Proceedings of the 2nd International Conference on Trends in Electronics and Informatics (ICOEI 2018)
IEEE Conference Record: # 42666; IEEE Xplore ISBN:978-1-5386-3570-4

Framework For Image Forgery Detection And Classification Using Machine Learning

Shruti Ranjan, Prayati Garhwal, Anupama Bhan, Monika Arora, Anu Mehra

Department of Electronics and Communication, Amity University, Noida, Uttar Pradesh, India
shruti.ran9@gmail.com, prayatigarhwal@gmail.com, abhan@amity.edu, monika4dec@gmail.com, amehra@amity.edu

Abstract - In the recent times, the rates of cyber-crimes have been surging prodigiously. It has been proven incredibly easy to create fake documents with powerful photo editing soft-wares being as pervasive as ever. Documents can be scanned and forged within minutes with the help of these soft-wares that have tools readily available just to do that. While photo manipulation software is handy and ubiquitous, there are also means to deftly investigate these morphed documents. This paper lays a foundation on investigation of digitally manipulated documents and provides a solution to distinguish original document from a digitally morphed document. A Graphical User Interface (GUI) was created for detection of digitally tampered images. This method has accuracy of 96.4% and has proven to be efficient and handy.

Keywords: Artificial Neural Networks; GLCM features; Graphical User Interface; Machine Learning ; Support Vector Machine

I. INTRODUCTION

With the rapid technological advancement has strengthened the growth of every field imaginable, security being one of them, it has also become easy to breach it. Not only can legal documents be stolen and forged, criminal evidence- such as photographs and security footage can be easily tampered with. One may feel it is enough for an institution to check ID's at the front gate but they do not realize how menial of a task it is for a criminal to get their hands on fake ID's. Posing as someone else in a public setting is a trouble free task even for amateur criminals. As mentioned before, photo editing tools which on top being easily accessible are also extremely friendly. One can learn basic photo editing tips in a few hours, even if they have never seen an image editing software before. There is nothing too advanced about photo editing anymore, whereas forgery has become even more difficult to detect.

Image forgeries may be classified into many types- such as copy-move forgery, splicing and many more. Research has been going on in this field for years now and many effective methods have been proposed to detect such forgeries. Xudong Zhao et al. proposed a method for colour channel design to find the most inequitable channel, which they called the optimal chroma-like channel, for feature extraction [1]. Another process to detect counterfeited documents, mainly tampered with using a photocopier, is through superimposition [2]. However, such techniques have now become obsolete since forgery these days is digital, clean and indistinguishable to the human eye. Therefore, machines are a more viable option now. Most of the techniques used to detect those manipulations employ machine learning and pattern recognition [3]. Region duplication can be detected by calculating the scale invariant feature transform (SIFT) key-points and then finding all the pixels within the duplicated region [4]. Digital documents that have been rotated, scaled or resized can also be detected easily using image processing tools [5]. Research has been done so far to detect duplicated regions in a document tampered using copy-move forgery with the help of block-based and traditional key-point based methods [6]. Since all the databases in a security system are digital, people mostly rely on the image features that can be extracted easily. For instance, gradient based texture features, with the help of a machine, can easily be calculated and compared [7]. Another devised scheme is to divide the image into overlapping blocks, thinking of them as vectors and find the manipulated region through radix sorting [8]. Image forgery detection can also be done using only image processing and without any embedded security information. This method makes use of Fuzzy Transform (F-Transform) and Ring Projection Transform (RPT) to detect forgeries. These transforms convert the data to a single dimension significantly reducing the computational capacity [9]. Various studies have also been done weighing down the pros and cons of the prevalent copy-move forgery detection (CMFD) techniques [10].Image processing algorithms such as DWT (Discrete Wavelet Transform) and SVD (Singular Value Decomposition) are one of the many feature extraction methods that are used today to detect forged images [11]. Another approach to detect

tampered images is to make use block based methods, but by using the non-overlapping texture blocks as a base for the smooth blocks, thus reducing the computational capacity [12]. Copy-move forgery (CMF) can also be detected using algorithm based on Stationary Wavelet Transform (SWT), which is able to accurately detect the duplicated blocks [13]. CMF can also be detected easily if the feature vector generated is based on colour perception and object representation [14]. Reflective SIFT based algorithms are also proficient in detecting duplicated blocks in copy-move forgeries [15].

In this paper, another such method for digital forgery detection is proposed. This method uses a Graphical Interface (GUI), designed to detect whether an image has been morphed or not. The GUI is designed to load the images, do its pre-processing before determining whether it has been tampered with. This is a novel methodology that allows the user to simply load the image onto the interface. Then the image is pre-processed before being analysed. This includes global contrast enhancement. Then the image is partitioned into three segments using k-means clustering, which separates the data into three sections, each having a dissimilar set of data. Then the segment containing the most information (which can be determined by calculating the mean of the segment) is chosen for further analysis. Then the selected segment's GLCM features are calculated and cross-validated with those of the scan in the database to determine whether the image has been morphed or not. The accuracy of proposed method was initially calculated using linear SVM and then using ANN as the accuracy was less on implementation of SVM.

II. METHODLOGY

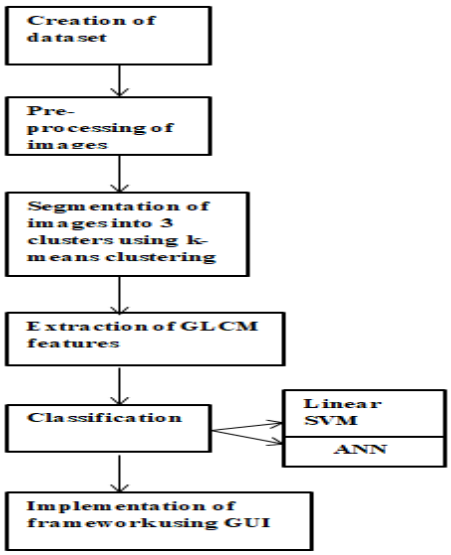

Fig 1: Flow chart showcasing methodology adopted

1. Creation of Dataset: The images used for this project were collected from various internet sources and morphed using photo editing tools. These images were edited using Adobe Photoshop CC 2017 to create a dataset with 120 pairs of images- one original and its edited version. These images (total 240 in count) were used in further analysis using MATLAB R2015a.

2. Pre- Processing of the Images: To make the details of the images stand out more, the query image was enhanced using *histogram equalization*. It is a necessary step because sometimes minute forgeries go undetected through the entire process. It is important that the machine gets most of the details in one go. Histogram equalization, as the name suggests, is a method, where the intensities are adjusted using the histogram of the image. This technique is used here for contrast enhancement.

Another essential stage in the pre- processing of an image is the removal of noise. De-noising is again done so that the details of the image are sharper and are not missed while extracting the features of the image. In this paper, de-noising is done using the median filter in MATLAB.

3. Segmentation: The image is segmented into 3 clustering by using *k-means* clustering. K-means clustering is a technique for quantizing vectors. This method divides the image into *k* segments, each containing mutually exclave data. This is a common method when it comes to pattern recognition and machine learning. One of the segmented images is chosen on the basis of the information contained in it. To determine this, the GLCM features of each segment are calculated and the segment with the highest mean is chosen. The GLCM of the segmented image are then compared with the original image using cross-validation, which gives another array, which is studied to determine whether an image is morphed or not, and function for the final result is added on the basis of that.

4. Extraction of GLCM Features: Out of all the methods to analyse an image, extraction of GLCM features has proven to be efficient time and time again. The gray level co-variance matrix is a tabulation that provides with statistical measures for texture analysis. This method takes into account the spatial relationship between the intensities of pixels in a gray-level image. In this paper, the GLCM features were calculated to study the differences in the original image and the digitally forged image. This gave 22 texture values (for each image) to work with, most of which were similar when it came to an image and its fraudulent counterpart. In practice, this would lead to

redundancy and would also increase the time to run the algorithm. Also, the histogram of oriented gradient (HOG) features was calculated which gave another set of features for the original and the morphed image. The HOG values of the original and the morphed images were reasonably apart from each other, which meant that these values will be useful in differentiating the original document from the morphed one. However, the order of matrix generated by HOG algorithm is too large to be successfully fed into an SVM so it could also not be of practical use.

5. Classification: Initially, the classifier used for classification of dataset into two parts as original or morphed was linear kernel SVM. A linear kernel SVM is the most suitable classifier for two-class classification problems. It finds an equivalent hyper-plane which separates the whole data by a specific criterion that depends on the algorithm applied. It tries to find out a hyper-plane which is far from the closest samples on the other side of the hyper plane while still classifying samples. It gives the best generalization techniques because of the larger margin. The accuracy obtained through linear SVM was low in this case (87.6%). So, Artificial Neural Network (ANN) classifier was applied on the dataset. ANN networks are basically a system of interconnected neuron like layers. The interconnection of the network can be adjusted based on the number of available inputs and outputs making it ideal for a supervised learning. Hidden layers were chosen as 5 for our dataset. The ANN model was trained by providing 220 images (contains both original and morphed documents). The ANN model itself separated the training, validation and the testing sets into 70%, 15% and 15% of the total dataset respectively. It gave a higher accuracy of 96.4% as compared to linear SVM.

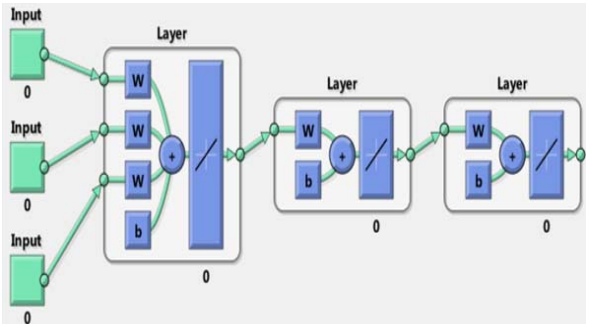

Fig 2: Block diagram of Artificial Neural Network

6. Creation of GUI: Designing of a Graphical User Interface (GUI) was deemed necessary because one had to repeatedly check whether an image had been tampered or not. To do this, the drag-and-drop GUIDE Layout Interface in MATLAB was used. Once, the front-end design was complete, a modified the back-end code of the interface was coded, which allowed to program the functions into the push buttons to give the required results. For example, the first push button was used to load the image onto the GUI; therefore, few lines were added to the code which allowed us to upload an image at the front-end.

III. EXPERIMENTAL RESULTS

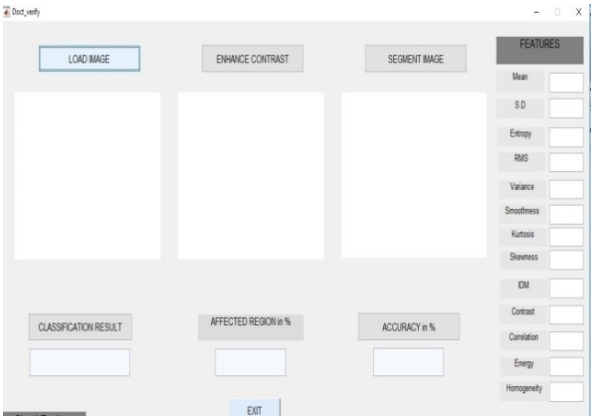

Fig 3: GUI design layout

Fig 4: Enhanced contrast of the query image

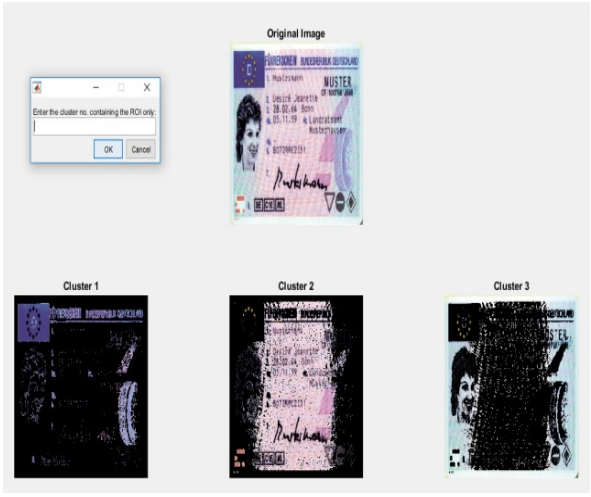

Fig 5: Image segmented into 3 clusters

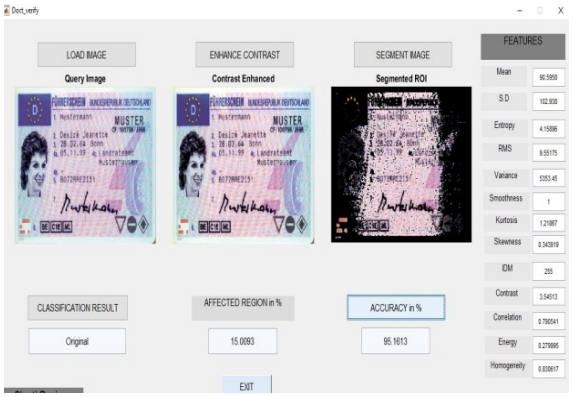

Fig 6: Selected segmented image

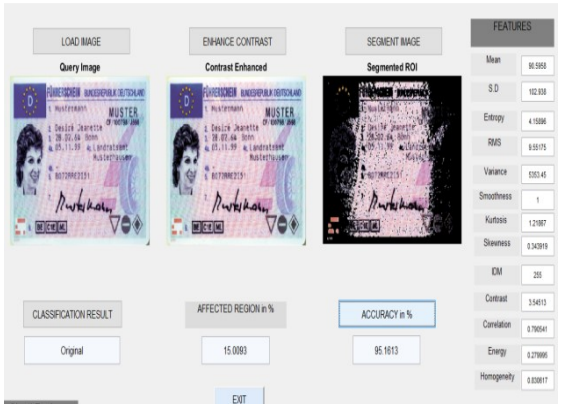

Fig 7: Result being showed as "original"

Fig 8: Result being showed as "morphed"

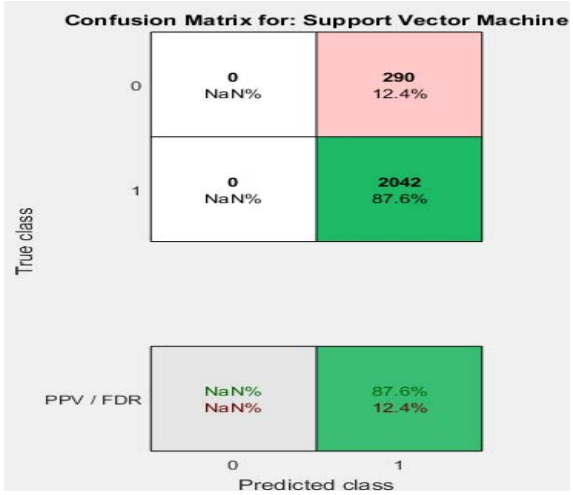

Fig 9: Confusion plot by Linear SVM

Since, the accuracy percentage of the Linear SVM was less (87.6%) thus, Artificial Neural Network (ANN) was applied on the dataset which gave higher accuracy than the linear SVM of 96.4%.

Fig 10: Confusion plot by ANN

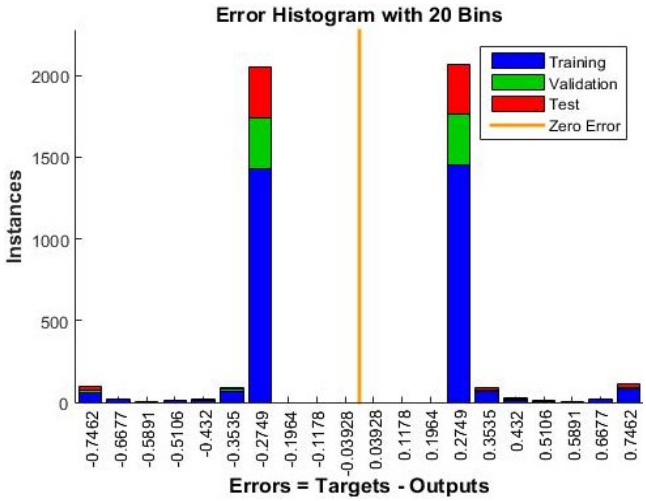

Fig 11: Error histogram by ANN

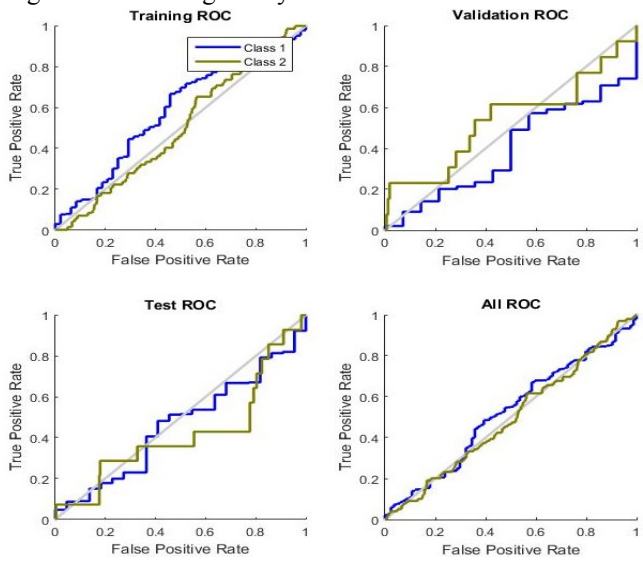

Fig 12: ROC plot by ANN

TABLE I: Classifiers used and their results

PARAMETERS	ANN	SVM
Accuracy	96.4%	87.6%
Sensitivity	97.0%	89.0%
Specificity	95.0%	86.0%

IV. RESULTS AND DISCUSSION

The approach discussed in this paper has been implemented successfully. The main objective of the project was to examine the computer manipulated legal documents for forensic testing using image processing techniques which have been implemented. The dataset required for the project was created adequately by manipulating the acquired images so that no naked human eye could identify the manipulations in the images. Thereafter, feature extraction by GLCM implemented in MATLAB Image Processing Toolbox R2015a also gave the required results for examination and comparison between the original and morphed legal documents. Also, the HOG values were computed but since the order of the matrix produced were very large to be trained in the SVM so it was excluded. Usage of SVM one to one classifier was to classify the difference between morphed and original legal documents on the basis of different categories would have resulted in more appropriated distinctive properties between these two documents. Also, by other mathematical expressions would have resulted in more refined results to distinguish morphed documents more easily. The use of ANN classifier gave a higher accuracy of 96.4% as compared to linear SVM which gave less accuracy.

V. CONCLUSION AND FUTURE SCOPE

Starting from acquisition of images for the creation of dataset and, then morphing it and later on scientifically examining the obtained results after application of transformation techniques on the morphed and original image and finally acquiring the texture features of the same was executed the same way it was planned initially. In the present time, with the advancement in the field of science and technology, the introduction of various advance images editing tools are also surging up. These advanced image editing tools have multitudinous features. We can use these advanced image editing tools in our further extension of the project to implement the required results more easily and instantly. While these tools are mostly used in the creative design related areas, criminals also can easily get access to them and as a result, can exploit them to create fake identities to hide themselves in public, or to commit a crime. Research has been going on for the past few decades to come up with a fool-proof method to detect these forged documents which do not look any different to the human eye. Most of the forgery detection methods rely on feature extraction and texture analysis of the scanned document, and the detection program is created through pattern recognition and machine learning. Our purpose was to propose one such method with good efficiency and accuracy. We will continue to refine the methodology so that there are lesser loop-holes in the analysis and will hopefully come up with a better method in future.

VI. REFERENCES

[1] Zhao, X., Li, S., Wang, S., Li, J., & Yang, K. (2012). Optimal chroma-like channel design for passive colour image splicing detection. EURASIP Journal on Advances in Signal Processing, 2012(1), 240.

[2] Joshi MC, Kumar A, Thakur S. Examination of digitally manipulated-machine generated document, a case study elucidating the issue of such unwanted progenies of modern technology. Prob Forensic Science 2011;56:162–73.

[3] Qureshi, Muhammad Ali, and Mohamed Deriche. A bibliography of pixel-based blind image forgery detection techniques, Signal Processing: Image Communication 39 (2015): 46-74.

[4] Xunyu Pan, Siwei Lyu, "Region Duplication Detection Using Image Feature Matching", Information Forensics and Security IEEE Transactions on, vol. 5, pp. 857-867, 2010, ISSN 1556-6013.

[5] Anil Dada Warbhe, Rajiv V. Dharaskar, Vilas M. Thakare, "Digital image forensics: An affine transform robust copy-paste tampering detection", Intelligent Systems and Control (ISCO) 2016 10th International Conference on, pp. 1-5, 2016.

[6] Mohsen Zandi, Ahmad Mahmoudi- Aznaveh, Alireza Talebpour, "Iterative Copy-Move Forgery Detection Based on a New Interest Point Detector", Information Forensics and Security IEEE Transactions on, vol. 11, pp. 2499-2512, 2016, ISSN 1556-6013.

[7] Xia, Z., Lv, R., Zhu, Y., Ji, P., Sun, H., & Shi, Y. Q. (2017). Fingerprint liveness detection using gradient-based texture features. Signal, Image and Video Processing, 11(2), 381-388.

[8] Lin, Hwei-Jen & Wang, Chun-Wei & Kao, Yang-Ta. (2009). Fast copy-move forgery detection WSEAS Transactions on Signal Processing. 5. 188-197.

[9] Ansari, Mohd Dilshad & Prakash Ghrera, Satya. (2018). Copy-move image forgery detection using direct fuzzy transform and ring projection. International Journal of Signal and Imaging Systems Engineering. 11. 44. 10.1504/IJSISE.2018.10011742.

[10] Badal Soni, Pradip K. Das, Dalton Meitei Thounaojam. (2018) CMFD: a detailed review of block based and key feature based techniques in image copy-move forgery detection. IET Image Processing 12:2, pages 167-178.

[11] Jawadul H. Bappy, Amit K. Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, B.S. Manjunath. Exploiting Spatial Structure for Localizing Manipulated Image Regions. (2017) IEEE International Conference on Computer Vision (ICCV), pages 4980-4989.

[12] Hajihashemi, Vahid & Gharabagh, Abdorreza. (2018). A Fast, Block Based, Copy-Move Forgery Detection Approach Using Image Gradient and Modified K-Means. 298-307. 10.1007/978-3-319-68385-0_25.

[13] Mahmood, Toqeer & Nawaz Tabassam & Mehmood, Zahid & Khan, Zakir & Shah, Mohsin & Ashraf, Rehan. (2016). Forensic analysis of copy-move forgery in digital images using the stationary wavelets.578-58310.1109/INTECH.2016.7845040.

[14] Kushol, Rafsanjany & Salekin, Md Sirajus & Hasanul Kabir, Md & Alam Khan, Ashraful. (2016). Copy-Move Forgery Detection Using Colour Space and Moment Invariants-Based Features.1-6. 10.1109/DICTA.2016.7797027.
[15] Agarwal, Vanita & Mane, Vanita. (2016). Reflective SIFT for improving the detection of copy-move image forgery.84-88. 10.1109/ICRCICN.2016.7813636