ALINE MARIA MALACHINI MIOTTO AMARAL

IDENTIFICAÇÃO DE AUTORIA DE DOCUMENTOS MANUSCRITOS UTILIZANDO CARACTERÍSTICAS GRAFOMÉTRICAS

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

CURITIBA

2014

ALINE MARIA MALACHINI MIOTTO AMARAL

IDENTIFICAÇÃO DE AUTORIA DE DOCUMENTOS MANUSCRITOS UTILIZANDO CARACTERÍSTICAS GRAFOMÉTRICAS

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

Área de Concentração: *Ciência da Computação*

Orientadora: Profa. Dra. Cinthia Obladen de Almendra Freitas

Coorientador: Prof. Dr. Flávio Bortolozzi

CURITIBA

2014

Dados da Catalogação na Publicação

Pontifícia Universidade Católica do Paraná

Sistema Integrado de Bibliotecas – SIBI/PUCPR

Biblioteca Central

A485i
2014

Amaral, Aline Maria Malachini Miotto
Identificação de autoria de documentos manuscritos utilizando características grafométricas / Aline Maria Malachini Miotto Amaral ; orientadora, Cinthia Obladen de Almendra Freitas ; coorientador, Flávio Bortolozzi. – 2014.
133 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2014
Bibliografia: f. 127-132

1. Autoria - Identificação. 2. Escrita - Identificação. 3. Linguística forense. 4. Informática. I. Freitas, Cinthia Obladen de Almendra. II. Bortolozzi, Flávio. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título.

CDD 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

PUCPR
GRUPO MARISTA

ATA DE DEFESA DE TESE DE DOUTORADO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ÁREA DE CONCENTRAÇÃO: CIÊNCIA DA COMPUTAÇÃO

DEFESA DE TESE DE DOUTORADO Nº 023/2014

Aos 11 dias de Março de 2014 realizou-se a sessão pública de Defesa da Tese de Doutorado intitulada **"Identificação de Autoria de Documentos Manuscritos Utilizando Características Grafométricas"** apresentada pela aluna **Aline Maria Malachini Miotto Amaral** como requisito parcial para a obtenção do título de Doutor em Informática, perante uma Banca Examinadora composta pelos seguintes membros:

Profª. Drª. Cinthia O. de A. Freitas PUCPR (Orientadora)	Cinthia	APROVADA
	(assinatura)	*(aprov/reprov.)*
Prof. Dr. Flávio Bortolozzi CESUMAR (co-orientador)		APROVADO
Prof. Dr. Alceu Souza de Britto Jr PUCPR		APROVADO
Prof. Dr. Jacques Facon PUCPR		Aprovado
Prof. Dr. Sérgio Scheer UFPR		Aprovado
Prof. Dr. Celso Antonio Kaestner UTFPR		APROVADO

Conforme as normas regimentais do PPGIa e da PUCPR, o trabalho apresentado foi considerado APROVADO (*aprovado/reprovado*), segundo avaliação da maioria dos membros desta Banca Examinadora. Este resultado está condicionado ao cumprimento integral das solicitações da Banca Examinadora registradas no Livro de Defesas do programa.

Prof. Dr. Mauro Sergio Pereira Fonseca
Coordenador do Programa de Pós-Graduação em Informática

Dedico este trabalho aos dois grandes amores da minha vida, minha **filha Marcela** e **meu marido Marcelo**, sem os quais eu não teria nem força nem motivação para continuar.

“O dever de um perito é dizer a verdade; no entanto, para isso é necessário: primeiro saber encontrá-la e depois querer dizê-la. O primeiro é um problema científico, o segundo é um problema moral". Nerio Rojas

Agradecimentos

Primeiramente a Deus, porque sem ele nada seria possível.

A minha grande incentivadora e orientadora Profa. Dra Cinthia Obladen de Almendra Freitas, por todos os ensinamentos e pela orientação segura.

Ao prof. Dr. Flávio Bortolozzi pelas orientações e importantes contribuições dadas à realização deste trabalho.

Aos meus pais Laert e Lourdinha pelo amor e confiança a mim dedicados.

Ao meu querido marido Marcelo Augusto pelo amor, pela força e pela compreensão.

A minha pequena grande menina Marcela, por me mostrar o que realmente importa na vida.

Ao meu amigo Arthur, pela amizade sincera e constantes incentivos.

Ao UniCesumar pelo apoio recebido durante a realização deste trabalho.

Ao PPGIa pela oportunidade e suporte oferecidos ao desenvolvimento deste trabalho.

A CAPES pelo apoio financeiro.

A todos que direta ou indiretamente colaboraram na execução deste trabalho.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRAFICOS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SÍMBOLOS

A_{geral}	*O ângulo médio de todos os laços ascendentes e descendentes do documento em análise*
alturaDocumento	*Altura do documento em análise*
AlturaLaco	*Altura de um laço*
alturaPalavra	*Altura da primeira palavra de cada linha do documento em análise*
ALVOT	*Votacion Algorithm*
$AM_{fragmento}$	*Angulo médio dos laços de um fragmento*
AM_{laco}	*Angulo de cada laço de cada fragmento*
CEDAR	*Center of Excellence for Document Analysis and Recognition*
DPA_{geral}	*Desvio padrão médio do ângulo de todos os laços ascendentes e descendentes do documento em análise*
DPH_{geral}	*Desvio padrão médio da altura de todos os laços ascendentes e descendentes do documento em análise*
DPT_{geral}	*Desvio padrão médio do tamanho de todos os laços ascendentes e descendentes do documento em análise*
DPW_{geral}	*Desvio padrão médio da largura de todos os laços ascendentes e descendentes do documento em análise*
EMD	*Empirical Mode Decomposition*
FISH	*Forensic Information System for Handwriting*
GS	*Goodness Subset*
hmax	*Valor da altura máxima de uma palavra, empiricamente definido*
$Hm_{fragmento}$	*Altura média dos laços de um fragmento*
HM_{geral}	*Altura média de todos os laços ascendentes e descendentes do documento em análise*
Hm_{laco}	*Altura média dos laços de cada fragmento*
HMM	*Hidden Markov Models*
HMT	*Hidden Markov Tree*
IAM	*Informatik und Angewandte Mathematik*
KAS	*K-Adjacent Segments*
KNN	*K Nearest Neighborhood*
larguraDocumento	*Largura do documento em análise*
LarguraLaco	*Largura de um laço*

NLPR	*National Laboratory of Pattern Recognition*
nrolinhasCarta	*Identificador que contém o número total de linhas de um documento*
nroPixelsPretos	*Identificador que contém o número de pixels pretos de uma linha do documento*
NumeroPixelsLaco	*Número de pixels de um laço*
PDM	*Distribution Point Model*
posicaoMargemDireita	*Identificador que contém a distância da margem direita do documento em análise*
posicaoMargemEsquerda	*Menor distância da margem esquerda do documento em análise*
posicaoMargemInferior	*Posição do ultimo pixel preto da última palavra do documento em análise*
posicaoMargemSuperior	*Posição do primeiro pixel preto da primeira palavra do documento em análise*
RIMES	*Reconnaissance et Indexation de données Manuscrites et de fac similÉS / Recognition and Indexing of handwritten documents and faxes*
SVM	*Support Vector Machine*
$TM_{fragmento}$	*Tamanho médio dos laços de um fragmento*
TM_{geral}	*Tamanho médio de todos os laços ascendentes e descendentes do documento em análise*
$TM_{laço}$	*Tamanho médio dos laços de cada fragmento*
VSM	*Vector Space Model*
WANDA	*Forensic Information System Handwriting*
WED	*Weigthed Euclidean Distance*
$WM_{fragmento}$	*Largura média dos laços de um fragmento*
WM_{geral}	*Largura média de todos os laços ascendentes e descendentes do documento em análise*
$WM_{laço}$	*Largura média dos laços de cada fragmento*
X	*Largura da caixa (bounding box) na qual a 1ª palavra do documento é inserida*
XML	*Extensible Markup Language*
Y	*Altura da caixa (bounding box) na qual a 1ª palavra do documento é inserida*

Resumo

A escrita como elemento biométrico tem sido alvo de muitas pesquisas. Neste contexto, diferentes soluções computacionais para identificação de autoria em documentos manuscritos vêm sendo apresentadas na literatura, cada uma delas focando em aspectos específicos da escrita, bem como em aspetos referentes à imagem dos manuscritos. Esta pesquisa propõe um método computacional baseado em características grafométricas que visa auxiliar e agilizar o processo, realizado pelos peritos, de identificação de autoria em manuscritos. O estudo conta com uma base de cartas forenses (modelo PUCPR) as quais são pré-processadas para extrair-se um conjunto de características grafométricas para que subsequentemente o processo de identificação de autoria seja realizado. Foram realizados dois grupos de experimentos, o primeiro grupo teve como objetivo selecionar e validar as características grafométricas implementadas (análises individuais e em grupo foram realizadas). Com a melhor combinação de características selecionada, taxas de acerto de 84% para um grupo de 100 diferentes escritores foram obtidas. O segundo grupo de experimentos teve como foco principal identificar o número de escritores que atinge uma convergência assintótica dos resultados dos experimentos. Pode-se observar que com 200 diferentes escritores nenhum ganho ou perda pode ser observado nos resultados obtidos considerando as características grafométricas utilizadas.

Palavras-chave: identificação de autoria, características grafométricas, classificadores, cartas forenses.

Abstract

Handwriting as a biometric element has been the subject of researches. In this context, different computational solutions of writer identification have been described in the literature, each focusing on specifics aspects of handwriting as well as aspects relating to the manuscript image. The objective of this PhD project is defining a method that aims to assist and improve the process of writer identification held by the experts through a computational solution. This study uses forense letters (in the PUCPR forense letter database format) which are preprocessed and a set of graphometric features is extracted. Subsequently, the process of writer identification is performed. It were realized two groups of experiments, the first aimed at selecting and validating the implemented graphometry features (individual and combining analyses were realized). With the best ensemble of features, accuracy of 84% was obtained whit 100 different writers. The second group of experiments has with objective to obtain the number of writers which stabilizes the results of the experiments. It can be observed that gradually the relation between the number of writers and accuracy is stabilized, and with 200 writers the results are maintained.

Keywords: writer identification, graphometric features, classifier, forensic letter.

Capítulo 1

Introdução

De acordo com Mendes (2003), a "documentoscopia é a parte da criminalística que estuda os documentos para verificar se são autênticos e, em caso contrário, determinar a sua autoria". Observa-se nesta área um grau elevado de subjetividade, uma vez que diferentes peritos podem chegar a diferentes conclusões sobre os mesmos documentos. Dessa forma, o uso de ferramentas computacionais que automatizem e padronizem todo ou parte do processo de identificação adotado pelos peritos tem se tornado campo de interesse da computação.

Em relação a identificação de autoria, esta pode ser dividida em duas grandes áreas de pesquisa sendo elas: verificação e identificação de autoria. A verificação tem como objetivo principal avaliar dadas duas amostras de manuscrito, se as mesmas são ou não de um mesmo escritor (1:1). Enquanto a identificação tem como objetivo, dentre um conjunto de escritores candidatos, identificar o autor de um documento questionado (1:N).

No contexto da identificação de autoria várias pesquisas (LUNA et al., 2011, HELLI; MOGHADDAM, 2010; SIDDIQI; VICENT, 2008; HE et al., 2008; BENSEFIA et al., 2005; BLANKERS et al., 2007; BULACU et al., 2007; PERVOUCHINI; LEEDHAM, 2007; SCHOMAKER et al., 2007) foram propostas com o objetivo de apresentar métodos que automatizam todo ou parte do processo de extração, análise das características e classificação da escrita humana. Os principais aspectos que diferenciam tais pesquisas são: a natureza das características utilizadas, as bases de dados aplicadas, os métodos de classificação e por fim as taxas de acerto obtidas. Pode-se destacar dois principais grupos de características, aquelas que utilizam em sua definição os mesmos aspectos utilizados pelos peritos, chamadas de

características grafométricas (LUNA et al., 2011; CHEN et al., 2010; PERVOUCHINI; LEEDHAM, 2007; BLANKERS et al., 2007; SCHLAPBACH; BUNKE, 2004; HERTEL; BUNKE, 2003; ZOIS; ANASTASSOPOULOS, 2000), e aquelas que utilizam informações da imagem do documento, normalmente informações relativas à textura do documento ou geração de *codebooks*. Características texturais e *codebooks* (HELLI; MORGHADDAM, 2010; HE et al., 2008; SIDDIQI; VICENT, 2008; BULACU et al., 2007; SCHOMAKER et al., 2007) normalmente apresentam taxas de acerto melhores do que características grafométricas. No entanto, seu processo de extração não é natural a um perito, e as mesmas têm aceitação limitada nos tribunais de justiça.

Dessa forma, o desafio desta Tese de doutorado é propor um método, suportado por uma solução computacional, para o problema de identificação de autoria de documentos manuscritos que utilize apenas características grafométricas (que possam ser aceitas, entendidas e utilizadas pelos peritos e demais operadores da Justiça). Deve-se destacar que os resultados esperados a partir do método proposto tem como meta atingir taxas comparáveis com aquelas apresentadas em métodos que utilizam características não grafométricas, isto levando-se em consideração o número de escritores para treinamento e teste do método proposto.

1.1. Objetivos

Este trabalho tem como objetivo geral propor um método para identificação de autoria em cartas forenses utilizando características grafométricas.

Como objetivos específicos pode-se destacar:

- realizar um levantamento bibliográfico acerca dos seguintes temas: escrita humana e autoria de documentos manuscritos;
- estender a base de cartas forenses PUCPR para 600 diferentes escritores;
- propor um método de modo a selecionar um conjunto de primitivas, método de classificação e, ainda, técnicas de extração de primitivas em cartas forenses;
- implementar um cenário de teste como prova de conceito para avaliar e validar o método proposto, analisando os resultados obtidos.

1.2. Justificativa

Como já destacado anteriormente, diferentes abordagens para identificação de autoria em documentos manuscritos vêm sendo apresentadas na literatura cada uma delas focando em aspectos específicos da escrita, bem como em aspetos referentes à imagem dos manuscritos.

Levando-se em consideração a relação entre número de escritores/taxa de acerto (ver Tabela 2.3 – Capítulo 02) pode-se observar que normalmente os trabalhos que utilizam apenas características grafométricas apresentam taxas de identificação menores que os trabalhos que utilizam características não grafométricas.

Diante do exposto, entende-se que desenvolver um método confiável, que utilize apenas características grafométricas, e que obtenha resultados comparáveis com métodos que não utilizam características grafométricas apresenta-se como um campo de investigação em aberto. Acredita-se que com o desenvolvimento do método proposto torna-se possível diminuir o espaço de busca a um autor de um documento questionado por um perito. Isto porque, num espaço amostral muito grande, mesmo que o índice de acerto oferecido pelo método não seja de 100%, pode-se reduzir consideravelmente este espaço, uma vez que é possível eliminar um grande grupo de escritores que não são autores do documento em análise, deixando para a análise manual apenas um pequeno grupo de escritores.

Também é importante ressaltar que desde 1997 a Pontifícia Universidade Católica do Paraná - PUCPR, por meio do Programa de Pós-Graduação em Informática - PPGIa, vem desenvolvendo pesquisas em Análise e Reconhecimento de Documentos, e este trabalho de doutorado visa contribuir com a evolução destas pesquisas. Tais pesquisas vêm ao longo desses anos, contribuindo significativamente para a elevação da qualidade e produção científica gerada pelos Grupos de Pesquisa associados ao tema. Em 2001 foi criado o Laboratório de Computação Forense e Biometria vinculado ao Grupo de pesquisa de mesmo nome, pertencente ao PPGIa. Em 2005, com a evolução das pesquisas do Grupo na área de Ciências Forenses e com o intuito de integrar as duas áreas de conhecimento relacionadas com o tema (Direito e Informática), foi criado também o Laboratório de Direito e Tecnologia - LADITEC, junto ao Programa de Pós-Graduação em Direito – PPGD. Os grupos de pesquisa envolvidos são: Direito do Consumo e Sociedade Tecnológica (PPGD) e Computação Forense e Biometria (PPGIa).

Com a criação deste Laboratório foi possível integrar os Grupos de Pesquisa do PPGIa e do PPGD. Hoje o laboratório atende aos alunos oriundos dos dois Programas de Pós-Graduação (Informática e Direito, ambos da PUCPR). Essa integração vem propiciando a ampliação dos temas de pesquisa, assim como criando a sinergia necessária entre os professores dos diferentes Programas de Pós-Graduação.

Além dos Programas de Pós-Graduação, o Laboratório possui uma forte integração com os Cursos de Graduação da PUCPR, destacando os Cursos de Ciência da Computação, Engenharia da Computação e Direito. Essa integração abre espaços aos alunos de graduação para participar dos Projetos de Pesquisa desenvolvidos pelos grupos de pesquisa. Os alunos de graduação são integrados ao laboratório em duas modalidades: os Projetos de Iniciação Científica e Projetos de Conclusão de Curso. Deve-se destacar também a integração dos alunos de pós-graduação por meio de projetos de pesquisa que contemplam áreas de estudo e desenvolvimento da Informática e do Direito.

1.3. Ineditismo do Trabalho

Este trabalho apresenta um método confiável, resultante de uma combinação de características grafométricas, para o problema de identificação de autoria. Este método atingiu resultados comparáveis a métodos que utilizam características texturais e/ou de *codebooks*, que normalmente apresentam resultados superiores. A confiabilidade do método proposto concentra-se no fato que experimentos com diferentes números de escritores foram realizados de forma a obter o número máximo de escritores que são necessários para estabilizar as taxas de identificação obtidas.

Além disto, não foram identificadas outras pesquisas envolvendo uma base de cartas forenses na língua Portuguesa contendo 600 autores, a qual possibilitou a realização de testes como prova de conceito, variando o número de escritores de 20 a 200 autores.

1.4. Motivação

A principal motivação deste trabalho encontra-se na definição de um método suportado por uma solução computacional para o problema de **identificação de autoria em manuscritos** utilizando apenas características grafométricas.

Deve-se ressaltar que a identificação de autoria em manuscritos é um problema extremamente desafiador, uma vez que o desenvolvimento de soluções computacionais que além de produzirem resultados estatisticamente comprobatórios (ou seja, taxas de identificação compatíveis com as encontradas na literatura) ofereçam procedimentos e resultados que sejam aceitos pela comunidade jurídica, é um campo de investigação em aberto. Nesse contexto, o uso exclusivo de características grafométricas pelo método proposto possibilita tal aceitação.

As características grafométricas embutem em sua formação aspectos que são únicos de cada escritor, ou seja, aspectos que permitem a individualização da escrita humana. Um fator extremamente desafiador é entender a relação entre a variabilidade na escrita da escrita de uma pessoa (ou seja, as variações presentes na escrita de um indivíduo) e os elementos gráficos que permitem identificar tal pessoa como única.

Outro aspecto muito interessante investigado neste trabalho refere ao fato de que cada pessoa, mesmo que inconscientemente, deixa suas "marcas" ao escrever. O estudo destes aspectos foi objetivo desta pesquisa, uma vez que o perito, assim como o método proposto, utiliza tais "marcas" para realizar o processo de identificação de autoria.

1.5. Contribuições

Pode-se destacar como contribuição deste trabalho o desenvolvimento de um método confiável, composto por uma **combinação de características grafométricas**, que atingiu resultados comparáveis com métodos de identificação de autoria que utilizam características não grafométricas.

Uma vez que as características que compõem tal método são puramente grafométricas, foi necessário um **entendimento profundo dos elementos que afetam a escrita de um indivíduo,** e consequentemente dos aspectos que promovem a individualização da sua escrita. Tais aspectos permitem a identificação, por métodos manuais ou computacionais, da escrita humana.

Deve-se ressaltar também como contribuição relevante desta pesquisa o uso de uma **base de cartas forense com foco na Língua Portuguesa, a base PUCPR** (FREITAS et al., 2008). Esta base foi **estendida** neste trabalho de doutorado e atualmente contém 1800 exemplares de 600 diferentes escritores (03 exemplares por escritor).

Outra importante contribuição foi a definição do número de escritores que atinge **uma convergência assintótica dos resultados**, ou seja, as taxas de identificação obtidas com o uso do método proposto. Pode-se observar que com uma amostra composta por 200 escritores nenhum ganho ou perda são observados nos resultados. Em trabalhos como o descrito nesta tese, uma das variáveis que afetam diretamente os resultados das soluções computacionais desenvolvidas é o número de escritores usados para treinamento e teste.

Deve ressaltar também, que para garantir a consistência matemática e estatística dos resultados obtidos, todas as características implementadas passaram por um **processo de normalização**.

1.6. Hipótese de Pesquisa

A hipótese básica desta pesquisa, comprovada com os resultados apresentados pelo método proposto, é que é possível desenvolver um método confiável, composto por uma combinação de características grafométricas, que produza resultados comparáveis aos apresentados por métodos que utilizam características não grafométricas.

1.7. Escopo do Trabalho

O método proposto neste trabalho apresenta uma abordagem para identificação de autoria em documentos manuscritos escritos no Alfabeto Latino, não podendo ser aplicado a outros Alfabetos como, por exemplo, o Alfabeto Grego. Deve-se destacar que os caracteres (símbolos) presentes nestes alfabetos são muito diferentes, e o método proposto se baseou em características que são fortemente influenciadas pelo sistema de escrita do escritor, ou seja, levam em consideração os símbolos presentes em um Alfabeto.

Para a validação da solução computacional que suporta o método proposto foi utilizada a base de cartas forenses PUCPR. O formato de carta padrão desta base foi concebido para a Língua Portuguesa. Dessa forma, os resultados obtidos com a utilização do método proposto são restritos a esta base de cartas forenses. No entanto, todas as características foram normalizadas o que torna possível a utilização deste método com outras bases de cartas (forenses ou não) que sejam escritas no Alfabeto Latino. Isto porque, a única limitação para aplicação deste método é que cada escritor redija três exemplares do mesmo padrão de carta, independente da Língua.

1.8. Metodologia Científica

Este trabalho consiste de uma pesquisa exploratória de natureza aplicada, pois tem como objetivo estudar a escrita humana e gerar conhecimento para a solução de um problema específico desta área que é a identificação de autoria em manuscritos.

Como método científico adotou-se o método dedutivo, uma vez que com base em um conhecimento técnico e científico já formalmente conhecido pelos peritos é possível desenvolver e avaliar uma solução computacional que ofereça suporte de maneira consistente com tais conhecimentos (ou seja, com tais premissas).

Trata-se de uma pesquisa quantitativa, uma vez que a abordagem adotada para análise do método proposto ocorrerá por meio dos resultados mensuráveis obtidos com os experimentos realizados.

Com relação aos procedimentos técnicos, foram realizados levantamentos bibliográficos, que fundamentaram o desenvolvimento do método proposto. Atividades experimentais foram realizadas subsequentemente neste método para análise do mesmo.

Como passos adotados para o desenvolvimento deste trabalho, inicialmente foram realizados estudos com o objetivo de entender os aspectos que influenciam a escrita humana. Também foram realizados estudos sobre métodos que automatizam o processo de identificação de autoria com o objetivo de definir as metas a serem atingidas nesta pesquisa, principalmente com relação às taxas de identificação de autoria que poderiam e deveriam ser obtidas pelo método proposto.

Como segunda etapa desta pesquisa, o método proposto foi definido e as características grafométricas que o compõem foram selecionadas, estudadas e implementadas.

Também, a base de cartas forenses utilizada para validação do método foi estendida e o protocolo dos experimentos a ser adotado foi especificado, incluindo a escolha do método de classificação.

Sequencialmente, análises individuais e em grupo das características implementadas foram realizadas, por meio de experimentos, e uma abordagem formal de seleção de características foi aplicada com o objetivo de identificar a combinação para o melhor conjunto de características que levasse as melhores taxas de identificação.

1.9. Estrutura do Trabalho

Este trabalho está organizado em 05 (cinco) capítulos. O Capítulo 02 apresenta os princípios básicos da escrita humana, destacando as características que fazem da escrita um elemento biométrico. Neste capítulo também se discute algumas das principais pesquisas sobre abordagens automáticas para identificação de autoria. Tais pesquisas foram fundamentais para a realização de uma análise crítica do método proposto neste trabalho, bem como para avaliarmos os resultados obtidos. No Capítulo 03 é apresentado o processo adotado para o desenvolvimento deste trabalho, que incluí a apresentação de algumas bases de cartas forenses, bem como a base PUCPR utilizada neste trabalho. E, ainda, detalha-se o método proposto, o protocolo adotado para a realização dos experimentos de validação deste método, bem como, os resultados obtidos com tais experimentos. O Capítulo 04 aborda uma análise e discussão dos resultados destacando aspectos grafotécnicos da escrita humana. Finalmente, no Capítulo 05 são apresentadas as conclusões e trabalhos futuros decorrentes desta pesquisa.

Capítulo 2

Pressupostos Teóricos

2.1 Considerações Iniciais

Este capítulo apresenta uma revisão bibliográfica de dois tópicos fundamentais da pesquisa desenvolvida. O primeiro trata da escrita humana e o segundo da autoria em documentos manuscritos. Na primeira parte são abordadas diversas características da escrita humana e analisadas de diferentes ângulos, tais como: as áreas de estudo da escrita humana, a fisiologia da escrita, os sistemas de escrita, classes e características, leis e princípios da escrita e seus elementos gráficos. A segunda parte visa entender os aspectos intrínsecos da autoria em documentos manuscritos, tais como: abordagens para identificação de autoria em documentos manuscritos e sistemas para identificação de autoria em documentos manuscritos. Finalizando este capítulo apresenta o problema abordado nesta pesquisa bem como uma proposta de solução.

2.2. Escrita Humana

De acordo com Mendes (2003), grafólogos, psicólogos, pedagogos e outros especialistas definiram a escrita. Alguns destes conceitos são: a escrita existe para perpetuar o pensamento; a escrita é a arte de traduzir palavras ou ideias por sinais convencionais; a escrita é uma harmonia da qual o grafólogo decompõe os acordes para reconstituí-los sob outra forma; e a escrita é a representação dos sons, nas palavras, com absoluta exatidão, da palavra material, a parte do significado que contém.

No entanto, do ponto de vista da grafotécnica[1], nenhum destes conceitos é satisfatório. Assim, de acordo com Mendes (2003), "a escrita é um gesto gráfico psicossomático que contém um número mínimo de elementos que possibilitam sua individualização". Outra definição de escrita relevante ao presente trabalho é "a escrita ou grafismo é um conjunto de gramas[2] e de traçados, não subordinados obrigatoriamente às formas convencionais dos alfabetos" (CAVALCANTI; LIRA, 1996).

Dessa forma, nas próximas seções são apresentados alguns dos aspectos fundamentais que caracterizam a escrita como um elemento biométrico e que servem de base para a realização de análises grafoscópicas.

2.2.1. Áreas de Estudo da Escrita

A escrita humana vem sendo alvo de pesquisa em diferentes áreas, cada qual com foco e objetivos distintos. Neste contexto, abaixo são apresentadas suas principais áreas de estudo.

Grafologia

Segundo Cavalcanti e Lira (1996), durante muitos anos, a grafologia, ciência que estuda os traços do temperamento, caráter e inteligência por meio da escrita, estiveram restritas a apenas algumas áreas como a medicina. Mais atualmente, essa ciência passou a ser utilizada nas mais diversas áreas como, por exemplo, na área de recursos humanos de empresas. O primeiro livro editado sobre grafologia foi em Capri, no ano de 1622, por Camilo Baldi, médico de Bolonhe. Em seu livro, Baldi iniciou a pratica da análise da escrita, procurando conhecer o indivíduo com base nos elementos da mesma.

Ainda segundo Cavalcanti e Lira (1996), o primeiro trabalho amplo sobre grafologia foi escrito por Jean-Hyppolyte Michon na França o que lhe atribuiu o título de precursor da grafologia. Enquanto que se considera o pai da Grafologia moderna o psicólogo alemão Ludwing Klags. Seus trabalhos datam de 1910, quando a grafologia alemã havia se desenvolvido bastante, ao contato com a psicologia vigente.

---

[1] Parte da Documentoscopia que estuda as escritas com a finalidade de verificar se são autênticas e, em caso contrário, determinar a sua autoria (CAVALCANTI; LIRA, 1996).

[2] "Grama é a unidade gráfica. É o registro resultante da execução de um gesto gráfico realizado sem mudança brusca de sentido" (CAVALCANTI; LIRA, 1996).

A grafologia não só permite conhecer o indivíduo, mas, também, possibilita avaliar seu desenvolvimento psíquico. Esta ciência pode auxiliar médicos, psicólogos, juristas, peritos criminais, empresários, etc. Dentro deste contexto, a análise da letra é encarada sobre doze perspectivas da escola simbólica da grafologia, sendo elas: tamanho, inclinação, largura, zonas, simetria, conexão das letras nas palavras, formas de conexão, pressão, linhas, espaços, margens e assinaturas (OLIVEIRA et al., 2005).

Ressalta-se que o foco deste trabalho não recai sobre a grafologia no que diz respeito ao estudo da psique humana por meio da escrita, mas sim lança-se mão das 12 (doze) perspectivas como elementos relevantes à identificação de autoria em manuscritos.

Grafoscopia

De acordo com Cavalcanti e Lira (1996), "a grafoscopia é a parte da Documentoscopia que se encarrega da verificação da autenticidade e da autoria dos grafismos".

Segundo Oliveira et al., (2005), a grafoscopia envolve o estudo da origem de um documento gráfico, ou seja, a partir de que punho escritor foi gerada uma escrita.

De acordo com Mendes (2003), 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.

Como a escrita está sujeita a inúmeras mudanças, decorrentes de causas variadas, ela exige conveniente interpretação técnica para o completo êxito dos exames grafoscópicos periciais (MENDES, 2003). Para a correta análise do perito grafotécnico, tanto para a identificação quanto para a autenticação de autoria, existe a necessidade de se entender os princípios básicos, ou seja, os elementos básicos do processo de aprendizado da escrita e também as Leis do Grafismo.

Ressalta-se que o foco do presente trabalho é a grafoscopia por meio da extração de características que decorrem do trato tradicional dos peritos com os documentos manuscritos questionados.

Paleografia

Segundo Berwange e Leal (2008), a Paleografia é o estudo de textos manuscritos antigos e medievais, independentemente da língua veicular do documento. Nesse contexto, a paleografia estuda a origem, a forma e a evolução da escrita, independentemente do tipo de suporte físico onde foi registrada, do material utilizado para proceder o registo, do lugar onde foi utilizada, do povo que a utilizou e dos sinais gráficos que adotou para exprimir a linguagem.

A paleografia é de fundamental importância para o entendimento da história e cultura das civilizações antigas. Por meio da decifração obtida pelos paleógrafos, os historiados e arqueólogos conseguem obter dados importantes a partir de documentos escritos. Ressalta-se que tais aspectos não fazem parte do foco deste trabalho de pesquisa.

2.2.2. A Fisiologia da Escrita

Conhecendo-se as três áreas de estudo da escrita pode-se passar ao entendimento da fisiologia da escrita, pois é a partir do seu estudo que se pode compreender desde a formação dos traços até elementos mais complexos como letras e palavras.

Assim, deve-se considerar o exposto por Morris (2000), o qual explica que os estudantes jovens aprendem a construir letras individuais desenhando uma linha por vez, enquanto que os estudantes mais velhos aprendem a conectar letras individuais e manter a legibilidade. Isto não é uma tarefa fácil. Em parte, o sucesso deste exercício depende do sistema de alfabetização e dos estilos de escrita ensinados, sua atenção nos detalhes, além de muitas horas para dominar a tarefa de escrever. Com o passar do tempo a ação de escrita individual de letras e conexão de traços torna-se habitual e o escritor atinge um aumento de velocidade enquanto mantém a legibilidade.

Um escritor com habilidade possui maturidade gráfica que envolve movimentos combinados dos dedos, punhos e braços (SAUDEK, 1978). Um escritor não produz uma letra ou grupo de letras exatamente da mesma forma toda vez, ou seja, é normal se esperar variações nas escritas dos escritores (variação intrapessoal).

Considerando-se o exposto, apresenta-se a seguir os princípios fisiológicos da escrita humana, bem como, os fatores que influenciam a formação das letras. Além disto, descrevem-se os princípios do movimento de escrita, visando o entendimento da complexidade do ato de escrever. Trata-se também da escrita natural e automática, ou

seja, aquela escrita que o autor apresenta depois de muito treino, pois se deve lembrar que a escrita é um gesto aprendido, não se nasce sabendo escrever.

I. Os Sete Princípios Fisiológicos da Escrita Humana

A escrita é normalmente feita com os dedos e mãos conectados ao corpo pelo braço[3]. Como os dedos e mão, o punho e braços possuem muitos nervos e músculos que podem afetar o escritor antes, durante e após a ação da escrita propriamente dita. Dessa forma, torna-se importante examinar os sete princípios fisiológicos da escrita descritos no trabalho de Saudek (1978). A seguir são apresentados tais princípios com base na discussão descrita no trabalho de Morris (2000), sendo a tradução livre indireta e de responsabilidade desta autora.

Princípio 01

A escrita é uma forma de se expressar e para o escritor graficamente maduro os movimentos envolvidos são habituais. Na maioria dos escritores, seus dedos, mãos e braços movem-se em ações rítmicas e irrestritas para acompanhar as instruções do cérebro. Se o escritor é graficamente maduro, ele será capaz de se concentrar no conteúdo em que está escrevendo e não em como cada movimento da caneta deve ser feito. Seu cérebro conhece cada movimento necessário para escrever uma letra, combinações de letras, conexões de traçados, etc., e como o resultado de tal escrita deveria parecer. No final do processo de escrita, o escritor é capaz de comparar conscientemente ou inconscientemente a imagem mental com o que realmente foi escrito.

Princípio 02

Os músculos da mão funcionam melhor quando fazem contrações rítmicas e movimentos relaxados. Eles tornam-se fatigados quando algum movimento é dominante. Assim, a escrita normal e natural ocorre quando os músculos não estão fatigados, mas sim estão funcionando de uma maneira rítmica. Ações de contração musculares são mais desenvolvidas do que ações de relaxamento musculares. Isto faz com que a força da mão do escritor na caneta seja mais forte nos traçados descendentes

[3] Não se pode esquecer das pessoas que utilizam os pés ou a boca para escrever. Disponível em: <http://www.apbp.com.br/> Acesso 09 dez. 2013.

do que nos ascendentes quando o escritor está segurando a caneta na chamada posição normal de escrita (SAUDEK, 1978). À medida que o escritor aumenta a pressão da caneta e move a caneta em direção a ele (traçado descendente), ele gradualmente aplica mais força na ponta da caneta em direção ao papel, resultando na escrita de uma linha mais escura. Quando traçados ascendentes são escritos, a pressão da caneta gradualmente diminui, e a ponta da caneta não é forçada contra a superfície do papel com muita pressão, assim resultando na escrita de uma linha mais clara.

Existem exceções, mas estas não são resultado de nenhuma exceção a este conceito. Existem escritores que escrevem traçados ascendentes que são mais escuros que traçados descendentes, mas isto acontece em função do escritor segurar a caneta em uma posição diferente da chamada "posição normal de escrita".

Princípio 03

Normalmente, o relaxamento do músculo requer mais tempo do que sua contração quando o escritor está fatigado, porque o relaxamento torna-se mais difícil. Contrações de músculos fatigados não aparecem em curto prazo, porque os músculos são mais fortes. No entanto, com o passar do tempo, contrações também são afetadas.

Princípio 04

Se a habilidade de relaxar o músculo do dedo é perdida, e a habilidade de contraí-lo não é perdida, os movimentos do dedo tornam-se limitados e espasmódicos. Esta condição de movimentos espasmódicos pode fazer com que o dedo atinja eventualmente a caneta. Se esta condição ocorrer no dedo indicador, resultando em uma variação de pressão da força da mão isto pode causar ocorrências não usuais de linhas claras e escuras em um único traçado manuscrito.

Princípio 05

Durante a ação de escrita, fatiga espasmódica pode ser localizada como definida no Princípio 04 ou os dedos, mãos e braço podem estar envolvidos. Isto não ocorre durante a constante alternância rítmica das contrações e relaxamentos musculares, exceto após períodos muito longos de escrita. Este ritmo idealmente controlado afeta a pressão da força na caneta e simultaneamente a pressão da escrita. Quando fatiga espasmódica ocorre, um ou mais dos hábitos de pressão podem ser afetados.

Princípio 06

Quando os primeiros sinais de fatiga ou câimbras ocorrem, os escritores normalmente fazem algum tipo de ajuste. Eles podem alterar a forma com que seguram a caneta, mudar o seu ângulo, etc.. Qualquer mudança afetará a escrita.

Princípio 07

É necessário estudar a forma com que a tensão e o relaxamento muscular se sucedem. Quando a contração e o relaxamento são rítmicos, a escrita é provavelmente mais normal. Se eles não são, precisa-se verificar se a atenção é desviada, mesmo se apenas momentaneamente.

A partir destes princípios, que relacionam a fisiologia da escrita, pode-se entender que existem diferentes fatores que influenciam a formação das letras. Este estudo é importante, devido ao fato de que qualquer análise grafoscópica deve considerar a existência destes fatores.

II. Fatores que Influenciam a Formação das Letras

De acordo com Saudek (1978), um escritor aprende a escrever pelo método do impulso. Os diferentes impulsos de escrita descritos no trabalho de Saudek (1978) são o impulso do (a):

- traçado - o escritor aprende a escrever desenhando traços individuais que quando conectadas juntas formam uma letra;
- letra - o escritor escreve letras inteiras como um único ato de escrita;
- sílaba - o escritor escreve sílabas ou várias letras conectadas juntas;
- palavra ou do nome - o escritor escreve palavras completas ou nomes como um simples ato de escrita e
- frase ou sentença - quando o escritor atinge maturidade gráfica, ele “pensa” e tenta escrever sentenças/frases como um único ato.

De acordo com Morris (2000), após praticar a formação de letras utilizando uma série de traçados individuais, o escritor aumenta a velocidade de escrita. Ele aprende a escrever letras inteiras, e então várias letras juntas em combinação, ou seja, palavras e, finalmente, consegue construir sentenças sem pensar em como desenhar cada traçado ou cada letra.

No trabalho de Saudek (1978), são apresentados doze fatores que influenciam a formação das letras.

O primeiro fator está associado aos *instrumentos e materiais utilizados para a escrita*, que exceto em casos extremos, o instrumento e o material utilizados para a escrita (tais como: caneta, tinta, papel, superfície do papel no qual a escrita ocorreu, etc.) tornam a escrita não identificável. Um escritor usando material e instrumento de escrita satisfatórios, juntamente com uma superfície para a escrita lisa, deveria ser capaz de escrever de uma forma natural e normal. Em raras ocasiões quando uma combinação de um ou mais destes fatores não é adequada, o ato de escrita pode ser afetado.

O segundo fator é o *Grau de maturidade gráfica do escritor*, visto que o escritor amadurece, ele se afasta do sistema ou estilo de caderno de caligrafia de escrita e incorpora mais individualidades em sua escrita. Independentemente da individualidade de escrita do escritor, deve-se destacar que o princípio da legibilidade sempre governa a escrita, ou seja, o escritor quer ser capaz de se comunicar com a pessoa que lerá o que ele escreveu, dessa forma, o que ele escreveu deve ser legível. Existem diferentes níveis de maturidade gráfica, cada uma, regida por um número de diferentes fatores trabalhando juntos. O sistema de impulso é o conceito básico para o entendimento da maturidade gráfica. Um escritor imaturo, ou seja, uma criança aprendendo a escrever usa traçados separados de caneta/lápis para desenhar uma letra. À medida que seu nível de maturidade aumenta, ele passa a escrever com o impulso da letra. Neste nível, cada letra é uma unidade completa, porque não existe nenhuma dúvida sobre como esta letra deve parecer ou como ela deve ser escrita. O próximo nível são os impulsos das sílabas e palavras. Nesta fase o escritor sabe como escrever combinações de letras e sílabas como unidades únicas. Eventualmente ele aprende como escrever palavras completas e uma combinação limitada de palavras. Ele sabe como soletrar as palavras e a combinação de letras necessárias para escrevê-las. Ele escreve as palavras confortavelmente porque ele não precisa pensar em quais letras compõem uma palavra ou sílaba. O nível mais alto é o impulso da sentença ou frase. O escritor neste nível pensa e procura escrever sentenças/frases completas.

O terceiro fator é a *velocidade relativa de escrita do escritor*, pois não é possível escrever rapidamente na maioria dos sistemas. A maioria dos sistemas necessita de pressão uniforme de escrita. Quando um escritor aumenta sua velocidade de escrita, ele percebe que não pode escrever traços ascendentes e descendentes com uma pressão

uniforme. Um aumento na velocidade de escrita tipicamente resulta em hábitos de pressão relativos. Hábitos de pressão relativa são diferentes nos traços ascendentes e descendentes. Eles são resultantes da alteração normal rítmica entre contrações e relaxamentos musculares. Hábitos de pressão relativa não são parte da maioria dos sistemas de escrita humana. Além disso, letras no estilo de "caderno de caligrafia" são redigidas de tal forma que não podem ser feitas sem parar em algum ponto durante sua escrita, ou redesenhando algum traço que já foi feito. Sempre que um escritor para ou redesenha alguma linha, ele perde tempo. Ninguém escreve na mesma velocidade, assim a velocidade de escrita deve ser um elemento significante no processo de identificação do escritor.

O quarto fator é o *sistema de escrita aprendido* (descrito em mais detalhes na próxima Seção) e o quinto fator *é a nacionalidade do escritor*, uma vez que embora muitos países utilizem símbolos do alfabeto iguais, poucos têm projetos de letras idênticos (HILTON, 1982).

Já o sexto fator é o *grau de sensibilidade visual e impressionabilidade do escritor* que esta relacionada ao que um escritor conhece, ou seja, uma letra ou como uma composição de letras deve parecer. Sua visão confirma que a imagem no papel está de acordo com a que estava em sua mente. Se uma pessoa pôde ver durante seus anos de formação, e logo após atingir maturidade gráfica perder sua visão, durante um tempo esta pessoa ainda consegue escrever neste nível de maturidade. A medida que o escritor fica mais velho, gradualmente perde sua visão, e passa a não ser mais capaz de relacionar o que escreveu com sua imagem mental, sua escrita irá mudar. Outros fatores no processo de envelhecimento irão afetar a maturidade gráfica.

O sétimo fator é o *poder de expressão gráfica do escritor* e o oitavo são as *características pessoais do escritor,* tais como vaidade, afetividade, e desejo de imitar outros. A personalidade e características pessoais de um escritor podem influenciar como uma pessoa escreve. Por exemplo, autoconfiança, sinceridade, falta de cuidado, imaginação e gosto estético, podem ter um impacto na natureza artística do escritor. Alguns exemplos disto são assinaturas estilizadas, desenhos, embelezamentos, detalhes em letras individuais para garantir legibilidade. Um exemplo muito interessante disto é a assinatura de John Hancock (Figura 2.1) na Declaração de Independência Americana. Também, deve-se considerar características referentes a fatores psicopatológicos. Um escritor que sofre de problemas psicopatológicos severos pode produzir textos que

contém características e elementos que possuem significado apenas para ele, por exemplo: não usuais e bizarros formatos de texto e letras, pontuações não usuais, etc. Além disso, quando uma pessoa está temporariamente excitada ou sobre estresse emocional, sua escrita pode ser adversamente afetada. Normalmente quando uma pessoa está excitada, ela não tem o mesmo nível de controle sobre a caneta e a escrita aparece mais randômica, maior e possivelmente menos legível.

Figura 2.1. Assinatura de John Hancock

[Fonte: http://www.americaslibrary.gov/jb/colonial/jb_colonial_hancock_2.html]

Os últimos quatro fatores são: *o conhecimento do escritor em linguagens estrangeiras, treinamento especial, etc.; as condições fisiológicas do escritor; impedimentos físicos crônicos que o escritor possa ter e finalmente se a forma da letra permanece sozinha, ou no início, meio ou final da palavra.*

Percebe-se que todos estes fatores vão fazer com que um escritor produza traçados diferentes de outro escritor, por exemplo, para uma mesma letra ou palavra. Além disto, um escritor poderá escrever uma mesma letra ou palavra de modo diferente (forma, inclinação, tamanho, etc.) dependendo dos fatores apresentados anteriormente. Assim, percebe-se que a escrita é complexa e que qualquer trabalho, seja pericial ou de desenvolvimento de sistemas computacionais, enfrentará diversos desafios. Tais fatores atuam diretamente sobre o resultado, ou seja, o traço, a escrita.

Neste contexto, cabe apresentar os princípios do movimento de escrita para que se possa entender como tais fatores atuam no movimento realizado durante a escrita.

Princípios do Movimento de Escrita

De acordo com Morris (2000), para a escrita normal e natural, algumas regras de execução da escrita devem ser destacadas, pois tais regras são observadas de perto pelos peritos diante da análise de escritas questionadas. As regras são as seguintes:

- cada escritor adota um “relativamente constante” e distinguível grau de inclinação na sua escrita;
- quando o ângulo médio de inclinação é alterado até certos limites, o seguinte pode acontecer:
    - a simetria da estrutura da curva pode ser comprometida;
    - a uniformidade do tamanho da letra pode ser destruída;
- cada escritor tem uma velocidade média de escrita. Ele pode mudar arbitrariamente sua velocidade média, mas sua escrita irá apresentar traços desta mudança de velocidade;
- “lentidão” induzida artificialmente é diagnosticada por uma redução na firmeza e harmonia durante a execução do ato de escrita. Uma indicação desta mudança está no tamanho de exemplos recorrentes da mesma letra;
- para um determinado conjunto de condições de escrita, todo escritor tem um determinado tamanho médio das letras. Normalmente, ele não desvia deste tamanho médio sem um esforço deliberadamente consciente;
- qualquer tentativa deliberada de variação do tamanho médio de uma letra é acompanhada por inconsistência no tamanho dos exemplos sucessivos da mesma letra;
- quanto maior a velocidade de escrita, maior a dificuldade de formar ângulos agudos entre dois traçados sucessivos executados perto e em direções opostas;
- para cada escritor, a localização de pontos de pressão relativa é involuntária;
- nenhum escritor pode deliberadamente alterar a localização de hábitos de pressão relativa exceto em detrimento da velocidade de escrita, simetria e qualidade das linhas.

Assim, observa-se que qualquer modificação voluntária ou involuntária da escrita deixará marcas e, portanto, despertará o interesse dos peritos durante a análise grafoscópica. Cabe ao perito, identificar os elementos da grafia que permitem estabelecer que uma escrita é natural e automática. Neste contexto, um dos elementos da grafia que deve ser observado, pois apresenta um alto poder discriminatório é a inclinação axial. Esta característica foi incluída em nossa base de características grafométricas e apresentou um grande impacto no processo de identificação de autoria, como pode ser observado na Tabela 3.2.

Escrita Natural e Automática

Segundo Morris (2000), a escrita natural e automática é qualquer escrita executada normalmente sem uma tentativa de controlar ou alterar seus hábitos de identidade e sua qualidade usual de execução. Dessa forma, pode-se observar que um escritor que escreve natural e automaticamente possui maturidade gráfica.

Neste sentido, existem algumas condições que devem ser atingidas antes que um escritor possa escrever naturalmente e automaticamente, são elas:

- o escritor não tem nenhuma dúvida sobre a forma e os movimentos necessários para escrever uma letra;
- o escritor tem um controle completo da caneta e da superfície de escrita e não existe nenhum problema mecânico relacionado a tal controle;
- não existe nenhum fator transitório, ou permanente, afetando a habilidade de escrita do escritor;
- não existem dúvidas do escritor com relação à legibilidade da escrita, hábitos de pressão relativa usados para escrever as letras, hábitos de espaço relativo entre letras, palavras, sentenças, linhas, parágrafos, e tamanho do espaço das margens, etc.;
- o escritor encontra-se confortável com a linguagem e o sistema de escrita no qual ele escreve;
- o escritor não muda de uma linguagem ou sistema de escrita para outro dentro de um mesmo texto.

Resumindo, qualquer ato ou ocorrência de qualquer evento que faça com que o escritor preste mais atenção na sua forma de escrita, configura um situação na qual seu nível de maturidade gráfica encontra-se comprometido.

Pode parecer que este conjunto configura muitas exigências, pois o escritor não está atento a todos estes elementos a cada instante. Isto faz do escritor um agente natural e automático quando decide ou necessita escrever. Quando o escritor precisa prestar atenção a estes elementos, sua escrita deixará de ser natural e automática, portanto, cabe estudar os sistemas de escrita, pois o autor somente atingirá a maturidade gráfica a partir do esforço e treino necessários para apreender o sistema de escrita a que for submetido.

2.2.3. Sistemas de Escrita

De acordo com Morris (2000), linguagens são sistemas de símbolos, e a escrita é o sistema para representar estes símbolos. Assim, um sistema de escrita pode ser definido como qualquer sistema convencional de marcas ou sinais que representam os enunciados da linguagem. Pode-se dizer que todos os sistemas de escrita representam algum estágio em uma direção progressiva para o sistema de escrita ideal, "o alfabeto".

Ainda de acordo com Morris (2000), a escrita humana é pensada e aprendida, por uma pessoa usando ou um padrão de caligrafia, ou observando e adotando uma letra, combinação de letras, ou símbolos escritos por alguém.

Hilton (1982) define "formato de uma letra de caderno de caligrafia" como o projeto de letras que é fundamental para o sistema de escrita. Ao invés do termo "caderno de caligrafia" Hilton (1982) utiliza o termo estilo de características e características nacionais. Então, estilo de características pode ser definido como o estilo de escrita manual que é adquirido pelo aprendiz e é praticado em um determinado local e em um determinado momento.

De acordo com Hilton (1982), embora muitos países utilizem símbolos do alfabeto iguais, poucos têm projetos de letras idênticos. Isto significa que embora os estilos de características não sejam úteis para a identificação de escrita, eles podem ser utilizados para determinar a nacionalidade do escritor, ou mais corretamente o país de origem do mesmo. No entanto de acordo com Morris (2000), isto não é mais uma declaração verdadeira, pois em função do grande influxo de imigrantes, por exemplo, nos Estados Unidos e a integração de muitas culturas na sociedade americana, as pessoas assimilaram aspectos de outras culturas. Por exemplo, com a integração das escolas americanas nos anos de 1960, muitos estudantes brancos e negros começaram a frequentar as mesmas escolas. Antes disso, existiam formatos de letras frequentemente encontrados na escrita dos americanos africanos e raramente encontrados na escrita de outros grupos raciais.

Segundo Morris (2000), os sistemas de escrita são valorosos porque pessoas que falam a mesma linguagem também desejam ser capazes de se comunicar por meio da escrita. Para atingir este objetivo, elas devem ter um alfabeto de símbolos que possam ser reconhecidos. Com o passar do tempo, muitos sistemas de escrita humana foram desenvolvidos, e com isto muitos dos formatos de letras usados hoje podem ser encontrados presentes no mundo ocidental.

Na região da Ásia, o formato da linguagem de escrita é muito diferente do sistema romano usado em muitos países do Ocidente e é frequentemente único para um pais ou grupo étnico particular.

Além disto, existem sistemas de escrita, como por exemplo, o sistema de Palmer (PALMER, 1935). O Método de Palmer (Figura 2.2) priorizava o domínio do ritmo da escrita, para o que ele chamava de *the writing machine*, e para isso usava os tradicionais adestramentos com ovais e retas ascendentes e descendentes. O sistema de Palmer se tornou o principal sistema de escrita norte-americano do século XX (VILLELA, 2009).

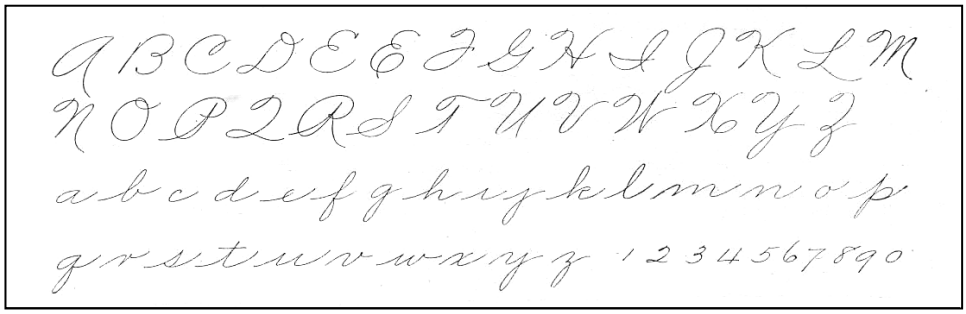


Figura 2.2. Sistema de Palmer

Fonte: [(VILLELA, 2009)]

Cada um destes sistemas é construído em função da cultura de comunicação desejada. O sistema de escrita manual permite que esta comunicação de escrita seja realizada, ou seja, os estudantes primeiramente aprendem a escrever de acordo com as regras definidas para os formatos das letras em um determinado sistema de escrita. Logo, eles variam de um sistema para alguma extensão ou eles praticam a escrita de letras conforme aquelas escritas por alguém que eles admiram e que vive em sua sociedade (MORRIS, 2000).

Para o perito os sistemas são a base sobre a qual pessoas aprendem a escrever. Estes devem saber que escritores não irão aderir fielmente à aqueles sistemas por um longo período de tempo, pois estes normalmente são impacientes e desejam completar a tarefa de escrita tão logo possível. Quando fazem isto, sua escrita passa a ter uma especial singularidade conhecida como individualidade (MORRIS, 2000). Neste contexto, cabe entender o que são classes de características e como estas classes podem ser trabalhadas pelos peritos e, também, como os sistemas computacionais podem extrair elementos que possam auxiliar nas análises grafoscópicas.

2.2.4. Classes de Características

O estudo das classes de características é útil ao perito visto que este precisa conhecer este conceito para não considerar um traçado como individualidade sendo que este traçado nada mais é do que um elemento de classe de características que ele, por inexperiência ou desconhecimento, nunca observou antes.

Definição de Classe de Características

De acordo com Hilton (1982), nem todas as características encontradas em um exame de documento questionado são particulares a uma única pessoa, e àquelas que são comuns a um grupo podem ser descritas como classe de características.

Segundo Morris (2000), a escrita de qualquer pessoa consiste de uma combinação de classes e características individuais, sendo que a extensão e a combinação são dependentes de cada indivíduo. Esta é uma das razões básicas pela qual a escrita manual é identificável, ou seja, pode-se atribuir autoria a um manuscrito. A literatura sobre a identificação de escrita humana não possui informações sobre o tamanho da população de escritores que aprende um sistema de escrita humana e quais variações comuns na forma padrão das letras são frequentemente encontradas em escritas randômicas. Observa-se que a maioria dos sistemas de escrita na América Latina, em particular, a língua portuguesa, deriva do braço lingüístico Indo-Europeu e subseqüentemente do Latim, também chamadas de idiomas/línguas Itálicas ou Românticas, visto serem originariamente utilizadas para escrever os romances (FREITAS et al., 2004). Não obstante sua fonte comum, a principal diferença observada nos sistemas hoje em dia, são detalhes nas letras. Estes detalhes são resultado de preferências estéticas e na confiança de que a escrita da letra é mais fácil com determinadas alterações do que no formato padrão do sistema de origem.

Ainda segundo Morris (2000), classes de características aprendidas pelo escritor são importantes porque elas, em parte, determinam, ou têm certa influência, sobre como a pessoa escreve. Assim, cabe dar continuidade ao estudo por meio da construção das letras e de como as letras podem ser conectadas, formando palavras.

Como as Letras são Construídas e Conectadas

De acordo com Morris (2000), as letras são construídas de acordo com:

- instruções dadas sobre o formato padrão das letras (tal como definida em um caderno de caligrafia);

- as instruções do(a) professor(a);
- a influência de amigos ou membros da família do estudante, entre outros.

Como pode ser observado, o estudante recebe muitas influências e instruções distintas e tenta satisfazê-las. A construção de letras seguindo alguma sequência de instruções (como a realizada em cadernos de caligrafia) ensina o estudante sobre como escrever em um determinado estilo de caligrafia, ou seja, em uma determinada classe de característica.

Nesse contexto, uma questão que precisa ser discutida é se as classes de características no formato que seguem o padrão de "caderno de caligrafia" são as mesmas para pessoas que escrevem com a mão direita (destras) e para pessoas que escrevem com a mão esquerda (canhotas). Ensinar canhotos a escrever, especialmente se o(a) professor(a) é destro(a), pode ser um desafio muito grande.

Segundo Morris (2000), estudantes canhotos precisam de necessidades especiais que requerem atenção do(a) professor(a). Se tais necessidades não forem atendidas, ou se alguma inconsistência no ensino da caligrafia ocorrer durante os anos de formação de tais estudantes, o que deveria ser uma classe de características acaba sendo uma característica individual.

Parte da pesquisa de Zaner-Bloser (2010), desenvolvida nos Estados Unidos, envolve o processo de seleção de um programa de escrita para todos os estudantes. Para isto foram definidas 06 (seis) questões, que devem ser respondidas antes da escolha de um programa de ensino de escrita, a saber:

1. Qual alfabeto é apropriado de acordo com o nível de desenvolvimento mental?
2. Qual alfabeto é mais fácil de escrever?
3. Qual alfabeto é mais fácil de ler?
4. Qual alfabeto é mais facilmente integrado?
5. Qual alfabeto é mais fácil de ensinar
6. Os manuscritos com letras inclinadas ajudam os estudantes na transição para letras cursivas?

Uma classe de características de um sistema é a direção da escrita, ou seja, a direção que ele deseja seguir quando escreve. Por exemplo, o escritor move a caneta, mão e braço da esquerda para a direita, porque é assim que a escrita e a leitura acontecem, ou ele move da direita para a esquerda como no sistema de escrita Arábico ou Hebreu. Outro aspecto que é uma classe de características é o comprimento das

hastes/projeções inferiores/superiores. Alguns sistemas de caligrafia declaram que as projeções inferiores devem ser aproximadamente dois terços da altura das projeções superiores. Em outro sistema de caligrafia, o comprimento das projeções inferiores deve ser o mesmo da altura das letras minúsculas que possuem loops superiores, tais como a letra "h"' (MORRIS, 2000).

Isto não é tão simples assim, pois além da construção das letras, deve-se levar em consideração a forma como as letras são conectadas. Basicamente existem três tipos de conexões: traços que conectam as letras na "posição de baixo" (como nas sílabas, "de" e "er"), traços que conectam as letras na "posição de cima" (como nas sílabas, "bu" e "ou") e traços que conectam as letras em uma combinação de posições (como nas sílabas, "ea" e "ma").

Todos estes elementos farão com que a escrita de cada pessoa seja individualizada e, portanto, passível de identificação.

2.2.5. Individualidade e Características Individuais

A individualidade da escrita é importante, pois caberá ao perito apontar o conjunto de elementos gráficos que caracterizam tal propriedade a cada escrita sob suspeita ou questionamento. Portanto, deve-se entender as características individualizadoras da escrita, bem como, a influência em destas sobre a construção das letras, dos traços de conexão e, ainda, sobre os traços iniciais (ataques) e traços finais (remates).

Definição

Hilton (1982) define características individuais de escrita como: "... mais ou menos peculiar a um específico escritor... constituem os elementos fundamentais de uma identificação ...".

De acordo com Morris (2000) apud Osborne (1929), qualquer característica na escrita ou qualquer hábito de escrita pode ser modificado e individualizado por diferentes escritores de muitas diferentes formas e em uma grande variedade de intensidade. E, assim, a individualidade de escrita de qualquer escritor particular é composta de todas estas características e hábitos comuns e incomuns. Como no processo de identificação de uma pessoa, sempre se deve levar em consideração uma combinação de particularidades que permitem a identificação, e necessariamente quanto

mais numerosos e não usuais forem os elementos e características desta pessoa, mais certeza se terá na identificação, ou seja, quanto maior o número de particularidades de uma pessoa com maior precisão será possível identificar sua escrita.

Para Saudek (1978), individualidade faz parte do "ritmo do escritor" e está associado a isto o desenvolvimento da sua maturidade gráfica. O ritmo de uma escrita manual é caracterizado pela individualidade formada pelas variações entre movimentos e imobilizações durante o ato da escrita.

De acordo com Morris (2000), individualidade também está relacionada com o nível de maturidade gráfica do escritor. Isto não significa que quanto maior o nível de maturidade gráfica maior a individualidade de um escritor. Embora exista uma relação entre estes dois aspectos, uma pessoa com baixo nível de habilidades de escrita pode ter tanto quanto, ou até mais individualidade em sua escrita do que um escritor com um alto nível de maturidade gráfica. Uma pessoa que escreve com um nível alto de maturidade gráfica, por exemplo, no "nível de impulso da sentença/frase", deve escrever com um determinado grau de legibilidade ou sua escrita irá consistir de uma série de linhas sem sentido ou movimentos de caneta.

De acordo com o Grupo de Trabalhos Científicos em Documentos (*Scientific Working Group for Questioned Documents* - http://www.swgdoc.org/), características de identificação são marcos ou propriedades que servem para individualizar a escrita, tais como: formatações e tamanhos relativos das letras.

À medida que um escritor amadurece graficamente, ele se afasta de algumas classes de características que aprendeu e introduz em sua escrita sua individualidade. Com o passar do tempo, estas características individuais tornam-se parte natural de sua escrita porque seus punhos, braços e dedos repetem movimentos que se tornaram habituais. É por meio da combinação destes dois elementos, classes de características e características individuais, que a escrita torna-se identificável (MORRIS, 2000).

Deve-se destacar que por meio da implementação de uma combinação de características individuais é possível desenvolver abordagens computacionais que ofereçam suporte ao processo de identificação de autoria. Cabe, portanto, apresentar algumas dessas características individuais estudadas pela grafoanálise e pertinentes ao estudo de identificação de autoria (MORRIS, 2000):

- nível de habilidade: definido como sendo uma avaliação de beleza aplicada na formação da letra;

- 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;
- forma caligráfica: é a representação pictórica da escrita;
- movimento: a direção do movimento dos instrumentos de escritura (lápis ou caneta) pode ser determinada, através da observação das variabilidades na densidade da tinta ou traço do lápis;
- proporções: refere-se geralmente à simetria das letras individualmente;
- relações de altura: são comparações ou correlações da altura de uma letra ou segmento de letra em relação à outra letra, normalmente dentro da mesma palavra;
- mínimos gráficos: são formados pelos pontos finais, vírgulas, os pingos nos "i´s", acentos (crase, circunflexo, til e agudo) e cedilhas. Essa característica é de grande relevância para os grupos de escritores da língua portuguesa;
- corte da letra "t": podem contribuir mais significativamente na caracterização do escritor que o pingo da letra "i";
- ascendentes e descendentes: são laçadas comuns nas letras cursivas. Podem apresentar formas arredondadas ou pontiagudas, simétricas ou assimétricas;
- pressão: representa a variabilidade de largura do traçado e a deposição de material em uma dada região do traço que pode ser tinta ou grafite;
- alinhamento em relação à linha de base: está associada à capacidade do escritor produzir linhas de textos alinhadas com uma linha guia horizontal fictícia (texto não pautado) ou real (texto pautado);
- velocidade: é frequentemente uma característica essencial para a identificação do autor, pois os movimentos rápidos do objeto de escrita são difíceis de ser duplicados por um falsificador;
- embelezamento: localiza-se usualmente no começo de uma letra, mas pode estar presente ao longo do material escrito;
- entradas e golpes de saída: podem ser movimentos habituais, podendo representar características identificadoras de um escritor;
- retraço: considerada uma característica do escritor, enquanto representa um comportamento natural da experiência de escrita;

- erros de ortografia e espaçamento: escrever incorretamente as palavras pode ser um indicativo de uma característica individual do escritor. Existem escritores que interrompem o curso da escrita entre combinações de letras específicas;
- formato: o formato de um documento questionado pode conter adicionalmente uma característica identificadora.

  A Figura 2.3, abaixo apresentada, exemplifica alguns exemplos de classes de características.

Figura 2.3. Exemplo de características individuais

Observa-se que a lista de características individuais não é pequena e que à medida que cada autor lança mão destes elementos, a escrita humana se torna mais complexa e desafiadora. Além disto, não se pode deixar de lado a influência da individualidade sobre a construção das letras, dos traços de conexão, dos ataques[4] e remates.

Influência na Construção das Letras, Traços de Conexão, Ataques e Remates

De acordo com Morris (2000), a individualidade e o ato de escrita são inseparáveis. Nem todas as escritas têm um alto grau de individualidade, tal como uma assinatura, mas a maioria tem pelo menos algumas características perceptíveis do escritor. Durante o processo de identificação de autoria de um documento manuscrito, o trabalho do perito é determinar quais características, quando tomadas coletivamente correspondem à individualidade do escritor. Então, a maneira como o escritor constrói suas letras é uma característica que deve ser considerada. Assim se, por exemplo, a letra é legível ou o leitor tem que supor qual letra é baseado no contexto da palavra ou da frase. Ocasionalmente, o perito terá que determinar qual característica, ou combinação de características é um acidente e não a escrita normal do escritor.

Ainda segundo Morris (2000), uma vez que movimentos de escrita são atividades habituais, os hábitos de escrita de um escritor definem como uma letra é redigida e como as letras de uma palavra são conectadas. Dependendo destes hábitos, traços de conexão podem ser individuais ou muito comuns, pois traços de conexão assim como a formação de letras possuem padrões. Dentro deste contexto, traços iniciais (ataques) ou finais (remates) podem ser muito simples ou "embelezados". O embelezamento dos traços iniciais não é algo que se aprende na escola, e sim é definido por hábitos de escrita individuais do escritor.

Entende-se com o exposto que o gesto gráfico é complexo, mas passível de análise. Neste sentido, as leis e princípios fundamentais da escrita permitem estabelecer referenciais para as análises grafoscópicas.

2.2.6. Leis e Princípios Fundamentais da Escrita

Solange Pellat em seu livro *La lois de Lécriture* (PELLAT,1927), ditou dois princípios fundamentais da grafia e quatro leis do grafismo que regem o gesto gráfico.

---

4 Todo lançamento tem um início (ataque) e um fim (remate) (MENDES, 2003).

Princípios Fundamentais da Escrita

De acordo com Mendes (2003), existem dois princípios fundamentais da grafia. O primeiro princípio é a *escrita é individual*, ou seja, é resultante de estímulos cerebrais que determinam movimentos e estes criam as formas gráficas. Dentro deste contexto, muito embora os cérebros de todos sejam anatomicamente iguais, a sua função varia de pessoa para pessoa. Já o segundo princípio afirma que *as leis da escrita independem do alfabeto utilizado*. A escrita é resultante de estímulos cerebrais que determinam a criação de fórmulas alfabéticas. Dentro deste contexto, os estímulos são particulares a cada punho e, por isso, os movimentos também são. As formas alfabéticas variam de tipo para tipo. Nessas condições, o que interessa ao perito é a movimentação do punho e não a forma gráfica.

Cabe comentar que tais princípios regem as análises realizadas pelos peritos, sejam estas realizadas sobre textos ou assinaturas manuscritas.

Leis do Grafismo

Ainda de acordo com Mendes (2003) e Cavalcanti e Lira (1996), existem quatro leis do grafismo.

A Primeira Lei, esta relacionada ao "*O gesto gráfico está sob a influência imediata do cérebro. Sua manifestação não é modificada pelo órgão escritor, se este funcionar normalmente e estiver suficientemente adaptado à sua função*". Dentro deste contexto, todas as atividades humanas têm sua origem no cérebro que por sua vez recebe informações por meio dos sentidos, armazenando-as e processando-as. Para produzir o grafismo a pessoa passa por três fases: a *evocação*, a *ideação* e a *execução*. Segundo Cavalcanti e Lira (1996), *evocação* significa buscar informações, imagens, sons ou emoções dentro da memória. *Ideação* é a fase em que o escritor, após ter evocado a ideia, põe em funcionamento a sua criatividade, acrescentando à simbologia sua característica individual. Enquanto que *execução* é a emissão da ordem do cérebro para que o órgão que porta o instrumento escritor execute os movimentos sobre o papel, já evocados e ideados. Se o escritor não possui boa evocação, não poderá idear bem e, portanto, não fará boa execução. Tudo isso depende de vários fatores, inclusive da memória.

A segunda Lei refere-se a "*Quando alguém escreve, o seu eu está em ação, mas o sentimento quase inconsciente dessa ação passa por alternativas contínuas de intensidade entre o máximo, onde existe um esforço a fazer, e o mínimo, quando este*

esforço segue o impulso adquirido". Segundo Cavalcanti e Lira (1996), esta lei é a que regula o automatismo do gesto gráfico. Inicialmente a escrita é o ato consciente, mas a seguir os movimentos se sucedem sem requerer a atenção do escritor. O máximo de intensidade se refere à ação do consciente, e o mínimo, à expressão do subconsciente (MENDES, 2003).

A terceira Lei refere-se ao fato de que *"Não se pode modificar voluntariamente a escrita em dado momento, senão introduzindo no traçado a própria marca do esforço despendido para obter a modificação"*. De acordo com Mendes (2003), "como a escrita é produto do subconsciente, não pode ser controlada pelo consciente". Quando o escritor procura, conscientemente, alterar a sua escrita, provocará um conflito, e esse conflito deixará no registro a marca dessa luta, seja em um pequeno desvio de traço, seja numa hesitação, ou em uma parada anormal do instrumento escrevente ou em um tremor. Complementando Cavalcanti e Lira (1996) afirmam que "sendo o grafismo natural produto das três fases do mecanismo fundamental do gesto gráfico, a interferência dolosa da atenção representa uma violência que inevitavelmente ficará registrada no traçado".

Finalmente a quarta Lei esta relacionada a "*Quando, por qualquer motivo, o ato de escrever se torna particularmente difícil, o escritor instintivamente dá às letras as formas que lhe são mais familiares e mais simples, esquematizando-as de modo que lhe seja mais fácil executar"*. De acordo com Mendes (2003), esta é a lei do mínimo esforço, que pode ocorrer em qualquer outro gesto do homem, uma vez que é um recurso ditado pelo subconsciente.

Assim, cabe ressaltar os comentários feitos por Volpi e Freitas (2013) sobre tais leis, a saber:

- deve-se lembrar de que o ato de escrever exige uma harmonia entre cérebro, cabeça, pescoço, braço, mão e, finalmente, dedos. Todo este sistema, tal qual uma orquestra, precisa aprender a escrever, ou seja, dominar a escrita como forma, como geração de formas e, ainda, de significados. Assim, uma lesão, ou falta de um dos instrumentos na orquestra, pode acarretar a não realização do gesto gráfico;
- o "eu" escritor passa por uma evolução, desde os primeiros rabiscos, até a formação das primeiras letras isoladamente do alfabeto. Esta evolução comprova que a escrita é um ato aprendido, o qual segue nos primeiros anos um modelo apresentado pelo professor ou pais, seja na escola ou em casa. Este aprendizado segue um sinuoso

caminho, entre diferentes estilos, tamanhos e inclinações; até que na fase adulta cada pessoa estabelece um padrão gráfico. Este padrão é memorizado, armazenado no cérebro, e a partir deste momento passa a ser executado sem dificuldades. Não é mais necessário lembrar o modelo aprendido na escola, nem mesmo ficar em dúvida como realizar tal letra ou conjunto de letras, o modelo está automatizado;

- deste modo, de tão automático que se torna o ato de escrever é que não se pode fugir ou fingir ser outro "EU" sem que se deixem marcas gráficas dessa tentativa. Pois, não se está tentando somente movimentar a mão e os dedos de forma diferente, para gerar outro traçado, mas sim, se está exigindo do cérebro que não produza o modelo automático;
- finalmente o último comentário é que este modelo automático, aceita variações e simplificações, quando por diferentes motivos, o escritor está impedido de realizar o modelo completamente. Um exemplo disto é a rubrica, simplificação da assinatura. Por sua vez, estas variações e simplificações não são executadas aleatoriamente, elas possuem uma lógica intrínseca ao ato de escrever particularizado de cada pessoa.

Neste ponto, Volpi e Freitas (2013) apontam que "somente o entendimento destes elementos permite ao perito capturar não somente a complexidade, mas também a beleza da escrita humana".

2.2.7. Elementos Gráficos da Escrita

Existem basicamente dois tipos principais de elementos gráficos, os estáticos e dinâmicos. A grande maioria dos autores da área de grafoscopia apresenta discussões levando em consideração esta classificação. No entanto, em trabalhos como de Plamondon e Lorrete (1989) e Oh e Suen (2002) os elementos estáticos se referem e são nomeados como características globais de um documento questionado, enquanto que os elementos dinâmicos se referem e são nomeados com as características locais do mesmo. Nesse contexto, 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); e locais (utilizam características locais, é feita a partir de letras e palavras, segmentadas do documento manuscrito). Uma discussão mais detalhada sobre esta divisão é apresentada na Seção 2.3, bem como o posicionamento adotado neste trabalho perante tal classificação.

Outra nomenclatura que também pode ser encontrada na literatura é a de elementos genéricos e elementos genéticos. A análise dos atributos genéticos e genéricos é feita tendo como base a dinâmica do traçado da escrita. Esta dinâmica representa um conjunto de fenômenos gráficos, usualmente produzidos de forma inconsciente pelo escritor, muitas vezes denominados de “gestos característicos” (MENDES, 2003).

A seguir são apresentados os elementos básicos da grafia, pois deles decorrem as primitivas que os sistemas computacionais buscam extrair dos documentos manuscritos.

Elementos Básicos da Grafia

De acordo com Cavalcanti e Lira (1996); Mendes (2003); Oliveira et al. (2005); Freitas e Volpi Neto (2007), existe um conjunto de elementos básicos da grafia que são importantes para a identificação de autoria, tal qual apresentado na Tabela 2.1.

Tabela 2.1. Elementos básicos da grafia

Elemento	**Descrição**
Campo gráfico	Corresponde ao espaço bidimensional onde a escrita é feita.
Movimento gráfico	Corresponde a todo o movimento de dedos que o indivíduo faz para escrever, sendo que cada movimento gráfico gera um traço gráfico.
Traço	Corresponde ao trajeto que o objeto da escrita descreve em um único gesto executado pelo autor.
Traço descendente, fundamental, pleno, ou grosso	Corresponde a todo o traço descendente e grosso de uma letra.
Traço ascendente ou perfil	Corresponde ao traço ascendente e fino de uma letra.
Ovais	Corresponde aos elementos em formas de círculo das letras “a, o, g, q”, dentre outras.
Hastes	Corresponde a 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	Corresponde a todos os plenos (descendentes) do “g”, “j”,”y”, “f”, etc. a partir da zona média até embaixo.
Bucles	Corresponde a 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	Corresponde ao esqueleto da letra, a parte indispensável da sua estrutura.
Parte secundária ou acessória	Corresponde ao revestimento ornamental ou parte não necessária à sua configuração.

Ainda de acordo com Cavalcanti e Lira (1996), tais elementos podem estar posicionados em zonas distintas e denominados de acordo com a Tabela 2.2. Esta classificação das zonas ou regiões facilita o processo de extração de primitivas quando se pretende desenvolver um sistema computacional, seja para identificação de autoria, seja para reconhecimento de palavras manuscritas.

Tabela 2.2. Áreas de posicionamento das letras

Posicionamento	**Descrição**
Zona inicial	Corresponde a área onde se encontra o ponto no qual se inicia a letra.
Zona final	Corresponde a área onde se encontra o ponto no qual termina a letra.
Zona superior	Corresponde 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	Corresponde 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	Corresponde a zona baixa da escrita a partir da base de todos os ovais.

Cabe, portanto, estudar e distinguir os elementos dinâmicos dos elementos estáticos da grafia, visto que isto possibilita a categorização das primitivas a serem extraídas pelos diferentes autores que já desenvolveram estudos relacionados ao tema tratado neste trabalho.

Elementos Dinâmicos da Escrita

Pode-se descrever a gênese gráfica como sendo a sequência de movimentos determinados pelos impulsos cerebrais que dão origem a forma. Dessa maneira, a gênese gráfica é materialização dos impulsos que emanam do centro nervoso da escrita, e, portanto é o elemento dinâmico, específico e inerente de cada pulso (MENDES, 2003).

Assim, de acordo com os trabalhos de Mendes (2003) e Calvacanti e Lira (1996); os elementos dinâmicos da escrita podem ser descritos pela pressão, gesto gráfico ou característico, movimento, mínimos gráficos e velocidade, descritos a seguir;

- **Pressão:** corresponde a força ou intensidade do traço, por sua característica dinâmica, possui estreita relação com a rapidez, com a continuidade e com a irradiação do impulso gráfico. A pressão pode ser resumida em quatro características:

- escrita tensa: na qual os movimentos gráficos são retos, firmes e seguros; e
- escrita frouxa: na qual existe um déficit de tensão nos movimentos, os quais são mais ou menos sinuosos, ondulados ou torcidos em qualquer de seus sentidos direcionais;
- deposição de tinta de caneta esferográfica e
- deposição de grafite;

- **Gesto gráfico ou característico:** na escrita existem modalidades de traços ou letras que chamam a atenção, pois imprimem ao traçado uma fisionomia especial que nenhuma outra pessoa poderia reproduzir da mesma maneira. Os elementos gráficos que podem formar o gesto característico são:
  - gancho: consiste de um movimento de regressão encontrado nos finais das letras ou na barra da letra "t";
  - clave: carrega todo o golpe de energia sobre a zona final do traçado que fica em forma de ponta quebrada;
  - golpe de sabre: refere-se ao movimento promovido por um impulso da caneta, que pode afetar as barras da letra "t" e as partes inferiores das letras (lançadas inferiores);
  - movimento em triângulo: são produzidos principalmente nas laçadas inferiores dos "t", "g", "y" e na barra da letra "t", podem também aparecer na circunferência das letras da zona média;
  - bucle: se apresenta na circunferência das letras da zona média, nas maiúsculas e nas ligações;
  - guirlanda: consiste num movimento em forma de arco aberto para cima, presente nos traços iniciais e finais e nas barras das letras "t";
  - arco: é encontrado preferencialmente nas zonas inicial, superior e média (nas ligações);
  - espiral: está presente nas letras maiúsculas;
  - inflação: apresenta um tamanho exagerado, presente nas maiúsculas;
  - laço: é uma espécie de movimento de retorno ao ponto de partida e
  - serpentina: afeta especialmente os traços iniciais e finais e as letras "m" e "n";
- **Movimento:** duas escritas que apresentam semelhanças nas formas podem ser diferentes quando analisadas sob a ótica do movimento da caneta. A direção do

movimento dos instrumentos de escrita (lápis ou caneta) pode ser determinada, frequentemente, através da observação das variabilidades na densidade da tinta ou traço do lápis;

- **Mínimos Gráficos:** são formados pelos pontos finais, vírgulas, os pingos nas letras "i", acentos (crase, circunflexo, til e agudo) e cedilhas e, ainda, por algum tipo de embelezamento ou gesto característico ao escrever. Uma pequena porção de escritura como um mínimo gráfico pode, em alguns casos, se tornar uma característica identificadora relevante;
- **Velocidade:** é frequentemente uma característica essencial para a identificação do autor. Um movimento rápido do objeto de escrita é difícil de ser duplicado por um falsificador. Os seguintes elementos de análise descrevem alguns indicadores da escrita, podendo classificá-la em rápida e lenta. A classificação rápida esta relacionada ao traçado tenso, sem tremor; ao alongamento na finalização das letras "e" e corte das letras "t"; as palavras ou letras conectadas; a aparência aplainada do texto, a redução da legibilidade. Já a classificação lenta aponta para a vacilação, o tremor, a escrita mais angular; o cruzamento da letra "t" em posição correta; a parada e o recomeço abrupto (clave); a escrita é feita de letras individuais e legíveis; o movimento pode apresentar ornamentos.

Alguns destes elementos podem ser claramente observados na Figura 2.3. É importante observar que os elementos dinâmicos da escrita contribuem para que a gênese seja um elemento específico da escrita, visto que depende das condições psicossomáticas de cada indivíduo. Assim como as características físicas, fisiológicas e químicas variam de indivíduo para indivíduo, o gesto gráfico também sofre esta variação. Uma vez que não existem duas pessoas de movimentos iguais, não podem existir grafismos idênticos.

Elementos Estáticos da Escrita

Segundo discussões apresentadas no trabalho de Mendes (2003), "a forma gráfica é o desenho, o feitio da escrita criado pelo movimento – a gênese. Em razão disso, é o elemento estático do grafismo".

Ainda segundo Mendes (2003), a forma pode ser considerada o elemento genérico, uma vez que é comum a todos que escrevem usando o mesmo tipo de alfabeto. A forma não é um elemento individualizador, mas sim as alterações a que esta

está sujeita (seja qual for a razão da alteração, tais como modificações para imitar uma pessoa, modificações para disfarçar a própria escrita, etc.).

Não se deve confundir a gênese com a forma gráfica. Dentro deste contexto, Söderman e O´Connell (1953) em seu livro *Manuel d´enquête criminelle modern* afirmam que um falsificador imita a forma e não a gênese, ou seja, imita o "desenho" e não o "movimento".

Considerando-se o exposto, pode-se mencionar como elementos estáticos todas as formas geradas pelos movimentos durante o ato de escrever.

Elementos Formais da Escrita

De acordo com Mendes (2003), os elementos formais da escrita podem ser classificados em objetivos e subjetivos. Nesse contexto, os elementos objetivos são:

- calibre: se refere ao tamanho das letras (por exemplo: macrografia - letras grandes, micrografia – letras pequenas de difícil leitura, alteração do calibre, gladiolagem – redução do tamanho das letras a medida que estas são escritas);
- inclinação axial: como já descrito anteriormente, é o comportamento dos eixos gramaticais (por exemplo: verticalizada, inclinada à direita, inclinada à esquerda, reversão de eixo – quando a inclinação das letras sofre uma mudança no sentido);
- espaçamentos gráficos: se referem às distâncias que guardam entre si os gramas, letras, vocábulos e linhas de escrita (por exemplo: intergramaticais, interliterais, intervocabulares, interlineares);
- andamentos gráficos: se referem aos momentos gráficos, ou seja, são os grupos de letras, de um mesmo vocábulo, que se interligam, separando-se de outros que lhe seguem. As silabações possuem normas gramaticais e, no andamento gráfico, são divisões arbitrárias, adotadas pelo escritor por força do habito sem obedecer a qualquer norma. Cada grupo de letras constitui um momento gráfico (por exemplo: um momento, dois momentos, três momentos);
- alinhamentos gráficos: corresponde ao comportamento da escrita em função de uma linha de pauta, ideal ou impressa (por exemplo: ascendente, descendente, acima da linha, apoiado na linha, sobre a linha de pauta, irregular);
- valores angulares e curvilíneos: se referem à predominância dos ângulos ou das curvas no grafismo (por exemplo: valores angulares, valores curvilíneos, misto)
- relações de proporcionalidade gramatical representam a relação do tamanho que as letras de uma palavra guardam entre si.

Pode-se observar que alguns dos elementos acima expostos podem ser observados na Figura 2.3. “Os elementos formais objetivos não são identificadores. No entanto, alguns deles podem ter alguma significação” (MENDES, 2003). Também deve-se destacar, que alguns elementos formais objetivos revelam hábitos inconscientes do escritor que os falsificadores, às vezes, não reproduzem nas imitações, pois se prendem a detalhes não aparentes.

Segundo Mendes (2003), os elementos subjetivos, não podem ser demonstrados, embora percebidos pelo perito. Dentro deste contexto, os elementos subjetivos mais apontados pelos especialistas consistem no *aspecto geral da escrita*, representada pela qualidade do traço, que por sua vez, depende do *grau de habilidade de punho*, do *ritmo de escrita*, da *velocidade* e do *dinamismo gráfico*.

Os elementos formais objetivos são de ordem genérica, por isso, comum a muitos grafismos, e consequentemente não são decisivos quando da conclusão sobre a autoria da escrita. No entanto, eles podem apresentar alguns aspectos que assumem maior significação, como por exemplo, o calibre dos traçados e a inclinação axial (ver Tabela 3.2, na qual pode-se observar a importância discriminatória da característica inclinação axial nos resultados obtidos pelo método proposto neste trabalho). Algumas vezes determinado grama é grafado em calibre maior que os demais. Os falsários podem não perceber esse detalhe ao reproduzir o modelo de calibre predominante. Na inclinação axial, também, as reversões de alguns passantes nem sempre são imitadas pelos falsificadores (MENDES, 2003).

Dentro deste contexto, os elementos formais subjetivos são mais importantes, do ponto de vista da identificação de autoria, uma vez que normalmente, os elementos formais objetivos são de mais fácil reprodução, pelos falsários, do que os elementos subjetivos.

2.2.8. Considerações Finais sobre a Escrita Humana

A Seção 2.2 apresentou uma fundamentação teórica da escrita humana e dos elementos que fazem desta um elemento biométrico. Ao longo da fundamentação apresentada foram descritos aspectos que levam em consideração elementos fisiológicos do escritor, ou seja, que são natos a cada escritor, bem como elementos obtidos durante seu processo de formação. Ambos os elementos são importantes para o processo de individualização da escrita. Também foram abordados os principais aspectos que

influenciam a construção do traçado de um escritor. Uma discussão sobre os elementos básicos da grafia também foi apresentada, uma vez que estes elementos são avaliados pelos peritos durantes suas análises grafotécnicas.

Esta fundamentação foi necessária para entender o mecanismo da escrita humana, os quais serão os pilares para a definição do método computacional que tem como objetivo automatizar o processo de identificação de autoria. Deve-se ressaltar também, que no método proposto são utilizadas características grafométricas, o que reforça a necessidade por um conhecimento profundo dos aspectos discutidos ao longo deste capítulo.

Para complementar a fundamentação necessária para o desenvolvimento deste trabalho, a próxima Seção apresenta um levantamento sobre métodos computacionais propostos na literatura que tem como objetivo, tal como neste trabalho, auxiliar e agilizar o processo de identificação da escrita humana. Deve-se ressaltar que trabalhos como de Hertel e Bunke (2003), Schlapbach e Bunke (2004), Chen et al. (2010), Luna et al. (2011), Zois e Anastassopoulos (2000) e Pervouchine e Leedham (2007) (descritos nas próximas Seções) também fazem uso de características grafométricas para automatizar o processo de identificação de autoria.

2.3. Autoria em Documentos Manuscritos

Segundo Srihari e Shi (2004), assim como em outros campos relacionados à Ciência Forense, o exame clássico de identificação de autoria em manuscritos é baseado no conhecimento e experiência do perito. Tem-se então a seguinte situação problema diante de medidas não objetivas e decisões que nem sempre podem ser reproduzidas. Portanto, tentativas de oferecer suporte automático e/ou semiautomático aos métodos tradicionais vem sendo estudadas.

Ainda, segundo Srihari e Shi (2004), as modernas tecnologias oferecem mecanismos para a construção de grandes bancos de dados criminais, visto que bases de dados digitais para aplicações forenses têm um importante papel na investigação criminal. Nesse contexto, sistemas de computadores eficientes e grandes quantidade de dados, a perícia forense se torna mais eficiente e precisa. Vários tipos de dados podem ser coletados de registros criminais ou evidências e, portanto, as bases de dados forenses são de muitos tipos, tais como: bases de impressões digitais, bases de registros criminais e coleções de registros multimídias incluindo imagens, vídeos, documentos e textos.

Estas bases são, também, de suma importância para a validação dos sistemas automáticos e semiautomáticos de identificação de autoria.

Dentro deste contexto, nas próximas Seções é apresentado um levantamento bibliográfico sobre as principais abordagens e sistemas para identificação de autoria presentes na literatura. Este levantamento, apresentado em Amaral et al. (2012a), é fundamental para que se possa fazer uma análise crítica sobre as características grafométricas utilizadas como parte do método proposto neste trabalho, bem como, sobre as taxas de acerto obtidas com o uso de tais características.

2.3.1. Classificações para Autoria em Documentos Manuscritos

Segundo Bensefia et al.(2005), existem duas diferentes classificações para a autoria de documentos manuscritos, são elas: a identificação e a verificação.

No processo de verificação de autoria procura-se avaliar dadas duas amostras de manuscrito, se as mesmas são ou não de um mesmo escritor (1:1). Nestas abordagens, normalmente, apenas duas classes são assumidas: autoria (associação) e não autoria (dissociação).

Por outro lado, no processo de identificação de autoria procura-se avaliar com base em um conjunto de escritores e um documento questionado, a qual escritor pertence o documento questionado, ou seja, atribui-se ao documento questionado sua autoria (1:N). Nestas abordagens são construídas classes para o escritor do documento questionado e para cada um dos escritores presentes na base de dados que contém o conjunto de escritores. Devido à grande quantidade de classes geradas, o problema de identificação de autoria não pode ser considerado uma tarefa simples de classificação (BENSEFIA et al., 2005).

Ainda de acordo com Bensefia et al. (2005), um sistema para identificação de autoria deve fornecer um subconjunto de candidatos relevantes (ou seja, um *ranking*), nos quais análises complementares podem vir a ser realizadas por um perito. Foi possível observar ao longo da revisão bibliográfica realizada que o uso de *rankings*, nomeados pelos pesquisadores como *Top*, para classificar os resultados do processo de identificação de autoria é uma prática comum. Isto porque, uma redução no espaço de busca dos suspeitos promove uma redução no tempo dedicado à análise grafotécnica realizada pelo perito.

Este trabalho concentra-se no processo de identificação de autoria, usando para tanto um conjunto de características grafométricas. Dessa forma, a seguir é apresentado

o levantamento bibliográfico realizado das diferentes abordagens para identificação de autoria. Este levantamento foi organizado levando-se em conta a classificação para abordagens de identificação de autoria proposta por Sreeraj e Idicula (2011).

2.3.2. Abordagens para Identificação de Autoria em Documentos Manuscritos

De acordo com Bulacu et al. (2007), dois importantes fatores naturais estão em conflito direto na tentativa de identificar uma pessoa com base em amostras de manuscritos: variações entre diferentes escritores (variabilidade interpessoal) em contraposição a variabilidade na escrita de um único escritor (variabilidade intrapessoal). Nesse contexto, as abordagens automáticas para identificação de autoria consistem na extração de representações computacionais (características) com o objetivo de maximizar a separação entre diferentes escritores, enquanto apresentam um padrão de escrita nas amostras do mesmo escritor.

Segundo He et al. (2008), abordagens para identificação de autoria podem ser classificadas de diferentes formas, contudo a mais simples e direta é a divisão em *online* e *offline* que se refere ao processo de aquisição ou captura do documento manuscrito para sua posterior análise.

Outras categorizações que levam em consideração o tipo de características extraídas dos manuscritos são encontradas na literatura. Uma classificação bastante interessante é apresentada no trabalho de Sreeraj e Idicula (2011) na qual as abordagens *offline* e *online* são classificadas de acordo com o nível de granularidade da característica extraída (por exemplo: documento, parágrafo, linhas, palavras e caracteres). Trabalhos como o de Siddiqi e Vicent (2008) apresentam uma divisão entre características locais e globais levando em consideração as características extraídas de um manuscrito.

Neste trabalho adotaram-se dois níveis de granularidade para categorizar as características para identificação de autoria. Consideram-se como características Globais, aquelas que envolvam a extração de informações em nível de documento, parágrafo e linhas, enquanto que aquelas que envolvam a extração de informações em nível de palavra e caractere serão consideradas como Locais. Para uma melhor compreensão a Figura 2.3 apresenta este esquema de classificação.

Neste contexto, esta Seção apresenta uma revisão das principais abordagens para identificação de autoria propostas na literatura levando em consideração a classificação apresentada na Figura 2.3. Deve-se destacar que um maior enfoque será dado às

abordagens *offline* uma vez que o método proposto neste trabalho foi definido a partir desta abordagem.

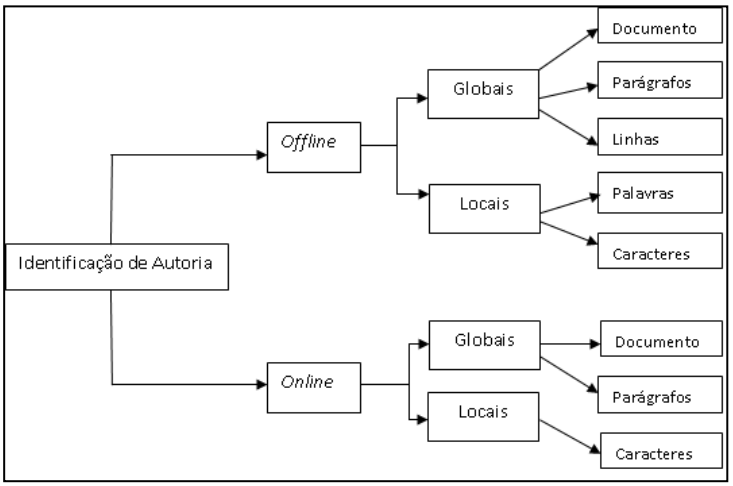


Figura 2.4. Classificação para abordagens de identificação de autoria
[Fonte: Adaptado de (SREERAJ; IDICULA, 2011)]

2.3.3. Identificação de Autoria *Online*

De acordo com He et al. (2008), nas abordagens *online* um equipamento é conectado ao computador, o qual converte o movimento de escrita em uma sequencia de sinais e os envia ao computador. A forma mais comum de equipamento é uma mesa digitalizadora (*tablet*) que consiste de uma caneta plástica ou eletrônica e uma superfície sensível a pressão, sobre a qual os usuários realizam seus manuscritos. Uma vez que informações dinâmicas do processo de escrita são capturadas, as quais representam importantes aspectos que individualizam a escrita humana, as abordagens *online* apresentam, em geral, taxas de classificação melhores que as abordagens *offline*.

De acordo com Sreeraj e Idicula (2011), a tarefa de identificação de autoria *online* é considerada menos difícil do que a *offline*, pois contém mais informações sobre o estilo de escrita de uma pessoa, tais como: velocidade, inclinação axial, pressão, progressão; todas capturadas diretamente durante a escrita. Infelizmente, sistemas *online* não são aplicados em muitos casos, uma vez que abordagens *online* não se prestam para auxiliar em análises grafoscópicas que visam encontrar o escritor de um manuscrito existente (HE et al., 2008). Nas próximas Seções são apresentados alguns dos trabalhos desenvolvidos e em desenvolvimento em cada uma das categorias das abordagens *online* apresentadas na Figura 2.3.

Abordagem Global

Na abordagem global de documentos *online*, diversas características podem ser extraídas de um documento manuscrito, entretanto duas delas são mais conhecidas: as relativas aos documentos e as relacionadas aos parágrafos.

No primeiro caso, ou seja, as ***características relativas ao documento***, o trabalho de Karthik et al. (2008) apresenta uma abordagem para classificação de série temporais multivariadas com aplicação para o problema de identificação de autoria *online*. O esquema de classificação resultante se aplica a muitas tarefas de discriminação de séries temporais e mostra resultados interessantes para a tarefas de reconhecimento de escrita manual *online*.

Hangai et al. (2000) propuseram uma abordagem *online* para identificação de autoria usando a altura e distância angular da caneta e também a pressão da escrita em tempo real. Os resultados experimentais com informações de 24 escritores atingiram taxas de autenticação de 98%.

Thumwarin e Matsuura (2004) apresentam um método *online* para reconhecimento de escrita tailandesa baseado na velocidade do baricentro da ponta da caneta. Experimentos de reconhecimento de escrita foram executados em uma base de dados composta de 6642 scripts escritos em tailandês por 81 escritores. As taxas de erros Tipo I (falsa rejeição) e Tipo II (falsa aceitação) foram 1.50% e 0.05%, respectivamente.

No trabalho de Tsai (2005) é apresentado um método para identificação *online* de escrita baseado no modelo de distribuição de pontos (PDM - *Distribution Point Model*). Este modelo fornece uma forma de descrever as variações das formas usando uma abordagem paramétrica. Foram realizados experimentos com 20 pessoas e as melhores taxas de identificação obtidas foram de 97,3%.

No segundo caso, ***as características extraídas dos parágrafos***, o trabalho de Jin et al. (2005) apresenta uma combinação de características estáticas e dinâmicas para identificação *online* de escrita. Experimentos foram conduzidos utilizando manuscritos de 55 pessoas da base NLPR (*National Laboratory of Pattern Recognition*). Os resultados mostraram que a combinação de métodos pode melhorar as taxas de identificação, bem como reduzir o número de características utilizadas. Neste trabalho, taxas próximas a 55% foram obtidas.

Namboodiri e Gupta (2006) propuseram um método para identificação de autoria que usa um conjunto específico de primitivas da escrita *online* que envolvem informações relativas ao formato das curvas do traçado O método permite que sejam apreendidas as propriedades do texto e dos escritores simultaneamente, uma vez que o mesmo pode ser usado com diferentes textos e em diferentes línguas. Os resultados obtidos chegaram a 87% com 10 diferentes escritores.

Abordagem Local

Na abordagem local de documentos *online*, diversas características podem ser extraídas de um documento manuscrito, entretanto a mais conhecida refere-se aos caracteres.

No caso das ***características extraídas dos caracteres***, o trabalho de Tan et al. (2009) apresenta um método para identificação de escrita independente de texto, integrado a um sistema industrial de reconhecimento de escrita que é usado para executar segmentação automática de um documento manuscrito em nível de caractere. Uma abordagem baseada na lógica Fuzzy foi adotada para estimar distribuições estatísticas de protótipos de caracteres em um alfabeto base. Estas distribuições modelam a individualidade do estilo de escrita de cada escritor. Taxas de 99,2% foram obtidas em bases de dados de 120 escritores, tendo como condição uma quantidade mínima de texto necessária (aproximadamente 160 caracteres ou aproximadamente 3 linhas de texto).

Nogueras e Zanuy (2012) propuseram um sistema para reconhecimento biométrico que pode ser adequadamente aplicado a pequenas sentenças de texto (palavras). Características relativas aos "golpes de entrada" (ataques) no traçado são consideradas como primitivas básicas do processo de identificação, levando-se em conta os movimentos dinâmicos de baixar e levantar da caneta. Taxas de identificação de 92,38% foram obtidas e uma função de custo de detecção mínima de 4,6 foi atingida com 370 usuários e apenas uma palavra. Resultados de 96,46% foram obtidos quando uma combinação de duas palavras foi utilizada.

2.3.4. Identificação de Autoria *Offline*

Abordagens para identificação de autoria *offline* utilizam manuscritos digitalizados após a escrita ter sido realizada, sendo obtidos arquivos que correspondem

a representações de imagens bidimensionais. Segundo Said et al. (2000), apesar dos contínuos esforços, a pesquisa na área de sistemas para identificação de autoria *offline* ainda é uma questão desafiadora.

De acordo com He et al. (2008), sistemas *offline* dependem de arquiteturas mais sofisticadas para executar a tarefa de identificação de autoria do que os sistemas *online*, e mesmo assim seus resultados ainda estão abaixo dos obtidos com as abordagens *online*.

Segundo Plamondon e Lorrete (1989), as abordagens para identificação de autoria *offline* podem ser divididas em dois grupos: dependentes do texto e independentes do texto. Identificação dependente do texto iguala um ou um pequeno grupo de caracteres/palavras e conseqüentemente requer que o escritor redija o mesmo texto fixo no manuscrito. Ainda segundo o mesmo autor, as abordagens de identificação independentes de texto, não fazem uso das características da escrita de alguns caracteres ou palavras específicas, uma vez que levam em consideração aspectos relativos ao *layout* do manuscrito, dentre outras características.

Deve-se ressaltar que neste trabalho apresenta-se uma abordagem para identificação de autoria que é dependente de texto, uma vez que tais estudos se fundamentam nos aspectos grafométricos da escrita, e para tanto, em geral, levam em consideração aspectos específicos dos parágrafos, palavras e letras. Também é importante destacar que a base utilizada neste trabalho (base de cartas forenses PUCPR, FREITAS et al. (2008)) é composta por 03 manuscritos de cada escritor, redigidos seguindo o padrão da carta forense PUCPR – ver Capítulo 3. Nas próximas Seções são apresentados alguns dos trabalhos desenvolvidos e em desenvolvimento em cada uma das categorias das abordagens *offline* apresentadas na Figura 2.3.

Abordagem Global

Na abordagem global de documentos *offline*, diversas características podem ser extraídas de um documento manuscrito, entretanto três delas são mais conhecidas: as relativas aos documentos, as relacionadas aos parágrafos e as atreladas às linhas.

No primeiro caso, ***as características extraídas do documento***, o trabalho de Said et al. (1998) propõe uma abordagem independente de texto para identificação de escrita que deriva de características relativas a textura usando filtros de Gabor e Matrizes de co-ocorrência de Escalas de Cinza. Resultados com taxas de 96% de acerto foram

obtidos utilizando 150 documentos teste de 10 diferentes escritores. Trabalhos similares podem ser observados em (TAN, 1998; ZHU et al., 2001).

Said et al. (2000) apresentam uma abordagem global baseada na análise de textura, na qual cada manuscrito de um escritor é considerado como uma diferente textura. Isto permitiu aos autores aplicar algoritmos padrões de reconhecimento de texto para a tarefa de identificação (por exemplo, a técnica de filtragem de multicanais de Gabor). Resultados de 96% de taxas de acerto foram obtidos em documentos teste de 40 diferentes escritores.

Ainda utilizando características relativas à textura, o trabalho de Helli e Morghaddam (2010) apresenta um método independente de texto para identificação de autoria na escrita Persa. Este método é baseado em características que são extraídas do manuscrito utilizando filtros de Gabor e de XGabor. As características extraídas de cada escritor são inseridas em um grafo de características relacionadas. Este grafo é construído usando relações entre as características extraídas por meio da aplicação do método *Fuzzy*. Os resultados experimentais apresentaram desempenho próximo a 100% de acerto para uma base de dados de 100 escritores.

He et al. (2008) apresentam um método para identificação de autoria em manuscritos Chineses independente de texto, baseado no modelo de Árvores Escondidas de Markov (HMT - *Hidden Markov Tree*). De acordo com os autores do trabalho, os resultados obtidos em seus experimentos apresentaram taxas de reconhecimento melhores quando comparados com modelos que utilizam filtros de Gabor multidimensionais e também reduziram significativamente o tempo de computação necessário para a realização dos experimentos.

O trabalho de Bulacu et al. (2007) apresenta uma combinação de características globais baseadas: na extração de características relacionadas a textura de um manuscrito; e em características alográficas. Segundo estes autores por meio da junção de tais características foi possível obter taxas de acerto de 82% tanto para o processo de identificação quanto para o processo de verificação de autoria.

Baranoski et al. (2005) propõem um método de extração global de características que utiliza a primitiva grafométrica denominada de inclinação axial. Esta primitiva representa o ângulo de inclinação da escrita em relação ao eixo vertical sendo o eixo horizontal representado por uma linha de base imaginária. Com base em uma análise de todo o documento é utilizada uma distribuição de tendências angulares de inclinação, o

que faz com que a avaliação seja similar a utilizada pelo perito forense. Foram realizados experimentos com 200 escritores distintos para a fase de treinamento e 115 escritores distintos para a fase de teste, obtendo resultados próximos a 95% de acerto para verificação de autoria, as comparações realizadas foram (1:1). Cabe destacar que a inclinação axial é uma das características grafométricas utilizadas no método proposto neste trabalho.

Karunakara e Mallikarjunaswamy (2011) propuseram uma abordagem para identificação de escrita, dependente de texto, na língua inglesa usando o modo de decomposição empírica (EMD - *Empirical Mode Decomposition*). Para estes autores, cada manuscrito de um escritor é considerado como uma diferente textura. Estas texturas são consideradas como uma única característica para a identificação de escrita. Os resultados obtidos atingiram taxas de acerto em torno de 94% com o classificador KNN e utilizando características registradas de manuscritos de 50 diferentes escritores.

No segundo caso, ***as características extraídas dos parágrafos,*** Siddiqi e Vincent (2008) propõem um método que leva em consideração as formas frequentes do traçado tal como definido no trabalho de Bensefia et al. (2005). A abordagem de Bensefia et al. (2005) utiliza a forma como as letras são desenhadas e segmentadas como se o objetivo desta tarefa fosse ler o texto. No entanto, no trabalho de Siddiqi e Vincent (2008) o reconhecimento do escritor é independente do que está escrito e se refere à forma com que fisicamente as linhas ou *loops* são produzidos. No trabalho de Bensefia et al. (2005) o documento manuscrito é dividido em um grande número de sub-imagens, e imagens morfologicamente similares são agrupadas em *clusters*. Com base em tais *clusters*, os padrões que ocorrem com frequência para um escritor são então extraídos. Nesse contexto, o escritor do documento questionado é definido encontrando-se a classe de estilo de escrita a qual seu documento pertence. Foram utilizados nos experimentos duas bases de dados a IAM (*Informatik und Angewandte Mathematik*) (650 escritores) e a RIMES (*Reconnaissance et Indexation de données Manuscrites et de fac similÉS/Recognition and Indexing of handwritten documents and faxes*) *(*375 escritores) e taxas de acerto em torno de 91% foram atingidas.

Schomaker et al. (2007) descrevem uma abordagem para identificação de autoria usando contornos de fragmentos de componente conectados a partir de manuscritos redigidos em estilo livre de escrita. O escritor é caracterizado como um gerador estocástico de padrões, produzindo uma família de fragmentos de caracteres (chamados

de *fraglets*). Usando um *codebook* desses *fraglets* a partir de uma base de treinamento independente, a distribuição de probabilidades dos *fraglets* é computada para a base de teste. Resultados revelaram uma alta sensibilidade do histograma dos *fraglets* para a identificação de escritores individuais. Experimentos em grande escala utilizando redes de Kohonen foram realizados, e os melhores resultados obtidos chegaram a atingir 97% de acerto.

Kumar et al. (2011) apresentam um método para extrair informações manuscritas de zonas de textos impressos de imagens a partir de documentos com conteúdo misto. Nesse trabalho foram usadas características que codificam formas locais de texto. Foram construídos dois *codebooks* das características de forma extraídas de um conjunto de manuscritos e dos textos de documentos impressos respectivamente. Para o processo de classificação foi utilizado o classificador SVM (*Support Vector Machine)* e taxas de reconhecimento de 98% foram obtidas, tendo sido utilizadas 732 imagens para a fase de treinamento e 625 imagens para a fase de teste.

No último caso, as ***características extraídas das linhas,*** Luna et al. (2011) propõem uma abordagem para resolver o problema de identificação de autoria de manuscritos com uma modificação no algoritmo da família de classificadores supervisionados ALVOT (*Votacion Algorithm*). A metodologia proposta nesse trabalho apresenta e exemplifica a possibilidade de identificação de autoria por meio de uma abordagem intermediária entre características de textura e estruturais. Características tanto em nível de linha quanto em nível de palavra são extraídas da imagem do manuscrito. Em nível de linha o espaço percentual da margem esquerda e margem direita, a separação entre linhas, a direção geral da escrita e o espaço entre palavras são considerados. Deve-se destacar que algumas destas características estão sendo utilizadas neste trabalho de doutorado para compor o conjunto de primitivas grafométricas.

No trabalho de Luna et al. (2011) um esquema não diferencial de pesos tradicional atribui um valor de peso para cada uma das características extraídas. Este mecanismo fornece flexibilidade suficiente para precisamente discriminar padrões pertencendo a classes onde o mesmo subconjunto de características é mais relevante, mas com diferentes proporções em cada caso. Foram realizados experimentos com 30 escritores atingindo resultados próximos a 92%.

O trabalho de Bensefia et al. (2005) propõe-se a demonstrar que a tarefa de identificação e verificação de autoria pode ser realizada usando características locais

como grafemas extraídos a partir da segmentação de manuscritos. Um modelo de recuperação baseado em texto é usado para o estágio de identificação de autoria. Isto permite o uso de um espaço de características particulares baseado nas freqüências das características. Documentos questionados são projetados neste espaço de características. Neste trabalho taxas de 95% de acerto foram obtidas na base PSI e 86% na base IAM. De acordo com Bensefia et al. (2005) o problema de identificação de autoria pode ser definido como um processo de encontrar conteúdos gráficos (conjuntos de grafemas do documento a ser identificado) em um grande conjunto de documentos (base de dados de treinamento). Os documentos recuperados devem ser classificados de acordo com uma similaridade com o documento questionado.

No trabalho de Schlapbach e Bunke (2004) um sistema independente de texto para identificação de autoria usando Modelos Escondidos de Markov (HMM - *Hidden Markov Models*) foi proposto. Para cada escritor conhecido foi construído um reconhecedor individual e treinado o mesmo com linhas de texto escritas por este escritor (cada linha de texto apresentada ao sistema é normalizada com relação à inclinação, altura e obliquidade e transformada em um vetor de primitivas). Uma linha de texto de origem desconhecida é apresentada para cada um destes reconhecedores identificados. Como resultado o reconhecedor conhecido que tiver maior *score* é definido como sendo o escritor da linha de texto cujo escritor é desconhecido. O sistema foi testado como 2200 linhas texto de 50 escritores e teve um taxa de reconhecimento de 94.4%.

Hertel e Bunke (2003) apresentam um sistema para identificação de escrita com características derivadas das linhas de um manuscrito (continuidade do traçado, regiões fechadas, contornos superiores e inferiores). Estas características são subsequentemente usadas no classificador KNN (*K Nearest Neighborhood*) que compara o vetor de características extraídas de um texto de entrada com um conjunto de vetores protótipos que pertencem a escritores com identidade conhecida. Este método foi testado com uma base de dados com páginas de textos manuscritos produzidas por 50 escritores. Taxas de reconhecimento próximas a 90% foram atingidas usando uma única linha do texto como entrada. A taxa de reconhecimento aumenta para quase 100% se a página toda do texto é fornecida como entrada para o sistema.

No trabalho de Chen et al. (2010) foi desenvolvido um método para detecção e remoção de linhas de referência (pré-impressas em papéis). Este método foi testado para

identificação de escritor na língua Árabe por meio de experimentos usando SVM, e tendo como características para identificação de escrita um conjunto de primitivas que levam em consideração informações relativas ao contorno de segmentos adjacentes. Os resultados iniciais, com 60 diferentes escritores e taxas de acerto em torno de 54,9% (com a remoção das linhas de referência pré-impressas) mostraram que em situações realísticas, nas quais linhas de referência são esperadas, removê-las melhora significantemente as taxas de identificação.

Abordagem Local

Na abordagem local de documentos *offline*, diversas características podem ser extraídas de um documento manuscrito, entretanto duas delas são mais conhecidas: as relativas a palavras e as relacionadas aos caracteres.

No caso das ***características extraídas das palavras,*** no trabalho de Luna et al. (2011), descrito na Seção anterior, um conjunto de características locais também é extraído das palavras, são elas: proporção da zona média das palavras comparada com as zonas ascendentes e descendentes e inclinação das palavras. Tais características são consideradas importantes do ponto de vista do perito grafoscópico e estão presentes na base de características do método proposto neste trabalho.

Zois e Anastassopoulos (2000) apresentam uma abordagem para o processo de identificação de autoria na qual o vetor de características de um escritor é derivado utilizando operadores morfológicos para obter o perfil horizontal das palavras (funções de projeções). As projeções são derivadas e processadas em segmentos de forma a aumentar a eficiência discriminatória do vetor de características. Para validar o método proposto no trabalho foram realizados experimentos tanto com classificador bayesiano quanto com redes neurais. As taxas de acerto chegaram próximas a 95%, em experimentos com 50 escritores usando palavras escritas em Inglês e em Grego.

Jain e Doermann (2011) propuseram um método para identificação *offline* de escritores usando a característica *K* segmentos adjacentes (KAS – *K-Adjacent Segments*). Esta abordagem apresentou taxas de acerto de 93% na base de dados IAM (*Informatik und Angewandte Mathematik)* (base de dados na língua Inglesa). Resultados obtidos de testes subsequentes demonstraram que as taxas de identificação melhoram quando o número de amostras de treinamento aumenta, e adicionalmente que o uso das características utilizadas pode ser estendido para a língua Árabe. K segmentos

adjacentes foi introduzido por Ferrari et al. (2008) como uma característica que representa o relacionamento entre conjuntos de bordas vizinhas em uma imagem por meio de detecção de objetos.

Finalmente, no caso das *características extraídas dos caracteres***,** Blankers et al. (2007) propõem um método para identificação de autoria por meio da extração e análise de características alográficas como loops e entrada (inicio) do traçado dos caracteres. Características sub-alográficas são computados a partir de partes das letras. Tais características são particularmente importantes para a identificação forense. Características relativas aos loops são encontradas em loops ascendentes que aparecem em letras como “l”, “k” e “b”, e descendentes que aparecem em letras como “g” e “j”. Características de entrada do traçado representam a primeira parte do caractere e são encontradas em quase todas as letras do alfabeto, especialmente na escrita manual cursiva. Foi usado neste trabalho o classificador KNN e obtida taxa de acerto próxima a 98% em uma base com 41 escritores.

Pervouchine e Leedham (2007) apresentaram um estudo sobre o uso de características extraídas de três caracteres, são eles: “d”, “y” e “f” e o grafema “th” como primitivas básicas no processo de identificação de autoria. De acordo os autores deste trabalho, a escolha por um conjunto previamente definido de caracteres e grafemas se justifica pelo fato de que é impossível considerar todos os caracteres e grafemas que ocorrem nos documentos manuscritos. Como classificador foram utilizadas Redes Neurais e a técnica de Algoritmo Genético foi utilizado para encontrar o conjunto de características ótimo (uma vez que um grande conjunto de primitivas foi extraído). É relevante destacar que nas conclusões do trabalho os autores reforçam a importância dos grafemas no processo de identificação de autoria, o que reforça a ideia de que o formato de um caractere é afetado pelos caracteres adjacentes a ele. As melhores taxas de classificação atingiram 58% em uma amostra com 165 escritores.

Bui et al. (2011) propõem um framework para identificação de autoria baseada em duas abordagens locais: caracteres e grafemas. Experimentos foram conduzidos usando o algoritmo *K-Means* e taxas de acerto próximas a 100% foram atingidas em uma base com 32 documentos escritos por 16 diferentes escritores.

2.3.5. Resumo das Abordagens para Identificação de Autoria *offline*

Com base no levantamento bibliográfico apresentado ao longo da Seção 2.3.4, a Tabela 2.3 apresenta um resumo do estado da arte considerando algumas das principais e mais atuais abordagens para identificação de autoria *offline*. Este classificação levou em consideração os seguintes aspectos: categoria (nível de granularidade da característica), características extraídas, número de escritores participantes dos experimentos, algoritmo de classificação utilizado, melhores taxas de acerto e se o trabalho utiliza características grafométricas. Este último aspecto foi avaliado uma vez que o presente trabalho apresenta uma abordagem para identificação de autoria centrada em características grafométricas da escrita e, para tanto, é importante destacar quais pesquisas utilizam esta mesma abordagem.

Tabela 2.3. Resumo estado da arte das abordagens de identificação de autoria *offline*

Categoria	Trabalho	Característica Extraída	Tamanho da Amostra	Classificador	Taxa de Acerto	Característica Grafométrica
Documento	Said et al. (1998)	Textura - Filtros de Gabor e Matrizes de Co-ocorrência	10 escritores	WED (Weigthed Euclidean Distance)	96%	Não
	Said et al. (2000)	Textura - Filtros de Gabor	40 escritores	WED	96%	Não
	Bulacu et al. (2007)	Textura	250 escritores	KNN	82%	Não
	He et al. (2008)	Textura	500 escritores	HMT	32%	Não
	Helli e Morghaddam (2010)	Textura - Filtros de Gabor e XGabor	100 escritores	Método Fuzzy	100%	Não
	Karunakara e Mallikarjunaswamy (2011)	Textura	50 escritores	KNN	94%	Não
Parágrafo	Schomaker et al. (2007)	Geração de codebooks	250 escritores	Redes de Kohonen	97%	Não
	Siddiqi e Vicent (2008)	Geração de codebooks	650 escritores -IAM 375 escritores – RIMES	WED	84% - IAM 74% - RIMES	Não
	Kumar et al. (2011)	Geração de codebooks	732 imagens (fase treinamento) 625 imagens (fase de teste).	SVM	98%	Não
Linha	Hertel e Bunke (2003)	Continuidade do traçado, regiões fechadas, contornos superiores e inferiores.	50 escritores	KNN	90%	Sim
	Schlapbach e Bunke (2004)	Inclinação, altura e obliqüidade das linhas de texto.	50 escritores	HMM	94,4%	Sim
	Bensefia et al. (2005)	Grafemas	88 escritores - PSI 150 escritores - IAM	VSM (*Vector Space Model*)	96% - PSI 86% - IAM	Não
	Chen et al. (2010)	Informações relativas ao contorno de segmentos adjacentes e remoção de linhas de referência pré-impressas.	60 escritores	SVM	54,9%	Sim
	Luna et al. (2011)	Espaço percentual da margem esquerda e margem direita, separação entre linhas, direção	30 escritores	Algoritmo ALVOT	88%	Sim

		geral da escrita, e espaço entre palavras.				
Palavras	Luna et al. (2011)	Proporção da zona média das palavras comparada com as zonas ascendentes e descendentes e inclinação das palavras.	30 escritores	Algoritmo ALVOT	88%	Sim
	Zois e Anastassopoulos (2000)	Uso de operadores morfológicos para obter o perfil horizontal das palavras.	50 escritores	Bayesiano e Redes Neurais	95%	Sim
	Jain e Doermann (2011)	K-segmentos adjacentes (KAS)	650 escritores	KNN	93%	Não
Caracteres	Blankers et al. (2007)	Características alográficas	41 escritores	KNN	98%	Não
	Pervouchine e Leedham (2007)	Uso de características extraídas de três caracteres, são eles: “d”, “y” e “f” e o grafema “th”.	165 escritores	Algoritmo DistAl – baseado no Perceptron	58%	Sim
	Bui et al. (2011)	Protótipos de caracteres e grafemas.	16 escritores	K-Means	100%	Não

2.3.6. Sistemas para Identificação de Autoria

Para oferecer uma visão mais completa do que a pesquisa vem proporcionando em situações reais aos peritos, serão brevemente apresentados sistemas para o processo de identificação de autoria em documentos manuscritos. São descritos quatro sistemas de identificação de autoria: FISH (*Forensic Information System for Handwriting*); WANDA (*Forensic Information System Handwriting*); CEDAR-FOX (*Center of Excellence for Document Analysis and Recognition*) e o *Write On*.

FISH é um Sistema de Informação para Análise Forense de Documentos Manuscritos (PHILLIP, 1996) baseado em missão crítica que permite que um perito digitalize, meça e armazene documentos de autores questionados e conhecidos com o propósito de identificar um escritor em uma base de escritores previamente armazenados.

Como resultado para o usuário final, o sistema FISH apresenta medições de características das letras, tais como altura, largura, laços/laçadas (*loops*) e distâncias. Também, uma lista de prováveis candidatos a autoria, bem como imagens digitalizadas de seus documentos. Segundo Van Erp et al. (2003), o sistema FISH é um sistema de identificação e análise de documentos manuscritos que capacita o usuário, um perito no exame de documentos questionados, a medir características do documento manuscrito por meio de um processo computacional e comparar estas características extraídas com a de outros documentos em um banco de dados.

O segundo sistema, conhecido como **WANDA,** é na verdade uma evolução do sistema FISH, fazendo alusão ao filme "Um peixe chamdo Wanda". Este sistema permite a análise de manuscritos e identificação de escritores e foi desenvolvido por Lambert Schomaker e sua equipe do Departamento de Inteligência Artificial da University of Groningen-Netherlands. De acordo com Van Erp et al. (2003), o sistema WANDA é uma ferramenta *desktop* que oferece suporte para um processo completo de extração de características encontradas em documentos manuscritos. Esta ferramenta possui uma série de características, tais como: utilização de tecnologias como *plug-ins*, XML (*Extensible Markup Language)* e ambiente cliente-servidor, que fazem com que este seja um sistema fácil de manter, portável e altamente adaptável.

Ainda de acordo com Van Erp et al. (2003), embora o FISH seja um sistema excelente com relação aos resultados obtidos para identificação de autoria de manuscritos, sua interface com o usuário possui algumas limitações. Além disso, uma

atualização em termos de tecnologia também foi necessária, para melhorar sua portabilidade e facilidade de manutenção.

De acordo com Franke et al. (2004), o sistema WANDA é um *framework* que oferece suporte para o processamento de amostras de documentos manuscritos no contexto de exames forenses de manuscritos e identificação de autoria. Este *framework* oferece suporte às seguintes funcionalidades: pré-processamento, anotação em documentos manuscritos, medição de características selecionadas, extração automática de primitivas e busca em um conjunto de dados.

Segundo Van Erp et al. (2003), a ferramenta WANDA possui um *plug-in* chamado WAN que permite que um usuário realize interativamente a medição de características selecionadas de documentos manuscritos. Este *plug-in* oferece suporte a 10 (dez) diferentes medidas. As medidas básicas consistem de informações extraídas da altura dos caracteres (ascendentes, descendentes, altura do corpo, e altura dos caracteres ovais), a inclinação dos caracteres, e a largura dos caracteres. Se presente no documento manuscrito, o WAN permite que o usuário extraia medidas de *loops* inferiores e superior de caracteres. Como única característica não baseada em caracteres, o WAN também permite a medição da distância média entre as linhas base em uma amostra do documento.

O terceiro sistema denominado **CEDAR-FOX**, segundo Srihari e Shi (2004), é um sistema completo para o gerenciamento de documentos manuscritos no contexto das Ciências Forenses, desenvolvido com o objetivo de oferecer suporte para análise, identificação, verificação e recuperação de documentos manuscritos. Ainda de acordo com estes autores, este sistema é focado na verificação de autoria de manuscritos.

Como um sistema de gerenciamento de documentos para análise forense, ele fornece três funcionalidades principais: pode ser utilizado como um sistema de análise de documentos; pode ser usado para a criação de uma biblioteca digital; e, pode ser usado como um sistema de gerenciamento de uma base de dados para recuperação de documentos e identificação de autoria.

Como um sistema de análise de documentos, o CEDAR-FOX fornece uma interface gráfica ao usuário final. Este pode carregar ou digitalizar a imagens do documento. O sistema irá então automaticamente extrair características baseando-se em um intensivo processo de processamento e reconhecimento de imagens (SRIHARI; SHI, 2004).

Esta ferramenta também oferece um suporte completo para o gerenciamento de uma biblioteca de documentos manuscritos. Para a criação da biblioteca, o sistema tem as seguintes funcionalidades: entrada de metadados do documento, por exemplo, numero de identificação do escritor e outras informações correlacionadas; criação de uma transcrição textual do conteúdo da imagem em nível de palavra; e extração automática de características em nível do documento tais como: largura do traço, inclinação, espaçamentos entre palavras; além de características mais especificas/locais que capturam características estruturais dos caracteres e palavras (SRIHARI; SHI, 2004).

De acordo com Srihari e Shi (2004), o CEDAR-FOX fornece suporte para recuperação de documentos nas seguintes modalidades de consulta: a imagem completa é o elemento questionado; uma imagem parcial (uma região de interesse do documento); a imagem de uma palavra; e uma palavra-chave. O usuário pode digitar palavras-chave em intervalos de palavras que vão desde palavras nos documentos, número do processo, nomes de pessoas, tempo, e palavras-chave pré-registradas tais como breves descrições de um caso.

O último sistema denominado **Write On**, segundo Pikaso Software Inc. (2000), é uma ferramenta de auxilio ao perito em exames forenses de documentos, pois oferece suporte estatístico para a análise forense de documentos. O software oferece um método eficiente e confiável para avaliar variação natural, encontrando todas as variações possíveis entre documentos questionados e documentos modelo. Ainda segundo Pikaso Software Inc. (2000), um índice de palavras serve como um ponto de partida na decisão sobre quais palavras ou combinações de caracteres devem ser procurados e analisados. Gráficos de ocorrência e mapas de pesquisa são gerados a partir destas buscas.

Os gráficos de ocorrência permitem uma fácil comparação passo a passo de todas as variações e os mapas destacam os locais de cada ocorrência no documento em análise. Quando cada pesquisa é finalizada uma tabela estatística é atualizada para refletir o histórico de cada pesquisa realizada e as respectivas ocorrências destas pesquisas nos documentos questionados e nos documentos modelo.

2.3.7. Considerações Finais sobre Autoria em Documentos Manuscritos

Na Seção 2.3 foram apresentados trabalhos relacionados a esta Tese de Doutorado. Pode-se observar que muitos trabalhos que apresentam boas taxas de acerto concentram-se em características texturais do documento manuscrito, bem como na

geração de *codebooks*. Embora muito relevantes estes trabalhos não apresentam processos de extração de características alinhados com as abordagens adotadas pelos peritos durante as análises forenses de documentos questionados.

Outra questão que deve ser destacada é que tais abordagens dependem de um alto poder computacional, uma vez que as mesmas envolvem a aplicações de filtros sobre as imagens dos manuscritos.

No presente trabalho as características extraídas dos manuscritos estão alinhadas com aquelas utilizadas pelos peritos, ou seja, envolvem aspectos grafométricos. Normalmente, trabalhos nesta linha apresentam taxas de acerto menores (como pode ser observado na Tabela 2.3), embora, como pode também ser observado no Capítulo 3 deste trabalho, foram atingidas taxas de acerto comparáveis àquelas obtidas por outros trabalhos apresentados na literatura.

2.4. Problema e Solução Proposta

Tendo como base a revisão apresentada ao longo deste capítulo e os desafios destacados no Capítulo 1 - Introdução, este trabalho de Doutorado propõe um método, baseado em características grafométricas e suportado computacionalmente, para o problema de identificação de autoria *offline.*

As características utilizadas no método proposto, apresentadas em destaque na Figura 2.4 e contextualizadas na classificação proposta por Seeraj e Idicula (2011), exigem uma manipulação direta, ou seja, pixel a pixel, da imagem do documento. Isto traz um alto nível de complexidade para o processo de extração, bem como distorções nos resultados. Para corrigir tais problemas, medidas corretivas, análises individuais das características, bem como, abordagens de seleção de primitivas foram realizadas de forma a obter a combinação de características que levasse às melhores taxas de acerto.

O conjunto de características definido neste trabalho é composto por doze características, no entanto, pode-se observar que quatro grandes grupos de características grafométricas estão sendo utilizados, são eles:

- **Hábitos do uso do espaço gráfico:** estas características são muito utilizadas pelos peritos, uma vez que alguns escritores podem fazer um bom uso da folha de papel, escrevendo até seus limites físicos, enquanto que outros podem deixar espaços em branco, usualmente regulares em todas as linhas ou margens.

Diferentes escritores iniciam e terminam suas escritas em diferentes posições. Assim, localizações tais como identação de sentenças, espaçamentos das margens, uso de espaços, pontos iniciais e finais são exemplos de "uso do espaço gráfico". Este grupo é formado pelas seguintes características: número de linhas e distância das margens (superior, inferior, direita e esquerda);

- **Tamanho das palavras:** outra importante característica utilizada pelos peritos é a definição do tamanho padrão das palavras. Dessa forma, a primeira palavra de cada linha foi "destacada" e sua altura e proporção de pixels pretos, em relação ao fundo, foram calculadas. Este grupo é formado pelas seguintes características: quantidade de pixels pretos da primeira palavra de cada linha e altura da primeira palavra de cada linha;
- **Inclinação axial:** esta característica grafométrica, também essencialmente utilizadas pelos peritos durantes suas análises, representa o ângulo geral de escrita e tem um importante poder discriminatório no processo de identificação de autoria (como pode ser observado nos resultados apresentados no Capítulo 3). Este grupo é formado pelo ângulo que a escrita apresenta em relação a uma linha horizontal de base (no caso das cartas forenses, esta linha é imaginária, visto que os escritores utilizam folhas não pautadas);
- **Hábitos de traçado de laços ascendentes e descendentes:** estas características, também observadas pelos peritos em suas análises, são extraídas das palavras e caracteres. Representam informações relativas ao tamanho e posição dos laços ascendentes e descendentes de caracteres tais como: "l", "t", "p", "g", h" e "f". Este grupo é formado pelas seguintes características: altura, largura, número de pixels e inclinação axial dos laços.

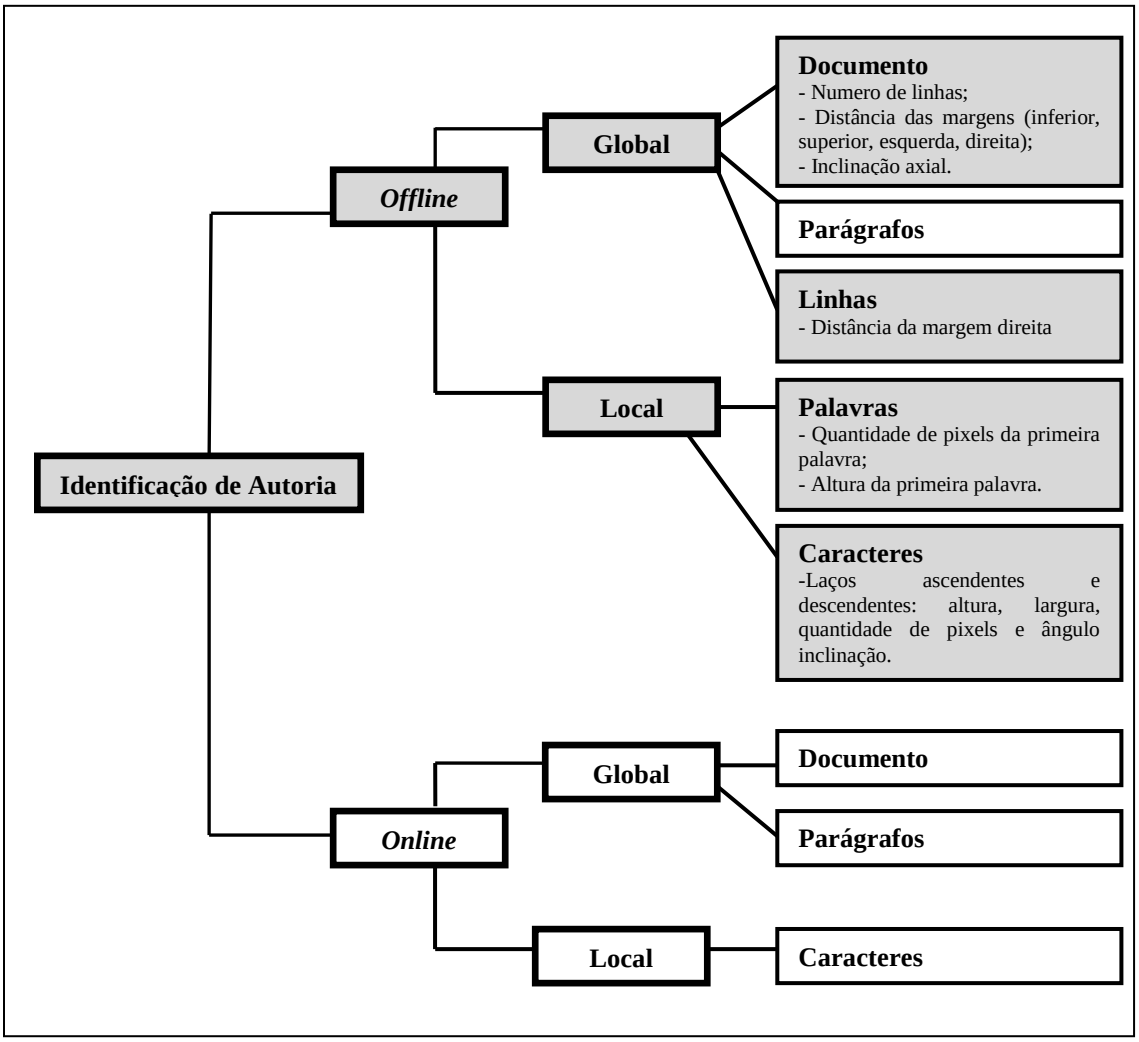


Figura 2.5. Esquema de abordagem para identificação de autoria adotado no método proposto

2.5. Considerações Finais

Este capítulo apresentou uma revisão bibliográfica com foco em dois assuntos principais, a escrita humana e os métodos automatizados para identificação de autoria. Com relação à escrita humana, foram estudados os aspectos que afetam a escrita de uma pessoa e que devem ser considerados por um perito durante o processo de identificação de autoria. Estes aspectos foram fundamentais para a definição e implementação do conjunto de características utilizado no método proposto. Com relação aos métodos automatizados para identificação da escrita, o estudo de tais métodos foi fundamental para se observar o que já foi produzido nesta área, bem como, os avanços que ainda podem ser obtidos. Finalmente, na última Seção, com base no cenário apresentado ao

longo deste Capítulo, foi introduzido o problema a ser resolvido nesta Tese bem como sua proposta de solução.

O próximo Capítulo apresenta detalhadamente o método proposto para o desenvolvimento da solução pretendida, incluindo descrição completa de todas as etapas do método proposto, bem como, do protocolo de experimentos e os resultados obtidos.

Capítulo 3

Método Proposto

3.1. Considerações Iniciais

Neste capítulo são apresentados os elementos necessários para o desenvolvimento deste trabalho. Neste contexto, na Seção 3.2 são descritas e discutidas a base de cartas forenses, uma vez que para a validação do método adotado neste trabalho são necessárias amostras de manuscritos definidas em um formato padrão. Na Seção 3.3 é apresentado detalhadamente o método proposto, destacando e descrevendo cada uma das etapas que o compõem. A Seção 3.4 aborda o protocolo adotado para a condução dos experimentos de validação do método proposto e os resultados obtidos com a realização destes experimentos.

3.2. Bases de Dados

Para a validação de qualquer abordagem que se proponha a auxiliar o processo de análise de documentos manuscritos, é necessário uma base com amostras de manuscritos. Com o objetivo de garantir que as amostras de manuscritos obtidas de um escritor possuam todas as letras do alfabeto (tanto com letras maiúsculas, quanto com letras minúsculas) bem como os numerais, acentos, símbolos especiais e símbolos de pontuação, formatos padrão de cartas forenses foram propostos na literatura (FREITAS et al., 2008) (MORRIS, 2000).

Dessa forma, nesta Seção é apresentada uma discussão sobre as principais bases de cartas forenses internacionais e sobre a base de cartas forenses PUCPR, utilizada nos experimentos realizados neste trabalho. Também é apresentada uma discussão sobre as diferenças entre bases de cartas que se baseiam em princípios forenses e aquelas que não se baseiam em princípios forenses.

3.2.1 Bases de Cartas Forenses Internacionais

Baseadas em Princípios Forenses

De acordo com Osborne (1929), o principal objetivo de uma carta forense é reproduzir a associação entre diferentes, letras, palavras, numerais e símbolos; sendo que as letras devem ser adaptadas à linguagem local. Nesse contexto, as Figuras 3.1 a 3.4 mostram alguns formatos de cartas forenses dedicados a colher o estilo de escrita manual, a individualidade e características estáticas e dinâmicas da escrita de um escritor. As Figuras 3.1 a 3.4 exemplificam os textos das cartas forenses de Londres, do Egito, de Idaho e do CEDAR.

Our London business is good, but Vienna and Berlin are quiet Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott Street and then goes to Turin and Rome and will join Colonel Parry and arrive at Athens, Greece, November 27th or December 2nd. Letters there should be addressed King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the 'Y. X.' Express tonight.

Figura 3.1. Carta Forense de Londres
[Fonte: (OSBORNE,1929)]

Dear Sam: From Egypt we went to Italy, and then took a trip to Germany, Holland and England. We enjoyed it all but Rome and London most. In Berlin we met Mr. John O. Young of Messrs. Tackico & Co., on his way to Vienna. His address there is 147 upper Zeiss Street, care of Dr. Quincy W. Long. Friday the 18th, we join C. N. Dazet, Esquire and Mrs. Dazet, and leave at 6:30 A.M. for Paris on the 'Q. X.' Express and early on the morning on the 25th of June start for home on the S. S. King. Very sincerely yours,

Figura 3.2. Carta Forense do Egito
[Fonte: (OSBORNE,1929)]

Dear Zach, Well, the old class of "16" is through at last. You ask where the boys are to be. Val Brown goes on the 24th to Harvard for law. Don't forget to address him as "Esquire." Ted Updyke takes a position with the N. Y. W. H.	From Nov 10, 1999 Jim Elder 829 Loop Street, Apt 300 Allentown, New York 14707 To Dr. Bob Grant 602 Queensberry Parkway

& H. R. R., 892 Ladd Ave., Fall River, Massachusetts, and Jack McQuade with the D. L. & W. at Jersey City, N. J. 400 E. 6th Street William Fellows just left for a department position in Washington; his address is 735 South G. St. At last account, Dr. Max King was to go to John Hopkins for a Ph.D. degree. Think of that! Elliott goes to Xenia, Ohio, to be a Y. M. C. A. secretary. I stay here for the present. What do you do next? How about Idaho? Yours truly, and goodbye.	Omar, 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. Organizing such an event is no picnic, and as President of the 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

Figura 3.3. Carta Forense de Idaho
[Fonte: (OSBORNE,1929)]

Figura 3.4. Carta Forense de CEDAR
[Fonte: (OSBORNE,1929)]

Segundo Morris (2000) existem diferentes formatos de cartas, e nenhum é melhor ou pior do que outro. Cada qual possui seu propósito específico e geralmente é projetado por um perito em análise de documentos manuscritos que acredita que tal formato lhe fornecerá as características de escrita do escritor que são mais úteis para conduzir suas análises.

Não Baseadas em Princípios Forenses

Diferentes bases de manuscritos vêm sendo apresentadas na literatura, tais como: IAM e RIMES com o objetivo de oferecer suporte para o processo de análise e reconhecimento de manuscritos. Estas bases foram desenvolvidas por pesquisadores, visando treinar e testar sistemas de reconhecimento de escrita manuscrita.

A base IAM (ZIMMERMANN; BUNKE, 2002) é composta de formulários com textos manuscritos de conteúdos variados escritos na língua inglesa. Um total de 650 diferentes escritores contribuiu para a construção desta base de dados, na qual 350 escritores escreveram em apenas uma página, 300 escritores escreveram em pelo menos duas páginas e 125 escritores escreveram em pelo menos quatro páginas.

A outra base de manuscritos, a base RIMES (GROSICKI et al., 2008) é uma base de dados que compreende cartas manuscritas na língua francesa compostas por textos enviados por indivíduos a empresas. Mais do que 1300 escritores contribuíram para a composição desta base de dados escrevendo até cinco cartas (totalizando 5600 cartas).

Estas bases são consolidadas e muito utilizadas em pesquisas na área de identificação de autoria, no entanto os textos definidos para a construção das mesmas não tiveram por fundamento os princípios forenses, ou seja, não levaram em consideração aspectos que são utilizados pelos peritos durante a realização de suas análises periciais.

Além disso, estas bases não utilizam um formato padrão único de texto, uma vez que as cartas presentes nestas bases possuem conteúdos variados. No contexto deste trabalho, em função do conjunto de características ter por base aspectos grafométricos, é fundamental uma base, tal como a base forense no modelo PUCPR (descrita na Seção 3.2.2), que tenha sido construída levando em consideração os princípios utilizados pelos peritos em grafoscopia.

3.2.2. Base de Cartas Forenses Modelo PUCPR

Segundo Freitas et al. (2008), a maioria dos trabalhos relacionados a modelos de documentos manuscritos encontrados na literatura são dedicados a língua inglesa e, dessa forma, eles não contemplam certas particularidades da língua portuguesa. Nesse contexto, no trabalho de Freitas et al. (2008), é apresentada a Base de Cartas Forenses Brasileira, ou Cartas Forenses PUCPR, que foi criada com o objetivo de suportar várias particularidades da língua portuguesa, tais como acentos (á, à, ã, ê, ü) e o símbolo especial (ç). Além da definição do modelo de carta (Figura 3.5), esforços foram conduzidos com o intuito de expandir a base de cartas modelo PUCPR.

A base de cartas forenses PUCPR possui atualmente 600 escritores, com 03 cartas por escritor, escritas em folha A4 não pautada. As cartas colhidas e então

digitalizadas em 300 dpi, com 256 escalas de cinzas. Uma versão binária da base de dados se encontra disponível, como apresentada no exemplo da Figura 3.6.

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

Figura 3.5. Carta Forense PUCPR

[Fonte: (FREITAS et al. ,2008)]

De acordo com Freitas et al. (2008), a carta forense PUCPR é concisa (131 palavras) e completa no sentido que contém todos os caracteres (letras e numerais) e certas combinações de caracteres interessantes. A Tabela 3.1 apresenta o número de ocorrências em cada posição de interesse no texto.

Tabela 3.1. Frequência posicional de ocorrências das letras. Fonte:

[(FREITAS et al., 2008)]

Maiúscula	Inicio	Minúscula	Inicio	Meio	Fim
A	3	**a**	12	55	20
B	1	**b**	1	4	1
C	4	**c**	5	17	---[5]
D	2	**d**	18	15	---
E	2	**e**	7	41	21

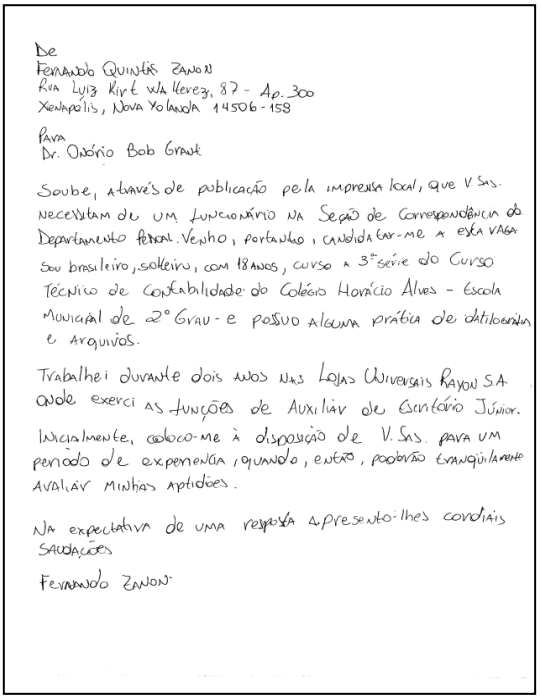
[5] Denota que esta letra não pode ser encontrada em uma posição específica na língua Portuguesa Brasileira.

F	2	**f**	2	1	---
G	2	**g**	0[6]	4	---
H	1	**h**	0	4	---
I	1	**i**	1	48	2
J	1	**j**	0	1	---
K	1	**k**	---	---	---
L	2	**l**	2	19	3
M	1	**m**	3	6	4
N	2	**n**	3	39	3
O	1	**o**	1	44	26
P	2	**p**	8	11	1
Q	1	**q**	1	2	1
R	2	**r**	1	38	5
S	6	**s**	3	20	20
T	2	**t**	1	24	2
U	1	**u**	3	18	2
V	4	**v**	1	7	---
W	1	**w**	---	---	---
X	1	**x**	0	4	0
Y	1	**y**	---	---	---
Z	2	**Z**	0	0	2

Além disso, a carta PUCPR também contém pontuações (“.”, “,”), dez classes de números, o símbolo especial (ç), e diferentes combinações tais como “nh”, “lh”, “qu”, e “00”.

Segundo Koppenhaver (2007), estas combinações são elementos gramaticais muito importantes que permitem estudos sobre a individualidade da escrita, uma vez que poucos “falsários” se preocupam em simular cuidadosamente as letras minúsculas. Normalmente, os “falsários” gastam muita energia duplicando letras maiúsculas acreditando que se eles reproduzem estas letras corretamente, as suas falsificações serão consideradas como genuínas.

[6] Denota que esta letra pode ser encontrada em uma específica posição na língua Portuguesa Brasileira, mas não existem palavras na carta forense PUCPR que incluem esta letra em alguma posição específica.



De
Fernando Quintas Zanon
Rua Luiz Kirt Walkerez, 87 - Ap. 300
Xenapolis, Nova Yolanda 14506-158

Para
Dr. Osório Bob Grant

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

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

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

Na expectativa de uma resposta apresento-lhes cordiais saudações

Fernando Zanon.

Figura 3.6. Carta PUCPR CF00001_01

3.2.3. Considerações Finais sobre Bases de Dados

A Seção 3.2 apresentou e caracterizou diferentes bases de dados de manuscritos que vem sendo utilizadas nos trabalhos que envolvem identificação e verificação de autoria. A definição e construção destas bases são de grande importância para trabalhos dessa natureza, assim como para esta pesquisa, uma vez que a "comprovação" da eficiência de um método de identificação/verificação depende da realização de experimentos com um grande número de documentos já digitalizados e em um formato padrão, ou seja, de documentos pertencentes a uma base formal de dados.

No contexto deste trabalho de Doutorado foi importante distinguir entre bases que se baseiam em princípios forenses daquelas que não utilizam tais princípios para sua concepção, pois o método proposto se baseia em características grafométricas, ou seja, necessita de uma base que levou em conta, durante sua definição, aspectos que são utilizados pelos peritos durante suas análises periciais. Dessa forma, a base PUCPR é utilizada como fonte de documentos manuscritos nesta pesquisa.

Na próxima Seção é apresentado o método proposto para o processo de identificação de autoria adotado neste trabalho. Deve-se ressaltar que este método foi validado por meio de experimentos que utilizaram manuscritos da base PUCPR.

3.3. Método de Identificação de Autoria utilizando Características Grafométricas

O método proposto neste trabalho é uma abordagem para identificação de autoria baseada em princípios grafométricos. Para tanto, foi definido, com base em trabalhos descritos na literatura, um método para automatização de tal processo.

Este método, como apresentado previamente na Figura 2.4, consiste em uma abordagem *offline* que envolve a extração de características Globais (em nível de documento e linha) e Locais (em nível de palavra e caractere). Dessa forma, nas próximas Seções é apresentada uma visão geral do método proposto, bem como, uma descrição detalhada sobre cada uma das etapas que o constituem.

3.3.1. Visão Geral

A Figura 3.7 apresenta uma visão geral do método proposto neste trabalho. Pode-observar que o método proposto segue o processo padrão adotado por abordagens de aprendizagem de máquina para aplicações que envolvam o reconhecimento de padrões. Neste trabalho o foco da aprendizagem de máquina é a classificação de "objetos" (cartas) dentro de um número de categorias ou classes.

As cartas (tanto àquelas dos escritores conhecidos, quanto àquela de autoria questionada) são submetidas aos processos de pré-processamento e extração de características, para que seus modelos possam ser criados e subsequentemente a autoria atribuída. Ao final da etapa de classificação um *ranking* com os prováveis autores do documento questionado é construído. O número de escritores incluídos neste *ranking* depende do número de escritores utilizados na base de treinamento. Nos experimentos (descritos em detalhes na Seção 3.4), foram utilizados, incrementados gradualmente (em intervalos de 20), 200 diferentes escritores para o treinamento (com 400 diferentes cartas – 2 por escritor) e 200 diferentes escritores para o teste (com 200 diferentes cartas – 1 por escritor). Assim, para cada documento questionado, pode-se produzir uma lista com o autor mais provável (lista com o escritor com a melhor posição no *ranking*). Quando o escritor correto, ou seja, o autor do documento questionado se encontra nesta lista a taxa de acerto é acrescida.

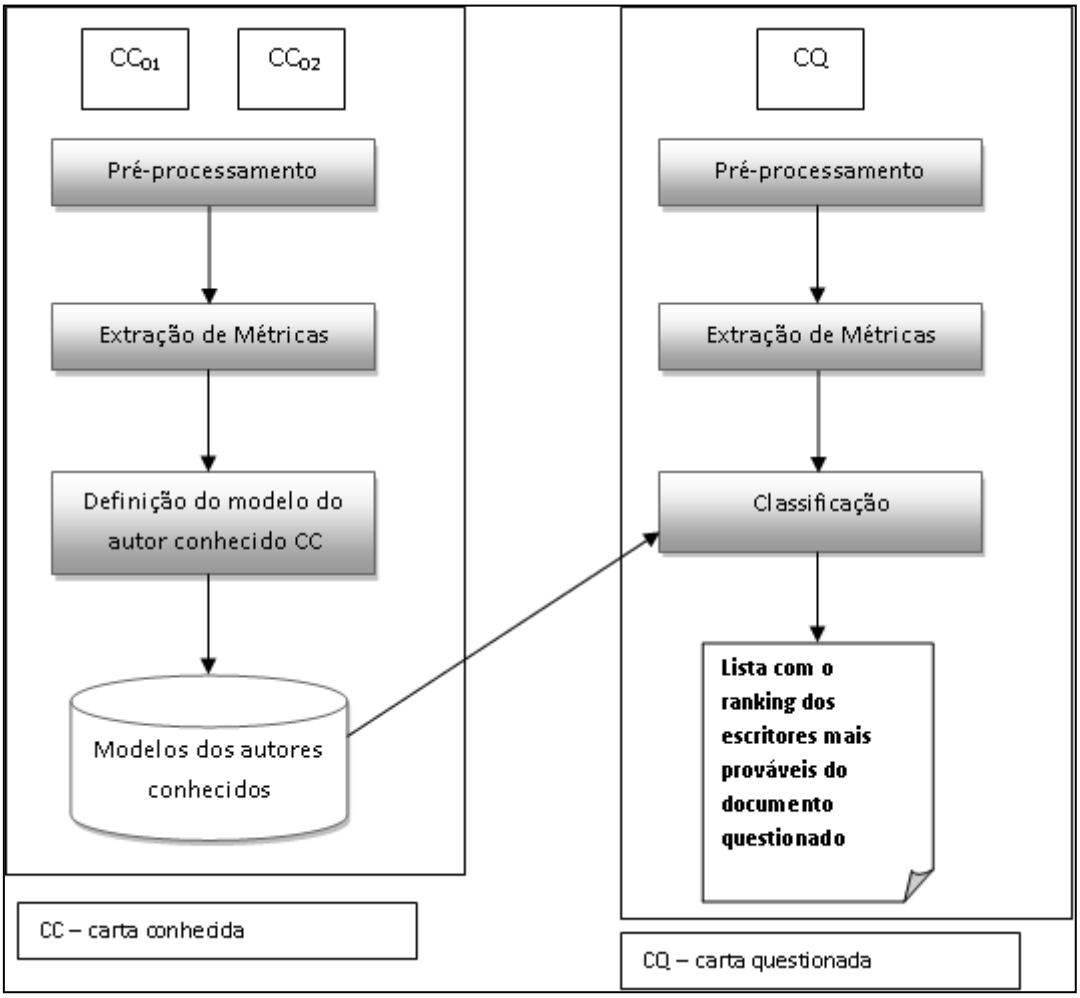


Figura 3.7. Visão geral do método proposto

3.3.2. Pré-processamento

A atividade de pré-processamento consiste em operações a serem realizadas na imagem a fim de possibilitar a extração das métricas para obtenção dos resultados finais. Esta atividade é dividida nas etapas descritas a seguir.

I. Binarização

O processo de binarização consiste em transformar as imagens que se encontram em tons de cinza (256 níveis de cinza) para uma imagem binária (preto e branco). Este processo além de reduzir o tamanho do arquivo, facilita o processamento computacional da imagem. Esta binarização foi realizada de duas diferentes formas:

- para a extração das características f_1 a f_7 (descritas na próxima Seção) foi utilizado o algoritmo de binarização global OTSU (OTSU, 1979);

- em função da característica f_8 (descrita na próxima Seção) necessitar da conservação de detalhes importantes no traçado original da imagem, utilizou-se para a extração desta característica o algoritmo de binarização local por Entropia de Abutaleb (ABUTALEB, 1989), pois esta técnica aplica limiares de acordo com a área da imagem.
- A Figura 3.9 apresenta o resultado da etapa de binarização da carta modelo apresentada na Figura 3.8.

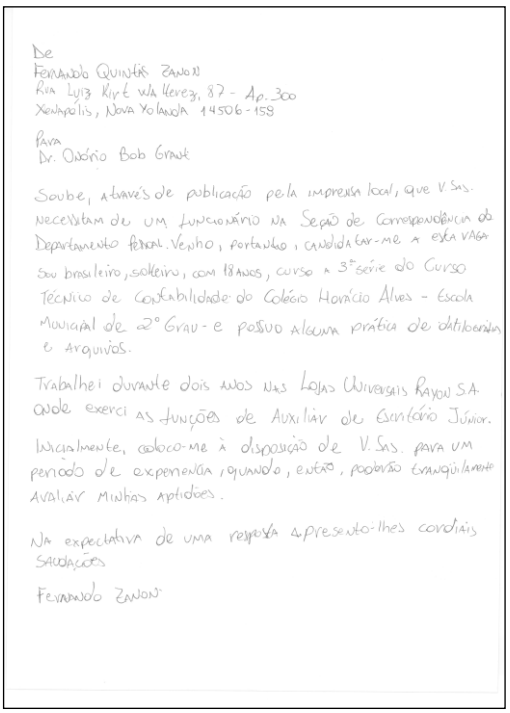


Figura 3.8. Exemplo de carta PUCPR original

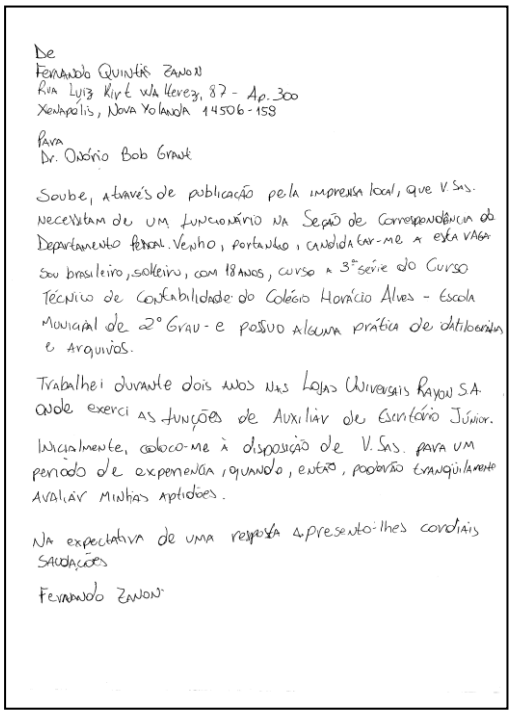


Figura 3.9. Exemplo de carta PUCPR binarizada

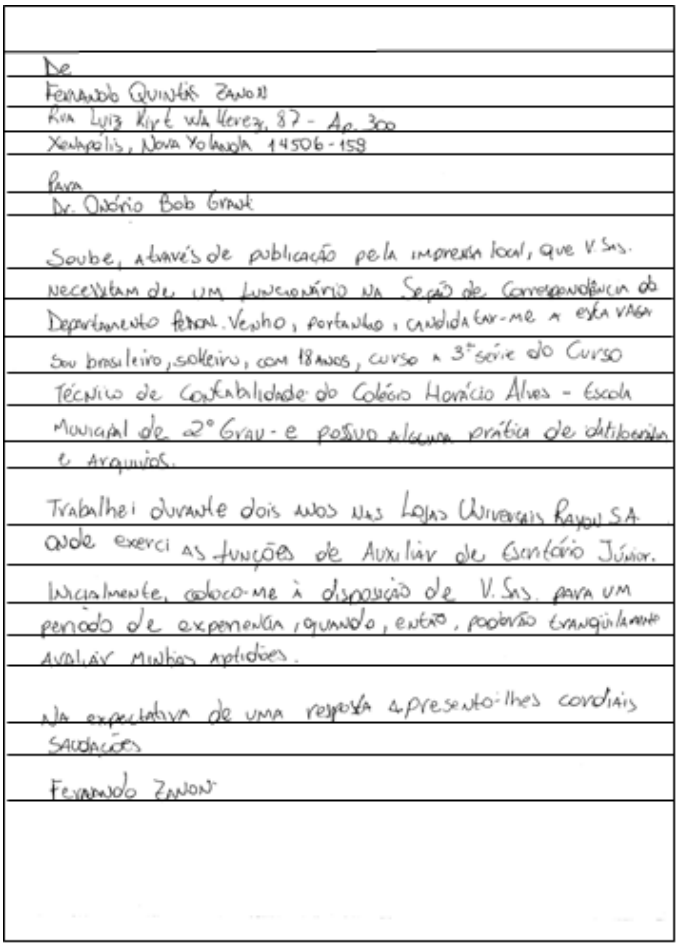
II. Segmentação

A segmentação é a etapa que tem como principal objetivo obter do documento em análise todas as informações necessárias para posteriormente proceder o cálculo das métricas. Dessa forma, são atividades de segmentação: (a) segmentação das linhas; (b) segmentação das palavras de cada linha; (c) divisão do documento em 24 fragmentos; (d) extração de contornos e bordas e (e) encontrar regiões nas palavras.

a) Segmentação de linhas

Esta etapa consiste em encontrar e segmentar as linhas do documento em análise, visto que as linhas devem ser separadas, pois são analisadas individualmente pelo método proposto. A técnica utilizada para esta tarefa é a de projeção horizontal de

pixels, ou seja, os espaçamentos verticais maiores ou iguais a média de todos os espaçamentos são considerados como um intervalo entrelinhas. Porém, sabe-se que existem problemas com escritores que possuem os comportamentos denominados *rising* (“escrever para cima”) ou *falling* (“escrever para baixo”), tal qual descrito por Freitas et al. (2008). A Figura 3.10 demonstra o resultado desta etapa.


De
Fernando Quintas Zanon
Rua Luiz Rivet Vallerez, 87 - Ap. 300
Xenopolis, Nova Yolanda 14506-158

Para
Dr. Osório Bob Grant

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

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

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

Na expectativa de uma resposta apresento-lhes cordiais saudações.

Fernando Zanon

Figura 3.10. Exemplo de carta PUCPR após o processo de separação de linhas

b) Segmentação das palavras de cada linha

Esta tarefa consiste em segmentar as palavras de cada linha para posteriormente processá-la. A técnica utilizada tem por base a determinação do histograma da projeção vertical de pixels de cada uma das linhas encontradas pela etapa anterior. A determinação do histograma de projeção vertical dos textos auxilia no processo de localização das diversas palavras que compõem o mesmo. Paralelamente, à determinação deste histograma são definidos dois tipos de espaçamentos: o entre palavras e o intra-palavras (FREITAS, 2001).

De acordo com Freitas (2001), o espaçamento entre palavras é por definição o espaço utilizado pelo escritor para separar duas palavras. E, o espaçamento intra-palavras é entendido como o espaço existente dentro de uma mesma palavra, entre dois caracteres disjuntos. O algoritmo de segmentação das palavras tem por base o cálculo médio de todos os espaçamentos horizontais existentes em uma linha do texto, obtidos pela análise do histograma de projeção vertical da imagem do documento. A Figura 3.11 demonstra o resultado desta etapa de pré-processamento.

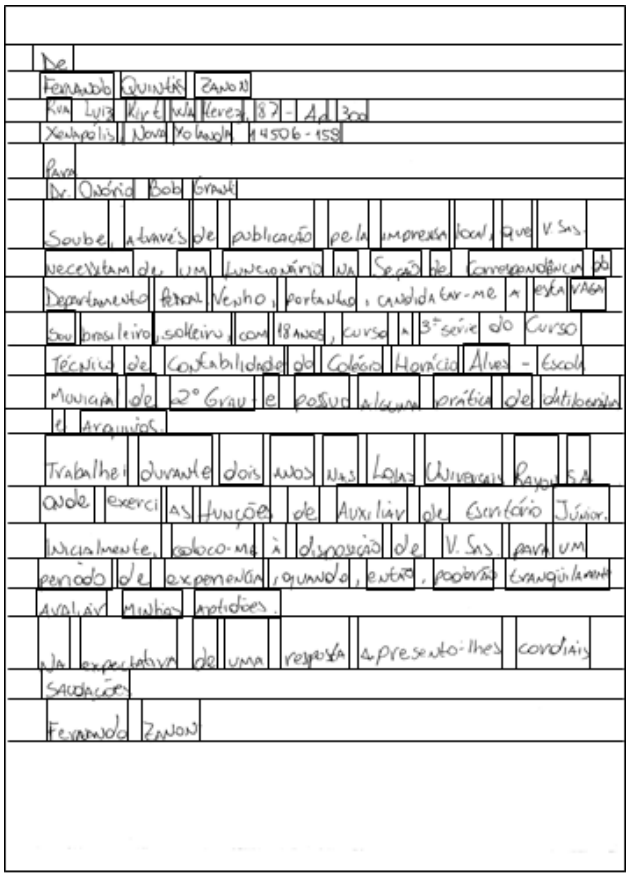


Figura 3.11. Exemplo de carta PUCPR após o processo de separação das palavras

c) Divisão do documento em 24 fragmentos

Este 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, tal qual proposto Baranoski (2005).

A escolha desta abordagem de segmentação justifica-se pelo fato de se disponibilizar um maior número de amostras do mesmo escritor, permitindo uma maior confiabilidade no processamento das características f_8, f_9, f_{10}, f_{11} *e* f_{12} de cada escritor

(uma vez que para o cálculo destas características 5 (cinco) diferentes fragmentos do mesmo escritor são usados). O resultado obtido com a divisão do documento é apresentado na Figura 3.12.

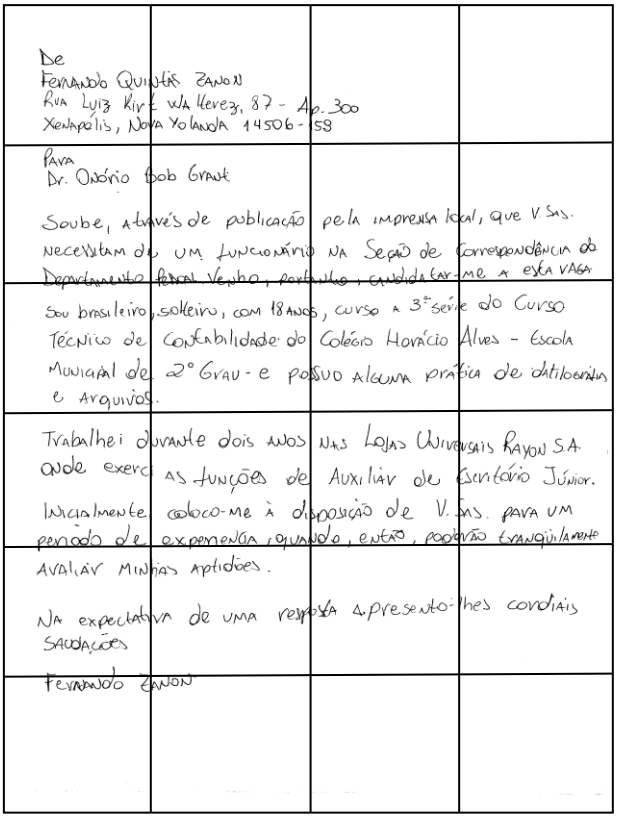


Figura 3.12. Exemplo de carta PUCPR após o processo de divisão em 24 fragmentos

d) Extração de Contornos e Bordas

Esta tarefa de pré-processamento foi realizada para a extração da característica f_8 e a abordagem adotada teve como base o trabalho de Baranoski (2005). Para tanto, foi utilizada Morfologia Matemática na detecção de bordas, aplicando-se os procedimentos de dilatação e erosão para delimitar os contornos dos traços (STENBERG, 1986).

O filtro morfológico de dilatação modifica a imagem de um manuscrito por meio de um elemento estruturante em cruz, deixando o traçado do autor mais espesso. De acordo com Facon (1996), o principal o objetivo da dilatação é aumentar o número de pixels nas bordas da imagem. Enquanto que o filtro morfológico da erosão realiza o processo inverso, deixando o traçado do escritor mais fino (por meio da redução do número de pixels nas bordas da imagem).

Neste contexto, após a aplicação dos processos de erosão e dilatação sobre a imagem da carta, as duas imagens resultantes (uma erodida e outra dilatada) são sobrepostas e é feita a subtração dos *pixels*, compatíveis nas duas imagens resultando na imagem de borda. Na Figura abaixo é apresentado um exemplo de carta PUCPR com os contornos e bordas extraídos.

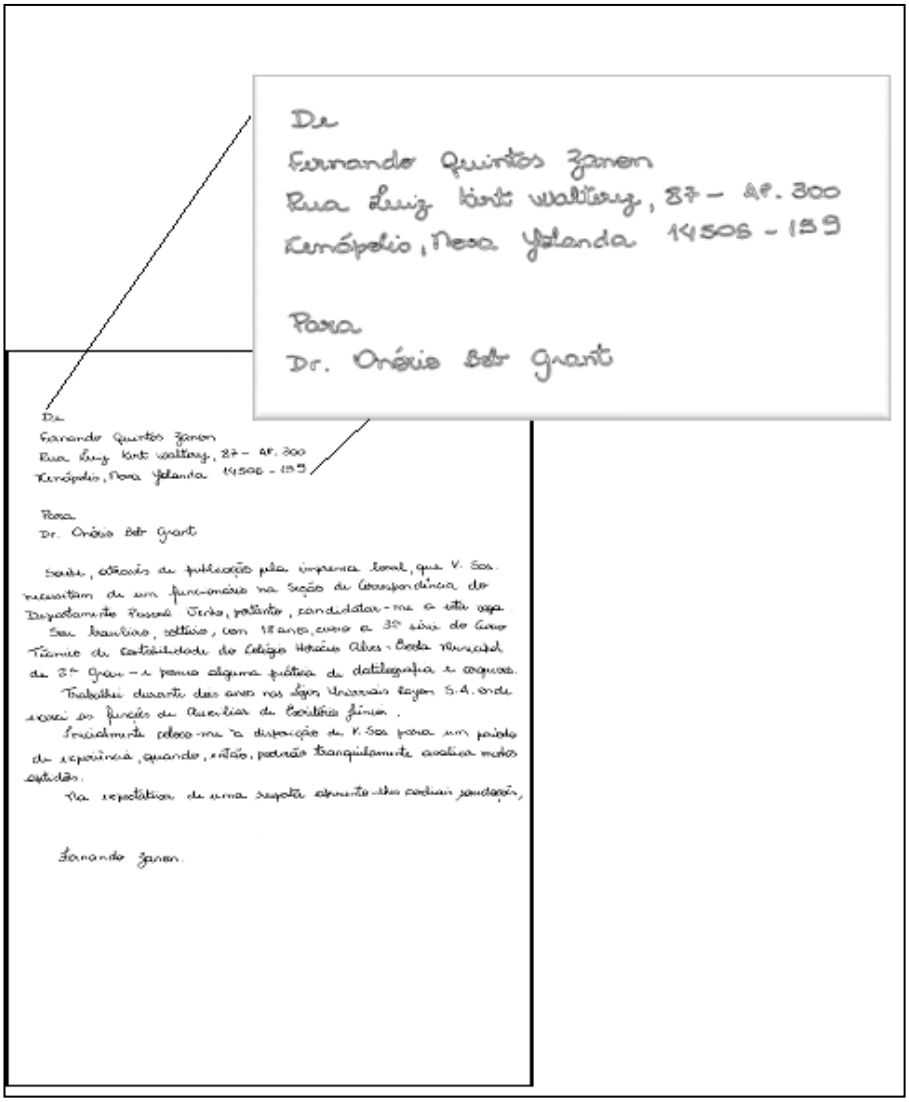


Figura 3.13. Exemplo de carta PUCPR após o processo de extração de contornos e bordas

e) Encontrar Regiões nas Palavras: Corpo da Palavra, Ascendente e Descendente.

A técnica empregada consistiu no histograma de transições branco/preto horizontal normalizado. As regiões são definidas por:

- Região ascendente: compreendida entre o limite superior máximo (LSM) da imagem e o limite superior do corpo da palavra (LS);

- Região corpo da palavra: compreendida entre o limite superior (LS) e inferior (LI) do corpo da palavra;
- Região descendente: compreendida entre o limite inferior do corpo da palavra (LI) e o limite inferior mínimo da palavra (LIM).

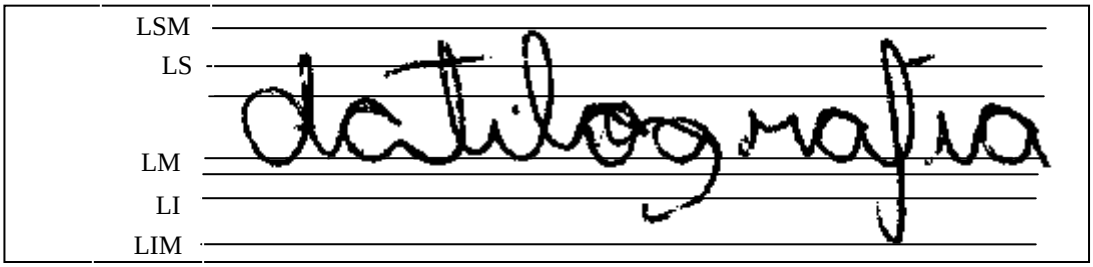


Figura 3.14. Regiões em uma palavra

Em seguida, foi necessário encontrar os laços fechados nas regiões ascendentes e descendentes (ver Figura 3.15). A técnica empregada para tanto foi baseada no algoritmo de determinação dos segmentos mais significativos em assinaturas, definido por Justino (2001).

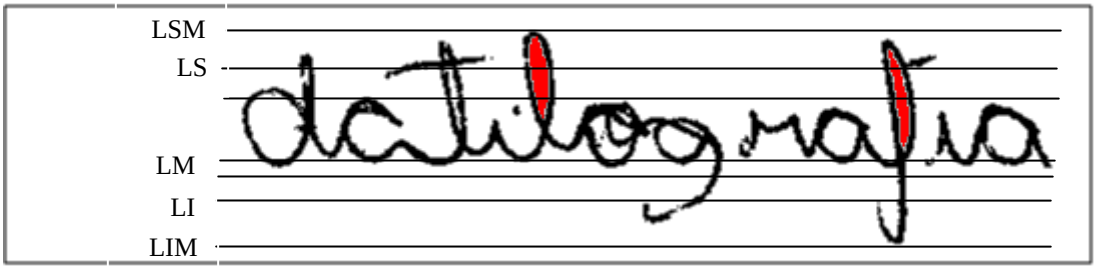


Figura 3.15. Laço na região ascendente

3.3.3. Extração de Características

Esta etapa refere-se à extração das características específicas, que se deseja avaliar da escrita do escritor. Com base nas características extraídas pode-se proceder a realização de cálculos/análises o que possibilita a identificação de autoria em um documento manuscrito.

O Quadro 3.1 apresenta as características grafométricas que compõem a base completa utilizada neste trabalho, bem como uma descrição do processo de extração das mesmas. Deve-se ressaltar que cada característica pode ser composta por uma única

informação (primitiva) ou por um conjunto de informações (primitivas) extraída(s) do documento manuscrito já pré-processado. Também foi incluída no Quadro 3.1 a posição em que as primitivas que compõem cada uma das características ocupam no vetor de primitivas que é submetido ao algoritmo de classificação utilizado.

De forma a garantir a consistência estatística e matemática dos resultados obtidos com o uso das características, as mesmas passaram por um processo de normalização, de forma que seus valores estão em uma escala entre 0 e 1.

O conjunto de características utilizado no método proposto conta atualmente com 85 (oitenta e cinco) primitivas agrupadas em 12 (doze) grupos de características (f_1-f_{12}). No entanto, pode-se observar, como destacado na Seção 2.4, que 04 (quatro) grandes grupos de características grafométricas estão sendo utilizadas, são elas: hábitos do usado espaço gráfico (f_1, f_3, f_4, f_5 *e* f_6), tamanho das palavras (f_2 *e* f_7), inclinação axial (f_8) e hábitos de traçado dos laços ascendentes e descendentes (f_9 - f_{12}).

Quadro 3.1. Características grafométricas do método proposto

Característica	**Posição no vetor de primitivas**	**Descrição**
f_1	1	*Número total de linhas da carta:* Levando-se em consideração que as cartas foram coletadas em papel A4, estabeleceu-se como padrão para a normalização que o número máximo de linhas da carta é 29, uma vez que ao se dividir a folha A4 em um tamanho padrão de linha (1 cm) o número total corresponde a 29 linhas. Também foram observadas visualmente cartas de 100 diferentes escritores, e em nenhuma dos casos o montante de linhas passou de 29. Dessa forma, esta característica foi computada da seguinte forma: $$f_1 = \frac{nrolinhasCarta}{29} \quad (3.1)$$

f_2	2 a 21	*Proporção de pixels pretos:* Para as 20 primeiras linhas da carta, com base na segmentação da primeira palavra de cada linha, é calculado o número de pixels pretos destas palavras. A 1ª palavra das 20 primeiras linhas foi inserida em uma *bounding box* e a altura e largura desta caixa foi computada. Dessa forma, depois de contados o número de pixels pretos dentro da caixa, os mesmos devem ser divididos pelo produto da altura e largura desta caixa. Por exemplo, dada a palavra apresentada abaixo:  Levando-se em consideração as medidas acima apresentadas, o calculo da proporção de pixels pretos normalizado seria da seguinte forma: $$f_2 = \frac{nroPixelsPretos}{X.Y} \quad (3.2)$$
f_3	22 a 41	*Posição da margem direita:* Para as 20 primeiras linhas da carta, a distância da margem direita é computada. Esta distância é definida usando-se uma linha de referência (linha imaginária que cruza a linha de texto ao meio de sua altura – esta linha pode ser melhor visualizada na Figura 3.16) e verificando o último pixel preto desta linha. Para a normalização desta característica, foi realizado o seguinte cálculo: $$f_3 = \frac{posicaoMargemDireita}{larguraDocumento} \quad (3.3)$$

f_4	42	*Posição da margem esquerda:* Esta distância é definida usando a linha de referência (já descrita anteriormente) de cada uma das linhas de texto da carta e identificando-se o menor valor de posição inicial. Para a normalização desta característica, foi realizado o seguinte cálculo: $$f_4 = \frac{posicaoMargemEsquerda}{larguraDocumento} \quad (3.4)$$
f_5	43	*Posição da margem superior*: Esta distância é definida pelo primeiro pixel preto da primeira palavra da carta (ou seja, da palavra segmentada da 1ª linha). Para a normalização desta característica, foi realizado o seguinte cálculo: $$f_5 = \frac{posicaoMargemSuperior}{alturaDocumento} \quad (3.5)$$
f_6	44	*Posição da margem inferior*: Esta distância é definida usando-se a linha de referência da última linha de texto da carta e identificando a posição final desta linha de referência. Para a normalização desta característica foi realizado o seguinte cálculo: $$f_6 = \frac{posicaoMargemInferior}{alturaDocumento} \quad (3.6)$$

f_7	45 a 64	*Altura da primeira palavra*: Para as 20 primeiras linhas da carta é calculada a altura da primeira palavra segmentada e teve seus valores limites extraídos na etapa de pré-processamento. Para a normalização desta característica foi realizada uma análise de 300 cartas de 100 diferentes escritores para a determinação da altura máxima a ser usada. Obteve-se o valor de h_{max} = 420 pixels (empiricamente definida). Dessa forma, tem-se a normalização desta característica: $$f_7 = \frac{alturaPalavra}{h_{max}} \quad (3.7)$$
f_8	65 a 81	*Inclinação axial:* Esta característica foi extraída com base na abordagem apresentada e validada no trabalho de Baranoski (2005), a saber: ✓ Cinco fragmentos são randomicamente selecionados da imagem segmentada, e já com os contornos de borda extraídos; ✓ Em seguida, verificam-se os fragmentos de borda em todas as direções, partindo do *pixel* central e conferindo os *pixels* posteriores com um operador lógico *AND*, finalizando nas extremidades do elemento estruturante (que possui tamanho 5) 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.

f_9	82	*Altura média dos laços ascendentes e descendentes*: Para computar esta característica foram realizados os seguintes procedimentos: ✓ A altura média de todos os laços ascendentes e descendentes de *toda* a carta é computada – HM_{geral}; ✓ O desvio padrão médio da altura de todos os laços ascendentes e descendentes de *toda* a carta é computado – DPH_{geral}; ✓ *Cinco* fragmentos são randomicamente selecionados da carta, e para *cada fragmento*, *três* laços são também randomicamente selecionados; ✓ A altura média de cada laço de cada fragmento é então computada: $$HM_{laco} = \frac{AlturaLaco - HM_{geral}}{DPH_{geral}} \quad (3.8)$$ ✓ Em seguida, procede-se o cálculo da altura média dos laços de um fragmento selecionado: $$HM_{framgmento} = \frac{\sum HM_{laco}}{3} \quad (3.9)$$ ✓ Finalmente, calcula-se a altura média dos laços de todos os fragmentos selecionados: $$f_9 = \frac{\sum HM_{fragmento}}{5} \quad (3.10)$$

f_{10}	83	*Largura média dos laços ascendentes e descendentes*: Para computar esta característica foram realizados os seguintes procedimentos: ✓ A largura média de todos os laços ascendentes e descendentes de *toda* a carta é computada – WM_{geral}; ✓ O desvio padrão médio da largura de todos os laços ascendentes e descendentes de *toda* a carta é computado – DPW_{geral}. ✓ *Cinco* fragmentos são randomicamente selecionados da carta, e para *cada fragmento*, *três* laços são também randomicamente selecionados; ✓ A largura média de cada laço de cada fragmento é computada: $$WM_{laco} = \frac{LarguraLaco - WM_{geral}}{DPW_{geral}} \quad (3.11)$$ ✓ Em seguida, procede-se o cálculo da largura média dos laços de um fragmento selecionado: $$WM_{fragmento} = \frac{\sum WM_{laco}}{3} \quad (3.12)$$ ✓ Finalmente, calcula-se a largura média dos laços todos os fragmentos selecionados: $$f_{10} = \frac{\sum WM_{fragmento}}{5} \quad (3.13)$$

f_{11}	84	*Tamanho médio (em número de pixels) dos laços ascendentes e descendentes*: Para computar esta característica foram realizados os seguintes procedimentos: ✓ O tamanho médio de todos os laços ascendentes e descendentes de *toda* a carta é computado – TM_{geral}; ✓ O desvio padrão médio do tamanho de todos os laços ascendentes e descendentes de *toda* a carta é computado – DPT_{geral}. ✓ *Cinco* fragmentos são randomicamente selecionados da carta, e para *cada fragmento*, *três* laços são também randomicamente selecionados; ✓ O tamanho de cada laço de cada fragmento é computado: $$TM_{laco} = \frac{NumeroPixelsLaco - TM_{geral}}{DPT_{geral}} \quad (3.14)$$ ✓ Em seguida, procede-se o cálculo do tamanho médio dos laços de um fragmento selecionado: $$TM_{fragmento} = \frac{\sum TM_{laco}}{3} \quad (3.15)$$ ✓ Finalmente, calcula-se o tamanho médio dos laços de todos os fragmentos selecionados: $$f_{11} = \frac{\sum NPM_{fragmento}}{5} \quad (3.16)$$

f_{12}	85	*Ângulo geral dos laços ascendentes e descendentes*: Para computar esta característica foram realizados os seguintes procedimentos: ✓ O ângulo médio de todos os laços ascendentes e descendentes de *toda* a carta é computado – A_{geral}; ✓ O desvio padrão médio do ângulo de todos os laços ascendentes e descendentes de *toda* a carta é computado – DPA_{geral}. ✓ *Cinco* fragmentos são randomicamente selecionados da carta, e para *cada fragmento*, *três* laços são também randomicamente selecionados; ✓ O ângulo de cada laço de cada fragmento é computado: $$AM_{laco} = \frac{AnguloGeralLaco - A_{geral}}{DPA_{geral}} \quad (3.17)$$ ✓ Em seguida, procede-se o cálculo do ângulo médio dos laços de um fragmento selecionado: $$AM_{fragmento} = \frac{\sum AM_{laco}}{3} \quad (3.18)$$ ✓ Finalmente, calcula-se o ângulo médio dos laços de todos os fragmentos selecionados: $$f_{12} = \frac{\sum AM_{fragmento}}{5} \quad (3.19)$$

As Figuras 3.16 (características f_1 a f_8) e 3.17 (características f_9-f_{12}) apresentam uma visão geral do processo de extração de características. Pode-se observar que após o pré-processamento da imagem do documento manuscrito, todas as características definidas pelo método proposto são computadas para serem incluídas no vetor de primitivas.

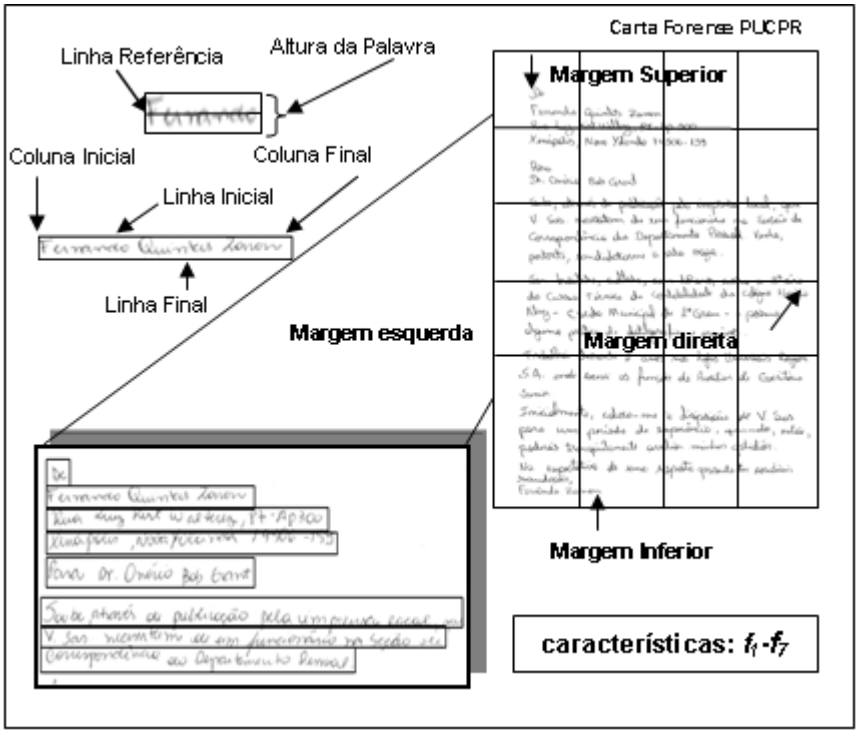


Figura 3.16. Visão geral do processo de extração de característica (f_1-f_7)

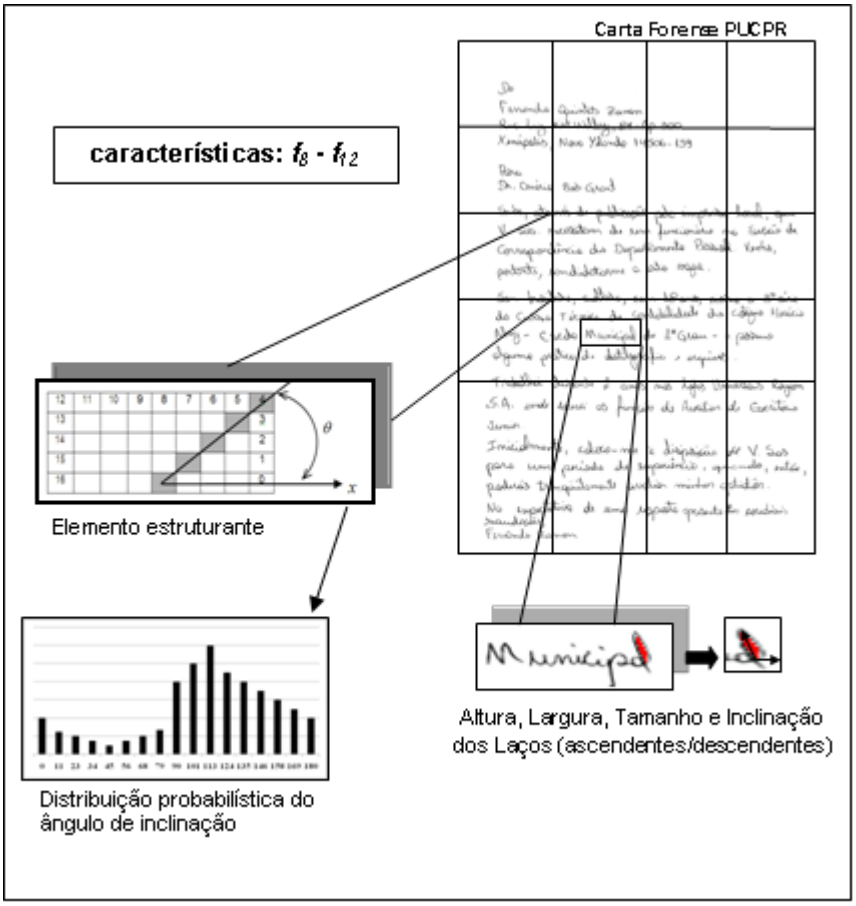


Figura 3.17. Visão geral do processo de extração de característica (f_8-f_{12})

3.3.4. Seleção de Características

De acordo com Amaral et al. (2013a), a seleção de características que melhor representam um problema a ser modelado, que são irrelevantes ou redundantes e àquelas que afetam negativamente os resultados (taxas de acerto) é fundamental.

Segundo Dash e Liu (1997), a busca manual das melhores características é exponencialmente proibitiva, mesmo com um número moderado de características. Dessa forma, no contexto desta pesquisa, foi aplicada, no conjunto de características grafométricas proposto, uma abordagem de seleção de características conforme apresentado em Amaral et al. (2013a). Por meio da aplicação de métodos de seleção de características em grandes conjuntos de primitivas, é possível se obter subconjuntos de menor dimensionalidade com as mesmas taxas de acerto, ou em alguns casos, com taxas de acerto até mesmo melhores.

De acordo com Liu e Yu (2005), um processo de seleção de características é composto por quatro passos, como pode ser observado na Figura 3.18. Estes passos consistem na: *geração do subconjunto*, *avaliação do subconjunto*, *definição do critério de parada* e *validação dos resultados*.

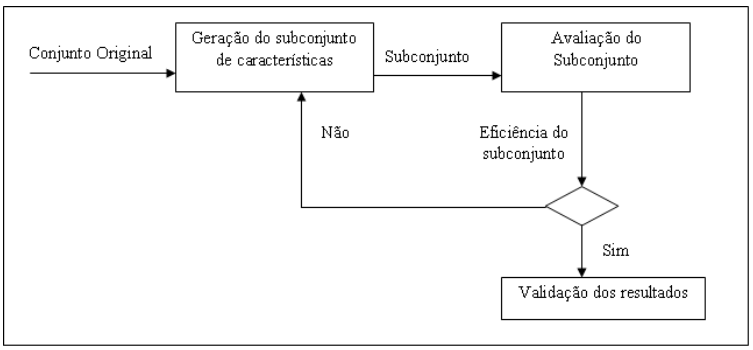


Figura 3.18. Processo geral de seleção de características

[Fonte: Adaptado de (LIU; YU, 2005)]

A *geração de subconjunto* é basicamente um processo de busca heurística no qual cada estado no espaço de busca especifica um subconjunto de características candidato para avaliação. Este processo é determinado pelo ponto de início da busca que por sua vez influencia a direção da busca (para frente - *forward*, para trás - *backward* ou bidirecional); e também pela estratégia de pesquisa. Analisar todos os subconjuntos candidatos, sendo que para um conjunto com N características existem 2^N subconjuntos candidatos, ou executar uma busca exaustiva, mesmo com um N pequeno é proibitivo. Consequentemente, diferentes estratégias de pesquisa tem sido exploradas, entre elas: completa, sequencial e randômica (LIU; YU, 2005).

Neste trabalho tem-se um conjunto de primitivas grafométricas de tamanho $N = 85$ *e, portanto,* foi aplicada a estratégia de pesquisa sequencial, iniciando com o conjunto de características vazio e com direção de busca "para frente" – *forward*. De acordo com Liu e Motoda (1998), a busca sequencial propicia que todo o conjunto de características seja avaliado reduzindo o risco de perder subconjuntos ótimos.

Cada novo *subconjunto gerado deve ser avaliado* de acordo com algum critério de validação (LIU; YU, 2005). Neste contexto, a eficiência de um subconjunto é determinada pelo critério empregado. Os critérios de avaliação podem ser categorizados em dois grupos com base em sua dependência de algoritmos de mineração que serão aplicados ao subconjunto de características selecionado. Neste trabalho de pesquisa adotou-se um critério de avaliação independente. Os critérios independentes são usados em algoritmos do modelo de filtros que avaliam a eficiência do subconjunto de características explorando seus aspectos intrínsecos sem envolver qualquer algoritmo de mineração. Exemplos de critérios independentes são medidas baseadas em distância, informação, dependência e consistência. O critério de avaliação independente adotado neste trabalho foi o de medidas baseadas na dependência, como discutido em Hall (1998). Segundo Liu e Yu (2005), as medidas baseadas em dependência são também conhecidas como medidas de correlação ou medidas de similaridade. Estas medem a habilidade de predizer o valor de uma variável a partir de outra.

O *critério de parada* determina quando o processo de seleção de características deve ser encerrado. Frequentemente, o critério de parada pode ser: busca completa; algum limite encontrado; adição ou remoção subsequente de alguma característica que falhou na produção do melhor subconjunto; e encontrar um conjunto suficiente bom (por exemplo, se a taxa de erro atingida é menor do que a taxa de erro permitida para uma determinada tarefa) (LIU; YU, 2005). Neste trabalho, o critério de parada utilizado foi o limite máximo no número de iterações, correspondendo a 5, sendo que este valor foi experimentalmente determinado.

O último passo do processo de seleção de características é a *validação dos resultados*. Esta tarefa corresponde à medida direta do resultado obtido, quando um conhecimento anterior dos dados é empregado. Nos experimentos realizados neste trabalho, são comparados os resultados (taxas de identificação) produzidos pelo conjunto completo de características (f_1-f_{12}) com os resultados produzidos com as características selecionadas.

Na Seção 3.4.2 são apresentados os resultados dos experimentos, incluindo os resultados obtidos com a aplicação do processo de seleção de características.

3.3.5. Método de Classificação

A etapa de classificação envolve a utilização de algoritmos que possibilitem a identificação do documento em análise. Como resultado deste processo, espera-se obter uma lista com o *ranking* dos escritores mais prováveis do documento em análise. Inicialmente foram realizados experimentos com os algoritmos de classificação SVM (VAPNIK, 1979) e KNN (DUDA; HART, 1973). Em função do melhor desempenho (em termos de taxa de acerto) apresentado pelo classificador SVM e por este ser uma abordagem de classificação usada em muitos outros estudos sobre identificação de autoria em documentos manuscritos descritos na literatura, SVM foi o classificador escolhido para o método proposto. Os resultados dos experimentos iniciais realizados com o classificador KNN foram incluídos no Apêndice A.

Os experimentos (descritos nas Seções 3.4) foram conduzidos de forma que, inicialmente, o vetor de primitivas foi avaliado integralmente, e sequencialmente foram realizados experimentos com características isoladas e grupos de características. Este tipo de análise foi essencial para que se pudesse identificar o "poder" discriminatório de cada uma das características, bem como, dos agrupamentos de características.

SVM *(Support Vector Machine)*

Segundo Burges (1998), embora a pesquisa sobre Máquinas de Vetores de Suporte (SVM) tenha sido iniciada no final dos anos 70, os SVM constituem uma técnica de aprendizado que vem recebendo nas últimas décadas crescente atenção da comunidade de Aprendizado de Máquina. De acordo com Scarpel (2005), SVM é um procedimento construtivo universal de aprendizagem baseado na teoria de aprendizagem estatística. O termo universal significa que o SVM pode ser utilizado para o aprendizado de várias representações como as redes neurais, as funções de base radial e funções polinomiais.

Enquanto técnicas tradicionais para reconhecimento de padrões são baseadas na minimização do *risco empírico*, isto é, tenta-se a otimizar o desempenho sobre o conjunto de treinamento, SVM minimizam o *risco estrutural*, isto é, a probabilidade de classificar de forma errada padrões ainda não vistos por uma distribuição de probabilidade de dados fixa e desconhecida (LIMA, 2002).

Algumas aplicações automatizadas com o uso de SVM podem ser citadas nas mais variadas áreas de pesquisa, 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 reconhecimento de cadeias de dígitos manuscritos e identificação de autoria (OLIVEIRA; SABOURIN, 2004); (CHEN et al., 2010).

Neste trabalho utilizamos o algoritmo SMO (*Sequential Minimal Optimization*) (PLATT et al., 2000) uma implementação de SVM disponibilizada na ferramenta WEKA (BOUCKAERT et al., 2009) que usa Kernel polinomial. Nesta implementação todos os valores que estão faltando no vetor de primitivas são substituídos globalmente e os atributos nominais são transformados em binários. Além disso, todos os dados de treinamento são normalizados[7] por *default.* No entanto, foi possível melhorar as taxas de acerto quando foi utilizado um filtro para padronizar[8] os dados de treinamento, ao invés de utilizar a normalização padrão.

3.4. Experimentos

Para validar os resultados obtidos com o método apresentado neste trabalho foi necessária a realização de experimentos. Tais experimentos seguiram um protocolo rigoroso para que os mesmos pudessem ser replicados em diferentes contextos. Sabe-se que diferentes variáveis afetam os resultados de pesquisas como a aqui apresentada, podendo-se citar: número de escritores, abordagem de classificação adotada e base de dados utilizada para o treinamento e teste. Dessa forma, a seguir encontra-se descrito o protocolo dos experimentos realizados.

3.4.1. Protocolo dos Experimentos

A Figura 3.19 apresenta uma visão geral do protocolo seguido ao longo dos experimentos realizados. De um modo geral, o vetor de primitivas obtido no processo de extração de características é usado como entrada para o algoritmo de classificação.

---

[7] Dados de treinamento normalizados são modificados para permanecerem em um intervalo de valores específico.

[8] Padronizar um vetor significa subtrair seus dados de uma medida local e dividi-los por uma medida de escala. Por exemplo, se o vetor possui valores randômicos com uma distribuição gaussiana, deve-se subtrair a média e dividir pelo desvio padrão, obtendo-se assim valores randômicos em um "padrão normal".

Os estágios de treinamento e teste foram realizados submetendo o vetor de primitivas extraído (das cartas forenses) ao classificador SVM, cuja implementação utilizada encontra-se disponível na ferramenta WEKA. Para o estágio de treinamento o vetor de primitivas é usado para compor o modelo e para o estágio de teste o vetor é usado para a identificação do escritor.

Cada escritor presente no conjunto de teste é comparado com cada um dos modelos definidos no estágio de treinamento, perfazendo um protocolo do tipo todos-contra-todos – *all-against-all*, sendo que o melhor resultado é selecionado por meio da técnica "vencedor leva tudo" – *winner-takes-all* e, portanto, indicado pelo classificador. Para os experimentos cujos resultados são apresentados na próxima Seção, foram selecionados aleatoriamente 200 diferentes escritores da base de cartas forenses PUCPR. Para o estágio de treinamento foram utilizadas duas cartas de cada escritor (totalizando 400 diferentes cartas), e para a fase de teste foi utilizada a 3ª carta de cada um dos 200 escritores utilizados na fase de treinamento (totalizando 200 diferentes cartas).

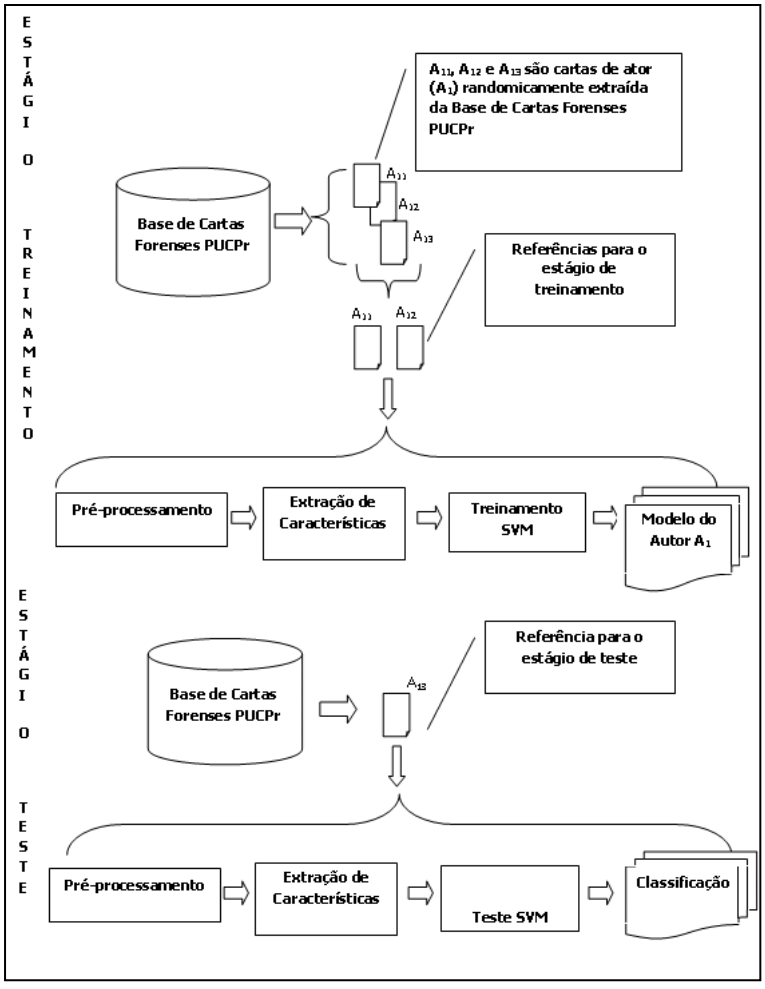


Figura 3.19. Protocolo dos experimentos

3.4.2. Resultados dos Experimentos

Foram realizados experimentos focando, principalmente, em dois aspectos:

- avaliar a eficiência individual e em grupo das características grafométricas extraídas de uma carta forense e que compõem o método proposto;
- obter o número de escritores necessários para validação do método, ou seja, o número de escritores no qual não se observa ganho ou perda nas taxas de acerto alcançadas.

Para tanto, experimentos com um número incremental de escritores (de 20 em 20) foram efetuados, levando-se em consideração todas as características e combinações de características.

A Tabela 3.2 apresenta o resultado individual, em termos de identificação do correto escritor, de cada uma das características grafométricas estudadas. Uma análise detalhada dos resultados obtidos nos experimentos é apresentada no próximo capítulo.

Tabela 3.2. Resultados de experimentos individuais com as características grafométricas

Característica	**Número de Escritores / Taxa de Acerto (%)**									
	20	**40**	**60**	**80**	**100**	**120**	**140**	**160**	**180**	**200**
f_1	15,00	7,50	3,33	5,00	4,00	5,00	4,30	3,75	3,89	4,00
f_2	25,00	30,00	18,33	20,00	17,00	15,83	15,00	11,88	13,89	14,00
f_3	20,00	20,00	16,67	13,75	15,00	14,17	11,42	10,62	8,33	9,00
f_4	10,00	10,00	8,33	3,75	5,00	4,17	2,86	1,25	1,67	1,50
f_5	10,00	7,50	6,67	6,25	5,00	3,33	4,30	1,88	2,22	1,00
f_6	30,00	15,00	5,00	5,00	8,00	3,33	3,57	1,88	2,22	1,00
f_7	35,00	35,00	30,00	26,25	25,00	22,50	21,43	18,13	18,33	19,50
f_8	85,00	85,00	76,67	73,75	68,00	71,66	67,15	65,00	62,78	60,50
f_9	5,00	5,00	5,00	5,00	4,00	3,33	2,14	1,88	1,67	1,00
f_{10}	5,00	5,00	5,00	3,75	3,00	3,33	2,14	2,50	1,67	1,00
f_{11}	6,00	5,00	5,00	3,75	3,00	3,33	1,43	1,88	1,00	1,00
f_{12}	10,00	7,50	5,00	3,75	4,00	3,33	2,14	1,25	1,67	2,00

Após a realização de experimentos individuais foram definidos empiricamente alguns agrupamentos de características e experimentos com estes grupos foram também realizados (ver Tabela 3.3). Deve-se ressaltar que para cada característica individual e

em grupo foram realizados 10 experimentos diferentes, um para cada grupo de escritores avaliado, totalizando 340 - 34 agrupamentos diferentes x 10 testes diferentes. Tais experimentos foram parcialmente apresentados em Amaral et al. (2012b).

Tabela 3.3. Resultados dos experimentos com agrupamentos empíricos de características

Grupo de Características	**Número de Escritores / Taxa de Acerto (%)**									
	20	**40**	**60**	**80**	**100**	**120**	**140**	**160**	**180**	**200**
f_1 & f_8	85,00	87,50	76,67	78,75	76,00	71,70	71,43	68,75	68,89	67,50
f_2 & f_8	60,00	67,50	51,67	48,75	49,00	42,50	42,14	41,25	40,00	40,00
f_3 & f_8	60,00	57,50	50,00	50,00	47,00	40,83	40,00	36,25	56,11	35,00
f_4 & f_8	80,00	77,50	73,33	72,5	69,00	65,00	63,57	60,00	56,11	56,50
f_5 & f_8	80,00	77,50	71,67	72,5	69,00	65,83	62,86	58,75	57,22	56,50
f_6 & f_8	90,00	90,00	83,33	80,00	78,00	75,00	71,43	69,38	67,78	68,50
f_7& f_8	75,00	80,00	63,33	62,50	59,00	60,00	56,43	50,63	50,56	51,50
f_1 & f_6 & f_8	85,00	90,00	85,00	83,75	80,00	78,33	77,14	75,00	71,11	69,00
f_1 & f_2 & f_3 & f_4 & f_5& f_6 & f_7	50,00	45,50	35,00	35,00	33,00	31,70	32,15	31,85	31,66	31,04
f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7& f_8	65,00	62,50	58,33	57,50	58,00	54,16	52,86	51,88	50,55	50,50
f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7&f_8 & f_9 & f_{10} & f_{11}& f_{12}	65,00	60,00	55,00	55,00	55,00	52,50	51,43	50,00	49,44	49,00
f_1 & f_6 & f_8 & f_9 & f_{10}& f_{11}& f_{12}	80,00	77,25	65,00	63,75	58,00	63,33	60,71	60,00	59,44	59,50
f_1 & f_6 & f_8& f_{11}	85,00	82,50	80,00	76,25	76,00	73,33	68,57	66,86	66,66	66,00
f_1 & f_6 & f_8& f_{12}	85,00	92,50	88,30	87,50	84,00	83,33	80,00	75,63	72,77	70,50
f_1 & f_6 & f_8& f_9	65,00	77,50	73,33	75,00	76,00	71,67	65,71	66,25	66,11	66,00
f_1 & f_6 & f_8& f_{10}	85,00	90,00	68,33	76,25	74,00	73,33	72,86	70,63	68,88	68,00
f_1 & f_6 & f_8& f_9& f_{10}& f_{11}	80,00	80,00	65,00	65,00	60,00	61,67	60,71	59,38	58,89	58,50
f_1 & f_6 & f_8& f_9 & f_{10}& f_{12}	65,00	75,00	68,33	67,50	64,00	65,00	62,15	61,88	61,11	61,00
f_1 & f_6 & f_8 & f_9& f_{11}	80,00	77,50	71,67	68,75	64,00	61,67	64,29	62,50	62,70	60,00
f_1 & f_6 & f_8& f_9& f_{10}	65,00	75,00	68,33	72,50	70,00	65,83	65,00	62,50	63,80	62,00
f_1 & f_6 & f_8& f_9& f_{11}& f_{12}	80,00	77,50	70,00	66,25	66,00	63,33	62,14	61,25	60,50	60,00
f_1 & f_6 & f_8 & f_{11}& f_{12}	85,00	85,00	76,67	76,25	73,00	68,33	67,14	65,00	62,70	63,00

Pode-se observar que a maioria dos agrupamentos avaliados incluiu a característica f_8 em sua composição. Esta decisão foi tomada em função do alto poder

discriminatório individual desta característica, como pode ser observado na Tabela 3.2. Uma discussão detalhada sobre estes resultados é apresentada no Capítulo 4.

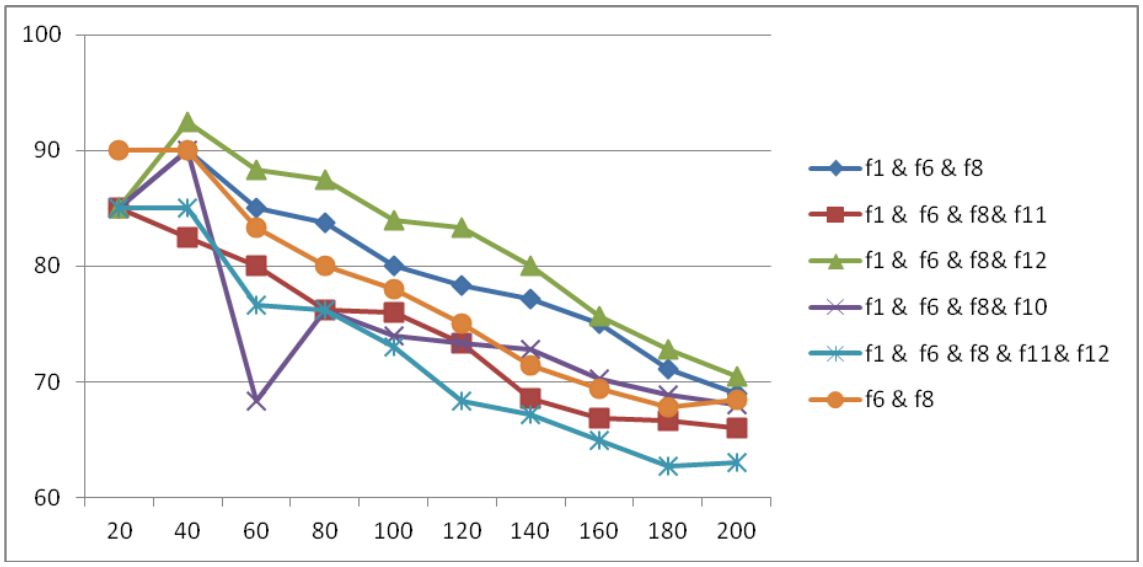
Com o objetivo de encontrar o melhor agrupamento de características, foi então aplicado um processo de seleção de características, conforme discutido na Seção 3.3.4 e apresentado em Amaral et al. (2013a). Assim, o melhor grupo de características (*goodness subset - GS*) é formado por: f_1 & f_6 & f_8 & f_{12}. Os resultados dos experimentos realizados com este grupo são apresentados na Tabela 3.4. É interessante ressaltar, que este grupo já havia sido experimentado empiricamente e que os resultados da seleção de características vieram a comprovar a eficiência deste agrupamento de características.

Tabela 3.4 Resultados dos experimentos com o melhor grupo de características - GS

Número de Escritores	Taxa de Acerto (%)
20	85,00
40	92,50
60	88,30
80	87,50
100	84,00
120	83,33
140	80,00
160	75,63
180	72,77
200	70,50

O Gráfico 3.1 apresenta, especificamente para os 6 melhores agrupamentos de características apresentados na Tabela 3.3, a relação entre o crescimento/redução das taxas de acerto em função do aumento do número de escritores. Pode-se observar que praticamente nenhum ganho ou perda é obtido quando se atinge um número de escritores próximo a 200. Este experimento, com números incrementais de escritores, foi discutido em Amaral et al. (2013b; 2013c).

Gráfico 3.1. Relação entre o número de escritores e as taxas de acerto



3.5. Considerações Finais

Este capítulo apresentou o método proposto, incluindo todas as etapas que o compõem, bem como, detalhes de como as características grafométricas foram implementadas. A abordagem de seleção de características adotada e que resultou na melhor combinação de características para o método proposto também foi discutida.

Os resultados e o protocolo experimental adotado para validação do método proposto também foram apresentados. Neste contexto, o próximo capítulo apresenta uma discussão detalhada sobre tais resultados.

Capítulo 4

Análise e Discussão dos Resultados

4.1. Considerações Iniciais

Este capítulo apresenta uma discussão detalhada sobre os resultados obtidos com experimentos realizados no método para desenvolvido nesta pesquisa voltado à identificação de autoria em documentos manuscritos. Dessa forma, a Seção 4.2 apresenta uma análise crítica de cada um dos grupos de características grafométricas estudados e que compõem o método proposto. Na Seção 4.3 é apresentada uma discussão sobre o nível de granularidade das características grafométricas presentes no método proposto. A Seção 4.4 aborda as principais diferenças, em termos de taxas de acerto e em termos do processo de extração, de características grafométricas e não-grafométricas. Finalmente, na Seção 4.5 é apresentada uma análise comparativa entre os melhores resultados obtidos com o método proposto e outras abordagens para identificação de escritores presentes na literatura que também utilizam características grafométricas.

4.2. Análise Crítica dos Resultados por Grupo de Características Grafométricas

A seguir é apresentada uma análise dos resultados dos experimentos levando-se em consideração os 4 (quatro) grandes grupos de características grafométricas utilizados no método proposto e apresentados na Seção 2.4.

4.2.1. Hábitos de Uso de Espaço Gráfico

As características grafométricas que compõem este grupo são: número de linhas da carta (*f1*), posição da margem direita (f_3), posição da margem superior (f_4), posição da margem esquerda (f_5) e posição da margem inferior (f_6). Tais características são muito utilizadas pelos peritos em análises forenses, uma vez que oferecem informações sobre a forma como o escritor faz uso do espaço disponível para a escrita, denominado de espaço gráfico. Individualmente, como pode ser observado na Tabela 4.1, estas características não apresentam taxas significativas de identificação. No entanto, quando utilizadas de modo associado com outra característica de maior poder discriminatório, como inclinação axial (f_8), consegue-se auxiliar o classificador na tomada de decisão entre possíveis escritores de um manuscrito. Estas situações podem ser observadas nas Tabelas 4.1 e 4.2.

Deve-se ressaltar que as características f_1 e f_6 fazem parte do conjunto que apresentou as melhores taxas de identificação (ver Tabela 3.4) e que é resultante da aplicação de um processo de seleção de características (conforme apresentado na Seção 3.3.4).

Tabela 4.1. Resultados de experimentos com as características grafométricas: hábitos de uso/posicionamento do texto na folha

Característica	**Número de Escritores / Taxa de Acerto (%)**									
	20	**40**	**60**	**80**	**100**	**120**	**140**	**160**	**180**	**200**
f_1	15,00	7,50	3,33	5,00	4,00	5,00	4,03	3,75	3,89	4,00
f_3	20,00	20,00	16,67	13,75	15,00	14,17	11,42	10,63	8,33	9,00
f_4	10,00	10,00	8,33	3,75	5,00	4,17	2,86	1,25	1,67	1,50
f_5	10,00	7,50	6,67	6,25	5,00	3,33	4,30	1,88	2,22	1,00
f_6	30,00	15,00	5,00	5,00	8,00	3,33	3,57	1,88	2,22	1,00

Na Figura 4.1 pode-se observar dois escritores com ângulos de inclinação semelhantes, no entanto com usos do espaço gráfico da folha de papel bem distintos, em especial a posição da margem inferior (f_6). Dessa forma, embora a característica inclinação axial permita que o classificador selecione os escritores com ângulos de inclinação similares ao escritor do documento questionado, foi por meio da característica f_6 que foi possível se efetivar a classificação correta.

A Figura 4.2, apresenta uma situação similar, na qual dois escritores embora com ângulos de inclinação de escrita semelhantes utilizaram um número diferente de linhas para escrever o conteúdo solicitado na carta forense PUCPR e esta característica foi fundamental para a tomada de decisão pelo classificador.

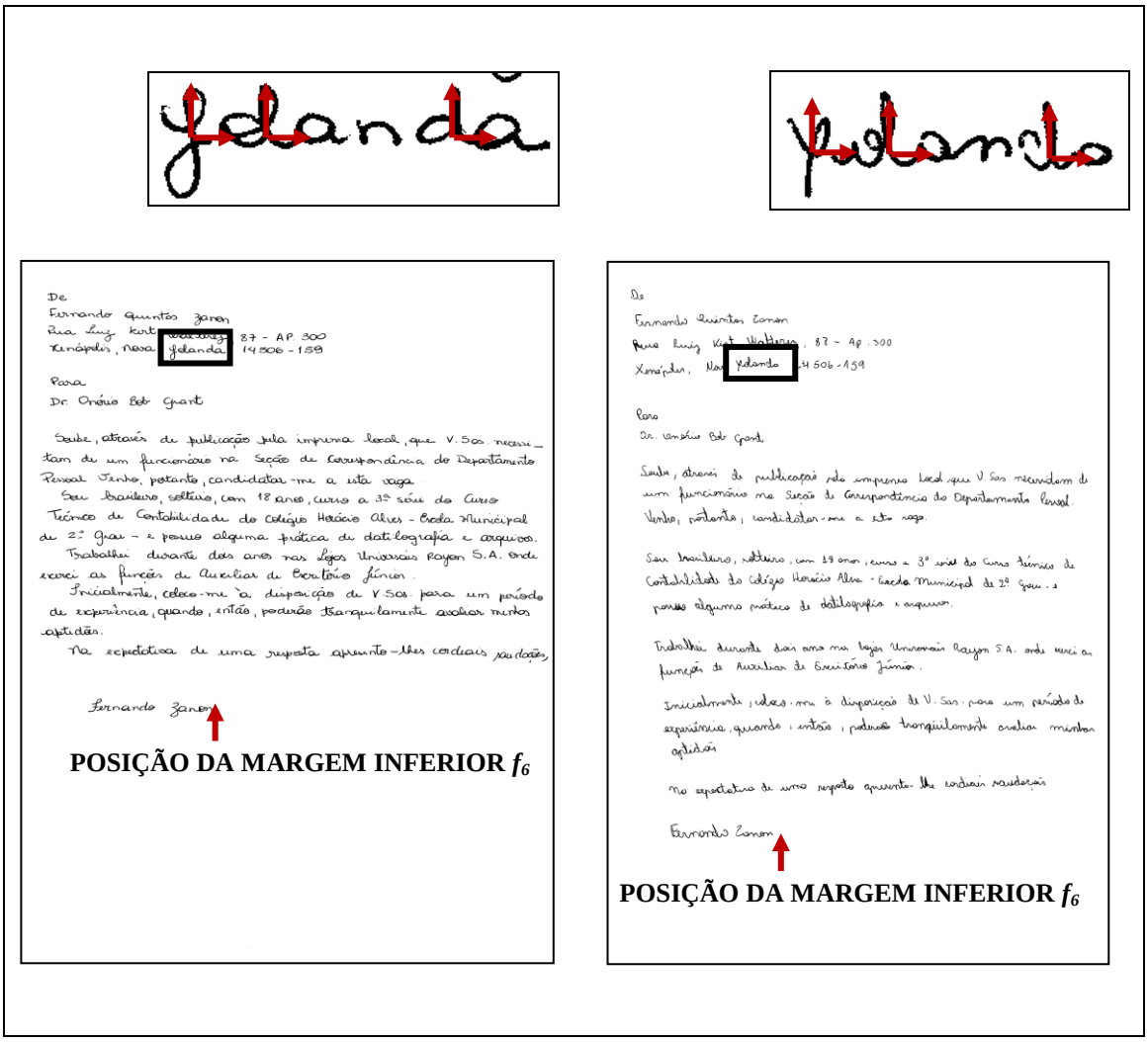


Figura 4.1. Uso da característica f_6 para a tomada de decisão

As demais características f_3, f_4 e f_5, embora úteis aos peritos forenses, mesmo quando combinadas com outras características grafométricas mais significativas, como é caso da inclinação axial – f_8, não apresentaram resultados interessantes. Isto pode ser observado na Tabela 3.3. Observa-se que ganhos não são obtidos quando estas características são inseridas na base e até mesmo em alguns casos ruídos são introduzidos nos dados, causando redução nas taxas de acerto.

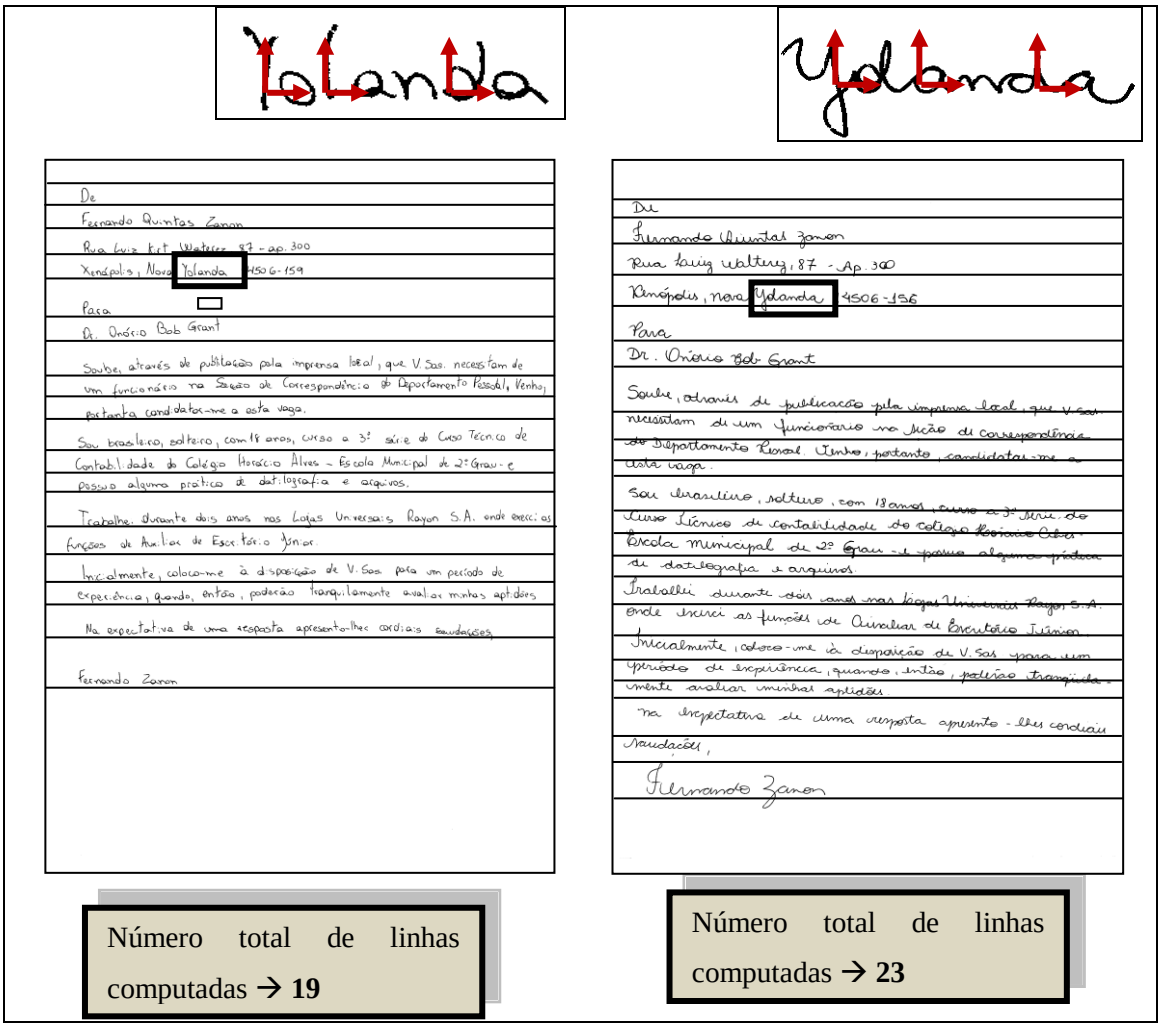


Figura 4.2. Uso da característica f_2 para a tomada de decisão

4.2.2. Tamanho das Palavras

As características grafométricas que compõem este grupo são: proporção de pixels pretos da primeira palavra de cada linha (f_2) e altura da primeira palavra de cada linha (f_7). Os resultados individuais obtidos com a utilização destas características podem ser observados na Tabela 4.2.

Tabela 4.2. Resultados de experimentos com as características grafométricas: tamanho das palavras

Característica	**Número de Escritores / Taxa de Acerto (%)**									
	20	**40**	**60**	**80**	**100**	**120**	**140**	**160**	**180**	**200**
f_2	25,00	30,00	18,33	20,00	17,00	15,83	15,00	11,88	13,89	14,00
f_7	35,00	35,00	30,00	26,25	25,00	22,50	21,43	18,13	18,33	19,50

Pode-se observar que estas características, principalmente, em função de limitações em seu processo de extração, não trazem ganhos nos resultados dos experimentos como pode ser observado na Tabela 3.3. Isto ocorre, normalmente, porque na demarcação das palavras das linhas, em função de características de escrita do escritor como "escrever para cima (*rising*)" ou "escrever para baixo (*falling*)", com o algoritmo de extração utilizado, não é possível se delimitar adequadamente a 1ª palavra destas linhas.

A Figura 4.3 apresenta o problema, no qual em função de um hábito comum da escrita do escritor, não é possível se computar corretamente as características f_2 e f_7.

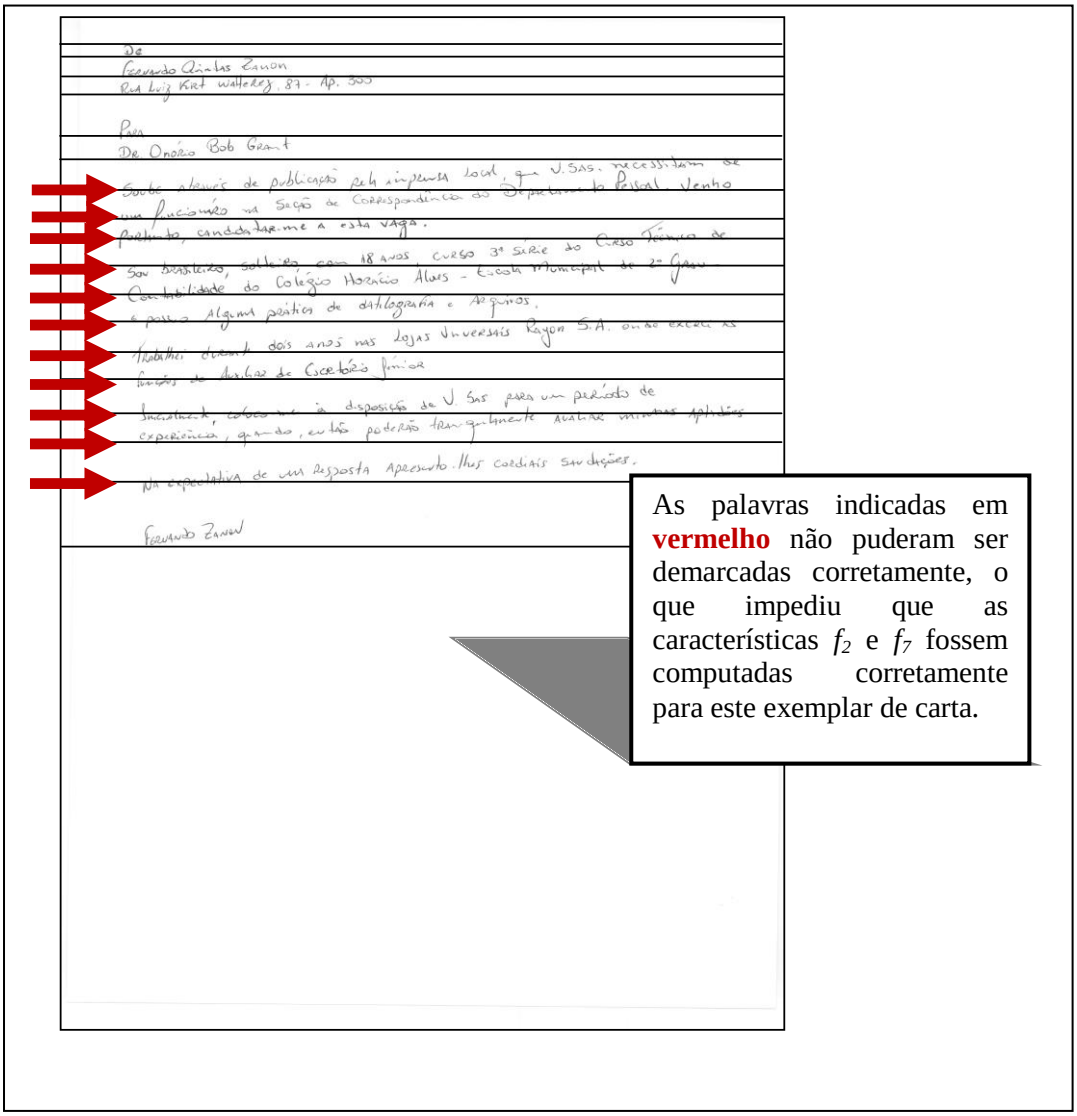


Figura 4.3. Demarcação das primeiras palavras de cada linha em um exemplar de carta presente na base de cartas forenses PUCPR

4.2.3. Inclinação Axial

A característica que compõe este grupo é a inclinação axial (f_8). Esta característica apresenta individualmente e em grupo a melhor taxa de acerto dentre todas as características grafométricas estudadas e incluídas no método proposto. Isto pode ser observado na Tabela 4.3 (análise individual) e também nas Tabelas 3.2 e 3.3.

Tabela 4.3. Resultados de experimentos com a característica inclinação axial

Número de Escritores	Taxa de Acerto (%)
20	85,00
40	85,00
60	76,67
80	73,75
100	68,00
120	71,66
140	67,15
160	65,00
180	62,78
200	90,50

Inclinação axial é uma característica grafométrica muito utilizada pelos peritos forenses e tem sido muito utilizada em abordagens automáticas para identificação de autoria (como pode ser observado na comparação apresentada na Seção 4.5).

Deve-se ressaltar que a característica f_8 faz parte do conjunto que apresentou as melhores taxas de identificação (ver Tabela 3.4) e que é resultante da aplicação de um processo de seleção de características (conforme apresentado na Seção 3.3.4).

A Figura 4.4 apresenta algumas variações no ângulo de escrita de exemplares de diferentes escritores presentes na base de cartas forenses PUCPR. Nestes exemplares foram colocadas demarcações do ângulo em pontos aleatórios em cada uma das linhas. Com tais demarcações, pode-se observar como a inclinação é uma característica marcante e constante em todo o documento de um escritor.

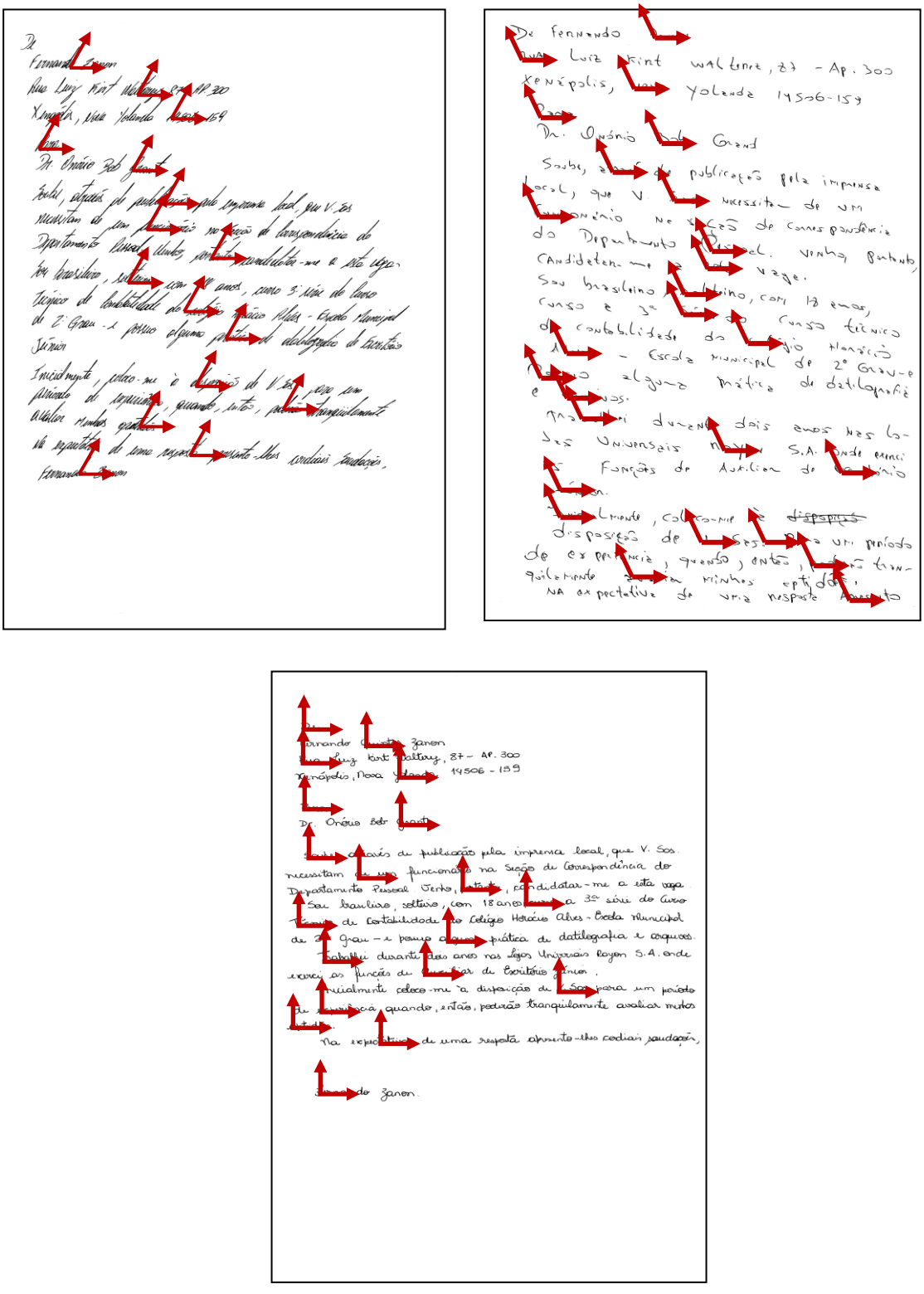


Figura 4.4. Exemplos de diferentes ângulos de escrita (esquerda, direita, vertical) para a característica f_8

4.2.4. Hábitos do Traçado dos Laços Ascendentes e Descendentes

As características grafométricas que compõem este grupo são: altura média dos laços ascendentes e descendentes (f_9), largura média dos laços ascendentes e descendentes (f_{10}), tamanho médio dos laços ascendentes e descendentes (f_{11}) e ângulo médio de inclinação dos laços ascendentes e descendentes (f_{12}). Estas características representam informações em um nível de granularidade menor do que as outras características grafométricas discutidas até o momento. Informações sobre os caracteres que possuem laços (*loops)* tais como: "l", "h", "t", "p", "g", "f" são coletadas para posteriormente as características f_9, f_{10}, f_{11} e f_{12} serem computadas.

A Tabela 4.4 apresenta os resultados individuais obtidos com estas características. Pode-se observar nas Tabelas 3.2 e 3.3 que, isoladamente, estas características não possuem um grande poder discriminatório, mas quando combina-se a característica f_{12} com as características f_1, f_6 e f_8 obtem-se o melhor agrupamento de características, ou seja, o agrupamento *GS* que apresenta as melhores taxas de acerto, tendo sido resultante do processo de seleção de características, conforme Seção 3.3.4.

Tabela 4.4. Resultados de experimentos com as características grafométricas: altura, largura, tamanho e ângulo de inclinação dos laços ascendentes e descendentes

Característica	**Número de Escritores / Taxa de Acerto (%)**									
	20	**40**	**60**	**80**	**100**	**120**	**140**	**160**	**180**	**200**
f_9	5,00	5,00	5,00	5,00	4,00	3,33	2,14	1,88	1,67	1,00
f_{10}	5,00	5,00	5,00	3,75	3,00	3,33	2,14	2,50	1,67	1,00
f_{11}	6,00	5,00	5,00	3,75	3,00	3,33	1,43	1,88	1,00	1,00
f_{12}	10,00	7,50	5,00	3,75	4,00	3,33	2,14	1,25	1,67	2,00

Deve-se destacar que, embora as características f_9, f_{10} e f_{11} apresentam bons resultados quando combinadas com a característica f_8, seus resultados não superam àqueles apresentados quando f_{12} é incluída ao melhor agrupamento de características (*GS*). Tais resultados ajudam a comprovar a eficiência da característica inclinação, seja ela em âmbito global (f_8), do documento como um todo, ou em âmbito local (f_{12}) para as palavras e caracteres que possuem laços ascendentes e descendentes.

A Figura 4.5 apresenta exemplos de laços ascendentes e descendentes extraídos de diferentes exemplares de cartas da base de cartas forenses PUCPR e que apresentam

diferentes ângulos de inclinação, altura, largura e tamanho, sendo que para um dos exemplos de laços, as características f_9, f_{10}, f_{11} e f_{12} são destacadas.

Figura 4.5. Exemplos de laços ascendentes e descendentes em diferentes palavras e caracteres de exemplares de cartas da base de cartas forenses PUCPR

4.3. Análise Crítica das Características Grafométricas de acordo com seu Nível de Granularidade

No conjunto de características grafométricas que compõem o método proposto foram incluídas, como pode ser observado na Figura 2.4, características globais e locais com diferentes níveis de granularidade (documento, linha, palavra e caractere). Nesta figura pode-se observar que características globais referem-se às informações retiradas do documento com um todo (f_1,f_3,f_4,f_5,f_6 e f_8), dos parágrafos e das linhas (f_2 e f_7). Enquanto que características locais referem-se a informações retiradas das palavras e caracteres (f_9,f_{10},f_{11},f_{12}). Apenas características em nível de parágrafo não foram estudadas e incorporadas no conjunto de características.

Quando se compara, no método proposto, o desempenho de acordo com a granularidade da característica, claramente observa-se que as características em nível de linha apresentam os maiores desvios nos resultados. Dessa forma, abordagens para correção de linha devem ser implementadas em trabalhos futuros para que tais características possam auxiliar no aumento das taxas de identificação. Deve-se ressaltar, que tais procedimentos de correção somente devem ser aplicados antes da extração das características de interesse, sem que sejam geradas modificações definitivas nas demais características da escrita, visto que modificações definitivas podem, por conseguinte, gerar erros na determinação de outras primitivas dependentes do aspecto original da escrita do escritor na carta forense. Deve-se também destacar que os peritos vêem com ressalvas este tipo de procedimento.

As características em nível de documento apresentaram-se como àquelas que mais contribuem para o aumento das taxas de identificação, como pode ser observado nas Tabelas 3.3 e 3.4, e são as características de maior relevância no grupo *GS*, resultante do processo de seleção de características. Isto reforça a ideia que em um primeiro momento, assim como procedimento normalmente adotado pelos peritos forenses, são informações "macro" àquelas que permitem identificar os principais candidatos dentre um conjunto de escritores para um documento questionado.

Em um segundo momento, características de granularidade mais fina, como é o caso das características em nível de palavras e caracteres, permitem refinar os resultados (ver Tabelas 3.3 e 3.4) e observar pontos específicos em um documento (como por exemplo, laços ascendentes e descendentes) que fornecem uma indicação mais precisa, dentre um conjunto de possíveis candidatos, do autor de um documento questionado.

4.4. Análise Crítica sobre o uso Exclusivo de Características Grafométricas

O conjunto de características presentes no método proposto neste trabalho de pesquisa inclui apenas características grafométricas, ou seja, que levam em consideração durante sua concepção os mesmo aspectos que são observados pelos peritos durante suas análises grafoscópicas. Dessa forma, neste trabalho, assim como ocorre na perícia forense, a análise de um documento questionado é dependente do texto do manuscrito em questão. Nesse contexto, necessita-se de uma base de manuscritos, tal como a base de cartas forenses PUCPR, que tenha levado em consideração durante sua

construção aspectos que são usualmente observados pelos peritos em análises de identificação de autoria.

O método proposto automatiza características que fornecem informações visuais para a tomada de decisão pelo perito, como por exemplo, o ângulo geral de inclinação da escrita de um escritor. Extrair manualmente estas informações, em grandes quantidades de exemplares de manuscritos, torna-se uma tarefa extremamente demorada, difícil e sujeita a erros, por isso, a automatização de todo ou parte deste processo é extremamente interessante. É importante ressaltar, que abordagens como a aqui proposta, ou seja, que utilizam características grafométricas são aceitas pela comunidade jurídica internacional nos tribunais, uma vez que o processo de análise do manuscrito em ambas abordagens segue os mesmos critérios, apenas o que muda é o método de aplicação (automatizado ou manual).

O processo de extração de características grafométricas, normalmente, envolve algoritmos de processamento de imagens e esse processo é realizado pixel a pixel na imagem do documento questionado. Tais algoritmos são complexos e, em função de uma série de particularidades do escritor, como por exemplo, o comportamento de "escrever para baixo/cima" (demonstrado na Figura 4.3) tem sua aplicabilidade limitada, trazendo resultados incorretos. Isto ocasiona reduções nas taxas de identificação, uma vez que ruídos são introduzidos nos resultados. Análises individuais das características foram realizadas (Tabela 3.2) e como trabalhos futuros deverão ser definidas estratégias para correção das limitações dos algoritmos utilizados. Destaca-se que mesmo obsrvando-se tais limitações, o método proposto alcançou os objetivos previstos inicialmente,

As abordagens que utilizam características não-grafométricas, como se pode observar na Tabela 2.3, normalmente envolvem a aplicação de transformações matemáticas sobre toda ou parte da imagem dos manuscritos, como por exemplo, o uso de filtros. Dessa forma, o conteúdo do manuscrito em análise não é relevante para o processo de identificação de autoria (e bases de textos variados podem ser usadas). Os algoritmos utilizados são menos limitados às particularidades de escrita e por isso taxas de identificação muito elevadas são obtidas com estas abordagens. No entanto, tais abordagens não são naturais aos peritos e aos juízes em um tribunal e seu uso em processos judiciais torna-se bastante limitado.

4.5. Comparação dos Resultado Obtidos com outros Trabalhos apresentados na Literatura

Para validar os resultados obtidos com o melhor grupo de características do método proposto (*GS*), na Tabela 4.5 apresenta-se uma comparação dos resultados obtidos pelo método proposto com outros estudos discutidos na literatura. Para garantir a confiabilidade dos resultados e a uniformidade da comparação realizada é necessário relacionar as taxas de identificação com o número de escritores, ou seja, com o tamanho da amostra utilizado nos experimentos.

Tabela 4.5. Comparação dos resultados obtidos com outros apresentados na literatura

Trabalho	Características Extraídas	Número de Escritores	Taxa de Acerto (%)
Schlapbach e Bunke (2004)	Inclinação, altura e obliquidade das linhas de texto.	50 escritores	94,40
Chen et al. (2010)	Informações relativas ao contorno de segmentos adjacentes e remoção de linhas de referência pré-impressas.	60 escritores	54,90
Luna et al. (2011)	Espaço percentual da margem esquerda e margem direita, separação entre linhas, direção geral da escrita, e espaço entre palavras.	30 escritores	88,00
Luna et al. (2011)	Proporção da zona média das palavras comparada com as zonas ascendentes e descendentes e inclinação das palavras.	30 escritores	88,00
Zois e Anastassopoulos (2000)	Uso de operadores morfológicos para obter o perfil horizontal das palavras.	50 escritores	95,00
Pervouchine e Leedham (2007)	Uso de características extraídas de três caracteres, são eles: “d”, “y” e “f” e o grafema “th”.	165 escritores	58,00
Método Proposto (*GS*)	Inclinação axial, número de linhas, posição da margem inferior, ângulo médio dos laços ascendentes e descendentes.	40 escritores	93,00
Método Proposto (*GS*)	Inclinação axial, número de linhas, posição da margem inferior, ângulo médio dos laços ascendentes e descendentes.	100 escritores	84,00
Método Proposto (*GS*)	Inclinação axial, número de linhas, posição da margem inferior, ângulo médio dos laços ascendentes e descendentes.	160 escritores	75,63
Método Proposto (*GS*)	Inclinação axial, número de linhas, posição da margem inferior, ângulo médio dos laços ascendentes e descendentes.	200 escritores	70,50

Obviamente, quanto maior o número de escritores, mais consistentes precisam ser as abordagens de identificação, e mais discriminatórios precisam ser os conjuntos de características utilizados por estas abordagens.

Deve-se ressaltar que na Tabela 4.5 apenas foram incluídas abordagens de identificação de autoria em documentos manuscritos que utilizam características grafométricas, tal qual o método proposto. Também para facilitar a comparação foram incluídos resultados do método proposto em experimentos com diferentes números de escritores (40, 100, 160 e 200).

Pode-se observar que foram obtidos resultados muito significativos quando comparados aos da literatura, uma vez que para um grupo de 100 escritores foram alcançadas taxas de 84%; e para 200 escritores taxas de 70.5%. Se comparada a quantidade de escritores com a dos demais trabalhos analisados, apenas o trabalho de Pervouchine e Leedham (2007) utiliza um número de escritores da mesma magnitude que o apresentado nesta pesquisa. No entanto, no trabalho de Pervouchine e Leedham (2007) o resultado obtido é de 58%, que é inferior aos resultados do método proposto com 200 escritores (70.5%). A melhor taxa de acerto encontra-se no trabalho de Zois e Anastassopoulos (2000) com 95%, mas neste caso foi aplicada uma quantidade de 50 escritores, tal taxa é praticamente a mesma obtida nos experimentos (93%) com o método proposto para um grupo com um número de 40 escritores.

Os trabalhos que apresentam os melhores resultados: Zois e Anastassopoulos (2000) e Schlapbach e Bunke (2004), assim como a abordagem proposta, utilizam características relacionadas à inclinação (do texto e de palavras), isto vem ressaltar a relevância desta característica grafométrica para o processo de identificação da escrita humana (seja ela automática ou não).

Ao comparar a abordagem proposta, com resultados de abordagens que utilizam características não-grafométricas, como apresentado na Tabela 2.3, pode-se observar que os resultados obtidos apresentam-se superiores aos de muitos trabalhos, tais como: Bulacu et al. (2007), He et al. (2008), Siddiqi e Vicent (2008). Ao se relacionar os resultados com o número de escritores dos experimentos, os resultados obtidos apresentam-se ainda mais promissores, uma vez que em muitos destes estudos um número relativamente pequeno de escritores foi aplicado, como, por exemplo, em Said et al. (1998) no qual um taxa de identificação de 96% foi obtida, no entanto apenas 10 escritores participaram dos experimentos.

Cabe ressaltar que foram realizados experimentos com um número incremental de escritores (ou seja, para 20, 40, 60, etc.), sendo que isto permitiu uma avaliação de quando o método proposto atinge a estabilidade, ou seja, quando pouca ou nenhuma alteração nos resultados pode ser observada. Dessa forma, com um número de 200 escritores quase nenhum ganho ou perda nas taxas de identificação é percebido (ver Gráfico 3.1). Isto reduz o tamanho da amostra necessário para garantir a confiabilidade dos resultados do método proposto em experimentos futuros. Além disto, sabe-se que em situações reais, dificilmente os peritos se deparam com um número de suspeitos tão elevado.

4.6. Considerações Finais

Este capítulo apresentou uma análise detalhada dos resultados obtidos com os experimentos de validação do método proposto levando-se em consideração diferentes aspectos. Além disso, uma comparação dos resultados obtidos com outros trabalhos apresentados na literatura também foi realizada.

No próximo Capítulo são apresentadas as principais conclusões deste trabalho bem como sugestões de trabalhos futuros.

Capítulo 5

Conclusão

A escrita humana como elemento biométrico vendo sendo alvo de muitas pesquisas. Muitas destas pesquisas apresentam soluções computacionais para o problema de identificação de autoria em documentos manuscritos. Nesse contexto, esta pesquisa teve por objetivo a definição de um método computacional para identificação de autoria em documentos manuscritos utilizando características grafométricas.

O método proposto envolveu a definição de um conjunto de características a serem extraídas de documentos manuscritos, mais especificamente de cartas forenses modelo PUCPR e, posteriormente, o uso destas características para o processo de identificação de autoria propriamente dito.

Assim, foi estabelecido um protocolo experimental de modo que cenários de testes para prova de conceito pudessem ser realizados a fim de se avaliar e validar o método proposto. Finalmente, realizou-se a análise dos resultados obtidos sob a ótica da grafoscopia sem esquecer os aspectos científicos e computacionais.

Portanto, apresenta-se a seguir as principais conclusões com base nas análises dos resultados, das características grafométricas e, ainda, do número de escritores considerados durante os experimentos realizados. Seguem as conclusões:

I. **Análises individuais com o grupo de características (f_1-f_{12}):**

- as características globais apresentam uma maior relevância no processo de identificação de autoria, considerando-se o conjunto de características estudadas e implementadas. Em um segundo momento, as características locais, em nível de palavra e caractere, auxiliaram o método a refinar os resultados;

- a característica inclinação axial apresenta-se como a característica que individualmente e em grupo tem a maior relevância no processo de identificação;
- características globais e locais que dependem da "demarcação" de linhas, introduziram ruído nos dados extraídos, pois algumas particularidades da escrita de um escritor (tais como, "escrever para cima/baixo") não foram suportadas pelos algoritmos de extração aplicados.

II. **Análises com agrupamentos de características:**

- agrupamentos de características empíricos foram construídos (ver Tabela 3.3) e pode-se observar que em todos os grupos com resultados significativos a característica inclinação axial estava presente;
- o grupo empiricamente composto que apresentou o melhor resultado foi o mesmo que àquele resultante da aplicação de um processo formal de seleção de características. Este grupo foi chamado de *GS* e é composto pelas características: f_1 -,*Número total de linhas da carta,* f_6 ,- *Posição da margem inferior,* f_8 - *Inclinação axial e* f_{12} *-Ângulo geral dos laços ascendentes e descendentes*

III. **Análises do número de escritores:**

- os resultados produzidos por métodos de identificação de autoria dependem diretamente do número de escritores utilizados nos experimentos. Quanto maior o número de escritores mais consistentes e confiáveis precisam ser as características que fazem parte do método de identificação, uma vez que mais discriminatórias as mesmas precisam ser;
- experimentos com um número incremental de escritores (20, 40, 60, 80, 100, 120, 140, 160, 180 e 200) foram realizados e com 200 escritores as taxas de identificação obtidas apresentaram uma convergência assintótica, ou seja, praticamente nenhum aumento ou redução nestas taxas de acerto foi observado. Isto demonstra a confiabilidade e estabilidade do método proposto, uma vez, que para novos experimentos não serão necessários mais do que 200 escritores para atingir resultados confiáveis.

Com base no contexto apresentado, pode-se concluir que o Método Proposto nesta pesquisa, composto apenas por características grafométricas, atingiu resultados comparáveis aos apresentados na literatura e em muitos casos superiores, tanto com abordagens que apresentam apenas características grafométricas quanto com abordagens que não utilizam características grafométricas (ver Seção 4.5).

Assim, pode-se propor trabalhos futuros que envolvam a revisão do processo de extração de todas as características que introduziram ruído nos resultados. Em especial um tratamento para corrigir logicamente a inclinação axial das linhas das cartas forenses antes do procedimento de extração de primitivas. Deste modo, características como o tamanho das palavras, bem como, a altura das mesmas poderão ter maior relevância nas taxas de identificação.

Outro trabalho bastante promissor para o refinamento dos resultados é o estudo e a implementação de novas características grafométricas em nível de palavras e caracteres. Pode-se observar, com os experimentos realizados, que tais características auxiliam no refinamento dos resultados e é neste sentido que novos estudos devem ser conduzidos.

O presente trabalho coloca-se ainda como mais uma contribuição na área de Análise e Reconhecimento de Documentos tendo como principal objetivo auxiliar a análise forense de manuscritos, ressaltando-se que a fundamentação técnico-científica tem por base os mesmos fundamentos utilizados pelos peritos em análises forenses de documentos questionados.

Referências

ABUTALEB, A. S. Automatic thresholding of gray level pictures using two dimensional entropy. **Computers Graphics & Image Processing**, n. 47, p.22-32, 1989.

AMARAL, A. M. M. M.; FREITAS, C. O. A.; BORTOLOZZI, F. Identificação de Autoria Offline em Documentos Manuscritos. In: SIMPÓSIO BRASILEIRO DE SEGURANÇA DA INFORMAÇÃO E DE SISTEMAS COMPUTACIONAIS, 2012. **Anais do Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais**. Porto Alegre: SBC, p. 598-610, 2012a.

AMARAL, A. M. M. M.; FREITAS, C. O. A.; BORTOLOZZI, F. The Graphometry Applied to Writer Identification. In: INTERNATIONAL CONFERENCE ON IMAGE PROCESSING, COMPUTER VISION AND PATTERN RECOGNITION, 2012. **Proceedings of International Conference on Image Processing, Computer Vision and Pattern Recognition** EUA: CSREA Press, v.1. p. 351-356, 2012b.

AMARAL, A. M. M. M., FREITAS, O. A., BORTOLOZZI, F. Feature Selection for Forensic Handwriting Identification. In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 2013. **Proceedings of International Conference on Document Analysis and Recognition (ICDAR)**, p. 922-926, 2013a.

AMARAL, A. M. M. M.; FREITAS, C. O. A.; BORTOLOZZI, F. Multiple Graphometric Features for Writer Identification as part of Forensic Handwriting Analysis**.** In: INTERNATIONAL CONFERENCE ON IMAGE PROCESSING, COMPUTER VISION, AND PATTERN RECOGNITION, 2013**. Proceedings of International Conference on Image Processing, Computer Vision, and Pattern Recognition,** v. 1. p. 10-17, 2013b.

AMARAL, A. M. M. M.; FREITAS, C. O. A.; BORTOLOZZI, F. Combining multiple features based on graphometry for writer identification as part of Forensic Handwriting Analysis. In: INTERNATIONAL DOCUMENT IMAGE PROCESSING, 2013. **Proceedings of International Document Image Processing**. Patras-Greece: International Association for Pattern Recognition (IAPR), v.1. p. 23-30, 2013c.

BARANOSKI, F. **Verificação da autoria em documentos manuscritos usando SVM**. 2005, 106p. Dissertação (Mestrado em Ciência da Computação) - Pontifícia Universidade Católica do Paraná, Paraná.

BENSEFIA, A.; PAQUET, T.; HEUTTE, L. A writer identification and verification system. **Pattern Recognition Letters**, v.26, n.13, p.2080-2092, 2005.

BERWANGE, A. R.; LEAL, J. E. F. **Noções de paleografia e diplomática**. Santa Maria: Editora da Universidade Federal de Santa Maria, 3.ed. Revista e Ampliada. 2008, 124p.

BLANKERS, V.; NIELS, R.; VUURPIJL, L. Writer identification by means of explainable features: shapes of loops and lead-in strokes. In: BELGIAN-DUTCH CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2007**. Proceedings of Belgian-Dutch Conference on Artificial Intelligence**, p. 17-24, 2007.

BOUCKAERT, R. R.; FRANKE, E.; HALL, M.; KIRKBY, R.; REUTEMANN, P.; SEEWALD, A.; SCUSE, D. **WEKA manual for version 3-6-1**. University of Waikato, Hamilton, New Zealand, 2009. 212p.

BUI, Q. A.; VISANI, M.; PRUM S.; OGIER, J. M. Writer identification using TF-IDF for cursive handwritten word recognition. In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 2011**. Proceedings of International Conference on Document Analysis and Recognition (ICDAR),** p. 844-848, 2011.

BULACU, M.; SCHOMAKER, L.; BRINK, A. Text-independent writer identification and verification on offline Arabic handwriting. In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 2007**. Proceedings of International Conference on Document Analysis and Recognition (ICDAR)**, p.769-773, 2007.

BURGES, C. J. C. A tutorial on Support Vector Machines for pattern recognition. **Data Mining and Knowledge Discovery**, p.121-167, 1998.

CAVALCANTI, A.; LIRA, E. **Grafoscopia essencial**. Porto Alegre: Sagra Luzzatto, 1996. 151p.

CHEN, J.; LOPRESTI, D.; KAVALLIERATOU, E. The impact of ruling lines on writer identification. In: INTERNATIONAL CONFERENCE ON FRONTIERS IN HANDWRITING RECOGNITION, 2010**. Proceedings of International Conference on Frontiers in Handwriting Recognition**, p.439-444, 2010.

DASH, M.; LIU, H. Fetaure selection for classification**. Intelligent Data Analysis**, p. 131-156, 1997.

DUDA, R. O.; HART, P. E. **Pattern classification and scene Analysis**. USA: Wiley-Interscience, 1973.

FACON, J. **Morfologia matemática**: teoria e exemplos. Editora Universitária Champagnat da Pontifícia Católica do Paraná: Curitiba, 1996. 304p.

FERRARI, V.; FEVRIER, L.; JURIE, F.; SCHMID, C. Groups of adjacent contour segments for object detection. **IEEE Transaction on PAMI**, n. 30, p.36 - 51, 2008.

FRANKE, K.; SCHOMAKER, L.; VEENHUIS, C.; VUURPIJL, L.; VAN ERP, M.; GUYON, I. WANDA: a common ground for forensic handwriting examination and writer identification**. ENFHEX news - Bulletin of the European Network of Forensic Handwriting Experts**, n.1/04, p. 23-47, 2004.

FREITAS, C. O. A. **Uso de modelos escondidos de Markov para reconhecimento de palavras manuscritas**. 2001, 188p. Tese (Doutorado em Informática) - Pontifícia Universidade Católica do Paraná, Paraná.

FREITAS, C. O. A.; BORTOLOZZI, F.; SABOURIN, R. Study of perceptual similarity between different lexicons. **International Journal of Pattern Recognition and Artificial Intelligence**. v.18, n.7, p.1321-1338, 2004.

FREITAS, C. O. A.; OLIVEIRA, L. S.; BORTOLOZZI, F.; SABOURIN, R. Brazilian Forensic Letter Database. In: INTERNATIONAL WORKSHOP ON FRONTIERS ON HANDWRITING RECOGNITION, 2008**. Proceedings of International Workshop on Frontiers on Handwriting Recognition**, v.1, p.64-69, 2008.

FREITAS, C. O. A.; VOLPI NETO, A. Técnicas forenses nos crimes de falsidade documental: assinatura realizada por robô?. **Âmbito Jurídico**, Rio Grande, 40, 30/04/2007. Disponível em:<http://www.ambito-juridico.com.br/site/index.php?n_link=revista_artigos_leitura&artigo_id=4033>. Acesso em: 21/10/2010.

GROSICKI, E.; CARRÉ, M.; BRODIN, J. M.; GEOFFROIS, E. RIMES: evaluation campaign for handwritten mail processing. In: INTERNATIONAL CONFERENCE ON FRONTIERS ON HANDWRITING RECOGNITION, 2008. **Proceedings of International Conference on Frontiers on Handwriting Recognition**, 2008.

HALL, M. A. **Correlation-based feature subset selection for machine learning**. 1998, 178 p. Tese (Doctor of Philosophi) - University of Waikato, Hamilton New Zealand.

HANGAI, S.; YAMANAKA, S.; HANAMOTO, T. On-line signature verification based on altitude and direction of pen movement. In: INTERNATIONAL CONFERENCE ON MULTIMEDIA & EXPO, 2000. **Proceedings of International Conference on Multimedia & Expo**, v.1, p. 489–492, 2000.

HE, Z.; YOU, X.; TANG, Y. Writer identification of Chinese handrwriting documents using hidden markov tree model. **Pattern Recognition**, v.41, p.1295-1307, 2008.

HELLI, B.; MOGHADDAM, E. A text-independent Persian writer identification based on feature relation graph (FRG). **Pattern Recognition**, v.43, p.2199-2209, 2010.

HERTEL, C.; BUNKE, H. A set of novel features for writer identification. In: INTERNATIONAL CONFERENCE ON AUDIO-AND VIDEO-BASED BIOMETRIC PERSON AUTHENTICATION, 2003. **Proceedings of International Conference on Audio- and video-based Biometric Person Authentication**, p. 679-687, 2003.

HILTON, O. **Scientific examination of questioned documents**. New York: Elsevier, 1982. 436p.

JAIN, R.; DOERMANN, D. Offline writer Identification using K-adjacent segments. In: **Proceedings of International Conference on Document Analysis and Recognition,** p.769-773, 2011.

JIN, W.; WANG, Y.; TAN, T. Text-independent writer identification based on fusion of dynamic and static features. In: INTERNATIONAL WORKSHOP BIOMETRIC RECOGNITION SYSTEMS, 2005**. Proceedings of the International Workshop Biometric Recognition Systems**, p. 197-204, 2005.

JUSTINO, E. J. R. **O Grafismo e os modelos escondidos de Markov na verificação automática de assinaturas**. 2001, 131p. Tese (Doutorado em Ciência da Computação) - Pontifícia Universidade Católica do Paraná, Paraná.

KARUNAKARA, K.; MALLIKARJUNASWAMY, B. P. Writer Identification based on offline handwritten document images in Kannada language using empirical mode decomposition method. **International Journal on Computer Applications**, v.30, n.6, p.31-36, 2011.

KARTHIK, K.; AGRAWAL, R.; BHATTACHARYYA, C. A large margin approach for writer independent online handwriting classification. **Pattern Recognition Letters**, v.29, p. 933-937, 2008.

KOPPENHAVER, K. M. **Forensic document examination**: principles and practices. Humana Press, 2007. 315p.

KUMAR, J.; PRASAD, R.; CAO, H.; ALMAGEED, W. A.; DOERMANN; NATARAJAN, D. Shape codebook based handwritten and machine printed text zone extraction. In: INTERNATIONAL CONFERENCE ON DOCUMENT RECOGNITION AND RETRIEVAL, 2011. **Proceedings of International Conference on Document Recognition and Retrieval**, 2011.

LIMA, A. R. G. **Máquinas de vetores suporte na classificação de impressões digitais**. 2002, 81p. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Ceará, Fortaleza, Ceará.

LIU, H. YU, L. Toward integrating feature selection algorithms for classification and clustering, **IEEE Transaction on Knowledge and Data Engineering**, vol. 17, n. 4, p. 491-502, 2005.

LIU, H.; MOTODA. H. **Feature selection for knowledge discovery and data mining**. Boston: Klwer Academic Publishers, 1998. 214p.

LUNA, E. C. H.; RIVERON, E. M. F.; CALDERON, S. G. A supervisoned algorihm with a new differentiated-weighting scheme for identifying the author of a handwritten text. **Pattern Recognition Letters**, v.32, p. 1139-1144, 2011.

MENDES, L.B. **Documentoscopia**. Campinas: Millennium, 2003. 344 p.

MORRIS, R. N. **Forensic handwriting identification**: fundamental concepts and principles. San Diego, California: Academic Press, 2000. 238p.

NAMBOODIRI, A.M., GUPTA, S. Text independent writer identification from online handwriting. In: INTERNATIONAL WORKSHOP ON FRONTIERS IN HANDWRITING RECOGNITION, 2006**. Proceedings of International Workshop on Frontiers in Handwriting Recognition**, p. 23–26, 2006.

NOGUERAS, E. S.; ZANUY, M. F. Biometric recognition using online uppercase handwritten text. **Pattern Recognition**, v. 45, p.128-144, 2012.

OLIVEIRA, L. S.; SABOURIN, R. Support Vector Machines for handwritten numerical string recognition. In: INTERNATIONAL WORKSHOP ON FRONTIERS IN HANDWRITING RECOGNITION, 2004. **Proceedings of International Workshop on Frontiers in Handwriting Recognition**, p.39- 44, 2004.

OLIVEIRA, L.S.; JUSTINO, E.; FREITAS, C. O. de A.; SABOURIN, R. The graphology applied to signature verification**.** In: CONFERENCE OF THE INTERNATIONAL GRAPHONOMICS SOCIETY, 2005. **Proceedings of Conference of the International Graphonomics Society**, v.1,p.286-290, 2005.

OH, I.S; SUEN, C.Y.A class-modular feedforward neural network for handwriting recognition, **Pattern Recognition**, v.35, p. 229-244, 2002.

OSBORNE, A. **Questioned documents**. 2.ed. Albany, New York: Boyd Printing Company, 1929. 230p.

OSUNA, E.; FREUD, R.; GIROSI, F. Support Vector Machines: training and applications. **MIT Artificial Intelligence Memo 1602**; MIT A. I.Lab, 1997.

OTSU, N. **A** threshold selection method from gray-level histograms. **IEEE Transations Systems, Man. and Cybernetics, SMC 9**, v.1, p.63-66, 1979.

PALMER, A. N. **The Palmer method of business writing**. New York: The A. N. Palmer Company, 1935. 98p.

PELLAT, SOLANGE. **Le lois de l'écriture**. Paris: Libraire Vuibert, 1927. 63p.

PERVOUCHINE, V.; LEEDHAM, G. Extraction and analysis of forensic document examiner features used for writer identification. **Pattern Recognition**, v.40, p.1004-1013, 2007.

PHILLIP, M. F. **FISH: das Forensische Informations-System Handschriften des bundeskriminalamtes – eine analyse nach ¨uber 5 Jahren Wirkbetrieb** (Tech. Rep.). Wiesbaden, Germany: Kriminaltechnisches Institut 53, Bundeskriminalamt, 1996.

PIKASO SOFTWARE INC. **Write-On 1.0 for windows user manual.** Pikaso Software Inc.Canada, 2000. 34p.

PLAMONDON, R.; LORRETE, G. Automatic signature verification and writer identification: the state of the art. **Pattern Recognition**, v. 37, n.2, p. 107-131, 1989.

PLATT, J.; CRISTIANINI, N.; SHAWE-TAYLOR, J. Large margin DAGs for multiclass classification. **Advances in Neural Information Processing Systems**. MIT Press. p.547-553, 2000.

SAID, H.E.S.; PEAKE, G.S, TAN, T; BAKER, K. Writer identification from non-uniformly skewed handwriting images. In: BRITISH MACHINE VISION CONFERENCE, 1998**. Proceeding of British Machine Vision Conference**, p. 478-487, 1998.

SAID, H. E. S.; TAN, T.; BAKER, K. Personal identification based on handwritings. **Pattern Recognition**, v.33, n.1, p.149-160, 2000.

SAUDEK, R. **Experiments with handwriting.** Great Britain: George Allen & Unwin, 1978. 389p.

SCARPEL, R. A. Utilização de Support Vector Machine em previsão de insolvência de empresas. In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 2005. **Anais do Simpósio Brasileiro de Pesquisa Operacional** Gramado, RS, p. 671- 677, 2005.

SCHOMAKER, L.; FRANKE, K; BULACU, M. Using codebooks of fragmented connected-component contours in forensic and historic writer identification. **Pattern Recognition Letters**, v 28, p.719–727, 2007.

SCHLAPBACH, A.; BUNKE, H. Off-line handwriting identification using HMM based recognizers. In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, 2004. **Proceedings of International Conference on Pattern Recognition**, v.2, p. 654-658, 2004.

SIDDIQI, I.; VINCENT N. Combining global and local features for writer identification. In: INTERNATIONAL CONFERENCE ON FRONTIERS IN HANDWRITING RECOGNITION, 2008. **Proceedings of International Conference on Frontiers in Handwriting Recognition,** p.48-53, 2008.

SÖDERMAN, H.; O'CONNELL, J. **Manuel d'enquête criminelle modern**. trad. Jacques David. Paris: Payot, 1953. 452 p.

SREEJAJ, M.; IDICULA, S. M. A survey on writer identification schemas. **International Journal of Computer Applications**, v. 26, n. 2, july, p.23-33, 2011.

SRIHARI, S. N.; SHI, Z**.** Forensic handwritten document retrieval system: document image analysis for libraries. In: INTERNATIONAL WORKSHOP ON PUBLICATION DATE, 2004**. Proceedings of International Workshop on Publication Date**, p.188- 194, 2004.

STENBERG, S. R. Grayscale morphology. **Computer Vision Graphics Image Process,** n.35, p.333-355, 1986.

TAN, T. Rotation invariant texture features and their use in automatic script identification. **IEEE Transaction on Pattern Analysis and Machine Intelligence**, v.20, n.7, p. 751-756, 1998.

TAN, G. X.; GAUDIN, C.V.; KOT, A. C. Automatic writer identification framework for online handwritten documents using character prototypes. **Pattern Recognition**, v.42, p.3313-3323, 2009.

THUMWARIN, P., MATSUURA, T. On-line writer recognition for Thai based on velocity of barycenter of penpoint movement. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING, 2004. **Proceedings of IEEE International Conference on Image Processing**, p.889–892, 2004.

TSAI, L. M. Y. Online writer identification using the point distribution model. In: INTERNATIONAL CONFERENCE ON SYSTEM, MAN AND CYBERNETICS, 2005. **Proceedings of International Conference on System, Man and Cybernetics**, v.2, p 1264-268, 2005.

VAN ERP, M.; VUURPIJL, L.G.; FRANKE, K.; SCHOMAKER, L. R. B. The WANDA measurement tool for forensic document examination. In: INTERNATIONAL CONFERENCE OF THE GRAPHONOMIC SOCIETY, 2003**. Proceedings of International Conference of the Graphonomic Society (IGS)**, Scottsdale, Arizona, USA, p.2-5, 2003.

VAPNIK, V**. Estimation of dependences based on empirical data**. Nauka, Moscow, 1979. (English translation: Springer Verlag, New York, 1982). 505p.

VILLELA, C. A. X. Escrita escolar brasileira: a escrita inglesa. **Revista Língua Escrita**. n.7, julho-dezembro, p.6-23, 2009.

VOLPI NETO, A.; FREITAS, C.O.A. Escrever para deixar sua marca. **Information Management**. São Paulo: Editora Guia, n. 39, 2013.

ZANER-BLOSER. **Handwriting research web site**. Disponível em: <http://zaner-bloser.com>. Acesso em: 20 outubro 2010.

ZHU, Y.; TAN, T.; WANG, Y. Font recognition based on global texture analysis. **Transaction on Pattern Analysis and Machine Intelligence**, v. 23, n.10, p.1192-1200, 2001.

ZIMMERMANN, M.; BUNKE, M. Automatic segmentation of the IAM off-line database forhandwritten English text. In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, 2002. **Proceedings of International Conference on Pattern Recognition,** p.35-39, 2002.

ZOIS, E.; ANASTASSOPOULOS, V. Morphological waveform coding for writer identification. **Pattern Recognition**, v. 33, n.3, p. 385-398, 2000.

Apêndice A

Resultados de Experimentos Iniciais com KNN

A tabela abaixo apresenta os resultados de experimentos iniciais realizados no método proposto (ainda com um conjunto parcial de características) com o classificador KNN.

Tabela A.1. Resultados dos experimentos realizados com KNN

Grupo de Primitivas	**Taxa de Acerto (%)**
f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7& f_8	42,00
f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7	29,00
f_1	2,00
f_2	17,00
f_3	17,00
f_4	3,00
f_5	3,00
f_6	2,00
f_7	20,00
f_8	67,00
f_1 & f_8	72,00
f_2 & f_8	49,00
f_3 & f_8	33,00
f_4 & f_8	60,00
f_5 & f_8	60,00
f_6 & f_8	75,00
f_7& f_8	50,00
f_1 & f_6 & f_8	76,00
f_1 & f_6	9,00