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

Regiane K. Hanusiak¹,

Edson J. R. Justino¹,

Luiz S. Oliveira²,

Robert Sabourin³

Resumo: A grafoscopia é uma área da ciência forense dedicada, entre outros fins, à identificação e a verificação da autoria de manuscritos, sejam eles contemporâneos ou antigos. Por se tratar de uma área onde predomina a subjetividade no uso da técnica pericial ou grafometria, a mesma tem se tornado campo de interesse da computação, na busca de soluções para a padronização e auxílio de tais procedimentos. Neste contexto, este artigo apresenta um método para análise e identificação da autoria de manuscritos, tendo como base os atributos genéticos e genéricos da escrita, componentes fundamentais para a identificação da autoria de manuscritos. O método proposto faz uso de uma abordagem global, na extração dos atributos grafoscópicos e de identificação.

1 Pontifícia Universidade Católica do Paraná, PUCPR, Curitiba, PR, Brasil
{regiane.hanusiak, edson.justino@pucpr.br}

2 Universidade Federal do Paraná, UFPR, Curitiba, PR, Brasil
{lesoliveira@inf.ufpr.br}

3 Ecole de Technologie Superieure, Université Du Quebec, Canadá
{robert.sabourin@etsmtl.ca}

Abstract: The graphometry is an area of forensic science dedicated to manuscript analysis and writer identification. The manuscripts can be ancient or contemporary. In the actuality, the examiner analysis is done based on the expert's subjectivity and, in most of the cases, the result can be different depends on the expert's point of view. Based on that, the computer science has investigated this area with the main purpose of to develop solutions to help the examination procedures. In this context, this paper presents a method to analysis the manuscript and author identification. To this purpose we are using graphometric feature based on genetic and generic attributes. The purposed method is based on global approach for the feature extraction and verification procedure.

1 Introdução

Em Ciência Forense existe uma área que se destaca pela complexidade que apresenta no que se refere à verificação e identificação da autoria. Conhecida como grafoscopia, a mesma abrange, entre outros aspectos da escrita, a verificação e a identificação da autoria de textos manuscritos. A complexidade do tema, em boa parte, encontra-se na subjetividade dos peritos na aplicação das técnicas grafométricas, as quais se destinam à determinação da autoria de um manuscrito qualquer. Essa determinação é obtida através da análise de um conjunto de atributos grafoscópicos [1]. Os atributos grafoscópicos se apresentam em grande número, tais como ataques, pressão, remates, entre outros [1,2,3]. Do conjunto de atributos grafoscópicos duas classes se destacam, os genéticos e os genéricos [1]. Os atributos genéricos possuem elementos relacionados com a forma geral do texto. Entre outras características estão o calibre, o espaçamento entre palavras e o comportamento pauta. Os genéticos traduzem a gênese da escrita do autor, apresentando aspectos do traçado elaborados de forma inconsciente. Os mesmos são de difícil ocultação pelo escritor e de difícil reprodução por terceiros. Características como a pressão, progressão, ataques, remates, entre outros, são alguns dos destaques dessa classe. Cada classe de atributos pode então se enquadrada, em um exame pericial, como convergência ou divergência mínima, média ou máxima.

Do ponto de vista de um perito, o resultado da análise é decorrência da evidenciação de um conjunto mínimo de atributos grafoscópicos e dos respectivos graus de convergência, encontrados nos manuscritos de referência (de autoria conhecida) e no questionado (de

autoria desconhecida). No que se refere ao grau de confiabilidade dos atributos, tanto os genéticos como os genéricos são importantes na tomada de decisão durante o exame pericial.

Do ponto de vista computacional, a identificação da autoria de manuscritos representa um grande desafio, pois a implementação computacional das classes de atributos grafoscópicos pode apresentar elevada complexidade. É possível encontrar dois tipos de abordagens computacionais básicas para o processo de extração de características da escrita, as locais e as globais [4]. As locais geralmente fazem uso de processos de segmentação contextual [7,10,11], segmentando o texto em linhas, palavras, letras e segmentos de letras ou traços. O produto gerado pelo processo de segmentação é posteriormente submetido a diferentes processos de extração de características, as quais podem ou não estar associadas às classes de atributos grafoscópicos vistos anteriormente [5,6]. As abordagens que fazem uso de processo de segmentação contextual apresentam elevada complexidade e, podem não atender a todos os escritores. Existem casos em que o conteúdo de linhas adjacentes se sobrepõe (comportamento pauta - Figura 1a), a inclinação das mesmas é elevada e ocorre de forma irregular (comportamento base - Figura 1b), e o texto é ilegível, impedido a identificação dos pontos de conectividade entre letras (momentos) e os espaçamentos entre palavras (espaçamentos - Figura 1c). Em decorrência disso, a solução usualmente adotada nesses casos é a segmentação manual ou semi-automática.

As abordagens globais utilizam segmentações não-contextuais, isto é, sem levar em consideração o teor do texto manuscrito, em que fragmentos ou subáreas da imagem do manuscrito são submetidos ao processo de extração de característica [7,8]. Uma terceira abordagem, denominada mista, também pode ser encontrada sendo esta uma combinação das duas anteriores [9].

Figura 1. (a) Texto com conteúdo conexo; (b) Texto com alinhamento irregular; (c) Texto ilegível.

Dentre as abordagens globais, a produção de texturas a partir do texto manuscrito, vem se mostrando uma alternativa promissora na extração de características. A remoção da predominância dos espaços em branco existentes entre linhas e palavras é o procedimento mais utilizado nesses casos. Em outras palavras, técnicas de segmentação contextuais são utilizadas para produzir imagens que serão analisadas de maneira global.

Said et al [8] apresentam uma abordagem contextual para a identificação da autoria de manuscritos. Inicialmente, é realizada uma normalização dos manuscritos removendo a inclinação das linhas, a normalização do tamanho das letras e a eliminação de espaços em branco entre palavras e linhas. Quando uma linha não é totalmente preenchida com texto, informações redundantes são incluídas. As características são extraídas diretamente dos valores das matrizes de co-ocorrência, obtida da imagem de textura gerada e pelos filtros de Gabor. Os autores reportam taxas de reconhecimento em torno de 96%. Busch et al [10] apresentam um método para a identificação de textos impressos em diferentes idiomas. O texto impresso é inicialmente submetido a um processo de correção da inclinação das linhas, a fim de eliminar imperfeições do processo de digitalização do documento. Em seguida, é feita a segmentação das linhas, com o objetivo de reduzir os espaços em branco existente entre as mesmas e entre palavras. Quando a linha apresenta pouco texto, assim como em Said, o restante da mesma é incrementado com conteúdo redundante obtido na mesma linha. Com base na textura gerada, várias abordagens para a extração de características são utilizadas, entre elas, descritores de textura, filtros de Gabor e *Wavelets*. Os índices de acerto ficaram entre 91% a 97% sendo que os melhores resultados foram alcançados com *Wavelets* e filtros de Gabor. Al-Dmour e Zitar [11] apresentam um método para a identificação de manuscritos arábicos. A textura é obtida seguindo um protocolo similar aos vistos anteriormente, isto é, corrigindo a inclinação das linhas de texto e retirando espaços entre linhas e entre palavras. Na extração das características foram utilizadas as matrizes de co-ocorrência, filtros de Gabor e as transformas de Fourier. Os resultados obtidos ficam em torno dos 90% de acerto. O melhor desempenho, nesse caso, ficou com as transformadas de Fourier.

Este artigo apresenta um método global de análise dos atributos genéticos e genéricos da escrita com base em imagens de textura utilizando uma abordagem de segmentação não-contextual. Isso propicia a análise de manuscritos em que o teor do texto não se apresenta legível e, portanto, de difícil segmentação (Figura 1). O processo proposto não exige correções na inclinação das linhas de texto e dispensa os processos de correção dos alinhamentos entre palavras de uma mesma linha. Das imagens de textura geradas, são calculadas as matrizes de co-ocorrência, as quais são utilizadas em conjunto com os descritores de textura de Haralick [12]. Os descritores de Haralick foram escolhidos devido às propriedades que os mesmos possuem no trato de textura onde a irregularidade dos padrões é predominante.

O protocolo de aprendizado e identificação se baseia nos procedimentos de análise grafoscópica, utilizando para tal o SVM [14] (*Support Vector Machine*) e uma abordagem independente de autor. Os resultados experimentais com taxas de acerto em torno de 95% em uma base de dados composta por mais de 300 escritores demonstram que o método de

segmentação proposto é capaz de manter as características genéricas e genéticas dos autores. O trabalho ainda faz uma análise da importância da quantidade de autores usados para o treinamento. Nesse trabalho se exclui a presença de falsificações.

2 Perícia grafoscópica em manuscritos

Os peritos classificam os manuscritos, em relação à autoria, como associação ou dissociação [2]. A associação indica que a grafia presente no manuscrito foi elaborada, de próprio punho, pelo escritor avaliado. A dissociação indica que o manuscrito não foi elaborado, de próprio punho, pelo escritor avaliado.

Durante a prova pericial, o perito utiliza um conjunto n de amostras de texto de autoria conhecida ou referência R , em comparação com a amostra de autoria desconhecida ou questionada Q . O perito observa um conjunto de atributos grafoscópicos nas n amostras de referência e na questionada, estabelece um índice de similaridade S entre elas e posteriormente, toma uma decisão D em função de S . O laudo pericial resultante depende da soma dos resultados obtidos das comparações individuais dos pares (referência/questionada), Figura 2.

Figura 2. Diagrama esquemático da análise pericial grafoscópica.

3 Atributos genéticos e genéricos

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”. Os atributos genéticos e genéricos podem ser considerados complementares. Na Tabela 1 é possível observar exemplos de atributos genéticos e genéricos e na Tabela 2, é possível ver ambas as classes e seus atributos relacionados.

Tabela 1. Exemplos de tributos genéticos e genéricos.

Remates	Momentos	Calibre

Inclinação	Valores angulares/curvilíneos	Proporcionalidade

Desenvolvimento	Ataque	Mínimos gráficos

O calibre determina a escala geral da escrita, tal como o tamanho das letras. O espaçamento determina o espaço médio entre palavras. O comportamento pauta determina a distância média do texto escrito, em relação à linha de base ou pauta. O comportamento base determina o ângulo médio de inclinação das linhas de texto, para cima ou para baixo, de uma linha de base imaginária. O comportamento base é enfatizado quando o texto é redigido em papel sem pauta. Os valores angulares e curvilíneos estabelecem o comportamento angular nos pontos de mudança de direção do traçado, tais como nas letras “L”, “M”, “G”, entre outras. Nesses casos, o traçado pode apresentar uma curvatura suave ou uma mudança brusca de direção. Os momentos determinam os pontos de interrupção do traçado, no decorrer da escrita de uma palavra. Estes estão presentes, usualmente, na conexão entre letras. A proporcionalidade determina o grau de variabilidade entre ascendente, descendentes e corpo das letras.

A pressão determina as variabilidades da força que o objeto de escrita exerce sobre o papel, durante a evolução do traçado. A progressão determina a velocidade imposta pelo escritor, ao objeto de escrita, durante o transcorrer da escrita. O ataque determina a forma do traçado no início de uma palavra. O remate determina a forma do traço na finalização de uma palavra. O desenvolvimento determina o grau de destreza do escritor na execução do traçado.

A inclinação axial determina o ângulo médio de inclinação da escrita. Os mínimos gráficos determinam as formas e angulações das acentuações e pontuações presentes no texto manuscrito.

Tabela2. Atributos genéticos e genéricos.

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

Como visto anteriormente, do ponto de vista computacional, o processo de segmentação e de extração das classes de características genéticas e genéricas é de elevada complexidade, tendo em vista restrições impostas pela segmentação do texto, como visto na Figura 1, e, por conseguinte, da extração das características. No entanto, a adoção de um método que aproveite as propriedades inerentes das duas classes de características, tanto no processo de segmentação como no de extração das características, podem favorecer a criação de um modelo computacional robusto e que atenda, de forma satisfatória, às expectativas requeridas em uma análise pericial.

4 Dissimilaridade na identificação da autoria de manuscritos

Os métodos computacionais destinados à identificação da autoria de manuscritos baseiam-se, usualmente, em duas abordagens, a abordagem independente do escritor ou WI (*Writer-Independent*) e a abordagem dependente do escritor ou WD (*Writer-Dependent*) [13]. A abordagem WD utiliza um modelo por escritor, enquanto que a outra faz uso de um modelo geral para todos os escritores. A abordagem WD possui a vantagem de modelar adequadamente os atributos genéticos e genéricos do escritor e como desvantagem, exige um conjunto elevado de exemplares de manuscritos, na geração de um modelo pessoal robusto. A abordagem WI possui a desvantagem da generalização. No entanto, possui a vantagem de necessitar um número reduzido de exemplares de manuscritos para cada escritor e de não necessitar de um novo treinamento do modelo, diante da análise de manuscritos de outros autores, que não tenha participado da geração do modelo inicial.

No treinamento do modelo WI, a classe w_1 representa a classe de manuscritos do mesmo escritor. A classe w_2 representa o conjunto de manuscritos de escritores distintos. Na

identificação, o modelo gerado é então utilizado para a comparação com os manuscritos de escritores desconhecidos. Os vetores de características são extraídos das imagens de referência e das imagens questionadas. O vetor de dissimilaridade, que nada mais é do que o módulo das diferenças entre os componentes dos vetores é calculado e usado para treinar o classificador. A idéia por traz disso é que se as amostras de referência e questionada pertencerem aos mesmos autores, o vetor de dissimilaridade terá componentes próximo a zero, caso contrário, os componentes serão bem maiores que zero. Mais detalhes sobre vetores de dissimilaridade estão disponíveis em [13].

5 Método proposto

O método proposto é baseado na abordagem de análise utilizada pela perícia grafoscópica (Figura 2). Em termos computacionais, vetores de dissimilaridade extraídos das imagens de textura são utilizados para treinar SVMs, os quais são utilizados como classificadores neste trabalho. Diferentes *kernels* foram testados e os que apresentaram melhores resultados foram os *kernels* linear e Gaussiano. O treinamento foi realizado validação cruzada (*5-fold cross-validation*). Os parâmetros C e γ foram determinados através de um *grid search*. A seguir são apresentados mais detalhes sobre o método proposto.

5.1 Segmentação do manuscrito com base nos atributos genéticos e genéricos

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

O processo de segmentação ocorre primeiramente, pela geração de uma imagem binária do texto manuscrito. Em seguida, percorre-se a imagem binária de cima para baixo e da esquerda para a direita, até que um componente de traço (pixel preto) seja encontrado. Retira-se da imagem binária todos os pixels conexos ao pixel encontrado. Transfere-se da imagem em níveis de cinza, todos os pixels obtidos pela imagem binária e os rearranja em uma nova área, usando com linha de base ou pauta, o ponto médio do conjunto extraído. Em seguida, retira-se o conjunto de pixels da imagem binária e busca-se por uma nova ocorrência, repetindo-se o passo inicial. A distância entre as linhas de base ou pauta é

determinada pela média das alturas encontradas nos segmentos da linha corrente. Dessa maneira reduz-se a presença de espaços em branco entre linhas. Portanto, a grafia que apresentar maior ocorrência de ascendentes e descendentes (proporcionalidade baixa), apresentara igualmente uma maior ocorrência de sobreposições, Figura 3.

Fig. 3: Um exemplo do processo de segmentação; (a) Exemplo de uma carta da base PUCPR; (b) Exemplo da carta segmentada.

A Tabela 3 mostra que os atributos genéricos e genéticos são preservados no método de segmentação proposto. Na Tabela 3a é possível observar variação no calibre para diferentes escritores. Na Tabela 3b se observam às mudanças de direção do traçado, referentes aos valores curvilíneos e angulares para escritores distintos. Na Tabela 3c se observa os diferentes graus de variabilidade entre ascendente, descendentes e corpo das letras. Na Tabela 3d se observa a diferença de pressão entre escritores distintos. Essa característica em particular, possui uma forte dependência do objeto de escrita e da textura do papel utilizado. Na Tabela 3e se observa o nível de progressão da escrita. A progressão estabelece que, quanto mais veloz for a execução da escrita, mais ilegível ela se tornar e quanto mais lento, mais legível se torna o texto. Na Tabela 3f se observam diferentes características relativas ao ataque e ao remate.

No primeiro caso, o escritor apresenta um traçado com início e final usualmente espessos, enquanto no outro, são ambos usualmente pontiagudos. Na Tabela 3g é possível observar o grau de destreza ou desenvolvimento apresentado por escritores distintos. O escritor com maior destreza apresentara trações firmes e regulares, já o com baixa destreza, apresentará um traçado heterogêneo e vacilante. Na Tabela 3h se observa as diferenças de inclinação axial entre textos de escritores diferentes. Essa angulação pode oscilar para mais

ou para menos no transcorrer da escrita, no entanto, sempre haverá uma predominância angular. Por fim, os mínimos gráficos, por serem componentes isolados, acabam sendo incorporados pelo processo de segmentação não contextual, juntamente com os momentos, confira na Tabela 3. Como se pode observar, os atributos genéricos e genéticos estão bem representados nas imagens de textura produzidas pelo método proposto.

Tabela 3. Exemplos dos atributos genéticos e genéricos observáveis após o processo de segmentação.

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

5.2 Extração de características

Os descritores propostos por Haralick [12] são descritores estatísticos de textura cujas propriedades buscam atender aos casos em que as mesmas apresentam padrões irregulares. O conjunto é composto por 13 (treze) descritores que buscam determinar estatisticamente as propriedades de distribuição e o relacionamento entre tons de cinza contidos em uma textura. Os descritores são calculados através do uso de matrizes de co-ocorrência P , obtidas da imagem da textura I em análise. Nesse estudo foram utilizados 4 (quatro) diferentes resoluções, 2, 4, 8 e 16, como o objetivo de estabelecer um comparativo entre resultados, dado o acréscimo de conteúdo proporcionado pelos diferentes níveis de cinza. Foram igualmente utilizadas 5 (cinco) distâncias para os pixels vizinhos ($d = 1, 2, 3, 4, 5$) e quatro angulações ($\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ$), sendo esses os valores mais utilizados pela literatura [6,10]. Do conjunto de treze descritores foram selecionados os sete que obtiveram os melhores resultados nos testes individuais preliminares. São eles: 2º. momento angular,

entropia, homogeneidade, dissimilaridade, variância inversa, energia e correlação. Veja a Tabela 4.

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

Tabela 4. Descritores de Haralick.

Descritor	Equação
2º. momento angular	$\sum_{i=0}^n \sum_{j=0}^m (p(i,j))^2 \quad (1)$
Entropia	$-\sum_{i=0}^n \sum_{j=0}^m p(i,j) \cdot \log(p(i,j)) \quad (2)$
Homogeneidade	$\sum_{i=0}^n \sum_{j=0}^m p(i,j) / (1 + i - j) \quad (3)$
Dissimilaridade	$\sum_{i=0}^n \sum_{j=0}^m p(i,j) \cdot i - j \quad (4)$
Variância Inversa	$\sum_{i=0}^n \sum_{j=0}^m p(i,j) / (i - j)^2 \quad i \neq j \quad (5)$
Energia	$\sqrt{\sum_{i=0}^n \sum_{j=0}^m (p(i,j))^2} \quad (6)$
Correlação	$\sum_{i=0}^n \sum_{j=0}^m p(i,j) ((i - \mu_i)(j - \mu_j) / \sqrt{\sigma_i^2 \sigma_j^2}) \quad (7)$

Os vetores de características $\mathbf{V}$, para cada descritor, utilizados nas fases de aprendizado e identificação, foram obtidos através das matrizes de co-ocorrências, com a variação das distâncias d e dos ângulos θ . Isto é, cada vetor de característica dos descritores é composto por 20 componentes normalizados entre zero e um.

5.3. Comparação

O processo de comparação é composto por duas fases, o treinamento e a identificação. No estágio de treinamento, as dissimilaridades entre os vetores de características são

calculadas entre pares de fragmentos de textos. Quando dois fragmentos pertencerem a um mesmo escritor, o vetor de característica é indicado com 1 (associação). Quando dois fragmentos pertencerem a escritores diferentes, o vetor de característica é indicado com -1 (dissociação). A dissimilaridade entre dois fragmentos de texto é considerada pequena, quando as amostras pertencerem a um mesmo escritor. O SVM é treinado então para separar pequenas dissimilaridades entre características (associação) e grandes dissimilaridades entre características (dissociação). É importante mencionar que os SVMs foram treinados para gerar estimações de probabilidade como saída. Desta maneira diferentes regras de combinação podem ser avaliadas.

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

5.5. Decisão

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

Com o objetivo de gerar a decisão final, o método proposto classifica as saídas do SVM utilizando como regra de fusão. A literatura apresenta várias possibilidades para o processo de fusão, tais como a média dos valores, o valor máximo, o valor mínimo, a soma dos valores e o voto majoritário [13]. Os experimentos demonstraram que nesse caso, os melhores resultados foram obtidos através da soma.

6. Base de Manuscritos

A base de manuscritos utilizada é composta por 315 escritores. Cada escritor redigiu três manuscritos de mesmo teor, totalizando 945 documentos. O conteúdo da carta foi elaborado de forma a contemplar o conjunto de letras do alfabeto da Língua Portuguesa, tanto minúscula quanto minúscula. Assim também procedendo com os símbolos utilizados nas acentuações, tais como o til, cedilha, acento circunflexo, acento grave, acento agudo e pingos nos *i*'s. A base foi dividida em dois conjuntos, a fim de permitir o treinamento dos modelos e a execução dos testes sem que os escritores de um conjunto participassem do outro. Assim procedendo, ficaram 115 autores para testes e 200 para o treinamento dos modelos. O conjunto de 200 autores para o treinamento foi novamente dividido em três

subconjuntos, cujo número de autores foi incrementado segundo um critério exponencial, isto é, 50, 100, 200 escritores. Este critério foi adotado, a fim de estabelecer a influência do número de autores no desempenho do modelo, por se tratar de um modelo WI. Os resultados demonstraram uma variação em torno de 1% de incremento do erro. Por essa razão, adotou-se o conjunto de 50 escritores, o primeiro, na apresentação dos resultados.

7. Resultados Experimentais

Num primeiro momento foram testados isoladamente cada descritor. Nesse caso, foram utilizados os fragmentos com 16 níveis de cinza, 50 escritores para o treinamento do modelo e 115 escritores para os testes. Na etapa de decisão foi utilizada a soma como regra de fusão dos valores resultantes do SVM. Os resultados obtidos estão na Tabela 5. É possível observar que a entropia, que descreve o grau de dispersão de níveis de cinza da textura, obteve o melhor resultado 93,6% de acerto, seguida da energia, que mede a ordenação da textura, com 92,6% de acerto. A proximidade dos resultados entre estes dois descritores já era esperado, uma vez que ambos são considerados características mutuamente dependentes.

Tabela 5. Resultados obtidos com os descritores individualmente, com as imagens dos fragmentos em 16 níveis de cinza.

Descritor	Erro tipo I	Erro tipo II	Erro Médio (%)
	Falsa rejeição (%)	Falsa aceitação (%)	
2º. momento angular	5,57	9,39	7,48
Entropia	4,00	8,70	6,35
Homogeneidade	5,74	9,22	7,48
Dissimilaridade	12,70	41,91	27,30
Energia	6,00	8,70	7,39
Variância inversa	7,48	24,87	16,17
Correlação	11,48	10,96	11,22

Por se tratar de um modelo global, outro teste foi elaborado a fim de determinar a estabilidade do modelo, dado a variação do número de escritores no treinamento do modelo. Este procedimento busca identificar a contribuição que o número de escritores pode oferecer ao modelo gerado. Para esse teste foram usados três conjuntos de treinamento com 50, 100 e 200 escritores, mantido o mesmo conjunto de 115 escritores nos testes, usados dois níveis de cinza e a regra de fusão pela soma. Os resultados estão descritos na Tabela 6. Com o acréscimo do número de amostras de treinamento, observou-se uma pequena variação, em torno de 1%, nas taxas de erro, demonstrado que isoladamente, os descritores apresentam um comportamento estável. O classificador treinado com os descritores de correlação não convergiu durante o treinamento.

Tabela 6. Resultados obtidos com a variação no número de escritores na geração do modelo.

Descritor	Média dos erros (%)		
Número de escritores no treinamento do modelo	50	100	200
2º. momento angular	5,39	4,70	5,48
Entropia	5,13	5,04	5,91
Homogeneidade	6,69	5,91	6,61
Dissimilaridade	5,91	6,00	5,56
Energia	5,47	4,69	5,21
Variância inversa	5,91	6,00	5,56
Correlação	-	-	-

A fim de se averiguar a influência da quantidade de informação gerada pelos níveis de cinza, na produção das matrizes de co-ocorrência, repetiram-se os testes, agora variando os níveis de cinza dos fragmentos em 2, 4, 8 e 16 níveis. Para tanto, foram utilizados 50 escritores no treinamento e 115 para os testes. Os resultados podem ser vistos na Tabela 7. Os mesmos demonstram que a redução no número de níveis de cinza produz uma redução nos erros em relação aos produzidos com 4, 8 e 16 níveis. Isto é decorrência da redução da complexidade da imagem, no que se refere aos níveis de cinza, e da consequente redução da dispersão dos valores na matriz de co-ocorrência.

Tabela 7. Resultados obtidos com a variação nos níveis de cinza dos fragmentos de manuscritos.

Descritor	Média dos erros (%)			
Níveis de cinza	2	4	8	16
2º. momento angular	5,39	6,69	6,78	7,48
Entropia	5,13	5,82	6,08	6,35
Homogeneidade	6,69	7,22	8,43	7,48
Dissimilaridade	5,91	6,96	27,04	27,30
Energia	5,47	6,61	6,69	7,39
Variância inversa	5,91	6,87	19,21	16,17
Correlação	-	11,74	13,04	11,22

Na Tabela 8 é possível observar estabilidade do modelo mesmo com a variação do *kernel*. Neste caso estão presentes os *kernel*s linear e o Gaussiano.

Tabela 8. Resultados obtidos com a combinação dos classificadores.

Descritor combinados	Erro tipo I		Erro tipo II		Média dos erros (%)	
	Falsa Rejeição (%)		Falsa Aceitação (%)			
	linear	gaussiano	linear	gaussiano	linear	gaussiano
2º. momento angular	3,48	4,00	7,30	7,82	5,39	5,91
Entropia	2,43	8,17	7,83	5,04	5,13	6,60
Homogeneidade	4,52	8,17	8,87	3,13	6,69	5,65
Dissimilaridade	3,30	6,60	8,52	5,21	5,91	5,91
Energia	3,30	4,00	7,65	8,00	5,47	6,00
Variância inversa	3,30	6,60	8,52	5,21	5,91	5,91
Correlação	-	24,86	-	12,17	-	18,52

A Figura 4 mostra a curva ROC (*Receiver Operating Characteristic*) gerada pelos descritores isoladamente, obtidos com os *kernels*, linear e o Gaussiano. No caso do kernel Gaussiano (Figura 4b) pode-se observar que o classificador baseado em Homogeneidade domina todos os outros classificadores para quase todos os pontos de operação da curva. O mesmo não acontece com os classificadores treinados com o kernel linear. Nesse caso, diferentes classificadores podem ser escolhidos em função de diferentes pontos de operação.

Isso nos levou a realizar um último experimento combinando os classificadores treinados com o kernel linear. Para isso foram testadas diferentes regras de fusão de classificadores e os melhores resultados foram conseguidos com a fusão das características baseadas no momento angular, entropia, dissimilaridade/variância inversa. A melhor regra de fusão nesses casos foi a soma. Os resultados da fusão estão descritos na Tabela 9.

Fig. 4: Curvas ROC geradas com o uso do *kernel* linear (a) para os descritores de Haralik.

Tabela 9. Resultados obtidos pela fusão dos resultados obtidos pelos 6 diferentes descritores.

Regra de fusão	Descritor		Média dos erros (%)		
	Voto	Soma	Máximo	Mínimo	Média
homogeneidade+ variância inversa+ dissimilaridade	6,43	5,91	10,61	10,43	8,00
2º. momento angular + entropia+energia	5,47	5,30	10,95	10,08	7,56
2º. momento angular + entropia+dissimilaridade	5,21	4,60	11,22	11,04	8,34
2º. momento angular + entropia+variância inversa	5,21	4,60	11,22	11,04	8,34

7. Conclusão e trabalhos futuros

O objetivo desse artigo foi apresentar um método para identificação de autoria de manuscritos baseado nos atributos genéticos e genéricos da escrita. Para esse propósito, foram utilizadas apenas duas classes, autoria e não autoria. O modelo WI (*Writer-Independent*) apresentado mostrou-se robusto na variação do número de autores de treinamento. O uso das características genéticas e genéricas, obtidos através do processo de segmentação não-contextual, mostrou-se promissor, com a obtenção de taxa de acerto da ordem de 95%. Os resultados obtidos assemelhassem ao obtidos por outros autores que utilizaram a textura como abordagem. No entanto, observou-se igualmente que os descritores de Haralick não apresentaram melhora significativa no desempenho, quando do acréscimo da base de autores para o treinamento, aproximadamente 1%, assim como no uso de diferentes *kernels* no SVM e nas combinações entre os descritores, através de diferentes mecanismos de fusão.

Como proposta para trabalhos futuros encontra-se a inclusão de outros descritores que permitem a extração de características de textura, tais como as *Wavelats* e filtros de Gabor. Outra proposta é testar uma abordagem de classificação diferenciada através de métodos de compressão de dados, aplicada em classificação de texturas. O objetivo é melhorar o processo de separação entre as classes autoria e não autoria.

Referências

- [1] Gobineau H. and Perron R., *Géénétique de l'écriture et étude de la personnalité: Essais de graphometrie*, Delachaux & Niestlée, (1954).
- [2] Morris, N.; *Forensic Handwriting Identification Fundamental Concepts and Principles*, Academic Press, p. 238, (2000).
- [3] Huder, R. A.; Headrick A. M. , *Handwriting Identification: Facts and Fundament*, CRC Press, New York, ISSN 0-8493-1285-X, pp. 434, (1999)
- [4] Shomaker, L.; *Advances in Writer Identification and Verification*, Proceedings of the International Conference on Document Analysis and Recognition (ICDAR) Curitiba, Brazil. (2007).

- [5] Cha, S. H., *Use of the Distance Measures in Handwriting Analysis*. Doctor Theses. State University of New York at Buffalo, EUA, p. 208, (2001).
- [6] Pervouchine, V.; Leedham, G.; *Study of Structural Features of Handwritten Grapheme 'th' for Writer Identification*, Proceedings of the Third International Symposium on Information Assurance and Security, Manchester, United Kingdom, (2007).
- [7] Srihari, S. N.; Cha, S. H.; Hina A.; Sangjilk, L.. *Individuality of Handwriting*, Journal of Forensic Sciences, (2002).
- [8] Said, H.; Peake, G.; Tan, T.; Baker, K.; *Personal Identification Based on Handwriting*, Patter Recognition, no. 33, p. 149-160, (1999).
- [9] Bulacu, M. , Shomaker, L. *Writer Identification Using Edge-Based Directional Features*, Proc. Of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003), IEEE Computer Society, Edinburgh, Scotland, vol. II, pp. 937-94, (1993).
- [10] Bush, A.; Boles, W.; Sridharan, S.; *Texture for Script Identification*, IEEE Transations on Pattern Analysis and Machine Inteligence, Vol. 27, no. 11, November, 1721-1732p., (2005).
- [11] Al-Dmour, A.; Zitar, R. A.; *Arabic writer identification based on hybrid spectral-statistical measures*, Journal of Experimental & Theoretical Artificial Intelligence, Volume 19, Number 4, December , 307-332p., (2007).
- [12] Haralick, R. M.; Shanmugam, K.; Dinstein, I.; *Textural Features for Image Classification*, IEEE Transactions on Systems, and Cybernetics, vol. smc-3, no. 6, november, (1973).
- [13] Bertolini, D., Oliveira, L. S., Justino, E., Sabourin, R., *Ensemble of Classifiers for Off-line Signature Verification*. IEEE International Conference on System, Man and Cybernetics (SMC 2008), 283-288p., Singapore, (2008).
- [14] Vapnik, V. *Statistical Learning Theory*, Wiley, N. Y, (1998).
- [15] Burges, C. J. C., *A Tutorial on Support Vector Machines for Pattern Recognition*, Data Mining and Knowledge Discovery 2, 121-167p, (1998).