EasyChair Preprint

№ 3731

A Simple Approach Towards Forgery Detection

Khawaja Muhammad Ali, Muhammad Shahzaib and Rida Nasir

July 3, 2020

A Simple Approach Towards Forgery Detection

Khawaja Muhammad Ali
Department of Electrical Engineering
Institute of Space Technology
Islamabad, Pakistan
khawaja07@ist.edu.pk

Muhammad Shazaib
Department of Electrical Engineering
Institute of Space Technology
Islamabad, Pakistan
mshahzaib07@gmail.com

Rida Nasir
Department of Electrical Engineering
Institute of Space Technology
Islamabad, Pakistan
ridanasir5@hotmail.com

Abstract

Forgery detection and investigation has been a relevant topic of interest for human beings since ages. Important messages written and transported by kings in old ages were sealed with signatures and stamps to achieve this purpose. But with the advent of digital technology, forgery detection has become even more important since tools for forgery have become vast as well. In this paper a technique based on pixel clustering has been introduced for detection of modification, alteration or forgery done with a different ink color pen. Hyperspectral images are used for ink mismatch detection in a handwritten note. We propose ink classification based on pixel intensities values present in all the bands of hyperspectral images of the handwritten note. Our proposed technique is quite simple yet effective in detecting ink mismatch with relatively high accuracy.

Keywords—Ink Mismatch, clustering, hyperspectral images

I. INTRODUCTION

Human visual system is limited as it is based on trichromatic nature for distinguishing between different colors and shape. As a result, it can be manipulated into believing that two very similar shades of a color are the same [1]. Forgery is done with the intention to deceive human vision. The forger not only tries to emulate the handwriting of the original writer, but also uses a pen that has a visually similar ink compared to the rest of the note. Hence, analysis of inks is of critical importance in questioned document examination. However, with the use of advance technology and image processing techniques available today, ink analysis can be performed that can help us in identifying forgery, fraud, backdating and ink age.

Hyperspectral imaging (HSI) has proved to be very effective in ink mismatch detection in the recent past and has therefore, become a topic of interest among researchers over the past decade. Not only ink mismatch detection but also to enhanced the visible spectrum of text-based documents that were corrupted with artefacts such as, corrosion, ink-bleeding and foxing etc [2].

Hyperspectral advantage lies in its nondestructive nature tool for detection and identification of forensic trace. A HSI analysis based forgery detection system was proposed that used well know, widely used k-means clustering algorithm with optimum number of HSI bands for classifying the questioned document contained different inks [3]. However, one big drawback of their proposed system that restrict for employing on real time scenarios of forgery documents was that their system assumes the questioned document to be composed in equal ratios of two types of inks. Brauns and Dyer developed a hyperspectral imaging system for forgery detection [4]. Padaon et al. based their ink mismatch detection using heat, along with narrowband tunable light source [5]. However, these solutions though effective but expensive and computationally complex.

Newer research techniques include state of the art classification techniques (such as machine learning) to identify the questioned documents. An accuracy of 98.2% and 88% have been achieved at a slightly simple convolutional neural network (CNN) with only 13 layers for blue and back color of inks respectively [6]. They had used five different brands of inks mixed in different proportions for testing. They achieved high accuracies at low resolution images. Authors in [7] used hybrid deep convolutional neutral network (CNN) which uses spatial information of data along with spectral features. Different brands of same Inks color were mixed with pairs of 2, 3, 4 and 5 in different proportions for testing. They achieved a maximum accuracy of 99.6% and 92.3% for blue and black inks respectively. Although they haven't done comparison with standard available CNNs. As seen by accuracy only blue color has produced maximum accuracy. Reference [8], the author further extends the research by including more number of combinations of ink mixing (up to ten) and increase the number of subjects (writers) to seven as well. Classification was done using five different deep CNN networks. However, their accuracy was drop down to 71.28% due to increase number of subjects and ink mixing ratios.

Reference [9] used clustering based on fuzzy C-mean, which is similar to k-mean clustering. They had used 14 images of HSI dataset each with 33 bands. They employed local thresholding Sauvola which solved non-uniform illumination problem in the HSI images. Merging of inks was used with ten different combinations in different ratios. They achieved an accuracy of 95% for black ink and around 99% for blue ink. The mixing of inks cased sudden decrease in accuracy for black inks as compared to blue inks with ink mixing ratios of 1:3 and 1:7. However, while incorporating the feature selection method the accuracies of both inks with different ratios tend to increase. Influenced Outlierness (INFLO) algorithm with focus on inks spectral responses instead of ink deposit traces or differentiating inks based on texture is

presented in [10]. The authors used point-to-point (P2P) distances criteria for local region and apply k-mean algorithm to find the average value of minimum distances between each data point of two clusters. Then applying the INFLO algorithm which is used when the cluster to be separated are in very close proximity. They compared the results with other algorithms as well which includes Local Outlier Factor (LOF) and Connectivity-based Outlier Factor (COF). They achieved maximum accuracy around 98% for blue ink and 99% for black ink with INFLO algorithm. However, with the mixing of different inks, black ink has greater accuracy than blue ink which is an opposite case while classification with deep CNN networks. Reference [11] proposed the use of different heuristic models for ink mismatch detection. Reference [12] and [13]worked on Sauvola and INFLO algorithms respectively.

In our paper we have proposed a simple and computationally light technique for identification of ink mismatch. It is based on the concept of K-means clustering.

The rest of the paper is organized in following sequence, Section II describes the Methodology, section III contains Algorithm followed by Result and Discussion in section IV and Conclusion and Future work in section V.

II. Methodology

A. Hyperspectral Image Dataset

The dataset used is available for download[*1]. It contains 33 bands each with a size of 81x627. The dataset was tempered with more than one inks of different brands.

B. Extraction of Text (foreground)

In first step, the text in the image is separated from the background using a global threshold value. Reason for using global threshold is that the text is clean from noise as by visual inspection so no need of adaptive thresholding. The minimum and maximum value of text in the image (other than background i.e 0) is 29 and 68 respectively. The binary threshold image is shown in Fig.1 with yellow background.

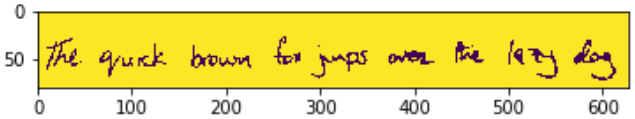


Fig.1 Threshold image band 1

So, a global threshold value was set to 20. We get 3190 pixels containing text out of total 50787 pixels. The spectral responses of these text values from all 33 bands were plotted and shown in Fig. 2. As the spread of this graph along y-axis shows some variation. Because if this image contains text with only single ink, the signature of every pixel would be same and overlapped each other. However, this suggests more than one inks.

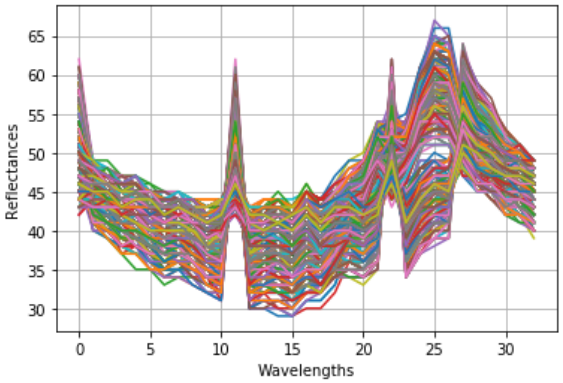


Fig. 2 Spectral responses of foreground

C. Clustering

For real time case scenarios of forgery detection in documents, it is preferable to have some algorithm that done unsupervised classification. In this study we choose K-mean clustering which is an unsupervised learning algorithm. It is widely used algorithm in ink mis match detection. But it has some drawback, while choosing the clusters (or groups) we must manually set the number of clusters. To cope with this problem, we start with three clusters and increase up to five and by using the elbow method shown in Fig. 3 we can find the optimal value of clusters. This figure indicates that there is no significant change in distortion/variance after 3 cluster value. It should be noted that the shown elbow diagram has been off set from K=2 to observe the clusters of text (as K=1 is obvious for background). However, this technique would give poor results if two clusters are in very close proximity.

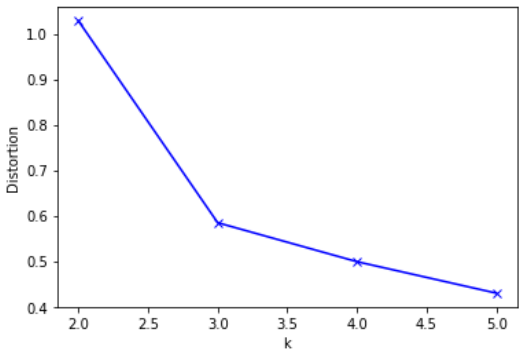


Fig. 3 Elbow Method showing optimum value of K=3

III. Algorithm

Our proposed algorithm is based on feature extraction based on pixels. Each pixel present in all the bands of the subject hyperspectral image serves as a unique feature. Each image has a size of 81×627 pixels. The whole hyperspectral image has 33 bands; thus, it has a size of 81×627×33 pixels. So, there are 1,675,971 features used in our approach.

*1 https://drive.google.com/file/d/1BIAnV7HFz7bOjIkyh22d63nQrweggj9E/view?usp=sharing

The programming language of our choice is Python owning to the availability of wide range of libraries and IDEs. Our IDE of choice is Spyder. We have used Python 3.7 on Intel(R) Core(TM) m3-7Y30 CPU @ 1.00 GHz 1.61 GHz with 8.00 GB RAM

We have coded our algorithm in Python efficiently using only a few libraries: Numpy, Panda, CV2, PIL, Sklearn.

A. *Steps*

Our proposed algorithm follows the following steps:

A. Read the images in a numpy array A of size 81×627×33

B. Resize the A into a matrix of size 50787×33. Each image is converted into a vector containing 50787 entries.

C. Select an initial value of K = 3 (number of clusters to be employed, assuming 2 clusters with 1 background) and make an elbow graph.

D. Apply simple K-mean clustering on B. For this purpose python inbuilt KMeans of Sklearn library is used. As a result, a 50787×1 vector of labels is produced

E. Convert the labels vector obtained in D, into an RGB image and plot.

F. Change the value of K to 4 and 5 separately and repeat from Steps D and E

The obtained images are compared and studied. It should be noted that one cluster belongs to the pixels present in the background. Thus, K=3 refer to two different ink pixels along with background pixels.

IV. Results And Discussion

We applied our proposed algorithm with three different values of K (clustering parameter). These are show in Fig. 4, Fig. 5 and Fig. 6 respectively.

Fig. 4 Output with K=3 clusters

Fig. 5 Output with K=4 clusters

Fig. 6 Output with K=5 clusters

In case of K=3, most optimum results are produced. Both inks used are clearly visible in green and red. Black portion represents the background.

In case of K=4 and K=5, It visible that the inks have been overlapped on each other which is unlikely in case of forgery. It is therefore concluded that the questioned images of handwritten notes include two different inks used for writing. Also, we can conclude that by observing Fig. 4 that the inks are mixed in fixed proportion.

V. Conclusion and Future Work

Our proposed algorithm based on clustering is simple yet quite effective in detection of ink mismatch in handwritten notes. We have used feature selection method to increase the accuracy of the algorithm. It is also important to note that our algorithm is based on data set containing balanced ink ratios. In case of unbalanced inks, CNN networks or other clustering techniques may be employed.

In addition to that, our work can further be extended with the use of discrete wavelet transforms and principle component analysis for identification of the bands of the hyperspectral image containing maximum relevant information. This will significantly reduce the features (number of pixels in this case) used for detection and make our algorithm more efficient and computationally less expensive. Furthermore, it is suggested that different combination of band along with transformations like HSV (hue, saturation, value) or HSL (hue, saturation, lightness) may be used to further distinguish the hyperspectral image from an RGB image. This transformation function would be of great impact as it will help to increase the overall accuracy for both blue and black inks.

VI. References

[1] E. H. Land and J. J. McCann, “Lightness and Retinex Theory,” *J. Opt. Soc. Am.*, vol. 61, no. 1, p. 1, Jan. 1971, doi: 10.1364/JOSA.61.000001.

[2] S. Joo Kim, F. Deng, and M. S. Brown, “Visual enhancement of old documents with hyperspectral imaging,” *Pattern Recognit.*, vol. 44, no. 7, pp. 1461–1469, Jul. 2011, doi: 10.1016/j.patcog.2010.12.019.

[3] Z. Khan, F. Shafait, and A. Mian, “Hyperspectral Imaging for Ink Mismatch Detection,” in *2013 12th International Conference on Document Analysis and Recognition*, Washington, DC, USA, Aug. 2013, pp. 877–881, doi: 10.1109/ICDAR.2013.179.

[4] K. Franke and S. Rose, “Ink-deposition model: The relation of writing and ink deposition processes,” *IEEE Workshop on Frontiers in Handwriting Recognition*, pp. 173–178, 2004.

[5] N. Otsu, “A threshold selection method from gray-level histograms,” vol. 11, pp. 23–27, 1975.

[6] M. J. Khan, “Deep learning for automated forgery detection in hyperspectral document images,” *J.*

*1 https://drive.google.com/file/d/1BIAnV7HFz7bOjIkyh22d63nQrweggj9E/view?usp=sharing

Electron. Imaging, vol. 27, no. 05, p. 1, Sep. 2018, doi: 10.1117/1.JEI.27.5.053001.

[7] M. J. Khan, K. Khurshid, and F. Shafait, "A Spatio-Spectral Hybrid Convolutional Architecture for Hyperspectral Document Authentication," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia, Sep. 2019, pp. 1097–1102, doi: 10.1109/ICDAR.2019.00178.

[8] A. ul I. 8., Muhammad Jaleed Khan, Khurram Khurshid, and Faisal Shafait, "Hyperspectral Image Analysis for Writer Identification using Deep Learning," presented at the DICTA, Australia, Dec. 2019.

[9] M. J. Khan, A. Yousaf, K. Khurshid, A. Abbas, and F. Shafait, "Automated Forgery Detection in Multispectral Document Images Using Fuzzy Clustering," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, Vienna, Apr. 2018, pp. 393–398, doi: 10.1109/DAS.2018.26.

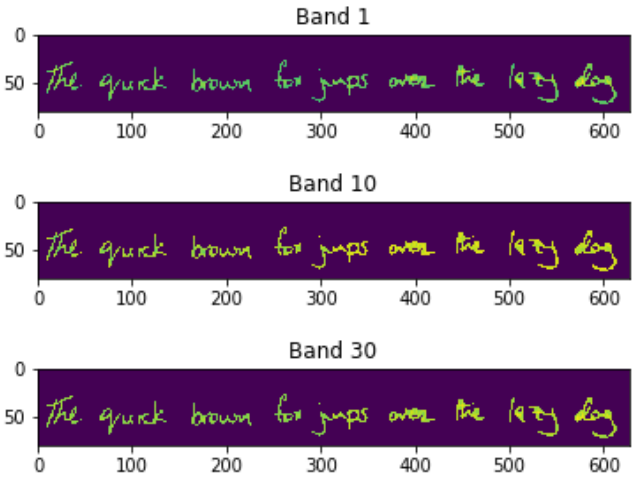
[10] Z. Luo, F. Shafait, and A. Mian, "Localized forgery detection in hyperspectral document images," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, Aug. 2015, pp. 496–500, doi: 10.1109/ICDAR.2015.7333811.

[11] A. Abbas, K. Khurshid, and F. Shafait, "Towards Automated Ink Mismatch Detection in Hyperspectral Document Images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, Nov. 2017, pp. 1229–1236, doi: 10.1109/ICDAR.2017.203.

[12] F. Shafait, D. Keysers, and T. M. Breuel, "Efficient implementation of local adaptive thresholding techniques using integral images," San Jose, CA, Jan. 2008, pp. 681510-681510–6, doi: 10.1117/12.767755.

[13] Jin, W., Tung, Han, and Wang, W, "Ranking outliers using symmetric neighborhood relationship," Berlin, Heidelberg, Apr. 2006, pp. 577–593.

APPEXDIX A (FIGURES)



*1 https://drive.google.com/file/d/1BIAnV7HFz7bOjIkyh22d63nQrweggj9E/view?usp=sharing