Intelligent Printing Technique Recognition and Photocopy Detection for Forensic Document Examination

Marco Schreyer, Christian Schulze, Armin Stahl and Wolfgang Effelsberg
mschreye@rumms.uni-mannheim.de
University of Mannheim and
German Research Center for Artificial Intelligence (DFKI)

Abstract: The detection of fraudulent documents is becoming an important issue in large scale office automation. In this context the recognition of a documents creation process is of valuable information in the examination of a questioned documents authenticity. We propose a novel method of printing technique recognition and photocopy detection based on statistical supervised machine learning. This is achieved by the classification of statistical document features that are obtained by an analysis of a given documents spatial and frequency domain. For the purpose of testing our methods we prepared a representative database of documents. Our results showed that (1) most of the documents are classified correctly (2) even at low scan resolutions.

GI-Topic: KI-BV (artificial intelligence - image understanding)

1 Introduction

The progress of digital printing and imaging technologies had a tremendous impact on the way we generate, publish and store information nowadays. This technological progress, as more and more applicable, is not only used for legitimate purposes but also for illegal activities. Especially in the case of banks, insurance companies and tax authorities, processing several thousand documents that are issued by a high number of invoicing parties each day. In such scenarios fully automatic document analysis systems are urgently needed to examine fraud and money laundry [FS07]. Important insights in a forensic document examination can be obtained by answering the questions: How was the suspected document at hand created? Is it an original or a photocopy? An opportunity to address these questions is offered by the appliance of techniques derived from digital image processing and statistical pattern recognition. Only a small number of publications concerning the classification of documents according to their creation process exists nowadays [LMB06, MAC^{+}05, Tch04]. To the best of our knowledge none of them is concerned with the detection of photocopied documents.

We will show that photocopied, laser and inkjet printed documents can be distinguished even at low scan resolutions. This is achieved by the examination of the unique characteristics that correspond exclusively to each document creation technique.

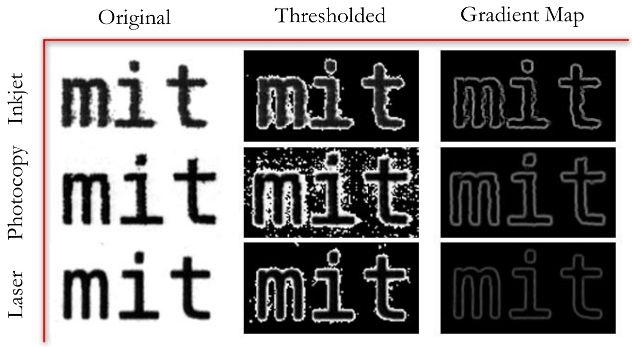

Figure 1: Representative document image extracts of an inkjet printed, laser printed and photocopied document scanned with a resolution of $2400dpi$.

2 Printing Processes and their Characteristics

Printing in a more general perspective can be seen as a “complex reproduction process“ in which printing ink is applied to a printing substrate in order to transmit information in a repeatable form [Kip01]. The examination of high resolution scan images of photocopied, laser and inkjet printed documents reveals unique characteristics corresponding to each of these document creation techniques. These characteristics or document image degradations are the main source of evidence in document forensics and can be understood as a documents *”every sort of less-than-ideal properties“* or *”the departure of an ideal version“* [Bia00]. Observing the exemplary document image extracts presented in Figure 1, four discriminative characteristics can be identified: (1) image noise and artifacts, (2) character edge roughness, (3) character edge contrast and (4) uniformity of printed character area.

3 Statistical Feature Extraction

To distinguish different printing techniques and photocopies according to their unique document image characteristics as illustrated in Figure 1 we performed four different types of document analyzes.

Noise Analysis We obtain a documents image noise by applying mean, median and Gaussian filtering techniques as described in [GSW07]. As statistical features the mean and standard deviation are obtained from the noise image. Furthermore, the correlation and mean squared error was calculated from the original and denoised document image.

Gradient Analysis To capture statistical information about fine image intensity variations that correspond to character edges and noisy image regions we apply different gradient filters as proposed by [Tch04]. Subsequently, the gradient histogram is calculated for each document image. As statistical features we obtain the mean and standard deviation is obtained for different histogram intervals.

DCT Frequency Analysis High frequencies are denoted by large graylevel alterations within a small image area and low frequencies are denoted by large areas of nearly constant graylevel values. Therefore, we utilized the concept of frequency using ”Discrete Cosine Transformation“ (DCT) for the purpose of document classification. This is done by calculating DCT coefficients mean and standard deviation as statistical features at certain frequency subbands as described in [SSSB09].

Multiresolution Wavelet Analysis To achieve independence of the transformations base functions we also utilized Multiresultion Wavelet Analysis [GW07]. This was done using the Haar and Daubechies Wavelets as well as Coiflets. At different scales of wavelet decomposition we obtain the mean and standard deviation as statistical features for document classification.

4 Experimental Setup and Results

Since none of the reviewed document image databases[1] is currently providing a document annotation of the printing technique used in its creation the necessity emerged to create a new document image database annotated with the needed ground truth information. In German speaking countries a document called the “Grauert” letter, implementing the DIN-ISO 10561 standard, is used for the test of printing devices. The “Grünert” letter, which is derived from this document, yields the same results in printer tests. Because of its high similarity in layout and content to regular written business letters, we used the ’Grünert’ letter as template for ground truth database creation. The created document database, consists of 49 different laser printouts, 14 different inkjet printouts as well as 46 photocopied documents. The variety of printing and photocopier device manufacturers covers all major brands typically present in (home) office environments, for example Hewlett-Packard^{TM}, Epson^{TM}, Canon^{TM}and Ricoh^{TM}. To comply with the “No Free Lunch Theorem” [Wol01], the subsequent feature classification is performed on the basis of two supervised machine learning techniques: a (1) Multilayer Perceptron (MLP) and a (2) Support Vector Machine (SVM).

The experimental results illustrated in Figure 2 confirm our initial hypothesis that printing technique recognition and photocopy detection can be achieved by the analysis of discriminative information obtained from scanned document images. Utilizing both classification techniques remarkable classification results are achieved for documents scanned with resolutions of $400dpi$ and $800dpi$. While the highest classification results 92.92% (99.08%) is achieved utilizing DCT Frequency Analysis and SVM classification at $400dpi$ ($800dpi$).

[1] UW English Document Image Database I - III, Medical Article Records System (MARS), MediaTeam Oulu Document Database and Google 1000 Books Project

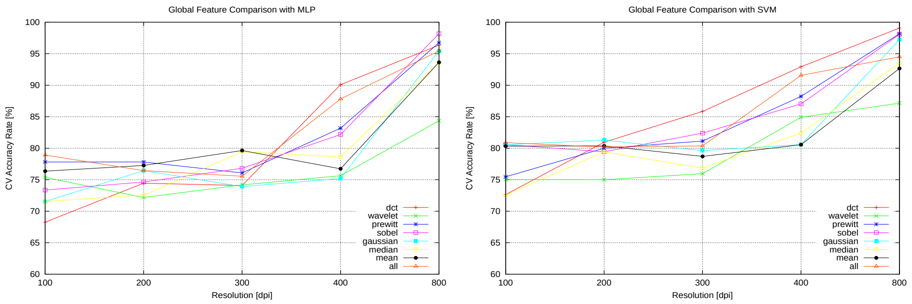

Figure 2: Feature evaluation accuracy results obtained by (left) multilayer perceptron and (right) support vector classification performing 10-fold stratified sampled cross validation.

References

[Bia00] Henry S. Biard. The State of the Art in Document Image Degradation Modeling. In *Proceedings of the 4th IAPR Workshop on Document Analysis Systems*, pages 1–13, 2000.

[FS07] Katrin Franke and Sargur N. Shrihari. Computational Forensics: Towards Hybrid-Intelligent Crime Investigation. In *Third International Symposium on Information Assurance and Security*, pages 383–386, 2007.

[GSW07] Hongmei Gou, Ashwin Swaminathan, and Min Wu. Robust Scanner Identification based on Noise features. In *Proceedings of the SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents IX*, 2007.

[GW07] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall International, 3 edition, 2007.

[Kip01] Helmut Kipphan. *Handbook of Print Media*. Springer, 1 edition, 2001.

[LMB06] Christoph H. Lampert, Lin Mei, and Thomas M. Breuel. Printing Technique Classification for Document Counterfeit Detection. In *Computational Intelligence and Security (CIS) 2006, Ghuangzhou, China*, 2006.

[MAC$^+$05] A.K. Mikkilineni, G.N. Ali, P.-J. Chiang, G.T.-C. Chiu, J.P. Allebach, and E.J. Delp. Printer identification based on graylevel co-occurance features for security and forensic applications. In *Proceedings of the SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents VII*, volume 5681, pages 430–440, San Jose, CA, 2005.

[SSSB09] Christian Schulze, Marco Schreyer, Armin Stahl, and Thomas M. Breuel. DCT Coefficient Analysis for Printing Technique and Document Copy Detection. In *Advances in Digital Forensics V (to appear)*. Springer, Boston, 2009.

[Tch04] Jack Tchan. The development of an image analysis system that can detect fraudulent alterations made to printed images. In *Proceesdings of the Photo-Optical Instrumentation Engineers (SPIE) Conference*, volume 5310, pages 151–159, 2004.

[Wol01] David H. Wolpert. The supervised learning no-free-lunch Theorems. In *In Proc. 6th Online World Conference on Soft Computing in Industrial Applications*, pages 25–42. Springer-Verlag, 2001.