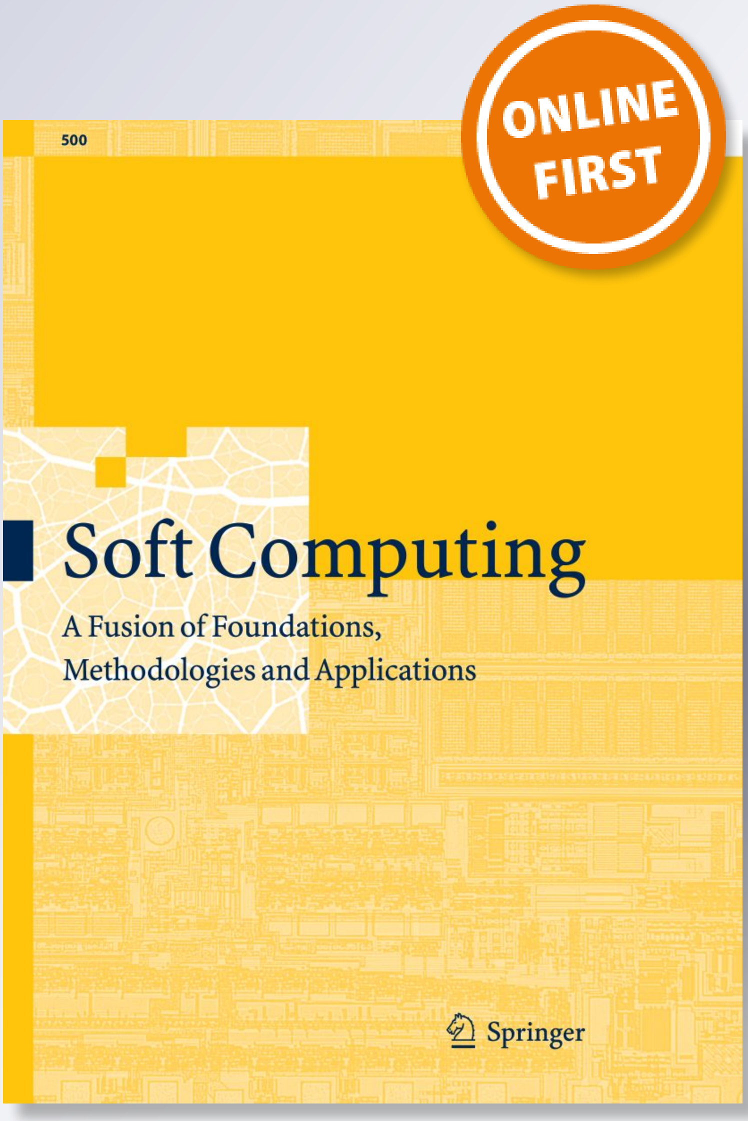
Forensic document examination system using boosting and bagging methodologies

Surbhi Gupta & Munish Kumar

Soft Computing
A Fusion of Foundations,
Methodologies and Applications

ISSN 1432-7643

Soft Comput
DOI 10.1007/s00500-019-04297-5



Soft Computing
https://doi.org/10.1007/s00500-019-04297-5

METHODOLOGIES AND APPLICATION

Forensic document examination system using boosting and bagging methodologies

Surbhi Gupta[1] · Munish Kumar[2]

Abstract
Document forgery has increased enormously due to the progression of information technology and image processing software. Critical documents are protected using watermarks or signatures, i.e., active approach. Other documents need passive approach for document forensics. Most of the passive techniques aim to detect and fix the source of the printed document. Other techniques look for the irregularities present in the document. This paper aims to fix the document source printer using passive approach. Hand-crafted features based on key printer noise features (KPNF), speeded up robust features (SURF) and oriented FAST rotated and BRIEF (ORB) are used. Then, feature-based classifiers are implemented using K-NN, decision tree, random forest and majority voting. The document classifier proposed model can efficiently classify the questioned documents to their respective printer class. Further, adaptive boosting and bootstrap aggregating methodologies are used for the improvement in classification accuracy. The proposed model has achieved the best accuracy of 95.1% using a combination of KPNF + ORB + SURF with random forest classifier and adaptive boosting methodology.

Keywords Document forensics · Printer forensics · KPNF · SURF · ORB · AdaBoost · Bagging

1 Introduction

Digital documents and their use have become increasingly dominant in the present era. It is almost impossible to avoid their use these days. These digital documents could be official contract images, bills and checks, etc. Paperless world is the objective behind these digital documents. Moreover, a digital document is easy, economical and efficient to maintain as compared to a hard copy, but its security is a challenge. Manipulation of digital documents has increased enormously due to the progression of information technology and image processing software. Document analysis and its authentication is a critical challenge. Important documents such as bank cheques, educational certificates, passports have watermarks on them which authenticate the digital document. Although active technologies dominate in this domain, still passive analysis for unprotected documents is always required. Active technologies add some security features in the digital document. Active approaches mainly use digital signatures or watermarks as defensive measures to protect the images. Originality and legitimacy of images and digital documents can be checked using the watermark or signature embedded in the image. Active techniques are used to protect copyright images and important documents. In active approach, a watermark or signature code is embedded in the image itself. It is embedded in the form of bits. These bits can later be extracted and checked to validate the image (Tayan et al. 2014). Forensics are basically divided into two categories, namely active or passive. These are briefly discussed in sub-sections.

1.1 Active forensics

The most popular active forensics approach for digital image authentication is watermarking. In watermarking

Communicated by V. Loia.

✉ Munish Kumar
munishcse@gmail.com

[1] Department of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India

[2] Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India

Published online: 14 August 2019

technique, some signal or data is embedded directly into the audio, image or video signal which needs protection. Embedded data are usually the extra information that travels with the digital signal. This extra information is usually embedded as discontinuous bits in the signal block. Whenever the original signal undergoes any alteration, the embedded data get lost or modified. Thus, watermarking provides an indication of ownership of the object. Embedded extra information is visible in the image in the case of visible watermarking. In the case of invisible digital watermarking, message is hidden as digital data in audio, picture or video as encoded bits. Invisible watermarks are not apparent, but their presence may be identified (Tao et al. 2014). A watermarking system usually has an encoder and a decoder. Host signal along with watermark and security key is fed to the encoder. The encoder inserts the watermark into the host signal. Various algorithms are available to perform the embedding of the watermark. The outputs of the encoder are the security key and the watermarked contents. A watermark extractor or detector involves a two-step process. First, some scrambling algorithm is applied to extract the watermark from the received signal. Then, the extracted unreliable watermark is analyzed and compared with the original one. Finally, confidential assessment is done to authenticate the host signal (Bianchi and Piva 2013; Tao et al. 2012). The main characteristics of a watermark are imperceptibility, capacity, security, robustness and false positives. A number of different watermarking techniques based on spatial and transform domain have been contributed by researchers. Spatial domain watermarking is attractive because of its simplicity and pre-assessment to robustness, capacity and imperceptibility. However, spatial domain-based solution is not so robust. A frequency domain-based solution like singular value decomposition (SVD) and discrete wavelet transform (DCT) are used as alternative techniques for decomposing images for watermarking (Cox et al. 2000). Another popular approach which is widely used for digital document authentication and copyright is digital signature. The digital signature is a mechanism to provide the proof of legitimacy of the originator. Digital signatures are also unique as of handwritten signatures. Digital signature needs a public and a private key. Private key is used to produce digital signatures by the signing authority, whereas the public key is used by the receiver of the signed document to decrypt the signature. The electronic signature is created using the private key owned by the signer. The signer needs to keep the private key secretly. Then, a hash algorithm is applied to create hash data corresponding to the signed document. Afterward, an encryption algorithm is used to encrypt the hash data. This encrypted data are known as a digital signature. A time stamp is also associated with the digitally signed document. If the document is modified after the signature, the digital signature gets distorted. For example, a service provider signs an agreement with his customer to provide some services. The service provider will be using his private key to generate the signature. Then, the customer receives the document and the public key. If the public key is not able to decrypt the signature, it means the signature is not authentic and the signature is considered invalid. Digital signature schemes are either symmetric or asymmetric-key systems. They ensure content legitimacy, reliability, and data privacy during transmission. The two most commonly used public-key digital-signature schemes are the Rivest–Shamir–Adleman (RSA) public-key encryption algorithm and the digital signature algorithm (Tayan et al. 2014; Subramanya and Yi 2006).

1.2 Passive forensics

Most of the available active methods embed security features in images/documents. These methods are costly and practically difficult to use. The need is to have easy, fast and low cost solutions to detect forged images/documents. A passive approach detects the image/document authenticity based on its intrinsic fingerprints. It does not use any preventive measure in advance. Most of these techniques are based on capturing the acquisition fingerprints of the digital document. These fingerprints utilize the traces left by acquisition device used to produce a digital document. A digital acquisition device has various components. These components tend to alter the input signal in some particular ways and leave intrinsic fingerprints in the image. Camera optical system, the image sensor and camera software have their unique fingerprints. Even if the acquisition step remains same, but still the fingerprints of sensors and camera may differ due to use of hardware from different manufacturers. Imaging sensors in source devices have various defects which may result in disturbances in the pixel intensity values. The sensor noise could be present due to pixel defects, fixed pattern noise or photo response non-uniformity (Bayram et al. 2005, 2008). The absence of coherence in sensor noise can be taken as a clue of printed document forgery. Some researchers presented texture features for printer recognition. Others used local features, such as line edge unevenness, area difference and relationship coefficients, for individual characters present in the image (Mikkilineni et al. 2004, 2005a; Lampert et al. 2006).

In the present study, the noise and other key features of printed documents are analyzed for the printer identification. The proposed technique used printer noise, ORB and SURF features to examine the forensic documents printed by different printer resources. Different combinations of these feature extraction methods are also explored. Three classification methodologies, namely the *k*-NN, decision tree, random forest and majority voting of these three classification techniques, are considered for the classification task. Further, the classification results are improved using adaptive booting and bootstrap aggregating methodologies. This paper is divided into nine sections. Introduction of present work is discussed in Sect. 1. Section 2 presents work related to document forensic examination. Feature extraction techniques and classification techniques are presented in Sects. 3 and 4, respectively. Section 5 depicts the introduction about adaptive boosting and bootstrap aggregating methodologies. Block diagram and working of the proposed system are presented in Sect. 6. Experimental results based on the proposed system are depicted in Sect. 7. Section 8 presents the comparative study between state-of-the-art work and proposed work. Finally, in Sect. 9, authors have presented the conclusion of the present work.

2 Related work

The main approaches for printed document classification are halftone-based detection, texture-based detection, and printer noise-based detection.

2.1 Halftone-based detection

Bulan et al. (2009) utilized association between geometric degradation due to laser printing for document forensics. This artifact was highlighted based on the difference of the region that a printer should have printed and the region that it actually printed. Geometric traces of printers were extracted by the dot positions in halftone documents. The positions of points in the test image were correlated. The performance of the proposed method was evaluated for the printer model identification. A database of printer signatures was generated by extracting geometric distortion signatures from several documents from each printer type. The mean of signatures was considered as printer signature. Geometric distortion signatures of test documents exhibited a high correlation with the corresponding printed signatures. A low correlation was exhibited for cases where the document was matched with other printer signatures. Wu et al. (2009) utilized the geometric degradation for the recognition of source printers. The proposed model performed feature extraction from the whole document image. A projective transformation was modeled to represent the geometric degradation. The center of letters was used to extract the degradation from scanned document and its .tiff image version. This model used singular value decomposition and removal of outliers. The resulting features were used to link the document and its source printer. A subset of the features was used as input feature vectors to a machine learning classifier. Twelve pages per printer (total 10), i.e., total 120 pages, were printed. The experimental results demonstrated the acceptable classification accuracy, but the considered dataset was small. Ryu et al. (2008) proposed identification of halftone texture in high resolution, scanned color documents. The histograms of angles from Hough transforms were calculated from each CMYK band. The document and its source were mapped based on high correlation. The evaluation of the proposed technique was performed on 9000 images obtained by 9 electrophotographic process (EP) printers. Forty different images were printed from each printer with 600 dpi.

2.2 Texture-based detection

Banding effects present on the document were studied by Ali et al. (2004). It was discussed that EP printers show signs of quasi-periodic banding effects. This approach was efficient for colored documents but not for text documents. These documents contain only a small range of gray levels. Mikkilineni et al. (2004) used low-cost printers for their experiments. The intrinsic signature was extracted from a high-resolution scanned image. It was then used to design and drive the extrinsic signature. The identifying information such as the printer serial number and date of printing was encoded during the document printing. These embedded watermarks acted as extrinsic signature of image. Mikkilineni et al. (2005a) used gray level co-occurrence matrices (GLCM) statistics based on textural features to identify the source of text documents. Documents were scanned at 2400 dpi with eight bits by pixel, and statistical features from GLCM were extracted for each ‘e’ character. Source printer classification was done using 5NN classifier. But the presented technique required the prior information about the printers in question. The limitation was that if the document source printer was not present in the classifier training data set, then it was mistakenly classified as one of the known printers. This technique was not robust to multiple font sizes, font types and different characters. Mikkilineni et al. (2011) used clustering and Euclidean distance to classify documents from different printers. Forensic printer identification was performed to find whether a document may or may not belong to a known set of printers. A classifier for printer identification was proposed using intrinsic signatures of the printers. The intrinsic signature was based on electromechanical properties of the printer. These signatures were unique to each printer and so it was difficult to forge them. Sequential feature selection and linear discriminant analysis were used to reduce the feature dimensionality. Tsai and Liu (2013) used GLCM statistics along with wavelet transform features. A specific character of the Chinese language was used for the texture pattern extraction from the scanned document. The Chinese printed resources were analyzed in order to find the source of printers. The feature selection technique and SVM were used to propose the source model of the documents. The average source identification rate was 98.64%. The proposed identification method was very useful for laser printer source identification. Ferreira et al. (2015) proposed three variant techniques for laser printer identification. The solutions used low-resolution scanned documents. First proposed method used two descriptors based on multi-directional and multiscale texture properties from micro-patterns. These descriptors were obtained from either letters or regions of interest. The inner part of printed letters was focused. Convolution texture gradient filter (CTGF) was proposed as a second descriptor. The CTGF is the histogram of low-level gradient filtered textures. Texture artifacts were investigated on segments of a document. These segments were called frames. The advantage of the third approach was that the printing source of a document was identified even if parts of it were unavailable. The accuracy of the first approach was 97.60%, 98.38% and 88.58% for characters, frames and documents, respectively. The accuracy of 94.19% and 88.45% was obtained for frames and documents, respectively. A new document dataset was proposed which is freely available for experimentation.

Tsai and Yuadi (2018) performed printed source identification using microscopic images. A detailed texture and structure information was obtained due to high magnification of the document image. It was stated that microscopic techniques could retrieve the shape and surface texture of a printed document. The proposed approach utilized image processing techniques and statistical features like local binary pattern (LBP), gray level co-occurrence matrix (GLCM), discrete wavelet transform (DWT), spatial filters, Haralick, and segmentation-based fractal texture analysis (SFTA) features. LBP approach achieved the highest source identification rate of 99.89%. Joshi and Khanna (2017) mentioned that while examining the documents, most of the approaches required the original/authentic documents to compare the character font. A local texture descriptor-based approach was proposed by them. The experimental results indicated that the techniques performed best for character printed in the same font setup. It achieved better recognition for printers of the same brand and model.

2.3 Printer noise-based detection

Khanna et al. (2007) performed camera image forensics based on scanner noise analysis. A unique noise pattern of each scanner brand was used for source device identification. First set of proposed features consisted of statistical properties such as mean, median, standard deviation, skewness and kurtosis. The periodicity between different rows of the fixed component of the sensor noise of a scanned image was detected using correlation. Second set of proposed features consisted of statistical properties of these correlations. 16D feature vector was obtained for each scanned image. These features captured the essential properties of the image and discriminated between different scanners. The second set of features represented the fixed component of the pattern noise. For low-quality scanners, a large amount of random noise was present in the document because of fluctuations in lighting conditions. The inter-row correlation in this case was quite small as compared to a high-quality scanner. This method obtained promising results for detecting spliced/forged documents made through the combination of two or more different document images. Elkasrawi and Shafait (2014) extracted features from the noise image, similar to Khanna et al. (2007). Ali et al. (2004) approach was extended with more number of text lines in the document. Low-resolution scanners were used for printer identification. The statistical features of pattern noise, produced by flatbed scanners, were used. The text lines were extracted using Tesseract-OCR (Smith 2007). Then the noise patterns were extracted from the filtered image subtracted from the original image. A binary image was obtained using Otsu threshold and median filtering. 15D feature vector was obtained based on mean, standard deviation, skewness and kurtosis. Features were extracted row- and column-wise from the gray image. The advantage of proposed statistical features was their independence on image content or size. Three sets of experiments were conducted. The average accuracy obtained for binary classification of inkjet and laser printer was 93:57% and 78:46%, respectively. The overall accuracy was low as the number of printers considered increased.

2.4 Other methods

Ryu et al. (2008) used image quality measures for document forensics. Different measures related to pixel differences, similarity between two images, frequency domain characteristics and human visual system characteristics were proposed. SVM print classifier was used for classifying the questioned document as real or fake. One laser printer, one inkjet printer and one scanner were used for experimentation. The SVM classifier was evaluated with the data sets prepared from the combination of all printers and scanner. The classifier achieved an accuracy of 80%. Kee and Farid (2008) modeled geometric deterioration resulted due to document printing. Frequently occurring letters were used for study. A simple scanner of 300 dpi resolutions was used for scanning the documents.

Document tampering detection and source identification were achieved. A model based on a set of degraded characters was presented. This printer profile was exploited for source printer recognition and local manipulation detection in a document. Proposed technique was capable to distinguish between printers of different make and model, but it could not differentiate printers of the same make. Schulze et al. (2008) examined the printed characters quality. It was assumed that different printing techniques leave different effects on the printed text. Textural and edge-based gray level features were used. Low scan resolution scanners were used that were usually used for high throughput scanning systems used for document management systems (DMS). The goal was to utilize the proposed features for low-resolution document scans. Other aspects like paper quality, ink type and document aging were ignored. Forty-nine laser and 13 inkjet printers were evaluated. Examined features performed better as compared to existing solutions. But the appropriate feature set needs to be chosen for different scan resolutions before use. The feature performance was dependent on the selection of classifier. Basic classification methods like decision trees provided high classification accuracy.

Schreyer et al. (2009) used discrete cosine transform (DCT) features to characterize photocopied, inkjet and laser printed documents. This technique extracted statistical features in the noise image, gradient image, the DCT image and the multi-resolution wavelet image. Document image noise was obtained by using mean, median and Gaussian filtering techniques. Mean, standard deviation, correlation and mean squared error were calculated from the original and de-noised document image. Gradient analysis was performed for extracting statistical information about fine image intensity variations corresponding to character edges and noisy image regions. Different gradient filters were applied, and gradient histogram was calculated for each document image. Mean and standard deviation are obtained for different histogram intervals. DCT frequency analysis was performed using mean and standard deviation of DCT coefficients. Multi-resolution wavelet analysis was performed using the Haar, Daubechies wavelets and Coiflets. Mean and standard deviation were obtained at different scales of wavelet decomposition. These features were used as feature vectors for machine learning classifiers. Image classification was done using multilayer perceptron (MLP) and SVM. The highest classification accuracy was achieved using DCT frequency analysis and SVM classification. It was 92.92% and 99.08% at 400 dpi and 800 dpi, respectively.

Choi et al. (2009) proposed forensic analysis of wavelet transform statistical analysis in the RGB and CMYK images to identify the source of color documents. Color laser printer identification scheme was presented for halftone images. Skewness, kurtosis and correlation features were used to train a SVM classifier. Results were obtained for a non-public dataset containing printouts from 9 different printers. Color printing techniques were different for each brand. The images were categorized into 4 image sets depending on the brand. The average classification accuracy achieved for color laser printer brand, color toner and color laser printer model was 97.89%, 92.28%, and 80.24%, respectively. Van et al. (2009) detected document irregularity. The text lines in questioned documents were examined for disarrangements to detect tampering. A line extraction algorithm was presented to detect the skew and orientation. The effectiveness of the method was illustrated on University of Washington-III (Phillips 1996) document dataset. A total of 159 images from the dataset were considered for experimentation. The proposed technique was evaluated against an available, open source, orientation detection technique. The proposed method was more efficient when tested for UW-I dataset and their own dataset. The main advantage of the proposed method was that orientation and skew were estimated in one step. Van et al. (2013a) extended this technique. An automated approach based on text-line rotation and alignment features was proposed for the verification of documents. Experimental evaluation of the proposed technique achieved the area under curve of 0.89%. The proposed approach was useful especially for high-volume documents due to its automatic nature.

Jiang et al. (2010) proposed Benford's law-based 9D feature vector for printer forensics. The printer's make and model were detected using the Benford's law. The first digit probability distribution of DCT coefficients were extracted from printed and scanned images. It was emphasized that a good forensic feature should be independent of the image content and it must be robust to the random noise. The ideal classifier must have high efficiency with less numbers of features. Proposed classifier used only nine forensic features based on the Benford's law characteristics. The obtained printer identification rate was 94% for five distinct printer brands. Tsai et al. (2011) used a similar strategy, but only the RGB color space was considered. Similarly, Bertrand et al. (2013) examined font similarity and deviations of characters in a questioned document to detect document forgery. The character shapes were compared, and the structural irregularities were detected in the document. Two indicators were used to identify forged documents. The extracted characters that were many similar or much different in shape were identified. The detection of copied and pasted region was done by character shape comparison. The difference between two character shapes was a distance obtained between their feature vectors. Binary low-resolution documents were manipulated roughly for evaluation of the proposed technique. The used intrinsic features were computed from the characters of the document. The second goal was to detect irregularities in the structure of the document. Character misalignment, different character size, character position or character orientation were examples of the inaccuracies. Software was generated to create fraudulent document images. Recall and precision values obtained for document forgery detection were 0.77% and 0.82%, respectively.

Van et al. (2013b) classified printers based on yellow point patterns in a document. These yellow dots were specific to a particular printer manufacturer. Printer class was detected by comparing two basic patterns from two different document printouts. It was verified whether the two printouts were from the same printer or not. Decoding was done automatically to find the serial number, the time and the date of the printed document. The document dataset aimed on tracking patterns. These tracking patterns were called Machine Identification Code. Proposed database contained 1264 images printed from 132 printers. Accuracy of 93% was achieved for printer classification. The proposed pattern tracking scheme achieved an accuracy of 91.3% and 98.3% for comparison and decoding, respectively. Gebhardt et al. (2013) used a similar approach to examine the character edges. The documents were characterized as either laser or inkjet based on the variance in the pixel gray-level. Edge roughness was taken as the major identity for a character printed by a printer. The character edges were checked for the fluctuations in gray levels. Local feature extraction based on optical character recognition (OCR) as a preprocessing step was also proposed. The Tesseract-OCR (Smith 2007) engine was used to extract the characters. Proposed features were aimed to identify a source for printed documents. These documents were scanned at a very low-resolution (400 dpi) scanner. No prior training was required for the classifier. A new dataset with documents printed from different inkjet and laser printers was generated. Each printer printed twenty different content pages of type contracts, invoices and scientific literature.

Li et al. (2018) have proposed a novel inkjet printer source identification using a print sample. They studied 15 low-cost inkjet printers at a microscopic level. They considered four printer intrinsic features, dot size, dot density, average distance to nearest dot and nearest dot sector. For classification, they considered support vector machine classifier and claimed to achieve reliable results.

3 Feature extraction techniques

Document authentication is usually based on feature extraction. In this case, the examined document is the input data. This data are processed to extract features for relevant information required to characterize the used printer. The classification of documents based on these features is the output of the system. Two basic categories in document examination are local and global features. Local features examine and analyze the connected components (CCs) or characters of the document (Amer and Goldstein 2012), while global features examine the whole document at once. We have used key printer noise features, oriented (FAST) rotated BRIEF and speeded up robust features.

3.1 Key printer noise features (KPNF)

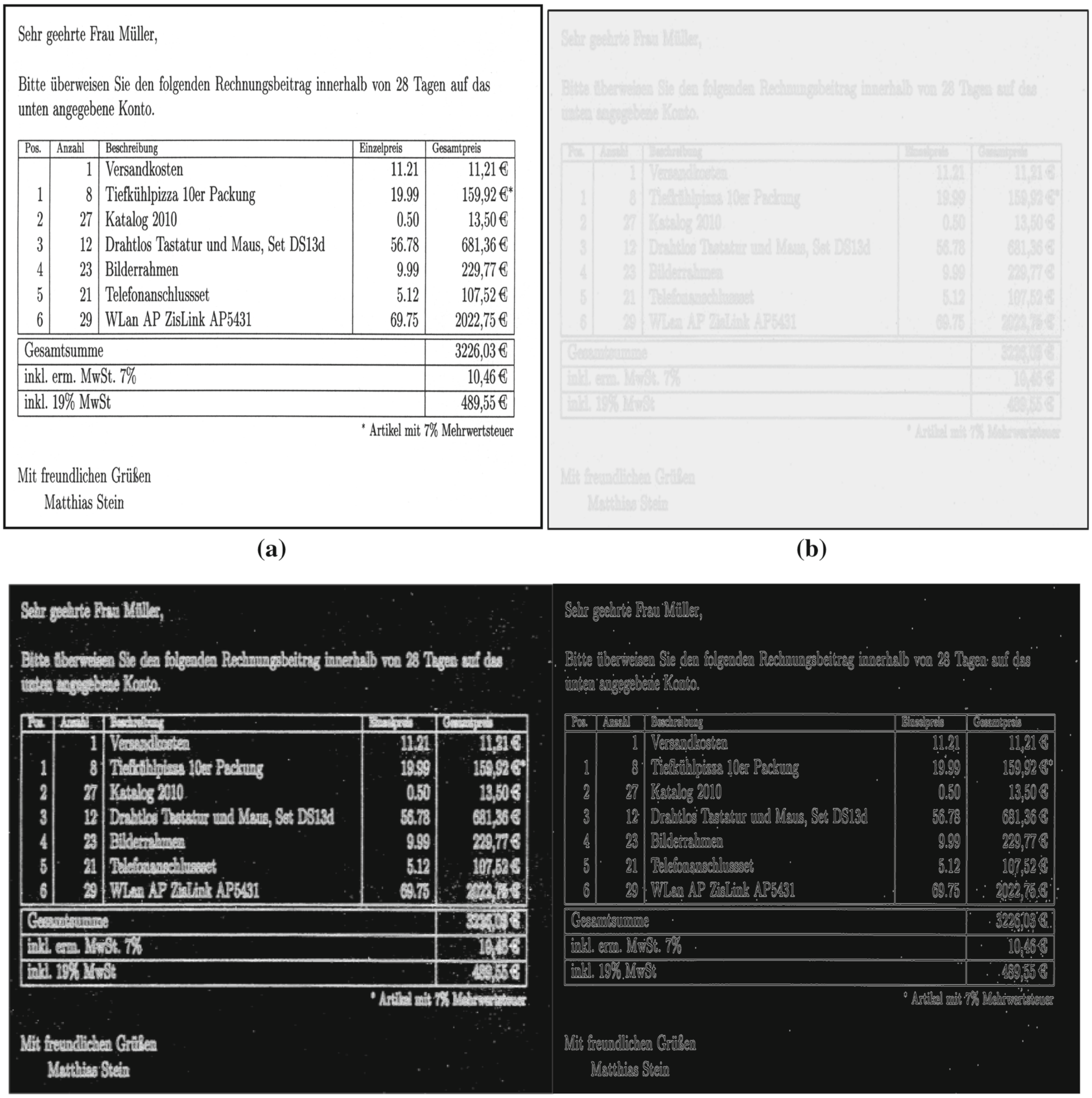
The proposed key printer noise features (KPNF) technique uses global features such as noise and texture to classify the documents printed by different printer resources. The noise present in the printed document is the inherent characteristic of the printer used. Figure 1 represents the noise and edge images obtained for a sample document. These images are used for feature extraction. This noise is extracted and used for extraction of features such as mean and standard deviation from the de-noising filers. Average, Median, Gaussian and Weiner filters are used as de-noising filters to obtain 8D features.

3.2 Speeded up robust features (SURF)

SURF is a local feature extraction method. For extracting image feature key points, it utilized local invariant fast key point detector and for extracting image feature descriptor, it utilized distinctive descriptor. Key point-based technique is very useful to overcome the limitations of the block-based methods. Speeded up robust features (SURF) are widely used to analyze images. SURF is a local feature extraction method. Its computational complexity is low as they follow a non-block approach. It utilized distinctive descriptor (Cedillo-Hernandez et al. 2013) for extracting image feature descriptor. It works by extracting the feature key point from an image as per application. Next, the orientation is assigned to these key points. The circular orientation is assigned with respect to the interested key points. Then, the squared area is tuned according to the selected orientation. Lastly, Haar wavelet responses are used to extract feature descriptor. An 8D feature vector is extracted as a descriptor vector.

3.3 Oriented FAST rotated BRIEF (ORB)

ORB is another popular local feature extraction method. It used FAST (Features from Accelerated Segment Test) key point detector for extracting image feature key points. In ORB, the Harris corner detector method is used to find out best interested points from those which are detected by FAST key point detector. It utilized Binary Robust Independent Elementary Features (BRIEF) descriptor for



Sehr geehrte Frau Müller,

Bitte überweisen Sie den folgenden Rechnungsbeitrag innerhalb von 28 Tagen auf das unten angegebene Konto.

Pos.	Anzahl	Beschreibung	Einzelpreis	Gesamtpreis
	1	Versandkosten	11.21	11,21 €
1	8	Tiefkühlpizza 10er Packung	19.99	159,92 €*
2	27	Katalog 2010	0.50	13,50 €
3	12	Drahtlos Tastatur und Maus, Set DS13d	56.78	681,36 €
4	23	Bilderrahmen	9.99	229,77 €
5	21	Telefonanschlussset	5.12	107,52 €
6	29	WLan AP ZisLink AP5431	69.75	2022,75 €

Gesamtsumme	3226,03 €
inkl. erm. MwSt. 7%	10,46 €
inkl. 19% MwSt	489,55 €

* Artikel mit 7% Mehrwertsteuer

Mit freundlichen Grüßen
Matthias Stein

(a) (b)

(c) (d)

Fig. 1 A sample of **a** original, **b** noisy, **c** logarithmic, **d** edge document image

extracting image feature descriptor (Vinay et al. 2015). The direction of patch is used for rotation on binary test patterns. The cost of computation is less as compared to SIFT and SURF, but the magnitude is faster than SURF. ORB extracts fewer but the topmost best features from an image. For finding stable interesting key points in image using FAST detector, it uses intensity of pixel value and threshold value. ORB extracts the best features from an image. Finally, an 8D feature vector is extracted as the length of the descriptor vector similar to SURF. The ORB algorithm has lower cost of computation and faster magnitude as compared to the SURF algorithm.

4 Classifier selection

A classifier is an algorithm which uses a feature set as input to train a model. A classifier generates a model after it has been successfully trained by the training dataset. This model is then used to classify the test data into their respective classes. Binary or multi-class classifier may be selected depending on the problem. Classifier parameters are chosen or obtained to maximize the efficiency of the classifier. Various classifier options available are (Kotsiantis et al. 2007) as decision tree; support vector machine (SVM); naive Bayes; Bayesian network (BN); artificial neural network (ANN); K-nearest neighbors (KNN); logistic regression; random forests, etc. The classifier is based on the requirement of the feature set to be classified. High bias and low variance classifiers, e.g., naive Bayes (Cestnik et al. (1987) are more suitable for smaller training set. But, low bias/high variance classifiers, e.g., KNN (Cover and Hart 1967) are beneficial when the training set is large. High bias classifiers are not adequate to provide exact models. A naive Bayes classifier will converge quickly as compared to the logistic regression model and thus require less data to train. It is fast and easy, but it can not learn interactions between the features. Logistic regression (Peng et al. 2002) is particularly useful when the training data will be expected to increase in the future, and it is to be quickly incorporated into the model. Decision trees (Swain and Hauska 1977) are easy to interpret and explain. They can handle feature interactions and are nonparametric. The most interesting feature of BNs (Jensen (1996), compared to decision trees or neural networks (Foody et al. 1995), is that it takes into account the prior information about a given problem, in terms of structural relationships between its features. SVMs (Vapnik 1995) offer high accuracy, avoid over fitting and can work well even if data is not linearly separable in the base feature space. But the choice of the kernel must be appropriate. Random forests (Breiman 2001) are most popularly used for problems in classification as they are fast, scalable and no tuning of parameters is required as in SVMs. But the conclusion is that better data often beat better algorithms. Moreover, if the dataset is very huge, then speed or ease of use must be the deciding parameters (Chen 2015). Some important parameters while choosing a classifier should be accuracy, efficiency, robustness, simplicity and size of the model. A classifier works with the features extracted from the image data. Initially, a large number of features are extracted, but a number of features could be reduced. This is called feature selection. Thus, feature preprocessing/reduction is done to reduce the computational cost of classification (Pereira et al. 2009). In present paper, the authors have considered, k-NN, decision tree, and random forest classifier for forensics documents examination. In the k-nearest neighbor (k-NN) classifier, difference between the candidate vector and stored vector are computed using Euclidean distance. It is computed as:

$$d = \sqrt{\sum\nolimits_{k=1}^{N} (x_k - y_k)^2}$$

Here, N denotes the number of features; x_k denotes the value of the stored feature, and y_k denotes the value of candidate feature. A decision tree learning algorithm uses the data characteristics for computing and decision making. Each node represents the data attributes, and the leaf node represents a classification. Usually these classifiers are used to classify various sub-samples within the dataset. A random forest is another classifier which eliminates the over-fitting problem of decision tree. The meta-estimator that fits the number of decision tree classifiers for an assembly design is called random forest. The random forest improves the recognition accuracy using mean values and control over-fitting. In this paper, the recognition accuracy has been further improved using a combination of classifiers and majority voting scheme.

5 Adaptive boosting and aggregate bootstrapping

5.1 Adaptive boosting (AdaBoost)

A classical problem faced during classification is the selection of appropriate classifier. Selection of classifier is a very critical task. Yoav Freund and Robert Schapire proposed the AdaBoost algorithm to overcome this problem (Freund and Schapire 1996). It is a technique for getting an efficient classifier out of weak classifiers. In this approach, one classifier from the pool of classifiers is extracted after M iterations to make a committee. All elements are assigned the same weight initially. The elements in the data set are weighted for each iteration, according to their relevance. Larger weights are assigned progressively where the committee performance degraded. Further new classifiers are added to the committee by predicting their possible contribution to solving the tedious problems (Rojas 2009). In this paper, we have used this methodology for improving the classification results of printed documents.

5.2 Bootstrap aggregating (bagging)

Another technique introduced by Breiman considers bootstrap samples of objects and then trains the classifiers on each sample. Then majority voting is used to make the decision based on combined classifier votes. Experiments proved that the classifier obtained using this technique converts a weak classifier into an efficient one. Bootstrap aggregating is an assembly algorithm that first generates different samples of the training data set and creates a classifier for each sample. The results of these multiple classifiers are then combined (such as averaged or majority voting). The aim of bagging is to estimate cluster label for each observation for a given number of clusters K. In the bootstrap aggregating (Breiman 1996), different learning sets of the same size are formed using random technique for replacement. Predictors are built for each new dataset and combined by majority voting. The confidence of predictions for individual observations is made (Dudoit and Fridlyand 2003). Kumar et al. (2018) have used adaptive boosting and bootstrap aggregating techniques for improving the recognition accuracy of medieval handwritten Gurmukhi manuscript.

6 Proposed system for forensic document examination

This section elaborates the system design and flowchart for the proposed Model. Figure 2 shows the system design for the proposed work. Initially, the image features are extracted. The image dataset is divided in training and testing dataset. Then, the supervised training is performed to obtain a classifier model and then testing is performed to evaluate this system.

Algorithm

Step 1 Input digital image of printed text document

Step 2 Extract feature descriptor vector using KPNF, ORB and SURF features for each image in dataset

Step 3 Use Average, Median, Gaussian and Weiner filter as de-noising filters. Mean and standard deviation from the de-noising filtered images are extracted to obtain 8D KPNF features

Step 4 Extract ORB and SURF features from the images

Step 5 Use K-means clustering algorithm to generate K numbers of clusters for every descriptor vector. Compute the mean of every cluster

Step 6 Use LPP dimensionality reduction algorithm to reduce the feature vector dimensions, 48-dimensional feature vectors is reduced to 8 for both ORB and SURF

Step 7 KPNF, SURF and ORB feature vectors are stored in a database for training and testing purpose

Step 8 Train the proposed system using features extracted in the previous step and apply *k*-NN, decision tree, random forest and majority voting for the classification task

Step 9 Predict the class of questioned documents, by submitting their KPNF, ORB and SURF features to the trained classifier

Step 10 Return the class of printer as output for the questioned document

7 Experimental results

The experimental results of the proposed model are obtained using an existing document dataset (Gebhardt et al. 2013) as shown in Fig. 2. This dataset contains printed documents from 20 inkjet and laser printers as listed in Table 1. Fifty documents per printer are taken into consideration. Document of three categories, i.e., contract, invoice and scientific papers, is included in the dataset. All documents printed by a printer are unique. A subset of 07

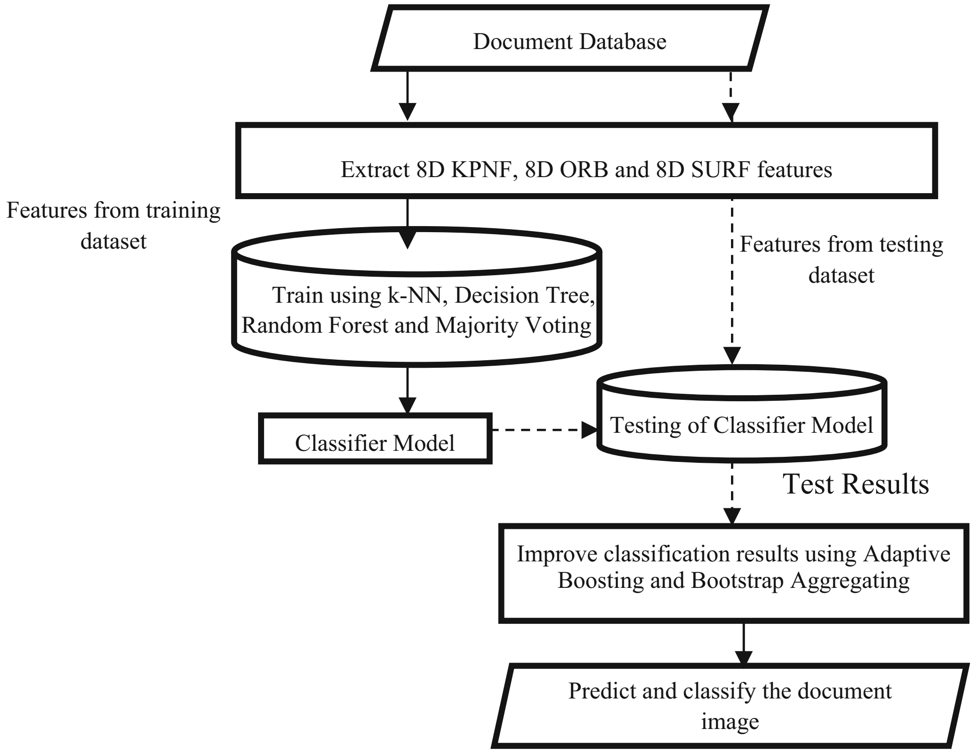

Document Database
Extract 8D KPNF, 8D ORB and 8D SURF features
Features from training dataset
Train using k-NN, Decision Tree, Random Forest and Majority Voting
Features from testing dataset
Classifier Model
Testing of Classifier Model
Test Results
Improve classification results using Adaptive Boosting and Bootstrap Aggregating
Predict and classify the document image

Fig. 2 Block diagram of proposed document forensics examination system

Table 1 Printers used for experimental work

Category	Inkjet/laser jet	Make
a	Inkjet	Officejet 5610
b	Inkjet	Epson Stylus Dx 7400
c	Inkjet	Unknown_1
d	Inkjet	Canon MX850
e	Inkjet	Canon MP630
f	Inkjet	Canon MP64D
g	Inkjet	Unknown_2
h	Laser	Samsung CLP 500
i	Laser	Ricoh Aficio MPC2550
j	Laser	HP LaserJet 4050
k	Laser	OKI C5600
l	Laser	HP LaserJet 2200dtn
m	Laser	Ricoh Afico Mp6001
n	Laser	HP Color LaserJet 4650dn
o	Laser	Nashuatec DSC 38 Aficio
p	Laser	Canon LBP7750 cdb
q	Laser	Canon iR C2620
r	Laser	HP LaserJet 4350
s	Laser	HP LaserJet 5
t	Laser	Epson Aculaser C1100

inkjet printers and 13 laser printers are considered for performance evaluation of the proposed system. For this purpose, the features are extracted from each document image. For experimental results, the entire dataset is partitioned into training dataset and testing dataset. In used partitioning strategy, 80% data are taken as training dataset and remaining data are taken as testing dataset. Fivefold cross-validation technique is also used to assess the effectiveness of the proposed system. Four classifiers, namely *k*-NN: C1, decision tree: C2, random forest: C3 and majority voting: C4, are considered in this work in order to classify the data. All experimental results are computed using i7 processing with 8 GB RAM. For classification, an open source WEKA tool is considered in the present work.

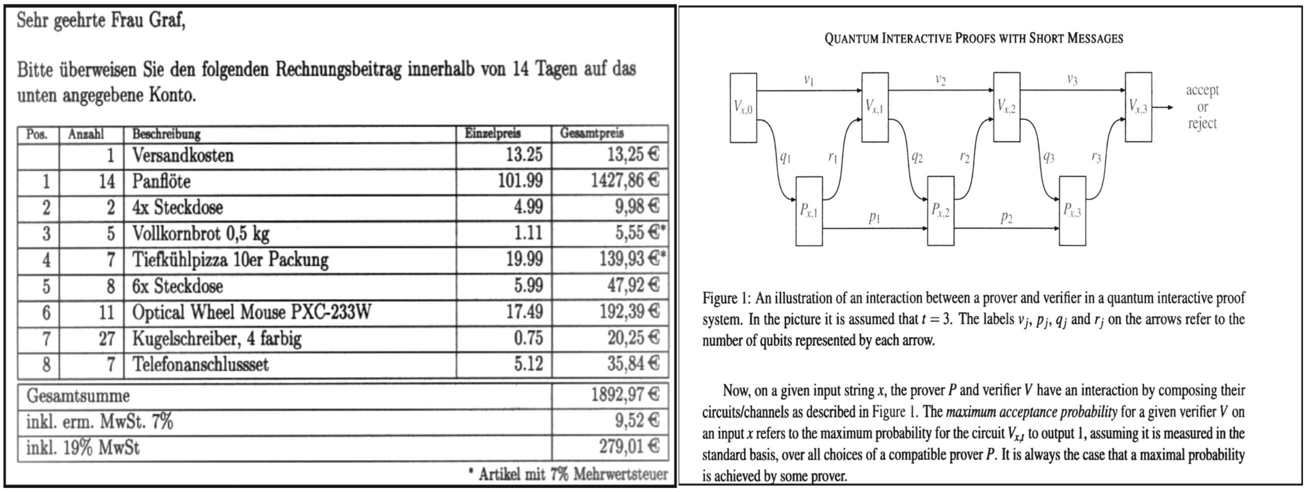
A few samples of documents printed using inkjet and laser printers are depicted in Figs. 3 and 4, respectively.

7.1 Recognition accuracy using various features and classifiers

The experiments are conducted using KPNF-, ORB- and SURF-based features independently and then using their integration. Classifier-wise recognition accuracy is depicted in Table 2. As depicted in this Table 2, maximum recognition precision of 91.5 has been achieved using random forest classifier with a combination of KPNF + ORB + SURF features. Experimental results based on precision rate, RMSE and ROC are graphically presented in Figs. 5, 6 and 7, respectively.

7.2 Recognition accuracy using bootstrap aggregating

In this sub-section, recognition results based on bootstrap aggregating are presented. Classifier-wise recognition results using bootstrap aggregating are depicted in Table 3. As shown in Table 3, maximum precision of 88.0% has been achieved using a combination of KPNF + SURF using majority voting as a classifier. Experimental results based on precision rate, RMSE and ROC using bootstrap aggregating are graphically presented in Figs. 8, 9 and 10, respectively.



Sehr geehrte Frau Graf,

Bitte überweisen Sie den folgenden Rechnungsbeitrag innerhalb von 14 Tagen auf das unten angegebene Konto.

Pos.	Anzahl	Beschreibung	Einzelpreis	Gesamtpreis
	1	Versandkosten	13.25	13,25 €
1	14	Panflöte	101.99	1427,86 €
2	2	4x Steckdose	4.99	9,98 €
3	5	Vollkornbrot 0,5 kg	1.11	5,55 €*
4	7	Tiefkühlpizza 10er Packung	19.99	139,93 €*
5	8	6x Steckdose	5.99	47,92 €
6	11	Optical Wheel Mouse PXC-233W	17.49	192,39 €
7	27	Kugelschreiber, 4 farbig	0.75	20,25 €
8	7	Telefonanschlussset	5.12	35,84 €

Gesamtsumme	1892,97 €
inkl. erm. MwSt. 7%	9,52 €
inkl. 19% MwSt	279,01 €

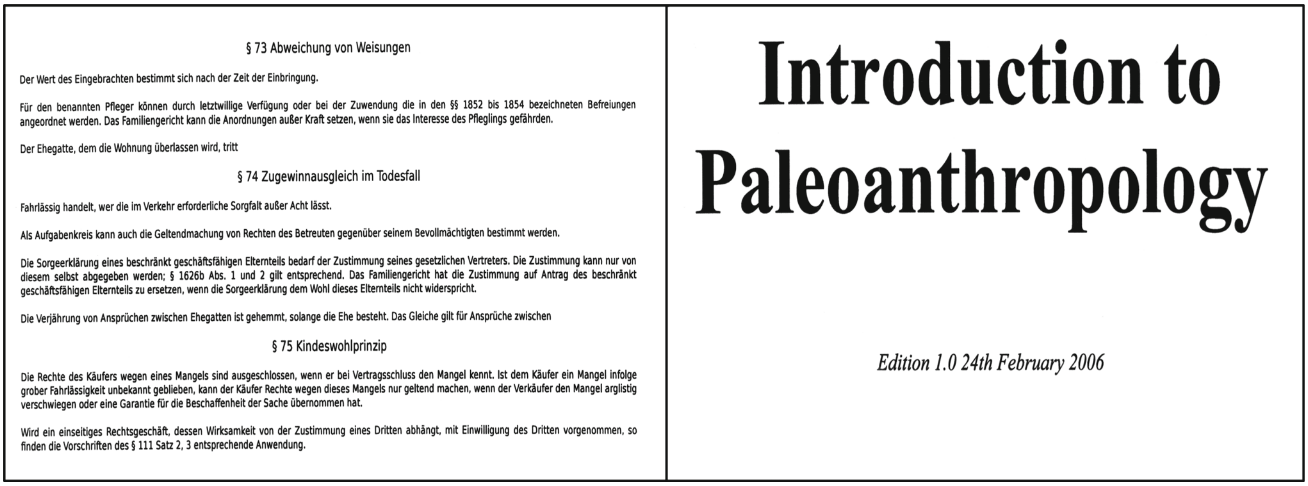
* Artikel mit 7% Mehrwertsteuer

QUANTUM INTERACTIVE PROOFS WITH SHORT MESSAGES

Figure 1: An illustration of an interaction between a prover and verifier in a quantum interactive proof system. In the picture it is assumed that $t = 3$. The labels v_j, p_j, q_j and r_j on the arrows refer to the number of qubits represented by each arrow.

Now, on a given input string x, the prover P and verifier V have an interaction by composing their circuits/channels as described in Figure 1. The *maximum acceptance probability* for a given verifier V on an input x refers to the maximum probability for the circuit $V_{x,t}$ to output 1, assuming it is measured in the standard basis, over all choices of a compatible prover P. It is always the case that a maximal probability is achieved by some prover.

Fig. 3 Samples of documents printed with inkjet printer



§ 73 Abweichung von Weisungen

Der Wert des Eingebrachten bestimmt sich nach der Zeit der Einbringung.

Für den benannten Pfleger können durch letztwillige Verfügung oder bei der Zuwendung die in den §§ 1852 bis 1854 bezeichneten Befreiungen angeordnet werden. Das Familiengericht kann die Anordnungen außer Kraft setzen, wenn sie das Interesse des Pfleglings gefährden.

Der Ehegatte, dem die Wohnung überlassen wird, tritt

§ 74 Zugewinnausgleich im Todesfall

Fahrlässig handelt, wer die im Verkehr erforderliche Sorgfalt außer Acht lässt.

Als Aufgabenkreis kann auch die Geltendmachung von Rechten des Betreuten gegenüber seinem Bevollmächtigten bestimmt werden.

Die Sorgeerklärung eines beschränkt geschäftsfähigen Elternteils bedarf der Zustimmung seines gesetzlichen Vertreters. Die Zustimmung kann nur von diesem selbst abgegeben werden; § 1626b Abs. 1 und 2 gilt entsprechend. Das Familiengericht hat die Zustimmung auf Antrag des beschränkt geschäftsfähigen Elternteils zu ersetzen, wenn die Sorgeerklärung dem Wohl dieses Elternteils nicht widerspricht.

Die Verjährung von Ansprüchen zwischen Ehegatten ist gehemmt, solange die Ehe besteht. Das Gleiche gilt für Ansprüche zwischen

§ 75 Kindeswohlprinzip

Die Rechte des Käufers wegen eines Mangels sind ausgeschlossen, wenn er bei Vertragsschluss den Mangel kennt. Ist dem Käufer ein Mangel infolge grober Fahrlässigkeit unbekannt geblieben, kann der Käufer Rechte wegen dieses Mangels nur geltend machen, wenn der Verkäufer den Mangel arglistig verschwiegen oder eine Garantie für die Beschaffenheit der Sache übernommen hat.

Wird ein einseitiges Rechtsgeschäft, dessen Wirksamkeit von der Zustimmung eines Dritten abhängt, mit Einwilligung des Dritten vorgenommen, so finden die Vorschriften des § 111 Satz 2, 3 entsprechende Anwendung.

Introduction to Paleoanthropology

Edition 1.0 24th February 2006

Fig. 4 Samples of documents printed with laser printer

Table 2 Experimental results using various features and classifiers

Number of Features	Precision rate (%age)				RMSE (%age)				ROC (%age)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
KPNF (8)	60.0	54.5	61.0	61.5	19.7	19.7	16.5	19.6	79.1	80.9	95.2	79.6
ORB (8)	60.5	48.5	66.5	65.0	19.7	20.8	15.4	18.7	75.9	78.2	91.1	81.6
SURF (8)	64.5	52.5	76.5	74.5	18.8	17.6	13.6	15.9	80.1	88.9	98.2	86.5
KPNF (8) + ORB (8)	51.0	62.5	72.5	67.0	21.8	17.9	14.7	18.2	74.2	83.5	96.9	82.6
KPNF (8) + SURF (8)	67.5	79.5	88.5	88.0	17.8	13.4	10.9	10.9	82.7	93.0	99.3	93.7
ORB (8) + SURF (8)	57.0	72.5	80.5	76.5	20.5	15.9	13.2	15.3	77.2	89.0	98.7	87.5
KPNF(8) + ORB (8) + SURF (8)	59.5	80.5	91.5	85.5	19.9	13.0	11.7	12.0	78.5	91.8	99.5	92.3

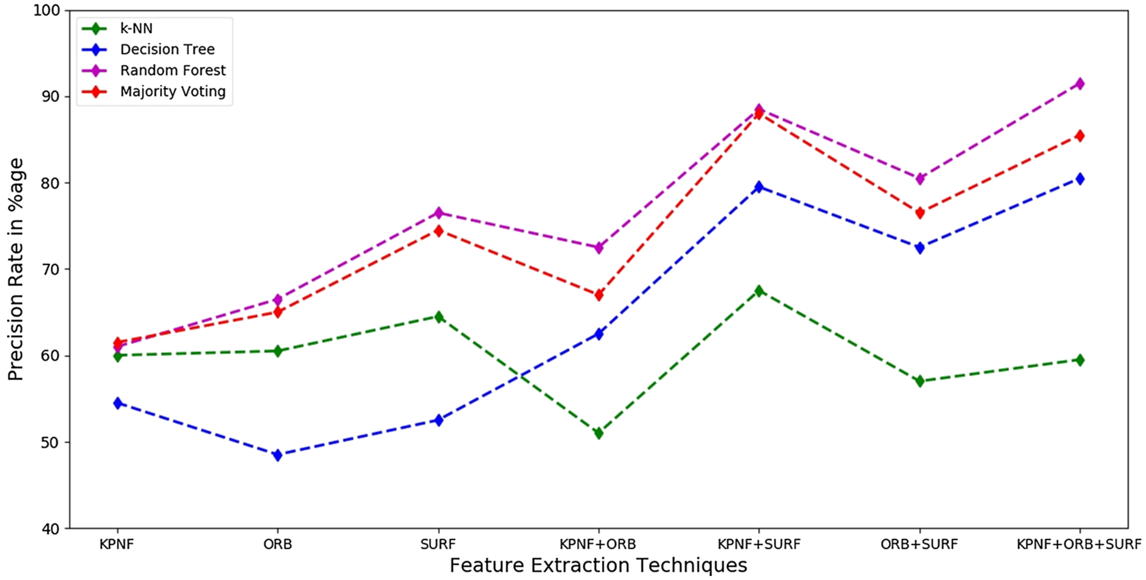


Fig. 5 Precision rate using various features and classifiers

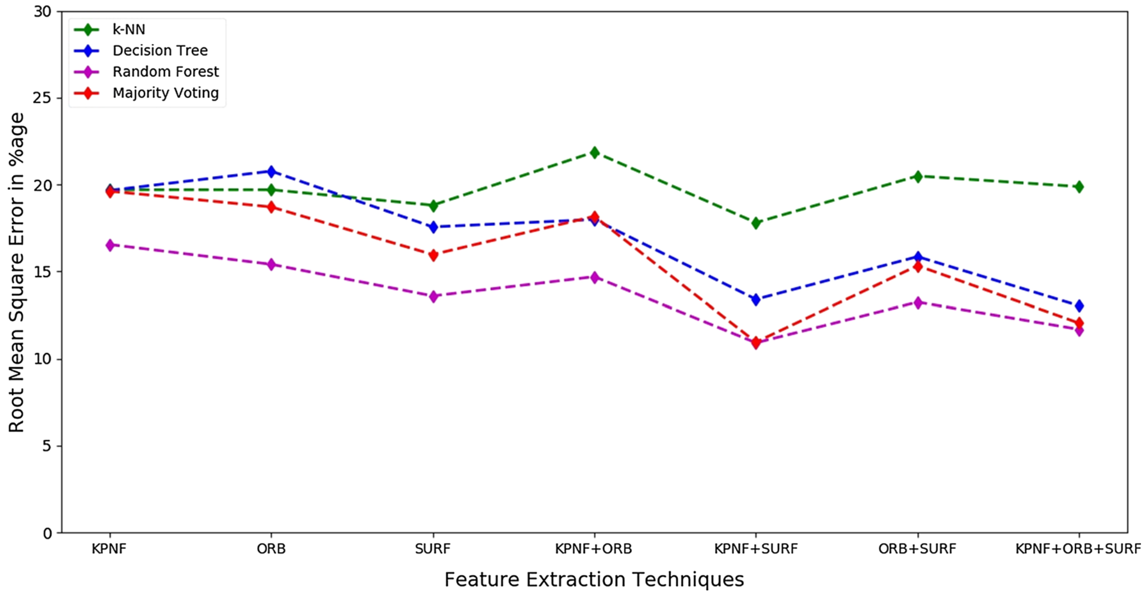


Fig. 6 RMSE using various features and classifiers

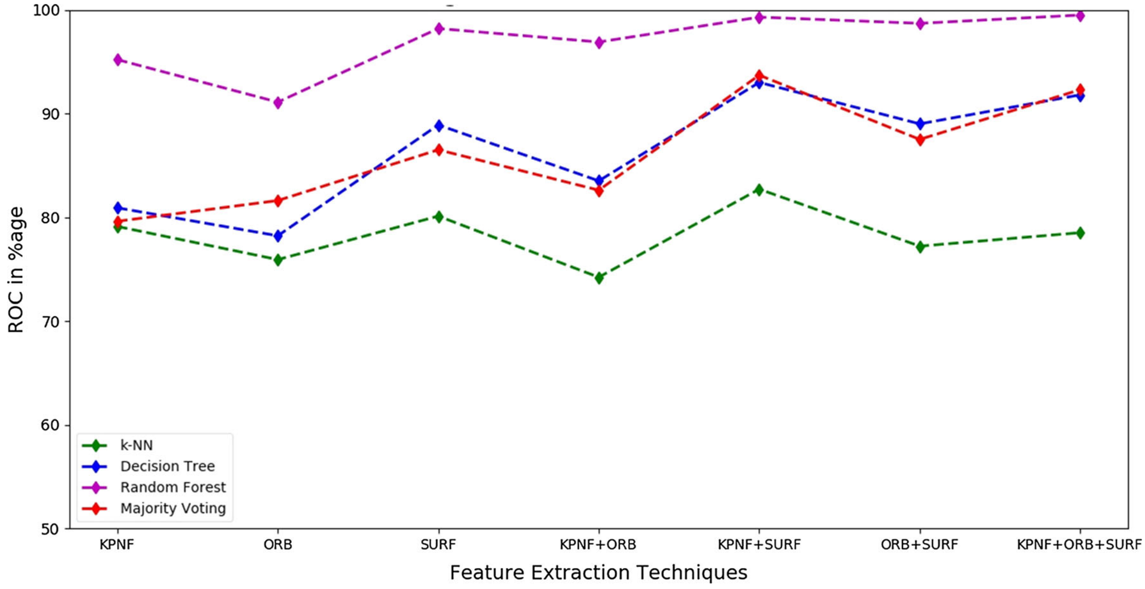


Fig. 7 ROC using various features and classifiers

7.3 Recognition accuracy using adaptive boosting

Next set of experiments has been carried out with adaptive boosting methodology as presented in Table 4. It has been observed that the random forest classifier-based integration of KPNF + SURF + ORB when combined with adaptive boosting methodology achieved 94.0% precision rate. Experimental results based on precision rate, RMSE and ROC using adaptive boosting are graphically presented in Figs. 11, 12 and 13, respectively.

Maximum recognition accuracy of 95.1% has been achieved using a combination of KPNF + SURF + ORB features and adaptive boosting methodology with random forest classifier. The confusion matrix for this case (a combination of KPNF + SURF + ORB features and adaptive boosting methodology with random forest classifier) is depicted in Table 5.

Table 3 Experimental results using bootstrap aggregating

Number of features	Precision rate (%age)				RMSE (%age)				ROC (%age)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
KPNF (8)	60.0	60.0	62.0	62.5	17.4	17.0	16.7	16.4	89.4	93.6	95.6	90.7
ORB (8)	58.0	57.5	63.5	65.0	18.9	17.2	16.2	16.8	79.1	88.0	90.8	86.2
SURF (8)	61.0	59.0	72.0	74.0	17.6	16.1	14.2	14.8	88.3	92.9	97.8	93.6
KPNF (8) + ORB (8)	52.5	69.0	64.5	69.5	18.8	15.1	15.8	14.9	83.3	93.9	93.7	91.2
KPNF (8) + SURF (8)	67.0	82.0	83.5	88.0	15.7	11.6	11.6	9.4	90.2	97.4	98.1	98.3
ORB (8) + SURF (8)	55.0	73.0	76.5	81.0	18.9	13.9	13.8	12.0	82.3	95.7	97.6	96.4
KPNF(8) + ORB (8) + SURF (8)	58.0	84.0	83.5	87.5	17.8	10.5	12.8	8.8	88.3	97.6	97.3	98.7

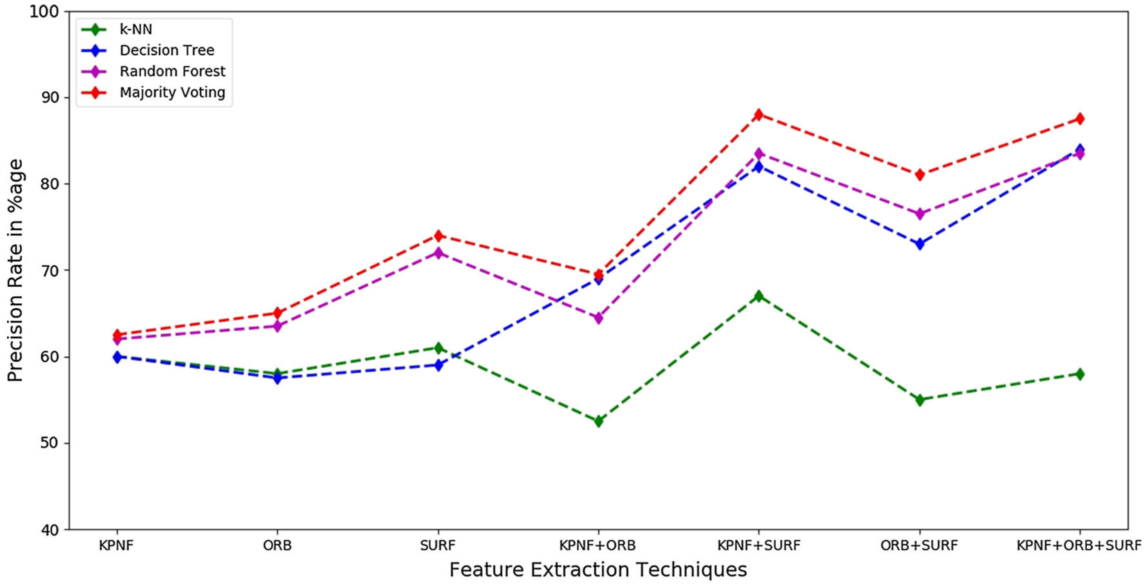


Fig. 8 Precision rate using bootstrap aggregating

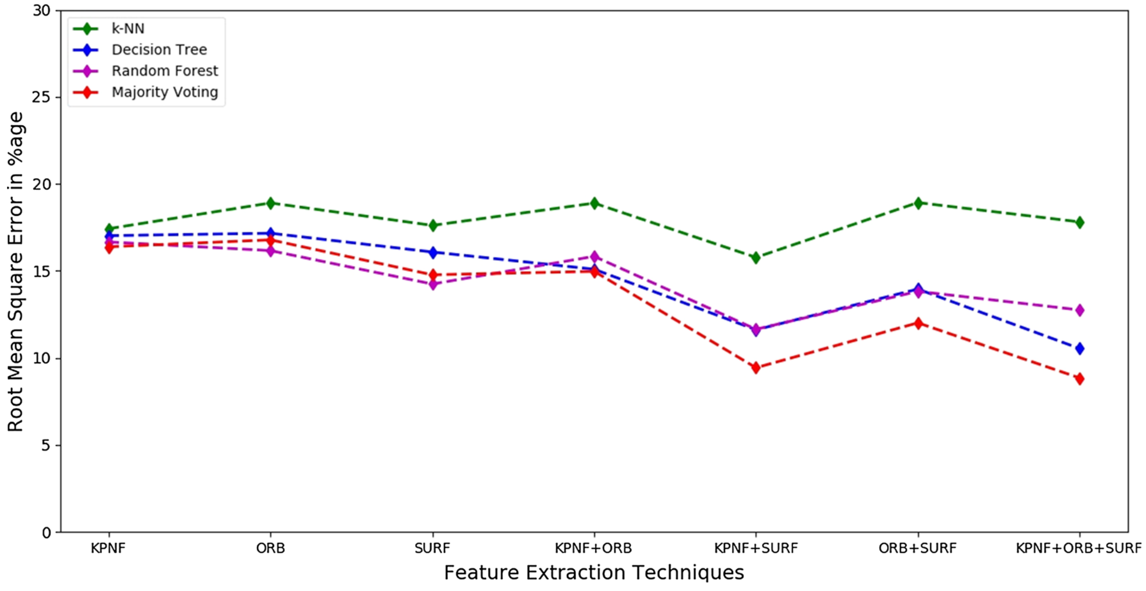


Fig. 9 RMSE using bootstrap aggregating

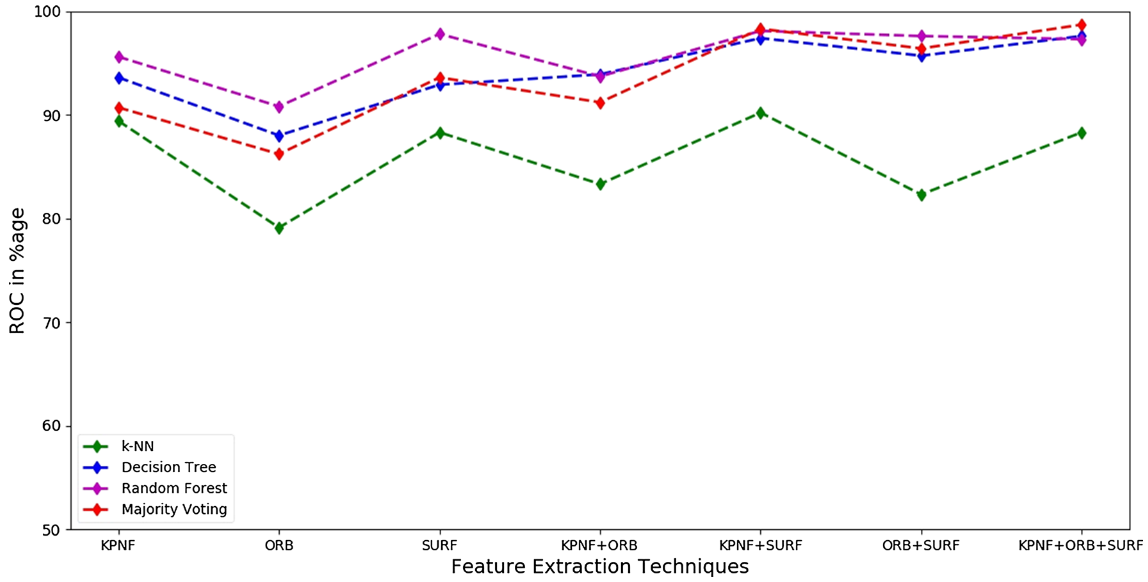


Fig. 10 ROC using bootstrap aggregating

Table 4 Experimental results using adaptive boosting

Number of features	Precision rate (%age)				RMSE (%age)				ROC (%age)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
KPNF (8)	59.5	61.0	62.0	62.5	19.9	17.54	18.8	17.9	77.3	91.3	86.6	91.8
ORB (8)	60.0	63.0	64.5	65.7	19.9	17.62	18.3	17.9	79.4	91.0	87.6	91.4
SURF (8)	62.0	52.5	75.5	70.5	19.0	20.12	14.9	16.6	79.8	89.8	93.1	91.6
KPNF (8) + ORB (8)	51.0	71.5	74.0	73.5	21.9	15.37	14.5	15.8	74.2	94.5	96.9	93.7
KPNF (8) + SURF (8)	67.5	85.0	91.0	91.0	17.8	11.94	10.8	9.3	82.7	98.0	99.3	99.0
ORB (8) + SURF (8)	57.0	77.5	80.0	83.0	20.5	14.61	13.1	12.6	77.2	94.9	98.7	97.4
KPNF(8) + ORB (8) + SURF (8)	59.5	85.0	94.0	91.5	19.9	11.63	11.5	8.6	78.5	98.8	99.7	99.5

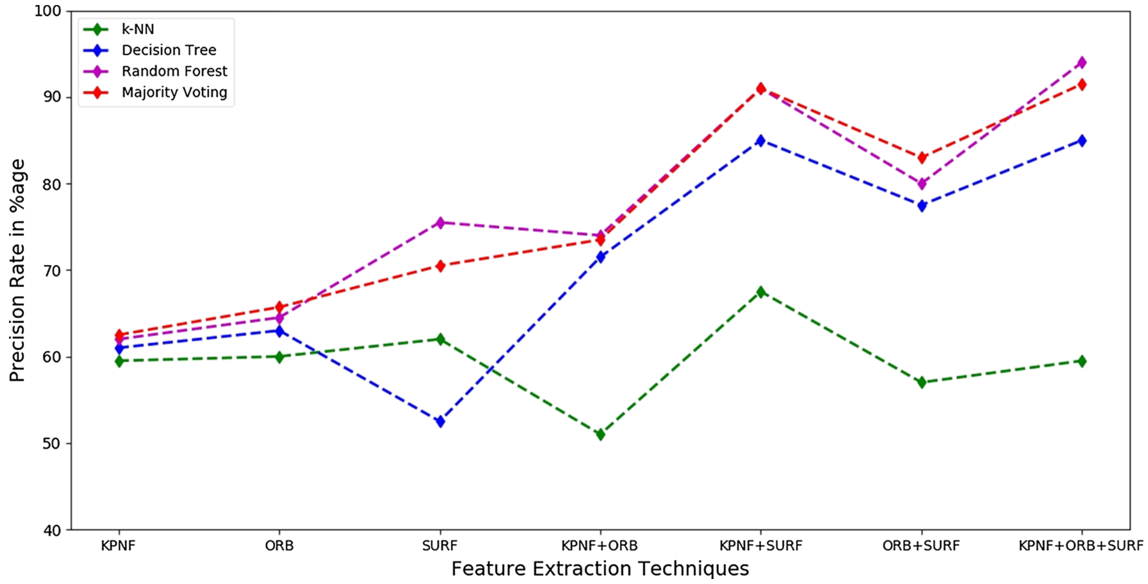


Fig. 11 Precision rate using adaptive boosting

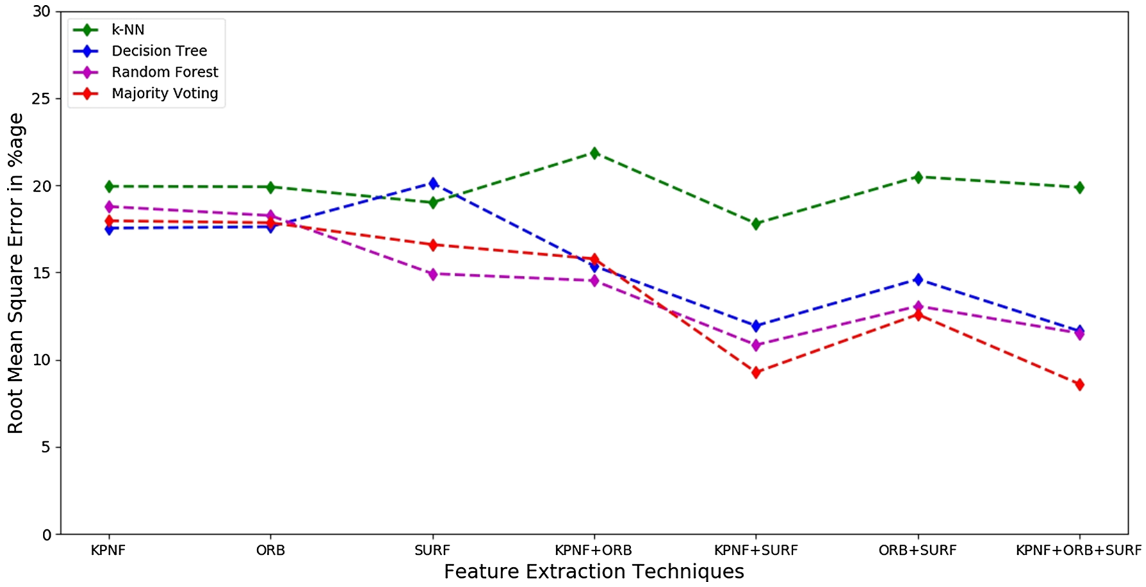


Fig. 12 RMSE using adaptive boosting

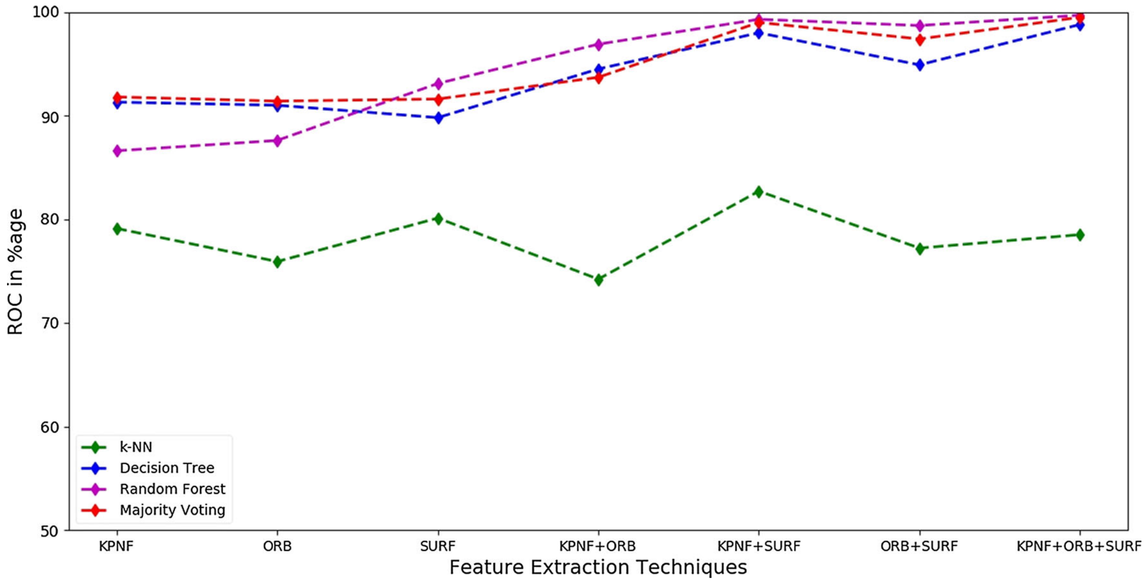


Fig. 13 ROC using adaptive boosting

8 Comparison with the state-of-the-art work

The comparison of the present work is done with the state-of-the-art methods based on GLCM (Mikkilineni et al. 2005a), DWT (Choi et al. 2009), GLCM-DWT (Tsai and Liu 2013), Cross Center-symmetric LTP (CCSLTP) (Fu and Yang 2012), multi-directional GLCM (GLCM MD) (Ferreira et al. 2015) and multi-directional multi-scale GLCM (GLCM MD MS) (Ferreira et al. 2015). The classification accuracy of these algorithms is listed in Table 6 and Fig. 14, respectively.

9 Inferences

In this paper, a passive model for source printer identification is proposed. It is based on key printer noise features (KPNF), speeded up robust features (SURF) and oriented fast rotated and BRIEF (ORB). Size of SURF and ORB descriptor require a high memory space for storing features. Therefore, a K-means clustering and LPP are used. K-means reduce the descriptor into 64 clusters and LPP reduce into 8 components each for SURF and ORB features. The proposed model based on KPNF + ORB + SURF can efficiently classify the questioned documents to

Table 5 Confusion matrix for KPNF + SURF + ORB features and adaptive boosting methodology with random forest classifier

Classified as		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
Canon_iR_C2620	a	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Canon_LBP7750_cdb	b	0	7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Canon_MP630	c	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Canon_MP64D	d	0	0	0	9	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
Canon_MX850	e	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Epson_Aculaser_C1100	f	0	0	0	0	0	8	0	1	0	0	1	0	1	0	0	0	0	0	0	0
Epson_Stylus_Dx_7400	g	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0
HP_Color_LaserJet_4650dn	h	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
HP_LaserJet_2200dtn	i	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
HP_LaserJet_4050	j	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0
HP_LaserJet_5	k	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0
Hp_LaserJet4350 o.4250	l	1	0	0	0	0	0	0	0	0	0	1	9	0	0	0	0	0	0	0	0
Nashuatec_DSC_38_Aficio	m	0	1	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
Officejet_5610	n	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0
OKI_C5600	o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
Ricoh_Aficio_MPC2550	p	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	11	0	0	0	0
Ricoh_Afico_Mp6001	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0
Samsung_CLP_500	r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0
Unknown_1	s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0
Unknown_2	t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8

Table 6 Comparison with state-of-the-art work

Feature extraction technique	Feature size (1-D)	Accuracy (%)
GLCM	12	86.2
DWT	22	90.6
GLCM-DWT	34	93.4
CCSLTP	128	57.8
GLCM MD	176	93.0
KPNF (8)	8	62.5
ORB (8)	8	65.9
SURF (8)	8	75.5
KPNF (8) + ORB (8)	16	74.4
KPNF (8) + SURF (8)	16	91.3
ORB (8) + SURF (8)	16	83.2
KPNF(8) + ORB (8) + SURF (8)	24	95.1

their respective printer class as compared to state of art. Experimental results have affirmed the viability of the proposed approach and proved the characteristic advantages

Four classifiers, namely *k*-NN: C1, decision tree: C2, random forest: C3 and majority voting: C4, are experimented for classification task. Authors improved the accuracy of 1.7% with the proposed system using adaptive boosting and bootstrap aggregating methodologies. Finally, a precision rate of 95.1% has been achieved using a combination of KPNF + SURF + ORB features and adaptive boosting methodology with random forest classifier. Integration of features has enhanced the accuracy and precision of the proposed system with added advantage of low dimensionality.

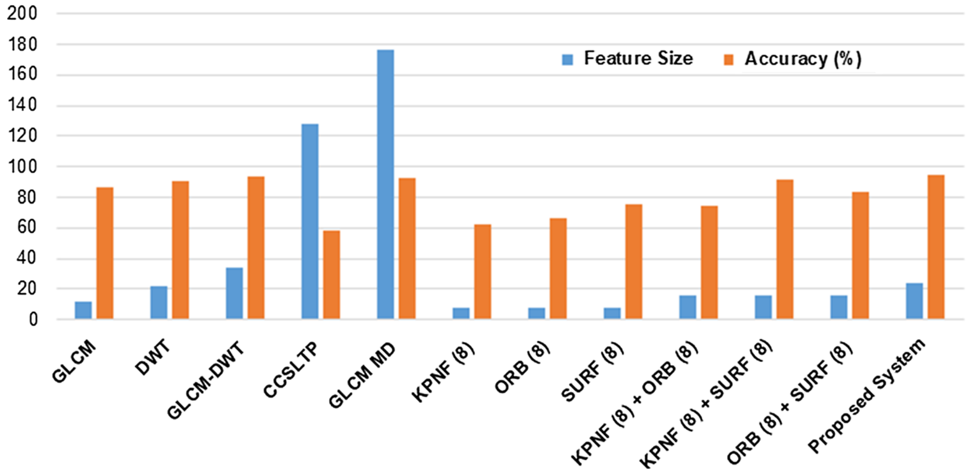


Fig. 14 Comparison with state-of-the-art work

Compliance with ethical standards

Conflict of interest Authors have no conflicts of interest in this work.

References

Ali GN, Mikkilineni AK, Delp EJ, Allebach JP, Chiang PJ, Chiu GT (2004) Application of principal components analysis and gaussian mixture models to printer identification. In: Proceedings of non-impact printing and digital fabrication conference, Salt Lake City, Utah, vol 1, pp 301–305

Amer M, Goldstein M (2012) Nearest-neighbor and clustering based anomaly detection algorithms for Rapidminer. In: Proceedings of 3rd Rapidminer community meeting and conference, Aachen, Germany, pp 1–12

Bayram S, Sencar H, Memon N, Avcibas I (2005) Source camera identification based on CFA interpolation. In: Proceedings of international conference on image processing, Genova, Italy, vol 3, pp 69–78

Bayram S, Sencar HT, Memon N (2008) Classification of digital camera-models based on demosaicing artifacts. Digit Investig 5(1):49–59

Bertrand R, Gomez-Kramer P, Terrades OR, Franco P, Ogier JM (2013) A system based on intrinsic features for fraudulent document detection. In: Proceedings of 12th international conference on document analysis and recognition, Washington, DC, pp 6–110

Bianchi T, Piva A (2013) Secure watermarking for multimedia content protection: a review of its benefits and open issues. IEEE Signal Process Mag 30(2):87–96

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Bulan O, Mao J, Sharma G (2009) Geometric distortion signatures for printer identification. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, Taipei, Taiwan, pp 1401–1404

Cedillo-Hernandez M, Garcia-Ugalde F, Nakano-Miyatake M, Perez-Meana H (2013) Robust object-based watermarking using SURF feature matching and DFT domain. Radio Eng 22(4):1057–1071

Cestnik B, Kononenko I, Bratko I (1987) Assistant 86: a knowledge elicitation tool for sophisticated users. In: Proceedings of 2nd European working session on learning, Bled, Yugoslavia, pp 31–45

Chen E (2015) Choosing a machine learning classifier. http://blog.echen.me/2011/04/27/choosing-a-machine-learningclassifier/. Accessed 13 March 2016

Choi JH, Im DH, Lee HY, Oh JT, Ryu JH, Lee HK, (2009) Color laser printer identification by analyzing statistical features on discrete wavelet transform. In: Proceedings of 16th IEEE international conference on image processing, Cairo, Egypt, pp 1505–1508

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27

Cox IJ, Miller ML, Bloom JA (2000) Watermarking applications and their properties. In: Proceedings of international conference on information technology: coding and computing, Las Vegas, Nevada, pp 6–10

Dudoit S, Fridlyand J (2003) Bagging to improve the accuracy of a clustering procedure. Bioinformatics 19(9):1090–1099

Elkasrawi S, Shafait F (2014) Printer identification using supervised learning for document forgery detection. In: Proceedings of 11th IAPR international workshop on document analysis systems, France, pp 146–150

Ferreira A, Navarro LC, Pinheiro G, dos Santos JA, Rocha A (2015) Laser printer attribution: exploring new features and beyond. Forensic Sci Int 247:105–125

Foody GM, McCulloch MB, Yates WB (1995) The effect of training set size and composition on artificial neural network classification. Int J Remote Sens 16(9):1707–1723

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proceedings of international conference on machine learning, vol 96, pp 148–156

Fu YR, Yang SY (2012) CCS-LTP for Printer Identification based on Texture Analysis. Int J Digit Content Technol Appl 6(13):250–264

Gebhardt J, Goldstein M, Shafait F, Dengel A (2013) Document authentication using printing technique features and unsupervised anomaly detection. In: Proceedings of 12th international conference on document analysis and recognition, Washington, DC, pp 479–483

Jensen FV (1996) An introduction to bayesian networks, vol 210. UCL Press, London, pp 22–25

Jiang W, Ho AT, Treharne H, Shi YQ (2010) A novel multi-size block Benford's law scheme for printer identification. In: Proceedings of Pacific-Rim conference on multimedia, Shanghai, China, pp 643–652

Joshi S, Khanna N (2017) Single classifier-based passive system for source printer classification using local texture features. IEEE Trans Inf Forensics Secur 13(7):1603–1614

Kee E, Farid H (2008) Printer profiling for forensics and ballistics. In: Proceedings of 10th ACM workshop on multimedia and security, Oxford, pp 3–10

Khanna N, Mikkilineni AK, Chiu GTC, Allebach JP, Delp EJ (2007) Scanner identification using sensor pattern noise. In: Proceedings of security, steganography, and watermarking of multimedia contents, electronic imaging, San Jose, CA

Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: a review of classification techniques. In: Proceedings of conference on emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies, pp 3–24

Kumar M, Jindal SR, Jindal MK, Lehal GS (2018) Improved recognition results of medieval handwritten Gurmukhi manuscripts using boosting and bagging methodologies. Neural Process Lett 1:1–14. https://doi.org/10.1007/s11063-018-9913-6

Lampert CH, Mei L, Breuel TM (2006) Printing technique classification for document counterfeit detection. In: Proceedings of international conference on computational intelligence and security, Guangzhou, China, vol 1, pp 639–644

Li Z, Jiang W, Kenzhebalin D, Gokan A, Allebach J (2018) Intrinsic signatures for forensic identification of SOHO inkjet printers. NIP Digit Fabr Confer 1:231–236

Mikkilineni AK, Chiang PJ, Ali GN, Chiu GTC, Allebach JP, Delp EJ (2004) Printer identification based on texture features. In: Proceedings of non-impact printing and digital fabrication conference, society for imaging science and technology, Salt Lake City, Utah, vol 1, pp 306–311

Mikkilineni AK, Chiang PJ, Ali GN, Chiu GTC, Allebach JP, Delp EJ (2005a) Printer identification based on graylevel co-occurrence features for security and forensic applications. In: Proceedings of security, steganography, and watermarking of multimedia contents, electronic imaging, California, pp 430–440

Mikkilineni AK, Khanna N, Delp EJ (2011) Forensic printer detection using intrinsic signatures. Media Forensics Secur 7880:78800–78805

Peng CYJ, Lee KL, Ingersoll GM (2002) An introduction to logistic regression analysis and reporting. J Educ Res 96(1):3–14

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and MRI: a tutorial overview. Neuroimage 45(1):S199–S209

Phillips IT (1996) User's reference manual for the UW English/technical document image database III. UW-III English/technical document image database manual

Rojas R (2009) AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting. Freie University, Berlin, Technical Report

Ryu SJ, Lee HY, Cho IW, Lee HK (2008) Document forgery detection with SVM classifier and image quality measures. Adv Multimed Inf Process 2008:486–495

Schreyer M, Schulze C, Stahl A, Effelsberg W (2009) Intelligent printing technique recognition and photocopy detection for forensic document examination. Informatiktage 8:39–42

Schulze C, Schreyer M, Stahl A, Breuel T (2008) Evaluation of graylevel-features for printing technique classification in high-throughput document management systems. Comput Forensics 28:35–46

Smith R (2007) An overview of the Tesseract OCR engine. In: Proceedings of 9th international conference on document analysis and recognition, Beijing, China, vol 2, pp 629–633

Subramanya SR, Yi BK (2006) Digital Signatures. IEEE Potentials 25(2):5–8

Swain PH, Hauska H (1977) The decision tree classifier: design and potential. IEEE Trans Geosci Electron 15(3):142–147

Tao H, Zain JM, Ahmed MM, Abdalla AN, Jing W (2012) A wavelet-based particle swarm optimization algorithm for digital image watermarking. Integr Comput Aided Eng 19(1):81–91

Tao H, Chongmin L, Zain JM, Abdalla AN (2014) Robust image watermarking theories and techniques: a review. J Appl Res Technol 12(1):122–138

Tayan O, Kabir MN, Alginahi YM (2014) A hybrid digital-signature and zero-watermarking approach for authentication and protection of sensitive electronic documents. Sci World J 8:1–15

Tsai MJ, Liu J (2013) Digital forensics for printed source identification. In: Proceedings of IEEE international symposium on circuits and systems, Melbourne, Australia, pp 2347–2350

Tsai MJ, Yuadi I (2018) Digital forensics of microscopic images for printed source identification. Multimed Tools Appl 77(7):8729–8758

Tsai MJ, Liu J, Wang CS, Chuang CH (2011) Source color laser printer identification using discrete wavelet transform and feature selection algorithms. In: Proceedings of IEEE international symposium on circuits and systems, Rio de Janeiro, Brazil, pp 2633–2636

Van BJ, Shafait F, Breuel TM (2009) Resolution independent skew and orientation detection for document images. In: Proceedings of SPIE-IS&T document recognition and retrieval, electronic imaging, San Jose, CA, pp 1–8

Van BJ, Shafait F, Breuel TM (2013a) Text-line examination for document forgery detection. Int J Doc Anal Recognit 16(2):189–207

Van BJ, Shafait F, Breuel TM (2013b) Automatic authentication of color laser print-outs using machine identification codes. Pattern Anal Appl 16(4):663–678

Vapnik V (1995) The nature of statistical learning theory. Springer, New York. Google Scholar. Accessed on 15 July 2015

Vinay A, Kumar CA, Shenoy GR, Murthy KB, Natarajan S (2015) ORB-PCA based feature extraction technique for face recognition. Procedia Comput Sci 58:614–621

Wu Y, Kong X, You XG, Guo Y (2009) Printer forensics based on page document's geometric distortion. In: Proceedings of 16th IEEE international conference on image processing, Cairo, Egypt, pp 2909–2912

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.