Evaluation of Graylevel-Features for Printing Technique Classification in High-Throughput Document Management Systems

Christian Schulze, Marco Schreyer, Armin Stahl, and Thomas Breuel

German Research Center for Artificial Intelligence (DFKI),
University of Kaiserslautern,
67663 Kaiserslautern, Germany
{christian.schulze,marco.schreyer,armin.stahl}@dfki.de
{tmb}@informatik.uni-kl.de
http://www.iupr.org

Abstract. The detection of altered or forged documents is an important tool in large scale office automation. Printing technique examination can therefore be a valuable source of information to determine a questioned documents authenticity. A study of graylevel features for high throughput printing technique recognition was undertaken. The evaluation included printouts generated by 49 different laser and 13 different inkjet printers. Furthermore, the extracted document features were classified using three different machine learning approaches. We were able to show that, under the given constraints of high-throughput systems, it is possible to determine the printing technique used to create a document.

Key words: feature evaluation, printing technique classification, counterfight detection, questioned document, document forensic, document management

1 Introduction

As with many new technologies, the opportunity to create printed documents in high quality has resulted in a more extensive usage of these technologies. However this progress, as more and more applicable, is not only used for legitmate purposes but also for illegal activities. With the digital imaging technique available today, it is simple to create forgeries or altered documents within short timeframes. Recent cases reported to the American Society of Questioned Document Examiners (ASQDE) reveal the increasing involvement of modern printing technologies in the production of counterfeited banknotes[1,2] and forged documents[3,4].

In particular within large companies and governmental organizations where paperless processing is aimed, many incoming documents and invoices are handled by large-scale automatic document management systems (DMS). Especially in the case of banks, insurances and auditing companies, processing several thousand documents each day, there is a major need for intelligent methods to determine if the processed documents are genuine or not. Observing the high number

of processed bills being related to payments and assuming that only a low percentage of these are forged or manipulated, it is easy to imagine that quite some disprofit could be prevented with the use of an authenticity verification system.

The examination of questioned documents usually progresses from the general to the specific[5]. It is a common practice for document examiners to step through their examinations attempting to first determine document class characteristics. Therefore important insights in the examination process can be obtained by answering the question: How the document at hand was created? The ability to investigate documents for consistence in printing technology can be a first useful observation, deciding if the given document is genuine. Furthermore, detecting if specific document regions have been printed with the same non impact printing technique, is an essential piece of information for making the decision.

Therefore, within this paper the performance of common textural and edge based graylevel features developed for digital printing technique recognition was evaluated in a real world scenario. The evaluation was carried out with respect to scan resolution constraints that commonly apply to high throughput scanning systems being used for DMS. The goal of the extensive investigation was to examine the tested features for their applicability to low resolution scans. For this study printouts generated by 49 different laser and 13 different inkjet printers have were evaluated. These printouts were based on a template document whose layout can usually be found in typical office environments. Aspects like the paper quality and ink type used, as well as the effect of document aging have not been taken into account, since the receiver of a document usually has no or only little influence on those details.

1.1 Related Work

Forensic document examiners are confronted on a daily basis with questions like by whom or what device a document was created, what changes have occurred since its original production, and is the document as old as it purports to be[5]. Therefore a variety of sophisticated methods and techniques have been developed since the prominent article [6] published by Albert S. Osborn and Albert D. Osborn in 1941.[1] The textbooks of Hilton[7], Ellen[8], Nickell[9], Kelly and Lindblom[5] offer excellent overviews of the state of the art in the techniques applied to questioned documents by forensic document examiners. These techniques can be divided into destructive and non destructive analysis determining physical and chemical document features.

Although the fact that the use of digital imaging techniques in the forensic examination of documents is relatively new, recent publications show promising

[1] "A document may have any one of twenty or more different defects that are not seen until they are looked for. Some of these things are obvious when pointed out, while others to be seen and correctly interpreted must be explained and illustrated", by Albert S. Osbornand Albert D. Osborn, co-founders of the American Society of Questioned Document Examiners published the year before the formal founding.

and interesting methods in terms of discriminating non impact printing techniques.

Therefore, a valuable source of information in the determination of a documents underlying printing technology can be gained by an assessment of the print quality. Oliver et al.[10] outlined several print quality metrics including line width, raggedness and over spray, dot roundness, perimeter and number of satellite drops. It is intuitively clear that for an evaluation of these metrics a high resolution scan of the document is inescapable. Another method proposed by Mikkilineni et al.[11] [12] traces documents according to the printing device by extracting graylevel co-occurrence features from the printed letter "e". But their method is based on $1200dpi$ high resolution scans and consequently also not feasible for high throughput systems.

The systems outlined by Tchan[13] exhibit high similarity to our approach. He captured documents with a camera at low resolution and differentiates printing technologies by measuring edge sharpness, surface roughness and image contrast. However, experimental results so far were shown for documents containing simple squares and circles but have not been tested on office documents.

Caused by the reduction in price of color laser printers in recent years, another dimension is added to the document feature space and is more and more recognized within the forensic science community. In [14] Dasari and Bhagvati demonstrated the capability to determine different printing substrates and therefore printing techniques by evaluating the documents hue component values within the HSV colour space. Another cutting edge approach investigated by Tweedy[15] and Li et al.[2,16] are yellow dotted protection patterns distributed on documents printed by color laser printers that are nearly invisible for the unaided human eye. It was demonstrated that these distinctive dotted patterns are directly related to the serial number and could be used for the identification of a particular laser printer. Nowadays, according to our observation, it can not be assumed that documents handled via high throughput systems, are exclusively printed by color laser printers. Therefore, this approach is not suitable for our purposes.

The physical characteristics of printing devices can beside the printing technology also leave distinctive fingerprints on printed documents. As recently shown by Akao et al. [17,18] the investigation of spur gears, holding and passing the paper through the printing device, can also be used to link questioned documents to suspected printers. Therefore the pitch and mutual distance of spur marks on documents was compared to already known printing devices. However this approach is also not applicable in our scenario since knowledge about spur mark distances of different devices is necessary to perform comparisons.

Judging from the literature we were able to review, so far no proper evaluation in real world scenarios of the proposed graylevel features is currently available.

2 Experiments

The evaluation described in the following, was based on four graylevel features as proposed by Lampert et al.[19] but was also covering three additional features originating from the work of Qu[20]. Within this chapter first an outline of the so far unpublished features of Qu will be presented. In a second step the experimental setup of the evaluation will be explained in detail.

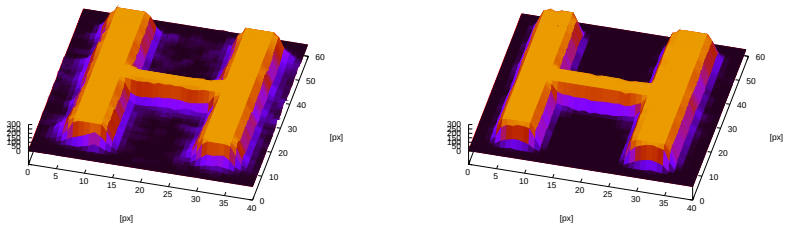

Fig. 1. Surface plots of (l) inkjet and (r) laser printed letter H, scanned with a resolution of 400*dpi*. Notice the sharper edges and cleaner surrounding of the laser printed character (r).

2.1 Features

A close look at the printed characters in Fig. 1 reveals the difference of printing signatures caused by specific printing techniques used in the creation of the document. As immediately obvious, laser and inkjet printed characters can be distinguished according to their differing edge sharpness and satellite droplets of ink. Furthermore, measuring the uniformity and homogenity of ink or toner substrate on printed areas is also a valuable feature in the detection of the used printing technique. The features of Lampert et al.[19] and Qu[20] were explicitly designed to elaborate on this observations. In the following the features proposed by Qu are explained in greater detail:

– **Perimeter Based Edge Roughness**
An approach for measuring the roughness of a character is to compare the perimeter difference of a binarized and a smoothed binarized image. For the binarization the first valley next to the lowest gray level found in the histogram of the original image is chosen as global threshhold. A character image binarized with threshhold T gives the perimeter p_b. After applying a smoothing with a median filter, the smoothed perimeter p_s can be obtained. The perimeter based edge roughness is then calculated as follows:

$$R_{PBE} = \frac{p_b - p_s}{p_s} \tag{1}$$

– **Distance Map Based Edge Roughness**
Instead of comparing simply the perimeter values of the binarized image I_b and its smoothed version I_s, this feature relates edge pixel locations via distance mapping. The distance map is initialized with the values taken from the smoothed binary image. Propagating the distances fills all entries of the distance map with the minimal distance to the nearest edge pixels of I_s.

$$DIST = \min\{d | d = \sqrt{(x-m)^2 + (y-n)^2}\}, \tag{2}$$

with $(x, y) \in I_b$ *and* $(m, n) \in I_s$. This information can be transformed into a distance histogram, where *mean, sample standard deviation, maximal and relative distance* can be calculated to form a feature vector with a relative distance defined as:

$$DIST_{rel} = \frac{\sum\limits_{x \in Edge} dist_{map}(x)}{|Edge|} \tag{3}$$

with *Edge* the set of edge pixels, and the maximal distance:

$$DIST_{max} = \max_{d \in dist_{map}} \{d - \overline{dist_{map}}\} \tag{4}$$

– **Gray Value Distribution on Printed Area**
As stated above, the differences in the uniformity of ink or toner coverage within printed regions can be used for the determination of the printing technique. To do this, a mask for the printed area is constructed by a variant of Min Max threshholding, only applied to regions containing black pixels. Afterwards a gray value histogram is extracted from the masked image. The coefficients a,b from a regression line, used to characterize the histogram, are used as the feature values.

$$b = \frac{\sum\limits_{i=0}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=0}^{n} (x_i - \overline{x})^2} \tag{5}$$

$$a = \overline{y} - b\overline{x} \tag{6}$$

with

$$\overline{x} = \sum_{i=0}^{n} x_i, \quad \overline{y} = \sum_{i=0}^{n} y_i, \quad n = 255$$

2.2 Experimental Setup

In a first step, well known *document image databases*[2] were evaluated regarding their usability for printing technique classification. Unfortunately, none of them

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

was providing an annotation scheme of the printing technique used for document creation. Therefore the necessity emerged to create our own document-database, annotated with needed *ground truth* information.

An important aspect in terms of ground truth generation is the selection or creation of a suited test-document. 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[3], 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 in database creation.

The prepared documents, consist out of 49 different laser and 14 different inkjet printouts, typically available in (home) office environments. The variety of device manufacturers exhibits all major printer manufacturers[4].

In a next step all documents were scanned using the 'Fujitsu 4860' high speed scanning device. This scanner is especially designed for high throughput scanning procedures and therefore the maximal scanning resolution is limited to $400dpi$. This is a common constraint for such devices due to reasons concerning processing performance and data storage. Therefore, all documents were scanned at $100dpi$, $200dpi$, $300dpi$, $400dpi$ and stored in the TIFF dataformat to avoid further information loss.

To perform classification on character level at least recognition and extraction of the connected components from the scanned document is necessary. After binarization of the scanned image data using Otsu's method[21], a regional growing algorithm as proposed in [22] was applied for detection. Subsequently, the minimal bounding box rectangle of each detected component was calculated and its content extracted.

For the classification of the extracted components the features outlined above were extracted from every image, giving multidimensional feature vectors for their representation. Each of the extracted features was examined through an exhaustive grid search within the features parameter space, to determine their optimal parameter setup.

According to the "no free lunch" theorem, it is desirable to evaluate classification problems with different classifier types. Therefore, the classification performance on the scanned documents was compared using implementations of the C4.5 decision tree[23], a Multilayer-Perceptron[24] and a Support Vector Machine[25]. Furthermore, the classifier parameters were also optimized in terms of high classification accuracy via corresponding grid searches.

As usual, the data was divided into a training and a test set. The classifiers were trained with features extracted from 6 randomly selected inkjet and laser printouts. For all scanned resolutions a ten-fold cross validation using stratified sampling, to avoid overfitting of the learned model, was performed. Let T be the training set and $c_1, ..., c_N$ the training vectors within T. For performance

[3] exhibiting the following characteristics, fonttype *'Courier New' normal*, fontsize *12 pt*, lineheight *12 pt*

[4] Hewlett-Packard, Epson, Canon, Ricoh, etc.

comparison the accuracy mean of all training runs M as defined in equation (7) was obtained for each training and resolution.

$$accuracy_T = \frac{\sum_j^M (\frac{\sum_i^N x_{i,T,cor}}{N_T} * 100)}{M} \tag{7}$$

Where $x_{i,T,cor}$ refers to a correct classified character in the training set T of training run j and N_T specifies the total amount of characters within T.

Within the testing phase the learned model of the corresponding classifiers was applied to 7 inkjet and 14 laser printouts. The test printouts were created by different printing devices and randomly selected. The performance of the classifiers was obtained in the following manner. Let D be a scanned document image and $c_1, ..., c_N$ the extracted characters of D. To compare classification performance on document level, the accuracy rate as stated in equation (8) was calculated for each test document.

$$accuracy_D = \frac{\sum_i^N x_{i,D,cor}}{N_D} * 100 \tag{8}$$

$x_{i,D,cor}$ refers to a correct classified character in D and N_D specifies the total number of characters within D.

3 Evaluation Results

In the following the evaluation results obtained from the experimental setup as described in Sec.2.2 are presented. Therefore, results of the training and the testing phase for each classifier are discussed.

3.1 Decision Tree classification

The decision tree classification evaluation was based on postpruned trees. Therefore, the confidence threshold for pruning was set to 25% and the minimum number of instances per leaf was set to 2.

Training results: Fig. 2 shows the classification performance of every single feature as well as for all features in the training phase. It can be observed, that features like *edge dist* reach higher classification accuracy with increasing scan resolution. However, features like *area diff*, achieve high classification accuracy at low resolutions. Interestingly an accuracy rate of nearly 95% at 100*dpi* using all features in combination was reached.

Testing results: Fig. 3 depicts the appliance of all features to the 21 test documents resulting in the quartile accuracy box plot on the left. Performing a pca on the test set, the three most discriminating features, namely *cooc lbp,*

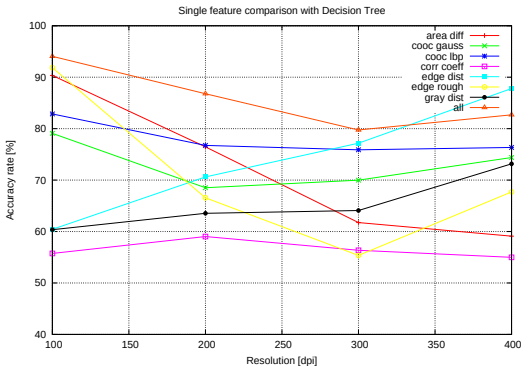

Fig. 2. Accuracy rate for all tested features with a C4.5 decision tree using optimized feature extraction and classification parameters.

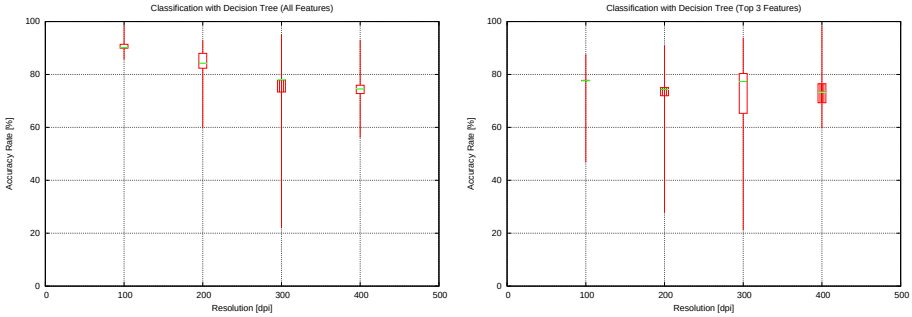

Fig. 3. Box plot of a C4.5 decision tree classification for a combination of all (l) and the 3 most discriminating features as identified by pca (r).

cooc gauss and *edge distmap*, could be identified. Classification results for these 3 features taken from the 21 test documents are visible at the right plot of Fig. 3.

Comparing the classification results, it can be observed that using fewer features leads to a decreasing accuracy for the resolutions. But still $75-80\%$ of a documents characters are recognized correctly for all tested resolutions.

3.2 Support Vector Machine (SVM) classification

Classification experiments using a SVM were based on a radial-basis kernel function using optimized parameters. The parameters for C and γ were obtained by coarse grid searching the SVM parameter space within the intervals $C = [2^{-5};2^{15}]$, $\gamma = [2^{-15};2^{3}]$ and for each of the scanned resolutions.

Training results: Overall, the classification results in Fig. 4 are slightly lower than for decision trees considering single features. Also a more constant devel-

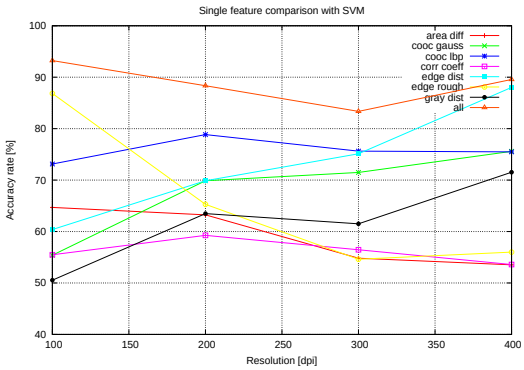

Fig. 4. Accuracy rate for all tested features, classified by a SVM using a rbf kernel and optimized parameters for feature extraction and classification.

opment of the curves for resolutions $> 200dpi$ can be observed (Fig. 2). Furthermore, a higher classification accuracy at $400dpi$ using all features is achieved.

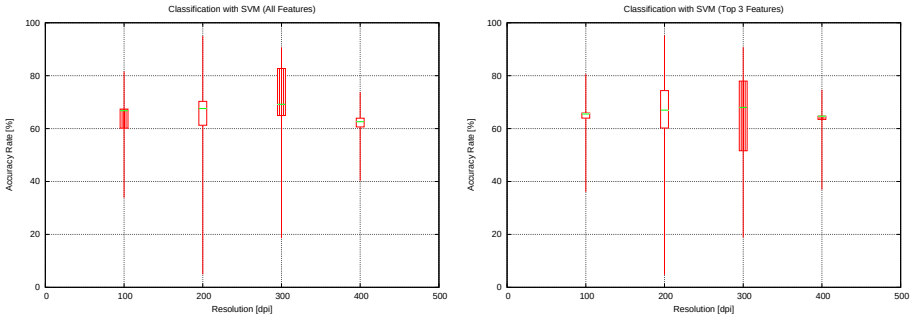

Fig. 5. Box plot of a SVM classification with a combination of all (l) and the 3 most discriminating features as identified by pca (r).

Testing results: As for the single feature evaluation, the box plots in Fig. 5 show a lower performance for all and the top 3 features. Only $60 - 70\%$ are classified correctly.

3.3 Multi-Layer Perceptron (MLP) classification

For the classification using a feed forward MLP, the learning rate of the backpropagation was set to 0.3, which lead to high classification accuracy. Furthermore, within every run the MLP has been trained with 500 training epochs.

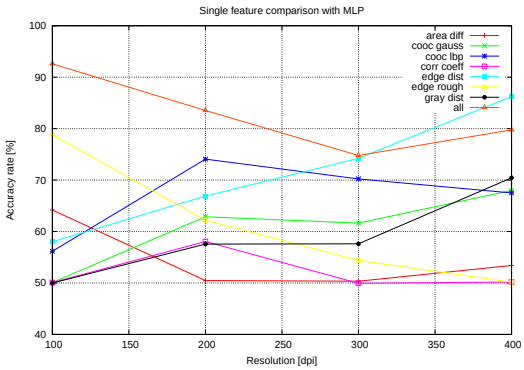

Fig. 6. Accuracy rate for classification of all tested features using a MLP.

Training results: Similar to the SVM, the accuracy rates achieved using a MLP for classification are slightly lower than using the C4.5 decision tree (Fig. 6). Even though some of the features reached higher accuracy rates at low scan resolutions, i.e. *area diff* and *edge rough*, while especially the *edge dist* feature is performing best at 400*dpi*. Surprisingly, using all features in combination yields a lower performance than using only the *edge dist* on its own refering to the "ugly duckling" theorem.

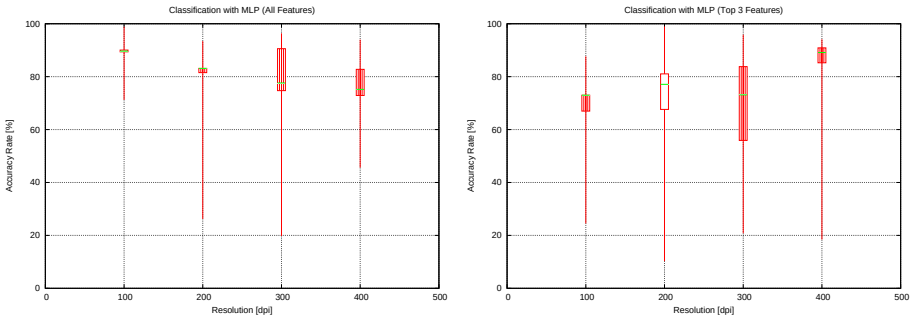

Fig. 7. Box plot of a classification with a multi-layer perceptron for a combination of all (l) and the 3 most discriminating features as identified by pca (r).

Testing results: The classification using all features shows a strong influence of the *edge rough* feature exhibiting only little variance at 100*dpi*. Since this feature is not among the top 3 pca features, the accuracy for this resolution drops down a lot (Fig. 7). Regarding results at 400*dpi* the top 3 features improved the accuracy rates significantly in comparison to experiments including all features.

4 Conclusion

We have presented a quantitative evaluation of common texture and edge based gray-level features for digital printing technique recognition, under the aspect of usability for high-throughput DMSs. The evaluation indicates that printing technique recognition is possible, even from low resolution scans which are specific to such systems. As the graphs for the single feature evaluation indicate, the examined features perform differently well for the tested resolutions. Therefore the appropriate feature set has to be picked for certain scan resolutions. It was also shown that the classifier used has influence on the feature performance. Furthermore, it could be demonstrated that even classification methods needing only short training, i.e. decision trees, were able to provide high classification accuracy. Due to the constraints of high-throughput document management systems so far only gray-scale scanned documents have been investigated. The examination of color properties of such documents will be part of our future work. Additionally, features capable for examining the differences between a laser print and a copy of a document will be developed. Furthermore the influence of the actually printed shape i.e. even edges versus round edges and its impact on classification accuracy has to be further elaborated.

References

1. Chim, J.L.C., Li, C.K., Poon, N.L., Leung, S.C.: Examination of counterfeit banknotes printed by all-in-one color inkjet printers. Journal of the American Society of Questioned Document Examiners (ASQDE) **7**(2) (2004) 69–75
2. Li, C.K., Leung, S.C.: The identification of color photocopiers: A case study. Journal of the American Society of Questioned Document Examiners (ASQDE) **1**(1) (1998) 8–11
3. Makris, J.D., Krezias, S.A., Athanasopoulou, V.T.: Examination of newspapers. Journal of the American Society of Questioned Document Examiners (ASQDE) **9**(2) (2006) 71–75
4. Parker, J.L.: An instance of inkjet printer identification. Journal of the American Society of Questioned Document Examiners (ASQDE) **5**(1) (2002) 5–10
5. Kelly, J.S., Lindblom, B.S.: Scientfic Examination of Questioned Documents. 2 edn. CRC Press (2006)
6. Osborn, A.S., Osborn, A.D.: Questioned documents. Journal of the American Society of Questioned Document Examiners (ASQDE) **5**(1) (2002) 39–44
7. Hilton, O.: Scientific Examination of Questioned Documents. 1 edn. CRC Press (1993)
8. Ellen, D.: The Scientific Examination of Documents. 2 edn. Taylor and Francis (1997)
9. Nickell, J.: Detecting Forgery: Forensic Investigations of Documents. 1 edn. University Press of Kentucky (2005)
10. Oliver, J., Chen, J.: Use of signature analysis to discrimnate digital printing technologies. In: Proceedings of the IS&T's NIP18: International Conference on Digital Printing Technologies. (2002) 218–222

11. Mikkilineni, A., Ali, G., Chiang, P.J., Chiu, G.C., Allebach, J., Delp, E.: Printer identification based on texture features. In: Proceedings of the IS&T's NIP20: International Conference on Digital Printing Technologies. Volume 20., Salt Lake City, UT (2004) 306–311
12. Mikkilineni, A., Ali, G., Chiang, P.J., Chiu, G.C., Allebach, J., Delp, E.: 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., San Jose, CA (2005) 430–440
13. Tchan, J.: The development of an image analysis system that can detect fraudulent alterations made to printed images. In van Renesse, R.F., ed.: Optical Security and Counterfeit Deterrence Techniques V, Proceedings of the SPIE, Volume 5310, pp. 151-159 (2004). Volume 5310 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference. (jun 2004) 151–159
14. Dasari, H., Bhagvati, C.: Identification of printing process using hsv colour space. In Narayanan, P.J., Nayar, S.K., Shum, H.Y., eds.: ACCV (2). Volume 3852 of Lecture Notes in Computer Science., Springer (2006) 692–701
15. Tweedy, J.S.: Class chracteristics of counterfeit protection system codes of color laser copiers. Journal of the American Society of Questioned Document Examiners (ASQDE) **4**(2) (2001) 53–66
16. Li, C.K., Chan, W.C., Cheng, Y.S., Leung, S.C.: The differentation of color laser printers. Journal of the American Society of Questioned Document Examiners (ASQDE) **7**(2) (2004) 105–109
17. Akao, Y., Kobayashi, K., Sugawara, S., Seki, Y.: Discrimnation of inkjet printed counterfeits by spur marks and feature extraction by spatial frequency analysis. In van Renesse, R.F., ed.: Optical Security and Counterfeit Deterrence Techniques V, Proceedings of the SPIE, Volume 5310, pp. 151-159 (2004). Volume 5310 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference. (jun 2004) 129–137
18. Akao, Y., Kobayashi, K., Seki, Y.: Examination of spur marks found on inkjet-printed documents. Journal of Forensic Science **50**(4) (July 2005) 915–923
19. Lampert, C.H., Mei, L., Breuel, T.M.: Printing technique classification for document counterfeit detection. In: Computational Intelligence and Security (CIS) 2006, Ghuangzhou, China. (2006)
20. Qu, L.: Image based printing technique recognition. Project Thesis (May 2006)
21. Otsu, N.: A threshold selection method from gray level histograms. IEEE Transactions on Systems, Man and Cybernetics (9) (1979) 62–66
22. Gonzalez, R.C., Woods, R.E. In: Digital Image Processing. 3 edn. Prentice Hall International (2007)
23. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
24. Rumelhart, D., Hinton, G., Williams, R.: Learning Internal Representations by Error Propagation in DE Rumelhart, JL McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations. Foundations MIT-Press (1986) 318–362
25. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.