FORGED ARTICLE RECOGNITION USING IMAGE PROCESSING AND INTELLECTUAL ACTIVITIES

1. Dr R N MUHAMMAD ILYAS
Assistant Professor
Department of Computer Science
The New College (Autonomous)
Chennai – 14 , India

2. *Dr M WASIM RAJA
Assistant Professor
Department of Computer Science
The New College (Autonomous)
Chennai – 14 , India

ABSTRACT

The forgery of articles has become common and this made a lot of trouble and difficulties to the officials. With new and classy digital printers and lot of software tools it became very simple to revise scanned articles. It becomes very difficult to differentiate from the original one and the forgery one. The present article identification is not so proficient, so some people construct forged articles and do unlawful activities. Manually validating the authenticity of the substantial articles enhances the administrative overhead and acquires more time.

Then image processing methods were used, but the majority of the images processing based forgery article identification methods were less accurate. To get better accuracy, our proposed system put forward an automatic article verification replica using Convolutional Neural Networks (CNN). Additionally, we make use of Optical Character Recognition (OCR) and Linear Binary Pattern (LBP) to extract the textual information and regional edges from the articles.

The proposed system also includes two methods to identify the forged article. First the QR-code scanner scans the code of the article and identifies that the article is original or forged. Second, the image processing techniques goes through three phases: Training phase, testing phase and Classification phase to identify the forged articles. Hence the uniqueness of article is examined and determined by making the forged article more robust and reliable.

Keywords: Article verification, convolution neural network, forgery article identification, article processing, LBP, OCR

1. INTRODUCTION

In recent world the articles can be manipulated effortlessly, Trust worthiness of articles is highly in demand. Officially, numerous technologies were developed and they were not as much as effective in countering the threat of identifying forged articles. So some new system should be created to control the threat. Several preventive methods have been taken by the government to end these forgery activities but still it is less effective.

The practice of using forged articles in the internet has enlarged. Now-a-days, every banking sector collects data from their clients. On the other side, the tradition of forged article in the internet is also rising swiftly. About 500,000 people in the United States are victims of forged articles every year. Identifying the forged article is a really difficult task [1]. Certifying that the entire articles are legal is very important because for a moment a tiny division of the article might have been forged and all collectively might prove legitimate. Several Multinational companies, Government and Private organizations are outsourcing the article authentication process to third party Business Process Outsourcing providers (BPO). By computerizing the article authentication process, enormous amount of money and manpower can be reduced.

The proposed system uses image processing techniques to identify forgery in article. The main objective of the proposed system is to create a fast and most efficient system for identifying forgery in articles. The proposed system also includes two methods to identify the forged article. First the QR-code scanner scans the code of the article and identifies that the article is original or forged. Second the image processing uses neural network concept.

In our proposed work, the uniqueness of article is discussed and focused on making the identification of forgery article more robust and reliable. Such systems are essential at the time of submission of articles on different web portals where it makes sure whether the article is real or not. After the advancement of computers and internet technologies, many organizations (Multi-national companies, Government Sectors, Banking Sectors, Private Sectors etc) have implemented a variety of computer based techniques to handle their day to day requirements. One of such essential implementation is digital document.

To translate the physical article into machine readable system, the OCR system changes the physical article into a gray scale image. Then the OCR system categorizes the dark and light areas of the gray scale image [2]. The dark areas are recognized as characters or letters and light area are recognized as background of the article.

Then it measures the authenticity of the article using Euclidian distance. In this method, the outcome extremely depends on the Euclidian distance between each pixel. Later, Pawel et al. [3], proposed a method to identify and classify genuine and forged articles. It uses a wide range of techniques from image processing, pattern recognition techniques. The complete article is transformed into binary images and then the Gray Level Co-occurrence Matrix (GLCM) is formed to identify the genuineness of the article. Similarly, Luiz et al. [4], used Convolution Neural Network along with suitable organization of adversarial examples to classify the article. Later, Nabil et al. [5], projected a method to validate the article which uses pattern matching and recognition methods to carry out the verification. In this research work, we have produced a Convolution Neural Network based on repeated article verification system to verify the uniqueness of entire article.

The proposed system allows the users to upload their article directly to the verification system. The system intake each and every article of the users automatically and pulls out all the key information in sequence and identifies the genuineness of the article. The proposed system uses Optical Character Recognition (OCR) and Linear Binary Pattern (LBP) to extract all the vital textual information from the articles with its regional edges.

Our contribution to the proposed automatic forgery identification system are, building a faster and accurate forgery article detecting system to detect all the possible forgeries in the article based on sliding window Convolutional Neural Network (CNN).

Also to obtain a high class texture feature extraction using Optical Character Recognition (OCR) method and Local Binary Pattern (LBP) system to examine the forged text in each neighbourhood pixel.

2. METHODOLOGY

The system which we put into operation first scans the QR-code of the article using Image processing techniques in deep learning. The Image Processing system mainly consists of two parts. They are Error Level Analysis and Neural Network. This combination helps us to identify whether the given article is manipulated by some resources or not. Deployment segment of our system plays a vital role in two ways. They are QR-code scanner components: The QR-code of the scanned article is verified so as to produce the encrypted data of the article and identify the forged data. Image Processing Component: The image processing component detects the article through the concept of neural network and error value analysis and identifies the article is forged or sole. Dealing with the challenges in the article authentication and deciding the forged article is extremely appreciated in research area. To diminish the challenges in forgery article authentication, numerous techniques were proposed.

The key security way used to guard the physical articles was holograms, seals and watermarks. However, when the physical article is scanned using Smartphone, scanners or any other devices, the quality of the holograms, seals and watermarks are reduced. Therefore, detecting the genuineness has become a difficult task. Later on years, 2D barcodes and QR code were launched. Barcodes were able to accommodate data's in a visual, machine readable design. The data's accommodated in the barcodes were used to detect the genuineness of the article. Based on the temperament of the forgery identification methods, they are classified into (a) textual forgery identification and (b) image forgery identification (holograms, seals and watermarks).

3. RELATED WORKS

The related works that have been performed on the texts and images based forgery identification methods are described below

3.1 Textual forgery identification in scanned articles:

Khan et al. [6], proposed a method to recognize the authenticity of the handwritten articles in a physical article by the hyper spectral imaging technique. To verify the authenticity of the article, the hyper spectral imaging technique were used which identifies the ink colour mismatch. If the words are written using the similar ink, then it declares that the whole article is non-forged article. If any of the 'alphabets' or 'expressions' gets mismatched with respect to the ink used, then the particular section of the data's are said to be as forged text. However the truthfulness of Khan et al. [6] is found to be not as much of expected. If the article along with forged texts are printed using the same printer, subsequently the proposed method does not succeed to detect the forgery.

Abramova et al. [7], proposed a new technique which identifies the near duplicates of the in the scanned article. To identify the textual forgeries in the articles, CopyMove Forgery Detection (CMFD) algorithm is used. Khan et al. [8], proposed a technique to improve the accuracy of ink mismatch recognition in Khan et al. [6], by means of fuzzy clustering and hyper spectral imaging technique. Maryam et al. [9], proposed a text forgery recognition using Convolution Neural Network (CNN) technique. It utilizes of a text free approach for efficient categorization of source printer by means of deep visual features of the letters, numbers and special characters in the article.

3.2 Image forgery identification in scanned articles:

In recent days, majority of the private and governmental sector depend on the digital articles. Certifying the originality of the articles is very much essential. Tsai et al. [10], proposed a technique to identify fake text in a printed article using the traditional image forgery method which can be classified into Copy Move Forgery Detection (CMFD) and Image splitting identification. In copy-move forgery, a particular section of an image is chosen and pasted into the other section of the same image. Now there will be a huge correlation among the two sections.

Mahmood et al. [11], proposed a novel technique to identify copy-move forgery assault by dividing the images into overlapping square blocks and DCT (Discrete Cosine Transform) and subsequently the Gaussian RBF kernel and PCA is applied to detect the forged sector in the image. An additional conventional technique of image forgery identification is, image splicing method. Here, a particular sector of an image is taken and they are included with few other image. The image splicing identification examines the clues that are left behind the tampering procedure on an image. The tampered sector is expected to contain an abnormal artifact in the edges. The major common abnormal activities of tampered sector are mismatch of letters due to brightness, discontinuity of edge, geometric place. Linear Discriminative Analysis (LDA) is used to distinguish the genuine and forged sector depending on the blur area.

3.3 Textual extraction using OCR and LBP in training set:

Segmentation is a procedure which finds out the component of an image. It detects the vital sectors of the image where data are written or printed. Segmentation method splits the text character from the figures, holograms, seals and watermarks images. In our proposed work the datas are segmented as a character and they are set as an input to the OCR unit. The proposed OCR algorithm segments the words into lonely characters that are recognized independently. The OCR carries out a planned analysis of the characters like texts or number etc to identify the patterns. It pulls out the topological and geometric pattern of each and every character. By utilizing these patterns and geometric values, this method tries to produce the physical form of the character. This planned analysis system gives enormous easiness to noisy articles.

Local Binary Pattern (LBP) is a proficient structural pattern which provides the pixels of a number and character by setting the threshold value on the circular neighbourhoods and provides the binary end result. In our proposed method, on behalf of every character, three sampling points are chosen and the binary values of their circular neighbourhood are calculated. To carry out LBP on behalf of every pixel in a scanned article is a time consuming procedure. Based on the significance of the characters, the sampling points can be customised.

4. PROPOSED MODEL

The core intention of the proposed forgery identification method is to examine every possibility of forgeries in an article with a smaller amount of time and produce an enhanced accuracy than existing systems. The first and foremost step after the article comes into the Forgery article identification method is preprocessing. The user uploads the article and it may include a few amounts of noise or defects based on the resolution of the scanned article. To decrease the noise in the article, preprocessing is carried out. Preprocessing utilizes filling and thinning method to enhance the quality of the character. Filling removes the minute holes, gaps, and breaks in the characters and thinning enhances the boundary line of the characters. When the scanned article is preprocessed, the next move is to pull out the textual and image information existing in the article. To pull out the information from the article, Optical Character Recognition (OCR) is utilized. It mines the textual information from the scanned article and translates into a planned data which is searchable and editable. OCR doesn't have the ability to identify the forged 'numbers' and 'letters' in a scanned article. To identify the edges of the forged letters, Local Binary Pattern (LBP) algorithm is utilized along with OCR. Algorithms utilized in the feature extraction segment is shown in Fig. 1

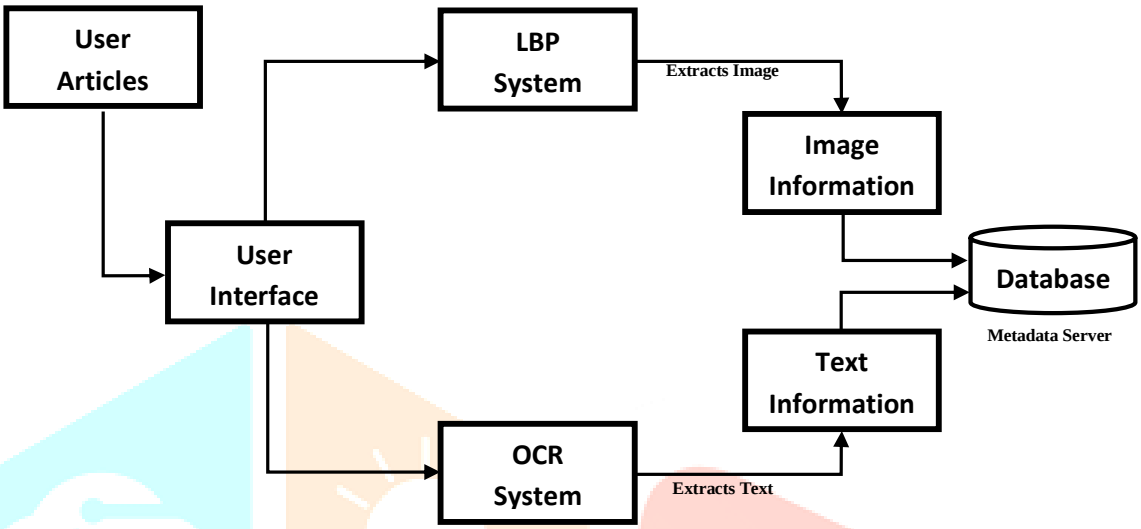


Fig 1: Flowchart of the proposed system for feature extraction

5. ALGORITHM:

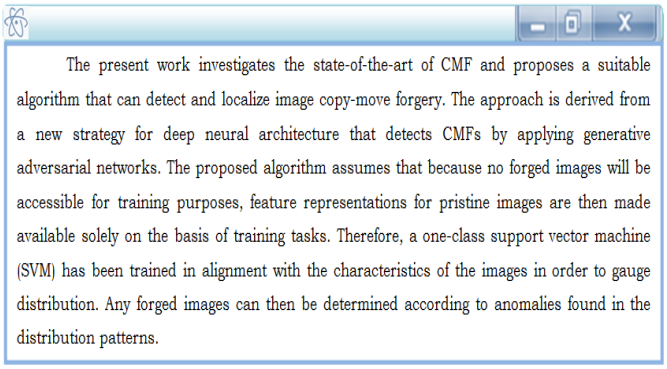
Proposed algorithm to for feature extraction using OCR and LBP

Step 1: Choose the scanned article from virtual dataset so as to perform pre-processing
Step 2: Convert the scanned image into gray level if the scanned article is coloured
Step 3: Perform Normalization to attain a standard size of image and remove noise from the scanned article
Step 4: Perform Rotation Correction if the scanned article is undergoes from wrap
Step 5: Save pre processed image to virtual dataset.
Step 6: Segment the pre processed image into three part (holograms, seals and watermarks)
Step 7: Then feature extraction is applied.
Step 8: Evaluate the performance based on Accuracy, Sensitivity and Specificity for analysing the proposed method
Step 9: Repeat Step 3 to Step 8 until we attain a better performance than the existing method

6. SYSTEM EVALUATION

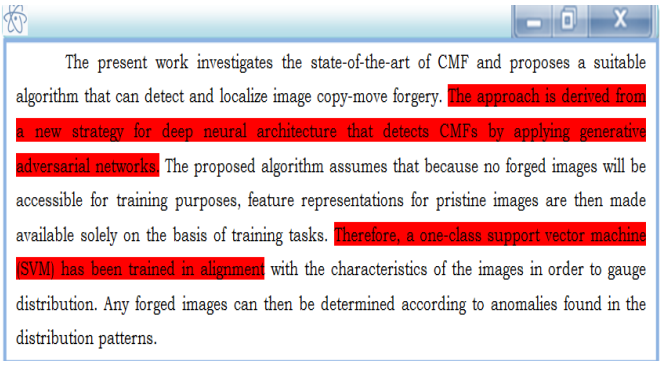
The first process contains QR-code Scanner application which is used for identifying the forged article. The QR-code includes encryption code of the article which is used to identify the forged articles. If the article is genuine, then the QR- codes in those articles provides the encrypted code of the articles. However, if the article is fake, then our proposed work does not produce any encrypted code and inform that the article is fake.

The research of the proposed forged article detection using LBP and OCR is performed using python programming language. Our planned OCR research is created to pull out the English letters. While utilizing the English text scanned articles, the accuracy is attained up to 96% even in low-brightness. The proposed structure is assessed using different criterions. The features we measured for assessing the performance are (a) Accuracy of OCR (b) Computation of time taken to complete feature recognition (c) Accuracy of the text forgery detection. (d) Accuracy of image (seal and hologram) forgery detection. The dataset utilized in the proposed work contains around 480 scanned images, with each and every image has the resolution of 1248 * 852 pixels and of 834Kb size.



The present work investigates the state-of-the-art of CMF and proposes a suitable algorithm that can detect and localize image copy-move forgery. The approach is derived from a new strategy for deep neural architecture that detects CMFs by applying generative adversarial networks. The proposed algorithm assumes that because no forged images will be accessible for training purposes, feature representations for pristine images are then made available solely on the basis of training tasks. Therefore, a one-class support vector machine (SVM) has been trained in alignment with the characteristics of the images in order to gauge distribution. Any forged images can then be determined according to anomalies found in the distribution patterns.

Fig 2(a): Scanned article taken as input



The present work investigates the state-of-the-art of CMF and proposes a suitable algorithm that can detect and localize image copy-move forgery. The approach is derived from a new strategy for deep neural architecture that detects CMFs by applying generative adversarial networks. The proposed algorithm assumes that because no forged images will be accessible for training purposes, feature representations for pristine images are then made available solely on the basis of training tasks. Therefore, a one-class support vector machine (SVM) has been trained in alignment with the characteristics of the images in order to gauge distribution. Any forged images can then be determined according to anomalies found in the distribution patterns.

Fig 2(b): Forged article extracted using OCR and LBP

To estimate the accuracy of the proposed methods, four key factor are uses such as true positive (TP), false positive (FP), true negative (TN) and false negative (FN). The true positive is a value where the identification of forged data is true, as the forgery is made to the article. The false positive is the case where the article is forged; however the prediction method identifies it. The true negative is the case where no forged data may be available however the expected result might be forged. The false negative is the case where the article is forged and it is not identified. The effectiveness of this method can be established by calculating the accuracy (A), sensitivity (S) and the specificity (P) as,

$$\text{Accuracy (A)} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{---------- (1)}$$

$$\text{Sensitivity (S)} = \frac{TP}{TP + FN} \quad \text{---------- (2)}$$

$$\text{Specificity (P)} = \frac{TP}{TP + FP} \quad \text{---------- (3)}$$

The performance evaluation of the text forgery identification in a scanned article using OCR and LBP is derived and compared with existing model (Maryam et al. [9] in Fig. 2(a) and Fig. 2(b)

S.No	Number of Characters in Article	Resolution	Size of Scanned Article in KB	CRC	WRC	Existing Method in %			Proposed Method in %		
						A	S	P	A	S	P
1	857	1248 * 852	834	720	137	87	76	88	90	79	91
2	935	1249 * 852	834	785	150	89	82	90	92	85	93
3	865	1250 * 852	834	727	138	86	77	87	89	80	90
4	896	1251 * 852	834	753	143	88	71	85	91	74	88
5	806	1252 * 852	834	677	129	90	80	86	93	83	89

Table 1: Feature Detection of Proposed and Existing method using OCR and LBP Patterns

Here, CRC – Correctly Recognised Characters, WRC – Wrongly Recognised Characters, A – Accuracy, S – Sensitivity, P – Specificity

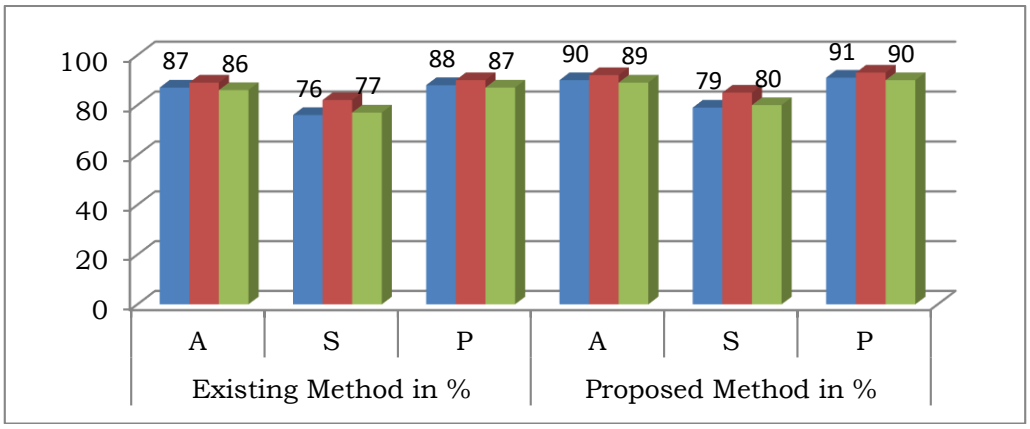


Fig 2: Performance Evaluation of Proposed and Existing method using OCR and LBP Patterns

7. CONCLUSION AND FUTURE SCOPE

In our research work, a forged article identification system is developed by utilizing Linear Binary Pattern (LBP) and Optical Character Recognition (OCR). The proposed method is capable of identifying whether the article is real, genuine and proficient. The core intention of our proposed method is to examine all the possible forgeries of a scanned article. Our proposed method utilizes an Optimal Character Recognition (OCR) technique to extract texts and Local Binary Pattern (LBP) algorithm to extract the non textual images (Seal, Hologram and Stamps) information from the scanned article. To educate our model, we used around 480 scanned images, with each and every image has the resolution of 1248 * 852 pixels and of 834Kb size. Our research investigation illustrates that the performance of our proposed work is better than the existing method. The accuracy of textual forged article identification and image forgery identification is above ninety percentages.

Future associated work would think about improving and testing the systems robustness by establishing a mixture of various forged categories of data like texture based images and raising the number of scanned images for execution by adding additional datasets. The Accuracy, Sensitivity and Specificity shall be included along with Euclidean Distance to estimate the distance among each pixel and setup its parameters. Grouping Image Processing technique and Machine Learning will be very efficient and the result attained will be accurate.

8. REFERENCES

[1] S. Teerakanok and T.Uehara,Copy-move forgery detection: A state of the art technical review and analysis, *IEEE Access* (2019), 40550–40568

[2] I. Noman, I. Zeeshan and Nazia, A Survey on Optical Character Recognition System, *J Information, Communication and Technology* (2016), 1–4.

[3] F. Pawel, and M. Andrzej, Stamps Detection and Classification Using Simple Features Ensemble, *Mathematical Problem in Engineering*, (2018), pp. 1–15.

[4] G.H. Luiz, S. Robert and S.O. Luiz, Characterizing and Evaluating Adversarial Examples for Offline Handwritten Signature Verification, *IEEE Transaction on Information Forensics and Security* (2019), 1–1.

[5] G. Nabil and A.M Awal, A New Descriptor for Pattern Matching: Application to Identity Document Verification, *IAPR Int. Workshop on Document Analysis System*, 2018, pp. 375–380.

[6] Z. Khan, F. Shafait and A. Mian, Hyperspectral Imaging forInk Mismatch Detection, *IAPR Int.Workshop on Document Analysis and Recognition*, 2013, pp. 877–891.

[7] Abramova, Svetlana and R. B¨ohme, Detecting Copy Move Forgeries in Scanned Text Documents, *Int. Conf.Media Watermarking, Security and Forensics*, (2016), pp. 1–9.

[8] M.J. Khan, A.Yousaf, K. Khurshid, A. Abbas and F. Shafait, Automated Forgery Detection in Multispectral Document Images Using Fuzzy Clustering, *IAPR Int.Work Document Analysis Systems*, (2018) pp. 393–398.

[9] B. Maryam, A. Hamid, M. Moetesum and I. Siddiqi, Document Forgery Detection using Printer Source Identification—A Text-Independent Approach, *Int Conf. Document Analysis and Recognition*, (2019), pp. 7–12.

[10] M.J Tsai, Y. Han and I Yuadi, Deep learning for printed document source identification, *Signal Processing: Image Communications* (2019), 184–198.

[11] T. Mahmood, T. Nawaz, A. Irtaza, R. Ashraf, M. Shah and M.T. Mahmood, Copy-Move Forgery Detection Technique for Forensic Analysis in Digital Images, *Mathematical Problems in Engineering*, (2016), pp. 1–13.