Automatic Ink Mismatch Detection in Hyper spectral Images Using K-means Clustering

Noman Raza Shah, Muhammad Talha, Fizza Imtiaz Aneeqah Azmat
190412008, 190412005, 190412009, 190411002
Department of Electrical Engineering, Institute of Space Technology, Islamabad 44000, Pakistan
noman14@ist.edu.pk, m.talha14@ist.edu.pk, fizza14@ist.edu.pk, aneeqah15@ist.edu.pk

***Abstract*— Hyper spectral imaging (HSI) is a technique that is used to obtain the spectrum for each pixel in the image. It helps in finding objects and identifying materials etc. Such an identification is very difficult using other imaging techniques. It allows the researchers to investigate the documents without any physical contact. Nowadays detection of unequal Ink mismatch based on HSI has shown vast improvement in distinguishing the inks. Detection of unequal Ink mismatch is an unbalanced clustering problem. This paper used K-means Clustering for ink mismatch detection. K-means Clustering find same subgroups in the data based on Euclidean distance. This paper demonstrates performance in unequal Ink mismatch based on HSI.**

***Keywords*- Hyper spectral Document Images, Non-destructive Analysis, Forensics Document, Ink Mismatch Detection, K-means Clustering**

The code and images can be downloaded from the following link: Code and Dataset

I. INTRODUCTION

One material can be differentiated from any other material by a unique spectral signature. The human eye can easily distinguish various colors. It is difficult to distinguish between two similar coloured inks as they lie close together in the visible spectrum [1][2]. However these unique inks have different spectral signatures. This spectral property is utilized to detect a forgery in document images as it tends to be recognized whether a document is unique or manipulated by applying automatic ink mismatch techniques.

The importance of inks analysis is to address significant issues about document images. Hyper spectral images consist of various spectral bands which are useful for automatic ink mismatch detection. It will also be helpful in forgery detection. On the basis of spectral signature inks can be used to accomplish many facts like forgery, ink aging and fraudulent document [3]. It is based on the assumption that how the forgery has been done with different ink or pen. Ink mismatch also plays a vital role in cheque verification in banks, degree testing in universities and different important papers of government offices as well. From past many years researchers have paid attention to propose different techniques for detection of ink mismatch using HSI and multi spectral images.

The ink mismatch technique has typically two methods i.e. destructive and non-destructive analysis [2][3]. The destructive analysis such as thin-layer chromatography [4] is a chemical solution-based analysis that has been used for the detection of ink mismatch by forensic documents experts. The destructive method has several drawbacks that includes failure to retrieve damage to the document. It is time consuming as it needs a large amount of measurement to be taken [3]. On the other hand, HSI is an efficient tool for non-destructive and non-contact examination of forensic documents to overcome such limitations [3][5]. HSI is a technique that combines spectroscopy and imaging. In this each image is acquired at a narrow band of the electromagnetic spectrum to capture detailed spectral data. Thus, HSI reveals the unseen details in an image without getting in direct contact with it. Hence non-destructive method is preferred .

Easton et al. [6] presented one of the primary works in multi spectral imaging. Spectral imaging framework i.e. Eureka Vision was established by Christens-Barry et al. [7] Such frameworks are helpful for forensic experts in ink investigation. Ink investigation using a band-by-band assessment of multi spectral images through visual evaluation is tedious. It requires to be physically seen by the analyst under each frequency of light. An advanced and complex HSI system for assessment of forensic document was made by the National Archives of the Netherlands [8]. It gave high spatial and high spectral resolution images which were captured from spectral range (from close to UV to IR). But such forensic documents require near fifteen minutes of exposure to be captured. This HSI system was very powerful but long acquisition time restricts the use of such a system [9]. Hedjam et al. [21] proposed a mathematical model for improving the meaningfulness of very crumbled text. A few techniques for ink mismatch identification dependent on HSI examination have been proposed in the most recent decade [19]. These techniques incorporate fuzzy c-means clustering [3], k-means clustering [5], localized hyper spectral image analysis [16], and deep convolutional network [22].

Abbas et al. [2] used HSI un mixing scheme for ink mismatch detection. It first emply Hyper spectral subspace identification by minimum error algorithm (HySime) [11] for dimensionality reduction .It approximates the number of signatures present in the HSI documents. Secondly, Minimum volume enclosing simplex (MVES) [12] algorithm is employed on a reduced dimensional data. It extracts the hidden end members and corresponding abundances from its hyper spectral observations. This paper used K-means Clustering technique for automatic ink mismatch detection.K-means is

a used as a partitional clustering algorithm. It divides the total number of samples into k different group with a condition that groups cannot be greater than number of samples. This employ unsupervised learning technique which find same subgroups in the data and then group the similar data in one cluster on the basis of its similarity, intensity or other features.

II. RELATED WORK

Hyper spectral images consist 100's of bands and each band is rich in information rich hence are useful than multi spectral images. The recent literature shows a high potential of HSI and Spatio-spectral features for document analysis and forensics.

In this context, Morales et al. [11] proposed an approach for ink analysis in pen verification and handwritten documents using Least Square Support Vector Machine (SVM) classification. Silva et al. [14] developed a non-destructive method to detect fraud in documents based on different chemo metric techniques. Khan et al. [15] proposed a joint sparse band selection based hyper spectral imaging document analysis technique to distinguish different metameric inks. Abbas et al. [2] proposed hyper spectral unmixing for ink mismatch detection. Our main focus is to distinguish visually similar inks which are mixed in varying proportions to form an unbalance clustering problem.

Jaleed et al. [3] proposed an efficient automatic ink mismatch detection technique using multi spectral image analysis. Ink pixels are segmented using local thresholding and Fuzzy C-Means Clustering (FCM) . Luo et al. [16] proposed a system for localized forgery detection using anomaly detection algorithm combined with unsupervised learning to handle the cases where the pixels belonging to different classes are highly unbalanced. Aythami et al. [20] proposed a system to detect forgeries in hand written documents particularly for bank cheques based on ink discrimination.

Braun et al. [17] proposed Fourier transform based HSI system to detect forgery. They used fuzzy clustering to group the similar ink spectra. The experiments show that inks can be qualitatively segmented into two clusters. One limitation of the system was the lack of quantitative information and slow imaging process. A visual comparison of black inks done by Hammond et al. [13] using multi spectral document imaging. George et al. [18] used HSI for visual enhancement of documents by separating the text written using two different inks in two directions. Khurshid et al. [19] proposed a CNN based ink mismatch detection method for HSDIs that employs a combination of spectral and spatial features of ink pixels for classification.

III. DATASET

The dataset consists of a hyper spectral cube which has a spatial resolution of 627 by 81 pixels with total 33 bands. Each band represent a gray scale handwritten image. The HSI is taken in the spectral range from 400 nm to 720 nm with a step size of 10 nm which results in 33 bands. A hyper spectral image consists of a phrase "The quick brown fox jumps over the lazy dog" written with either black ink pen or blue ink pen. Furthermore, each pen originated from various brand to ensure that they include varieties inside their ink regardless of whether they have externally same color. Fig.1. shows the 33 bands as 33 gray scale images having handwritten phrase "The quick brown fox jumps over the lazy dog" .

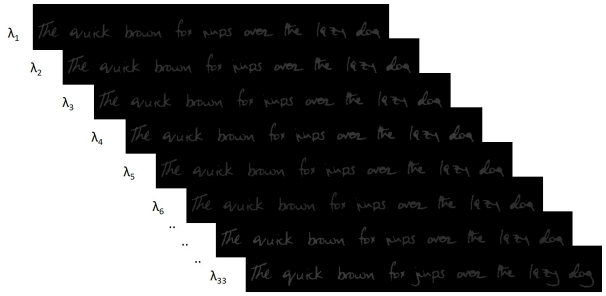

Fig. 1. Representation of 33 bands as gray scale images. λ_1,λ_2,...,λ_{33} shows the respective band number.

IV. METHODOLOGY

Clustering algorithm is an unsupervised algorithm that is used to partition the data set into clusters (subsets). K-means clustering is one of the most useful unsupervised algorithms. It is simple and less computationally expensive. It clusters the data into k clusters where k is the number of clusters which has to be assigned initially. As it is unsupervised algorithm it is used when un-labelled data is available which has not ground truth.

The clustering algorithm minimizes the squared error between a cluster centroid and its members. This implies that the k-means algorithm tries to optimize the objective function shown in equation 1. In each iteration has to result in better solution, the algorithm always converge. The number of clusters varies with the number of mixed inks. Previous work include the assumption that there are two inks in the image. An implication of this assumption could be that an image with more than two inks could still be grouped into two clusters. Selecting appropriate number of clusters play a vital part in correct segmentation.

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} ||x_n - \mu_k||^2 \tag{1}$$

The proposed methodology is illustrate in Fig. 2. First, we have images in form of gray scale so, we stacked the images and create a hyperspectral cube. In the next step we apply K-means algorithm on hyperspectral cube. By using the labels and cluster centres, we generate a color image which can easily classify text written with different inks.

V. EXPERIMENTS AND RESULTS

A. Visualization of Bands

The data consists of a hyper spectral cube which is of size 81x627x33 as 33 PNG images in place of the 33 bands of a hyper spectral image in gray scale form. The document

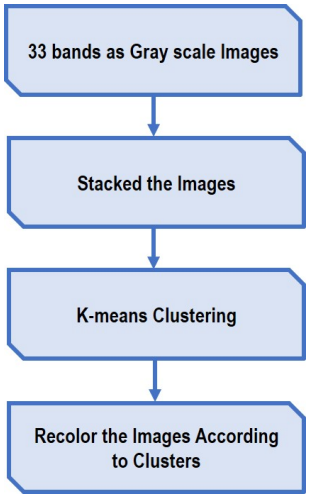

Fig. 2. Proposed Methodology

contains some handwritten text with one or more pens of different brands. First of all 1st , 10th and 30th bands of the hyper spectral image is displayed as gray scale image which is shown in Fig. 3.

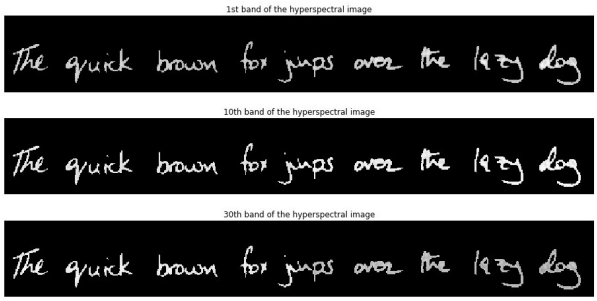

Fig. 3. 1st , 10th and 30th bands of the hyper spectral image

B. Spectral Signature

Signatures are one of the most widely recognized method to verify a record. Financial organizations use signatures for checking individual identity in financial and regulatory exchanges. The utilization of signature as a validating source in everyday life legitimizes the need of a complete authentication system. Spectral signature is the variation of reflectance of a material with respect to wavelengths (i.e., reflectance/ as a function of wavelength). Spectral response of three pixel of HSI image is shown in Fig. 4 as well as the whole spectral response is shown in Fig. 5

C. Applying K-means Clustering

In the next step K-means clustering is applied with number of cluster is 3 i.e. k=3. Each cluster is label with different color such as cluster 1,2 and 3 has red, blue and green colors respectively. Finally visualize the K-means clustering algorithm as shown in Fig. 6

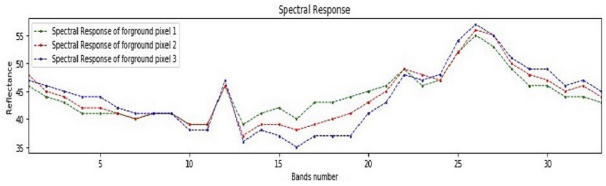

Fig. 4. Spectral response of three pixels of HSI

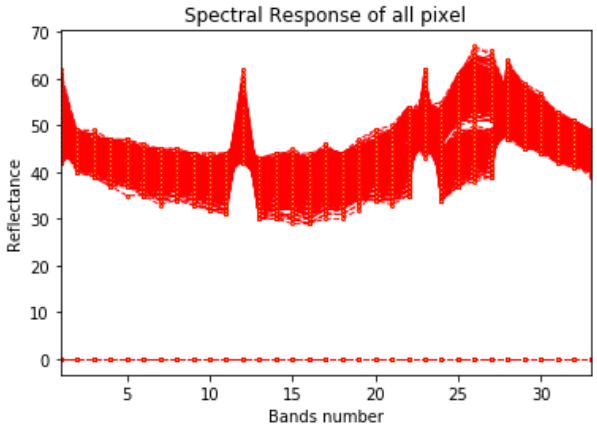

Fig. 5. HSI Spectral response

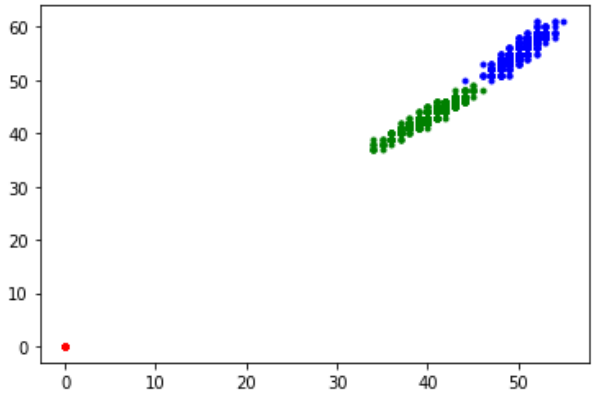

Fig. 6. Data visualization as clusters

D. Assigning Colors

Lastly, we access the labels regenerated from K-means Clustering which classify text written with different inks in the document images. Recolor image as shown in Fig. 7

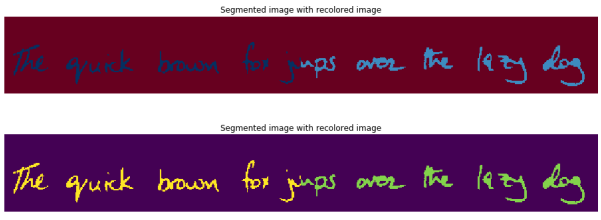

Fig. 7. Recolored image after K-means Clustering

VI. CONCLUSION

HSI has great potential for determining the legitimacy in forensic document examination. We used K-means clustering for automatic ink mismatch detection. This technique displayed great potential in unequal Ink mismatch detection. It is simple to implement, computationally faster and it guarantees convergence. One of the limitations of this technique is manually choosing the no of cluster i.e. "k" and it is sensitive to outliers as well.

VII. FUTURE WORK

In future implementation of different thresholding techniques i.e. Sauvola, Niblack and Global thresholding can improve the results. There is a high probably that changing value of k (clusters) can improve the performance of proposed technique. However, this limitation can also be overcome by using deep network or CNN based end to end manner network in future work. We hope that the results presented in this paper will be more motivating for the researchers to explore new and exciting challenges [23] towards study of forensic documents.

REFERENCES

[1] E. H. Land and J. McCann, "Lightness and retinex theory," JOSA, vol. 61, no. 1, pp. 1–11, 1971.

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

[4] V. Aginsky, "Forensic examinaton of "slightly soluble" ink pigments using thin-layer chromatography," Journal of Forensic Sciences, vol. 38, pp. 1131–1131, 1993.

[5] Z. Khan, F. Shafait and A. Mian, "Hyperspectral Imaging for Ink Mismatch Detection," 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, 2013, pp. 877-881, doi: 10.1109/ICDAR.2013.179.

[6] R. L. Easton Jr, K.T. Knox, and W.A. Christens-Barry, "Multispectral imaging of the Archimedes palimpsest". 32nd IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Oct 2003, pp. 111-116.

[7] W. A. Christens-Barry, K. Boydston, F. G. France, K. T. Knox, R. L. Easton Jr, M. B. Toth, "Camera system for multispectral imaging of documents". Proc SPIE Sensors, Cameras, and Systems for Industrial/Scientific Applications X, pp. 724908-724908, 2009.

[8] R. Padoan, T. A. Steemers, M. Klein, B. Aalderink, and G. de Bruin, "Quantitative hyperspectral imaging of historical documents: technique and applications," ART Proceedings, 2008.

[9] M. E. Klein, B. J. Aalderink, R. Padoan, G. De Bruin, and T. A. Steemers, "Quantitative hyperspectral reflectance imaging," Sensors, vol. 8, no. 9, pp. 5576–5618, 2008.

[10] A. Morales, M. A. Ferrer, M. Diaz-Cabrera, C. Carmona, and G. L. Thomas, "The use of hyperspectral analysis for ink identification in handwritten documents," in Security Technology (ICCST), 2014 International Carnahan Conference on. IEEE, 2014, pp. 1–5.

[11] J. M. Bioucas-Dias and J. M. Nascimento, "Hyperspectral subspace identification," Geoscience and Remote Sensing, IEEE Transactions on, vol. 46, no. 8, pp. 2435–2445, 2008.

[12] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," Signal Processing, IEEE Transactions on, vol. 57, no. 11, pp. 4418–4432, 2009.

[13] Havermans, John Aziz, Hadeel Scholten, Huib. (2003). Non Destructive Detection of Iron-Gall Inks by Means of Multispectral Imaging Part 2: Application on Original Objects Affected With Iron-Gall-Ink Corrosion. Restaurator-international Journal for The Preservation of Library and Archival Material - RESTAURATOR. 24. 88-94. 10.1515/REST.2003.88.

[14] C. S. Silva, M. F. Pimentel, R. S. Honorato, C. Pasquini, J. M. Prats-Montalban, and A. Ferrer, "Near infrared hyperspectral imaging for forensic analysis of document forgery," Analyst, vol. 139, no. 20, pp. 5176–5184, 2014

[15] Z. Khan, F. Shafait, and A. Mian, "Automatic ink mismatch detection for forensic document analysis," Pattern Recognition, 2015, (In Press).

[16] Z. Luo, F. Shafait, A. Mian, Localized forgery detection in hyperspectral document images, in: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, pp. 496–500.

[17] E. Brauns, B. Dyer, Fourier transform hyperspectral visible imaging and the nondestructive analysis of potentially fraudulent documents, Appl. Spectrosc. 60 (8) (2006) 833–840.

[18] S. George, J.Y. Hardeberg, Ink classification and visualization of historical manuscripts: application of hyperspectral imaging, in: Proceedings of the ICDAR, 2015, pp. 1131-1135.

[19] M. J. Khan, K. Khurshid and F. Shafait, "A Spatio-Spectral Hybrid Convolutional Architecture for Hyperspectral Document Authentication," 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 2019, pp. 1097-1102.

[20] Morales, Aythami Ferrer, Miguel Diaz, Moises Carmona-Duarte, Cristina Thomas, Gordon. (2014). The use of Hyperspectral Analysis for Ink Identification in Handwritten Documents. Proceedings - International Carnahan Conference on Security Technology. 2014. 10.1109/CCST.2014.6986980.

[21] R. Hedjam, M. Cheriet, and M. Kalacska, "Constrained energy maximization and self-referencing method for invisible ink detection from multispectral historical document images," in Pattern Recognition (ICPR), 2014 22nd International Conference on. IEEE, 2014, pp. 3026–3031.

[22] M. J. Khan, A. Yousaf, A. Abbas, and K. Khurshid, "Deep learning for automated forgery detection in hyperspectral document images," J. Electron. Imaging, vol. 27, no. 05, p. 1, Sep. 2018.

[23] Rizwan Qureshi, Muhammad Uzair, Khurram Khurshid, Hong Yan, Hyperspectral Document Image Processing: Applications, Challenges and Future Prospects, Pattern Recognition, 90(1): 12-22, 2019.