Kmeans Clustering Based Ink Mismatch Detection

SUBMISSION DATE / POSTED DATE

28-06-2020 / 29-06-2020

CITATION

ALI, KHAWAJA Muhammad; Shazaib, Muhammad; Nasir, Rida (2020): Kmeans Clustering Based Ink Mismatch Detection. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.12580295.v1

DOI

10.36227/techrxiv.12580295.v1

Kmeans Clustering Based Ink Mismatch 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 investigation and detection 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 shapes. As a result, it can be manipulated into believing that two very similar shades of a color are the same. 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 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. Its advantage lies in its nondestructive nature tool for detection and identification of forensic traces [1]. Brauns and Dyer [4] developed a hyperspectral imaging system for forgery detection. Padaon et al. [5]. based their ink mismatch detection on the use of heat along with narrowband tunable light source resource. However, these solutions though effective are expensive and computationally complex. 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.

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.

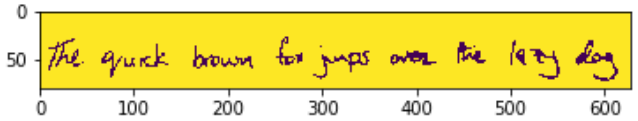


Fig.1 Threshold image band 1

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. 1. 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.

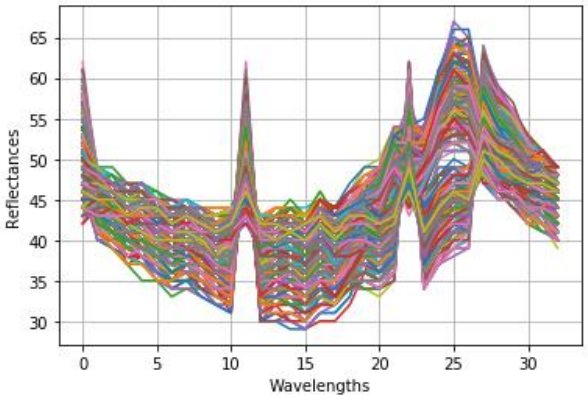


Fig. 2 Spectral responses of foreground

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 Kmean

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

clustering which is an unsupervised learning algorithm. It is widely used algorithm in ink mis match detection [1]. 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).

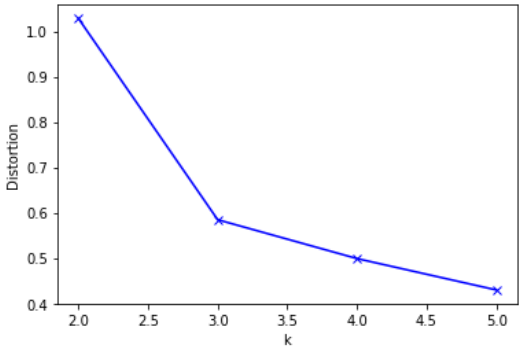


Fig. 3 Elbow Method showing optimum value of K=3

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 serve 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.

The programming language of our choice is Python due to its wide rage 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 with efficiency using only these libraries: Numpy, Panda, CV2, PIL, Sklearn []….

A. *Steps*

Our proposed algorithm follows the following steps:

A. Read the images in a numpy array of size of 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)

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 studied. It should be noted that one cluster belongs to the pixels present in the background. Thus, K=3 refer two different inks and one 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 ink have been overlapped on each other which is unlikely in case of forgery. It is therefore concluded that the subject images of handwritten notes include two different inks used for writing.

Conclusion

Our proposed algorithm based on clustering is simple yet quite effective in detection of ink mismatch in handwritten notes. Our work can further be extended by the use of discrete wavelet transforms and principle component analysis for identifying the bands of the hyperspectral image which contain maximum 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.

References

[1] Khan, Zohaib & Shafait, Faisal & Mian, Ajmal. (2013). Hyperspectral Imaging for Ink Mismatch Detection. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. 877-881. 10.1109/ICDAR.2013.179.

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

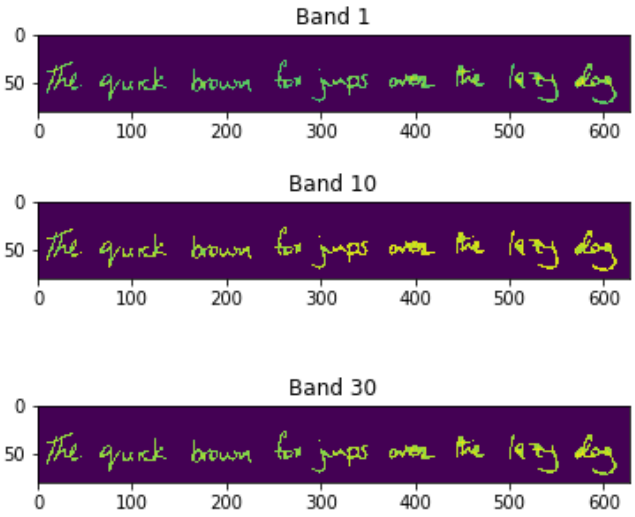
[3] Muhammad Jaleed Khan, Adeel Yousaf, Asad Abbas, and Khurram Khurshid "Deep learning for automated forgery detection in hyperspectral document images," Journal of Electronic Imaging 27(5), 053001(4 September 2018). https://doi.org/10.1117/1.JEI.27.5.053001

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

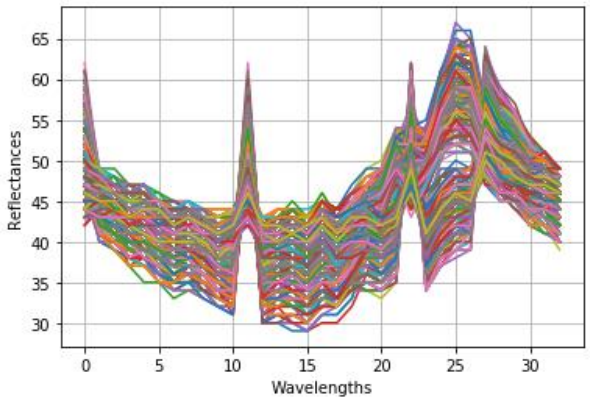
[5] N. Otsu, “A threshold selection method from gray-level histograms,” Automatica, vol. 11, no. 285-296, pp. 23–27, 1975.

APPEXDIX A (FIGURES)

Part 1(a)



Part 1(b)



Part 1(c)

Elbow Plot with K=5