FRANCIS LUIZ BARANOSKI

VERIFICAÇÃO DA AUTORIA EM DOCUMENTOS MANUSCRITOS USANDO SVM

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

2005

FRANCIS LUIZ BARANOSKI

VERIFICAÇÃO DA AUTORIA EM DOCUMENTOS MANUSCRITOS USANDO SVM

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

Área de Concentração: *Metodologias e Técnicas de Computação*

Orientador: Prof. Dr. Flávio Bortolozzi
Co-orientador: Prof. Dr. Edson J. Rodrigues Justino

CURITIBA
2005

Baranoski, Francis Luiz
Verificação da Autoria de Documentos Manuscritos Usando SVM. Curitiba, 2005. 88p.

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

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

Dedico este trabalho aos meus pais Luiz e Maria Helena, pelo apoio e compreensão e a minha namorada Simara pelo carinho e paciência.

Agradecimentos

A Deus que me deu forças nos momentos de desânimo nesta jornada.

A minha família e minha namorada que sempre me incentivaram a continuar lutando pelos meus ideais.

Ao Professor Orientador Dr. Flávio Bortolozzi pelas contribuições seguras. Ao Professor e amigo Co-orientador Dr. Edson Justino pela ajuda na fundamentação teórica do trabalho, pelos questionamentos e contribuições construtivas, pelo incentivo e ainda pelo laço de amizade construído.

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

Aos meus Colegas de estudos, Professores, e Funcionário do PPGIA.

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

SUMÁRIO

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AMQ	Amostra de Manuscrito Questionado
AMRF	Amostra de Manuscrito de Referência
AMTR	Amostra de Manuscrito de Treino
CEDAR	*Center of Excellence for Document Analysis and Recognition*
Dpi	*Dot per inch*
FEPI	Ferramenta de Processamento de Imagens
KNN	*K – Nearest Neighbors*
HMM	*Hidden Markov Model*
LADTEC	Laboratório de Direito e Tecnologia
MMH	Hiperplano de Margem Máxima
PPGDES	Programa de Pós-Graduação em Direito Econômico e Social
PPGIA	Programa de Pós-Graduação em Informática Aplicada
PUCPR	Pontifícia Universidade Católica do Paraná
RBF	Redes com Funções de Base Radial
S	Vetores Standard
SRM	Minimização de Risco Estrutural
SV	Vetores de Suporte
SVM	*Support Vector Machine*
T	Vetores de Teste
VC	Dimensão Vapnik Chervonenkis

Lista de Símbolos

d	Dimensão do vetor atributo
$f(x)$	Função densidade probabilidade
$w_.$	Grupo ou classe
x	Objeto, padrão de entrada ou atributo
y_i	Rótulo da saída do SVM
$K(x_i, x)$	Função do Kernel
C	Penalidade de erro no SVM
$\vec{p}$	Vetor de pesos do SVM
b	Bias
n	Número de classes
Φ	Hiperplano de separação ótima
ξ	Magnitude do erro de classificação
δ	Margem máxima
f_v	Vetor de características extraído de uma imagem
D_i	Decisão na identificação de manuscritos

Resumo

A verificação automática de autoria de manuscritos estáticos ou *off-line* apresenta-se como um problema em aberto, devido aos fatores como a variabilidade da escrita de um mesmo autor e a semelhança de escrita entre autores diferentes. A abordagem proposta nesta dissertação, relacionada à verificação da autoria de manuscritos, baseia-se na visão da grafoscopia. O método utiliza uma abordagem global de classificação baseada em: *Support Vector Machine (SVM),* características extraídas da grafoscopia e medidas de distâncias na geração dos vetores de comparação das mesmas. Nesse modelo somente duas classes são assumidas: autoria (associação) e não autoria (dissociação). Para validar o experimento 50 autores foram usados no treinamento, e 265 autores foram usados nos testes, compondo, assim, a base de dados para os experimentos. A abordagem apresentada propõe uma solução através do uso de um número reduzido de manuscritos por autor (em torno de 5), assim como a redução do número de classes de autores, problemas que são encontrados em outros métodos relacionados à autenticação de autoria em manuscritos [SRIHARI et al., 2002], [BULACU et al. 2003], [SCHLAPBACH & BUNKE, 2004]. As seguintes etapas constituem esta abordagem: aquisição de dados (colheita e digitalização dos manuscritos), pré-processamento (preparação da imagem para a extração de características aplicadas sobre as imagens digitalizadas), segmentação (imagem particionada em fragmentos para a extração de características), extração de características (propriedades relevantes que caracterizam a individualidade da escrita, características embasadas na grafoscopia), cálculo de distâncias entre as características (etapa na qual a distância entre os vetores de características, pertencentes a duas amostras, são primeiramente computadas e usadas para verificação do autor), produção de um modelo (um conjunto de referências é gerado para utilização no processo comparativo, que usa *SVM*), e finalmente o processo de decisão (saída do modelo produzido é avaliada, verificando se o manuscrito pertence à determinada classe ou não). As taxas erros de obtidos estão na faixa de 2,5 % para falsa rejeição e de 7,9% para falsa aceitação. O erro total foi de 10,4%.

Palavras-chave: 1. Grafoscopia 2. Autoria de Manuscritos. 3. *Support Vector Machine* 4. Reconhecimento de Padrões.

Abstract

The off-line automatic handwritten verification is an open problem due to factors like the handwriting variability from the same author, the handwriting similarity among different authors. The proposed approach in this dissertation, related to the handwritten authorship verification, is based on the Questioned Document Examination (QDE). The method is based on the *Support Vector Machine (SVM)*, the features extracted from the Questioned Document, and the distance measures in their comparison vector generation. Only two classes are assumed in this model: authorship (association) and non-authorship (dissociation). For the experiment validation, 50 authors were used in the training and 265 authors were used to testing. The presented approach proposes a solution through the use of a reduced number of handwritten for each author (around 5 manuscripts per writer), just as the reduction in the number of classes, problems found in other methods related to the authorship authentication in handwritten [SRIHARI et al., 2002], [BULACU et al.,2003], [SCHLAPBACH & BUNKE, 2004]. This approach consists of the following stages: data acquisition (requested standard and manuscript scanning), preprocessing (image preparation for the feature extraction applied over the digitized images), segmentation (image partitioned in fragments for the feature extraction), feature extraction (relevant properties which characterize the handwriting individuality, features based on the Questioned Document Examination), feature distance measurement (stage in which the distance between the feature vectors belonging to two samples is firstly computed and used to verify the author), a model production (a reference set is created for utilization in the comparative process which uses *SVM*), and, finally, the decision process (the output of the produced model is evaluated, verifying if the handwritten belongs to the specific class or not). The obtained error rates are around 2,5% for false rejection and 7,9% for false acceptance. The total error was 10,4%.

Key-words: 1. Questioned Document Examination (QDE). 2. Handwritten verification. 3. *Support Vector Machine (SVM).* 4. Extracted Questioned Document Examination features.

Capítulo 1

Introdução

A grafoscopia tradicional é o campo da Ciência Forense destinada a buscar respostas para as questões judiciais associadas a documentos manuscritos. Distintamente da documentoscopia, a grafoscopia visa tratar unicamente dos aspectos da escrita e sua autoria, não abordando os diferentes tipos de documentos ou materiais de suporte onde o manuscrito foi aposto, [MORRIS, 2000] e [DINES, 1998].

A grafoscopia, tradicionalmente utilizada na autenticação de documentos na área jurídica, vem sendo extensivamente utilizada como ferramenta destinada à identificação e verificação da autoria, auxiliando na solução de crimes ou na identificação de suspeitos [JUSTINO, 2002].

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, [SANTOS et al., 2004].

Para a grafoscopia são relevantes dois elementos de análise, o grafostático e o grafocinético, [MORRIS, 2000], [DINES, 1998] e [JUSTINO, 2002]. O primeiro aborda critérios mais globais de análise, tais como a altura, comprimento e forma. O segundo aborda elementos dinâmicos do traçado, tais como inclinação axial, pontos de ataque e remates.

A análise pericial de documentos manuscritos pode gerar discordâncias na determinação da autenticidade dos mesmos. Isto ocorre pelo fato de envolver um conjunto de procedimentos não normatizados e sujeitos a uma análise subjetiva do perito, e também pelo

fato de a verificação manual, para uma grande quantia de documentos, ser tediosa e facilmente influenciada pelos fatores físicos e psicológicos [XIAO & LEEDAM, 1999]. Desta maneira, a análise pericial torna-se uma tarefa complexa.

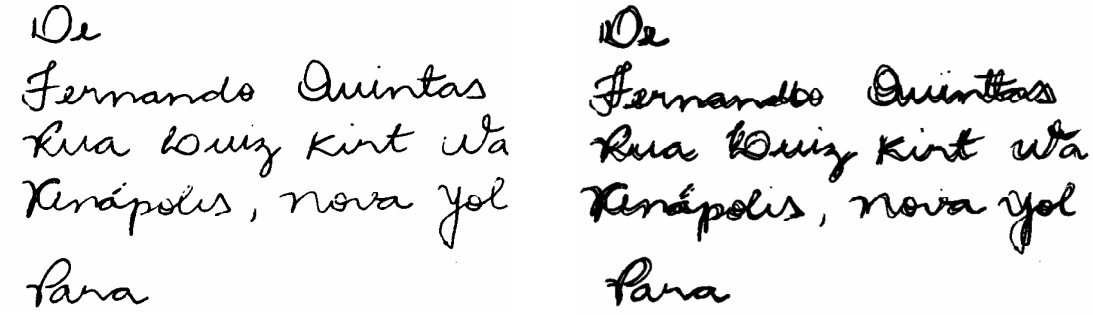
Baseando-se nestes princípios, a análise e autenticação de documentos manuscritos tornam-se alvo de muitas pesquisas no campo computacional, [SRIHARI, 2002], [BULACU et al., 2003], [LEEDAM & CHACHRA, 2003], utilizando técnicas de reconhecimento de padrões e aprendizado de máquina, buscando soluções computacionais automatizadas e semi-automatizadas que sejam possíveis de serem implementadas e que possam ser comprovadas cientificamente.

A identificação e representação do conhecimento necessário para abordar a problemática dos manuscritos, quanto à verificação de autoria, juntamente com a proposta para a sua solução, encontram-se apresentados no decorrer deste trabalho.

1.1. Desafio

O desafio da abordagem proposta é buscar soluções computacionais para minimizar a complexidade que envolve os fatores relacionados à escrita natural em manuscritos, como as variabilidades intrapessoais e as similaridades interpessoais.

As variabilidades intrapessoais, Figura 1.1(b), decorrem da instabilidade existente entre as escritas do mesmo autor. A escrita de uma pessoa pode mudar ao longo do tempo devido a diversos fatores tais como estado psicológico do autor e/ou tipos de caneta e texturas diferentes de papel. Já as similaridades interpessoais, Figura 1.2, são outros fatores de complexidade que representam semelhanças na escrita de autores distintos, tais como forma e estilo.



(a) (b)

Figura 1.1: (a) Exemplo de um manuscrito de um autor em específico, (b) variações intrapessoais observadas pela sobreposição de outro manuscrito do mesmo autor.

(a) (b)

Figura 1.2: Similaridades interpessoais (a) e (b)

A variabilidade da escrita é um problema complexo a ser tratado, pois pode ser influenciada por alguns fatores como alfabeto, sexo, idade, etnia da população analisada [SRIHARI & CHA, 2001].

No processo que envolve a verificação automática de documentos manuscritos, o objetivo é determinar se o manuscrito é de próprio punho do autor ou não. Portanto, dada duas classes w_1 e w_2,, em que w_1, representa a classe de exemplares genuínos (autores), e w_2 representa a classe dos exemplares de autores distintos (não autoria), o desafio é buscar computacionalmente a autenticação da autoria dos manuscritos, mesmo com as variabilidades presentes na escrita. A complexidade na verificação aumenta quando há muitas similaridades interpessoais e muitas variabilidades intrapessoais, dificultando a distinção entre as classes.

1.2.Motivação

A maior motivação desse trabalho encontra-se no caráter prático apresentado no mesmo, pois o método visa auxiliar e agilizar o processo de verificação da autoria de manuscritos realizado pelos peritos através de uma solução computacional. Permite-se ainda, através do método proposto, uma comparação segura entre o manuscrito conhecido com um outro questionado, retirando o fator da subjetividade aplicada pelos peritos no processo pericial.

A verificação da autoria em manuscritos, por métodos computacionais, não é um problema de fácil solução, pois uma solução aceitável deve passar por um rigoroso processo de avaliação, envolvendo não somente resultados estatisticamente comprobatórios, mas também compatíveis com os critérios aceitos pela comunidade jurídica internacional [JUSTINO, 2002].

Desta maneira, a abordagem proposta apresenta um método computacional baseado nos princípios da grafoscopia, técnica pericial utilizada na verificação de autoria de manuscritos e aceita pela comunidade jurídica internacional.

1.3.Objetivos

Este trabalho tem como objetivo apresentar uma abordagem para a verificação da autoria de documentos manuscritos. Dentro deste contexto, as seguintes metas são apresentadas:

- Apresentar uma solução computacional embasada em princípios jurídicos, para o auxílio na verificação da autoria de um manuscrito;
- Utilizar os preceitos da grafoscopia na análise das características da escrita;
- Utilizar um método de extração de características global, visando simplificar o processo de extração das características grafoscópicas;
- Utilizar um método de classificação com duas classes, autor e não autor;
- Utilizar características tolerantes às variações intrapessoais e similaridades interpessoais.

1.4.Contribuições

Este trabalho apresenta as seguintes contribuições:

- Implementação computacional da metodologia utilizada pela perícia grafotécnica, tanto para as características utilizadas quanto para o processo de análise e verificação da autoria.
- Formação de uma base de dados de manuscritos para a validação de procedimentos computacionais e que sirva como suporte para trabalhos futuros;
- Uso de uma abordagem global de treinamento e classificação que utilize somente dois modelos, autor e não autor, independente do autor e manuscrito analisado, ou ainda da inserção de novos autores;
- Publicação do artigo , “Identificação da Autoria em Documentos Manuscritos Usando SVM”, no 5º Encontro Nacional de Inteligência Artificial

1.5.Organização da Dissertação

Esta dissertação está organizada em sete capítulos. O primeiro contém uma introdução com uma contextualização sobre verificação automática de documentos manuscritos. No segundo capítulo são apresentadas as fundamentações teóricas sobre as Ciências Forenses e a verificação de manuscritos, bem como as técnicas computacionais em reconhecimento de padrões, relevantes para esse trabalho. No terceiro capítulo é feita uma revisão de trabalhos que abordam verificação e identificação da autoria de manuscritos por meios computacionais e que serve como suporte para este. No quarto capítulo é elucidada a metodologia. No quinto capítulo é apresentado o método juntamente com suas etapas. Os experimentos realizados são detalhados no sexto capítulo, os quais validam estatisticamente o método. E finalmente o sétimo capítulo, apresenta conclusões e propostas para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta a base teórica para o processo de verificação de autoria de documentos manuscritos. Apresenta breve descrição da grafoscopia, reconhecimento de padrões, como também tipos e métodos de abordagens relacionados ao problema.

2.1. Ciências Forenses

Ciências Forenses é o conjunto de ciências que envolve diferentes áreas do conhecimento humano, tais como a medicina, odontologia, fonética, documentoscopia, grafoscopia, biometria, administração, contábil, informática, entomologia, química, balística, toxologia, engenharias, psicologia, entre outras. Tais ciências apresentam ferramentas utilizadas para esclarecer questões associadas a alguma prova, no âmbito do judiciário. Tal esclarecimento confirmará a convicção do juiz sobre os elementos necessários em um processo judicial.

A documentoscopia é uma das ciências associadas à área forense que trata do estudo ou análise de documentos. 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 ou 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 [JUSTINO,2001].

Uma subárea da documentoscopia é a grafoscopia, que visa tratar unicamente dos aspectos da escrita e sua autoria. Nesse caso, a escrita pode estar relacionada a vários fatores, como por exemplo, à autenticidade da autoria e determinação da contemporaneidade do manuscrito. Seu estudo será de fundamental importância para o desenvolvimento da abordagem proposta neste trabalho, sendo, portanto, detalhado na seção posterior.

2.1.1 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].

2.1.2 Elementos básicos da grafia

Na análise grafotécnica pode-se encontrar alguns termos elementares da grafia que devem ser ressaltados, como segue [JUSTINO, 2001]:

- **Campo gráfico** é o espaço bidimensional onde a escrita é feita.
- **Movimento gráfico** é todo o movimento de dedos que o indivíduo faz para escrever, sendo que cada movimento gráfico gera um traço gráfico.
- **Traço** é o trajeto que o objeto da escrita descreve em um único gesto executado pelo autor.
- **Traço descendente, fundamental, pleno, ou grosso** é todo o traço descendente e grosso de uma letra.
- **Traço ascendente ou perfil** é o traço ascendente e fino de uma letra.
- **Ovais** são os elementos em formas de círculo das letras “a, o, g, q”, dentre outras.
- **Hastes** são todos os traços plenos (movimento de descanso) das letras “l”, “t”, “b”, “f”, etc. até a base da zona média. Também são consideradas hastes os traços verticais do “m” e do “n” maiúsculo e minúsculo.
- **Laçadas inferiores** são todos os plenos (descendentes) do “g”, “j”,”y”, “f”, etc. a partir da zona média até embaixo.
- **Bucles** são todos os traços ascendentes (perfis) das hastes das laçadas inferiores e, por extensão, todo o movimento que ascende cruzando a haste e unindo-se a ela formando círculo.
- **Partes essenciais** são o esqueleto da letra, a parte indispensável da sua estrutura.
- **Parte secundária ou acessória** é o revestimento ornamental ou parte não necessária à sua configuração.

Nas letras são distinguíveis algumas diferentes zonas, como segue:

- **Zona inicial** é a área onde se encontra o ponto no qual se inicia a letra.
- **Zona final** é a área onde se encontra o ponto no qual termina a letra.
- **Zona superior** é a área onde se encontra o ponto mais alto ocupado pelas hastes, pelos pontos e acentos, pelas barras do “t” e parte das letras minúsculas.

- **Zona média** é a área central ocupada por todas as vogais minúsculas (a, e, i, o, u) e pelas letras “m” e “n”, “r”, etc, cuja altura toma-se como base para medir o nível de elevação das hastes e o nível de descanso das laçadas inferiores.
- **Zona inferior** é a zona baixa da escrita a partir da base de todos os ovais descendentes, das letras maiúsculas ou de outras letras.

Os termos elementares da grafia podem ser observados nas Figuras 2.1 e 2.2.

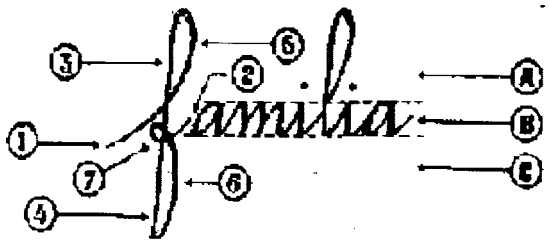


Figura 2.1: Exemplo de alguns elementos básicos da grafia: (1) Zona Inicial; (2) Zona final; (3) Haste; (4) Laçada; (5) Bucle da haste; (6) Bucle da laçada; (7) Bucle em forma de laço; (A) Zona superior; (B) Zona média; (C) Zona inferior. Adaptado de [JUSTINO, 2001].

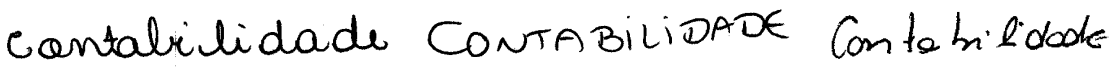
Figura 2.2 Exemplo das diferenças entre as partes de um elemento gráfico: (A) Partes essenciais; (B) Partes secundárias. Adaptado de [JUSTINO, 2001].

2.1.3 Características individuais do autor

As características individuais são de suma importância para a verificação de autoria, tanto no contexto da perícia grafoscópica convencional como nas abordagens computacionais, pois através de uma combinação de características individuais do autor e uma freqüência de ocorrências, faz com que a escrita de um autor seja diferençável de outros. Um estudo feito por Justino [JUSTINO, 2002], descreve algumas características particulares do autor usadas na grafoscopia, que serão apresentadas a seguir.

A forma caligráfica

A forma caligráfica é a representação pictórica da escrita sendo provavelmente a mais básica das características individuais do autor. Existem três tipos de formas caligráficas, a cursiva, caixa alta ou tipográfica e a mista, podendo ser observadas na Figura 2.3 respectivamente.

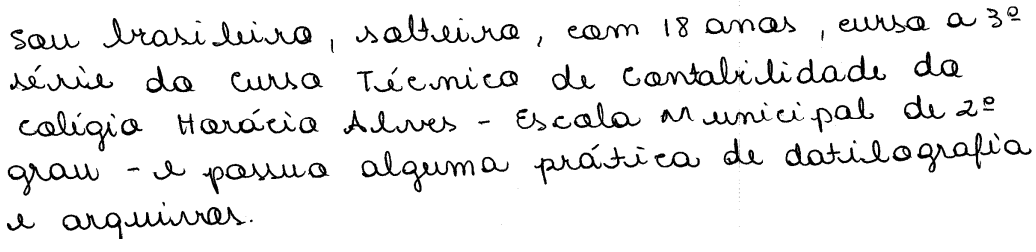


(a) (b) (c)

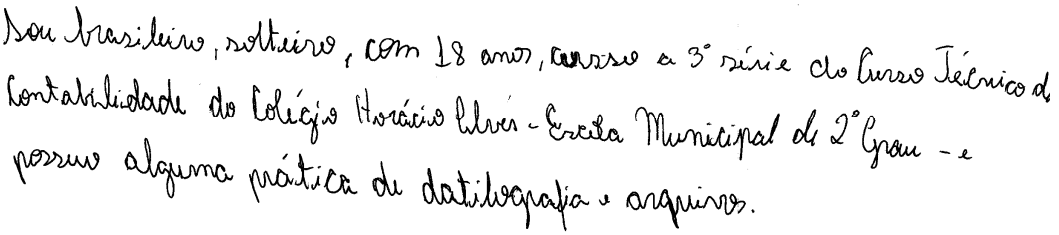
Figura 2.3 Exemplos das formas caligráficas: (a) cursiva; (b) tipográfica; (c) mista.

Nível de habilidade

Os autores podem ter tanto alto nível de habilidade quanto baixo nível de habilidade. Os de alto nível são capazes de produzir textos rítmicos bem traçados, podendo ser artisticamente embelezados. Já o escritor com baixo nível de habilidade produz textos com escrita vacilante, traçada lentamente. Esta característica é discriminatória para a identificação ou exclusão da autoria, sendo que o autor com baixo nível de habilidade é incapaz de escrever acima de seu próprio nível. Porém, o autor com alto nível de habilidade pode disfarçar a escrita escrevendo em um nível abaixo do seu.



(a)



(b)

Figura 2.4 (a) Escrita com alto nível de habilidade; (b) Escrita com baixo nível de habilidade.

Inclinação axial

A inclinação axial é 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 de 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.

(a) (b) (c)

Figura 2.5 Exemplos de escritas com inclinação axial: (a) à direita; (b) à esquerda; (c) nula.

Movimento

É a direção do movimento dos instrumentos de escritura, lápis ou caneta, podendo ser determinada através da observação das variabilidades na densidade de tinta da caneta ou do traço do lápis. Figura 2.6.

Figura 2.6 Imagem obtida por microscópio que mostra estrias produzidas por caneta esferográfica. Adaptado [JUSTINO, 2002].

Proporções

Referem-se geralmente às simetrias das letras individualmente. Este conceito normalmente desenvolve uma relação entre a proporção de uma letra em relação à outra, como por exemplo a letra "B", na qual o bulbo de topo não é do mesmo tamanho que o bulbo de base.

(a) (b)

Figura 2.7: Exemplos de proporção representados em (a) e (b) entre elementos de uma mesma letra para um mesmo autor.

Relações de altura

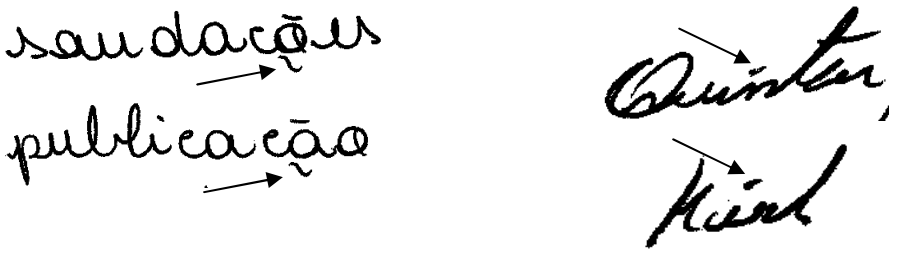
É a comparação ou correlação da altura de uma letra ou segmento de letra em relação à outra letra, normalmente dentro da mesma palavra. Espera-se que o autor mantenha tanto as letras maiúsculas como minúsculas no mesmo sistema de escrita, ou seja, na mesma altura ao longo de um corpo de escrita.

(a) (b)

Figura 2.8: Exemplos de relações de altura representados em (a) e (b) para um mesmo autor.

Mínimos gráficos

Os mínimos gráficos são pequenas porções de escritura, como pontos finais, vírgulas, acentos gráficos, cedilhas, que podem tornar-se características identificadoras devido ao estilo discriminante empregado pelo autor ao mínimo gráfico.



(a) (b)

Figura 2.9: Exemplos dos formatos de mínimos gráficos cedilha em (a) e pingos do da letra "i" em (b).

Corte da letra "t"

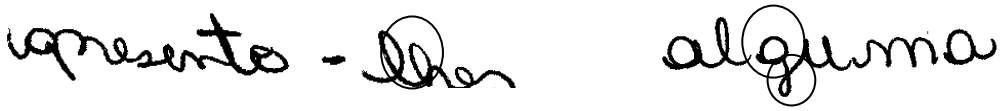
Variações na forma do corte da letra contribuem expressivamente para a distinção do autor, pois podem estar alinhadas na horizontal, apresentar inclinações, apresentar elevação do traço à direita ou esquerda ou ainda podem estar conectadas a um golpe de saída de uma letra terminal de uma palavra.

(a) (b)

Figura 2.10: Exemplos de corte da letra "t".

Laçadas

As laçadas ocorrem geralmente em letras cursivas possuindo elementos ascendentes e descentes. Elas podem ainda apresentar-se em formas pontiagudas ou arredondadas, simétricas ou assimétricas. Desta forma, a laçada é um traçado que apresenta um movimento de retorno para o ponto de partida.



(a) (b)

Figura 2.11: Exemplos de laçadas.

Pressão

Representa a variabilidade da largura do traçado e o acúmulo de material em uma determinada região do traço, dependendo da pressão imposta pelo autor e também da espessura da ponta da caneta. Esta característica pode indicar movimento.

Figura 2.12: Exemplos de diferenças de pressão.

Alinhamento em relação à linha base

Esta característica está associada à capacidade do autor de produzir linhas de textos alinhadas com uma linha guia horizontal imaginária em papel não pautado, ou linha real em papel pautado. As linhas de texto podem apresentar graus distintos de inclinação.

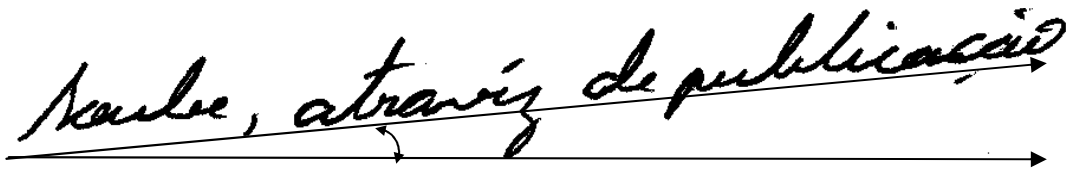


Figura 2.13 Exemplo de inclinação.

Descontinuidade do traçado

Descreve onde o objeto de escrita, caneta ou lápis, ergue-se do papel. Ocorre normalmente no meio de uma palavra provocando a descontinuidade do traçado.

(a) (b)

Figura 2.14: Exemplos de descontinuidade do traçado.

Velocidade

A velocidade da escrita é freqüentemente uma característica essencial para a identificação da autoria. Alguns elementos explicitam a ocorrência de escrita rápida e lenta, como segue:

Rápida

- Traçado tenso sem tremor;
- Alongamento e finalização das letras “e”, e cortes das letras “t”;
- Palavras ou letras conectadas;
- Aparência aplainada;
- Redução da legibilidade.

Lenta

- Vacilação e tremor;
- Traçado mais angular;
- Cruzamento das letras “t” em posição correta;
- Parada e começo abrupto;
- Escrita feita de letras individuais e legíveis;
- Movimento podendo apresentar ornamentos.

Figura 2.15: Exemplo de escrita rápida e lenta, respectivamente.

Embelezamento

Localiza-se usualmente no começo de uma letra, podendo estar presente ao longo do manuscrito.

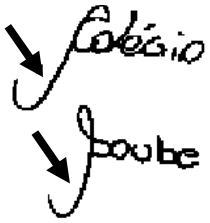


Figura 2.16: Exemplos de embelezamento da escrita.

Retraço

O retraço é o processo no qual o objeto de escrita repinta uma porção escrita da linha, normalmente em direção oposta, como um movimento descendente seguido por um movimento ascendente sobre a linha existente. Classifica-se como característica do autor enquanto representa um comportamento natural, pois quando acontecer em forma de correção da letra pode-se configurar em um indicativo de fraude.

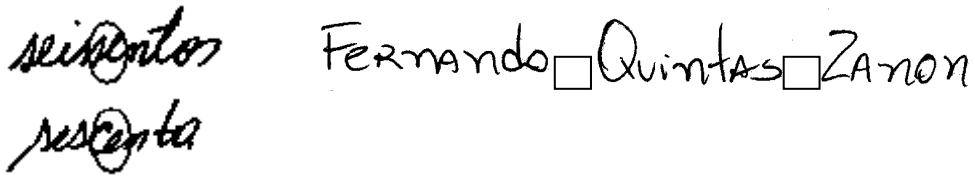
(a) (b)

Figura 2.17: Exemplos de retraço.

Erros de ortografia e espaçamento

A ortografia incorreta das palavras pode ser um indicativo de uma característica individual do autor.

Alguns escritores interrompem o curso da escrita entre combinações de letras específicas. Espaçamento entre letras adjacentes ou até mesmo entre palavras, podem descrever características habituais do autor.



(a) (b)

Figura 2.18: Exemplo de escrita incorreta (a), adaptado [JUSTINO, 2003a], e espaçamento entre palavras (b).

Formato

O formato de um documento questionado pode conter adicionalmente uma característica identificadora. Utilizando como exemplo o campo montante em um cheque bancário, um escritor pode usar elementos gráficos como "#120,00#", Figura 2.19(a), enquanto que outro pode usar "=120,00=", Figura 2.19(b). Em alguns a localidade pode aparecer abreviada "Ctba." e em outros por extenso "Curitiba".

(a) (b)

Figura 2.19: (a) e (b) exemplo de diferenças no formato [JUSTINO, 2003a].

Entradas e Golpes de Saída do Traçado

As entradas e golpes de saída de uma letra podem repetir-se em formações de letras semelhantes como nas letras "U", "V", "M" e "N". As entradas e golpes de saída podem ser movimentos habituais, podendo representar características identificadoras de um escritor. O mesmo pode ser dito em relação aos outros golpes entre uma letra e outra, na escrita cursiva, cujo objetivo é criar uma ligação entre as letras individuais. Golpes de conexão permitem que o escritor tenha mais criatividade, ressaltando as características individuais. Estas características podem não ter sido enfatizadas durante o processo de aprendizado, podendo ser, portanto, importantes na identificação de autoria.

Figura 2.20: Exemplo de entradas e golpes de saída.

2.1.4 Modelos para Análise de Manuscritos

Na maioria dos laboratórios forenses, um "exemplar genuíno" ou s*tandard* é simplesmente um item conhecido com o qual um item desconhecido pode ser comparado. Um exemplar genuíno normalmente possui uma quantidade suficiente de texto escrito para identificar características da individualidade do autor. A origem indubitável da escrita deve ser ligada, nos tribunais, à sua autenticidade. O perito deve confrontar a escrita dos exemplares genuínos com a escrita do documento questionado e com isso produzir um laudo técnico, no qual o parecer técnico demonstre sua autenticidade ou discordâncias [JUSTINO, 2003].

O exemplar original ideal a ser usado para a comparação da escrita manuscrita é aquele obtido sob as mesmas condições com a qual o documento questionado fora produzido. Ele contém as mesmas palavras, números e símbolos. Foi escrito usando-se aproximadamente o mesmo tempo e os mesmos tipos de recursos, como tipo de papel e caneta. O exemplar original deve, portanto, reproduzir suficientemente todas as variabilidades da escrita do autor. Adicionalmente, o mesmo deverá ser produzido sem que o autor conheça o propósito do seu uso. Obviamente, nem todos esses requisitos serão satisfeitos em todos os casos, mas é importante, sempre que possível, que a maioria dessas condições sejam satisfeitas. Se o exemplar original não se desviar demasiadamente da duplicação ideal do documento questionado, o perito pode utilizar toda a sua habilidade para produzir um laudo definitivo de identificação ou rejeição [JUSTINO, 2003].

Existem dois tipos básicos de exemplares utilizados como modelo, os colhidos e os coletados [JUSTINO, 2003]. Os exemplares coletados são aqueles 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. Eles estão, portanto,

livres da tentativa de disfarce. A desvantagem do exemplar coletado está na possível dificuldade em encontrar espécimes que reproduzam o formato e texto do documento questionado. Os exemplares coletados inadequadamente podem conduzir o perito a uma comparação inconsistente. A vantagem do exemplar coletado reside em eliminar a possibilidade do disfarce, o que freqüentemente supera as possíveis desvantagens.

Os exemplares colhidos são aqueles nos quais o indivíduo é intimado a reproduzir um material escrito específico. Essa classe de exemplares possui a vantagem de conter, aproximadamente, o formato e o conteúdo do documento questionado, produzido de acordo com as orientações do perito. Ele possui, contudo, a desvantagem de o autor conhecer a finalidade do documento, que pode ser usado contra seus interesses.

Modelos tradicionais

Existem vários modelos de documentos de coleta, que vêm sendo usados em vários países. Esses modelos, apesar de não duplicarem o conteúdo exato do documento questionado, possuem muitas associações de palavras, letras e símbolos encontradas em cartas comuns [JUSTINO, 2002].

Os modelos se adaptam aos padrões de grafia do idioma usado. Adicionalmente os mesmos apresentam todos os caracteres do alfabeto, maiúsculos e minúsculos, acentuações (características da linguagem usada), pontuação comum e os números de 0 a 9. A maioria das cartas modelo, clássicas na literatura, foram confeccionadas para grupos de escritores da língua inglesa, como por exemplo, a “Carta da Classe 16”, a “Carta do Egito” e a “Carta de Londres”, Figura 2.21 respectivamente. Portanto, podem vir a desconsiderar características individuais importantes em um idioma diferente, como por exemplo, os mínimos gráficos encontrados na língua portuguesa.

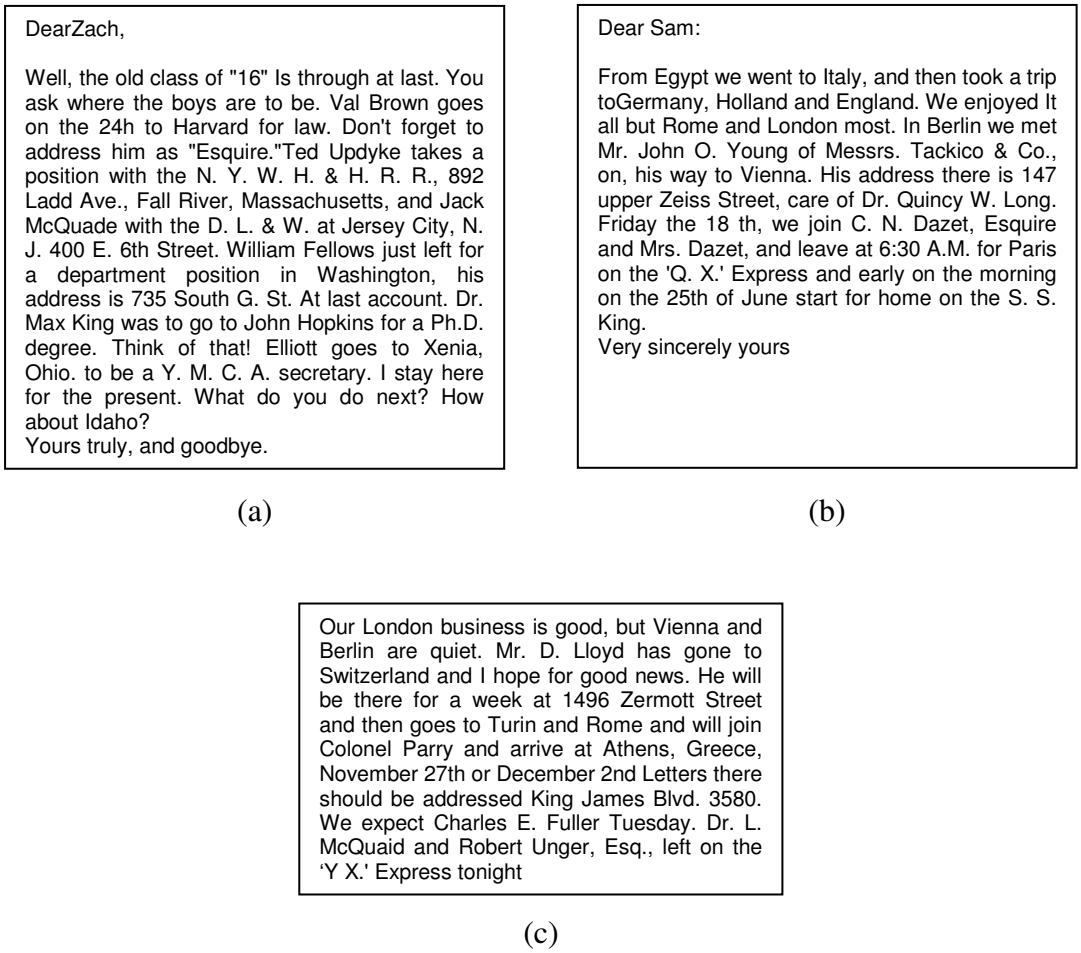


Figura 2.21: (a) Carta Classe "16"; (b) Carta do Egito; (c) Carta de Londres.

2.1.5 Laudos de documentos questionados

O laudo pericial consiste na formação de opinião do perito grafotécnico referente a um determinado caso, sendo expressa em um documento ou oralmente, freqüentemente baseada em uma análise complexa. O laudo concluído estará sujeito à revisão crítica por via de testemunhos, conferências de pré-julgamento, e talvez um interrogatório rigoroso elaborado por vários advogados.

É provável que o exame tenha sido empreendido a pedido do investigador, promotor ou do advogado. Provavelmente a opinião formada como resultado do exame pericial esteja escrita até certo ponto em conformidade com as necessidades e expectativas do submissor. Porém, o parecer final encontra-se nos jurados e juiz, [JUSTINO, 2002].

Durante o julgamento, o testemunho do perito pode expressar e demonstrar suas opiniões. No entanto, na maioria dos casos, civil e criminal, isto não ocorre. Desta maneira, o

relatório técnico ou laudo é extensamente usado e a interpretação acaba ficando a critério do juiz e dos jurados.

2.2. 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 [RASHA, 1994]. O processo de verificação da autoria de manuscritos pode ser enquadrado neste universo, onde o manuscrito é um exemplo de padrão que pode ser representado por uma matriz de *pixels*.

Reconhecimento de padrões pode ser definido como a categorização de dados de entrada dentro de classes identificáveis via extração de características significantes ou atributos com detalhes relevantes. Conseqüentemente, o objetivo fundamental de um sistema de reconhecimento de padrões pode ser a classificação, como também a regressão. Um sistema de reconhecimento de padrões pode ser dividido em algumas etapas: aquisição do sinal, pré-processamento, extração de características, classificação e pós-processamento [DUDA & HART, 1973].

A etapa da aquisição do sinal é feita através de um sensor, a qual está ligada diretamente a fase posterior, pré-processamento, em que serão retirados detalhes de pouca relevância no reconhecimento dos padrões desejados, portanto a qualidade do sensor é importante para um sistema de reconhecimento de padrões, pois o mesmo pode compromenter a eficiência na identificação de padrões.

As características são quaisquer medidas extraíveis de um padrão que podem contribuir para a classificação, sendo que as mesmas podem ser representadas por valores contínuos, que são valores mapeados de toda a população, ou discretos, valores mapeados de amostras da população.

A classificação é raramente executada usando uma simples característica de um padrão de entrada. Geralmente, diversas características são requeridas para serem capazes de distinguir entre diferentes classes. Selecionar estas características pode ser uma difícil tarefa que pode requerer significante esforço computacional. A seleção de características é o processo de entrada para o reconhecimento de padrões, a qual envolve geralmente um julgamento. A chave seria escolher e extrair características que sejam [RASHA, 1994]:

- Computacionalmente possíveis;
- Conduzam a um sistema de boa classificação com poucos erros de classificação;
- Reduzam a quantia de informação manipulada sem perda de informação relevante.

Desta forma, a classificação associa os dados de entrada dentro de uma ou mais classes pré-definidas, baseada na extração de características significantes ou atributos. Com base nestas características extraídas, a verificação de autoria de manuscritos consiste em estabelecer uma regra de decisão através da comparação com o modelo de referência devidamente armazenado em uma base de conhecimento, que descreve uma representação análoga. O modelo de referência é obtido em uma fase anterior chamada treinamento (produção de um modelo) [JUSTINO, 2001].

A fase de treinamento é uma etapa muito importante do sistema de verificação. Os modelos oriundos dessa fase possuem um conjunto rico de informações que permitem uma boa precisão do processo de identificação. Essas informações possuem a vantagem de possibilitar a eliminação de redundâncias, que por sua vez propiciam uma redução do tempo gasto no processo de decisão.

2.3. Abordagens de Reconhecimento de Padrões

A escolha do tipo de representação (os tipos de 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 que os dois métodos formais mais comuns são:

- **Métodos estruturais:** buscam descrever informações geométricas de maneira estrutural, representando formas complexas a partir de componentes elementares, chamadas primitivas. Os métodos estruturais distinguem-se basicamente em dois tipos [JUSTINO, 2001]:
  1. Métodos estruturais propriamente ditos, nos quais a estrutura utilizada é um grafo que permite representar as formas, as primitivas e as relações entre elas. A fase

de decisão consiste na comparação do grafo representativo da forma do modelo com o grafo da forma em teste;

2. Os métodos sintáticos, nos quais a estrutura é usada para codificar a forma em uma lista, utilizando um alfabeto cujos termos representam elementos da forma a descrever. A fase de decisão consiste na análise da lista com a ajuda de regras sintáticas, como as utilizadas em um texto escrito em uma linguagem natural.

• **Métodos estatísticos**: consistem em efetuar as medições do espaço métrico através da estatística. O aprendizado é executado através da separação de um conjunto de amostras em classes obedecendo a um conjunto de características comuns. São especialmente importantes nos sistemas cujas classes possuem uma elevada instabilidade entre os vários espécimes. Os principais são os chamados paramétricos e os não paramétricos [JUSTINO, 2001].

1. Paramétricos trabalham com hipóteses de que as classes em questão possuem uma distribuição de probabilidade com comportamento determinado. O método supõe o conhecimento prévio das leis que regem a probabilidade das classes envolvidas e que seus parâmetros de estimação possuem normalmente um comportamento gaussiano. Esses métodos exigem uma base de dados de aprendizado para uma correta estimação dos parâmetros.
2. Não paramétricos assumem que as leis de formação da probabilidade de uma classe são desconhecidas. O problema consiste em propor algoritmos de convergência que determinem o limiar ideal de decisão.

2.4. Tipos de Abordagens de Verificação

As abordagens relacionadas à verificação 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.

As abordagens relacionadas à identificação de autoria de manuscritos podem ser divididas em:

- Globais – utilizam características globais, é feita a partir de segmentos do manuscrito, como parágrafos, linhas, ou simplesmente pedaços da imagem.
- Locais – utilizam características locais, é feita a partir de letras e palavras, segmentadas do documento manuscrito.

As características locais, no caso da verificação de autoria de manuscritos, absorvem eficientemente características discriminantes inerentes ao autor, tais como cortes da letra "t", laçadas, dentre outras, vistas na Seção 2.1.3. Porém, para esta abordagem não há um método de segmentação automático eficiente, o qual consiga extrair da amostra do manuscrito apenas as letras ou palavras, sendo tal método feito manualmente, como consta na literatura [CHA, 2001], [SRIHARI et al., 2002]. Desta forma, o processo de segmentação torna-se uma tarefa penosa e demorada, acarretando em um problema quando transposto para sistemas práticos, em situações reais.

O diagrama da Figura 2.22 demonstra onde se situa a abordagem deste trabalho, no contexto de verificação automática de autoria em manuscritos (área em cinza).

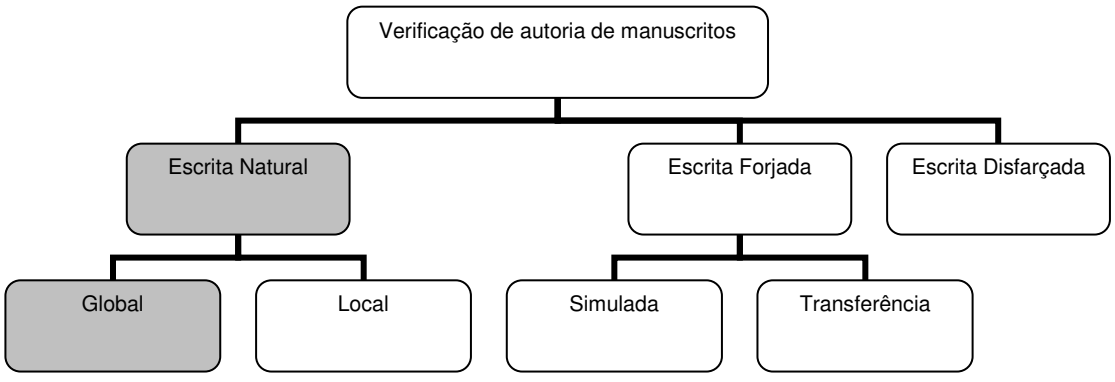


Figura 2.22 Diagrama hierárquico de classificação de métodos de verificação de autoria de manuscritos.

2.5. *Support Vector Machine* (*SVM*)

Os *SVM*s (*Support Vector Machine)* foram introduzidos recentemente como uma técnica para resolver problemas de reconhecimento de padrões. Esta estratégia de aprendizagem foi proposta por Vapnik [VAPNIK, 1995] e tem atraído a atenção dos pesquisadores devido as suas principais características, que são a sua boa capacidade de generalização e robustez diante de dados de grande dimensão.

Algumas aplicações automatizadas podem ser citadas nas mais variadas áreas de pesquisas, tais como, no reconhecimento de faces [OSUNA et al., 1997], classificação de impressões digitais [LIMA, 2002] e principalmente na área de manuscritos, como na verificação de assinaturas [JUSTINO, 2003b], reconhecimento de cadeias de dígitos manuscritos [OLIVEIRA & SABOURIN, 2004] e identificação de autoria [BARANOSKI et al., 2005]. A maioria das aplicações obtêm resultados comparáveis ou até mesmo superiores a outros algoritmos de aprendizado em algumas tarefas, como em Redes Neurais Artificiais [SANTOS, 2004].

No *SVM*, os padrões de entrada são transformados para um vetor de características de alta dimensionalidade, cujo objetivo é separar as características linearmente no espaço. Uma vez que o espaço adequado de características é definido, o *SVM* seleciona o hiperplano particular, chamado de hiperplano de margem máxima (MMH), o qual corresponde a maior distância de seus padrões no conjunto de treinamento. Estes padrões são chamados de vetores de suporte (SV), [SANTOS, 2004].

A idéia principal é separar as classes com superfícies que maximizem a margem entre elas. Um conjunto de dados de duas classes w_1 e w_2, linearmente separáveis com margem de separação máxima δ ou também chamado de hiperplano de separação ótima [OSUNA et al., 1997] são demonstrados na Figura 2.23.

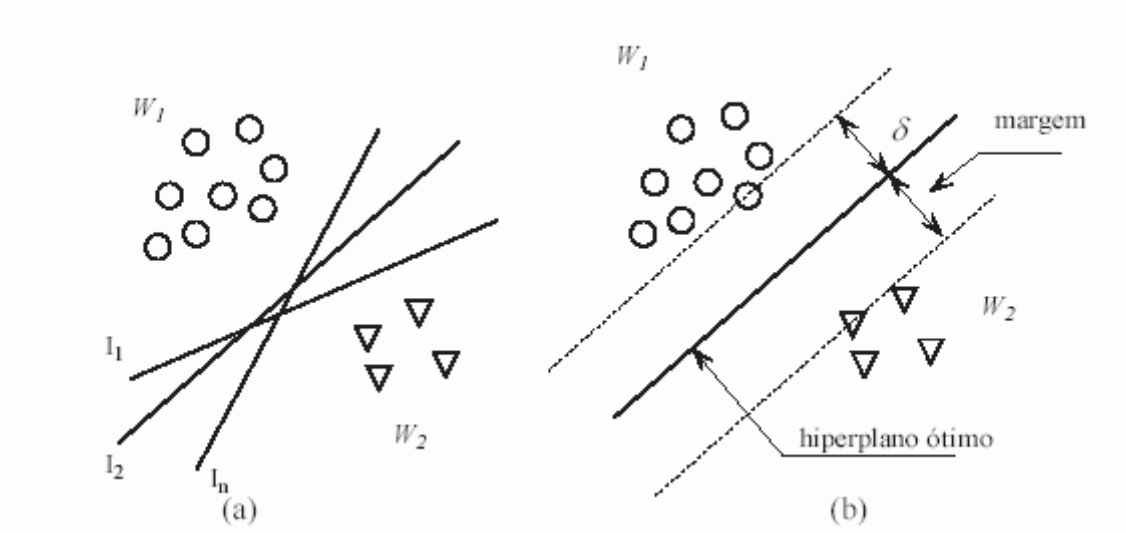


Figura 2.23: Classificação entre duas classes *W1* e *W2* usando hiperplanos: (a) Hiperplanos arbitrários *li* e (b) hiperplano com separação ótima, máxima margem para duas classes.

Para encontrar a superfície de decisão ótima, o algoritmo de treinamento o *SVM* tenta separar da melhor forma possível os pontos dos dados de ambas as classes. Os pontos mais próximos do limite entre as duas classes são selecionados, por serem mais importantes na solução, do que os pontos que estão mais distantes, os quais ajudam a definir a forma da melhor superfície de decisão que outros pontos.

Problemas complexos exigem funções mais complexas de classificadores para sua solução, como um classificador polinomial, que forma superfícies de decisão diferenciadas, conforme a Figura 2.24. Os vetores de suporte são representados por pontos com preenchimentos mais escuros.

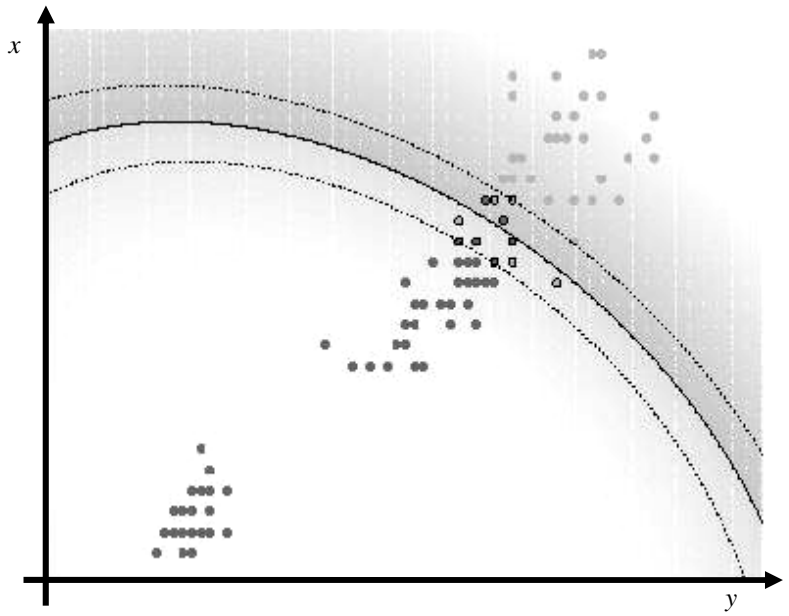


Figura 2.24: Superfície de decisão de um classificador polinomial. Adaptado [LIMA, 2002]

O *SVM* é baseado na idéia de minimização do risco estrutural, o qual minimiza o erro de generalização, isto é, erros verdadeiros em amostras não vistas. O número de parâmetros livres usado no *SVM* depende da margem que separa os pontos dos dados, mas não do número de características de entrada a fim de evitar o sobre-ajuste [MUKKAMALLA et al., 2002].

De acordo com Santos [SANTOS, 2004], o *SVM* provê um mecanismo genérico de preencher a superfície de hiperplano por dados através do uso de uma função de *kernel*. A literatura apresenta várias possibilidades de *kernels* para o *SVM*, [BURGES, 1998], [MULLER et al.,2001] e [JOACHIMS, 2002], dentre as quais o usuário pode prover uma função, tal como linear, polinomial, ou RBF para o *SVM* durante o processo de treinamento, o qual seleciona vetores de suporte ao longo da superfície desta função. Esta capacidade permite classificar uma faixa de problemas maiores.

O limite de decisão entre duas classes é definido pelo SVM:

$$f(x) = sign\left(\sum_{x_i \in sv} y_i \alpha_i^0 K(x_i, x) + b_o \right) \qquad (2.1)$$

onde x é um padrão de entrada; x_i é o *i-ésimo* vetor de suporte, SV é o conjunto de vetores de suporte; $y_i = \pm 1$ é o rotulo do padrão x_i ; b_o é o bias do hiperplano; α_i^0 é o *i-ésimo* multiplicador de Lagrange para o hiperplano ótimo; e finalmente $K(x_i , x)$ é a função do *kernel*, que pode mapear se necessário o dado de entrada para um alto espaço dimensional, conhecido como espaço de características. A função *kernel* é escolhida a *priori* e determina o tipo de classificador, (linear, polinomial ou RBF). Os *kernels* comumente usados são apontados com suas respectivas fórmulas na Tabela 2.1.

Tabela 2.1 *Kernels* do *SVM*

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

A idéia básica do *SVM* é mapear um espaço de entrada em um espaço de características de alta dimensionalidade. Este mapeamento pode ser feito linearmente ou não, de acordo com a função *kernel* usada para mapeamento. Neste novo espaço de características, o SVM constrói hiperplanos ótimos através dos quais as classes são separadas com o objetivo de estabelecer uma margem maior entre cada classe e um erro mínimo na classificação. O hiperplano ótimo pode ser escrito como uma combinação de poucos pontos de características cujos pontos são chamados de vetores de suporte do hiperplano ótimo [KHOLMATOV, 2003].

O SVM é baseado no principio de minimização do risco estrutural (SRM). O princípio de indução (SRM) tem dois objetivos principais. Primeiro, controlar o risco empírico no conjunto de dados de treinamento e segundo controlar a capacidade da função de decisão usada para obter o valor deste risco. A função de decisão do SVM treinada linearmente $f(\vec{x})$ é descrita pelo vetor de pesos $\vec{p}$, um limiar b e padrões de entrada $\vec{x}$:

$$f(\vec{x}) = sign(\vec{p}.\vec{x} + b) \qquad (2.2)$$

Dado um conjunto de treinamento S_l composto por duas classes separadas, w_1 ($y_1 = +1$) e w_2 ($y_1 = -1$), o SVM acha o hiperplano com a máxima distância euclidiana. De acordo com os princípios do SRM, haverá apenas um hiperplano ótimo com a margem máxima δ, definida como a soma das distâncias do hiperplano para os pontos mais próximos das classes. Este limiar do classificador do linear é o hiperplano ótimo separador, conforme demonstrado na Figura 2.23.

$$S_l = ((\vec{x}_1, y_1),...,(\vec{x}_l, y_l)), \vec{x}_1 \in \Re^n, y_1 \in \{-1, +1\} \tag{2.3}$$

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

$$f(\xi) = \sum_{i=1}^{l} \xi \tag{2.4}$$

A solução do SVM [JUSTINO et al., 2003] pode ser encontrada se mantiver o limite superior na dimensão VC (número de pontos máximo 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:

$$\min_{\vec{w},b,\xi} = \frac{1}{2}\vec{p}.\vec{p} + C\sum_{i=1}^{l} \xi_i \tag{2.5}$$

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 para os erros.

2.6.Comentários Finais

Neste capítulo foi descrita toda a teoria de fundamental importância para a elaboração do presente trabalho. Os tipos de abordagens e métodos relacionados à identificação de manuscritos fornecem uma visão clara do presente trabalho situando o mesmo no contexto de reconhecimentos de padrões. A teoria de *SVM* fornece subsídios para os experimentos realizados, descritos no Capítulo 6. No capítulo seguinte são descritas algumas abordagens relacionadas à identificação de autoria de manuscritos classificadas por tipos de abordagens.

Capítulo 3

Estado da Arte

Existe um número considerável de trabalhos relacionados à autenticação de manuscritos [CRETTEZ, 1995], [SRIHARI et al., 2002], [BULACU et al., 2003]. A verificação de autoria em documentos manuscritos apresenta diferentes abordagens, dependendo do método utilizado na extração de características e do modelo de classificação. As abordagens podem ser, basicamente, classificadas em duas, global e local.

Neste capítulo são apresentados alguns desses trabalhos que foram utilizados como referência.

3.1. Abordagens Locais

Crettez [CRETTEZ, 1995] não trata especificamente de verificação de autoria em seu método. Sua proposta é caracterizar o estilo de escrita do autor e separá-lo em estilos de escrita diferentes, baseada em *fuzzy clusterization*. No seu método não é realizada a análise semântica das palavras. As características utilizadas são: a espessura do traçado (dependendo da pressão exercida pelo escritor, bem como a espessura da ponta da caneta), o corpo principal da palavra (o centro de três zonas verticais da palavra), a densidade espacial de caracteres (estimada pelo número de traços verticais na palavra) e a inclinação axial, sendo esta última a característica principal na abordagem de Crettez [CRETTEZ, 1995].

A obtenção da inclinação axial é feita através do chamado diagrama direcional, que consiste em examinar cada palavra obtendo o histograma das linhas retas que fazem parte do traçado, desenhando um diagrama direcional da inclinação axial do autor. O experimento

utiliza uma base de dados de 3788 palavras, na qual os escritores são agrupados em famílias de estilos de escritores através de *fuzzy clusterization* .

Zois e Anastassopoulos [ZOIS & ANASTASSOPOULOS, 2000] propõem um método que usa somente palavras na identificação do autor. O sistema analisa o manuscrito em dois idiomas, em inglês e grego, para demonstrar que o sistema se adapta a qualquer situação, independente do idioma usado, caracterizando, desta maneira, uma abordagem não-contextual.

A base de dados é composta por 50 autores, os quais escrevem 45 amostras da palavra "característica" escritas em grego e inglês.

Para a retirada de informações desnecessárias das imagens, é aplicada uma fase de pré-processamento utilizando técnicas de binarização e afinamento, eliminando, desta forma, diferenças entre as canetas usadas pelo autor, e tratando problemas encontrados em amostras que não foram digitalizadas corretamente. A partir da imagem pré-processada, são extraídas as projeções horizontais, que são descritores de forma global que provêem um tipo de codificação das imagens em linhas [ZOIS & ANASTASSOPOULOS, 2000]. Estas projeções são processadas morfologicamente para a obtenção do vetor de características.

Para o processo de identificação e verificação de autoria são testados dois classificadores: redes MLP e classificação bayesiana com o método *leave-one-out*. Segundo Zois e Anastassopoulos [ZOIS & ANASTASSOPOULOS, 2000] os resultados obtidos são considerados satisfatórios tanto na identificação de autoria, com taxas de erro em torno de 8% usando classificação bayesiana e 3,5% com o uso de redes MLP, quanto na verificação da autoria alcançando baixas taxas de erro em torno de 5% aplicando classificação bayesiana e 2% usando redes neurais.

Enquanto que a maioria dos autores tratam dos problemas de verificação e identificação usando palavras, letras ou partes do manuscrito, Leedham e Chachra [LEEDHAM & CHACHRA, 2003], em sua abordagem, tratam especificamente de dígitos numéricos para a identificação e verificação da autoria, considerando a escrita natural e incluindo falsificações.

Nem todas as características grafoscópicas são compucionalmente possíveis. Leedham e Chachra [LEEDHAM & CHACHRA, 2003] identificam 11 características grafoscópicas como computacionalmente possíveis, e aplicam-nas na identificação e verificação de autoria de dígitos, citadas a seguir: relação de altura e largura, número de pontos finalizadores,

número de junções, número de *loops,* inclinação, distribuição de altura e largura, densidade de pixels, medida angular, centro de gravidade e gradiente.

A base de dados de dígitos é composta por 15 autores, com 10 linhas de dígitos, usando em cada linha dígitos escritos aleatoriamente de 0 a 9, digitalizados em 300 *dpi.*

Após a extração, as características são armazenadas em um vetor de 961 posições de cada dígito por autor. Para a identificação e verificação, a base de dados, agora convertida em vetor de características, é dividida em um conjunto de vetores s*tandard* (S) e vetores de teste (T) [LEEDHAM & CHACHRA, 2003]. Para qualquer vetor de teste escolhido é computada a distância de Hamming com cada vetor de cada autor no conjunto de vetores *standard* (S). O vetor de teste de dígitos é atribuído ao escritor que produzir a distância de Hamming mínima [LEEDHAM & CHACHRA, 2003]. Os resultados obtidos no processo de identificação alcançam taxas de acerto de 100% e na verificação 80% de acerto, descritos por Leedham e Chachra, [LEEDHAM & CHACHRA, 2003].

3.2.Abordagens Globais

Said [SAID et al., 1998] apresenta um sistema off-line para a identificação de autoria, não-contextual, baseado em reconhecimento de textura, utilizando filtros de Gabor multi-canal e matriz de co-ocorrência de escala de cinza no processo de extração de características. A normalização dos manuscritos ocorre com: a remoção da inclinação da escrita, padronização do tamanho da letra do autor e eliminação de espaços em branco entre palavras e entre linhas.

A base de dados usada por Said [SAID et al. 1998] é composta por 10 autores. O manuscrito de cada autor é dividido em 25 amostras (128 x 128 *pixels*), das quais são usadas: 10 amostras para o treinamento do modelo e 15 para testes, e vice-versa. As amostras que estão no conjunto de treinamento não aparecem no conjunto de testes.

A identificação do autor ocorre usando *K-Nearest Neighbor e Weight Euclidean Distance (WED)* como classificadores. Os melhores resultados foram obtidos usando *WED* chegando a resultados em torno de 95%. Este método foi aplicado posteriormente por Yong [ZHU et al., 1999] na identificação de autores chineses, alcançando resultados semelhantes aos obtidos por Said.

Um outro método, proposto por Bulacu [BULACU et al., 2003], utiliza duas características globais. Neste método são utilizadas somente características que podem ser automaticamente extraídas da imagem dos manuscritos, caracterizando o método como sendo totalmente automático. Bulacu [BULACU et al., 2003] usa características de distribuição angular-direcional sobre as bordas extraídas dos manuscritos, sendo a primeira distribuição de direção-de-borda, e a segunda distribuição de junção-de-bordas.

Na grafoscopia tradicional, as duas características representam respectivamente inclinação axial e movimento durante a escrita, obtendo resultados superiores aos comparados com outras características usadas em sistemas forenses de identificação de escritores [BULACU et al., 2003], tais como entropia de *pixel*s, regularidade da escrita e distribuição de comprimento.

A inclinação axial é obtida através de um diagrama direcional similarmente a [CRETTEZ, 1995], porém a abordagem adotada por Bulacu [BULACU et al., 2003] é global e o diagrama direcional é extraído a partir de bordas do traçado do autor, alcançado resultados melhores que os obtidos por Crettez [CRETTEZ, 1995], por sofrer uma menor influência na espessura do traçado. Já o movimento da escrita é calculado pela junção de dois fragmentos de borda.

Para a avaliação é usada uma base de dados chamada “Firemaker”, com 250 autores holandeses, contendo 2 amostras por autor, colhidas de maneira normatizada nas quais os autores usam mesma marca e tipo de caneta, papel pautado e letra de forma. Para a classificação é usado o *K-NN* em uma estratégia *leave-one-out,* computadando-se a distância euclidiana entre vetor de características de um autor escolhido contra todos os vetores de características dos outros autores. Neste método não há separação entre conjunto de treinamento e teste, sendo as 500 amostras usadas apenas em um conjunto de teste. Os resultados dos testes realizados alcançaram resultados em torno de 90% de acerto na identificação da autoria [BULACU et al., 2003].

3.3. Abordagens Globais e Locais

Um sistema que utiliza características globais e locais é proposto por Schlapbach & Bunke [SCHLAPBACH & BUNKE, 2004]. O método trata de imagens de linhas escritas por

autor, para o treinamento e classificação. As imagens são normalizadas removendo a inclinação das palavras e das linhas em relação a uma linha base imaginária.

Para a extração de características globais são extraídas: fração de *pixels* pretos da imagem, centro de gravidade e momentos de 2ª ordem. As características locais são: posição do *pixel* superior e inferior, transição entre *pixels* pretos e brancos locais, fração de *pixels* entre o *pixel* superior e inferior.

Neste trabalho foi usado o HMM como classificador para a tarefa de identificação e verificação de autor. Para cada autor o HMM é treinado, tendo para diferentes autores diferentes HMMs (*n* autores, *n* HMMs). Por serem treinados para cada autor, os HMMs possuem transições e probabilidades diferentes dependendo do autor, ficando, assim, o HMM especializado para cada autor treinado.

Na fase de classificação, um texto desconhecido é apresentado ao HMM. A saída do identificador é uma transcrição da entrada juntamente com um *score* de reconhecimento em termos de probabilidade. Desta forma, as saídas são ordenadas em ordem decrescente dos *scores* de reconhecimento.

Baseado no *ranking* pode-se identificar o autor da linha ou, no caso da verificação da autoria, dizer se a escrita pertence ao escritor. Considera-se que o correto reconhecimento das palavras tem um *score* mais alto em relação ao *score* de reconhecimento incorreto.

Para toda inclusão de novos autores o sistema é treinado. Para identificação e treinamento são usadas 4.307 linhas de 100 diferentes autores. São divididos em quatro conjuntos, 3 para treinamento e 1 para teste. Os conjuntos de treinamento não aparecem no conjunto de teste. Nos experimentos para a identificação de autoria Schlapbach e Bunke, obtiveram 96% de autores corretamente identificados, [SCHLAPBACH & BUNKE, 2004].

Para o processo de verificação, a probabilidade de o escritor questionado ser o autor é comparada com a média da melhores probabilidades geradas, usando um intervalo de confiança e atribuindo a autoria se o autor estiver acima do limiar. Nos experimentos para a verificação de autoria Schlapbach e Bunke, taxas de erro inferiores a 3%, [SCHLAPBACH & BUNKE, 2004].

Srihari [SRIHARI et al., 2002] propõe um método de identificação e autenticação de autoria embasado na grafoscopia. As características são divididas em macro características (globais) aplicadas aos documentos inteiros, parágrafos e palavras, podendo ser aplicadas também aos caracteres; e micro-características (locais) aplicadas sobre as letras. Para a

extração de características locais, os manuscritos são segmentados manualmente em parágrafos, palavras e letras.

As características globais são compostas pelas seguintes características: (i) entropia de níveis de cinza, (ii) limiar do valor de níveis de cinza, (iii) número de *pixels* pretos, (iv) número de contornos (exteriores e interiores), (v) números de *slope* (vertical, horizontal, negativo, positivo), (vi) inclinação axial, (vii) altura. As características locais são: gradiente e concavidade.

Srihari [SRIHARI et al., 2002] mostra uma equivalência entre as características computacionais globais e as características da grafoscopia tradicional, porém não cita nenhuma equivalência entre as características locais.

O método de identificação usa o *K-NN,* com a estratégia *leave-one-out,* assim confrontando um autor conhecido com uma base de autores desconhecidos, sendo considerado um problema de *n*-classes. Srihari [SRIHARI et al., 2002] realizou testes para os diferentes níveis: documento inteiro, parágrafo e palavra. Para características globais os melhores resultados alcançados ao nível de documento inteiro foram 96% de acerto considerando 10 autores e 60% de acerto para 900 autores. Para as características locais os resultados são similares aos das características globais, sendo que os melhores resultados são encontrados combinando características globais e locais: 98% de acerto para 10 autores e 89% para 900 autores.

No método de verificação a base de dados é dividida em treinamento, validação e teste. Neste método uma amostra de um autor desconhecido x é confrontada com uma amostra de autoria y, atribuindo autoria ou não autoria ao manuscrito x, caracterizando um problema de 2-classes. Para a classificação são usadas Redes Neurais Artificiais. As taxas de acertos encontradas são de 95% considerando o documento inteiro e 96% combinando características globais e locais.

Para a avaliação do método proposto é usada a base de dados CEDAR [SRIHARI et al., 2002], constituída de 1000 autores, 3 amostras por autor, totalizando 3000 amostras.

A Tabela 3.1 abaixo demonstra o resumo das abordagens apresentadas neste capítulo.

Tabela 3.1: Resumo do Estado da Arte

Referência	Características	Classificador
[CRETTEZ, 1995]	Local: Espessura do traçado, o corpo principal da palavra, densidade espacial de caracteres, inclinação axial.	Fuzzy Clusterization
[ZOIS & ANASTASSOPOULOS, 2000]	Local: Projeções horizontais	*K-NN*
[LEEDHAM & CHACHRA, 2003]	Local: Relação de altura e largura, número de pontos finalizadores, número de junções, número de *loops,* inclinação, distribuição de altura e largura, densidade de *pixels*, medida angular, centro de gravidade e gradiente.	Medidas de Distância
[SAID et al., 1998]	Global: Textura.	*WED* e *K-NN*
[ZHU et al., 1999]	Global: Textura.	*WED* e *K-NN*
[BULACU et al., 2003]	Global: Inclinação axial, mudança de movimento.	*K-NN*
[SCHLAPBACH & BUNKE, 2004]	Global e Local: Globais: fração de *pixels* pretos da imagem, centro de gravidade e momentos de 2ª ordem. Locais: número de transições entre *pixels* pretos e brancos locais, fração de *pixels* pretos entre o *pixel* superior e inferior.	*HMM*
[SRIHARI et al., 2002]	Global e Local: Globais: (i) entropia de níveis de cinza, (ii) limiar do valor de níveis de cinza, (iii) número de *pixels* pretos, (iv) número de contornos (exteriores e interiores), (v) números de *slopes* (vertical, horizontal, negativo, positivo), (vi) inclinação axial, (vii) altura. Locais : gradiente e concavidade.	*KNN e RNA*

3.4. Visão Crítica

A visão crítica do estado da arte busca contribuir para uma abordagem mais consistente e que possua uma conotação prática. Entretanto, uma comparação mais refinada das abordagens em termos de resultados estatísticos torna-se difícil, justamente pela heterogeneidade das bases de dados encontradas, e também por não existir uma base de dados internacional [LEEDHAM, 1994].

Algumas abordagens baseadas em características globais [SCHLAPBACH & BUNKE, 2004], [ZHU et al., 1999], [SAID et al., 1998], retiram dos manuscritos, no processo de normalização, aspectos importantes relacionados a características do autor, tais como espaçamento entre linhas, inclinação e tamanho da letra.

As abordagens baseadas em características locais possuem uma capacidade discriminatória maior com relação às variabilidades [CRETTEZ, 1995], porém o processo de segmentação é transformado em uma tarefa árdua e demorada por ser feito manualmente. Isso ocorre pelo fato de não existirem métodos completamente automáticos para a segmentação de palavras ou caracteres.

A quantidade de autores na base de dados em alguns métodos é bastante reduzida para uma validação estatística, [SAID et al. 1998], [LEEDHAM & CHACHRA, 2003]. Outras possuem quantidade suficiente para uma avaliação estatística. No entanto, há pouca variabilidade entre as palavras [ZOIS & ANASTASSOPOULOS, 2000], o que inviabiliza a aplicação real, pois em um caso real podem ocorrer variações na escrita como um todo. Em algumas abordagens [ZOIS & ANASTASSOPOULOS, 2000], [BULACU et al., 2003] a formação da base de dados é realizada restringindo o estilo de escrita do autor com caneta e papel pautado, prejudicando a implementação de novas características grafoscópicas, tais como inclinação em relação à linha base e restrição do espaço gráfico.

Comumente, a maioria dos métodos apresentados neste capítulo utilizam uma abordagem policotômica [SRIHARI et al.,2002],[BULACU et al., 2003] para a identificação. Nota-se que estas abordagens possuem taxas de erros baixas para poucos autores, sendo consideravelmente elevadas quando o número de autores é aumentado. A necessidade do modelo de conhecer todos os autores reside no fato da não-generalização do sistema caso novos autores sejam incluídos. Já a abordagem dicotômica, usada por [SRIHARI

et al.,2002], necessita apenas de um treinamento, criando assim, um processo de generalização do modelo.

3.5.Comentários Finais

Os trabalhos apresentados neste capítulo contribuíram para a elaboração do presente trabalho no que se refere ao tipo de abordagem empregada, escolha de características e classificadores. Desta forma, este trabalho apresenta uma abordagem baseada nos preceitos da grafoscopia, tendo como característica principal elementos grafocinéticos, em uma arquitetura de duas classes (autoria e não autoria), usando *Support Vector Machine* (SVM). No capítulo a seguir é descrita a metodologia.

Capítulo 4

Metodologia

Neste capítulo é detalhada a metodologia aplicada na abordagem proposta para a verificação da autoria de manuscritos. É apresentado um estudo do processo de geração e aquisição dos modelos de cartas forenses, fundamentais na análise de documentos questionados, assim como a extração de características, medidas de distância e o modelo global.

4.1. Requisitos

Uma análise mais profunda sobre a aplicação da maioria das abordagens, para verificação da autoria de manuscritos, impõe restrições limitadas a um problema em específico. A redução do escopo contribui para o alcance dos resultados desejáveis, porém, seu uso prático pode ficar comprometido quando o meio utilizado difere do meio proposto. Esta abordagem não visa solucionar de forma abrangente o problema em questão. No entanto, apresenta uma proposta viável, na prática, desde que respeitando alguns critérios.

Os requisitos principais observados na elaboração desse método foram:

- Ser tolerante às variações pessoais e intolerante as similaridades interpessoais;
- Respeitar os princípios determinados pela Grafoscopia;
- Reduzir a complexidade do processo de geração do modelo (autor e não autor).

4.2.Recursos

O objetivo desta seção é descrever detalhadamente a solução proposta para o problema da verificação da autoria de manuscritos, levando em conta cada um dos requisitos estabelecidos. A mesma subdivide-se em 5 partes principais, que vem ao encontro dos requisitos que a abordagem exige:

1. Produção do modelo da carta forense e colheita dos espécimes para a formação da base de dados utilizada para a validação de sistemas computacionais;
2. Uso de técnicas grafoscópicas para a extração de características, em uma abordagem global e simplificada;
3. Modelo de verificação com duas classes, autoria e não autoria;
4. Utilização de medidas de distância na análise do espécime conhecido e questionado;
5. E finalmente, a combinação de todos estes recursos, sendo abordada no Capítulo 5.

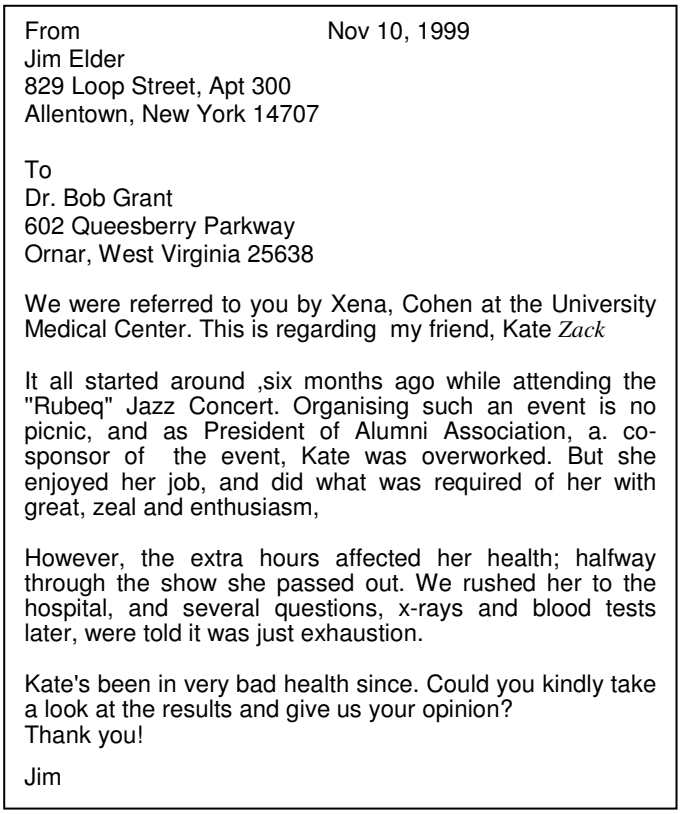
4.3.Modelos de Cartas Forenses

Atualmente alguns modelos de cartas foram criados, baseados nas características dos modelos tradicionais abordados na seção 2.1.4, tendo por objetivo a colheita e criação de bases de dados que permitam a avaliação e validação de sistemas automáticos e semi-automáticos.

O CEDAR, *Center of Excellence for Document Analysis and Recognition*, apresenta um modelo denominando Carta CEDAR, que é usada para a coleta de exemplares de 1000 autores de diferentes regiões dos Estados Unidos da América, na língua inglesa [CHA, 2001], Figura 4.1(a).

A ausência de um modelo para escrita latina, mais especificamente a de língua portuguesa, motivou a criação de um modelo pela PUCPR (Pontifícia Universidade Católica do Paraná), do Laboratório de Direito e Tecnologia (LADITEC), denominado Carta PUCPR, modelo aplicado para avaliação computacional pioneiro no Brasil. O modelo PUCPR contém todas as particularidades da escrita da língua portuguesa, tais como acentuação, mínimos

gráficos, contendo um léxico de 124 palavras e objetivando a coleta de manuscritos de 500 autores [JUSTINO, 2002].



From Nov 10, 1999
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Bob Grant
602 Queesberry Parkway
Ornar, West Virginia 25638

We were referred to you by Xena, Cohen at the University Medical Center. This is regarding my friend, Kate *Zack*

It all started around ,six months ago while attending the "Rubeq" Jazz Concert. Organising such an event is no picnic, and as President of Alumni Association, a. co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great, zeal and enthusiasm,

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?
Thank you!

Jim

(a)

De
Fernando Quintas Zanon
Rua Luiz Kirt Walterez, 87 - Ap. 300
Xenápolis, NovaYolanda 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 esta vaga. Sou brasileiro, solteiro, com 18 anos, curso a 3ª série do Curso Técnico de Contabilidade do Colégio Horácio Alves - Escola Municipal de 2° 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 â 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

(b)

Figura 4.1 (a) Carta CEDAR (b) Carta PUCPR.

O aspecto mais importante em relação às bases de dados é a viabilização das pesquisas na área de autenticação de documentos manuscritos e áreas afins para permitir a validação estatística.

4.4.Colheita

A aquisição dos manuscritos é feita através da colheita (Seção 2.1). A colheita foi realizada entre voluntários, na maioria durante seções programadas em instituições de ensino no decorrer dos anos de 2002 a 2005, aos quais é apresentando o modelo que deve ser transcrito na íntegra 3 vezes em folhas de papel A4(21 x 29,7 cm), sem pauta. A repetição na transcrição do modelo possibilita o mapeamento das variabilidades intrapessoais.

Juntamente com os manuscritos são colhidas informações sobre o autor, tais como nome, idade, sexo, destro ou canhoto e grau de escolaridade.

No processo de colheita algumas restrições são feitas aos autores:

- Uso de caneta esferográfica azul ou preta;
- Não hifenização das palavras, caso falte espaço na linha;
- Texto escrito sem auxílio de linhas guia.

As restrições não devem influenciar no processo de escrita do autor, pois no manuscrito ficarão todas as características de sua escrita que serão analisadas posteriormente para a verificação da autoria [JUSTINO, 2002].

Colhidos os manuscritos, o processo de criação da base digital, assim como etapas de pré-processamento serão descritos no Capítulo 5.

4.5.Extração de Características

Como visto no capítulo 2, características são quaisquer medidas extraíveis de um padrão, podendo contribuir para a classificação. A escolha de boas características é fundamental para a robustez de um método computacional de verificação de autoria. Desta forma, o estudo das técnicas de reconhecimento de padrões e o estudo da grafoscopia contribuíram diretamente para a construção do método proposto. Estes estudos auxiliaram no processo de identificação de características que conseguem mapear os padrões de escrita e extrair as características inerentes ao autor para a posterior verificação do manuscrito.

A grafoscopia tradicional engloba várias características no processo de análise e identificação e verificação da autoria, tais como estilo da escrita, embelezamento, inclinação axial, nível de habilidade, pressão exercida pelo autor no objeto de escrita, entre outras (seção 2.1.1). Porém, algumas destas características, atualmente avaliadas subjetivamente pelo perito, não são computacionalmente de fácil implementação. No entanto, soluções estão sendo desenvolvidas [JUSTINO, 2001], [SRIHARI et al., 2002], [BULLACU et al., 2003], [LEEDHAM & CHACRA, 2003]. Algumas dessas soluções serviram de base para este trabalho, já vistas no Capítulo 3.

4.5.1. Inclinação Axial

A inclinação axial é o ângulo de inclinação da escrita em relação ao eixo vertical, sendo o eixo horizontal representado por uma linha de base imaginária. Essa inclinação pode

ocorrer à direita, à esquerda ou ser nula (alinhada ao eixo vertical), conforme a Figura 2.5. A inclinação axial da escrita pode ocorrer desde o princípio de uma palavra até o final da mesma, ou desde o princípio de uma oração, parágrafo ou página até o final das mesmas. Não são raros os casos em que o escritor apresenta um misto dessas inclinações, ou seja, inclinações à esquerda, à direita e alinhada ao eixo vertical (inclinação axial nula).

Se essa mudança de inclinação é habitual, ela pode ser considerada uma característica identificadora. Embora esteja longe de ser uma regra definitiva, uma escrita com inclinação axial à direita ou à esquerda pode indicar a existência de um escritor destro ou canhoto, respectivamente. No entanto, é comum entre os escritores canhotos o posicionamento inclinado do documento, no momento da escrita. Esse procedimento tem como objetivo produzir uma inclinação axial nula, ou quase nula. Esse hábito pode ser considerado uma característica ou estilo do escritor [JUSTINO, 2002].

Na prova pericial, os peritos quantificam as variações na inclinação axial encontradas por todo o documento dando a essa inclinação uma avaliação aproximada. Para tanto, leva-se em consideração o comportamento em todo o texto analisado.

A abordagem global de análise visual executada pelo perito carrega consigo medições subjetivas e de pouca precisão. No entanto, essa característica se mostra extremamente rica em elementos grafocinéticos e que, devidamente avaliados, podem contribuir de forma decisiva na identificação do autor [JUSTINO, 2002]. Além disso, resultados já apresentados por outros autores [CRETTEZ, 1995], [BULACU et al., 2003], demonstram computacionalmente o potencial dessa característica. Assim sendo, a mesma foi escolhida como primeira a ser testada no método proposto por este trabalho.

A inclinação axial é uma característica grafoscópica computacionalmente viável, sendo algumas soluções encontradas na literatura [JUSTINO, 2001], [CRETTEZ, 1995], [BULACU et al., 2003]. Porém, duas delas se destacam. Na primeira a inclinação é combinada com várias outras características extraídas do autor, usando-se apenas um ângulo predominante da inclinação geral da escrita e apenas uma direção angular [SHIHARI et al., 2003]. Na segunda é utilizada uma distribuição de tendências angulares de inclinação [BULACU et al., 2003].

O uso de apenas um ângulo na inclinação axial não abrangeria todos os elementos grafocinéticos da escrita de um autor, redundando em uma avaliação similar a da utilizada

pelo perito. Assim sendo, foi adotada neste método a distribuição de tendências angulares de inclinação.

4.6.Medidas de Distância

Existem dois pontos cruciais no desenvolvimento de um sistema automático de verificação da autoria de manuscritos computadorizado. O primeiro é alcançar uma representação que pode maximizar a distância entre manuscritos de indivíduos diferentes. O outro seria manter os manuscritos de mesma pessoa constantes, uma vez que uma medida de distância seja corretamente definida.

4.6.1. Distância Euclidiana

Distância Euclidiana é derivada do cálculo da distância geométrica entre dois pontos. A Distância Euclidiana é adequada para o tratamento de classes cujos elementos tendem a se agrupar próximos à média, ou seja, possuem variância desprezível. Problemas em que as classes apresentam comportamento semelhante quanto à forma da função de distribuição de probabilidades e valores de variância também são indicados para o uso da Distância Euclidiana.

A Distância Euclidiana entre duas amostras x_i e x_j é a raiz quadrada do somatório das diferenças entre os valores de x_i e x_j para todas as variáveis, ou seja, para k = 1,..., L [BARBEAU et al., 2002]:

$$d_{ij} = \sqrt{\sum_{k=1}^{L} (x_{ik} - x_{jk})^2} \qquad (4.1)$$

4.7.Modelo Pessoal e Modelo Global

Os sistemas automáticos de verificação da autoria de textos manuscritos baseiam-se usualmente em duas abordagens de modelos para classificação, a pessoal e a global [JUSTINO et al., 2003].

Algumas abordagens utilizam o modelo pessoal [SRIHARI et al, 2002], [SCHLAPBACH & BUNKE, 2004], usando um modelo por autor e baseando-se no conceito da policotomia, ou seja, a classificação do problema em n-classes. Nesse modelo, cada autor representa uma classe. Usualmente, o modelo pessoal exige um conjunto elevado de exemplares genuínos para sua geração, pois para cada autor será gerado um modelo específico e que descreve adequadamente as características do mesmo. Este modelo apresenta a vantagem de descrever adequadamente as variabilidades intrapessoais do autor, apresentando, porém, a desvantagem da geração de um novo modelo a cada inclusão de um novo autor.

Já o modelo global utiliza o conceito da dicotomia, ou seja, a divisão do modelo em 2 classes, sendo elas 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. O modelo global possui a desvantagem da generalização. No entanto, possui a vantagem de necessitar de um número reduzido de exemplares de cada autor e de não necessitar de um novo treinamento do modelo, na inclusão de novos autores.

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.

O modelo global foi o escolhido para esse trabalho por apresentar as características de simplicidade do processo de treinamento e geração do mesmo e por ser desnecessário o treinamento na inclusão de novos autores, o que viabiliza sua aplicação em situações reais.

4.8.Comentários Finais

A metodologia proposta apresentada neste capítulo relacionado ao problema de verificação da autoria de documentos manuscritos foi detalhada em pontos essenciais para a abordagem como: características, medidas de distância e modelo global.

Será apresentada no próximo capítulo tanto a visão do perito na análise de manuscritos questionados como as etapas que envolvem o método proposto.

Capítulo 5

Método Proposto

Neste capítulo é detalhado o método proposto para a verificação de autoria de manuscritos. É apresentada a visão da perícia grafoscópica aplicada à análise de manuscritos questionados em comparação com o modelo computacional proposto, assim como suas fases.

5.1. Perícia Grafoscópica em Manuscritos

A grafoscopia convencional utilizada na análise forense de documentos questionados é essencial para o desenvolvimento de uma abordagem computacional automática nessa área (Seção 2.1). Com base neste estudo, procura-se montar uma estrutura que utilize todos os recursos da grafoscopia procurando assemelhar-se à visão que o perito tem no processo de análise de um documento manuscrito, dentro de um método computacional.

Os peritos grafotécnicos classificam os textos manuscritos em relação à autoria como: associação w_1 ou dissociação w_2 [JUSTINO, 2002]. A associação indica que a grafia presente no manuscrito foi elaborada, de próprio punho, pelo autor avaliado. Já a dissociação indica que o manuscrito não foi elaborado, de próprio punho, pelo autor avaliado.

A idéia é baseada em um processo de comparação, no qual, o perito, de posse de algumas amostras genuínas do autor analisado, as compara com a amostra de autoria desconhecida ou questionada, com base nas características grafoscópicas. O processo de decisão é baseado em uma métrica de similaridade observada na amostra questionada, em relação a todas as amostras genuínas, permitindo ao perito atribuir ou não a autoria do documento em questão.

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 conhecida (referência) M_{Ki} *(i=1,2,3...n)*, em comparação com 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 amostra 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 5.1, [SANTOS, 2004].

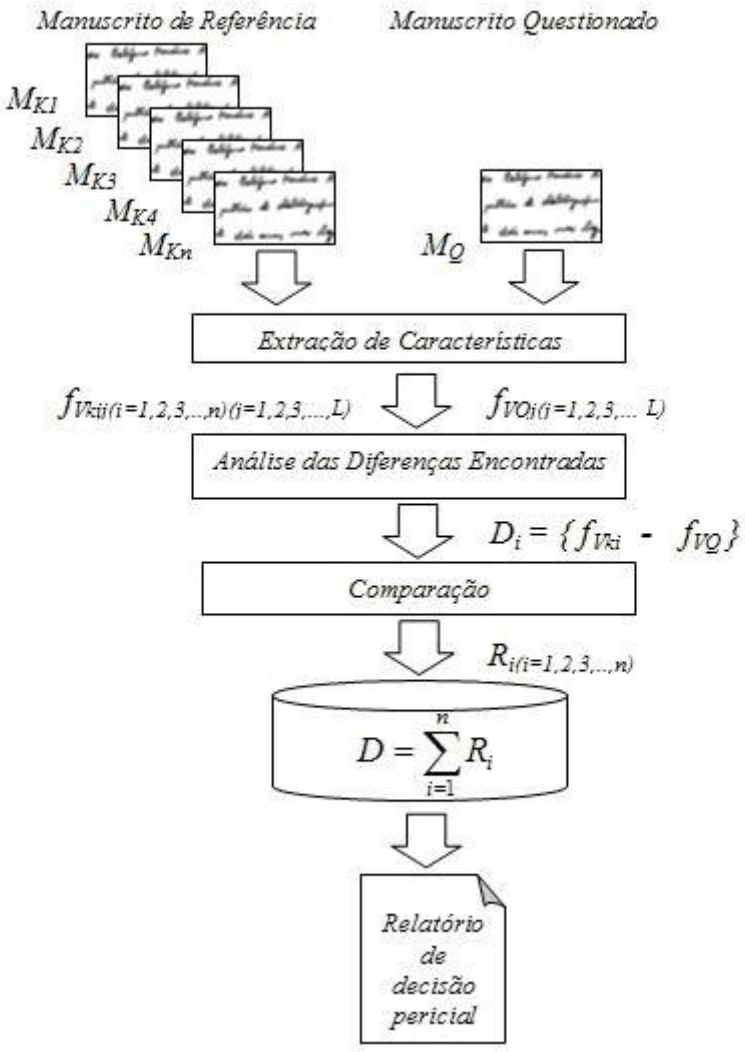
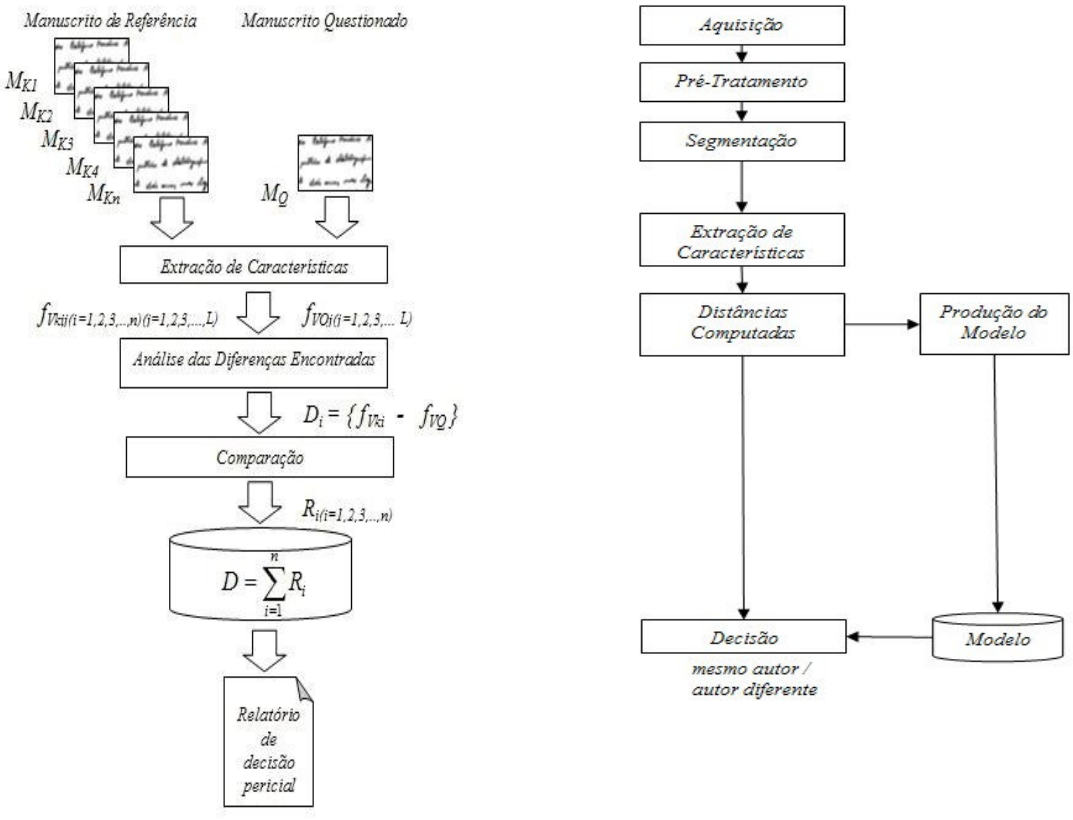


Figura 5.1: Esquema do processo de decisão na verificação de manuscritos baseado na visão pericial.

Os procedimentos utilizados pelos peritos propiciam o desenvolvimento de uma abordagem automática incorporando características da grafoscopia, medidas de distância, e processo de decisão usando voto majoritário, já proposto por Santos para autenticação de assinaturas [SANTOS, 2004]. As etapas desse processo estão descritas nas seções posteriores.

5.2. Etapas do Processo de Verificação de Autoria em Manuscritos

O desenvolvimento de um sistema automático de verificação de autoria de manuscritos requer alguns procedimentos adicionais, conforme a Figura 5.2(b) demonstra, em comparação com a análise pericial, Figura 5.2(a).



(a) (b)

Figura 5.2: Um comparativo das etapas no processo de verificação de manuscritos: (a) processo de análise e decisão pericial; (b) processo computacional proposto.

No desenvolvimento do sistema de verificação de manuscritos, tornam-se indispensáveis algumas etapas, como segue:

- **Aquisição dos dados:** imagem do manuscrito produzida a partir de um *scanner*;
- **Pré-processamento:** preparação da imagem para a extração de características através de binarização e extração do contorno;
- **Segmentação:** divisão da imagem do manuscrito em diversos fragmentos usados na fase de extração de características;
- **Extração de características:** extração, a partir dos manuscritos, de características inerentes ao autor;
- **Distância entre as características:** diferença entre os vetores de características extraídas, usada 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;
- **Processo de decisão**: avaliação da saída do modelo produzido, verificando se o manuscrito caracteriza associação (autoria) ou dissociação (não autoria).

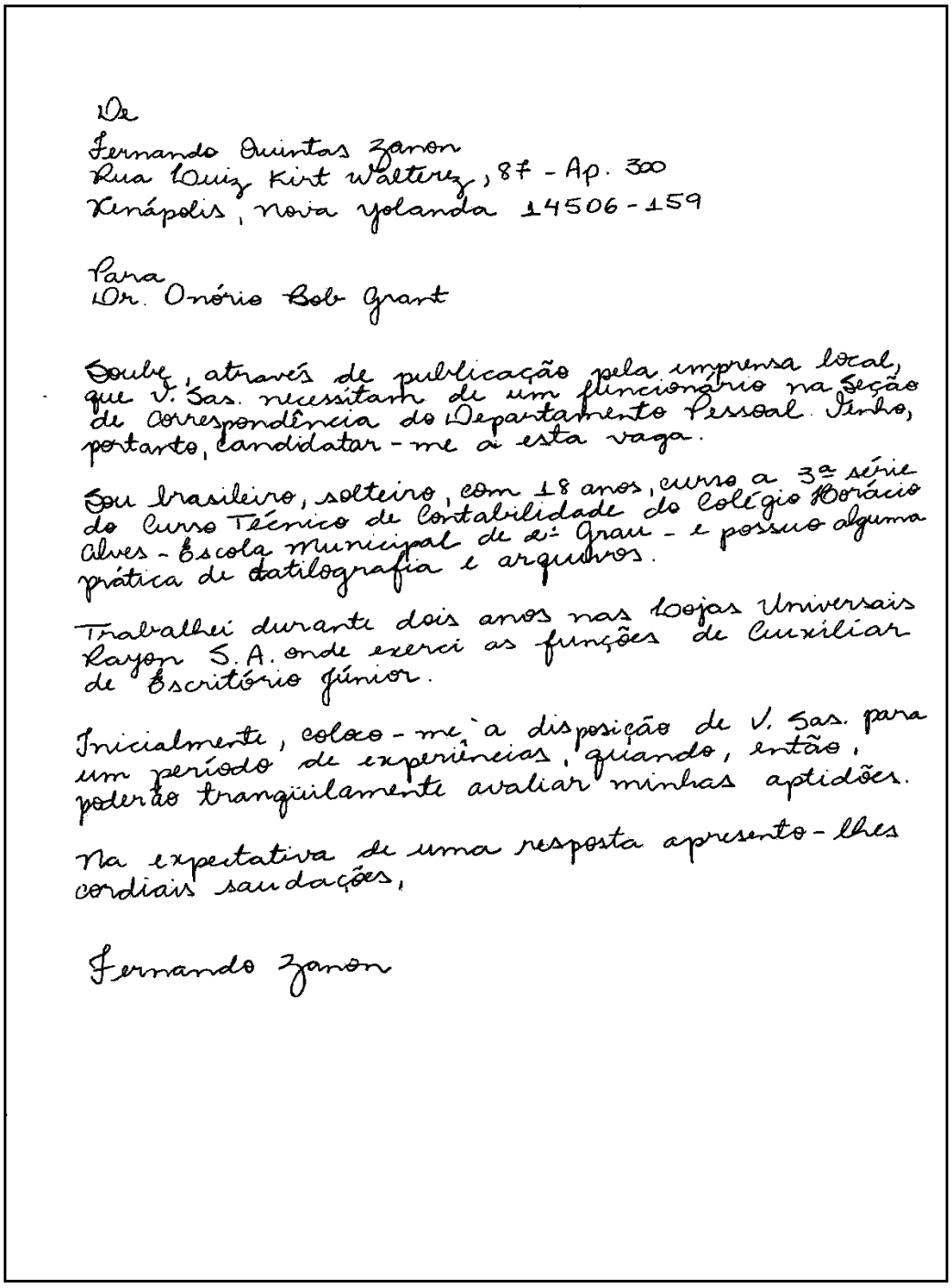
A seguir, cada etapa é detalhada demonstrando os aspectos genéricos e as particularidades encontradas em termos do método de verificação de autoria de manuscritos proposto.

5.2.1. Aquisição dos Dados

Para a avaliação do desempenho de um método de verificação da autoria de manuscritos, um fator importante é a análise da composição da base de dados usada para validar os procedimentos de aprendizado e de testes. Esta deve conter um número suficiente de autores, permitindo a validação estatística. O número de espécimes por autor a ser utilizado é outro fator relevante, pois deve representar satisfatoriamente as variações intrapessoais de cada autor.

No Capítulo 4 foram vistos os procedimentos para a obtenção dos exemplares físicos (manuscritos) da base de dados. Posteriormente a esta fase, é realizada a digitalização dos manuscritos em um *scanner Hewllet-Packward* modelo *HP Deskjet 5550c*, com 256 níveis de cinza, densidade de 300 *dpi*, formato *Bitmap* (.BMP) sem interferência de ruídos ou imagens pré-impressas.

Atualmente, a base de dados é composta por 315 autores, sendo 3 amostras por autor, totalizando 945 imagens. A Figura 5.3 exemplifica uma amostra de manuscrito. As bases de dados dos manuscritos, física e digital, encontram-se sobre os cuidados do LADITEC (Laboratório de Direito e Tecnologia) ligado ao Programa de Pós-Graduação em Informática (PPGIA) e ao Programa de Pós-Graduação em Direito Econômico e Social (PPGDES).



De
Fernando Quintas Zanon
Rua Luiz Kirt Walter, 87 - Ap. 300
Xinápolis, 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. Tenho, 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 experiências, quando, então, poderão tranqüilamente avaliar minhas aptidões.

Na expectativa de uma resposta apresento-lhes cordiais saudações,

Fernando Zanon

Figura 5.3: Amostra de manuscrito digitalizado e armazenado na base de dados.

A base de dados PUCPR, como visto anteriormente, possui três amostras por autor visando o mapeamento das variabilidades da escrita dos autores. Porém, um problema que não pode ser considerado é a alteração da escrita devido a alguns fatores como, por exemplo, estado psicológico e idade cronológica [LEEDHAM & CHACHRA, 2003].

Uma forma de evitar esse problema seria obter uma estimativa real das variações da escrita. Os manuscritos deveriam ser colhidos em sessões programadas por um longo período [FANG, 2003]. Entretanto, esta colheita se mostrou inviável em função das dificuldades em se conseguir autores voluntários num volume suficiente para esse trabalho.

O problema de colheitas regulares é, porém, minimizado, tendo em vista a abordagem proposta, pois esta requer apenas um único treinamento do modelo, podendo generalizar as classificações para inclusão de novos autores e pequenas variações na escrita.

Manuscritos colhidos em períodos regulares ou irregulares tendem a possuir distâncias semelhantes de suas características se comparados com manuscritos da mesma época da colheita, fazendo com que os mesmos se enquadrem na classe autoria.

No Capítulo 6, será detalhado o protocolo de divisão da base de dados, tanto para treinamento quanto para teste. Nesta abordagem as amostras dos manuscritos passam a ser representadas pelas distâncias computadas entre características de dois manuscritos, e não mais pelas características extraídas destes. Ocorre assim uma combinação entre as distâncias de características extraídas de cada fragmento de manuscrito com o objetivo de se obter um número maior de amostras intra e interclasse.

5.2.2. Pré-Processamento

O pré-processamento é uma etapa primordial, pois prepara a imagem do manuscrito retirando detalhes desnecessários para as etapas posteriores, tais como, segmentação e extração de características.

Binarização

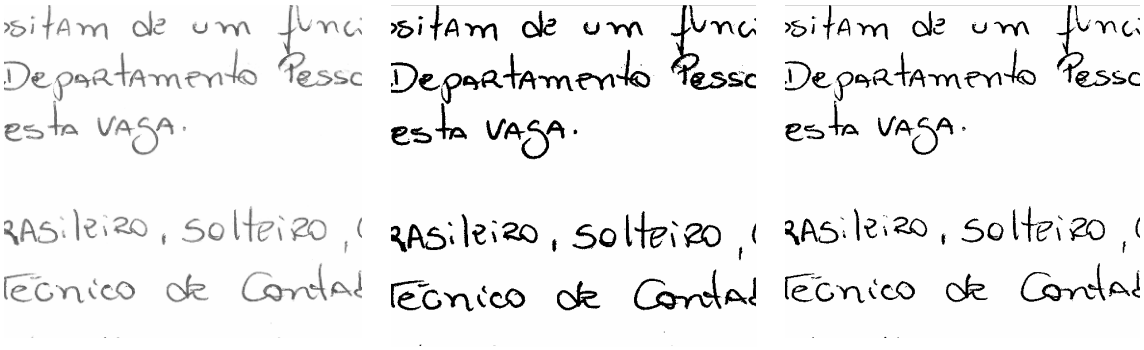
A binarização consiste no processo de transformação de uma imagem em 256 níveis de cinza, a partir de um valor limiar, em uma imagem binária, ou seja, preto e branco. Sendo assim, há uma redução na quantidade de dados a serem tratados, eliminando ruídos e

facilitando a extração de componentes relevantes da imagem além de reduzir drasticamente o tamanho da imagem em *bytes*.

São duas as principais abordagens de binarização, a global e a local. A binarização global é utilizada em imagens onde os valores dos *pixels* dos componentes da imagem e os do fundo são razoavelmente consistentes em seus respectivos valores sobre a imagem inteira. Então, um simples valor de limiar pode ser encontrado para esta imagem e usado para todos os *pixels* da imagem. A binarização local é usada em imagens que possuem variação de elementos de primeiro e segundo plano. Para tanto os limiares variam de acordo com a área.

Através do software FEPI[1] , testes foram realizados para constatar qual abordagem, global ou local, apresentaria resultados satisfatórios.

Dentre as várias técnicas utilizadas, as que apresentaram melhores resultados, sem causar perda de informação nas imagens, foram a binarização global de Otsu [OTSU, 1979] e a binarização local por entropia bidimensional de Abutaleb [ABUTALEB, 1989], que podem ser observadas na Figura 5.4.



(a) (b) (c)

Figura 5.4 (a) imagem em 256 níveis de cinza; (b) imagem binarizada por Entropia de Abutaleb; (c) imagem binarizada por Otsu.

Para o trabalho proposto assume-se a perspectiva da binarização local. Em função do uso de uma abordagem global, para a extração de características, foi utilizada a binarização local por entropia de Abutaleb, pois esta técnica aplica limiares de acordo com a área da imagem. Desta forma, ela conserva detalhes importantes no traçado original da imagem, os quais não se mantinham na binarização global de Otsu, (Figura 5.4).

[1] FEPI: Ferramenta de Processamento de Imagens, http://www.ppgia.pucpr.br/~facon/MaterialGraduacao2005.html, 2005

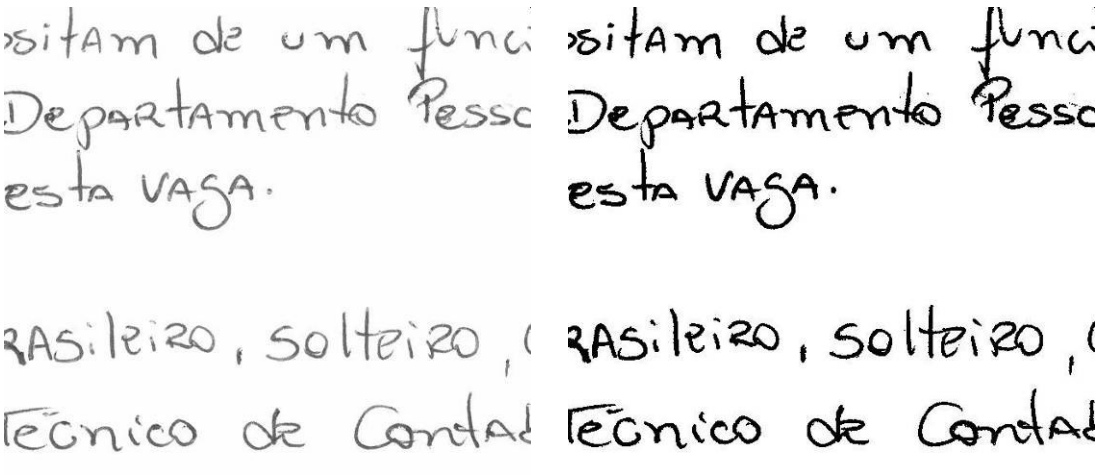
Detecção de Bordas por Dilatação e Erosão

A utilização de morfologia matemática na detecção de bordas por dilatação e erosão é um processo que faz extração de contornos bem definidos, usados posteriormente para a extração da inclinação axial, detalhada na seção 5.2.3.

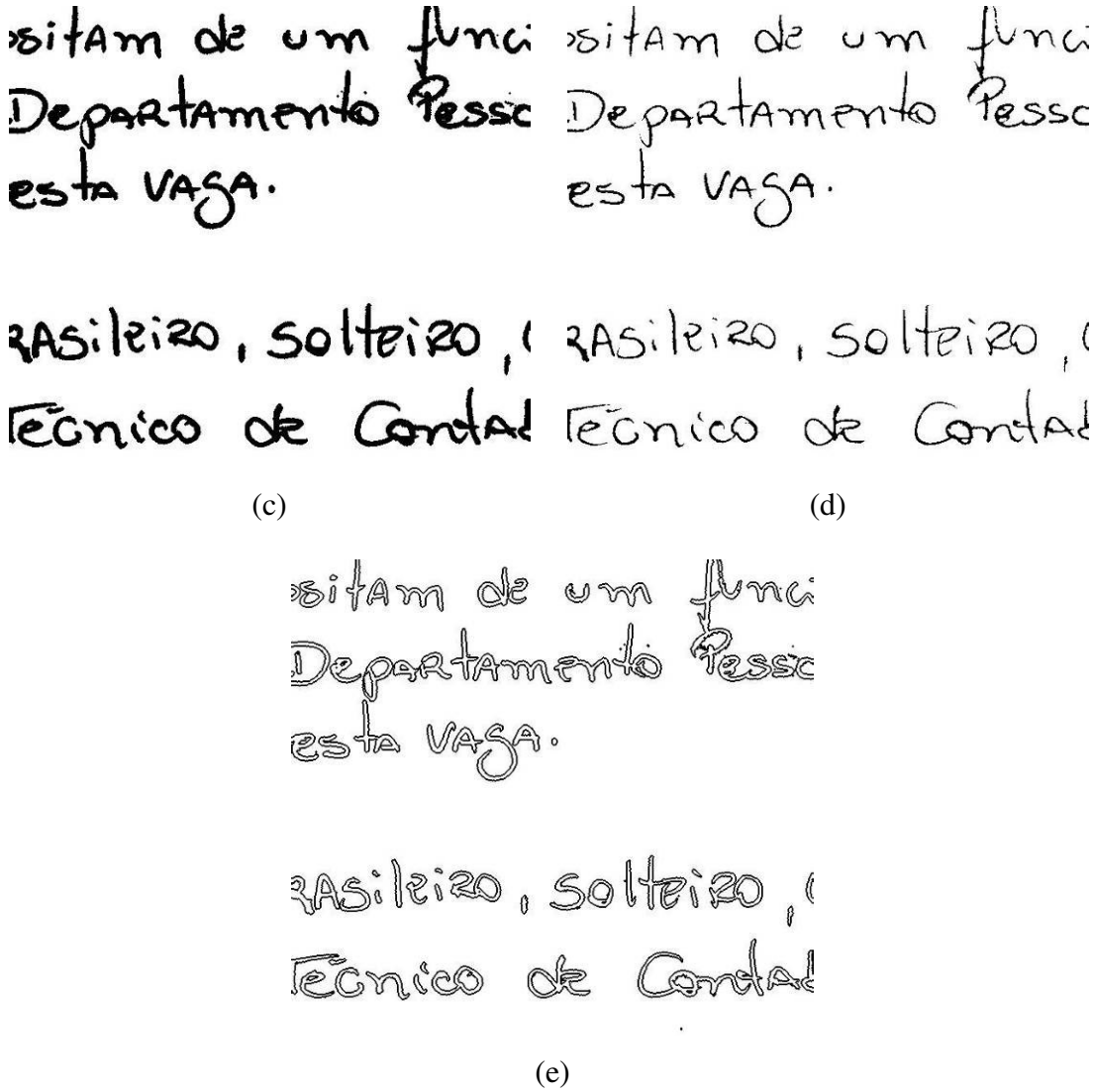
O filtro morfológico de dilatação modifica a imagem de um manuscrito através de um elemento estruturante em cruz, deixando o traçado do autor mais espesso, Figura 5.5(c). O principal efeito da dilatação é incrementar *pixels* nas bordas da imagem [STERNBERG, 1986].

Já o processo do filtro morfológico de erosão faz o processo inverso, aplicando o elemento estruturante em cruz e deixando o traçado do autor mais fino, Figura 5.5(d). O principal efeito da erosão é decrementar *pixels* nas bordas da imagem [STERNBERG, 1986]

Após a geração da imagem dilatada e imagem erodida, estas são sobrepostas e é feita a subtração dos *pixels*, compatíveis nas duas imagens resultando na imagem de borda. As imagens de bordas apresentam contornos bem definidos com espessura de 1 a 3 *pixels,* Figura 5.6(e).



(a) (b)



(c) (d)

(e)

Figura 5.5: (a) imagem em 256 níveis de cinza; (b) imagem binarizada; (c) imagem dilatada; (d) imagem erodida; (e) imagem de contorno resultante.

Segmentação

Conforme visto no Capítulo 3, a segmentação pode ocorrer em vários níveis para o processo de identificação e/ou verificação de autoria [SRIHARI et al., 2002], [ZOIS & ANASTASSOPOULOS, 2000], [SAID & BAKER, 1998], [BULACU et al., 2003], dependendo da abordagem de extração de características adotada, podendo ser global ou local (Seção 2.3).

Geralmente, nas abordagens locais, a segmentação consiste em dividir o manuscrito visando retirar partes de interesse, como palavras ou letras, [SRIHARI et al., 2003], [ZOIS & ANASTASSOPOULOS, 2000]. A vantagem neste processo de segmentação consiste na análise de particularidades da escrita, consideradas discriminantes pela grafoscopia, podendo ser observadas apenas localmente, tais como pingos das letras “i”, cortes das letras “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 [SRIHARI et al., 2003].

As abordagens globais de extração de características geralmente dividem automaticamente a imagem em fragmentos de textos para a extração de suas características [SAID & BAKER, 1998], [BULACU et al., 2003]. 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 e letras).

Neste trabalho adotou-se a abordagem global não contextual, ou seja, abordagem na qual o contexto da escrita não é relevante ao estudo. Para tanto, desenvolveu-se um algoritmo para a segmentação das imagens dos manuscritos digitalizados, simulando situações reais em que o perito depara-se com fragmentos de manuscritos. O processo consiste em dividir a imagem em 24 fragmentos regulares, ou seja, o algoritmo divide a imagem de entrada, particionando a amostra do manuscrito em 4 fragmentos horizontais e em 6 fragmentos verticais, conforme a Figura 5.6.

A escolha desta abordagem de segmentação justifica-se pelo fato de se disponibilizar um maior número de amostras dos autores na base de dados, permitindo uma maior variabilidade na escrita do autor, quando estes forem comparados.

Os segmentos extraídos do manuscrito serão armazenados, e a partir deles serão extraídas as características.

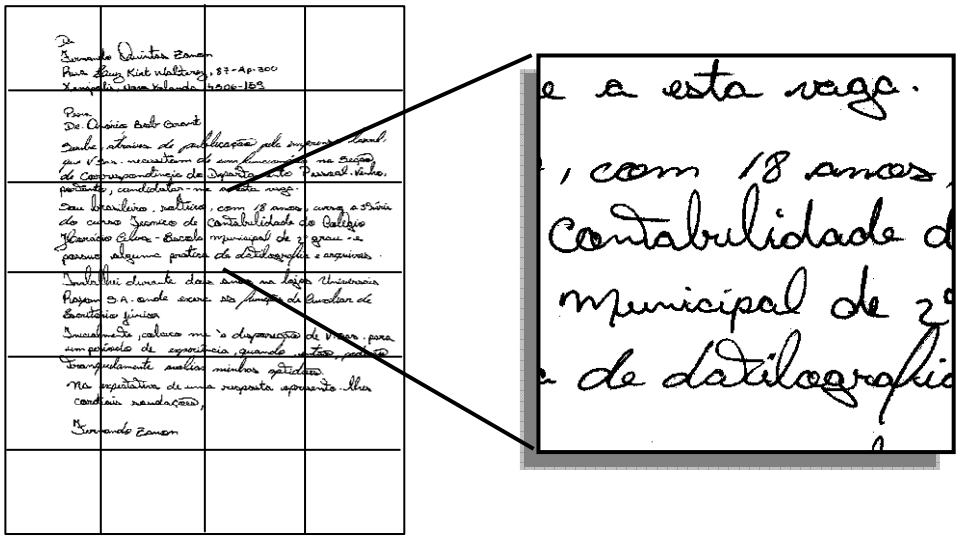


Figura 5.6 Exemplo da segmentação do manuscrito.

5.2.3. Extração de Características

Como visto nos Capítulos 2 e 4, a escolha de primitivas significativas constitui-se em uma das fases essenciais na elaboração do método de verificação. A representação computacional reflete diretamente nos resultados obtidos, nos quais a robustez do método é diretamente proporcional à qualidade das características, caracterizando-se, assim, em uma fase de suma importância.

Inclinação axial

A inclinação axial, como visto no Capítulo 2, é uma característica grafocinética que descreve o aspecto dinâmico do traçado e o ângulo de inclinação da escrita.

A inclinação axial foi implementada em função de alguns fatores associados a esta característica, descritos na seção 4.5.

A inclinação axial vem sendo amplamente utilizada tanto para o reconhecimento do estilo da escrita do autor [CRETTEZ, 1995] como para identificação e verificação de autoria [BULACU et al., 2003], [SRIHARI et al., 2002]. Como visto no Capítulo 3, o primeiro utiliza o cálculo de um diagrama direcional diretamente sobre o traçado, enquanto o segundo utiliza o cálculo do diagrama direcional sobre as bordas do traçado. Para a extração desta característica, na abordagem proposta, foi utilizada a técnica da distribuição de borda direcional. Nesta técnica, consideram-se as bordas, pois obtiveram melhores resultados na

discriminação da inclinação axial do autor em relação à inclinação extraída diretamente do traçado [BULACU et al., 2003]. Por serem mais finas, reduzem a influência da espessura do traçado sobre o cálculo.

Distribuição de borda direcional

A imagem pré-processada e segmentada é representada por uma imagem de borda na qual apenas os *pixels* desta borda estarão em preto. A imagem é então percorrida considerando-se o *pixel* da borda do traçado no centro do elemento estruturante retangular, conforme a Figura 5.7. 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 nas 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 de borda e computa-se a posição do fragmento em um vetor de posições para a construção do histograma.

O vetor de posições é normalizado pela distribuição de probabilidade $p(\theta)$ que dá a probabilidade de encontrar na imagem um fragmento de borda orientado em um ângulo θ em relação ao eixo horizontal, gerando um vetor de características de 17 posições.

O algoritmo implementado utiliza, sobre o segmento da imagem, elementos estruturantes com k = 3, 4 ou 5 *pixels* ao longo do fragmento de borda, no qual para cada elemento estruturante são quantificadas respectivamente em L = 9, 13 e 17 direções de inclinação que também representam a dimensionalidade do vetor final de características. A Figura 5.7 exemplifica o elemento estruturante com k = 5 e conseqüentemente L = 17.

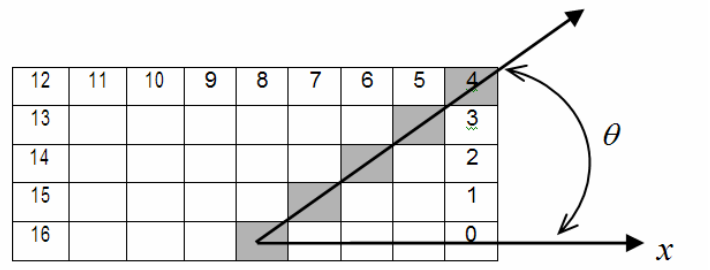
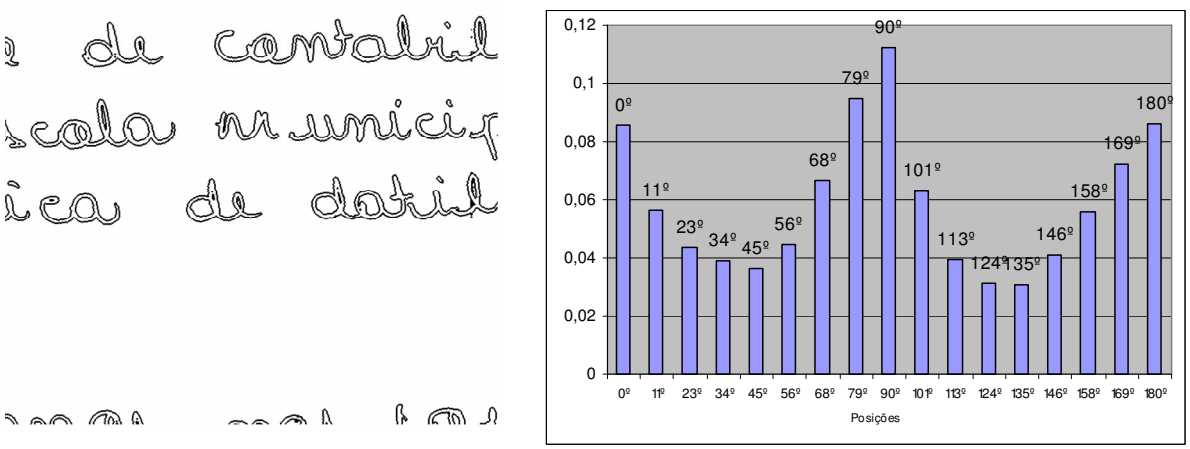


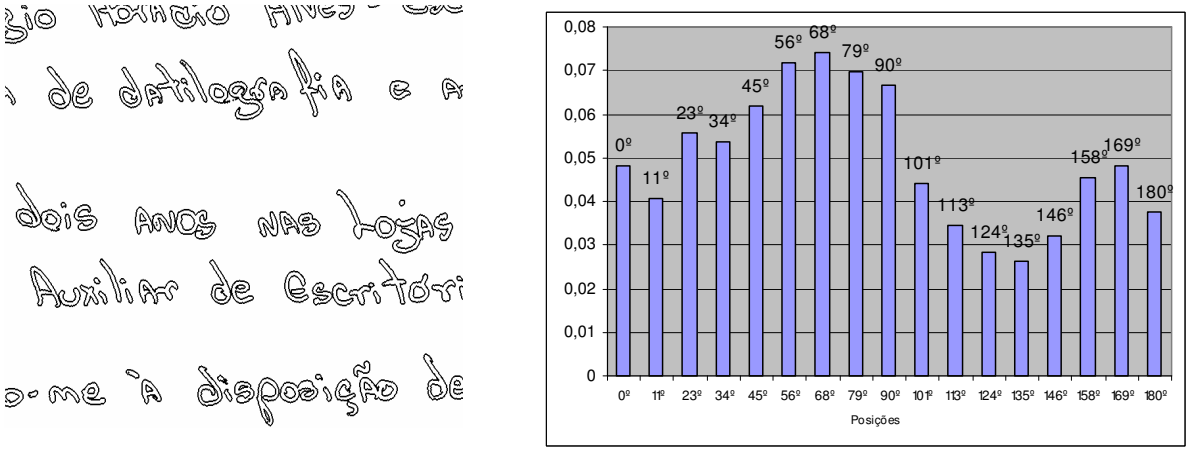
Figura 5.7 Exemplo do elemento estruturante.

Na abordagem proposta utiliza-se a distribuição de borda-direcional com elemento estruturante $k = 5$, gerando a quantidade de posições de $L = 17$, os quais apresentam resultados mais satisfatórios na detecção da inclinação do manuscrito em relação aos elementos estruturantes $k = 3$ e $k = 4$. Isto ocorre pois os elementos estruturantes menores e, conseqüentemente, com menos posições a serem computadas, são menos precisos em comparação ao experimento realizado o com elemento estruturante $k = 5$.

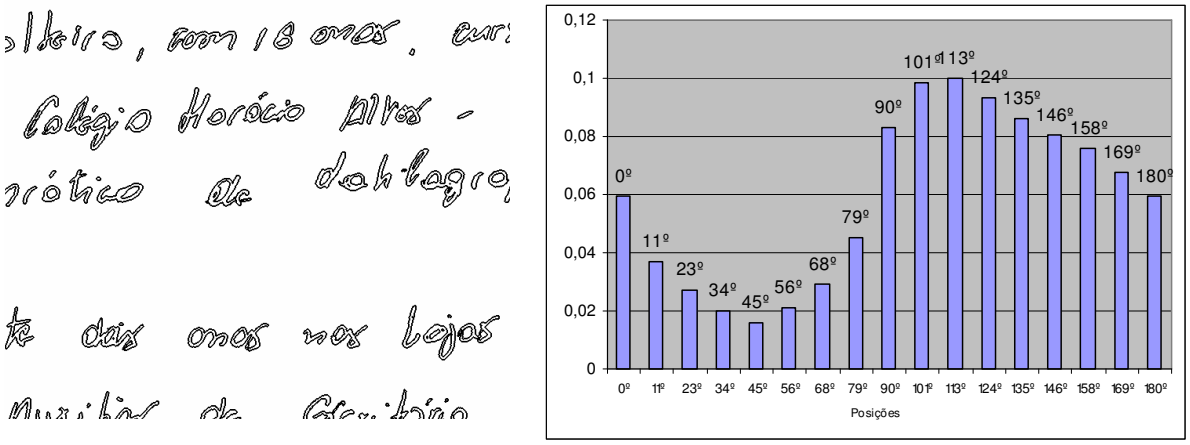
A detecção da inclinação axial é demonstrada na Figura 5.8, na qual pode ser observado o comportamento axial da escrita através dos ângulos de inclinação no histograma. As imagens representam respectivamente a inclinação axial nula (Figura 5.8(a)), inclinação axial à esquerda (Figura 5.8(b)) e inclinação axial à direita (Figura 5.8(c)).



(a)



(b)



(c)

Figura 5.8: Exemplo de inclinação axial do manuscrito: (a) inclinação axial nula, (b) inclinação axial à esquerda e, (c) inclinação axial à direita.

5.2.4. Cálculo das Distâncias entre as Características

O processo de classificação depende significativamente da métrica de distância, ou seja, é necessária a escolha de uma medida de distância adequada ao problema proposto [CHA, 2001].

Para a verificação de assinaturas, Santos [SANTOS, 2004] realizou testes com três tipos distintos de medidas de distância, bloco cidade, euclidiana quadrática e euclidiana, sendo esta última fixada para todos os seus experimentos por apresentar melhores resultados em relação às outras.

Com base nos testes de Santos, foram realizados experimentos com a distância Euclidiana. Os resultados se mostraram satisfatórios em relação aos manuscritos, sendo essa distância, portanto, também utilizada neste trabalho. Novas medidas de distâncias estão sendo propostas como trabalhos futuros.

No cálculo das distâncias Euclidianas, primeiramente, toda a base de dados foi convertida, dentro de um conjunto de vetores de características *fv*, as quais foram extraídas dos fragmentos dos manuscritos de referência M_{ki} *(i= 1,2,3...n)* e do manuscrito questionado M_Q (Equações (5.1 e 5.2),

$$fv_{Ki_{(i=1,2,\ldots,n)}} = (f_1, f_2, \ldots, f_L) \qquad (5.1)$$

$$fv_Q = (f_1, f_2, ..., f_L) \tag{5.2}$$

sendo *fv* os conjuntos de características e *L* o número máximo de células de cada característica. O vetor de distâncias $D_{i\ (i=1,2,3,...,n)}$ entre os fragmentos dos manuscritos de referência e o manuscrito questionado será computado para se obter a entrada do classificador, nesta abordagem o *SVM*, no treinamento, validação e verificação.

$$D_{i_{(i=1,2...n)}} = \sqrt{\left(fv_{Ki} - fv_Q\right)^2} \tag{5.3}$$

Dado um fragmento do manuscrito genuíno *x* e outro questionado *y*, aplica-se o extrator de características sobre as duas imagens, gerando um vetor de características, num total de 17 posições para cada um dos manuscritos, representando o número de células. O vetor da característica contendo a inclinação axial do manuscrito *x* e *y* fica representado respectivamente por $(INC_1^x .....INC_{17}^x)$ e $(INC_1^y .....INC_{17}^y)$. Extraídas as características, são computadas as distâncias entre os vetores de manuscritos, que serão usados como entrada na produção do modelo. A Figura 5.9 e 5.10 ilustram o processo descrito acima.

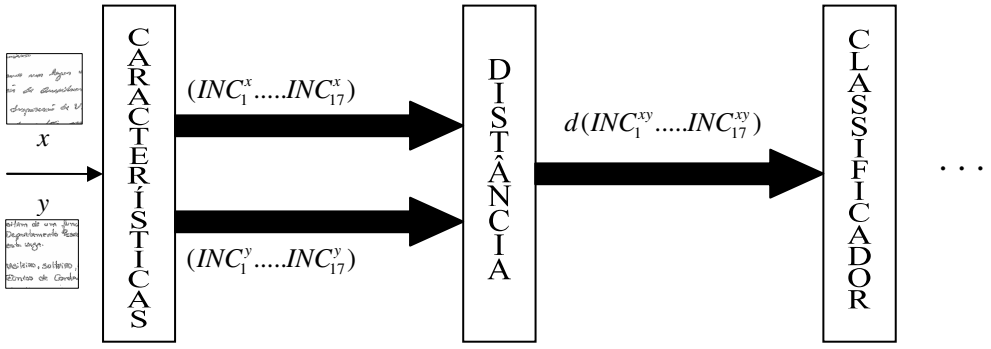


Figura 5.9: Resumo do processo de cálculo das medidas de distância.

5.2.5. Produção do Modelo

Há dois estágios na fase de produção de um modelo, o de treinamento e o de verificação. As distâncias das características extraídas são calculadas usando pares de

amostras de manuscritos, como será abordado no Capítulo 6. Na abordagem, para suprir este estágio, será usado o *SVM*.

Para a produção do modelo de treinamento, se o cálculo usar duas amostras de mesmo autor, será gerado um vetor de distâncias pertencente à classe w_1 (autoria ou associação), Figura 5.10.

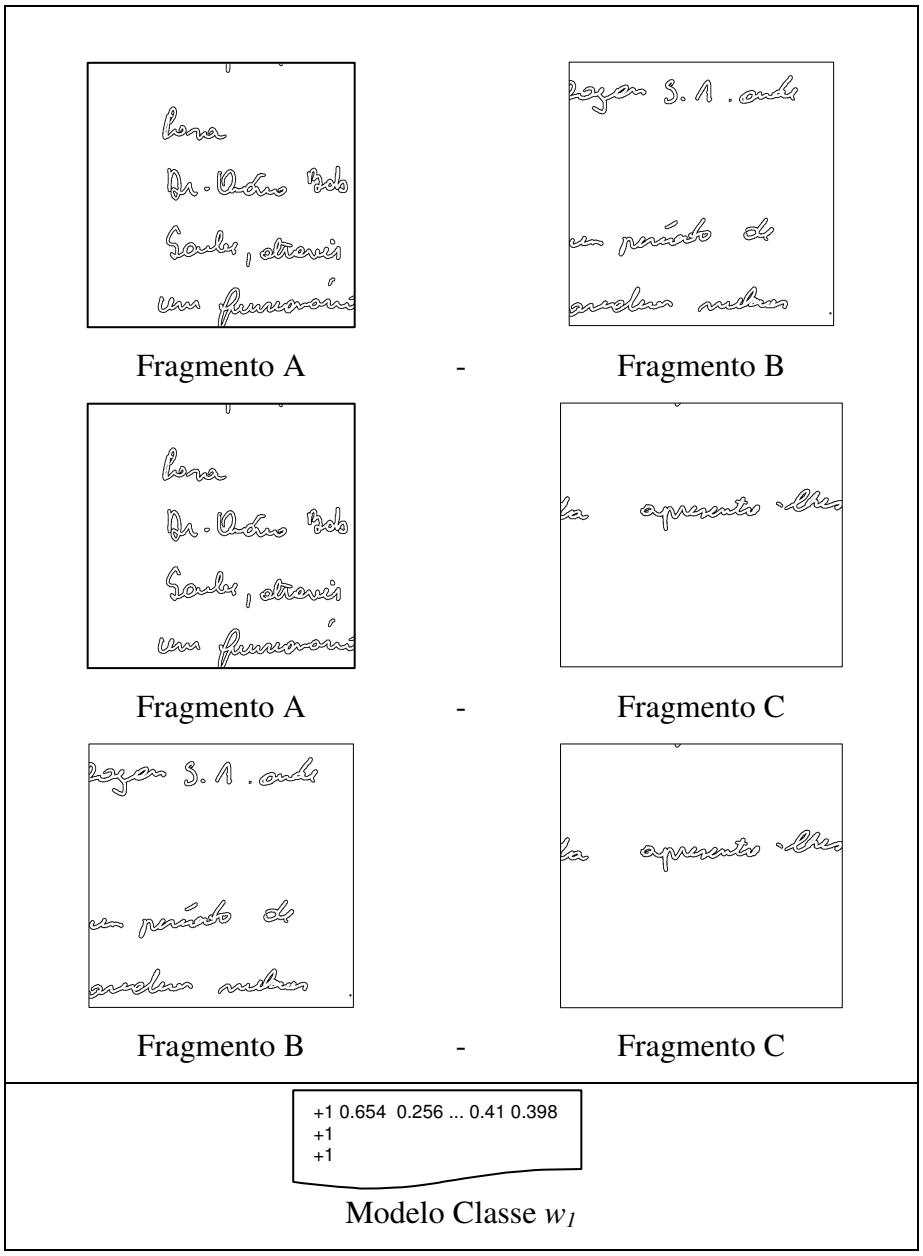


Figura 5.10. Ilustração do processo para a produção do modelo de treinamento usando fragmentos distintos do mesmo autor, dos quais serão computadas as Distâncias Euclidianas, gerando a (classe w_1).

Caso o cálculo seja efetuado entre duas amostras de diferentes autores o vetor de distâncias é pertencente à classe w_2 (não autoria ou dissociação), Figura 5.11. Portanto, para a entrada do *SVM* a classe w_1 é indicada como +1, e a classe w_2 como -1.

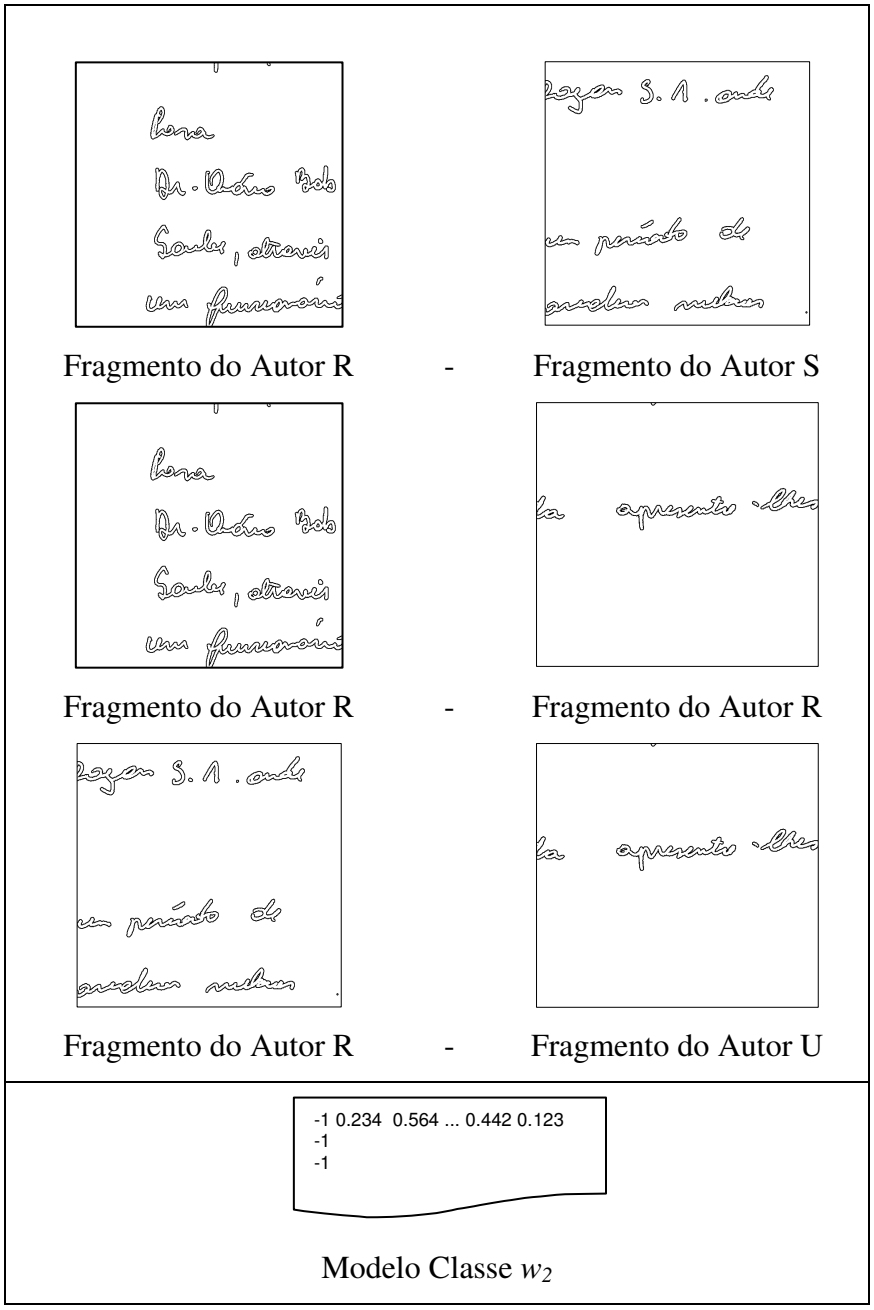


Figura 5.11. Ilustração do processo para a produção do modelo de treinamento usando fragmentos de autores distintos, dos quais serão computadas as Distâncias Euclidianas, gerando a (classe w_2).

Partindo da hipótese de que a medida de distância entre os vetores de características extraídos de amostras de 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.

5.2.6. Processo de Decisão

No processo de decisão foi avaliada a saída do modelo produzido verificando se o manuscrito deve ser considerado como pertencente ou não à determinado autor.

Conforme visto no Capítulo 3, em algumas abordagens, o processo de decisão em relação à autoria é gerado na saída do classificador [BULACU et al., 2003], [SRIHARI et al, 2002]. Porém, o método proposto é baseado na visão pericial, isto é, simula o processo da prova pericial na qual a saída do classificador é apenas uma entrada parcial para o módulo de determinação do voto majoritário (Seção 5.1).

No método proposto, cada amostra genuína conhecida (referência) será comparada com amostras de manuscritos questionados (teste), Figuras 5.12 e 5.13. Para este propósito, um conjunto de manuscritos do mesmo autor foi usado como referência ou modelo com o objetivo de produzir uma decisão final.

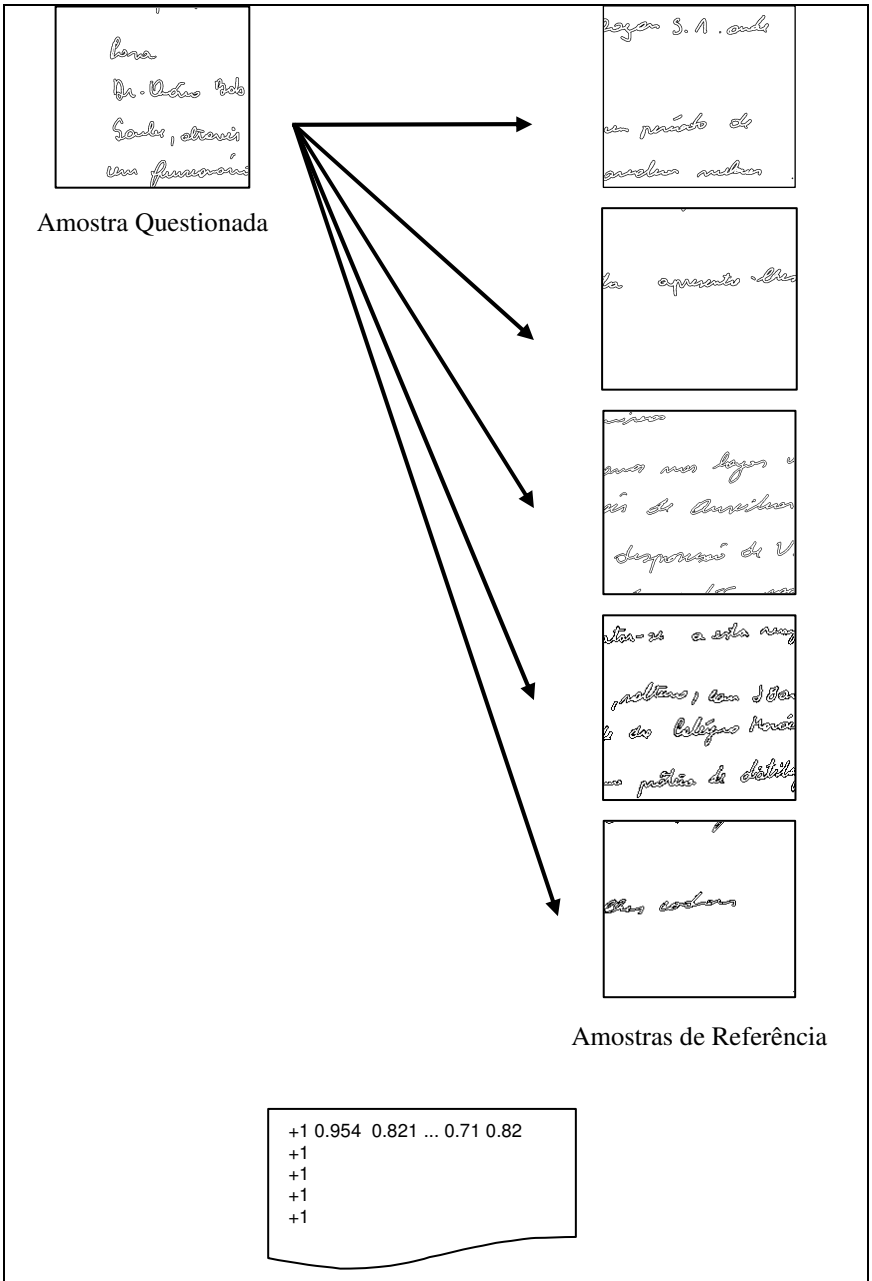


Figura 5.12. Processo de comparação entre o manuscrito questionado e os manuscritos conhecidos, fragmentos de manuscritos de mesmo autor (classe w_1).

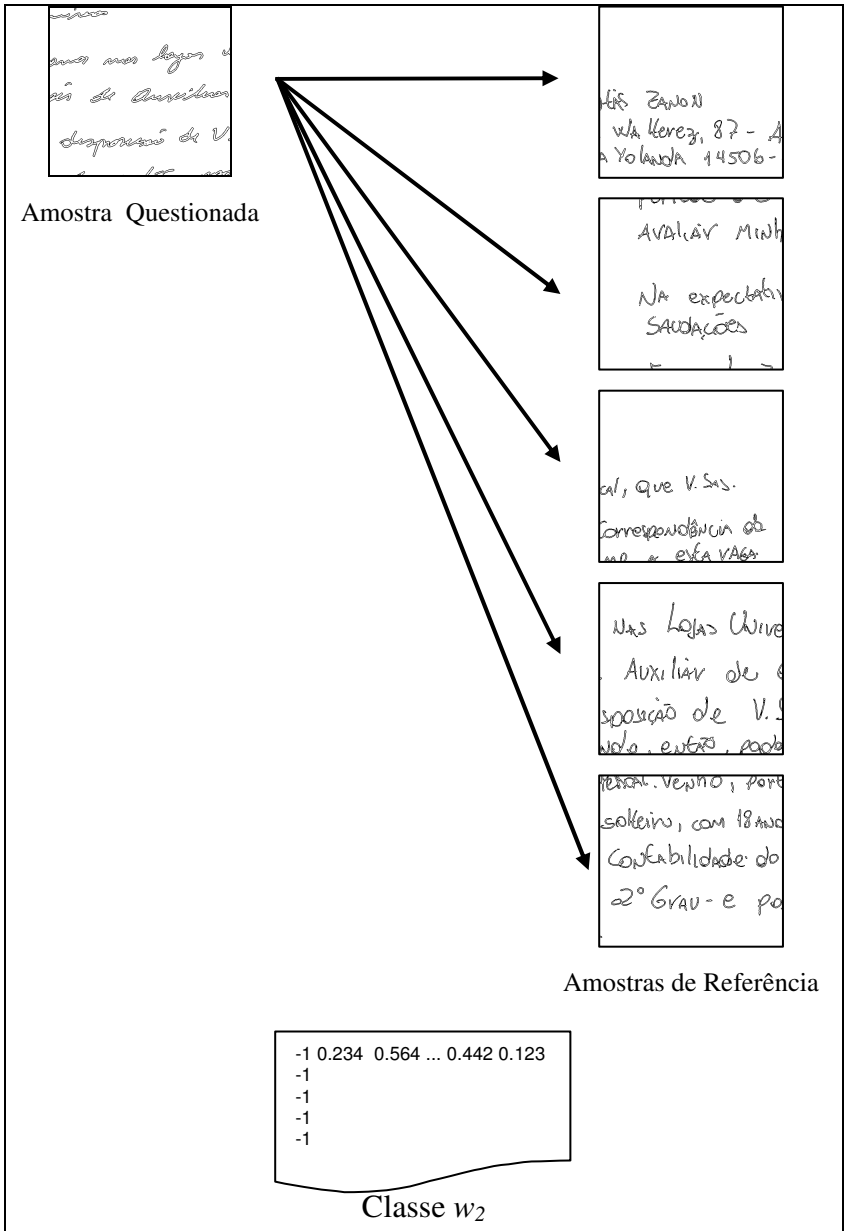


Figura 5.13. Processo de comparação entre o manuscrito questionado e os manuscritos conhecidos, fragmentos de manuscritos de autores diferentes (classe w_2).

O sistema proposto combina todas as saídas de classificação em uma regra de decisão baseada no voto majoritário [SANTOS, 2004], o qual é aplicado sobre as saídas do classificador. Desta forma a fórmula do voto majoritário é dada por:

$$Vm = \frac{Nref + 1}{2} \tag{5.4}$$

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

5.3.Comentários Finais

Neste capítulo foi apresentado o método proposto para a verificação da autoria de manuscrito. A visão do perito e as características Grafoscópicas foram fundamentais para a definição do mesmo. No capítulo seguinte serão mostrados e comentados os resultados apresentados pelo método.

Capítulo 6

Experimentos Realizados e Análise de Erros

Este capítulo contém o detalhamento do protocolo experimental, apresentando a divisão da base de dados e sua utilização. É abordado também o uso do *SVM* como ferramenta de classificação, assim como os resultados obtidos e análise dos erros.

6.1. Protocolo Experimental

Como visto no Capítulo 4, a base de dados é composta por 3 amostras de manuscritos por autor, respectivamente, AMTR(Amostra de Manuscrito de Treinamento), AMQ(Amostra de Manuscrito Questionado) e AMRF(Amostra de Manuscrito de Referência), as quais são usadas para fins diferentes nos experimentos realizados.

Nos experimentos iniciais, a base de dados contava com 145 autores, totalizando 435 amostras. Posteriormente, a mesma foi incrementada para um total de 315 autores, totalizando 945 amostras.

A base de dados de 145 autores foi segmentada em 24 fragmentos regulares (Seção 5.2.2), gerando um total de 10440 fragmentos de amostras. Já a base de 315 autores, depois de segmentada, totalizou 22680 amostras nos testes finais. Os fragmentos em branco ou com pouca informação foram descartados. Considerou-se como contendo pouca informação as imagens de fragmentos com quantidade de *pixels* pretos inferiores a 1% do total de *pixels* da imagem. Esse percentual foi estimado através da análise de um subconjunto da base de dados.

6.1.1. Divisão da Base de Dados

A divisão da base de dados consiste na separação de dois grupos de manuscritos distintos, o conjunto de dados de treino e conjunto dos dados de testes.

A base de treino é utilizada para a geração do modelo de aprendizado do classificador, enquanto que a base de teste é usada para a validação do método proposto.

É importante salientar que por se tratar de um modelo global de classificação, o conjunto de dados usado pela fase de treinamento e produção do modelo não participa do conjunto de dados usado nos testes. Ou seja, no conjunto de testes não há em hipótese alguma autores do conjunto de treinamento. Desta maneira, como visto no Capítulo 4, o classificador busca autenticar autores nunca vistos.

Outro fator a ser destacado é o balanceamento do modelo de treinamento, tanto na divisão das classes w_1 e w_2 quanto no número de espécimes por autor. Cada classe é representada por 50% do conjunto total de amostras.

Treinamento

Para a etapa de treinamento, por se tratar de uma abordagem que visa a redução de espécimes, usaram-se apenas três espécimes combinadas para a geração das classes w_1 (Figura 5.10) e w_2 (Figura 5.11). 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].

Para a geração da classe w_1, considerando a base segmentada, três espécimes de cada autor são aleatoriamente retirados do conjunto AMTR, sendo 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 w_2, para um dado autor, são escolhidos três fragmentos de manuscritos de autores

diferentes, selecionados aleatoriamente do conjunto AMTR. Desta forma, gera-se um arquivo do modelo que será a entrada no classificador *SVM*.

Teste

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 [SANTOS, 2004].

Desta maneira, a base de testes usa um conjunto com cinco espécimes de referência para cada autor e cinco espécimes de manuscritos questionados, retirados respectivamente dos conjuntos AMRF e AMQ, para gerar o modelo de teste. Para a geração da classe $w_{1,}$ são usados espécimes de mesmo autor aleatoriamente (Figura 5.12), enquanto que para a geração da classe w_2 são usados autores distintos (Figura 5.13). No processo de verificação é comparado 1 (um) dos espécimes do autor desconhecido contra 5 referências de um autor conhecido, selecionados aleatoriamente na base de dados. Um total de 50 combinações por autor é gerado, divididas entre as classes w_1 e w_2.

6.1.2. Protocolo de Testes

O protocolo de teste inicial é demonstrado na Tabela 6.1, em que são usados 75 autores para o treinamento e 70 autores para os testes.

Tabela 6.1: Protocolo do número de manuscritos utilizados nos experimentos iniciais.

Manuscritos				
Processos	**75 autores**	**70 autores**		**145 autores**
	AMTR	AMRF	AMQ	**Total**
	3 manuscritos por autor	5 manuscritos por autor	5 manuscritos por autor	
Treino	225			225
Voto Majoritário		350	350	700
Total				925

As combinações entre os valores das distâncias das características, utilizadas nos experimentos iniciais tanto para treinamento quanto para teste, são apresentadas na Tabela 6.2 a seguir:

Tabela 6.2: Protocolo de amostras utilizadas nos experimentos iniciais.

Manuscritos					
Processos	**75 autores**		**70 autores**		**145 autores**
	AMTR		AMRF e AMQ combinados		**Total**
	Classe w_1	Classe w_2	Classe w_1	Classe w_2	
Treino	225	225			550
Voto Majoritário			1750	1750	3500
Total					4050

Nos testes seguintes foram incluídos novos autores com o objetivo de comparar os resultados com os testes anteriores, como mostrado nas Tabelas 6.3 e 6.4.

Tabela 6.3: Protocolo de manuscritos utilizados nos experimentos com inclusão de 170 novos autores.

Manuscritos				
Processos	**75 autores**	**240 autores**		**315 autores**
	AMTR	AMRF	AMQ	**Total**
	3 manuscritos por autor	5 manuscritos por autor	5 manuscritos por autor	
Treino	225			225
Voto Majoritário		1200	1200	2400
Total				2625

Tabela 6.4: Protocolo de amostras utilizadas nos experimentos com inclusão de 170 novos autores.

Manuscritos					
Processos	**75 autores**		**240 autores**		**145 autores**
	AMTR		AMRF e AMQ combinados		**Total**
	Classe w_1	Classe w_2	Classe w_1	Classe w_2	
Treino	225	225			550
Voto Majoritário			6000	6000	12000
Total					12550

Nos experimentos finais ocorreram dois treinamentos com o objetivo de observar o comportamento do método proposto. Primeiro, com aumento no número de amostras de treino e depois, reduzindo o número de amostras de treino. No primeiro, a base de treinamento continha 200 autores e a base de testes 115 autores, Tabelas 6.5 e 6.6. No segundo, o treinamento foi elaborado com uma base composta por 50 autores para treino e 265 autores para os testes, Tabelas 6.7 e 6.8.

Tabela 6.5: Protocolo de manuscritos utilizados nos experimentos finais, treinamento com número elevado de manuscritos.

Manuscritos				
Processos	**200 autores**	**115 autores**		**315 autores**
	AMTR	AMRF	AMQ	**Total**
	3 manuscritos por autor	5 manuscritos por autor	5 manuscritos por autor	
Treino	600			600
Voto Majoritário		575	575	1150
Total				1750

Tabela 6.6: Protocolo de manuscritos utilizados nos experimentos finais, treinamento elevando o número de autores.

Manuscritos					
Processos	**200 autores**		**115 autores**		**145 autores**
	AMTR		AMRF e AMQ combinados		**Total**
	Classe w_1	Classe w_2	Classe w_1	Classe w_2	
Treino	600	600			1200
Voto Majoritário			2875	2875	5750
Total					6950

Tabela 6.7: Protocolo de manuscritos, amostras utilizadas nos experimentos finais, treinamento com diminuição de autores.

Manuscritos				
Processos	**50 autores**	**265 autores**		**315 autores**
	AMTR	AMRF	AMQ	**Total**
	3 manuscritos por autor	5 manuscritos por autor	5 manuscritos por autor	
Treino	150			150
Voto Majoritário		1325	1325	2650
Total				2800

Tabela 6.8: Protocolo de amostras utilizadas nos experimentos finais, treinamento com diminuição de autores.

Manuscritos					
Processos	**50 autores**		**265 autores**		**145 autores**
	AMTR		AMRF e AMQ combinados		**Total**
	Classe w_1	Classe w_2	Classe w_1	Classe w_2	
Treino	150	150			300
Voto Majoritário			6625	6625	13250
Total					13550

6.2. Experimentos

A seguir são apresentados os resultados obtidos pelos experimentos. Os testes foram divididos em grupos considerando os seguintes aspectos:

- Experimento inicial e aumento na quantidade de autores;
- Aumento na quantidade de autores no treinamento;
- Diminuição na quantidade de autores no treinamento;
- Método proposto vs. análise pericial.

O objetivo da realização dos testes subdivididos, conforme os itens acima, foi a melhoria na taxa de erro total. A taxa de erro total é dividida em dois tipos de erros: o de falsa rejeição, também chamado erro Tipo I, e o de falsa aceitação, erro Tipo II.

O erro de falsa rejeição (erro Tipo I) caracteriza-se quando o manuscrito de entrada é membro da classe (w_1) e é incorretamente classificado como não membro da classe.

O erro de falsa aceitação (erro Tipo II) caracteriza-se quando o manuscrito de entrada não é membro da classe (w_2) e é incorretamente classificado como membro da classe.

6.2.1. SVM

O pacote freeware SVM*light* foi utilizado para as etapas de treinamento e teste. Com o objetivo de estabelecer uma regularização empírica para o melhor valor da constante C, os valores testados variaram de 0 a 10000, porém os melhores resultados apresentados foram encontrados usando o valor default para C. O ajuste do parâmetro d do kernel polinomial é iniciado com 1 e os parâmetros –r e –s em 0,01. Maior detalhamento sobre configurações e parâmetros dos SVM podem ser encontrados em Joachims [JOACHIMS, 2000].Com relação ao kernel utilizado, experimentos realizados com o kernel linear geraram resultados superiores ao kernel polinomial como demonstrados na Tabela 6.9. Para a escolha da melhor função de kernel forma feitos experimentos com os protocolos da Tabelas 6.1, 6.2 e 6.3.

Tabela 6.9 – Experimento realizado para a determinação do melhor kernel

Autores	**kernel linear**			**kernel polinomial**		
	Falsa Rejeição Erro Tipo I (%)	**Falsa Aceitação (Erro Tipo II) (%)**	**Erro Total (%)**	**Falsa Rejeição Erro Tipo I (%)**	**Falsa Aceitação (Erro Tipo II) (%)**	**Erro Total (%)**
35	6,3	12,7	19,0	6,6	12,5	19,1
70	6,7	10,5	17,2	8,0	10,2	18,2
240	5,1	11,3	16,4	7,3	10,4	17,7

6.2.2. Experimento Inicial e Aumento na Quantidade de Autores

No primeiro teste, a fim de otimizar alguns parâmetros no processo de classificação, foi utilizado o protocolo da Tabela 6.2 para treinamento e teste. Inicialmente, são usados para testes 35 autores do total obtendo-se uma taxa de erro de 15,5% (Tabela 6.10), divididos em taxa de erro Tipo I e erro Tipo II, conforme a Tabela 6.10. Com o objetivo de avaliar o desempenho do método, com o aumento da base de teste, utilizou-se a base completa de 70 autores. Nesse caso observou-se uma queda no erro de falsa aceitação e falsa rejeição (erro total 13,4%), indicando que o modelo estava adaptado para o problema em questão.

Tabela 6.10: Primeiro experimento, resultados diferentes com aumento de autores.

Kernel Linear	**Voto Majoritário**		
Autores	**Falsa Rejeição Erro Tipo I (%)**	**Falsa Aceitação (Erro Tipo II) (%)**	**Erro Total (%)**
35	4,2	11,3	15,5
70	4,2	9,2	13,4

Como visto na seção 6.1, houve um aumento significativo na base de dados de manuscritos, sendo incorporados 170 novos autores. Assim, foi realizado um novo experimento usando o protocolo da Tabela 6.4, com o mesmo treinamento dos testes demonstrados na Tabela 6.10. Observou-se também nova queda das taxas de erro, desta vez, num percentual inferior ao anterior. Isto indica que o método proposto está próximo do limite

máximo de performance propiciado pelo conjunto de soluções do método proposto (arquitetura do classificador e características).

Tabela 6.11: Resultados com aumento da base de dados.

***Kernel* Linear**	**Voto Majoritário**		
Autores	**Falsa Rejeição Erro Tipo I (%)**	**Falsa Aceitação Erro Tipo II (%)**	**Erro Total (%)**
35	4,2	11,3	15,5
70	4,2	9,2	13,4
240	3	9,1	12,1

6.2.3. Aumento da Quantidade de Autores no Treinamento

Depois de constatado que, com o aumento da base de dados de autores para teste os resultados melhoraram, havendo uma diminuição na taxa de erro total, conforme visto na seção anterior (seção 6.3.1), aumentou-se a base de dados de treinamento com o objetivo de melhorar o erro médio, protocolo da Tabela 6.6. Obteve-se uma redução do erro do Tipo I (falsa rejeição), porém o erro do Tipo II (falsa aceitação) sofreu pouca alteração. Os resultados são demonstrados na Tabela 6.12. Esse fenômeno ocorre em decorrência da flexibilidade absorvida pelo modelo, absorvendo melhor as variabilidades intrapessoais.

Tabela 6.12: Resultados obtidos com aumento na base de treinamento 200 autores.

Kernel Linear	**Voto Majoritário**		
Autores	**Falsa Rejeição Erro Tipo I (%)**	**Falsa Aceitação Erro Tipo II (%)**	**Erro Total (%)**
115	1,7	10,8	12,5

6.2.4. Diminuição da Base de Treinamento

Finalmente, um último experimento foi realizado utilizando o protocolo da Tabela 6.8 com a diminuição da base de treino, usando somente 50 autores. Observaram-se melhorias nos resultados em termos de erro de falsa rejeição, 2,5 % e falsa aceitação, 7,9 %, gerando um

erro total de 10,4 %. Isso demonstra que com a combinação de um número elevado de amostras para teste, em relação ao conjunto de treino e um número relativamente baixo de amostras de treino, em relação ao conjunto teste, o método tende a gerar resultados melhores, podendo ser observados na Tabela 6.13.

Tabela 6.13: Resultados obtidos com 50 autores na base de treinamento.

Kernel Linear	**Voto Majoritário**		
Autores	**Falsa Rejeição Erro Tipo I (%)**	**Falsa Aceitação Erro Tipo II (%)**	**Erro Total (%)**
265	2,5	7,9	10,4

6.2.5. Método Proposto vs. Análise Pericial

Neste trabalho, apesar da abordagem ser totalmente baseada na visão pericial, não será realizado um comparativo, em termos de taxas de acerto e erro, pois no processo pericial é usado um conjunto de características complementares para a determinação da autoria (associação) ou não autoria (dissociação), enquanto que o método proposto utiliza apenas a inclinação axial.

A Figura 6.1 ilustra um exemplo de verificação complexa que geralmente ocorre nos casos em que documentos de autores distintos possuem muita similaridade na grafia. Assim, a análise da inclinação não seria suficiente para a discriminação entre autores, pois esta similaridade pode ser confundida com a variabilidade intrapessoal.

(a) (b)

Figura 6.1: Similaridade interpessoal (a) e (b).

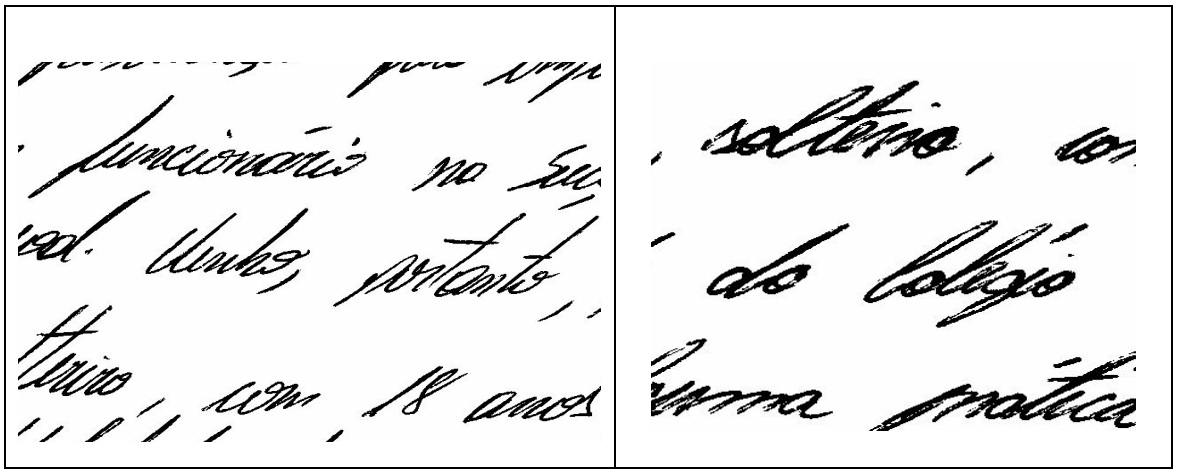
No método proposto a inclinação axial apresentou um bom desempenho, mapeando os ângulos de inclinação da escrita dos autores analisados em 17 possíveis posições, como visto na seção 5.2. Desta forma, distinguiram-se manuscritos que apresentam variabilidade intrapessoal e similaridade interpessoal, com taxas de erros aceitáveis, podendo ser

melhoradas com a inclusão de outras características. O método proposto demonstrou-se, portanto, eficiente como uma possível ferramenta de auxílio ao perito.

6.3. Comentários Finais

No primeiro experimento os resultados mostraram-se promissores, porém alguns fatores podem ser melhorados, tais como o acréscimo de outras características grafoscópicas.

A maior taxa de erro observado foi a de falsa aceitação. Um fator que influencia nesta taxa de erros é a similaridade interpessoal. Em uma análise entre as amostras usadas nos experimentos, constataram-se ocorrências de similaridades interpessoais levando a uma classificação incorreta do manuscrito, Figura 6.2.



(a) (b)

Figura 6.2: Similaridades entre amostras de autores distintos.

Outro fator que contribui para o aumento de taxas de erro consiste na quantidade de texto do fragmento, tanto em manuscritos de mesmo autor como em autores diferentes. Desta maneira, o manuscrito com quantidade de informação menor não representará satisfatoriamente a variação angular da inclinação axial, podendo gerar erros de falsa rejeição ou falsa aceitação. Este fenômeno foi observado nos experimentos, porém já relatados em experimentos realizados por [BULACU et al., 2003]. As quantidades de informações diferentes entre manuscritos podem ser observadas na Figura 6.3.

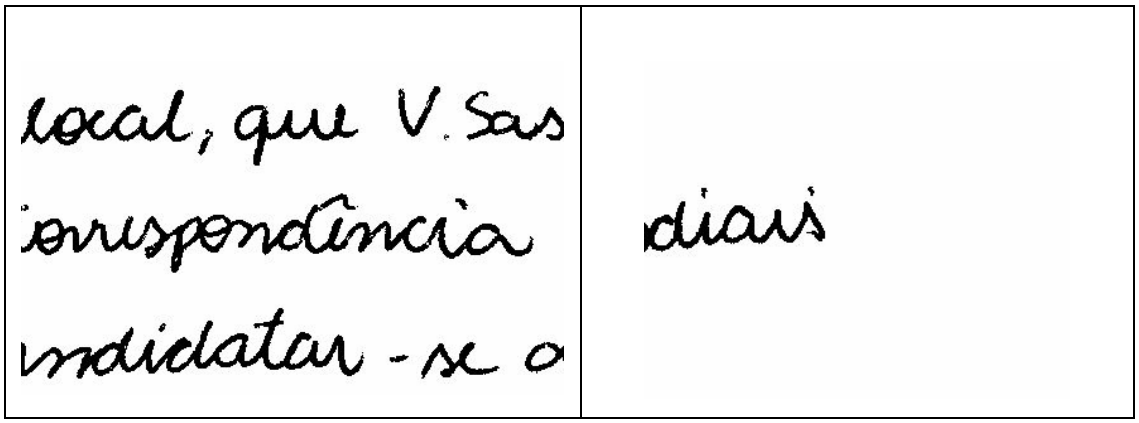


Figura 6.3: Quantidade de informações diferentes entre manuscritos.

A variabilidade intrapessoal é outro fator que provoca erros de classificação, porém com menor incidência se comparada com as similaridades interpessoais, Figura 6.4.

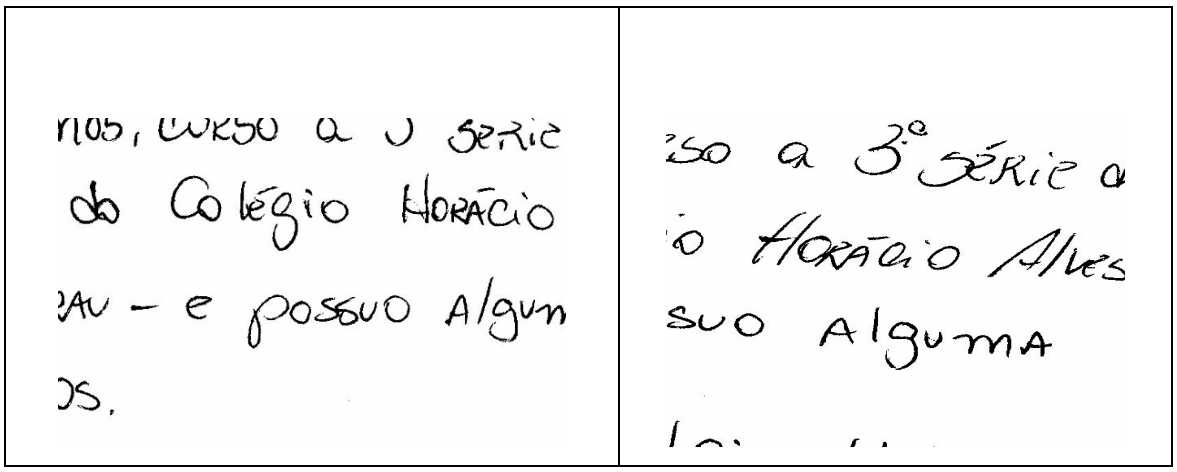


Figura 6.4: Variabilidades Intrapessoais.

O método mostrou-se robusto para a distinção entre manuscritos com forma caligráfica tipográfica comparados aos de formas caligráfica cursiva, devido as letras cursivas possuírem ligaduras entre as letras sendo menos comuns na forma tipográfica influenciando na variação angular no momento da extração da inclinação axial, Figura 6.5..

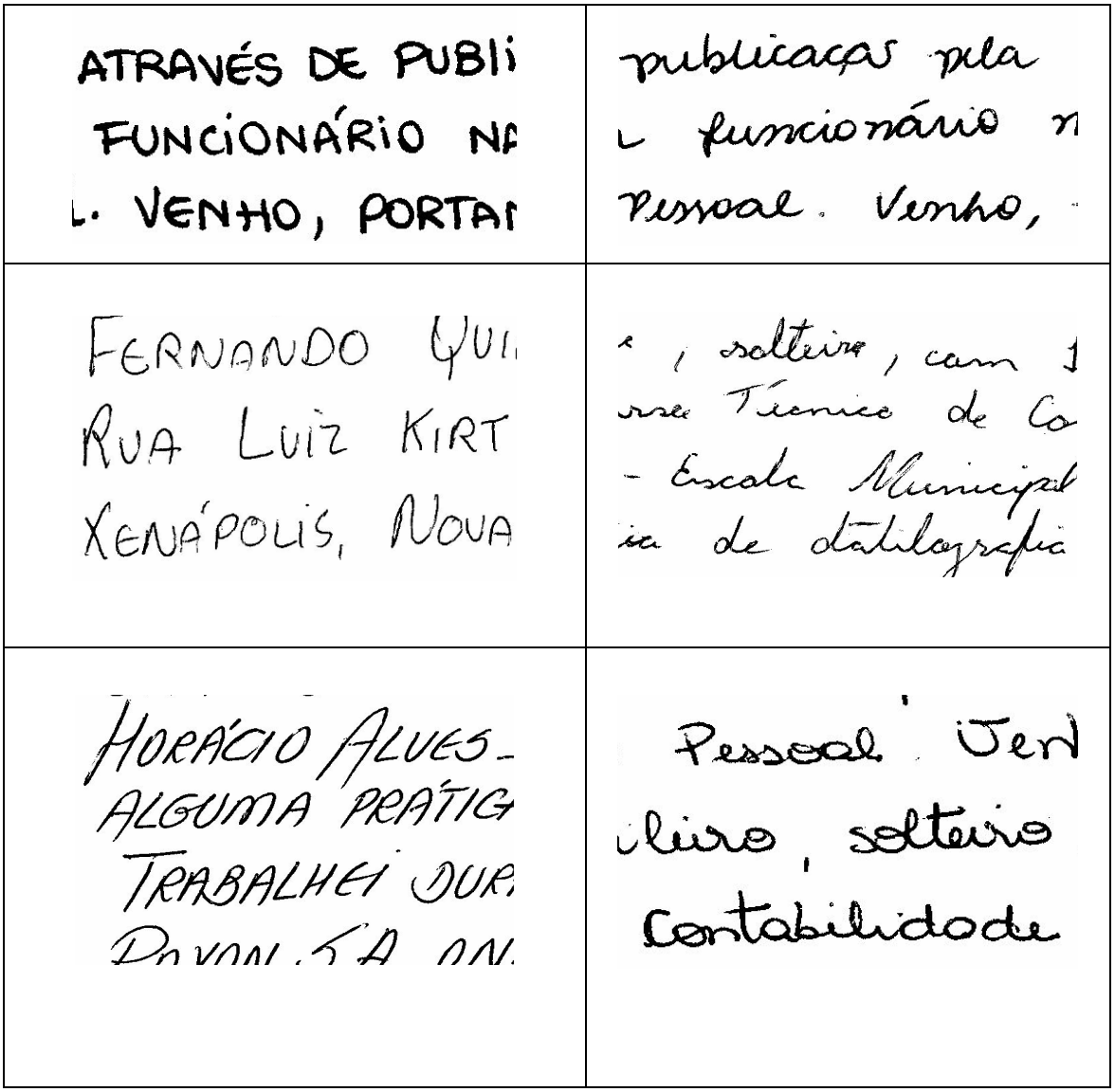


Figura 6.5: Distinção entre diferentes formas caligráficas.

Capítulo 7

Conclusão

O trabalho proposto apresenta uma abordagem para a verificação de manuscritos estáticos, baseados nos princípios da grafoscopia. Depois de detalhada a metodologia de extração de características, o uso da medida de distância Euclidiana aplicada sobre as primitivas para a verificação de similaridade, a produção do modelo e processo de decisão baseado na visão do perito grafotécnico, assim como os experimentos para validação estatística, estas são as conclusões observadas:

- O objetivo principal do método proposto, que era a verificação automática da autoria de documentos manuscritos, apresenta resultados promissores com taxas de falsa rejeição em torno de 2,5% e falsa aceitação de 7,9%, no melhor caso;
- O método mostrou-se robusto para o problema de escrita natural, ou seja, escrita com ausência de falsificações;
- Por se tratar de um modelo genérico, o mesmo gera um menor esforço computacional por não necessitar de novos treinamentos à medida que se incluem novos autores. Esta característica minimiza a complexidade que envolve as etapas de treinamento do modelo;
- A problemática envolvendo o número excessivo de manuscritos por autor no processo de treinamento, em abordagens pessoais, pode ser superada com a abordagem proposta. No trabalho proposto foram usados somente 3 (três) fragmentos de manuscritos por autor, para o treinamento, e 5 (cinco) fragmentos de manuscritos por autor no processo de decisão;
- A abordagem proposta simula a análise grafotécnica pericial, procurando simular um ambiente real;

- A característica grafocinética, inclinação axial, demonstrou uma boa capacidade discriminatória com uma taxa de acerto na ordem de 90%.

O principal propósito deste trabalho foi reportar um método robusto de verificação de autoria de documentos manuscritos embasado nos princípios da Grafoscopia. Dois pontos importantes devem ser ressaltados. O primeiro é o potencial apresentado pelo método em reduzir o número de amostras de manuscritos necessárias no treinamento. O segundo é a habilidade do modelo de absorver novos autores sem a necessidade de um novo treinamento. Em termos de taxa de erro, os resultados mostrados são promissores, principalmente em termos de identificação de mesmo autor. Foi possível perceber a grande capacidade do *SVM* em classificar novos autores, sem conhecimento a priori.

Como proposta para trabalhos futuros encontram-se:

- A inclusão de outras características grafoscópicas, permitindo ao modelo absorver mais adequadamente as variabilidades intrapessoais dos autores e propiciar tanto a redução da taxa de falsa aceitação como também um teste comparativo com a análise humana em termos de taxas de acerto e erro;
- A implementação de outros cálculos de medidas de distâncias além da medida de distância Euclidiana, permitindo a verificação do desempenho de outras medidas sobre as mesmas condições;
- Estender as aplicações de escrita natural para falsificações;
- Incrementar a base de dados e incluir falsificações.

Referências Bibliográficas

[ABUTALEB, 1989] ABUTALEB, A. S. Automatic Thresholding of Gray Level Pictures Using Two Dimensional Entropy. Computers Graphics & Image Processing, No. 47, 1989, 22-32 p.

[BARANOSKI et al., 2005] BARANOSKI, F. L.; JUSTINO, E. J. R.; BORTOLOZZI, F. *Identificação da Autoria em Documentos Manuscritos Usando SVM.* 5º ENIA, São Leopoldo, 2005, 544-552p.

[BARBEAU et al., 2002] BARBEAU, J.; VIGNES-LEBBE, R; STAMON, G. *A Signature based on Delaunay Graph and Co-occurrence Matriz.* In Proceedings of the ICCVG'2002, Pologne, 2003, 1-8p.

[BULACU, et al., 2003] BULACU, M.; SHOMAKER, L.; VUURPIJL L.. *Writer Identification Using Edge-Based Directional Features.* In Proceedings of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003), IEEE Press, 2003, 937-941p, vol. II, 3-6 August, Edinburgh, Scotland.

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

[CHA, 2001] CHA, SUNG HYUK. *Use of the Distance Measures in Handwriting Analysis.* Doctor Theses. State University of New York at Buffalo, EUA, 2001.

[CRETTEZ, 1995] CRETTEZ, Jean Pierre. *A set of handwriting families: style recognition.* Int. Conf. on Document Analysis and Recognition (ICDAR 1995), IEEE Press, , 1995 489-494p.

[DINES, 1998] DINES, J. E., *Document Examiner Textbook*, Pantex Intl Ltd, p. 566.

[DUDA & HART, 1973] DUDA, R. O. and HART, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley-Interscience, USA, 1st edition

[FACON, 1996] FACON, Jacques. *Morfologia Matemática: Teoria e exemplos*, Editora Universitária Champagnat da PUCPR, Curitiba, PR Brasil.

[FANG et al., 2003] FANG, B.; LEUNG, Y.Y.; TANG, K. W.; TSE, P. C. K; KWORK, Y. K.; WONG, Y. K. *Off-line Signature Verification by the tracking of features and stroke positions*. Pattern Recognition, Vol. 36, 2003, 91-101p.

[JOACHIMS, 2002] JOACHIMS T., *Optimizing Search Engines Using Clickthrough Data*, ACM Conference on Knowledge Discovery and Mining (KDD), 1-10p.

[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 Paraná, Brasil, 2001

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

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

[JUSTINO et al, 2003b] JUSTINO, E. J. R. SABOURIN, R, BOTOLOZZI, F. *An Off-Line Signature Verification Method Based on SVM Classifier and Graphometric Features. ICAPR, 2003*.

[KHOLMATOV, 2003] KHOLMATOV, A. *A Biometric Identity Verification Using On-Line & Off-Line Signature Verification.* Master's Dissertation. Sabanci University, 2003.

[LEEDHAM, 1994] LEEDHAM, C. G., *Historical Perspectives of Handwriting Recognition Systems.* University of Essex, London 1994.

[LEEDHAM & CHACHRA, 2003]LEEDHAM, G.; CHACHRA, S.. *Writer Identification using Innovative Binarised Features of Handwritten Numerals.* Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003. 413-416p. vol.1.

[LIMA, 2002] LIMA, A. R. G. *Máquinas de Vetores Suporte na Classificação de Impressões Digitais.* Dissertação de Mestrado. Universidade Federal do Ceará, Brasil, 2002

[MUKKAMALLA et al., 2002] MUKKAMALA, S.; JANOSKI, G; SUNG, A.H. *Intrusion Detection Using Neural Networks and Support Vectors Machines.* Proceedings of IEEE International Joint Coference on Neural Networks, IEEE Computer Society Press, EUA, 2002, 1702-1707p.

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

[OSUNA et al., 1997] OSUNA, E.; FREUD, R.; GIROSI, F. *Support Vector Machines: Training and Applications.* MIT Artificial Intelligence Memo 1602; MIT A. I. Lab,1997.

[MULLER et al., 2001] MÜLLER, K., MIKA, S.,RÄTSCH, G., TSUDA, K. and SCHÖLKOPF, B. *An Introduction to Kernel-Based Learning Algorithms*, IEEE Transactions on Neural Networks, Vol. 12, No. 2, March, 2001, 181-202p.

[OLIVEIRA & SABOURIN, 2004] OLIVEIRA, L. S.; SABOURIN, R.. *Support Vector Machines for Handwritten Numerical String Recognition.* In: IWFHR – 9, IEEE Computer Society, Tokyo, 2004, 39- 44p. .

[OTSU, 1979] OTSU, N. *A Threshold Selection Method from Gray-Level Histograms.* IEEE Transactions on Systems, Man and Cybernetics, v. SMC 9, 1979, No.1 62-66 p.

[RASHA, 1994] RASHA, ABAS. *A Prototype System for Off-Line Signature Verification Using Multilayered Feedforward Neural Networks.* Msc. Dissertation, Department of Computer Science RMIT, 1994, 6-42p.

[SAID et al, 1998] SAID, H. E. S.; PEAKE G. S.; BAKER, K. D.. *Writer Identification from Non-uniformly Skewed Handwriting Images.* British Machine Vision Conference, 1999.

[SANTOS, 2004] SANTOS, César R. *Análise Automática de Assinaturas Manuscritas Baseada nos Princípios da Grafoscopia.* Dissertação de Mestrado, Pontifícia Universidade Católica do Paraná, Brasil, 2004.

[SANTOS et al, 2004] SANTOS, C. R., JUSTINO, E. J. R., BORTOLOZZI, F. SABOURIN, R. *An Off-Line Signature Verification Method based on the Questioned Document Expert's Approach and a Neural Network Classifier*, In: The Ninth International Workshop on Frontiers in Handwriting Recognition, Tokyo, 10-14p. 2004

[SCHLAPBACH & BUNKE, 2004] SCHLAPBACH, A.; BUNKE, H. *Using HMM Based Recognizers for Writer Identification and Verification*. In: IWFHR – 9, IEEE Computer Society, Tokyo, 2004.

[SRIHARI et al, 2002] SRIHARI, S. N.; CHA, S. H.; HINA A.; SANGJIK, L.. *Individuality of Handwriting,* Journal of Forensic Sciences, 2002.

[SRIHARI & CHA, 2001] SRIHARI, S. N., CHA, S. H., AORA , H, LEE, S.. *Individuality of Handwriting: A Validation Study*. Center of Excellence for Document Analysis and

Recognition. University at Buffalo, State University of New York. IEEE, 2001 106 – 109 p. 2001

[STERNBERG, 1986] STERNBERG S.R., Grayscale morphology, Computer Vision Graphics Image Process. 35 (1986) 333-355 p..

[VAPNIK, 1995] VAPNIK, V . *The nature of statistical learning theory.* Springer Verlag, 1995.

[XIAO & LEEDHAM, 1999] XIAO, X; LEEDHAM, G.. *Signature Verification by Neural Networks with Selective Attention and a Small Training Set. Applied Intelligence.* Vol. 11, No. 2, 1999, 213-223 p.

[ZHU et al., 1999] ZHU Y.; TAN, T.; WANG, Y. *Biometric Personal Identification Based on Handwriting*. Chinese Academy of Sciences, China, 1999.

[ZOIS & ANASTOSSOPOULOS, 2000] ZOIS, E.N.; ANASTASSOPOULOS, V. M*orphological Waveform Coding for Writer identification,* Journal of the Pattern Recognition Society Vol. 33, 2000, 385-398 p.