Anais Eletrônico
VI Mostra Interna de Trabalhos de Iniciação Científica **ISBN 978-85-8084-413-9**
23 a 26 de outubro de 2012

ESTUDO E AVALIAÇÃO DE UMA FERRAMENTA PARA IDENTIFICAÇÃO DE AUTORIA

Patrícia Mateus Saramela[1], *Aline Maria Malachini Miotto Amaral*[2]

RESUMO: Documentos com frequência são alvos de fraudes para diversos fins, por isso a análise de sua autenticidade é de suma importância. Documentos manuscritos são alvos ainda maiores e mais fáceis de fraudes já que não é necessário o uso de alta tecnologia para executá-las. A análise para a identificação dos autores desse tipo de documento ainda é feita, em larga escala, por processos manuais e/ou químicos. Processos esses, antiquados e imprecisos que podem degradar e danificar o documento em questão, podendo ele até mesmo perder seu valor jurídico. Diversas ferramentas foram proposta na literatura com o objetivo de auxiliar o processo de identificação de autoria. O trabalho de Campos e Lopes (2010) apresenta uma ferramenta computacional que tem como objetivo extrair e avaliar características de documentos manuscritos digitalizados e posteriormente fazer uma análise para identificação de autoria dos mesmos. Dessa forma, o objetivo deste trabalho é o estudo e validação da ferramenta para identificação de autoria proposta no trabalho de Campos e Lopes (2010). Para tanto, foram realizados experimentos, utilizando-se 60 amostras de manuscritos de 20 diferentes escritores da Base de Cartas Forenses PUCPR, e obteve-se como melhor resultado uma taxa de 80% de acerto.

PALAVRAS-CHAVE: identificação de autoria, taxa de acerto, base de cartas forenses.

1 INTRODUÇÃO

Em situações de litígio, muitas vezes surgem questionamentos quanto à autoria de documentos apresentados como provas. O problema se agrava quando se tratam de documentos manuscritos, uma vez que as tentativas de fraude são em grande número e variedade, sendo que os fraudadores utilizam-se de métodos simples ou de métodos com alta tecnologia para executá-las.

Atualmente, o processo de análise de manuscritos é realizado pela perícia através de métodos manuais (aparelhamento ótico) e/ou químicos. Segundo Sheikholesmani et al. (1996), a extração manual de características de manuscritos é tediosa e sujeita a erros.

Além disso, diferentes grafologistas podem extrair as mesmas características do mesmo do manuscrito de forma diferente. É também possível que o conteúdo do manuscrito influencie a análise grafológica. Umas das alternativas para auxiliar os

[1] Acadêmica do Curso de Sistemas de Informação do Centro Universitário de Maringá, CESUMAR, Maringá-PR. Programa Institucional de Bolsas de Iniciação Científica (PIBIC/CNPq – CESUMAR) - patricia.saramela@bol.com.br
[2] Orientadora, Docente Mestre do Centro Universitário de Maringá, CESUMAR, Maringá-PR; amiotto@cesumar.br

grafologistas a superarem estes problemas é a automatização de todo ou parte do processo de extração e análise das características.

Pesquisas como as de Bulacu e Shomaker (2003), Srihari e Shi (2004), Oliveira et al. (2005) e Campos e Lopes (2010) dentre outras, foram realizadas para permitir o desenvolvimento de sistemas automáticos ou semi-automáticos de análise que auxiliem os grafotécnicos na verificação de autoria de documentos manuscritos. Um problema particular desta verificação é estabelecer uma relação entre as características dos textos manuscritos de um determinado autor.

O trabalho de Campos e Lopes (2010) apresenta uma ferramenta computacional que tem como objetivo extrair e avaliar características de documentos manuscritos digitalizados e posteriormente fazer uma análise para identificação de autoria dos mesmos. A validação dos resultados de ferramentas computacionais tais como a apresentada no trabalho citado é extremamente importante e muitas vezes muito trabalhosa, uma vez que depende de grandes quantidades de dados de entrada (documentos digitalizados) e da análise estatística dos resultados produzidos.

Neste contexto, o objetivo deste trabalho é o estudo e validação da ferramenta para identificação de autoria proposta no trabalho de Campos e Lopes (2010). Para tanto, foram realizados experimentos, utilizando-se 60 amostras de manuscritos de 20 diferentes escritores da Base de Cartas Forenses PUCPR (FREITAS et al., 2008).

Este artigo esta organizado da seguinte forma: a Seção 02 discute os principais aspectos que devem ser considerados para a automatização do processo de identificação de autoria, na Seção 03 é apresentado o processo adotado para a validação da ferramenta em estudo. Na Seção 04 são apresentados os principais resultados obtidos com os experimentos de validação realizados. Finalmente, na Seção 05 são apresentadas a considerações finais desta pesquisa.

2 IDENTIFICAÇÃO DE 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 primariamente baseado no conhecimento e experiência do perito forense. Devido a esta

situação problemática, de medidas não objetivas e decisões que nem sempre podem ser reproduzidas, tentativas de oferecer suporte automático e semi-automático aos métodos tradicionais vem sendo realizadas.

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, com 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 fotos, documentos e textos. Estas bases são, também, de suma importância para a validação dos sistemas automáticos e semi-automáticos de identificação de autoria.

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 (primitivas) 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 pelo qual o manuscrito é capturado para sua posterior análise. Neste trabalho, faremos a análise de uma ferramenta que se baseia em amostras de manuscritos já digitalizados, ou seja, de um sistema para identificação *offline* de autoria.

ISBN 978-85-8084-413-9

3 DESENVOLVIMENTO

Para a condução dos experimentos de validação, inicialmente a ferramenta de Campos e Lopes (2010) foi estudada de forma a identificar as características que são extraídas dos manuscritos e utilizadas no processo de identificação.

O conjunto de características extraídas nesta ferramenta conta com sete características, no entanto, pode-se observar que dois grandes grupos de características estão sendo utilizadas, são elas:

- **Hábitos do uso/posicionamento do texto na folha**: estas características são muito utilizadas pelos peritos forenses, 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. 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 “posicionamento do texto na folha”. 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 forenses e a definição do tamanho padrão das palavras. Em nossos estudos, a primeira palavra de cada linha foi “destacada” e sua altura e proporção de pixels pretos 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.

A Tabela 1 apresenta uma descrição de cada uma das características implementadas na ferramenta de Campos e Lopes (2010).

ISBN 978-85-8084-413-9

Tabela 1. Características extraídas (Amaral et al., 2012)

Característica	Descrição
f_1	*Número total de linhas da carta.*
f_2	*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.
f_3	*Posição da margem direita*: para as 20 primeiras linhas da carta, a distância da margem direita é identificada. Esta distância é definida usando-se uma linha de referência (linha imaginária que cruza a linha no meio de sua altura – esta linha pode ser melhor visualizada na Figura 3.14) e verificando o último pixel preto desta linha.
f_4	*Posição da margem esquerda:* esta distância é definida usando a linha de referência (já descrita anteriormente) de cada uma das linhas da carta e identificando o menor valor de posição inicial.
f_5	*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).
f_6	*Posição da margem inferior*: esta distância é definida usando-se a linha de referência da última linha, e identificando a posição final desta linha de referência.
f_7	*Altura da primeira palavra*: para as 20 primeiras linhas da carta é calculada a altura da primeira palavra que foi segmentada e teve seus valores limites extraídos na etapa de pré-processamento.

A Figura abaixo apresenta uma visão geral das características implementadas na ferramenta de Campos e Lopes (2010).

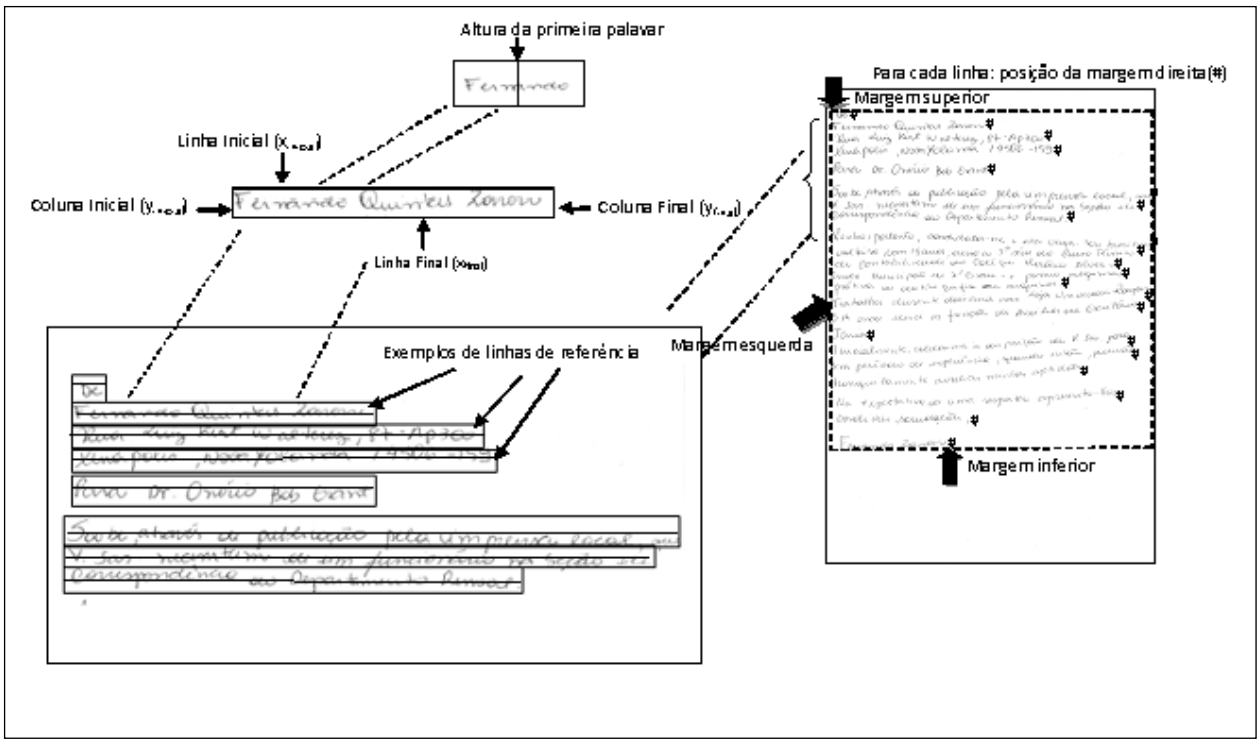


Figura 1. Características extraídas da ferramenta de Campos e Lopes (Amaral et al., 2012)

ISBN 978-85-8084-413-9

Para validar os resultados obtidos com é necessária a realização de experimentos. Segundo Amaral et al. (2012), tais experimentos devem seguir um protocolo rigoroso para que os mesmos possam ser replicados em diferentes contextos. Diferentes variáveis afetam os resultados de trabalhos como o nosso, são elas: número de escritores, abordagem de classificação adotada e base de dados utilizada para o treinamento e teste.

Neste contexto, a Figura 2 apresenta uma visão geral do protocolo seguido ao longo dos experimentos realizados. Este protocolo seguiu a abordagem definida no trabalho de Amaral et al. (2012).

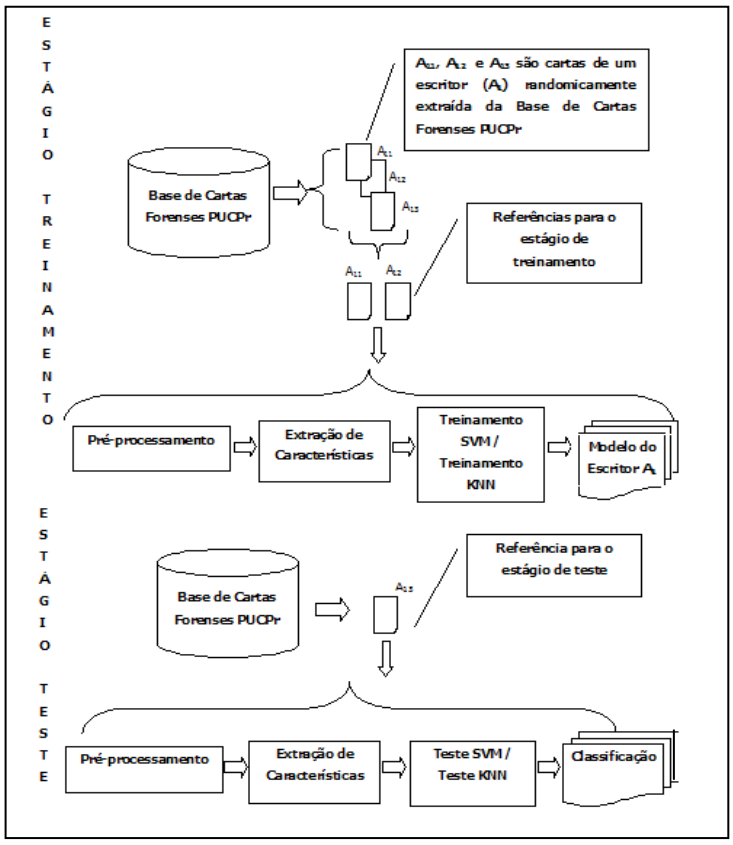


Figura 2. Protocolo dos experimentos (AMARAL et al., 2012)

O vetor de primitivas obtido no processo de extração de características (métricas) por meio da ferramenta de Campos e Lopes (2010), ver Figura 3, é usado como entrada para o(s) algoritmo(s) de classificação.

ISBN 978-85-8084-413-9

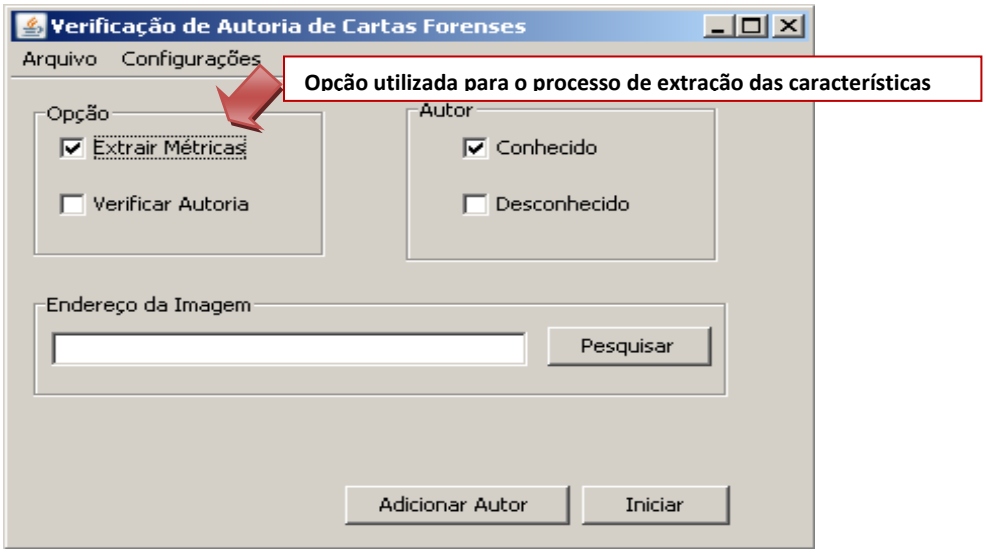


Figura 3. Tela Principal da Ferramenta e Campos e Lopes (2010)

Os estágios de treinamento e teste foram realizados submetendo o vetor de primitivas extraído (das cartas digitalizadas) aos classificadores. Para o estágio de treinamento o vetor de primitivas é usado para compor o modelo, a para o estágio de teste o vetor é usado para a identificação de escrita.

Cada escritor presente no conjunto de teste é comparado com cada um dos modelos definidos no estágio de treinamento (todos-contra-todos – *all-against-all*), assim o melhor resultado (“vencedor leva tudo” – winner-takes-all) obtido é escolhido pelo classificador. Para os experimentos cujos resultados são apresentados na próxima seção, foram selecionados aleatoriamente 20 diferentes escritores da base de cartas forenses modelo PUCPR (ver Figura 3 e 4).

Para o estágio de treinamento foram utilizadas duas cartas de cada escritor (totalizando 40 diferentes cartas), e para a fase de teste foi utilizada a 3ª carta de cada um dos 20 escritores utilizados na fase de treinamento (totalizando 20 diferentes cartas).

O arquivo contendo as métricas extraídas para o estágio de treinamento foi submetido para o classificador SVM (*Support Vector Machine)* (VAPNIK, 1982) gerando modelo de cada escritor. Finalmente, o arquivo de teste extraído é submetido para os modelos previamente gerados na etapa de teste, e os processo de classificação resultante é computado.

ISBN 978-85-8084-413-9

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.

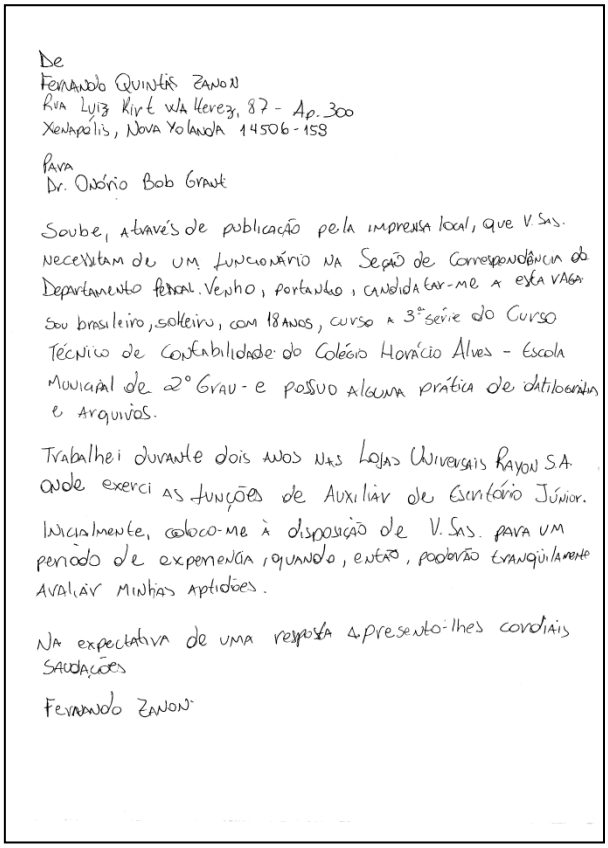
Venho, portanto, candidatar-me a esta vaga. Sou brasileiro, solteiro, com 18 anos, curso a 3ª série do Curso Técnico de Contabilidade do Colégio Horácio Alves - Escola Municipal de 2° Grau - e possuo alguma prática de datilografia e arquivos.

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 4. Modelo de carta Forense PUCPR (FREITAS et al., 2008)



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

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

Na expectativa de uma resposta apresento-lhes cordiais saudações

Fernando Zanon

Figura 5. Carta PUCPR CF00001_01 (amostra presente na base PUCPR)

ISBN 978-85-8084-413-9

4 RESULTADOS OBTIDOS

Os resultados obtidos nestes experimentos atingiram taxas de **80%** de acerto, utilizando como algoritmo de classificação SVM e com 20 diferentes escritores (totalizando 60 cartas diferentes). A Tabela 1 apresenta avaliações individuais das características extraídas.

Tabela 2. Resultados dos experimentos.

Grupo de características	(%) Performance
f_1 . Número de linhas	25%
f_2 . Proporção de pixels pretos da primeira palavra de cada linha	55%
f_3 . Posição da margem direita	55%
f_4 . Posição da margem esquerda	15%
f_5 . Posição da margem superior	15%
f_6 . Posição da margem inferior	15%
f_7 . Altura da primeira palavra de cada linha	50%
f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7	**80%**

Quando comparado com outros trabalhos presentes na literatura, ver Tabela 3, pode-se observar que os resultados obtidos com os experimentos realizados apresentam-se promissores. Deve-se destacar que novos experimentos, com um número maior de escritores também devem ser realizados, para conseguirmos obter resultados mais confiáveis.

Tabela 3. Comparação de taxa de identificação da ferramenta de Campos e Lopes (2010) com outros trabalhos da literatura

Autores	Número de Escritores	(%) Taxa de Identificação
Luna et al. (2011)	30	92%
Blankers et al. (2007)	41	98%
Siddiqi e Vincent (2008)	100	92%
Pervouchine e Leedham (2007)	165	58%
Campos e Lopes (2010)	20	80%

5 CONSIDERAÇÕES FINAIS

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. Nesse contexto, o objetivo deste trabalho é a validação da ferramenta de Campos e Lopes (2010) que se propõe a automatizar parte do processo de identificação de autoria.

Pode-se observar com os experimentos realizados (ver Tabela 2), que individualmente as características extraídas dos documentos manuscritos não possuem um alto poder discriminatório, no entanto, quando agrupadas elas permitem uma taxa de reconhecimento de 80% em um experimento com 20 diferentes escritores (contabilizando 60 manuscritos, sendo 3 manuscritos de cada escritor).

Como trabalhos futuros, pretende-se estudar e incluir novas características na ferramenta de Campos e Lopes (2010). Assim, acredita-se que será possível aumentar as taxas de acerto obtidas até o presente momento.

REFERÊNCIAS

AMARAL, A. M. M. M; FREITAS, C. O. A.; BORTOLOZZI, F. **The graphometry applied to writer identification**. To be publishes in: Proceedings of the 2012 International Conference on Image Processing, Computer Vision, and Pattern Recognition. july, 2012, Las Vegas, USA.

BLANKERS, V.; NIELS, R.; VUURPIJL, L. **Writer identification by means of explainable features: shapes of loops and lead-in strokes.** In: Proceedings of the 19th Belgian-Dutch Conference on Artificial Intelligence, p. 17-24, 2007.

BULACU, M.; SCHOMAKER, L.; BRINK, A. **Text-independent writer identification and verification on offline Arabic handwriting**. In: Proceedings of the 9th Conference on Document Analysis and Recognition (ICDAR), 2007.

CAMPOS, C. S. de; LOPES A. L. **Identificação de autoria em Documentos Manuscritos**. 2010, 32p. Projeto Físico (Graduação em Engenharia da Computação) - Pontifícia Universidade Católica do Paraná, Paraná.

FREITAS, C. O. A.; OLIVEIRA, L. S.; BORTOLOZZI, F.; SABOURIN, R. **Brazilian Forensic Letter Database.** In: Proceedings of the 11th International Workshop on Frontiers on Handwriting Recognition, 2008.

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.

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.

OLIVEIRA, L.S.; JUSTINO, E.; FREITAS, C. O. de A.; SABOURIN, R. **The graphology applied to signature verification**. In: Proceedings of the 12th Conference of the International Graphonomics, 2005.

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

SHEIKHOLESLAMI, G.; SRIHARI, S. N.; GOVINDARAJU, V. **Computer aided graphology**. In: Proceedings of the Fifth International Workshop on Frontiers in Handwriting Recognition. Essex, England. p.457-460. September, 1996.

SIDDIQI, I.; VINCENT N. **Combining global and local features for writer identification**. In: Proceedings of the Eleventh International Conference on Frontiers in Handwriting Recognition, p.48-53, 2008.

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

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