A computational approach for printed document forensics using SURF and ORB features

Munish Kumar, Surbhi Gupta & Neeraj Mohan

Soft Computing
A Fusion of Foundations,
Methodologies and Applications

ISSN 1432-7643

Soft Comput
DOI 10.1007/s00500-020-04733-x

ONLINE FIRST

Springer

Soft Computing
https://doi.org/10.1007/s00500-020-04733-x

METHODOLOGIES AND APPLICATION

A computational approach for printed document forensics using SURF and ORB features

Munish Kumar[1] · Surbhi Gupta[2] · Neeraj Mohan[3]

Abstract

Document forgery is quite common nowadays due to the availability of cost-effective scanners and printers. Important documents like certificates, passport, identification cards, etc., are protected using watermarks or signatures. These are made secured with a protective printing mechanism with extrinsic fingerprints. Therefore, it is easy to authenticate such documents. Other documents required a passive approach for their authentication. These approaches look for document inconsistencies for chances of modification. Some of these attempt to detect and fix the source of the printed document. This paper proposes a classifier-based model to identify the source printer and classify the questioned document in one of the printer classes. A novel approach of utilizing Speeded Up Robust Features and Oriented Fast Rotated and BRIEF feature descriptors is proposed for printer attribution. Naive Bayes, *k*-NN, random forest and different combinations of these classifiers have been experimented for classification. The proposed model can efficiently classify the questioned documents to their respective printer class. An accuracy of 86.5% has been achieved using a combination of Naive Bayes, *k*-NN, random forest classifiers with a simple majority voting scheme and adaptive boosting methodology.

Keywords Document forensics · Printer forensics · SURF · ORB · Voting scheme · AdaBoost

1 Introduction

It is a digital world where everything is going paperless. But, even nowadays many important documents are still on paper. Popular examples include certificates, receipts, official documents, etc. These documents are vulnerable as they lack the required security features. This limitation has invited manipulations in documents. These manipulations in documents are termed as document tampering and can be performed easily using economical devices like scanners and printers. Usually, the document to be manipulated is first scanned and then the scanned image of the original document is manipulated easily. Therefore, before relying on a document, one must check its authenticity. Generally, the document authentication is done using active techniques. The techniques such as a watermark or signature are widely used to protect the digital documents. These techniques embed some additional extrinsic fingerprints to the document so that any manipulation will disturb these fingerprints and hence can be traced easily. But it is not possible to use such technology for all the documents as its costly and time-consuming. Manipulators exploit this weakness and attempt the desired changes in the document. Such unprotected documents require authentication using passive techniques. Such techniques are based on document image intrinsic features. Intrinsic features are the fingerprints of hardware and/or software used for the production of the authentic/manipulated document. While examining printed documents for manipulations, the identification of source printer can be extremely helpful. Therefore, there is a requirement for techniques that can identify the source printer. It has many industrial applications. In developing countries, every piece of information

Communicated by V. Loia.

✉ Munish Kumar
munishcse@gmail.com

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

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

[3] Department of Computer Science and Engineering, I.K.G Punjab Technical University Mohali Campus, Mohali, Punjab, India

Published online: 31 January 2020

can not be digitized, displayed and made available on digital devices. Due to that a lot of information is still managed using hard copies. Further, all such documents can not be protected using special features called extrinsic fingerprints as it is costly. So passive technique will authenticate such documents when required without the presence of external fingerprints. There may be cases, where one has to identify odd documents among piles of printed documents. The original printer source is known in such cases. The need is to identify the documents which are printed using printers other than the original printer. The presence of the countless number of economical printers has made their identification challenging. So, accurate and robust printer attribution techniques are very significant. This paper proposes feature-based classification of source printer using scanned images of printed documents. Most of the approaches for printed document classification are based on the analysis of halftone, texture or printer noise. But no effort has been done till date to utilize the key-point-based features to analyze the document images. Feature extraction using SIFT and SURF is common for analyzing images for object classification and detection. But, its application to document images has not been explored. So, this paper aims to analyze the use of key-point-based feature extraction using SURF and ORB for the classification of source printer using printed documents. Thus, we may fix the print technology and the printer made for printed documents and conclude whether the suspicious document is genuine or manipulated. In this paper, first, we have discussed the need and application of the presented work. Related work is exhibited in Sect. 2. The mathematical model for feature extraction is presented in Sect. 3. Section 4 covers the experimentation, comparison and discussion. Section 5 has concluding remarks.

2 Related work

There are many approaches to document tamper detection. Most of these approaches identify the document source and checks whether the document has been printed by the authorized printer. Other approaches look for document inconsistencies for the probability of modification. Printer attribution for document examination is based on either local and global features. Local features examine and analyze the connected components (CCs) or characters of the document. These techniques will study and analyze the statistics of some particular, frequently occurring characters like ‘e’ or ‘a’ for clues of modification. While global features examine the whole document at once. These techniques will analyze statistical features like noise across the document to identify manipulations. Some of the major contributions in printer attribution based on local features are as follows. Initially, Ali et al. (2003) used signal projection from text letters and classified the source printer based on this signal. The tests used seven printers and the documents contained approximately 10 lines with 40–100 words. Mikkilineni et al. (2004) proposed a technique to print documents securely even on low-cost printers. Intrinsic and extrinsic features obtained by the printer modeling process were used. Mikkilineni et al. (2005) proposed texture feature-based descriptors to discover the document source for document forensics. The technique was based on gray-level co-occurrence matrices (GLCM) statistics. Scanned text documents at 2400 dpi were considered. All ‘e’ letters were used for experimentation. Twenty-two statistical features from GLCM were extracted per character.

Mikkilineni et al. (2011) used a clustering-based approach to classify documents from different printers. Forensic printer identification was performed to fix the source printer of the document. Tsai and Liu (2013) combined GLCM statistics with sub-bands of wavelet transform. A specific character of the Chinese language was used for the texture pattern extraction from the scanned document. The average source identification rate was 98.64%. Laser printer source identification was even better. Similarly, Bertrand et al. (2013) examined font similarity and deviations of characters in a questioned document to detect document forgery. The detection of copied and pasted region was done by character shape comparison. Recall and precision values obtained for document forgery detection were 0.77% and 0.82%, respectively. Gebhardt et al. (2013) examined 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. Joshi and Khanna (2018) 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. Similar pixel structures were located and used for comparison. The experimental results indicated that the technique performed best for characters printed in the same font setup. It achieved better recognition for printers of the same brand and model. Recently, Kim (2017) used sentence clustering for improved document classification. Research contributions based on global features are as follows. Foremost, Ali et al. (2004) used banding effects present on the document for printer identification. The author discussed that EP printers exhibited quasiperiodic banding artifacts. These artifacts were used as an effective intrinsic signature. This approach worked well for colored printouts but was not suitable for text-only documents. Khanna et al. (2007) performed camera image forensics based on scanner noise analysis. A unique noise pattern of each scanner brand was extracted in the form of 16-D feature vector for source device identification. These features captured the essential properties of the image and discriminated between different scanners. Ryu et al. (2008) developed an image quality measures-based classifier for document forensics. Different measures related to pixel differences and image similarity were proposed. Further, frequency domain and vision characteristics were added. The classifier achieved an accuracy of 80%. Van-Beusekom et al. (2013) classified printers based on yellow point patterns in a document. These yellow dots were specific to a particular printer manufacturer. Patterns from two different document printouts were compared to detect the source printer class. Accuracy of 93.0% was achieved for printer classification. The proposed pattern tracking scheme achieved an accuracy of 91.3% and 98.3% for comparison and decoding, respectively. Elkasrawi and Shfait (2014) extracted features from the noise image, similar to Khanna et al. (2009). Ali et al. (2004) approach was extended and in their extended approach, Low-resolution scanners were used for printer identification. The statistical features based on noise formed by scanners were used. 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.

Jiang et al. (2018) propose a novel multi-channel intelligent attack detection method based on LSTM-RNNs. They introduced a voting algorithm to decide whether the input data is an attack or not. Olakanmi and Dada (2019) presented a morphism approach for the client to efficiently perform the proof of correctness of its outsourced computation without re-computing the whole computation. Ferreira et al. (2017) proposed three different techniques for laser printer identification. The solutions used low-resolution scanned documents. First, the proposed method used two descriptors based on multi-directional and multi-scale 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. 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 et al. (2018) have performed printed source identification using microscopic images. A detailed texture and structure information was obtained due to the 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%. Li et al. (2018) have proposed a novel inkjet printer source identification. Fifteen low-cost inkjet printers were analyzed at a microscopic level. They considered four printer intrinsic features, dot size, dot density, average distance to nearest dot and nearest dot sector. A support vector machine classifier was used and claimed to achieve reliable results.

Most of the contributions for printer attribution either worked on character's local features or printer intrinsic fingerprints. Other explored textural features based on GLCM matrix. None of them has explored the possibility of utilization of key-point-based features like SIFT, SURF or ORB except Gupta and Kumar (2019). Moreover, the SVM classifier was widely used and the other classifier's performance was not compared. This paper presents novelty in terms of feature extraction technique, classifiers explored and use of adaptive boosting for performance improvement. In the present study, the SURF and ORB as feature extraction methodologies and three classification methodologies, namely Naïve Bayes, *k*-NN, and random forest as classification systems, are considered for the printer identification. Different combinations of these classification methods and AdaBoost (Adaptive Boosting) methodology are also explored to improve accuracy. The main contributions of this paper, in this regard, are:

- To study the global features of printed text documents.
- To fix the print technology and the printer make for printed documents.
- To propose and implement a document classifier that can identify an odd document out of a number of questioned documents. 'Odd' here means a document printed from a different printer.

3 Mathematical modeling of the proposed algorithm

It is evident that every printer leaves some fingerprints on the printed document. These fingerprints are unique to every printer. The print technique for various categories of printers is different too. Two main technologies used for printing are Inkjet and Laser. Both techniques work differently and hence have characteristics features. Figure 1 shows the printer fingerprints present in the printed document.

These fingerprints are the characteristic of the printer. Such fingerprints are traced using key-point-based descriptors, i.e., SURF and ORB. A key point is the position where the feature has been detected, while the descriptor is an array containing numbers to describe that feature. In this section, we will discuss the mathematical aspects of the proposed descriptor methodology. All the steps used in the final algorithm are elaborated in the following subsections.

3.1 SURF (Speeded Up Robust Features) descriptor

SURF is a local feature extraction method. It uses a local invariant fast key-point detector for extracting image feature key points. It utilized a distinctive descriptor for extracting the image feature descriptor. It is a fast and robust computational method as compared to the SIFT feature extraction method. It works by extracting the feature key point from an image based on the requirements. The next step is to assign the orientation to the key points. The orientation is assigned in circular motion 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 descriptors.

SURF uses wavelet responses in horizontal and vertical directions for feature extraction. A neighborhood of size 20 s × 20 s is taken around the key point where s is the size. It is divided into 4 × 4 subregions.

$$v = \left(\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right) \tag{1}$$

For each subregion, horizontal and vertical wavelet responses are taken and a vector is formed as shown in Eq. (1). Hence, the SURF feature descriptor with a total 64 dimensions is obtained. But a higher speed of computation and matching can be obtained, if the dimensionality is reduced.

3.2 ORB (Oriented FAST Rotated BRIEF) descriptor

ORB is an emerging local feature extraction method. It utilized FAST (Features from Accelerated Segment Test) key-point detector for extracting image feature key points and BRIEF (Binary Robust Independent Elementary Features) descriptor for extracting image feature descriptor (Vinay et al. 2015). These two are used because of their performance and low computational cost. They are robust to illumination, blur and affine. It is rotation invariant as well as faster than SIFT. For ORB feature extraction, first, the FAST key-point detector is used to detect the possible interested points. Then, the best-interested points are further filtered using the Harris corner detector method. Orientation is applied to corners for providing orientation, and the direction of the patch is used for rotation on binary test patterns.

ORB adds an orientation component to FAST by utilizing an intensity centroid cloud mechanism. The centroid is found by moments of patch as in Eq. (2).

$$m_{ab} = \sum_{xy} x^a y^b I(x, y) \tag{2}$$

where m_{ab} represents the $(a + b)$th order moment of image with intensity values $I(x, y)$.

Further, the centroid is obtained as in Eq. (3) and a vector is obtained.

$$C = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right) \tag{3}$$

Then, the orientation is calculated as in Eq. (4)

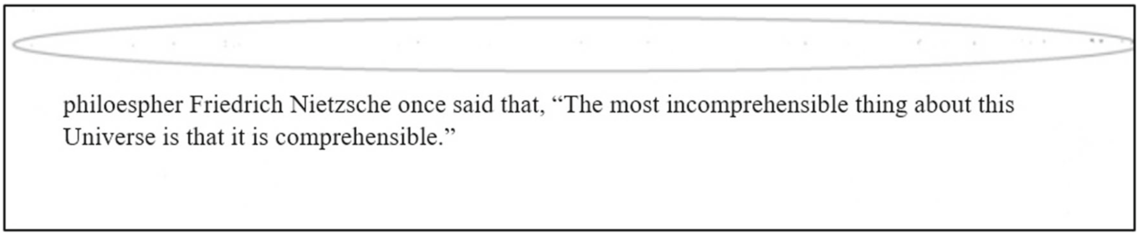
$$\theta = a\tan 2(m_{01}, m_{10}) \tag{4}$$

where a tan 2 is the quadrant aware version of arctan.

It makes the BRIEF rotation invariant by using steered BRIEF.

3.3 *K*-means clustering for clustering of descriptors

Clustering is a very popular image processing technique that groups similar descriptors together. *K*-means



philoespher Friedrich Nietzsche once said that, "The most incomprehensible thing about this Universe is that it is comprehensible."

Fig. 1 Printer fingerprints in a printed document

clustering an unsupervised learning approach in which, group n-dimensional descriptor vector into K number of groups. Clustering technique utilized for assignment of the collection and arrangement of objects such that items in a similar gathering (called a group) are more compared to each other than to those in different gatherings (Rasli et al. 2012).

In this paper, K-means clustering is used to group the similar descriptor obtained using SURF and ORB. It works as follows:

Step 1 Initially choose random K input vectors (data points) cluster initialization

Step 2 Find the cluster center that is closest using Euclidean distance, and assign that input vector to the corresponding cluster for each input vector

Step 3 Update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster

Step 4 Repeat steps 2 and 3 until no more change in the value of the means

3.4 LPP (locality preserving projection) dimensionality reduction

The feature dimensionality reduction is achieved by the LPP method. It reduces the computational space of the algorithm to increase the performance. LPP focuses on the neighborhood connection among the information. It achieves reduction by discarding unimportant parts and reduce the information but preserves the important information (Zhuo et al. 2014). LPP algorithm can be used as an alternative to PCA as PCA fails to capture underlying data structures that lie on a nonlinear manifold. The projections of the algorithm are obtained by firstly building a graph that incorporates neighborhood information of the dataset. Then, a transformation matrix that maps the data points to a subspace is computed using Laplacian of the graph. The main steps are as follows:

Step 1 Construct a neighborhood graph G with n nodes, where n corresponds to the number of variables in the original dataset. Using the k-nearest neighbor algorithm, an edge is placed between node i and j, if i is among k-nearest neighbors of j and vice versa. LPP will consider this graph while choosing projections

Step 2 Choose weights using a Gaussian kernel given graph G, an $m \times m$ weight matrix W is constructed by assigning a weight W_{ij} based on Eq. (5) if a connection (edge) exists between node i and j. A weight of zero is assigned if there is no connection between the nodes. This results in the weight matrix being sparse and symmetric

$$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}} \tag{5)}$$

Step 3 Compute the Laplacian matrix L as in Eq. (6)

$$L = D - W \tag{6}$$

where D is a diagonal matrix whose entries are column sums of weight matrix W as in Eq. (7)

$$D_{ii} = \sum_{i} W_{ji} \tag{7}$$

Compute the eigenvalues and eigenvectors for the generalized eigenvector problem as in Eq. (8)

$$XLX'a = \lambda XDX'a \tag{8}$$

The eigenvector decomposition algorithm yields a full matrix 'a' where the columns correspond to eigenvectors, and a diagonal matrix of generalized eigenvalues, λ. The column vectors of 'a' are ordered according to their eigenvalues in ascending order.

Step 4 Apply linear mapping. The transformation vector $A = (a_0, a_1, \ldots, a_{d-1})$ is then embedded in the linear Eq. (9) to output the transformed data matrix y.

$$y = A'x \tag{9}$$

3.5 Classification

In this section, classification techniques considered in the present work have been discussed. The classification phase utilizes the features extracted for the classification of the objects in a particular class. In this study, Naïve Bayes, k-NN, and random forest classifiers are investigated for classification. Further, their combination is explored with a voting scheme for recognition.

3.5.1 Naïve Bayes

The Naïve Bayes classifier is a classifier method based on clear semantics to represent probabilistic knowledge (John and Langley 1995). This classifier considers the most important information and makes simple assumptions for the same. Its working is based on the fact that predictive characteristics are independent in a given class. Another assumption is that the prediction process is not influenced by any hidden or latent attributes.

3.5.2 *k*-NN

k-nearest neighbor (*k*-NN) is based on the study of neighboring samples in the training feature set. *k*-NN is a lazy machine learning algorithm. In this technique, first, the locations and labels of the training samples are used to divide the space into regions. The most frequent class among the *k*-nearest training samples is assigned a position in the space. Usually, Euclidean distance is used to calculate the distance between the stored feature vector and candidate feature vector in *k*-nearest neighbor algorithm.

3.5.3 Random forest

Random forest is an ensemble algorithm that combines many algorithms together for classification problems. Random forest eliminates the problem of over-fitting experienced in the case of a decision tree. A random forest classifier collects the majority votes from different decision trees and then predicts the classification results. The random forest uses mean values to improve perceptive accuracy. Achieved accuracy of random forest is outstanding among existing supervised learning algorithms. They are remarkably efficient on large databases is remarkable (Breiman 2001).

3.6 Adaptive boosting

Boosting is a way to manage machine learning in light of making a precise expectation rule by combining many less efficient and inaccurate rules. The AdaBoost algorithm of Freund and Schapire was the most efficient boosting algorithm (Freund and Schapire 1999). This algorithm is widely used for numerous applications in multiple domains. Many efforts have been made to clarify why it works, how it works, and what are its capacities. In this paper, we have used this methodology for improving the classification results of printed documents. AdaBoost is a classifier with high precision. It gives the structure to order and makes the building up of sub-classifiers simple. In the present paper, we have considered three classifiers, namely Naïve Bayes, *k*-NN and random forest as discussed above in Sect. 3.5. Experimental outcomes based on the above-mentioned classifiers, their combination and AdaBoost algorithm are discussed in the next section.

Proposed Algorithm

The proposed methodology (Fig. 2) and the algorithm are discussed as follows:

Proposed Algorithm:

Step 1. Input digital image of printed text document

Step 2. Extract feature descriptor vector using ORB and SURF features for each image in the dataset as discussed in Sect. 3.1 and 3.2

Step 3. Use the *K*-means clustering algorithm on the feature descriptor vector as discussed in Sect. 3.3. *K* numbers of clusters are generated for every descriptor vector. Compute the mean of every cluster

Step 4. Use the LPP dimensionality reduction algorithm to reduce the feature vector dimensions as discussed in Sect. 3.4. A 48-dimensional feature vector is reduced to 8-D for both ORB and SURF

Step 5. Combines both SURF and ORB feature vectors, store them in database image features for training and testing purposes

Step 6. Train the proposed system using Naïve Bayes, *k*-NN and random forest classifiers as discussed in Sect. 3.5

Step 7. Apply AdaBoost to further enhance the accuracy of the model as discussed in Sect. 3.6

Step 8. Predict the class of questioned documents by submitting their ORB and SURF features to the trained classifier

Step 9. Return the class of printer as output for the questioned document

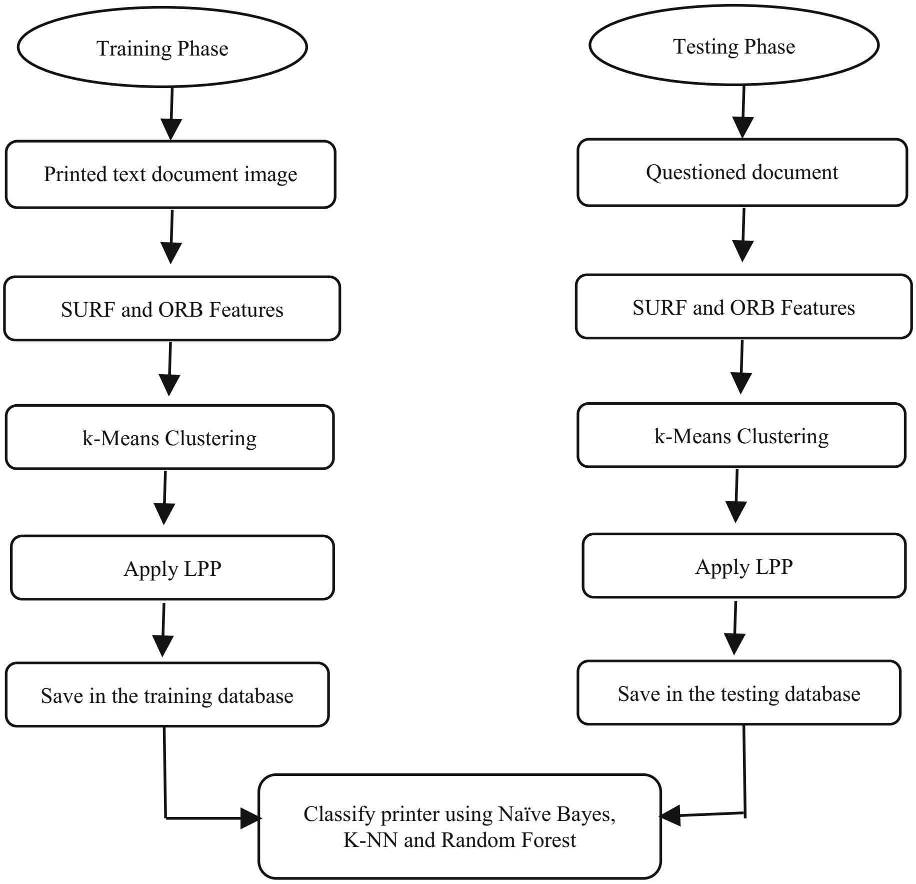
4 Experimental results and discussion

This section includes the details of experiments conducted, their analysis and comparison with other parallel techniques for printer attribution.

4.1 Dataset

The experimental results for the proposed model-based classifier are obtained using a public dataset proposed by Khanna et al. (2007). This dataset contains printed documents from 20 inkjet and laser printers. Fifty documents per printer are taken into consideration. All documents printed by a printer are unique. Document of three categories, i.e., contract, invoice and scientific papers are included in the dataset. This diversity features unique challenges for the feature extraction and anomaly detection process. For every printer, a unique dataset has been created in order to ensure a content-independent feature extraction system. The following document types are included.

Fig. 2 Proposed methodology


Training Phase
Printed text document image
SURF and ORB Features
k-Means Clustering
Apply LPP
Save in the training database
Testing Phase
Questioned document
SURF and ORB Features
k-Means Clustering
Apply LPP
Save in the testing database
Classify printer using Naïve Bayes, K-NN and Random Forest

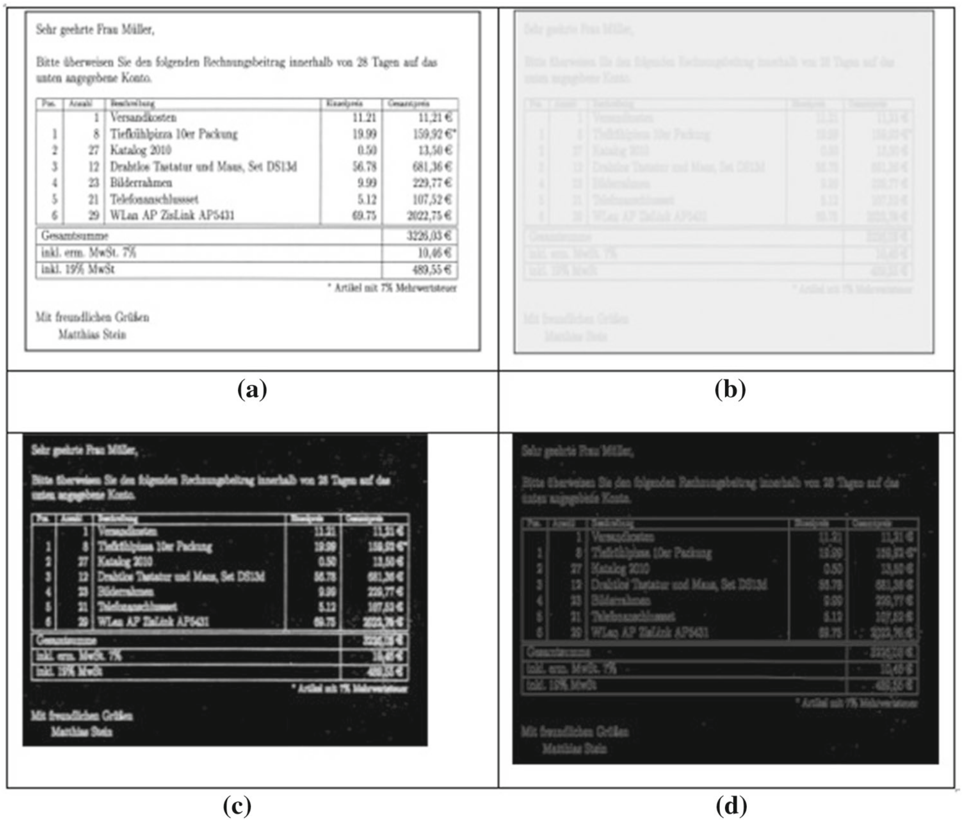
1. Contracts: The contract only contains text but in different font types and sizes. A contract will never contain pictures, lines and diagrams. The contracts were created automatically using a Python script.
2. Invoices: The invoices feature different font sizes and variety as well as vertical and horizontal ruling lines. It has logos, composed of a small picture and colored text. Like the contracts, these documents are also created using a Python script.
3. Scientific Literature: The last type contains real-world examples, pages taken from existing scientific papers and books. They feature a large variety of content, e.g., different font types, and sizes as well as pictures, diagrams and formulas.

This dataset is the first of its kind and has the variety and richness. It features realistic document types of varying difficulty. A subset of 07 inkjet printers and 13 laser printers are considered for performance evaluation of the proposed classification system. For this purpose, the features are extracted from each document image. Each image of the dataset is resized into a size of 320 × 240. Figure 3 depicts the close view of printed text by two different printers. Figure 4 depicts the noise and edge images obtained for a sample document used during analysis. The names of source printers considered in this work are depicted in Table 1.

Fig. 3 Samples taken from Ink Officejet 5610 and Laser Samsung CLP 500

4.2 Experimental setup

For experimental results, the entire dataset is partitioned into a training dataset and testing dataset. In used partitioning strategy, 80% data are taken as a training dataset and remaining data are taken as a testing dataset. Fivefold cross-validation technique is also used for assessing the effectiveness of the proposed system. Three classifiers, namely Naïve Bayes: C1, *k*-NN: C2 and random forest: C3 are considered in this work in order to classify the data. A performance analysis is carried out with 80% data as training data and the remaining 20% data as a testing dataset.



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	Telefonanschlusset	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

Fig. 4 Samples of **a** original, **b** noisy, **c** logarithmic, **d** edge document image

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 Laserjet4350
s	Laser	HP Laserjet 5
t	Laser	Epson Aculaser C1100

4.3 Discussion of results

A recognition rate of 82.1% and 82.5% has been achieved for partitioning strategy and fivefold cross-validation technique with a combination of Naïve Bayes, *k*-NN and random forest classifiers as depicted in Table 2, Figs. 5, 6 and 7. The results of various experiments demonstrated that our algorithm can accomplish a higher correct rate of 86.5% and 83.2% for partitioning strategy and fivefold cross-validation technique, respectively, with AdaBoost methodology as depicted in Table 3, Figs. 6 and 8. The confusion matrix for the accuracy of five-fold cross-validation (83.2%) is presented in Fig. 9.

4.4 Comparison with other techniques

In this paper, authors have presented a passive model for printer attribution based on Speeded Up Robust Features (SURF) and Oriented Fast Rotated and BRIEF (ORB). The size of SURF and ORB descriptors requires a high memory space for storing features. Therefore, a *K*-Means clustering algorithm and LPP has also been considered. *K*-means algorithm will cluster and thus reduce the descriptor into 64 clusters and LPP reduces them to 8 components each for SURF and ORB both features. Three classifiers, namely *k*-NN, Naïve Bayes, random forest and their combination are

Table 2 Results achieved using partitioning strategy (80% training data set and 20% testing data set) and using fivefold cross-validation technique

Classifier	Partitioning Strategy			fivefold cross-validation		
	SURF (%)	ORB (%)	SURF + ORB (%)	SURF (%)	ORB (%)	SURF + ORB (%)
Naive Bayes	48.0	37.0	80.0	52.6	41.2	77.4
k-NN	64.5	60.5	57.0	66.9	56.7	55.2
Random Forest	76.5	66.5	79.8	75.0	66.8	81.0
Naive Bayes + K-NN	65.0	60.5	68.0	67.4	56.9	65.3
Naive Bayes + Random Forest	74.5	62.5	82.5	70.6	64.7	82.6
K-NN + Random Forest	64.0	60.5	57.0	55.4	56.7	55.4
Naive Bayes + K-NN + Random Forest	72.9	64.8	82.5	72.9	64.8	82.1

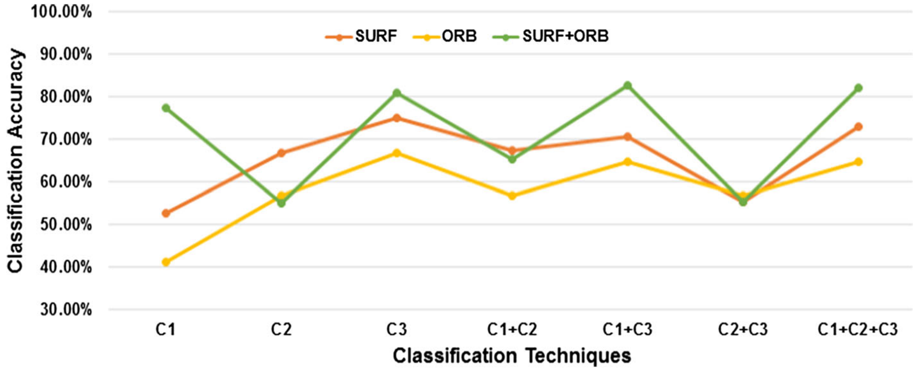


Fig. 5 Results achieved using partitioning strategy (80% training data set and 20% testing data set)

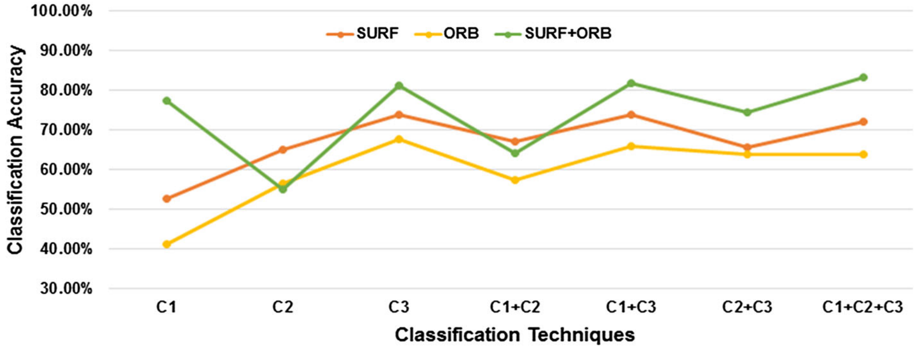


Fig. 6 Results achieved using fivefold cross-validation technique

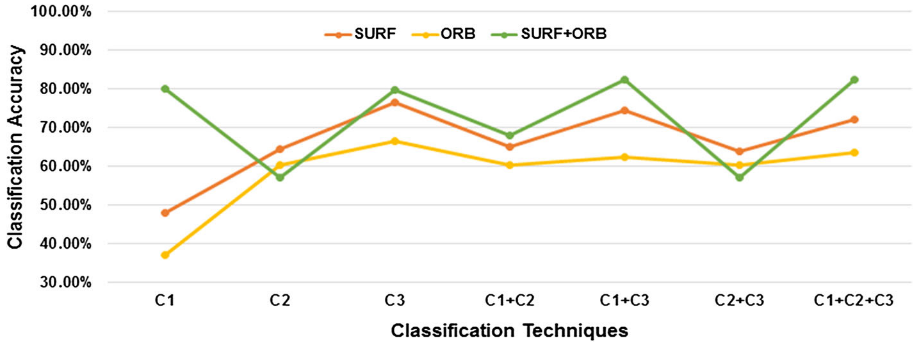


Fig. 7 Results achieved using partitioning strategy with adaptive boosting methodology

Table 3 Results achieved using partitioning strategy and fivefold cross-validation with adaptive boosting methodology

Classifier	Partitioning strategy with adaptive boosting methodology			Fivefold cross-validation with adaptive boosting methodology		
	SURF (%)	ORB (%)	SURF + ORB (%)	SURF (%)	ORB (%)	SURF + ORB (%)
Naive Bayes	48.0	37.0	80.0	52.6	41.2	77.4
k-NN	62.0	60.0	57.0	65.2	56.6	55.2
Random Forest	75.5	64.5	84.0	73.9	67.6	81.2
Naive Bayes + K-NN	65.0	58.5	67.5	67.1	57.5	64.1
Naive Bayes + Random Forest	73.5	63.5	82.0	73.9	66.1	81.8
K-NN + Random Forest	64.0	58.5	76.5	65.7	64.0	74.4
Naive Bayes + K-NN + Random Forest	70.5	64.0	86.5	72.2	64.0	83.2

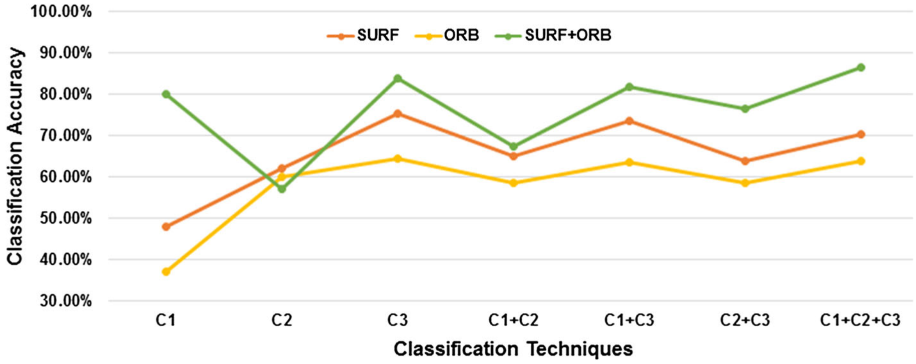
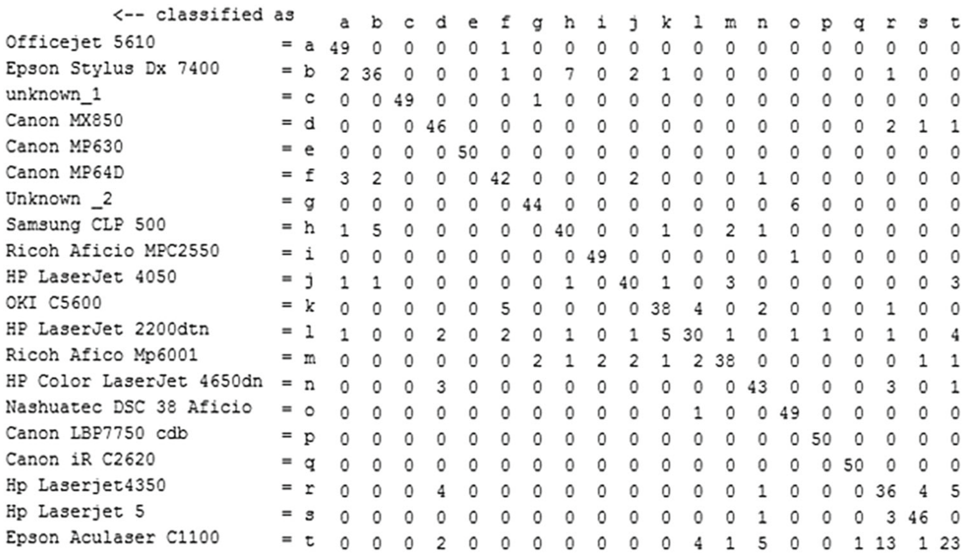


Fig. 8 Results achieved using fivefold cross-validation with adaptive boosting methodology



```
                <-- classified as    a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t
Officejet 5610                = a   49  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
Epson Stylus Dx 7400          = b    2 36  0  0  0  1  0  7  0  2  1  0  0  0  0  0  0  1  0  0
unknown_1                     = c    0  0 49  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0
Canon MX850                   = d    0  0  0 46  0  0  0  0  0  0  0  0  0  0  0  0  0  2  1  1
Canon MP630                   = e    0  0  0  0 50  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
Canon MP64D                   = f    3  2  0  0  0 42  0  0  0  2  0  0  0  1  0  0  0  0  0  0
Unknown _2                    = g    0  0  0  0  0  0 44  0  0  0  0  0  0  0  6  0  0  0  0  0
Samsung CLP 500               = h    1  5  0  0  0  0  0 40  0  0  1  0  2  1  0  0  0  0  0  0
Ricoh Aficio MPC2550          = i    0  0  0  0  0  0  0  0 49  0  0  0  0  0  1  0  0  0  0  0
HP LaserJet 4050              = j    1  1  0  0  0  0  0  1  0 40  1  0  3  0  0  0  0  0  0  3
OKI C5600                     = k    0  0  0  0  0  5  0  0  0  0 38  4  0  2  0  0  0  1  0  0
HP LaserJet 2200dtn           = l    1  0  0  2  0  2  0  1  0  1  5 30  1  0  1  1  0  1  0  4
Ricoh Afico Mp6001            = m    0  0  0  0  0  0  2  1  2  2  1  2 38  0  0  0  0  0  1  1
HP Color LaserJet 4650dn      = n    0  0  0  3  0  0  0  0  0  0  0  0  0 43  0  0  0  3  0  1
Nashuatec DSC 38 Aficio       = o    0  0  0  0  0  0  0  0  0  0  0  1  0  0 49  0  0  0  0  0
Canon LBP7750 cdb             = p    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 50  0  0  0  0
Canon iR C2620                = q    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 50  0  0  0
Hp Laserjet4350               = r    0  0  0  4  0  0  0  0  0  0  0  0  0  1  0  0  0 36  4  5
Hp Laserjet 5                 = s    0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  3 46  0
Epson Aculaser C1100          = t    0  0  0  2  0  0  0  0  0  0  0  4  1  5  0  0  1 13  1 23
```

Fig. 9 Confusion matrix of results using a combination of Naive Bayes + K-NN + Random Forest and AdaBoost methodology with five-fold cross-validation

considered for the classification task. The proposed model can efficiently classify the questioned documents to their respective printer class. Experimental results have affirmed the viability of the proposed approach and proved the characteristic advantages.

The comparison of the present work is done with texture-based GLCM technique by Mikkilineni et al. (2011) and Cross Center-symmetric LTP (CCSLTP) by Fu and Yang (2012). The classification accuracy of these algorithms is listed in Table 4. The best accuracy for the proposed system has been obtained by using an adaptive boosting methodology. The best precision rate of 85.6% has been achieved using a combination of SURF + ORB features and adaptive boosting methodology.

Table 4 Comparison with state-of-the-art work

Feature extraction technique	Feature size (1-D)	Accuracy (%)
GLCM	12	86.2
CCSLTP	128	57.8
SURF	8	72.9
ORB	8	64.8
SURF + ORB	16	82.5
SURF + ORB with ADABoost	16	86.5

5 Conclusion

A computational approach for printed document forensics has been proposed using global features such as SURF and ORB to classify the documents printed by different printer resources. An effective classifier model is proposed. The classifier aims to fix the print technology and the printer make for printed documents. The document classifier can identify an odd document out of a number of questioned documents. This paper presents novelty in terms of feature extraction technique, classifiers explored and use of adaptive boosting for performance improvement. In the present study, the SURF and ORB as feature extraction methodologies and three classification methodologies, namely Naïve Bayes, *k*-NN, and random forest as classification systems are considered for the printer identification. A public database of printed documents using different printers is used to validate the results. Classification accuracy of 86.5% has been obtained using a combination of Naïve Bayes, *k*-NN and random forest classifiers with AdaBoost methodology.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Ali, G.N., Mikkilineni, A.K., Allebach, J.P., Delp, E.J., Chiang, P.J., Chiu, G.T.: Intrinsic and extrinsic signatures for information hiding and secure printing with electrophotographic devices. In: Proceedings of the Non-impact Printing and Digital Fabrication Conference, New Orleans, Louisiana, vol. 2, pp. 511–515 (2003)

Ali, G.N., Mikkilineni, A.K., Delp, E.J., Allebach, J.P., Chiang, P.J., Chiu, G.T.: Application of principal components analysis and gaussian mixture models to printer identification. In: Proceedings of the Non-impact Printing and Digital Fabrication Conference, Salt Lake City, Utah, vol. 1, pp. 301–305 (2004)

Bertrand, R., Gomez-Kramer, P., Terrades, O.R., Franco, P., Ogier, J.M.: A system based on intrinsic features for fraudulent document detection. In: Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington DC, USA, pp. 106–110 (2013)

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

Elkasrawi, S., Shafait, F.: Printer identification using supervised learning for document forgery detection. In: Proceedings of the 11th IAPR International Workshop on Document Analysis Systems, France, pp. 146–150 (2014)

Ferreira A, Bondi L, Baroffio L, Bestagini P, Huang J, dos Santos J, Tubaro S, Rocha A (2017) Data-driven feature characterization techniques for laser printer attribution. IEEE Trans. Inf. Forensics Secur 12(8):1860–1873

Freund Y, Schapire RE (1999) A Short Introduction to Boosting. J. Jpn. Soc. Artif. Intell. 14(5):771–780

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.: Document authentication using printing technique features and unsupervised anomaly detection. In: Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, US, pp. 479–483 (2013)

Gupta S, Kumar M (2019) Forensic document examination system using boosting and bagging methodologies. Soft Comput. https://doi.org/10.1007/s00500-019-04297-5

Jiang F, Fu Y, Gupta BB, Lou F, Rho S, Meng F, Tian Z (2018) Deep learning based multi-channel intelligent attack detection for data security. IEEE Trans. Sustain. Comput. https://doi.org/10.1109/TSUSC.2018.2793284

John, G.H., Langley, P.: Estimating Continuous distributions in bayesian classifiers. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, pp. 338–345 (1995)

Joshi S, Khanna N (2018) Single classifier-based passive system for source printer classification using local texture features. IEEE Trans. Inf. Forensics Secur. 13(7):1603–1614

Khanna, N., Mikkilineni, A.K., Chiu, G.T.C., Allebach, J.P., Delp, E.J.: Scanner identification using sensor pattern noise. In: Proceedings of the Security, Steganography, and Watermarking of Multimedia Contents, Electronic Imaging, San Jose, CA, US, 65051K1-K11 (2007)

Khanna N, Mikkilineni AK, Delp EJ (2009) Scanner identification using feature-based processing and analysis. IEEE Trans. Inf. Forensics Secur. 4(1):123–139

Kim M (2017) Simultaneous learning of sentence clustering and class prediction for improved document classification. Int. J. Fuzzy Logic Intell. Syst. 17(1):35–42. https://doi.org/10.5391/IJFIS.2017.17.1.35

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

Mikkilineni, A.K., Chiang, P.J., Ali, G.N., Chiu, G.T.C., Allebach, J.P., Delp, E.J.: Printer identification based on graylevel cooccurrence features for security and forensic applications. In: Proceedings of the Security, Steganography, and Watermarking of Multimedia Contents, Electronic Imaging, California, USA, pp. 430–440 (2005)

Mikkilineni, A.K., Chiang, P.J., Ali, G.N., Chiu, G.T.C., Allebach, J.P., Delp, E.J.: Printer identification based on texture features. In: Proceedings of the Non-impact Printing and Digital Fabrication Conference, Society for Imaging Science and Technology, Salt Lake City, Utah, vol. 1, pp. 306–311 (2004)

Mikkilineni AK, Khanna N, Delp EJ (2011) Forensic printer detection using intrinsic signatures. In: SPIE proceedings, media watermarking, security, and forensics III, vol. 7880. 78800R. https://doi.org/10.1117/12.876742

Olakanmi OO, Dada A (2019) An efficient privacy-preserving approach for secure verifiable outsourced computing on untrusted platforms. Int. J. Cloud Appl. Comput. 9(2):79–98

Rasli, R.M., Zalizam, T., Muda, T., Yusof, Y., Bakar, J.A.: Comparative analysis of content based image retrieval techniques using color histogram: a case study of GLCM and K-Means clustering. In: Proceedings of the Third International Conference on Intelligent Systems Modelling and Simulation, pp. 283–286 (2012)

Ryu SJ, Lee HY, Cho IW, Lee HK (2008) Document forgery detection with SVM classifier and image quality measures. In: Proceedings of the 9th pacific rim conference on multimedia (PCM'08), pp 486–495

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

Tsai MJ, Yuadi I, Tao YH (2018) Decision-theoretic model to identify printed sources. Multimed. Tools Appl. 77:27543–27587

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

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

Zhuo L, Cheng B, Zhang J (2014) A comparative study of dimensionality reduction methods for large-scale image retrieval. Neurocomputing 141:202–210

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