.A. onde exerci as funções de Auxiliar de Escritório Júnior.

Inicialmente, coloco-me à disposição de V. Sas. para um período de experiência, quando, então, poderão tranqüilamente avaliar minhas aptidões.

Na expectativa de uma resposta apresento-lhes cordiais saudações.

Fernando Zanan.

Figura 9. Exemplo de um manuscrito digitalizado da base PUC-PR.

Os manuscritos foram obtidos por meio de uma colheita feita com voluntários, na maioria durante as seções programadas em instituições de ensino no decorrer dos anos 2002 e 2005, aos quais é apresentado o modelo que deve ser transcrito na íntegra 3 vezes em folhas de papel A4 (21 x 29,7 cm), sem pauta. Esta transcrição foi realizada de acordo com regras estabelecidas que não devem influenciar o processo de escrita do autor: uso

de caneta esferográfica azul ou preta; não hifenização das palavras caso falte espaço na linha e o texto deveria ser escrito sem auxílio de linhas guia.

Após a colheita dos manuscritos, estes foram digitalizados por meio de um *scanner Hewllet-Packward* modelo *HP Scanjet 5550*c, com 256 níveis de cinza, densidade de 300 dpi e formato *Bitmap* (.BMP), sem interferência de ruídos ou imagens pré-impressas.

4.3 Pré-tratamento

O pré-tratamento das imagens consiste em prepará-las através de transformações para a etapa seguinte, de extração de características.

4.3.1 Segmentação do manuscrito com base nos atributos genéticos e genéricos.

A abordagem proposta utiliza uma análise da textura do traçado da escrita, e devido a isto, toda a base original de cartas PUC-PR foi transformada para eliminar espaços em branco, pois a aplicação da extração das características de textura na base original trouxe resultados com alta taxa de erro, já que a presença dos espaços em branco contém informações redundantes que dificultam a classificação.

O critério de segmentação se baseia nas propriedades inerentes dos atributos genéticos e genéricos da grafia do escritor. Isto é, independentemente da forma com que o texto tenha sido redigido (independente do contexto), o processo de segmentação procederá à separação dos segmentos de traço não ligados ou conexos (presença de momentos ou espaçamentos), mantendo seu ângulo de inclinação original (comportamento base), reorganizando cada segmento encontrado em um novo alinhamento, reduzindo os espaços em branco entre as linhas de texto, entre palavras e entre segmentos de palavras. Com isso, quanto maior for a inclinação em relação à linha de base e quanto maior for a presença de momentos entre letras de uma palavra, mais compacto e mais fragmentado será o texto resultante.

A transformação envolve um algoritmo desenvolvido que realiza o processo de segmentação (extração e compactação do conteúdo da escrita) criando-se novas imagens,

eliminando os espaços em branco. Assim, foi criada uma nova base de imagens com o texto compactado da escrita.

O processo de segmentação é executado da seguinte forma:

1. A imagem de um texto manuscrito é carregada em 256 níveis de cinza e aplica o pré-tratamento da binarização;
2. A imagem binarizada é gravada separadamente;
3. Em seguida, percorre-se a imagem binária de cima para baixo e da esquerda para a direita, até que um componente de traço (*pixel* preto) seja encontrado.
4. Retira-se da imagem binária todos os *pixels* conexos ao *pixel* encontrado, através da lógica de preenchimento da área (*fill area*), aplicado de forma recursiva, para marcar um grupo de *pixels* interligados a partir do que foi encontrado, nas oito direções do *pixel* principal encontrado;
5. O grupo ou trecho é marcado com outra cor na imagem binarizada (em memória *buffer*);
6. A partir dos valores de altura e largura obtidos após item 5, o algoritmo busca na imagem original o trecho, e o transcreve para uma imagem auxiliar (Figura 10);

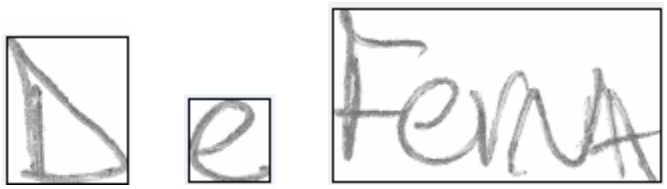


Figura 10. Exemplos de trechos selecionados pelo algoritmo *fill area.*

7. O algoritmo compara o trecho da carta binarizada, gravado separadamente, com o trecho recortado do original, para selecionar apenas o texto que corresponde ao grupo preenchido pelo algoritmo *fill area* (Figura 11). Este passo é realizado para eliminar possíveis traços que no recorte do *bounding box*, estejam contidos na caixa, mas que não fazem parte do grupo selecionado pelo algoritmo *fill area*;

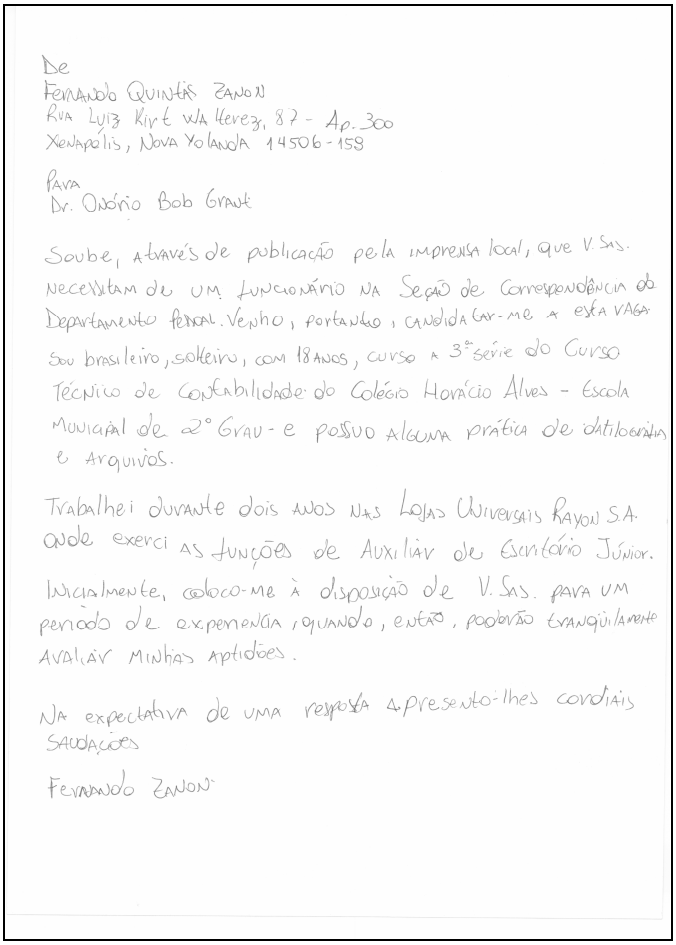
Figura 11. Exemplo de trecho que ao ser recortado pelo *bounding box*, trouxe consigo parte da escrita não selecionada pelo *fill area* e é eliminada pelo algoritmo.

8. Transfere-se da imagem em níveis de cinza, todos os *pixels* obtidos pela imagem binária e os rearranja em uma nova área (nova imagem), usando com linha de base ou pauta, o ponto médio do conjunto extraído;
9. Em seguida, retira-se o conjunto de *pixels* da imagem binária e busca-se por uma nova ocorrência, repetindo-se o passo inicial.

Abaixo é descrito o algoritmo de segmentação em pseudo-código:

```
METODO BINARIZA IMAGEM()
TRANSFORMAA IMAGEM EM PRETO E BRANCO

METODO MAIN()
MATRIZ_IMAGEM_BINARIZADA = MATRIZ INTEIRO[TAMANHO_IMAGEM][TAMANHO_IMAGEM]
MATRIZ_IMAGEM_ORIGINAL = MATRIZ INTEIRO[TAMANHO_IMAGEM][TAMANHO_IMAGEM]
LINHA_BASE = 0 ;
        PARA (I=0; I<MATRIZ.TAMANHO; I++)
        PARA (J=0; J<MATRIZ.TAMANHO; J++)
        SE MATRIZ[I][J] = PRETO
                FILLAREA()
                EXTRAI_BOUNDING_BOX()
                NOVO FRAGMENTO EM BRANCO(MENOR_I, MAIOR_I, MENOR_J, MAIOR_J)
                SE (HÁ ESPAÇO NA MESMA LINHA PARA INSERIR NOVO FRAGMENTO)
                        PARA (X=MENOR_I; X<MAIOR_I; X++)
                                PARA (Y=MENOR_I; Y<MAIOR_J; Y++)
                                        SE(MATRIZ_IMAGEM_BINARIZADA[X][Y]=OUTRA COR)
                                                INTEIRO VALOR=MATRIZ_IMAGEM_ORIGINAL[X][Y]
                                                GRAVA PIXEL NOVO FRAGMENTO(VALOR, X, Y)
                        VETOR_ALTURAS = ALTURA NOVO FRAGMENTO
                        ESCREVE_FRAGMENTO_CARTA_CAMPACTADA(LINHA_BASE, FRAGMENTO)
                SENAO
                        LINHA_BASE = LINHA_BASE + MÉDIA VETOR ALTURAS / 2 ;
                        VETOR_ALTURAS = ZERADO;

METODO FILLAREA()
        PINTA DE OUTRA COR OS PIXELS CONECTOS A PARTIR DO PIXEL ENCONTRADO


METODO EXTRAI BOUNDING BOX ()
        DE ACORDO COM OS PIXELS DE OUTRA COR, EXTRAI:

        MAIOR_I = VALOR DA LINHA DO PIXEL QUE POSSUI MAIOR I;
        MENOR_I = VALOR DA LINHA DO PIXEL QUE POSSUI MENOR I;
        MAIOR_J = VALOR DA COLUNA DO PIXEL QUE POSSUI MAIOR J;
        MENOR_J = VALOR DA COLUNA DO PIXEL QUE POSSUI MENOR J;

METODO ESCREVE_FRAGMENTO_CARTA_CAMPACTADA()

        A PARTIR DA MAIOR COLUNA DO ULTIMO FRAGMENTO GRAVADO, ESCREVE O FRAGMENTO POSICIONANDO

SEU CENTRO NA LINHA BASE;
```

A distância entre as linhas de base ou pauta é determinada pela média das alturas encontradas nos segmentos da linha corrente. Dessa maneira reduz-se a presença de espaços em branco entre linhas. Portanto, a grafia que apresentar maior ocorrência de ascendentes e descendentes (proporcionalidade baixa), apresentará igualmente uma maior ocorrência de sobreposições Figura 12.



De
Fernando Quintas Zanon
Rua Luiz Kiyt Wa Herezi, 87 - Ap. 300
Xenapólis, Nova Yolanda 14506-158

Para
Dr. Osório Bob Grant

Soube, através de publicação pela imprensa local, que V. Sas. necessitam de um funcionário na Seção de Correspondência do Departamento Pessoal. Venho, portanto, candidatar-me a esta vaga. Sou brasileiro, solteiro, com 18 anos, curso a 3ª série do Curso Técnico de Contabilidade do Colégio Horácio Alves - Escola Municipal de 2º Grau - e possuo alguma prática de datilografia e arquivos.

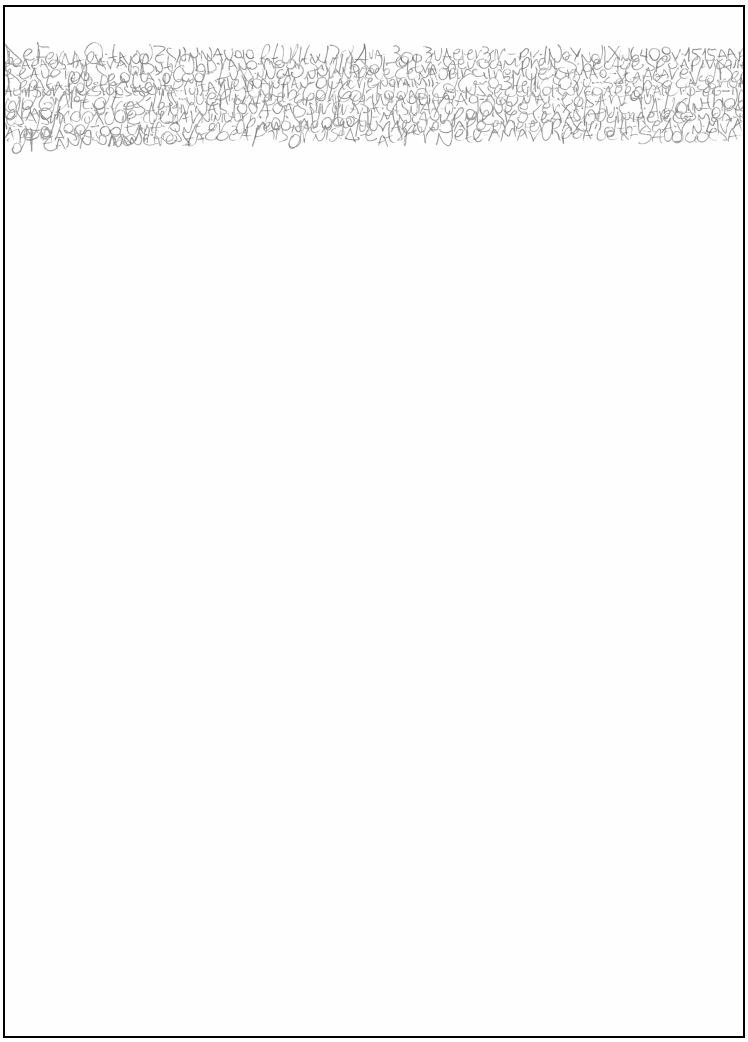
Trabalhei durante dois anos nas Lojas Universais Rayon S.A. onde exerci as funções de Auxiliar de Escritório Júnior.

Inicialmente, coloco-me à disposição de V. Sas. para um período de experiência, quando, então, poderão tranquilamente avaliar minhas aptidões.

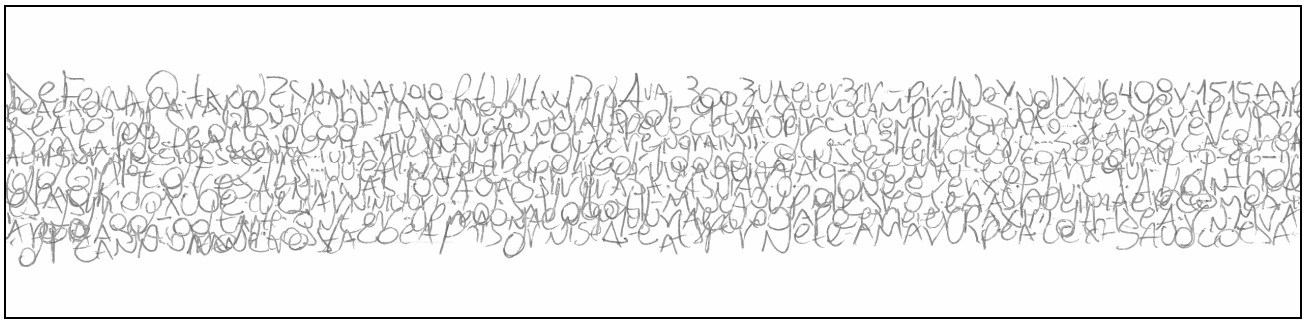
Na expectativa de uma resposta apresento-lhes cordiais saudações

Fernando Zanon

(a)



(b)



(c)

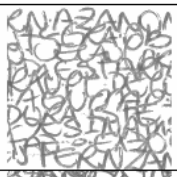
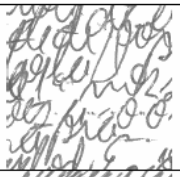
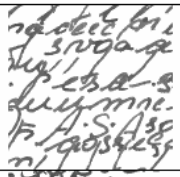
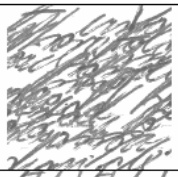
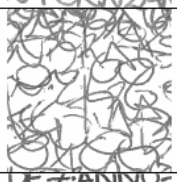
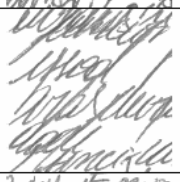
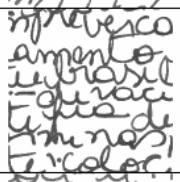
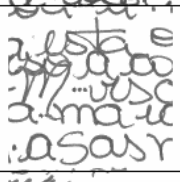
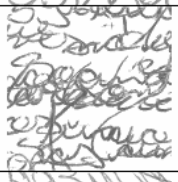
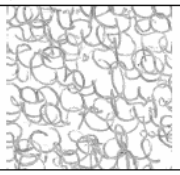
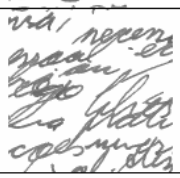
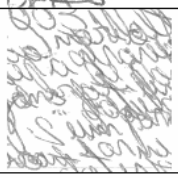
Figura 12. Um exemplo do processo de segmentação; (a) Exemplo de uma carta da base PUCPR; (b) Exemplo da carta segmentada; (c) melhor vizualização do texto segmentado.

Na imagem gerada pelo algoritmo de segmentação do texto, os trechos são adicionados conforme estes são encontrados, na busca de cima para baixo, da esquerda para a direita, e isto implica na não-formação de uma sintaxe. A falta de sintaxe do texto não influência na análise da textura que depende apenas do conteúdo do traçado.

A transformação que tem como objetivo eliminar os espaços em brancos e agrupar a parte escrita na nova imagem, gerar um padrão de textura da escrita para ser analisado, sendo que os espaços em branco não contém informação que contribua para a autenticação de manuscritos, prejudicando os resultados.

Além dos atributos grafoscópicos vistos anteriormente, como parte intrínseca do processo de segmentação, outros são passíveis de serem observados como resultantes do mesmo processo. Na Tabela 5(a) é possível observar variação no calibre para diferentes escritores. Na Tabela 5(b) se observam às mudanças de direção do traçado, referentes aos valores curvilíneos e angulares para escritores distintos. Na Tabela 5(c) se observa os diferentes graus de variabilidade entre ascendente, descendentes e corpo das letras. Na Tabela 5(d) se observa a diferença de pressão entre escritores distintos. Essa característica em particular, possui uma forte dependência do objeto de escrita e da textura do papel utilizado. Na Tabela 5(e) se observa o nível de progressão da escrita. A progressão estabelece que, quanto mais veloz for a execução da escrita, mais ilegível ela se torna e quanto mais lento, mais legível se torna o texto. Na Tabela 5(f) se observam diferentes características relativas ao ataque e ao remate. No primeiro caso, o escritor apresenta um traçado com início e final usualmente espessos, enquanto no outro, são ambos usualmente pontiagudos. Na Tabela 5(g) é possível observar o grau de destreza ou desenvolvimento apresentado por escritores distintos. O escritor com maior destreza apresentará trações firmes e regulares, já o com baixa destreza, apresentará um traçado heterogêneo e vacilante. Na Tabela 5(h) se observam as diferenças de inclinação axial entre textos de escritores diferentes. Essa angulação pode oscilar para mais ou para menos no transcorrer da escrita, no entanto, sempre haverá uma predominância angular. Por fim, os mínimos gráficos, por serem componentes isolados, acabam sendo incorporados pelo processo de segmentação não contextual, juntamente com os momentos.

Tabela 5. Exemplos dos atributos genéticos e genéricos observáveis após o processo de segmentação.

a) Calibre			e) Progressão		
b) Valores curvilíneos e angulares			f) Ataques e remates		
c) Proporcionalidade			g) Desenvolvimento		
d) Pressão			h) Inclinação axial		

4.3.2 Processamento para 16, 8, 4 e 2 níveis de cinza

As cartas digitalizadas originais da base PUC-PR possuem 256 níveis de cinza. Após as imagens novas serem geradas a partir da original, em que o conteúdo da escrita está concentrado, também em 256 níveis de cinza, serão criadas a partir destas, novas imagens em 16, 8, 4, 2 níveis de cinza, para que sejam trabalhadas as matrizes de co-ocorrência. O aplicativo *IrfanView* [SKILJAN, 2008] irá auxiliar esta etapa e realizar a conversão dos níveis de cinza através de um processo *batch*.

Os quatro valores de níveis de cinza foram determinados com o objetivo de comparação de resultados. Após a etapa de experimentos, foi possível obter melhores resultados com dois níveis de cinza, devido a redução da complexidade da imagem em relação aos níveis de cinza que conseqüentemente reduzem a dispersão dos valores na matriz de co-ocorrência.

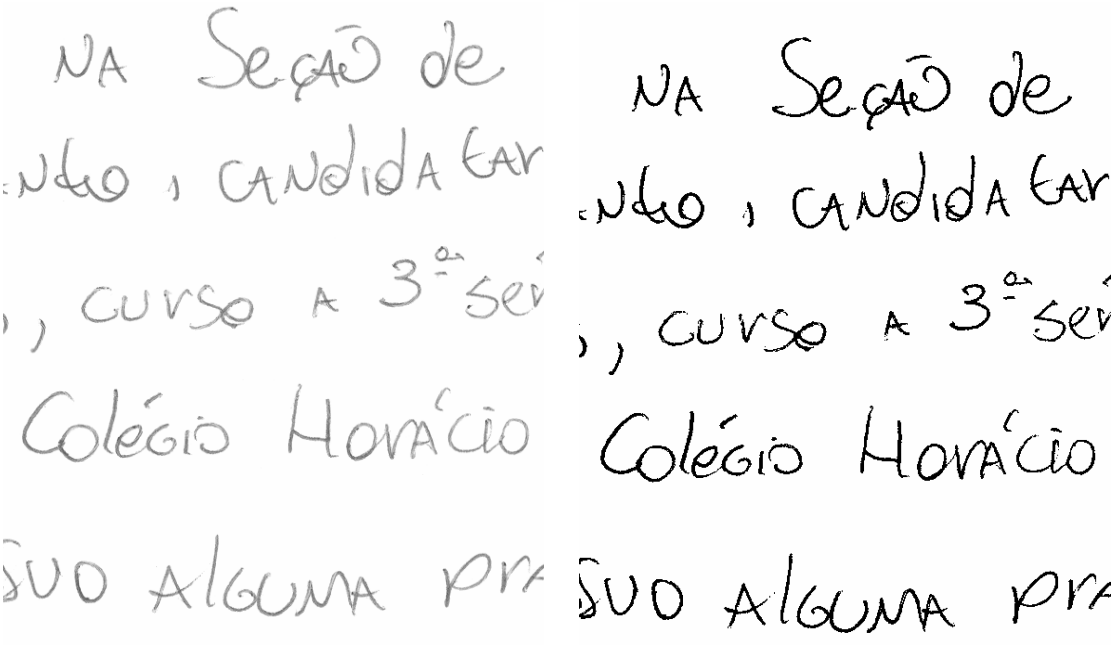
Para as imagens em 16 níveis de cores corresponde à uma imagem de 4 bits, em que o valor 0 é a cor preta e o valor 15 corresponde a cor branca. No cálculo da matriz de co-ocorrência, seu tamanho é indicado pela quantidade de níveis de cinza. Por exemplo, para os 16 níveis de cinza, um fragmento possui uma matriz de co-ocorrência de tamanho

16 x 16. Este pré-tratamento é aplicado anteriormente à etapa de análise do traçado por meio da matriz de co-ocorrência.

4.3.3 Binarização

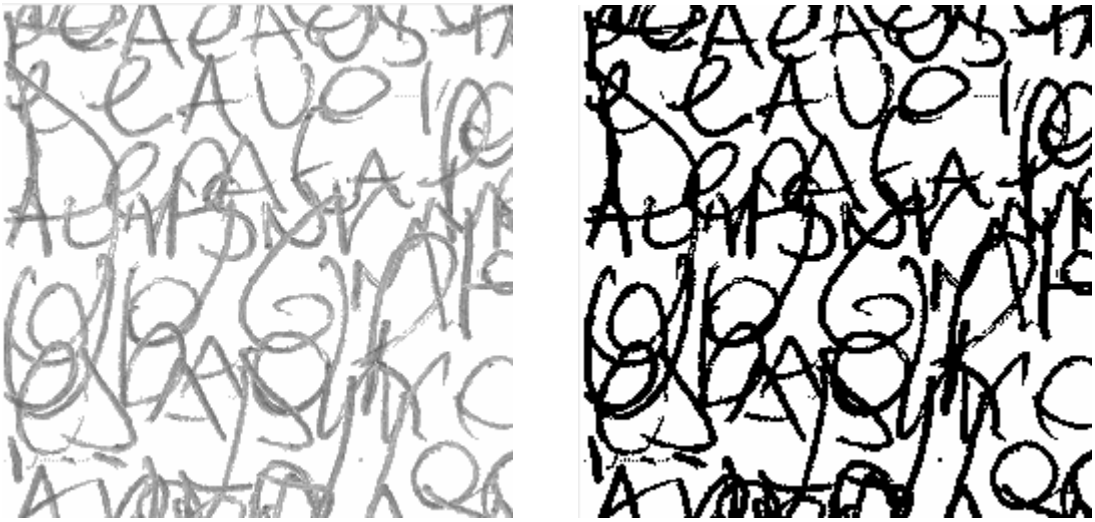
A binarização consiste no processo de transformação de uma imagem em 256 níveis de cinza, a partir de um novo limiar, em uma imagem binária, ou seja, preto (*pixel* com valor 0) e branco (*pixel* com valor 255) (Figura 13 e Figura 14). Esta transformação implica na redução da quantidade de dados a serem tratados, eliminando ruídos e facilitando a extração de componentes relevantes da imagem alem de reduzir drasticamente o tamanho da imagem em *bytes* e foi aplicada nesta abordagem, tanto para a extração de característica de textura, como a de inclinação axial.

Para esta abordagem a binarização global de Otsu foi a selecionada, pois dentre as várias técnicas utilizadas, foi a que apresentou melhor resultado melhores resultados, sem causar a perda de informação nas imagens [BARANOSKI, 2005].



(a) (b)

Figura 13. Fragmento de um manuscrito da base original PUC-PR; (a) Fragmento selecionado em 256 níveis de cinza; (b) O mesmo fragmento após ser binarizado pelo método Otsu.



(a) (b)

Figura 14. Fragmento de um manuscrito da base compactada PUC-PR; (a) Fragmento selecionado em 256 níveis de cinza; (b) O mesmo fragmento após ser binarizado pelo método Otsu.

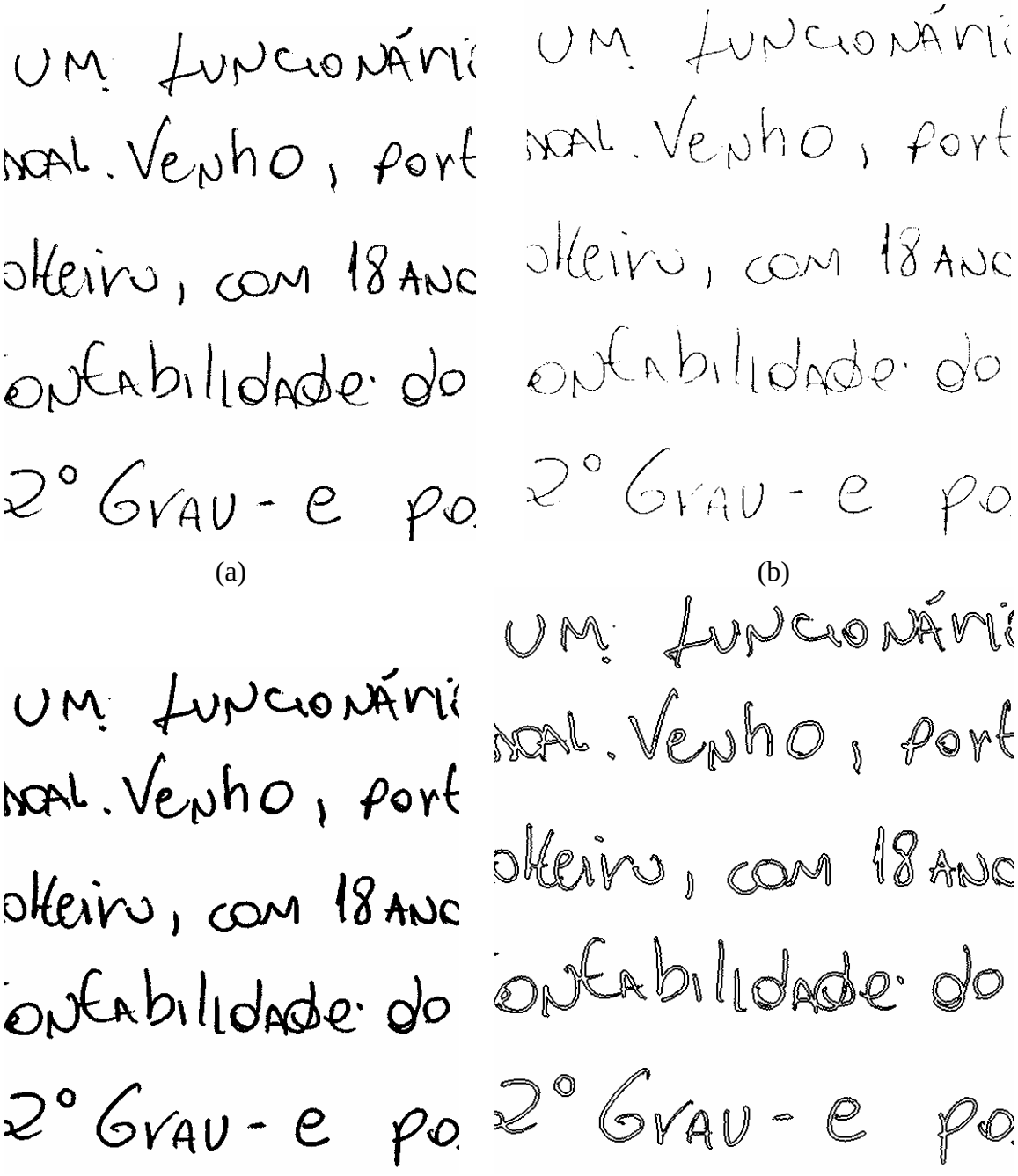
A binarização é utilizada apenas para a etapa de preparação da base, em que são criadas as imagens com conteúdo da escrita concentrado. Para os cálculos das matrizes de co-ocorrência e extração de características apenas o pré-tratamento de redução para 16, 8, 4 e foi realizado pelo aplicativo *IrfanView* [SKILJAN, 2008] e a binarização para dois níveis de cinza foi utilizado o algoritmo implementado no trabalho (método Otsu).

4.3.4 Detecção das bordas da escrita por Dilatação e Erosão

O pré-tratamento de detecção de bordas por dilatação e erosão utiliza a morfologia matemática, e a partir deste, são extraídos contornos bem definidos do traçado da escrita.

Primeiramente as imagens passam pelo pré-tratamento de binarização. Após, o filtro morfológico de dilatação modifica a imagem de um manuscrito através de um elemento estruturante em cruz, deixando o traçado da escrita mais espesso, devido a incrementação dos *pixels* nas bordas das imagens, enquanto o processo morfológico de erosão faz o processo inverso, deixando o traçado do autor mais fino, devido a

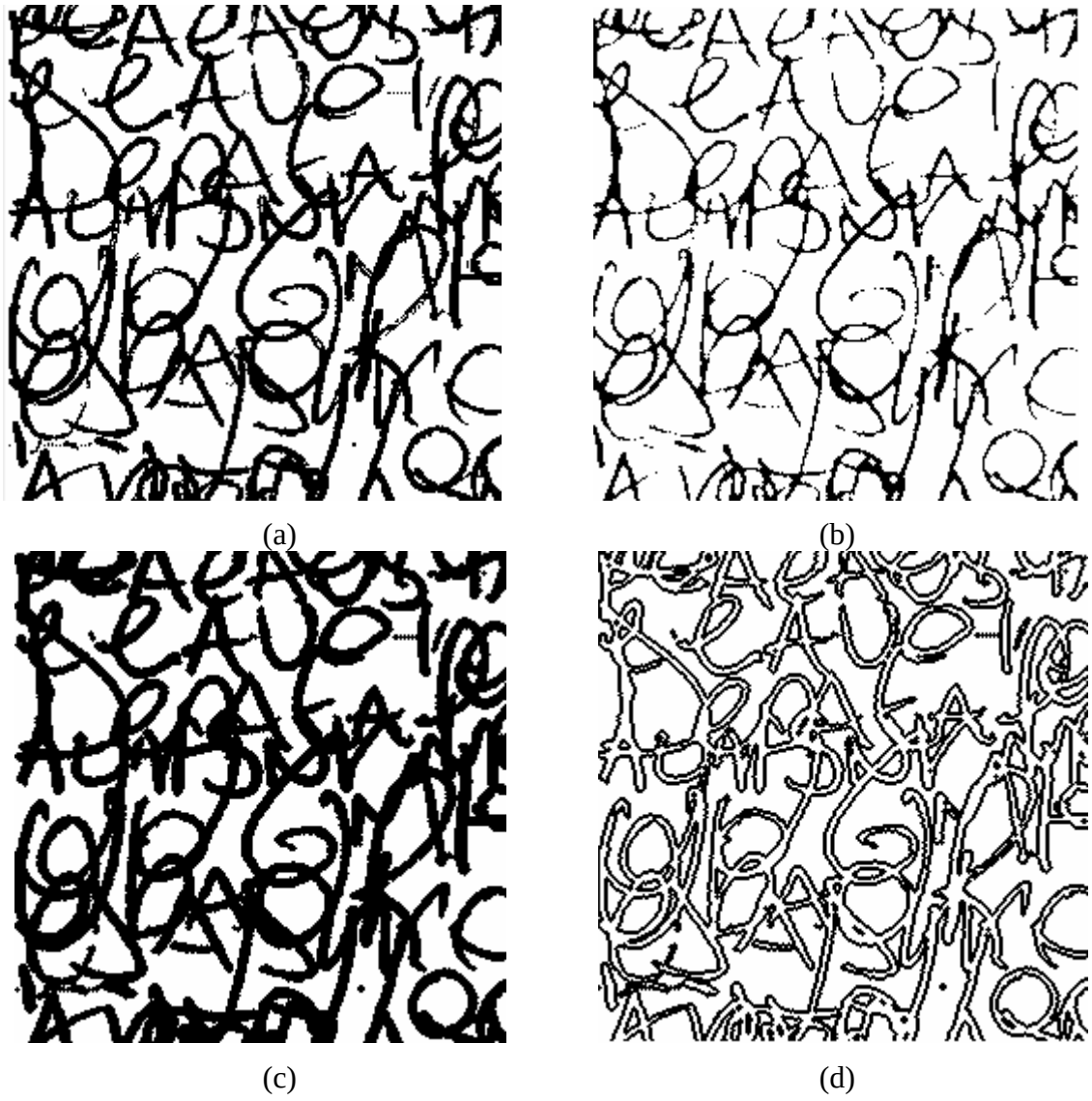
decrementação de *pixels* nas bordas da imagem. A imagem com a borda é obtida através da sobreposição da imagem dilatada com a imagem erodida em que é feita a subtração de *pixels*. (Figura 15 e Figura 16).



(a) (b)

(c) (d)

Figura 15. Fragmento de um manuscrito da base original PUC-PR (a) Fragmento binarizado; (b) fragmento erodido; (c) fragmento dilatado; (d) fragmento com bordas extraídas.



(a) (b)

(c) (d)

Figura 16. Fragmento de um manuscrito da base compactada PUC-PR (a) Fragmento binarizado; (b) fragmento erodido; (c) fragmento dilatado; (d) fragmento com bordas extraídas.

A detecção das Bordas por Dilatação e Erosão é aplicada somente para a extração da inclinação axial.

4.3.5 Divisão do manuscrito em fragmentos

Nas abordagens locais a segmentação consiste em dividir o manuscrito visando retirar partes de interesse, como palavras ou letras. Possui a vantagem de possibilitar a análise de particularidades da escrita, consideradas discriminantes pela grafoscopia, podendo ser observadas apenas localmente, tais como pingos da letra “i, cortes da letra “t, dentre outras, vistas no Capítulo 2. Porém a desvantagem encontra-se na dificuldade de se implementar uma solução computacional para essa finalidade, a qual é feita usualmente de forma manual.

Para as abordagens globais, geralmente a divisão é feita automaticamente, em fragmentos de textos para a extração de características. A vantagem deste tipo de segmentação encontra-se na simplicidade do processo, enquanto que a desvantagem encontra-se no fato de não permitir uma abordagem contextual de extração de características (uso de palavras ou letras).

Este pré-tratamento é aplicado anteriormente à etapa de análise do traçado por meio da matriz de co-ocorrência e extração de características de inclinação axial e descritores de Haralick.

Para a aplicação da extração de característica de inclinação axial, o manuscrito da base original PUC-PR, de tamanho 2448 x 3760, é divido em 24 fragmentos de tamanho 637 x 589 (Figura 17).

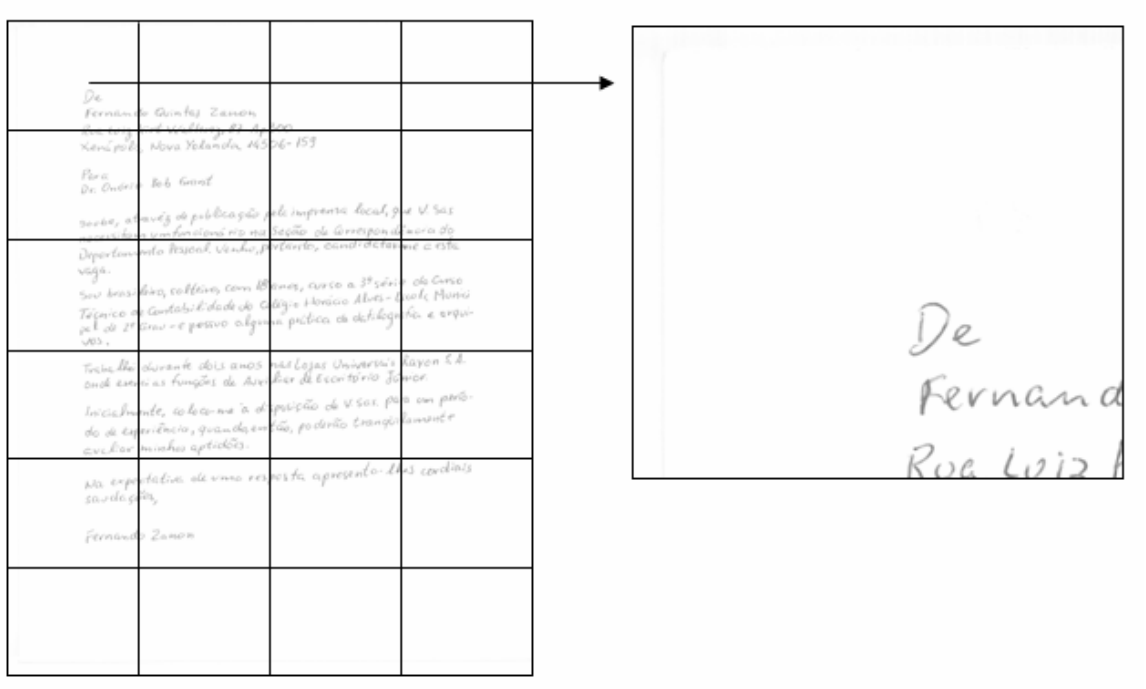


Figura 17. Exemplo de manuscrito segmentado da base original PUC-PR.

Já para a aplicação da extração de característica a partir da base compactada, foi determinado o menor valor da altura do bloco compactado, entre todas as imagens da base transformada. Depois de encontrado o valor da altura, foi determinado o tamanho do fragmento, de 256 x 256, sendo possível retirar 9 fragmentos de cada bloco de escrita compactada de uma imagem de manuscrito (Figura 18).

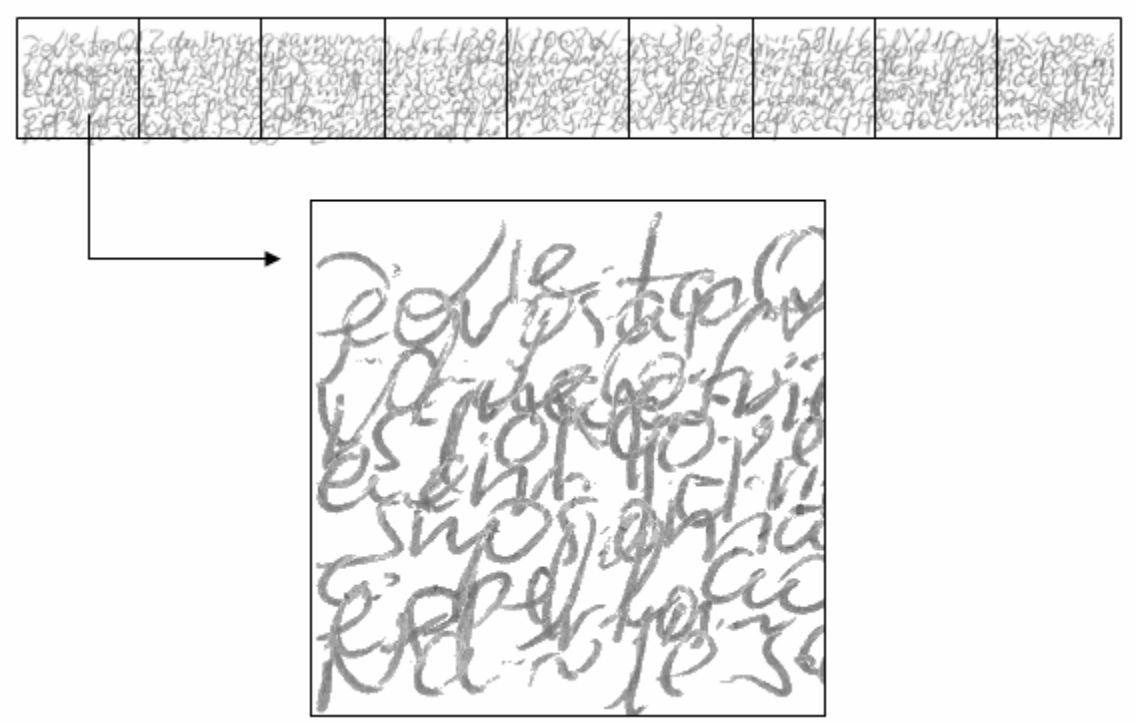


Figura 18. Exemplo de manuscrito segmentado da base compactada PUC-PR.

4.4 Extração de Características

A abordagem deste trabalho tem como foco a implementação da extração de característica de textura do traçado da escrita, para verificação de autoria de manuscritos, que vem sido aplicada nas abordagens recentes [FRANKE,2002], [SAID, 1998], [AL-DMOUR, 2005], [BUSCH, 2005]. Devido às características de textura, foi aplicado o pré-tratamento de segmentação do texto na base original PUC-PR. Com isto, a técnica de extração da característica de inclinação axial de Baranoski [BARANOSKI, 2005] foi re-implementada para ser aplicada na nova base, em busca de resultados melhores.

4.4.1 GLMC e Descritores de Haralick

Os descritores de Haralick foram escolhidos devido às propriedades que os mesmos possuem no trato de textura onde a irregularidade dos padrões é predominante. Para a extração das características, que se tratam dos descritores de Haralick [HARALICK, 1973] é realizada a análise da textura do traçado que implica primeiramente na obtenção das matrizes de co-ocorrência, a partir das imagens em níveis de cinza.

Primeiramente, para cada fragmento é criada uma matriz da imagem *MI* de acordo com o tamanho do fragmento (256 x 256). Cada posição contém o valor do nível de cinza do *pixel* da imagem.

De acordo com a quantidade de níveis de cinza, o tamanho da matriz de co-ocorrência é determinado, ou seja, se um fragmento está em 4 níveis de cinza, este terá um matriz correspondente de tamanho 4 x 4. Após a divisão dos blocos de textura em fragmentos, são extraídas vinte matrizes de co-ocorrência para cada fragmento, através da combinação de quatro direções θ = 0°, 45°, 90° e 135° e cinco distâncias d=1, 2, 3, 4, 5, sendo estas resoluções as mais utilizadasm pela literatura [PERVOUCHINE, 2007][BUSCH, 2005].

Uma matriz de co-ocorrência é obtida através dos seguintes cálculos:

1. Sendo os 4 níveis de cinza da imagem, uma matriz de co-ocorrência $M_{i,j}$ de tamanho i=4 x j=4 é criada.
2. A matriz de co-corrência é preenchida realizando uma comparação: seja, por exemplo, determinado ângulo 0° e distância 1: na matriz *MI* original dos níveis de cinza, para cada *pixel*, é capturado o valor do *pixel* e o valor do *pixel* na direção 0° e distância 1 em relação ao *pixel* capturado.
3. O valor (x) do nível de cinza do *pixel* principal e o valor do nível de cinza (y) de seu vizinho na direção 0° e distância 1, corresponde a um ponto adicionado na posição de i e j respectivamente na matriz de co-ocorrência:

$$M[i][j] = M[i][j] + 1 \qquad (18)$$

4. Para cada matriz calculada para as direções θ = 0°, 45°, 90° e 135°, logo em seguida, o mesmo cálculo é aplicado novamente a partir da matriz original de níveis de cinza, mas com base no complemento do ângulo de θ°, para a mesma distância, incrementando os valores na mesma matriz de co-ocorrência $M_{i,j}$, conforme a combinação de níveis de cinza encontrados, que indicam uma posição em $M_{i,j}$.

Nesse estudo foram utilizados 4 (quatro) diferentes resoluções, ou níveis de cinza, 2, 4, 8 e 16, com o objetivo de estabelecer um comparativo entre resultados, dado o acréscimo de conteúdo proporcionado pelos diferentes níveis de cinza. Abaixo segue o pseudo-código para obter a matriz de co-ocorrência:

```
METODO MAIN()

MATRIZ_IMAGEM_ORIGINAL = MATRIZ INTEIRO[TAMANHO_IMAGEM][TAMANHO_IMAGEM]

PARA (I=0; I<MATRIZ.TAMANHO; I++)

        PARA J (J<MATRIZ.TAMANHO; J++)

                MATRIZ_IMAGEM_ORIGINAL[I][J] = VALOR PIXEL[I][J] DA IMAGEM

                PARA CADA ANGULO (0, 45, 90, 135)

                        PARA CADA DISTANCIA(1, 2, 3, 4, 5)

                        MATRIZ_CO_OCORRENCIA = MATRIZ INTEIRO[QTDE_NIVEIS_CINZA][QTDE_NIVEIS_CINZA]

                        SE(ANGULO = 0)

                                CALCULA_MATRIZ_ANGULO_0 (DISTANCIA)

                        SE(ANGULO = 45)

                                CALCULA_MATRIZ_ANGULO_45 (DISTANCIA)

                        SE(ANGULO = 90)

                                CALCULA_MATRIZ_ANGULO_90 (DISTANCIA)

                        SE(ANGULO = 135)

                                CALCULA_MATRIZ_ANGULO_135 (DISTANCIA)

                        NORMALIZA_MATRIZ()


METODO CALCULA_MATRIZ_ANGULO_0(DISTANCIA)

        PARA (I=0; I<MATRIZ_IMAGEM_ORIGINAL.TAMANHO; I++)

                PARA  (J=0; J<MATRIZ_IMAGEM_ORIGINAL.TAMANHO - DISTANCIA; J++)

                        INTEIRO A = MATRIZ_IMAGEM_ORIGINAL [I][J]

                        INTEIRO B = MATRIZ_IMAGEM_ORIGINAL [I][J+DISTANCIA]

                        PREENCHE_MATRIZ_CO_OCORRENCIA(A, B)

        //CALCULO PARA COMPLEMENTO DO ANGULO 0, 180

        PARA (I=0; I<MATRIZ_IMAGEM_ORIGINAL.TAMANHO - DISTANCIA; I++)

                PARA (J=MATRIZ_IMAGEM_ORIGINAL.TAMANHO - 1; J>=DISTANCIA; J--)

                        INTEIRO A = MATRIZ_IMAGEM_ORIGINAL [I][J]

                        INTEIRO B = MATRIZ_IMAGEM_ORIGINAL [I][J-DISTANCIA]

                        PREENCHE_MATRIZ_CO_OCORRENCIA(A, B)
```

```
METODO CALCULA_MATRIZ_ANGULO_45(DISTANCIA)
        PARA (I=DISTANCIA; I<MATRIZ_IMAGEM_ORIGINAL.TAMANHO; I++)
                PARA  (J=0; J<MATRIZ_IMAGEM_ORIGINAL.TAMANHO - DISTANCIA; J++)
                        INTEIRO A = MATRIZ_IMAGEM_ORIGINAL [I][J]
                        INTEIRO B = MATRIZ_IMAGEM_ORIGINAL [I-DISTANCIA][J+DISTANCIA]
                        PREENCHE_MATRIZ_CO_OCORRENCIA(A, B)
        //CALCULO PARA O COMPLEMENTO DO ANGULO 45, 225
        PARA (I=0; I<MATRIZ_IMAGEM_ORIGINAL.TAMANHO - DISTANCIA; I++)
                PARA (J=DISTANCIA; J<MATRIZ_IMAGEM_ORIGINAL.TAMANHO; J++)
                        INTEIRO A = MATRIZ_IMAGEM_ORIGINAL [I][J]
                        INTEIRO B = MATRIZ_IMAGEM_ORIGINAL [I+DISTANCIA][J-DISTANCIA]
                        PREENCHE_MATRIZ_CO_OCORRENCIA(A, B)


METODO CALCULA_MATRIZ_ANGULO_90(DISTANCIA)
        PARA (I=DISTANCIA; I<MATRIZ_IMAGEM_ORIGINAL.TAMANHO; I++)
                PARA  (J=0; J<MATRIZ_IMAGEM_ORIGINAL.TAMANHO; J++)
                        INTEIRO A = MATRIZ_IMAGEM_ORIGINAL [I][J]
                        INTEIRO B = MATRIZ_IMAGEM_ORIGINAL [I-DISTANCIA][J]
                        PREENCHE_MATRIZ_CO_OCORRENCIA(A, B)
        //CALCULO PARA O COMPLEMENTO DO ANGULO 90, 270
        PARA (I=0; I<MATRIZ_IMAGEM_ORIGINAL.TAMANHO - DISTANCIA; I++)
                PARA (J=0; J<MATRIZ_IMAGEM_ORIGINAL.TAMANHO; J++)
                        INTEIRO A = MATRIZ_IMAGEM_ORIGINAL [I][J]
                        INTEIRO B = MATRIZ_IMAGEM_ORIGINAL [I+DISTANCIA][J]
                        PREENCHE_MATRIZ_CO_OCORRENCIA(A, B)
```

```
METODO CALCULA_MATRIZ_ANGULO_135(DISTANCIA)
        PARA (I=DISTANCIA; I<MATRIZ_IMAGEM_ORIGINAL.TAMANHO; I++)
                PARA  (J=DISTANCIA; J<MATRIZ_IMAGEM_ORIGINAL.TAMANHO; J++)
                        INTEIRO A = MATRIZ_IMAGEM_ORIGINAL [I][J]
                        INTEIRO B = MATRIZ_IMAGEM_ORIGINAL [I-DISTANCIA][J-DISTANCIA]
                        PREENCHE_MATRIZ_CO_OCORRENCIA(A, B)
        //CALCULO PARA O COMPLEMENTO DO ANGULO 135, 270
        PARA (I=0; I<MATRIZ_IMAGEM_ORIGINAL.TAMANHO - DISTANCIA; I++)
                PARA (J=0; J<MATRIZ_IMAGEM_ORIGINAL.TAMANHO - DISTANCIA; J++)
                        INTEIRO A = MATRIZ_IMAGEM_ORIGINAL [I][J]
                        INTEIRO B = MATRIZ_IMAGEM_ORIGINAL [I+DISTANCIA][J+DISTANCIA]
                        PREENCHE_MATRIZ_CO_OCORRENCIA(A, B)

METODO PREENCHE_MATRIZ_CO_OCORRENCIA()
        MATRIZ_CO_OCORRENCIA[I][J] = MATRIZ_CO_OCORRENCIA[I][J]+1

METODO NORMALIZA_MATRIZ()
        REALIZA SOMA DOS VALORES MATRIZ_CO_OCORRENCIA
        REALIZA A MÉDIA PARA CADA UM DOS VALORES DA MATRIZ_CO_OCORRENCIA, A PARTIR DA SOMA
```

Após obtenção das 20 matrizes de co-ocorrência, são extraídas as características de textura através dos descritores. Os descritores propostos por Haralick [HARALICK, 1973], são descritores estatísticos de textura cujas propriedades buscam atender aos casos em que as mesmas apresentam padrões irregulares. O conjunto de descritores é composto por 14 (quatorze) descritores que buscam determinar estatisticamente as propriedades de distribuição e o relacionamento entre tons de cinza contidos em uma textura. Os descritores são calculados através do uso de matrizes de co-ocorrência *P*, obtidas da imagem da textura *I* em análise. Do conjunto de quatorze descritores foram selecionados os 6 (seis) que obtiveram os melhores resultados nos testes individuais preliminares. São

eles: 2º. momento angular, entropia, homogeneidade, dissimilaridade, variância inversa, energia.

Sendo *n* e *m*, respectivamente, o número de linhas e colunas da matriz de co-ocorrência e *p(i,j)*, o valor normalizado da célula da matriz de co-ocorrência, obtido através da divisão de todos os componentes *P(i,j)* pelo somatório de todos os valores das células da matriz de co-ocorrência. Os valores de σ_i e σ_j representam respectivamente, o desvio padrão na direção *i* e na direção *j*. Os valores μ_i e μ_j representam respectivamente, a média na direção *i* e na direção *j*.

Os vetores de características **V**, para cada descritor, utilizados nas fases de aprendizado e verificação, foram obtidos através das matrizes de co-ocorrências, com a variação das distâncias *d* e dos ângulos θ. Isto é, cada vetor de característica dos descritores é composto por um conjunto de 20 valores e esse vetor é normalizado pela média dos seus valores.

A representação de um método computacional reflete diretamente nos resultados obtidos, nos quais a robustez do método é diretamente proporcional à qualidade das características, caracterizando-as, assim, em uma fase de grande importância.

A combinação de características é um estudo muito relevante, pois através destes, podemos identificar o comportamento de uma característica: se ela traz bons ou maus resultados ou aplicada isoladamente ou em conjunto. Nem sempre a combinação de características que apresentam bons resultados isoladamente pode trazer maiores resultados devido a combinação, sendo que podem ser obtidos resultados de baixo valor. Nos experimentos realizados foram combinadas características de textura através do classificador SVM. Os resultados podem ser observados no capítulo cinco.

4.4.2 Inclinação Axial

A técnica aplicada para extração da característica de inclinação axial, realizada por Baranoski [BARANOSKI, 2005], foi re-implementada e aplicada neste trabalho, porém utilizou-se a base PUC-PR com texto compactado, a mesma utlilizada para a extração de características de textura. A intenção foi verificar as diferenças de resultados entre as bases e determinar resultados mais satisfatórios.

A inclinação axial é uma característica grafocinética que descreve o aspecto dinâmico do traçado e o ângulo de inclinação da escrita. Esta técnica considera as bordas do traçado, pois estas obtiveram melhores resultados na discriminação da inclinação axial do autor em relação à inclinação extraída diretamente do traçado.

A extração de características envolve o seguinte processo: uma imagem pré-processada e segmentada é representada por uma imagem de borda na qual apenas os *pixels* desta borda estão em preto. A imagem então é percorrida considerando-se o *pixel* da borda do traçado no centro do elemento estruturante retangular (Figura 19). Em seguida, verificam-se os fragmentos de borda em todas as direções, partindo deste *pixel* central e conferindo os *pixels* posteriores com um operador lógico AND, finalizando as extremidades do elemento estruturante apenas se houver a presença de um fragmento de borda inteiro. Ou seja, se todos os *pixels* vizinhos forem pretos, considera-se o fragmento da borda e computa-se a posição do fragmento em um vetor de posições para a construção do histograma que determina à inclinação que pode ser à esquerda, à direita ou nula.

O algoritmo implementado para a proposta implicará elemento estruturante $k = 5$ (distância de cinco *pixels* a partir do central, incluindo-o) ao longo do fragmento da borda, no qual para cada elemento estruturante são quantificadas 17 direções de inclinação (ângulos: 0°, 11°, 23°, 34°, 45°, 56°, 68°, 79°, 90°, 102°, 113°, 124°, 135°, 146°, 158°, 169°, 180°) que também representam a dimensionalidade do vetor final de características.

Por exemplo, para um elemento estruturante com comprimento k=5 e L = 17 direções, partindo de um *pixel* central de posição i=4 e j=4, com ângulo $\theta = 45°$, caso os próximos 4 *pixels* interligados sejam pretos, isto indica uma nova quantidade para a direção correspondente ao ângulo de 45° (Figura 19).

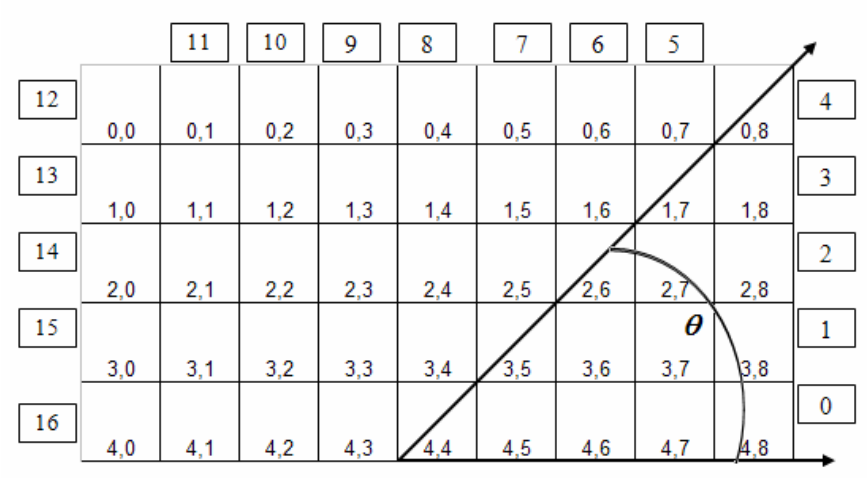


Figura 19. Exemplo de um elemento estruturante com comprimento *k*=5 e L = 17 direções.

Desta forma, é possível identificar a direção da escrita, de acordo com a quantidade de pontos das direções. Se uma imagem possui uma maior quantidade de pontos nos ângulos obtusos, isto significa que o autor tem sua escrita com inclinação à esquerda, caso contrário, à direita.

Esta característica grafocinética demonstrou uma boa capacidade discriminatória na abordagem de Baranoski, e melhores resultados nesta abordagem, quando aplicada a nova base de texto compactado.

4.5 Comparação

O processo de comparação é composto por duas fases, o treinamento e a verificação. Este processo depende significativamente da métrica de distância, ou seja, é necessária a escolha de uma medida de distância adequada ao problema proposto. Os resultados da distância Euclidiana mostraram resultados satisfatórios em relação aos manuscritos, sendo que esta foi selecionada para ser calculada utilizando as características extraídas a partir da inclinação axial e descritores de Haralick.

Um conjunto de características é obtido de acordo com os fragmentos do manuscrito, sendo esse conjunto um vetor f_v e os fragmentos do manuscrito de referência representados por $M_{ki} = (f_1, f_2, ..., f_L)$ e do manuscrito questionado M_Q, equações:

$$fv_{ki} = (f_1, f_2, ..., f_L) \qquad (19)$$

$$fv_Q = (f_1, f_2, ..., f_L) \qquad (20)$$

em que f_v são os conjuntos de características e L o número máximo de células de cada característica. O vetor de distâncias $D_{i(i = 1,2,..., n)}$ entre os fragmentos dos manuscritos de referência e o manuscrito questionado será computado para se obter a entrada do classificador, nesta abordagem proposta, o SVM, no treinamento e verificação.

$$D_{i(i = 1,2,..., n)} = \sqrt{(fv_{ki} - fv_Q)^2} \qquad (21)$$

No estágio de treinamento, as medidas das distâncias entre os vetores de características são calculadas entre pares de fragmentos de textos. Dado um fragmento do manuscrito genuíno *x* e o outro questionado *y*, aplica-se a técnica de extração de características de textura sobre as duas imagens, gerando um vetor de características, de 20 posições (1 característica x 20 matrizes de co-ocorrência) para cada fragmento. O vetor de características contendo em ordem, as características para cada matriz de co-ocorrência do fragmento *x* e *y* é representado por $(V_1^x ....V_{20}^x)$ e $(V_1^y ....V_{20}^y)$. Após, extraídas as características, são computadas as distâncias entre os vetores de fragmentos, que serão utilizados como entrada na produção de um modelo (Figura 20).

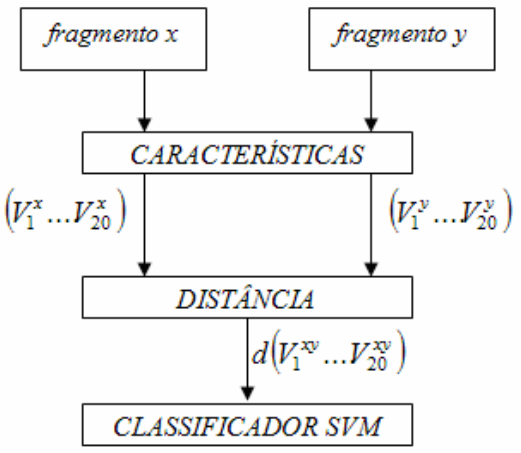


Figura 20. Diagrama do processo de cálculo da distância Euclidiana.

Para a característica de inclinação axial, o vetor de características possui 17 posições que correspondem aos 17 ângulos determinados.

Neste trabalho foram utilizadas 3 amostras por autor para treinamento e 5 amostras por autor para os testes.

Quando dois fragmentos pertencerem a um mesmo escritor, o vetor de característica é indicado com 1 (associação). Quando dois fragmentos pertencerem a escritores diferentes, o vetor de característica é indicado com -1 (dissociação). A distância **E** entre dois fragmentos de texto é considerada pequena, quando as amostras pertencerem a um mesmo escritor. O SVM é treinado então, para separar pequenas distâncias entre características (associação) e grandes distâncias entre características (dissociação).

No estágio de verificação, o SVM possui duas saídas. A primeira é composta pelos fragmentos pertencentes a um mesmo escritor, classe w_1. A segunda é composta por fragmentos pertencentes a escritores distintos, classe w_2.

4.6 Decisão

Partindo da hipótese de que as medida das distâncias entre os vetores de características extraídos de amostras de um mesmo autor são menores entre si, e de que a medida de distância entre os vetores de características de autores diferentes são maiores entre si, o *SVM* é então, treinado para separar pequenas distâncias entre características considerando-as como associação, e distâncias maiores entre características como dissociação.

O processo de decisão é avaliado conforme o modelo produzido, verificando se o manuscrito deve ser considerado como pertencente ou não a um determinado autor. O processo de decisão em relação à autoria é gerado na saída do classificador. Porém, para que o método proposto seja baseado na visão pericial, a saída do classificador torna-se apenas uma entrada parcial para a etapa de determinação do voto majoritário, ou de fusão de resultados. Com o objetivo de gerar a decisão final, o método proposto classifica as saídas do SVM utilizando como regra de fusão. A literatura apresenta várias possibilidades para o processo de fusão, tais como a média dos valores, o valor máximo, o valor mínimo, a soma dos valores e o voto majoritário [BERTOLINI, 2008]. Os

experimentos demonstraram que nesse caso, os melhores resultados foram obtidos através da soma.

Usualmente, em uma prova pericial, o perito utiliza um conjunto de amostras de textos de origem conhecida. Cada amostra conhecida, pertencente ao conjunto de referência *R* (4 a 10 amostras), é comparada com a amostra de autoria desconhecida ou questionada *Q*. Nos experimentos, foi usado um conjunto com 5 (cinco) amostras de referência para cada escritor.

Na abordagem proposta, conforme os modelos vistos, uma comparação é feita entre o manuscrito questionado e os manuscritos conhecidos, com fragmentos do mesmo autor, gerando a classe w_1 e com fragmentos de autores diferentes, gerando a classe w_2.

Um algoritmo que realiza uma regra de decisão do voto majoritário é aplicado sob as saídas do classificador SVM. Sua fórmula é dada por:

$$Vm = \frac{Nref + 1}{2} \qquad (22)$$

em que *Nref* é o número de manuscritos de referência e *Vm* é o valor do voto majoritário, no qual a decisão final baseia-se em $Vm >= 3$ para 5 amostras, dependendo da classe w_1 ou w_2.

4.7 Comentários finais

Este capítulo apresentou as etapas de um método proposto neste trabalho para a um processo de verificação de autoria de manuscritos. Mostra-se que a etapa de pré-tratamento é muito relevante, com a técnica de segmentação do texto para as características de textura, e as técnicas de binarização, erosão, dilatação do texto escrito e extração da borda da escrita, para a característica de inclinação axial. As outras etapas como extração de características, comparação e decisão, fazem parte do método e faz com quem este se assemelhe a um modelo de visão pericial e que possa ser considerado seguramente como uma técnica computacional semi-automática.

Capítulo 5

Experimentos e Análise de Erros

Este capítulo contém os resultados obtidos através dos testes, realizados de acordo com as regras de divisão de base PUC-PR, original e compactada, e o protocolo utilizado para quantidade de escritores nos treinamentos e testes. A classificação feita pelo SVM e análise de resultados pelos votos são apresentados e comparados, para as duas diferentes características, de textura e inclinação axial.

No mérito de uma acusação pericial, deve-se tomar a decisão com a maior certeza possível, garantindo a confiabilidade da mesma. Partindo desta premissa, a rejeição é mais importante que a aceitação. A abordagem deste trabalho busca reduzir as taxas de falsa rejeição e falsa aceitação, sendo mais relevante que as taxas de falsa rejeição sejam preferencialmente maiores que as de falsa aceitação.

5.1 Base de Manuscritos

A base de manuscritos utilizada é composta por 315 escritores. Cada escritor redigiu três manuscritos de mesmo teor, totalizando 945 documentos digitalizados. O conteúdo da carta foi elaborado de forma a contemplar o conjunto de letras do alfabeto da Língua Portuguesa, tanto minúscula quando minúscula. Assim também procedendo com os símbolos utilizados nas acentuações, tais como o til, cedilha, acento circunflexo, acento grave, acento agudo e pingo nos i´s.

A divisão da base de dados consiste na separação de dois grupos de manuscritos distintos, o conjunto de autores para treino e o conjunto de autores para teste. A base de treino é utilizada para geração do modelo de aprendizado do classificador, enquanto a base de testes é utilizada para validação do método proposto.

A divisão da base foi realizada com objetivo de permitir o treinamento dos modelos e a execução dos testes sem que os escritores de um conjunto participassem do

outro, por se tratar de um modelo global de classificação. Desta forma o classificador busca autenticar autores nunca vistos anteriormente.

Assim procedendo, ficaram 115 autores para testes e 200 para o treinamento dos modelos. O conjunto de 200 autores para o treinamento foi novamente dividido em três subconjuntos, cujo número de autores foi incrementado segundo um critério exponencial, isto é, 50, 100, 200 escritores. Este critério foi adotado, a fim de estabelecer a influência do número de autores no desempenho do modelo, por se tratar de um modelo WI. Os resultados demonstraram uma variação em torno de 1% de incremento do erro.

Para os testes envolvendo as características de textura, cada manuscrito da base compactada PUC-PR foi dividido em 9 fragmentos, sendo um total de 27 fragmentos por autor. Para os testes envolvendo a característica de inclinação axial, o manuscrito da base original PUC-PR foi dividido em 24 fragmentos, sendo um total de 72 fragmentos por autor. Foram utilizadas 3 fragmentos para treinamento, 5 para referência e 5 para teste, sem que nenhum se repita em um dos três tipos de seleção.

5.1.1 Treinamento

.

A escolha de três fragmentos de manuscritos por autor decorre da abordagem adotada na qual o classificador deverá ser treinado para generalizações, sendo sensível às variações intrapessoais e intolerante às similaridades interpessoais. Esse processo simula uma situação real em que o perito depara-se com um número restrito de exemplares para observar as particularidades de cada autor. O valor três seria o mínimo necessário para que o mesmo pudesse executar o processo de análise com o mínimo de confiabilidade [JUSTINO, 2002].

Assim como na abordagem de Baranoski [BARANOSKI, 2005], para a geração da classe *w1*, considerando a base segmentada, três espécimes de cada autor são selecionadas e são computadas as distâncias Euclidianas dos mesmos, dois a dois. A escolha aleatória entre espécimes do mesmo autor visa conseguir uma representação da variabilidade intrapessoal no modelo. Já para a geração da classe *w2*, para um dado autor, são escolhidos três fragmentos de manuscritos de autores diferentes. Um total de 6 combinações por autor é gerado, divididas entre as classes *w*1 e *w*2.

Desta forma, gera-se um arquivo do modelo que será a entrada no classificador *SVM*.

5.1.2 Testes

Na análise pericial é comum o perito deparar-se com poucas amostras no processo de comparação. Usualmente, o perito utiliza amostras de textos de autoria conhecida. Cada amostra conhecida, pertencente ao conjunto de referência (usualmente de 4 a 10 amostras), é comparada com a amostra da autoria questionada.

Desta maneira, a base de testes usa um conjunto com cinco espécimes de referência para cada autor e cinco espécimes de manuscritos questionados para gerar o modelo de teste. Para a geração da classe *w1*, são usados espécimes de mesmo autor, em para cada um dos 5 fragmentos de teste são calculadas as distâncias euclidianas em relação aos 5 fragmentos de referência do mesmo autor. Já para a geração da classe *w2* são usados autores distintos. No processo de verificação é comparado cada um dos fragmentos de teste do autor desconhecido contra 5 referências de um autor conhecido. Um total de 50 combinações por autor é gerado, divididas entre as classes *w*1 e *w*2.

Para a classificação, o pacote *freeware SVMlight* é utilizado para etapas de treinamento e teste, gerando modelos e saídas. O objetivo é reduzir a taxa de erro total dos testes que se dividem em dois tipos:

- Erro de falsa rejeição: é quando o manuscrito de entrada é membro da classe (w_1) e é incorretamente classificado como membro da classe.
- Erro de falsa aceitação: caracteriza-se quando o manuscrito de entrada não é membro da classe (w_2) e é incorretamente classificado como membro da classe.
-

5.2 Experimentos com descritores de Haralick

Utilizando a base compactada PUC-PR foram testados os descritores de Haralick separadamente. Nesse caso, foram utilizados os fragmentos com 16, 8, 4 e 2 níveis de cinza. Os testes foram realizados com 25, 50, 100 e 200 escritores para o treinamento do

modelo e 115 escritores para os testes. Na etapa de decisão foi utilizada a soma como regra de fusão dos valores resultantes do SVM.

A Tabela 7 exibe os resultados de cada descritor para os 16 níveis de cinza. É possível observar que a entropia, que descreve o grau de dispersão de níveis de cinza da textura, obteve o melhor resultado 93,6% de acerto, seguida da energia, que mede a ordenação da textura, com 92,6% de acerto. A proximidade dos resultados entre estes dois descritores já era esperado, uma vez que ambos são considerados características mutuamente dependentes.

Tabela 7. Resultados obtidos com os descritores individualmente, com as imagens dos fragmentos em 16 níveis de cinza.

Descritor	**Erro Tipo I Falsa Rejeição (%)**	**Erro Tipo I Falsa Aceitação (%)**	**Erro Médio (%)**
2.o Momento Angular	5,57	9,39	7,48
Entropia	**4,00**	**8,70**	**6,35**
Homogeneidade	5,74	9,22	7,48
Dissimilaridade	12,70	41,91	27,30
Energia	**6,00**	**8,70**	**7,39**
Variância Inversa	7,48	24,87	16,17

Por se tratar de um modelo global, outro teste foi elaborado a fim de determinar a estabilidade do modelo, dado a variação do número de escritores no treinamento do modelo. Este procedimento busca identificar a contribuição que o número de escritores pode oferecer ao modelo gerado. Para esse teste foram usados três conjuntos de treinamento com 25, 50, 100 e 200 escritores, foi mantido o mesmo conjunto de 115 escritores nos testes, foi usado 2 níveis de cinza e a regra de fusão pela soma. Os resultados podem ser vistos na Tabela 8. Com o acréscimo do número de amostras de treinamento, observou-se uma pequena variação, em torno de 1%, nas taxas de erro, demonstrado que isoladamente, os descritores apresentam um comportamento estável.

Outra conclusão obtida de acordo com os resultados da Tabela 8, é que conforme aumenta-se o número de escritores para o treinamento, é possível verificar um aumento no erro tipo I, de falsa rejeição, porém uma redução no erro tipo II de falsa aceitação. Isto indica uma situação de um *trade-off* a ser escolhido, ou seja, de acordo com o tipo de escolha, perde-se em um aspecto e ganha em outro. Abaixo são apresentados os resultados para seis descritores divididos nas Tabelas 16.1 e Tabela 16.2, para dois níveis

de cinza, com a fusão de soma aplicada. A Figura 21 exibe a curva ROC, mostrando que a utilização de 200 escritores para o conjunto de treinamento, em dois níveis de cinza, apresenta menor taxa de média de erros, em relação às menores quantidades de escritores para treinamento.

Tabela 8. Resultados obtidos com a variação no número de escritores na geração do modelo para dois níveis de cinza, com regra de fusão pela soma.

Descritor	**Média dos Erros (%)**			
Número de escritores no treinamento do modelo	**25**	**50**	**100**	**200**
2.o Momento Angular	**5,65**	5,39	4,70	5,48
Entropia	5,83	5,13	5,04	5,91
Homogeneidade	6,35	6,69	5,91	6,61
Dissimilaridade	5,82	5,91	6,00	5,56
Energia	5,73	5,47	4,69	5,21
Variância Inversa	5,82	5,91	6,00	5,56

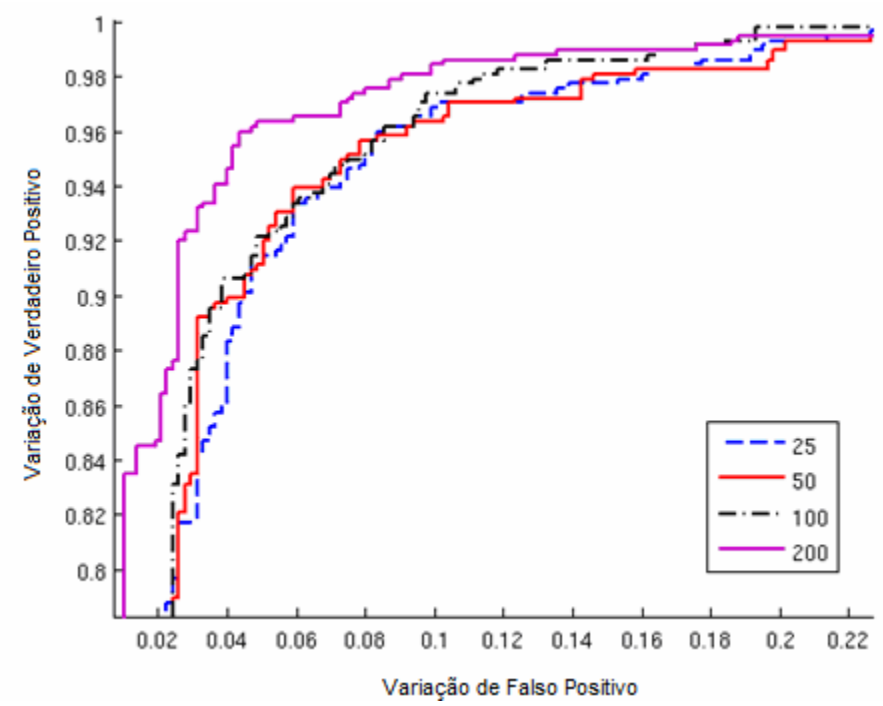


Figura 21. Curvas ROC comparando as diferentes quatidades de escritores para o conjunto de treinamento.

A fim de se averiguar a influência da quantidade de informação gerada pelos níveis de cinza, na produção das matrizes de co-ocorrência, repetiram-se os testes, agora variando os níveis de cinza dos fragmentos em 2, 4, 8 e 16 níveis. Para tanto, foi utilizado 50 escritores no treinamento e 115 para os testes. Os resultados podem ser vistos na Tabela 9. Os mesmos demonstram que a redução no número de níveis de cinza produz uma redução nos erros em relação aos produzidos com 4, 8 e 16 níveis. Isto é decorrência da redução da complexidade da imagem, no que se refere aos níveis de cinza, e da consequente redução da dispersão dos valores na matriz de co-ocorrência. A comparação entre os descritores para 2 níveis de cinza, utilizando 50 escritores para treinamento e 115 escritores para teste, é apresentada pela curva ROC na Figura 22.

Tabela 9. Resultados obtidos com a variação nos níveis de cinza dos fragmentos manuscritos para 50 escritores de treinamento e 115 para testes, com regra de fusão pela soma.

Descritor	**Média dos Erros (%)**			
Níveis de Cinza	**2**	**4**	**8**	**16**
2.o Momento Angular	**5,39**	6,69	6,78	7,48
Entropia	**5,13**	5,82	6,08	6,35
Homogeneidade	**6,69**	7,22	8,43	7,48
Dissimilaridade	**5,91**	6,96	27,04	27,30
Energia	**5,47**	6,61	6,69	7,39
Variância Inversa	**5,91**	6,87	19,21	16,17

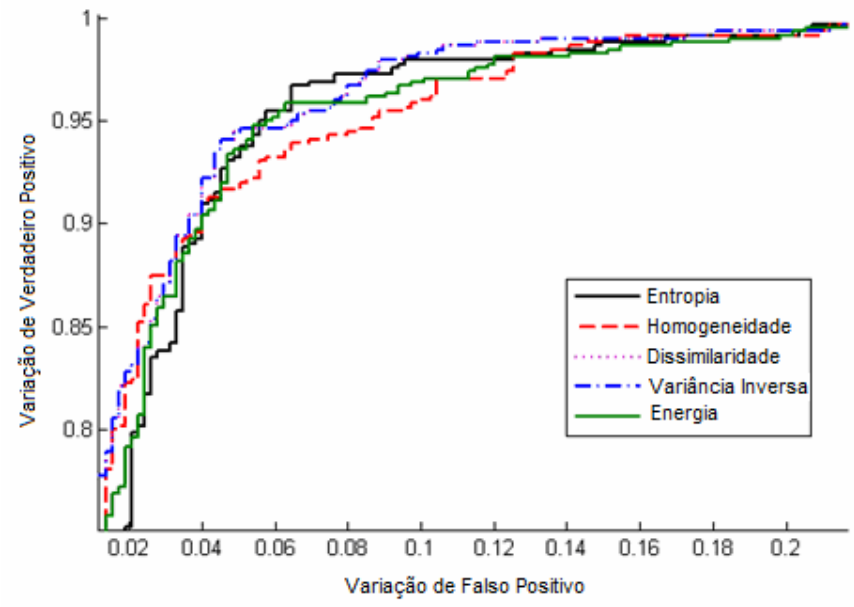


Figura 22. Curvas ROC comparando os descritores para dois níveis de cinza.

Na Tabela 10, pode-se observar que os resultados da característica de entropia, com a utilização de fragmentos em dois níveis de cinza, utilizando os diferentes tipos fusão e voto majoritário.

O resultado que o classificador SVM apresenta, é um valor escalar próximo de zero. Quanto mais discriminatória for a classificação mais distante de zero estará o valor. A fusão por soma utiliza o peso desse valor na contabilização do resultado. Quanto mais valores distantes de zero o classificador gerar, maior será o valor da soma, portanto, mais certeza terá no resultado. Por este motivo, comparando com o resultado de saída direto do SVM, a fusão por soma, foi a mais adequada para encontrar o melhor resultado.

Tabela 10. Resultados obtidos com os diferentes tipos de votos para o descritor entropia, com fragmentos em 2 níveis de cinza.

Descritor	**Média dos Erros (%)**		
Entropia	**SVM**	**Voto Majoritário**	**Fusão por Soma**
50 para treinamento, 115 para testes	7,77	6,08	**5,13**
100 para treinamento, 115 para testes	8,41	5,47	**5,04**
200 para treinamento, 115 para testes	8,96	5,82	**5,91**

Após fixar a quantidade de testes, foi estabelecido um protocolo para comparar os resultados com a quantidade de escritores para treinamento fixo e variação da quantidade de escritores para teste. A Tabela 11 mostra os resultados os descritores com a utilização de fragmentos em dois níveis de cinza, variando a quantidade de testes em 35, 70 e 240 escritores, com número fixo de 75 escritores para treinamento.

Tabela 11. Resultados obtidos com variando a quantidade de escritores para teste, sendo 75 escritores fixos para treinamento.

Descritor	**Média dos Erros (%)**		
Número de escritores para teste	**35**	**70**	**240**
2.o Momento Angular	3,71	4,42	6,12
Entropia	2,86	3,71	5,79
Homogeneidade	4,57	5,00	6,66
Dissimilaridade	3,42	5,28	6,29
Energia	3,43	4,71	6,21
Variância Inversa	3,42	5,28	6,29

A Tabela 12 também mostra a variação da quantidade de escritores para testes, porém, utilizando 50 escritores para treinamento, e para testes a variação de 50, 150 e 265 escritores.

Tabela 12. Resultados obtidos com variando a quantidade de escritores para teste, sendo 50 escritores fixos para treinamento.

Descritor	**Média dos Erros (%)**		
Número de escritores para teste	**50**	**150**	**265**
2.o Momento Angular	2,80	6,33	5,92
Entropia	3,60	5,40	5,92
Homogeneidade	3,20	5,93	6,15
Dissimilaridade	3,20	5,53	6,04
Energia	2,80	6,46	6,07
Variância Inversa	3,20	5,53	6,04

Na Tabela 13, é possível observar estabilidade do modelo mesmo com a variação do *kernel*. Neste caso estão presentes o *kernel* linear e o gaussiano.

Tabela 13. Resultados obtidos com a combinação de classificadores.

Descritor Combinado	**Erro Tipo I Falsa Rejeição (%)**		**Erro Tipo II Falsa Aceitação(%)**		**Média dos Erros (%)**	
kernel	linear	gaussiano	linear	gaussiano	linear	gaussiano
2º. momento angular	**3,48**	4,00	**7,30**	7,82	**5,39**	5,91
Entropia	**2,43**	8,17	**7,83**	5,04	**5,13**	6,60
Homogeneidade	**4,52**	8,17	**8,87**	3,13	**6,69**	5,65
Dissimilaridade	**3,30**	6,60	**8,52**	5,21	**5,91**	5,91
Energia	**3,30**	4,00	**7,65**	8,00	**5,47**	6,00
Variância inversa	**3,30**	6,60	**8,52**	5,21	**5,91**	5,91

Na Figura 23 é possível observar a curva ROC *(Receiver Operating Characteristic)* gerada para o *kernek* gaussiano. Nesse caso, é possível observar os dois descritores que apresentaram, isoladamente, o melhor poder de discriminação, isto é, a homogeneidade e o 2º. momento angular.

Por fim, para determinar o desempenho dos descritores quando combinados numa fusão de classificadores, selecionaram-se dois conjuntos de três descritores que apresentaram o melhor potencial discriminatório, segundo a curva ROC, no primeiro conjunto são eles: a dissimilaridade; a homogeneidade; e a variância inversa, e no segundo: a energia; a entropia; e o momento. Os resultados podem ser vistos na Tabela

14. O conjunto de descritores que apresentou o melhor resultado originou-se da combinação de dois anteriores. Isto é, de dois descritores do segundo conjunto com um do primeiro. Sendo que, para essa combinação os resultados foram os mesmos, demonstrado a estabilidade de comportamento, na participação da entropia e do segundo momento angular. Tendo em vista que os mesmos vinham apresentando, nos testes anteriores, desempenho superior aos demais, a fusão dos descritores veio a ratificar essa tendência.

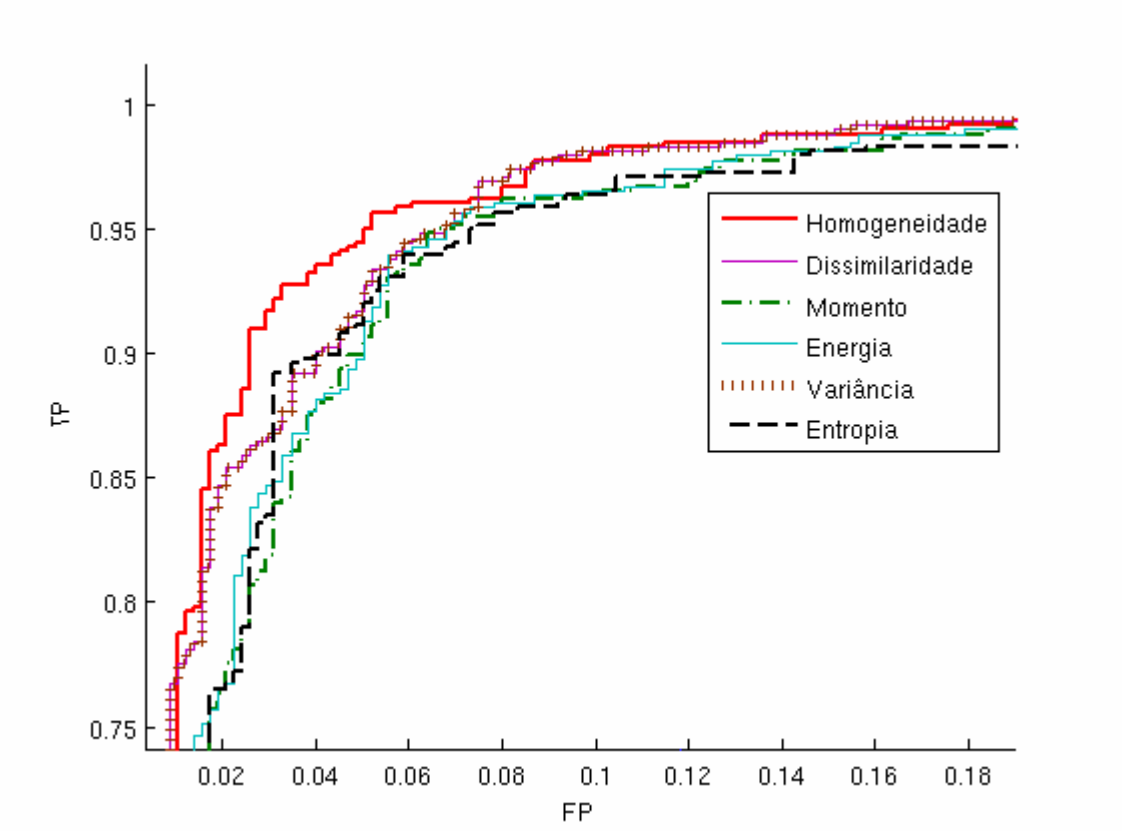


Figura 23. Curvas ROC geradas com o uso do *kernel* gaussiano (b) para os descritores de Haralick.

Tabela 14. Resultados obtidos pela fusão dos resultados obtidos pelos 6 diferentes descritores.

Descritor	**Média dos Erros (%)**				
Regra de Fusão	Voto	Soma	Máximo	Mínimo	Média
Homogeneidade + variância inversa + dissimilaridade	6,43	5,91	10,61	10,61	8,00
2º. momento angular + entropia + energia	5,47	5,30	10,95	10,95	7,56
2º. momento angular + entropia + dissimilaridade	**5,21**	**4,60**	**11,22**	**11,22**	**8,34**
2º. momento angular + entropia + variância inversa	**5,21**	**4,60**	**11,22**	**11,22**	**8,34**

Com o objetivo de comprovar que os resultados da base compactada PUC-PR em relação à base original, para as características de textura, foram melhores, a entropia foi selecionada e aplicada à base original, com 16 níveis de cinza, para 50 escritores de treinamento e 115 de testes, com a fusão de soma. Com estes resultados é possível observar uma melhoria em torno de 50% de acerto quando utiliza-se características de textura aplicadas aos fragmentos com texto compactado, com a eliminação dos espaços em branco (Tabela 15).

Tabela 15. Resultados obtidos das diferentes bases, utilizando a característica de entropia.

Entropia	**Erro Tipo I Falsa Rejeição (%)**	**Erro Tipo II Falsa Aceitação(%)**	**Média dos Erros (%)**
Base Original	15,30	57,04	36,17
Base Compactada	2,43	7,83	5,13

5.3 Experimentos com inclinação axial

Este trabalho re-implementou a técnica para extração de característica de inclinação axial da abordagem de Baranoski [BARANOSKI, 2005], com o objetivo de reaproveitar a base compactada PUC-PR e comparar com os resultados da base original.

Porém os testes abaixo não podem ser comparados à porcentagem de acerto da abordagem de Baranoski, que foi em torno de 90%, pois o protocolo utilizado pelo autor na classificação reutiliza amostras de treinamentos entre as amostras de teste.

Abaixo seguem as tabelas (Tabela 16 e Tabela 17) com os testes relativos à inclinação axial, aplicada na base original e compactada, com 115 escritores para testes e variação da quantidade de escritores para treinamento em 25, 50, 100 e 200. Os fragmentos possuem os 256 níveis originais de cinza, sem sofrer alteração após digitalização.

A redução da taxa de erro da base compactada em relação à base original foi em torno de 45%.

Tabela 16. Resultados obtidos da base original PUC-PR, utilizando a característica de inclinação axial.

Inclinação Axial – base original	**Média dos Erros (%)**		
escritores p/ treinamento	**SVM**	**Voto Majoritário**	**Fusão por Soma**
25	22,73	19,22	21,48
50	25,16	22,61	22,61
100	22,85	19,82	20,17
200	21,63	18,69	20,17

Tabela 17. Resultados obtidos da base compactada PUC-PR, utilizando a característica de inclinação axial.

Inclinação Axial – base compactada	**Média dos Erros (%)**		
escritores p/ treinamento	**SVM**	**Voto Majoritário**	**Fusão por Soma**
25	10,68	9,30	9,13
50	10,50	9,83	9,56
100	10,33	9,21	9,30
200	10,76	9,56	8,78

5.5 Comentários finais

Com estes resultados podemos concluir que as características de textura são mais relevantes em relação à característica angular de inclinação axial, quando aplicadas de acordo com os protocolos estabelecidos nesta abordagem envolvendo a base PUC-PR.

Capítulo 6

Conclusão

A abordagem deste trabalho apresentou um processo para verificação de manuscritos de conteúdo estático, com base em atributos grafoscópicos genéricos e genéticos da escrita, buscando simular a análise grafotécnica pericial. Para esse propósito, foram utilizadas apenas duas classes, autoria e não autoria. As etapas do processo foram detalhadas: aquisição da base, pré-tratamento, extração de características, produção do modelo pelo cálculo da distância euclidiana e decisão. De acordo com experimentos realizados com intenção de obter validações estatísticas, foi possível concluir:

- O objetivo principal de trazer resultados com alta taxa de precisão foi obtido para o método proposto de verificação semi-automática de autoria de manuscritos com média de erro em torno de 5%.
- O uso das características genéticas e genéricas, obtidos através do processo de segmentação não-contextual, mostrou-se promissor, com a obtenção de taxa de acerto da ordem de 95%.
- O modelo WI (*Writer-Independt*) apresentado, mostrou-se robusto: na variação do número de autores de treinamento; em relação ao problema de escrita natural, ou seja, escrita com ausência de falsificações; e por se tratar de um modelo global em que se houver inclusão de novos autores, não há necessidade de novo treinamento, evitando a complexidade e esforço computacional;
- Os resultados obtidos assemelhassem ao obtidos por outros autores que utilizaram a textura como abordagem. No entanto, observou-se igualmente que os descritores de Haralick não apresentaram melhora significativa no desempenho, quando do acréscimo da base de autores para o treinamento, aproximadamente 1%, assim como no uso de diferentes *kernels* no SVM e nas combinações entre os descritores, através de diferentes mecanismos de fusão.

- A característica grafocinética, inclinação axial apresentou uma taxa de precisão em torno de 90 obtida também através do processo de segmentação não-contextual. A redução da taxa de erro da com o uso do processo de segmentação não-contextual, em relação à base original foi em torno de 45%.
- Através dos testes com diferentes níveis de cinza, é possível concluir que a redução no número de níveis de cinza produz uma redução nos erros, devido à redução da complexidade da imagem, no que se refere aos níveis de cinza, e da conseqüente redução da dispersão dos valores na matriz de co-ocorrência. Podemos observar os resultados em dois níveis de cinza, que obtiveram menor taxa de erro.
- O processo de compactação da escrita, em que o texto foi segmentado e agrupado com o objetivo de reduzir os espaços em branco, trouxe resultados mais precisos em relação à base original, comprovando que os espaços em branco prejudicam os valores de extração de características, confundindo o processo de decisão. Desta forma, a remoção foi mais eficiente porque não descaracterizou o autor.
- Apesar da remoção dos espaços em branco, os resultados das características de textura em relação à característica de inclinação axial, apresentaram menor taxa de erro, pois as características de textura se destacam mais que características angulares em imagens com texto agrupado.

Como proposta para trabalhos futuros encontra-se a inclusão de outros descritores que permitem a extração de características de textura, tais como as *Wavelets* e Filtros de Gabor, utilizando a nova base gerada com ausência dos espaços em brancos.

Outra proposta é testar uma abordagem de classificação diferenciada através de métodos de compressão de dados, aplicada em classificação de texturas.

Além das características de textura, outras características baseadas em atributos grafoscópicos podem ser aplicadas na nova base, em busca de métodos computacionais cada vez menos complexos que gerem melhores resultados.

O objetivo é melhorar o processo de separação entre as classes autoria e não autoria.

Referências Bibliográficas

[AL-DMOUR, 2007] Al-Dmour, A.; Zitar, R. A.; *Arabic writer identification based on hybrid spectral-statistical measures*, Journal of Experimental & Theoretical Artificial Intelligence, Volume 19, Number 4, December , 307-332p., (2007).

[ALVES, 2006] ALVES, W. A. L.; ARAÚJO, S. A.; LIBRANTZ, A. F. H.; *Reconhecimento de padrões de texturas em imagens digitais usando uma rede neural artificial híbrida.* Exacta, São Paulo, v.4, n.2, p. 325-332, (2006).

[BARANOSKI, 2005] BARANOSKI, F. L.; *Verificação da autoria em documentos manuscritos usando SVM*. Dissertação de Mestrado, Pontifícia Universidade Católica do Paraná, (2005).

[BERTOLINI, 2008] BERTOLINI, D.; OLIVEIRA, L. S.; JUSTINO, E.; SABOURIN, R.; *Ensemble of Classifiers for Off-line Signature Verification*. IEEE International Conference on System, Man and Cybernetics (SMC 2008), 283-288p., Singapore, (2008).

[BRINK, 2007] BRINK, A.; SCHOMAKER, L.; BULACU, M; *Towards Explainable Writer Verification and Identification Using Vantage Writers.* (ICDAR 2007), 23-26 September 2007, Curitiba, Paraná, Brazil, 824-828p.

[BULACU, 1993] BULACU, M.; SHOMAKER, L; *Writer Identification Using Edge-Based Directional Features,* Proc. Of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003), IEEE Computer Society, Edinburgh, Scotland, vol. II, pp. 937-94, (1993).

[BURGES, 1998] BURGES, C. J. C.; *A Tutorial on Support Vector Machines for Pattern Recognition*, Data Mining and Knowledge Discovery 2, 121-167p, (1998).

[BUSH, 2005] BUSH, A.; BOLES, W.; SRIDHARAN, S.; *Texture for Script Identification,* IEEE Transations on Pattern Analysis and Machine Intelligence, Vol. 27, no. 11, November, 1721-1732p., (2005).

[CHA, 2001] CHA, S. H.; *Use of the Distance Measures in Handwriting Analysis*. Doctor Theses. State University of New York at Buffalo, EUA, p. 208, (2001).

[FRANKE, 2002] FRANKE, K.; BÜNNEMEYER, O.; SY, T.; *Ink Texture Analysis for Writer Identification.* Fraunhofer Institute for Production Systems and Design Technology, Berlin, Germany, 2002.

[GOBINEAU, 1954] GOBINEAU, H.; PERRON, R.; *Géenéetique de léecriture et étude de la personnalité: Essais de graphometrie*, Delachaux & Niestlée, (1954).

[GOMIDE & GOMIDE, 1995] GOMIDE, T.; GOMIDE, L.; *Manual de Grafoscopia,* Editora Saraiva, São Paulo, Brasil, 106 p., (1995).

[HARALICK, 1973] HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I.; *Textural Features for Image Classification,* IEEE Transactions on Systems, and Cybernetics, vol. smc-3, no. 6, November, (1973).

[HE, 2005] HE, Z. M.; FANG, B.; DU, J.; TANG, Y. Y.; YOU, X.; *A Novel Method for Off-line Handwriting-based Writer Identification,* Eight International Conference on Document Analysis and Recognition (ICDAR'05), IEEE, 2005.

[HUBER & HEADRICK, 1999] HUBER, R. A.; HEADRICK, A. M.; *Handwriting Identification: Facts and Fundament,* CRC Press, New York, ISSN 0-8493-1285-X, pp. 434, (1999).

[IMDAD, 2007] IMDAD, A.; BRES, S.; EGLIN, V.; *Writer Identification using Steered Hermite Features and SVM* . (ICDAR 2007), 23-26 September 2007, Curitiba, Paraná, Brazil, 839-843p.

[JIANGSHENG, 2002] JIANGSHENG, Y.; *Method of k-nearest neighbors*. Institute of Computational Linguistics, Peking University, China, 2002.

[JUSTINO, 2001] JUSTINO, E. J. R.; *O Grafismo e os Modelos Escondidos de Markov na Verificação Automática de Assinaturas.* Tese de Doutorado, Pontifícia Universidade Católica do Paran, Brasil, 2001.

[JUSTINO, 2002] JUSTINO, E. J. R.; *A análise de documentos questionados*. Produção Bibliográfica de Cunho Técnico para obtenção de grau de Professor Titular. Pontifícia Universidade Católica do Paraná, Brasil, 2002.

[JUSTINO, 2003] JUSTINO, E. J. R.; A *Autenticação de Manuscritos Aplicada à Análise Forense de Documentos*. In: TIL - 1°. Workshop em Tecnologia da Informação e Linguagem Humana, 2003, São Carlos. TIL - 1°. Workshop em Tecnologia da Informação e Linguagem Humana, 2003, v. 1. p. 102-106.

[MORRIS, 2000] MORRIS, N.; *Forensic Handwriting Identification Fundamental Concepts and Principles*, Academic Press, p. 238, (2000).

[PERVOUCHINE, 2007] PERVOUCHINE, V.; LEEDHAM, G.; *Study of Structural Features of Handwritten Grapheme ´th´ for Writer Identification*, Proceedings of the Third International Symposium on Information Assurance and Security, Manchester, United Kingdom, (2007).

[ROCHA & LEITE, 2007] ROCHA, A. R.; LEITE, N. J; *Classificação de texturas a partir de vetores de atributos e função de distribuição de probabilidades*. Universidade Estadual de Campinas, Instituto de Computação, Campinas, SP, Brasil, 2007.

[SAID, 1998] SAID, H. E. S.; PEAKE, G. S.; TAN, T. N.; BAKER, K. D. *Writer Identification from Non-uniformly Skewed Handwriting Images.* Department of Computer Science, University of Reading, UK, 1998.

[SAID, 1999] SAID, H.; PEAKE, G.; TAN, T.; BAKER, K.; *Personal Identification Based on Handwriting*, Patter Recognition, no. 33, p. 149-160, (1999).

[SHEN, 2002] SHEN, C.; RUAN, X.; MAO, T.; *Writer identification using Gabor wavelet, Intelligent Control and Automation*, 2002. Proceedings of the 4th World Congress on Volume 3, 10-14 June 2002 Page(s):2061 - 2064 vol.3.

[SCHOMAKER, 2007] SCHOMAKER, L.; *Advances in Writer Identification and Verification*, Proceedings of the International Conference on Document Analysis and Recognition (ICDAR) Curitiba, Brazil. (2007).

[SRIHARI, 2002] SRIHARI, S. N.; CHA, S. H.; HINA A.; SANGJILK, L.; *Individuality of Handwriting*, Journal of Forensic Sciences, (2002).

[SKILJAN, 2008] SKILJAN, I.; *http://www.irfanview.com/*, 2008.

[VAPNIK, 1998] VAPNIK, V. *Statistical Learning Theory*, Wiley, N. Y, (1998).