Atribuição Forense de Impressoras

Giuliano R Pinheiro e Anderson Rocha
Instituto de Computação - UNICAMP
giuliano.pinheiro@students.ic.unicamp.br
rocha@ic.unicamp.br

Resumo

De forma geral, atribuir um documento digital ao seu dispositivo gerador envolve sua descrição de forma única que permita um casamento unívoco com tal dispositivo. Por exemplo, atribuir uma foto a uma câmera, um documento digitalizado a um scanner, um documento impresso a uma impressora, etc. Até que ponto se pode enganar uma análise forense quando seu objeto de observação não é informação contida no próprio documento, mas a própria textura e características visuais daquilo que foi impresso? Este trabalho propõe um método multidirecional e multiescala de atribuição forense de impressoras com base no estado da arte com resultados promissores.

Introdução

Falsificações em documentos com intenções criminosas são situações complicadas de investigar. Com a crescente facilidade de edição de documentos digitais, essa dificuldade só tem aumentado. A análise forense de impressoras, em Computação Forense, utiliza de técnicas computacionais para caracterizar uma impressora por aquilo que ela imprime, tornando possível a atribuição de um documento apontado como evidência de um crime a uma impressora de um suspeito, por exemplo.

Desenvolvemos uma maneira de apontar a impressora fonte de um documento digitalizado utilizando um conjunto de dados produzido para imitar cenários reais de texto impresso, sem haver nenhum controle estatístico ou de estilo gráfico sobre o texto.

Metodologia

Criamos um *dataset* para avaliação dos resultados e um esquema de caracterização e classificação de documentos impressos por um rol de 10 impressoras.

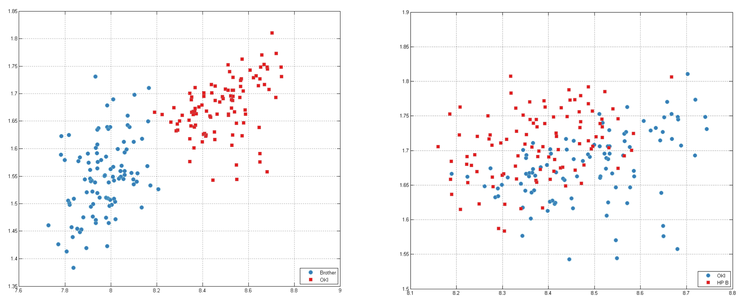

Figura 1: Separabilidade das características: à esquerda, é visível a separabilidade de duas classes por duas características, o que não ocorre sempre, como é o caso à direita, das mesmas duas características para outro par de classes.

Dataset

Produzimos um conjunto de dados de documentos retirados da *Wikipedia* em formato PDF. Estes documentos estão igualmente divididos em 4 categorias advindas de 2 qualidades: em inglês e em português, com figuras e sem figuras.

	Fabricante	Modelo	Apelido
1	Brother	HL4070CDW	Brother
2	Canon	D1150	Canon D
3	Canon	MF3240	Canon MF
4	Canon	MF4370DN	Canon MFDN
5	HP	CP2025	HPA
6	HP	CP2025	HPB
7	HP	CP1518	HPCP
8	Lexmark	E260D	Lex
9	OKI	C330	OKI
10	Samsung	CLP315	Sams

Tabela 1: Impressoras usadas no trabalho

Os documentos foram impressos nas impressoras indicadas na tabela 1 e, em seguida, um subconjunto deles foi escaneado a 600 dpi (*dots per inch*). Esses foram os documentos adotados para uso e validação dos experimentos realizados.

Caracterização e classificação

Com o conjunto de dados gerado, extraímos todos os caracteres 'e' do documento, o mais frequente. Com apenas 30 documentos por impressora, isso resultou em cerca de 245 mil caracteres. Em seguida, foi feita a caracterização, seguindo dois modelos:

Por caractere , onde assumimos que um caractere contém, sozinho, informação suficiente para identificar a impressora fonte.

Por documento , onde assumimos que não um caractere não contém informação suficiente da impressora, mas o conjunto dos caracteres de um documento, sim.

Os descritores são estatísticas da Matriz de Co-Ocorrência de Tons de Cinza, ou GLCM. Essa matriz tem a seguinte propriedade: o elemento $glcm_k, l$ conta quantas vezes um pixel de intensidade k era vizinho de um de intensidade l, sendo essa vizinhança indicada por um vetor de deslocamento.

Nos baseamos em [1] e estendemos o descritor, obtendo uma versão multidirecional, onde usamos todas as 8 direções adjacentes a um pixel ao invés de uma, e multiescala, onde construímos uma pirâmide de 4 escalas da imagem original.

Esses vetores descritores foram retirados ou de GLCMs individuais dos caracteres ou de GLCMs acumuladas por documento, de acordo com os modelos de caracterização apresentados, e treinam um classificador SVM (*Support Vector Machine*).

Resultados e Discussão

A figura 2 apresenta os resultados do descritor final, multidirecional e multiescala, por caractere, aplicado ao nosso *dataset*. Já a figura 3 mostra a o mesmo descritor usando a descrição por documento.

Figura 2: Resultado de experimento de classificação por caractere.

Essa matriz mostra, em porcentagem, o quanto as classes foram confundidas entre si. Repare que a situação se aproxima do ideal quanto mais negra é a diagonal principal.

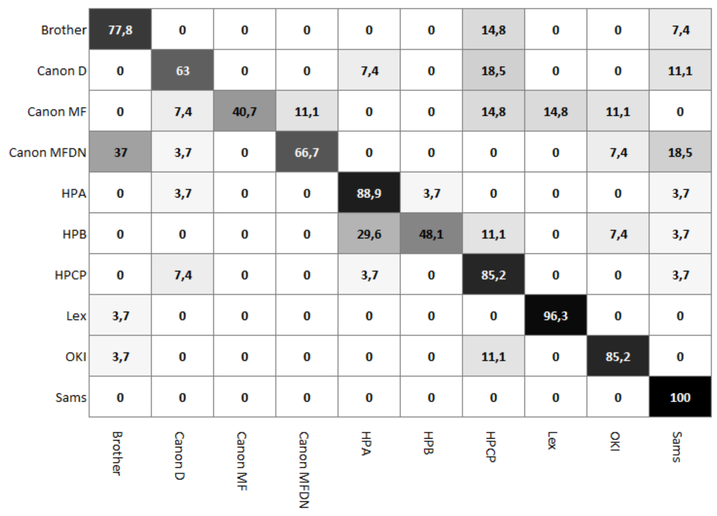

Figura 3: Resultado de experimento de classificação por documento.

Comparando as figuras, vemos que a abordagem por documento é ligeiramente melhor. Há confusão entre algumas classes (e.g.: canto superior direito), mas os valores são pequenos comparados ao acerto da classe verdadeira, na diagonal.

Em relação à abordagem da literatura no nosso *dataset*, nossa melhoria se mostrou mais eficaz que o algoritmo usado por Mikkilineni et al.[1], com taxa de acerto média de 75%, contra 55% do original, um aumento de 36%.

Conclusões

A abordagem multidirecional e multiescala proposta superou a literatura em nosso *dataset*, feito para manter o conteúdo impresso o mais próximo da realidade. Há muito que se pode melhorar, como a dimensionalidade do descritor ou um estudo das características mais discriminativas.

O método original não lida bem com a escassez de dados. Usamos resolução de 600 dpi, 4 vezes menor que a original, o que nos deu 6,25% dos dados que o estado da arte. Ainda assim, obtivemos acerto 36% maior que o método original para o nosso *dataset*.

Referências

[1] A. K. Mikkilineni, P.-J. Pei-Ju Chiang, G. N. Ali, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp. Printer identification based on textural features. In *Intl. Conference on Digital Printing Technologies*, pages 306–311, 2004.