

# Identificação da Autoria em Documentos Manuscritos Usando SVM

Francis L. Baranoski, Edson J. R. Justino, Flávio Bortolozzi

Pontifícia Universidade Católica do Paraná – Rua Imaculada Conceição,  
1515, CEP 80215-901, Curitiba, PR, Brasil  
Laboratório de Ciência Forense

francis@ppgia.pucpr.br,  
{edson.justino, flavio.bortolozzi}@pucpr.br

***Abstract.** The author's manuscript identification based on handwriting patterns has been used in forensic science to assist a criminal solution and the suspect identification. The manuscript pattern database and an efficient search process can help to identify an unknown author. This paper shows a global approach to identify the manuscript's author based on graphometric features and a (SVM) Support Vector Machines classifier.*

***Resumo.** A identificação da autoria de um manuscrito, através de bancos de dados de padrões gráficos, tem sido utilizada em ciência forense, para auxiliar na solução de crimes e na identificação de suspeitos. As bases de padrões gráficos, juntamente com um processo de busca eficiente, podem ajudar a localizar e identificar suspeitos com antecedentes criminais e também, auxiliar na solução de crimes de autoria desconhecida. Esse artigo apresenta um método global de identificação da autoria de manuscritos, com base em características grafotécnicas e num classificador (SVM) Support Vector Machines.*

## 1. Introdução

A grafoscopia tradicional é o campo da Ciência Forense destinada a buscar respostas para as questões judiciais associadas a documentos manuscritos (criminal, cível). Distintamente da documentoscopia, a grafoscopia visa tratar unicamente dos aspectos da escrita e sua autoria, não abordando os diferentes tipos de documentos ou materiais de suporte onde o manuscrito foi apostado, Morris (2000) e Dines (1998).

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

No contexto da grafoscopia, dois objetos de análise se apresentam, os manuscritos e as assinaturas. Mesmo possuindo características distintas, ambos mantêm uma estreita relação entre si, possuindo a mesma raiz ou origem no processo de aprendizado do escritor. Isto é, carregam consigo as experiências adquiridas pelo escritor, durante o seu processo de aprendizado e posteriormente, através do aperfeiçoamento do estilo pessoal de escrita, Santos et al. (2004).

Para a grafoscopia são relevantes dois elementos de análise, o grafostático e o grafocinético, Morris (2000), Dines (1998) e Justino (2002). O primeiro aborda critérios mais globais de análise, tais como a altura, comprimento e forma. O segundo aborda elementos dinâmicos do traçado, tais como inclinação axial, pontos de ataque e remates.

## 2. Perícia Grafotécnica em Manuscritos

Os peritos grafotécnicos classificam os textos manuscritos, em relação à autoria, como associação ou dissociação, Justino (2002). A associação indica que a grafia presente no manuscrito foi elaborada, de próprio punho, pelo autor avaliado. A dissociação indica que o manuscrito não foi elaborado, de próprio punho, pelo autor avaliado.

Durante a prova pericial, o perito utiliza um conjunto  $n$  de amostras de texto de autoria conhecida (referência)  $M_{ki}$  ( $i=1,2,3,\dots,n$ ), em comparação com a amostra de autoria desconhecida (questionada)  $M_Q$ . O perito observa, tendo como base características grafotécnicas  $f_{Vki}$  ( $i=1,2,3,\dots,n$ ) e  $f_{VQ}$ , diferenças de medição entre as amostras conhecidas e a desconhecida  $D_i$  ( $i=1,2,3,\dots,n$ ) e posteriormente, toma uma decisão  $R_i$  ( $i=1,2,3,\dots,n$ ). O laudo pericial resultante  $D$  depende da soma dos resultados obtidos das comparações individuais dos pares (referência / questionada), (Figura 1).

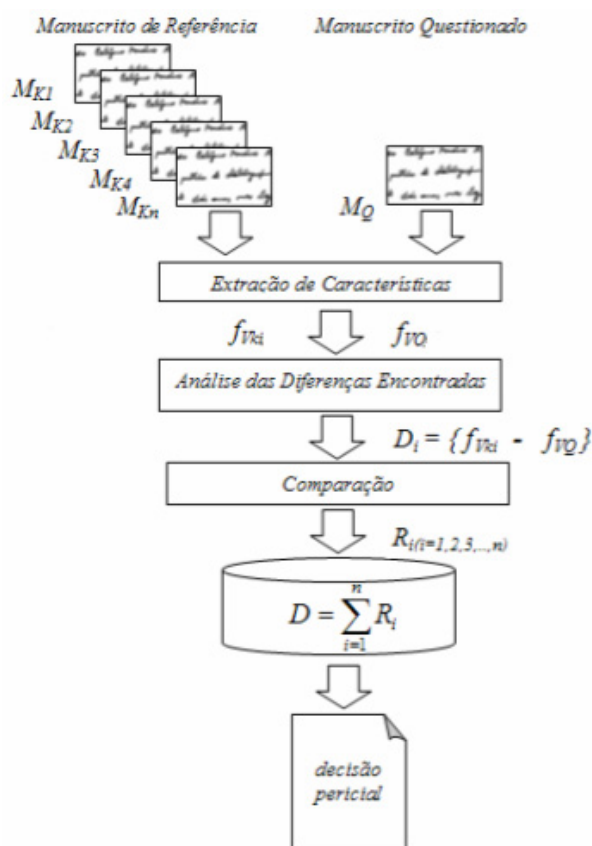


Figura 1 Diagrama esquemático do procedimento de perícia grafotécnica.

### 3. Autenticação de Manuscrito e SVM

Os métodos automáticos de verificação da autoria de textos manuscritos baseiam-se usualmente em duas abordagens, global e pessoal, Justino et al. (2003). A abordagem pessoal utiliza um modelo por autor, enquanto que a abordagem global faz uso de um modelo geral para todos os autores. O modelo pessoal, usualmente, exige um conjunto elevado de exemplares genuínos, para a geração de um modelo robusto e apresenta a vantagem de modelar adequadamente as variabilidades intrapessoais do autor. O modelo global possui a desvantagem da generalização. No entanto, possui a vantagem de necessitar um número reduzido de exemplares de cada autor e de não necessitar de um novo treinamento do modelo, na inclusão de novos autores.

No treinamento do modelo global, a classe  $W_1$  representa a classe de exemplares genuínos dos autores usados para o treinamento. A classe  $W_2$  representa o conjunto de exemplares pertencentes a outros autores. Na verificação, o modelo gerado é então utilizado para a comparação com o espécime desconhecido (Figura 2).

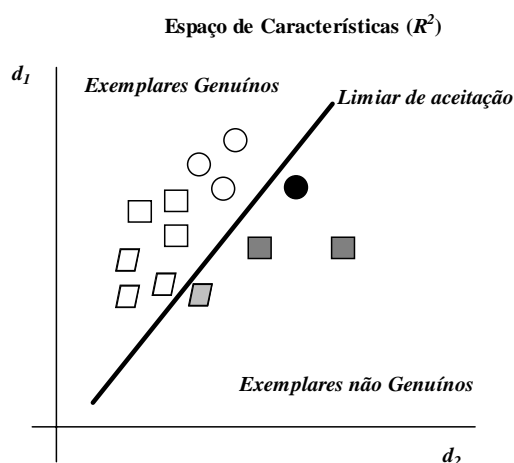


Figure 2. Modelo global de verificação da autoria de manuscritos.

*Support Vector Machine (SVM)* foi desenvolvido por V. Vapnik (1998) e é uma nova técnica no campo teórico do aprendizado estatístico. A técnica se baseia no princípio da Minimização do Risco Estrutural (MRS). O princípio da indução do MRS possui dois objetivos. O primeiro é controlar o risco empírico no conjunto de treinamento. O segundo é controlar a capacidade da função de decisão usada para obter esse valor de risco. A Função de decisão do *SVM* linear é descrito por um vetor de peso  $\bar{w}$ , um limiar  $b$  e um padrão de saída  $\bar{x}$  (Equação 1).

$$f(\bar{x}) = \text{sign}(\bar{w} \cdot \bar{x} + b) \quad (1)$$

Dado um conjunto de vetores de treinamento  $S_l$  (Equação 2) pertencente a duas classes separáveis,  $W_1$  ( $y_i = +1$ ) e  $W_2$  ( $y_i = -1$ ), o *SVM* encontra o hiperplano com a máxima distância Euclidiana do conjunto de treinamento. De acordo com o princípio do MRS, existirá somente um hiperplano com a margem máxima  $\delta$ , definida como a soma das distâncias do hiperplano até o ponto mais próximo das classes. Esse limiar do classificador linear é a separação ótima do hiperplano (Figura 3).

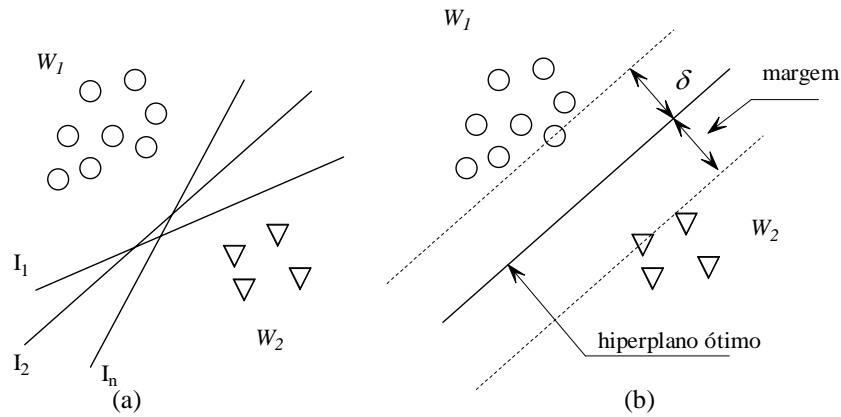
$$S_l = ((\bar{x}_1, y_1), \dots, (\bar{x}_l, y_l)), \bar{x}_i \in \mathfrak{R}^n, y_i \in \{-1, +1\} \quad (2)$$

Nos casos de conjuntos de treinamento não separáveis, o  $i$ -ésimo ponto possui uma variável  $\xi_i$ , que representa a magnitude do erro de classificação. A função de penalidade  $f(\xi)$  representa a soma dos erros de classificação (Equação 3)

$$f(\xi) = \sum_{i=1}^l \xi_i \quad (3)$$

A solução do SVM pode ser encontrada através da minimização dos erros de treinamento (Equação 4).

$$\min_{\bar{w}, b, \xi} = \frac{1}{2} \bar{w} \cdot \bar{w} + C \sum_{i=1}^n \xi_i, \quad (4)$$



**Figura 3. Classificação entre duas classes  $W_1$  e  $W_2$  usando hiperplanos: (a) Hiperplanos arbitrários  $I_i$  e (b) hiperplano com separação ótima, máxima margem.**

A literatura apresenta várias possibilidades de *kernels* para o SVM em aplicações envolvendo o reconhecimento de padrões, Burges (1998), Muller et al. (2001) e Joachims (2002). Nesse estudo inicial foi utilizado apenas o *kernel* linear (Equação 5).

$$K(\bar{x}, \bar{y}) = (\bar{x} \cdot \bar{y}) \quad (5)$$

#### 4. Base de Dados de Cartas Forenses

Para a formação da base de dados de cartas forenses, Justino (2002) e Cha (2001), foram colhidas três amostras de cartas de 145 autores distintos (Figura 3a), num total de 435 cartas. Do conjunto total de autores, foram selecionados 75 para o treinamento do modelo e 70 para os testes. As imagens das mesmas foram digitalizadas em 300dpi e 256 níveis de cinza. Uma amostra das três cartas foi usada para criar a base de treinamento.

Cada carta foi subdividida em 24 fragmentos regulares, para formação das bases de treinamento, referência e teste. Os fragmentos sem texto foram retirados (Figura 3b).

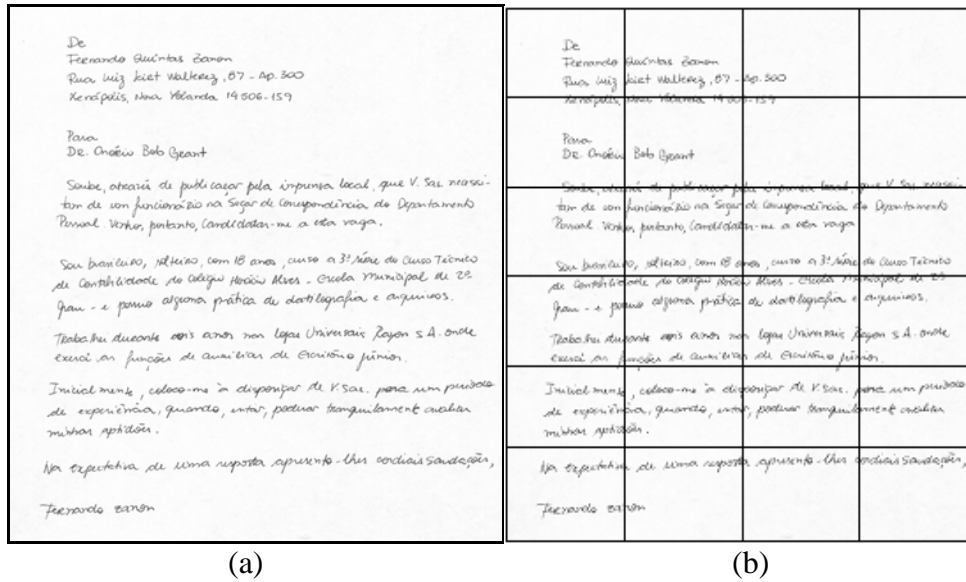


Figura 3. (a) Exemplo de carta forense. (b) Exemplo de segmentação.

Para cada autor foram usados 3 fragmentos para treinamento, 5 fragmentos para referência e 5 fragmentos para teste. A base de treinamento é composta por duas classes, a de mesmo autor  $W_1$  (225 amostras) e de autores diferentes  $W_2$  (225 amostras). Sendo a primeira formada pela comparação, dois a dois, entre os três fragmentos de um dado autor e a última, formada por fragmentos de autores diferentes, selecionados aleatoriamente na base. Os 5 fragmentos de referência de cada autor são utilizados no processo de verificação do exemplar questionado (Figura 1). Para os testes foram usados os fragmentos de referência e teste, numa comparação dois a dois (1750 amostras). Para autores diferentes foram combinados os fragmentos de referência, de um dado autor, com amostras de teste de outros, selecionados aleatoriamente na base (1750 amostras).

## 5. Método Proposto

O método proposto se baseia nos princípios da grafoscopia (Figura 1). Os segmentos de texto manuscritos são submetidos a um conjunto de fases descritas a seguir.

### 5.1. Pré-tratamento

As imagens, dos segmentos de texto, foram convertidas em imagens binárias, utilizando-se a limiarização por entropia bidimensional, Abutaleb (1989), (Figura 4a). Os contornos ou bordas dos traços foram obtidos através da aplicação do filtro morfológico de dilatação, seguido de uma erosão, com elemento estruturante em cruz Gonzalez (1992), (Figura 4b).

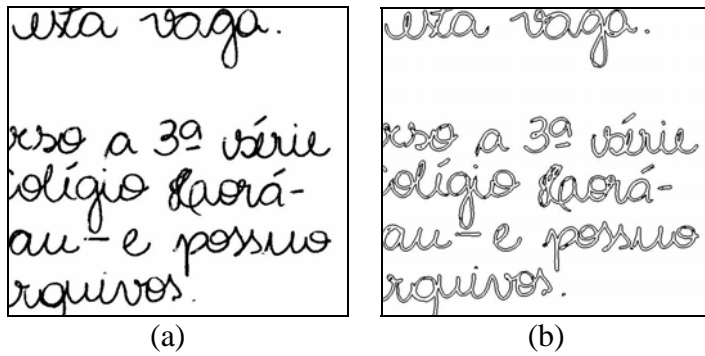


Figura 4. (a) Exemplo de imagem binária. (b) Exemplo do contorno do traço.

### 5.2. Extração de Características

O procedimento de comparação é baseado na análise da inclinação axial, uma característica grafocinética bem conhecida em grafoscopia e que vem sendo extensivamente utilizada em métodos para a identificação da autoria em manuscritos, Crettez (1995), Bulacu e Shomaker (2003). O Processo consiste em percorrer a imagem, considerando-se o pixel da borda do traço no centro de um elemento estruturante quadrado. Em seguida, verifica-se em todas as direções partindo deste pixel central e conferindo os pixels posteriores, finalizando nas extremidades do elemento estruturante, apenas se houver a presença de um fragmento de borda inteiro. Ou seja, se todos os pixels vizinhos forem pretos, considera-se o fragmento de borda, calculando-se a posição do fragmento em um vetor de posições, para a construção do histograma. Este vetor de posições é finalmente normalizado pela distribuição de probabilidade  $P(\theta)$ , que é a probabilidade de encontrar na imagem um fragmento de borda orientado em um ângulo  $\theta$  em relação ao eixo horizontal.

O algoritmo implementado utiliza, sobre o segmento da imagem, elementos estruturantes com  $k = 3, 4$  ou  $5$  pixels, ao longo do fragmento de borda onde, para cada elemento estruturante são quantificadas respectivamente em  $L = 9, 13$  e  $17$  direções de inclinação, que também representam a dimensionalidade do vetor final de características (Figura 5)

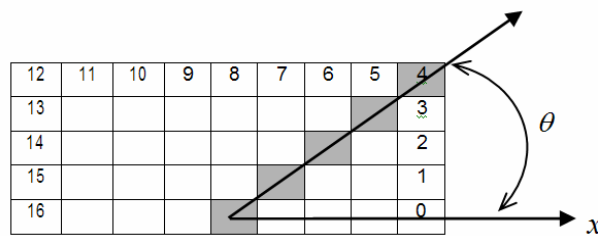


Figura 5. Exemplo do elemento estruturante com  $k = 5$  e consequentemente  $n = 17$ .

Na abordagem proposta utilizou-se a distribuição de borda-direcional com elemento estruturante  $k = 5$ , devido ao tamanho do elemento estruturante, e a quantidade de posições  $L = 17$ . O teste com a base em questão, demonstraram que os valores de  $k$  e  $L$  apresentam resultados satisfatórios na detecção da inclinação do

manuscrito, comparados aos elementos estruturantes  $k = 3$  e  $k = 4$ . A (Figura 6) apresenta o comportamento da inclinação axial para diferentes autores.

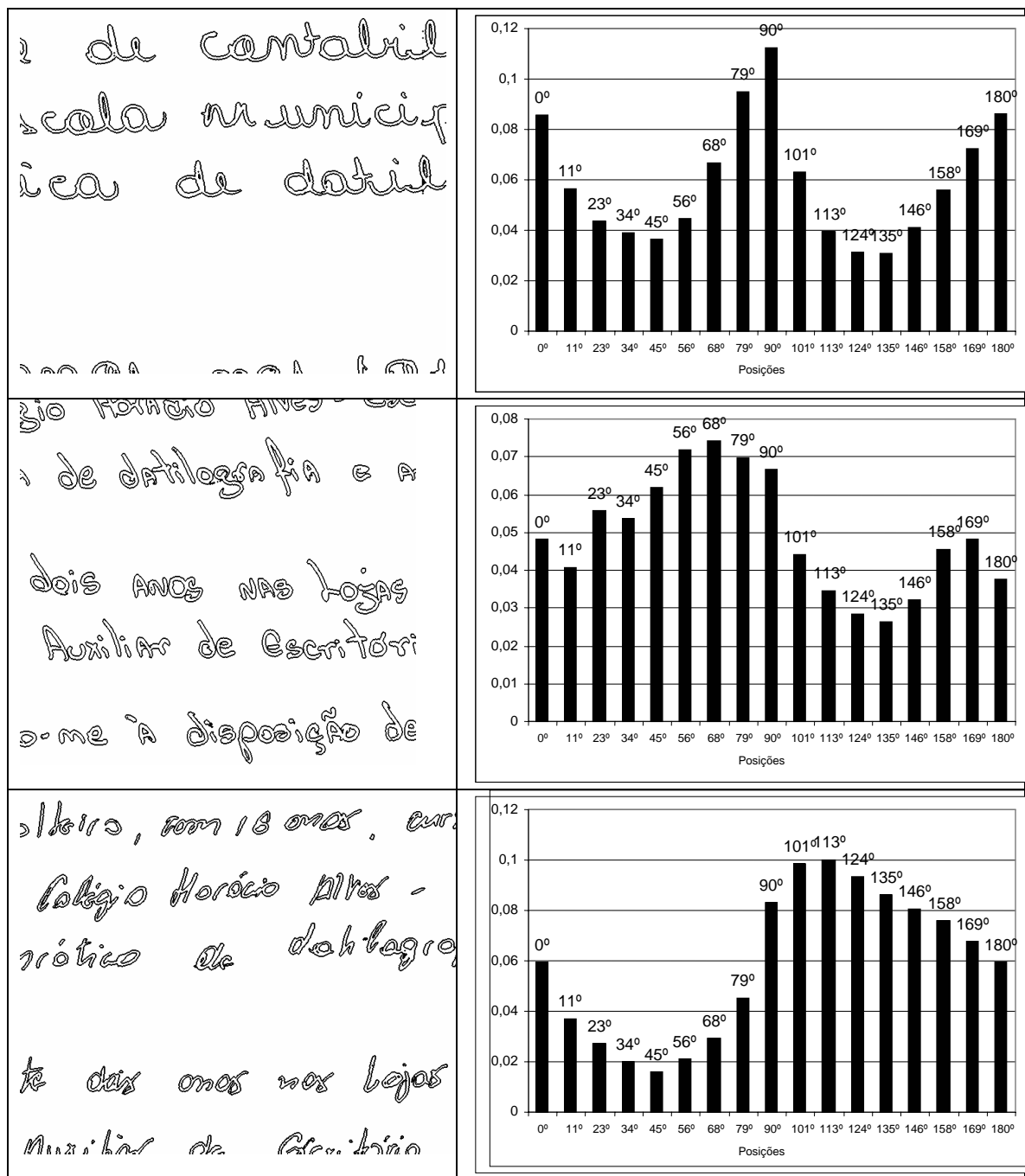


Figura 6. Exemplo de inclinação axial do manuscrito.

### 5.3. Medida das Distâncias entre Características

As bases de dados foram convertidas em vetores de características, Cha (2001). Os vetores de característica  $f_V$  são extraídos dos fragmentos  $M_{ki}$  ( $i=1,2,3...n$ ) e  $M_Q$  (Equações 6 e 7).

$$fv_{Ki(i=1,2,...,n)} = (f_1, f_2, \dots, f_L) \quad (6)$$

$$fv_Q = (f_1, f_2, \dots, f_L) \quad (7)$$

O vetor das distâncias Euclidianas  $D_i$  ( $i=1,2,3,...,n$ ) entre os fragmentos são calculados para se obter o conjunto de treinamento e testes (Equação 8).

$$D_{i=1,2,...,n} = \sqrt{(fv_{Ki} - fv_Q)^2} \quad (8)$$

### 5.4. Comparação

O processo de comparação é composto por duas fases, o treinamento e a verificação. No estágio de treinamento, as medidas das distâncias entre as características  $D_i$  ( $i=1,2,3,...,n$ ), são calculadas entre pares de fragmentos de textos. Quando dois fragmentos pertencerem a um mesmo autor, o vetor de característica é indicado com 1 (associação). Quando dois fragmentos de texto pertencerem a autores diferentes, o vetor de característica é indicado com -1 (dissociação). A distância entre dois fragmentos de texto é considerada pequena, quando as amostras pertencerem a um mesmo autor. O SVM é treinado então, para separar pequenas distâncias entre características (associação) e grandes distâncias entre características (dissociação).

No estágio de verificação, o SVM possui duas saídas. A primeira é composta pelos fragmentos pertencentes a um mesmo autor  $W_1$ . A segunda é composta por fragmentos pertencentes a autores distintos  $W_2$ .

### 5.5. Decisão

Usualmente, em uma prova pericial, o perito faz uso de um conjunto de amostras de textos de origem conhecida. Cada amostra conhecida, pertencente ao conjunto de referência (4 a 10 amostras), é comparada com a amostra de autoria desconhecida ou questionada. Nesse experimento foram utilizadas 5 amostras de referência, para cada autor.

Com o objetivo de gerar a decisão final, o método proposto classifica as saídas através de um somatório dos resultados. Esse último estágio representa a decisão final do perito  $D$  (Figura 1).

## 6. Resultados

A (Tabela 1) mostra os resultados obtidos usando *kernel* linear. Os resultados demonstram a capacidade discriminatória da característica grafocinética (inclinação axial), mesmo sendo utilizada em uma abordagem global. A taxa de erro de falsa

aceitação, ainda apresenta valores elevados, se comparado com a falsa rejeição. Em termos de taxa de erro total, o método apresentou resultados promissores, 86,429%.

**Tabela 1. Resultados obtidos usando SVM e kernel linear**

<b>Voto Majoritário</b>	<b>Falsa Rejeição (Erro Tipo I) (%)</b>	<b>Falsa Aceitação (Erro Tipo II) (%)</b>	<b>Erro Total (%)</b>
<b>SVM linear</b>	4,286%	9,286%	13,571%

## 7. Conclusão e Trabalhos Futuros

O objetivo principal desse artigo foi apresentar um método para identificação de autoria de manuscritos baseado nos princípios da grafoscopia. Para esse propósito, foram utilizadas apenas duas classes (associação e dissociação). O modelo global adotado tem se mostrado promissor na redução do número de exemplares por autor usado no treinamento do modelo e na eliminação da necessidade de um novo treinamento, quando da inclusão de novos autores.

Como proposta para trabalhos futuros encontra-se a inclusão de outras características grafotécnicas, permitindo ao modelo absorver mais adequadamente as variabilidades intrapessoais dos autores e propiciar a redução da taxa de erro de falsa aceitação. Uma outra proposta é testar outros *kernels*, com o objetivo de melhorar o processo de separação das classes.

## 8. Referências

- Santos, C. R., Justino, E. J. R., Bortolozzi, F. Sabourin, R. (2004) “ An Off-Line Signature Verification Method based on the Questioned Document Expert’s Approach and a Neural Network Classifier”, In: The Ninth International Workshop on Frontiers in Handwriting Recognition, Tokyo, 10-14p.
- Justino, E. J. R., Bortolozzi, F., Sabourin R. (2003) “An Off-line Signature Verification Method Based on SVM Classifier and Graphometric Features”, The 5th International Conference on Advances in Pattern Recognition, 2003, Calcutta , 200-204p.
- Bulacu, M. , Shomaker, L. (2003) “Writer Identification Using Edge-Based Directional Features”, Proc. of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003), IEEE Computer Society, pp. 937-941, vol. II, 3-6 August, Edinburgh, Scotland.
- Joachims, T., (2002) “Optimizing Search Engines Using Clickthrough Data”, ACM Conference on Knowledge Discovery and Mining (KDD), 1-10p.
- Justino, E. J. R. (2002) “A Análise de Documentos Questionados”, Monografia para concurso de professor titular, Pontifícia Universidade Católica do Paraná, 74p.
- Cha, S. H.”(2001) “Use of the Distance Measures in Handwriting Analysis. Doctor Theses. State University of New York at Buffalo, EUA, p. 208.

- Müller, K., Mika, S., Rätsch, G., Tsuda, K. and Schölkopf, B. (2001) "An Introduction to Kernel-Based Learning Algorithms", IEEE Transactions on Neural Networks, Vol. 12, No. 2, March, 181-202p.
- Morris, N. (2000) "Forensic Handwriting Identification Fundamental Concepts and Principles", Academic Press, 2000, p. 238.
- Vapnik, V. (1998) "Statistical Learning Theory", Wiley, N. Y.
- Dines, J. E. (1998), "Document Examiner Textbook", Pantex Intl Ltd, p. 566.
- Burges, C. J. C., (1998) "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery 2, 121-167p.
- Crettez, J. P. (1995) "A set of handwriting families: style recognition", In Proc. of the 3th. International Conf. on Document Analysis and Recognition, pages 489-494, Montreal, August 1995. IEEE Computer Society Press.
- Gonzalez, R. C., Woods, R. E., "Digital Image Processing", Addison-Wesley Publishing Company, 1992.
- Abutaleb, A. S. (1989), "Automatic Thresholding of Gray-level Pictures using Two Dimensional Entropy", Computers Graphics & Image Processing, no. 47, 22-32p.