 doi:10.20944/preprints202007.0002.v1

Ink Detection Using K-Mean Clustering in Hyperspectral Document

Saifullah Saeed
Department of Electrical Engineering
Institute of Space Technology
Islamabad Highway, Islamabad, Pakistan
saifullahsaeed617@gmail.com

Muhammad Reshail Raza
Department of Electrical Engineering
Institute of Space Technology
Islamabad Highway, Islamabad, Pakistan
reshailraza937@gmail.com

Rizwan Sharif
Department of Electrical Engineering
Institute of Space Technology
Islamabad Highway, Islamabad, Pakistan
miftikhar.msee17seecs@seecs.edu.pk

***Abstract*—In document forensic, Ink mismatch relays very important information about forgeries in this way we can find out the authenticity of documents. Finding out and distinguishing these unique inks from the multispectral document is very challenging task. In this paper we proposed the method to identify the inks using clustering. We used K-Mean clustering instead of widely known Fuzzy C-Means Clustering (FCM) and successfully identity the number of inks. For the purpose of optimizing and improving our results we used two optimization techniques such as Elbow and silhouette optimization techniques.**

***Index Terms*—Forensic, Ink Mismatch, Clustering, K-means algorithm, Elbow, Silhouette**

I. Introduction

It is known that that every material has properties to reflect, transmit light or even absorb it in particular way, that gives every material their distinctive color [12]. This process is called as Spectral response. Distinctive material color can be recognized from light by eye, but eye can't differentiate between two visibly close colors even though as per their spectral information is concerned, they are very distant from each other.

As it is told that human eye can't differentiate between two spectrally distinct but visibly colors so it is very important to check whether the documents have been forged or not by forensics using ink mismatching techniques. If the document has more than one type of ink then this means that document has been forged. As it can't be seen by naked eye because inks used are visibly close but they are in fact very distinct from each other on the basis of spectral responses.

So, analysis of inks is of great importance to probe the questioned documents for their validity. Some people might manipulate the information to use for blackmailing, fraud, forgeries, backdating and illegal purposes. As documents used as evidences in courts, but authenticity and dignity of evidence is very important. To validate the document and distinguish ink, there are two main approaches one is destructive while other one is non-destructive examinations. In destructive approach chemical analysis such as Thin Layer Chromatography (TLC) is utilized to separate different components of ink's mixture [16]. But this approach holds variables that can affect the process and don't give us optimal results like it is sensitive to temperature also many other factors should be taken care of to get the results. So, it is very time consuming. On the other hand, non-destructive approach is soft and has a lot of great aspects. The approach prospects are great and so does its potential as the one proposed with spectral imaging.

Hyperspectral imaging is emerged as efficient non-destructive tool for detection, comparison, enhancement and identification of forensic traces. Such systems are being used in separating inks of documents by forensics. But this is very difficult and time consuming considering that document needs to be manually observed by the examiner under each wavelength of time and then based on qualitative analysis the examiner can make decisions which are very time consuming. Numbers of observable wavelengths is mainly dependents upons the spectral resolution of imaging system. Thus, for efficient and effective questioned document examination ink analysis plays vital role. Captured hyperspectral images are manually analyzed by Examiners on a band-by-band basis and attempt to distinguish the distinctions through visual assessment. This procedure is difficult time consuming, and rather subjective. Therefore, for document examiners automatic ink analysis system would be essentially useful.

In this paper, an efficient K-Means Clustering based automatic ink mismatch detection technique is proposed. Ink pixels are segmented out using local thresholding and K-Means is used to divide the spectral response vectors of ink pixels into different clusters. Experiments are carried out on ink combinations using different techniques to find out the optimal Inks.

II. Related Work

Efficiency of forgery detection systems for Hyperspectral image analysis has been significantly improved over the recent years. For automated forgery detection systems [5] various HIS based models and techniques have been employed and used. Hyperspectral sensing system for non-destructive forgery detection was developed by E.B. Brauns et al [2]. by employing interferometer in questioned documents that are potentially deceitful documents. The interferometer depends on different moving parts for frequency tuning or slowing down the acquisition procedure and subsequently moderation. To distinguish between different writing inks this work fills in as a concept. In the National Archives of Netherlands, a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2020 doi:10.20944/preprints202007.0002.v1

comparatively complex and advanced Hyperspectral Imaging framework for historic documents examination was developed that provides high resolution and good outcome in spatial as well as spectral domain [7] [3], i-e. ranging from near ultraviolet to near infrared range. The acquisition time of hyperspectral sensors was very long even though for analysis of ancient documents it was very good, efficient and robust. In the past for writing analysis in document images Ink-deposition traces and texture [4] have been used. Our proposed work exclusively concentrates on finding the Ink- spectral responses as well as finding optimal number of Inks that are used in counterfeiting the documents.

Accurate differentiation of materials based on their spectral signatures can be found out using Hyperspectral imaging. So, HIS has very good potential to differentiate the inks used based on their spectral analysis. Extensively rich information can be found out by even a spectrum of one pixel about the surface of material in comparison to color image. HSI analysis for refurbishment of ancient documents was used by R.J. Hejdam et al [17]., while targeting the practical scenarios where ink could not be detected or differentiated via naked human eye. Image enhancement for degraded images was proposed by F. Hollaus et al [6]. In which he proposed that degraded images can be enhanced via utilizing spectral as well as spatial information. Chemometrics was employed by C.S. Silva et al [8]. and to detect forgery in questioned documents using HSI analysis that used the age information of various inks. In an embedded system, HSI analysis and least square SVM classification for ink analysis is employed by A. Morales et al [11]. The method proposed by Z. Khan et al [1]. for forgery detection system was based on Hyperspectral image analysis that for discrimination among inks uses k-means clustering after selection of optimum bands in a potentially fake document. However, it is assumed for this experiment to hold that two type of inks has been used in equal proportions for the document, due to which this technique made it not practical and applicable in real life scenarios. Based on localized HSI analysis to detect forgery in documents was proposed by K. Khurshid et al [9]., by enhancing the visually negligible differences which can't be perceived by naked human eye between the quality of inks in forged as well as original documents.

III. Proposed Methodology

The hyperspectral document image $I \in K^{MxNxB}$ as input the M is number of rows , N is number of column and B is number of bands.First we process the data and convert $I \in K^{MxNxB}$ into $A_{i,j} \in K^{MxN}$. We applied k-means algorithm[13] with different cluster 'Q'.

$$w = \sum_{q=1}^{Q} \sum_{i=0,j=0}^{K} (A_{i,j}, C_q) \tag{1}$$

In k-means every cluster (Q) gave information of inks.But we didn't find exact number of number of inks with k-means.so we use k-means with elbow method[14].In elbow method, describe the explained variation of clusters function and get the elbow and gave optimal number of clusters.if we were increasing number of clusters and improve explained variation.if we didn't find elbow and gave over fitting.we use point to line distance formula.we find the projection required point on given line.The maximum distance told elbow.Pinitial and Pend is clusters distances.

We used another method Silhouette method[15] with k-means clustering.Silhouette value measure how similar point cohesive with own cluster and separate with other cluster.its range [-1 +1],-ve shows the point assign wrong cluster and +ve value shows the point assign own cluster.The Silhouette Value sli(i) for data point *i*:

$$sli(i) = \frac{b(i) - a(i)}{max\{b(i), a(i)\}} \tag{2}$$

subject to

$$-1 \leq sli(i) \leq 1 | Q_i | > 1 \tag{3}$$

If sli(i) is zero, its mean minimum one point in cluster and b(i) calculate the dissimilarity of each point to other clusters.

$$b(i) = min_{i \neq j} \frac{1}{|Q_j|} \sum_{j \in Q_j} d(i,j) \tag{4}$$

a(i) calculate similarity of each point to its own clusters and calculate average distance.

$$b(i) = \frac{1}{|Q_i| - 1} \sum_{j \in Q_j, i \neq j} d(i,j) \tag{5}$$

Silhouette measure optimal number of clusters at global maximum point.we compared both method results and find number of inks.

IV. Experimental Analysis

After we perform experimental analysis on document images with different ratios of inks. The UWA Writing Inks Database has been used for experiment. The database contains 33 bands for each hyperspectral images in the visible range. First we extract three bands for our experiment from all bands of hyperspectral image bands.

We separate out the foreground and background pixels from image and then plotted the spectral responses foreground.and spectral responses of foreground pixels to show the number of unique values of image and variations. Spectral responses of first band 1 and all bands are shown below 2 .

After finding spectral responses we found out the optimal number of inks using K-Means as well as other optimization techniques such as elbow and silhouette method to get optimal number of inks. The elbow method runs k-means clustering on the data set for a range of values for k and then for each value of k computes an average score for all clusters. When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the "elbow" (the point of inflection on the curve) is the best value of k. It is also noted that Elbow

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2020 doi:10.20944/preprints202007.0002.v1

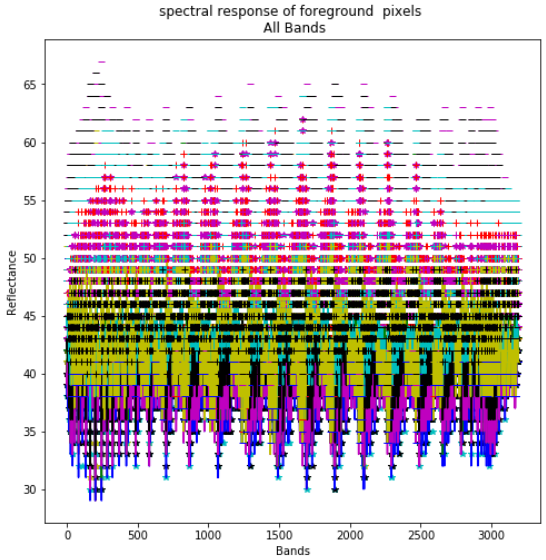

Fig. 1. Spectral responce of all bands.

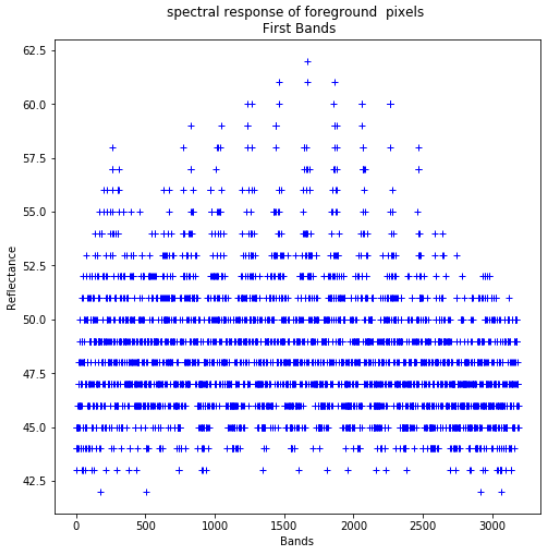

Fig. 2. Spectral responce of one band.

method cannot be used if data is not very clustered, in this case the curve will be smooth and value of k becomes unclear.

For Elbow method first we defined Distortion and Inertia for k=21 clusters. We iterate the values of k from 1 to 21 and calculate the values of distortions for each value of k and calculate the distortion and inertia for each value of k in the given range. To determine the optimal number of clusters, we have to select the value of k at the “elbow” i-e the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we concluded that the optimal number of clusters for the data is 5 as shown in Figure 3.

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or

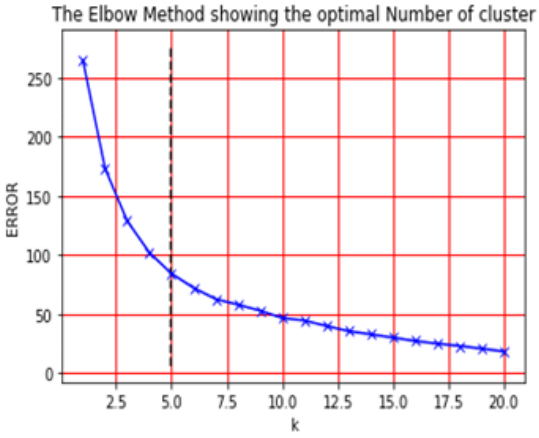

Fig. 3. Graphical Representation of Elbow Method.

very close to the decision boundary between two neighboring clusters and negative values indicate that those sample might have been assigned to the wrong cluster. The highest peak of curve is selected as optimal values of K and it can be seen that the highest value of curve occurs at 4 that means the optimal value of clusters suggested by this method is 4 as shown in Figure 4 while for Elbow method it was five as it was seen in above mentioned Figure 3.

After that we applied inks(clusters) on given hyperspectral document. It was applied on image1 only but it can be used on all pictures.

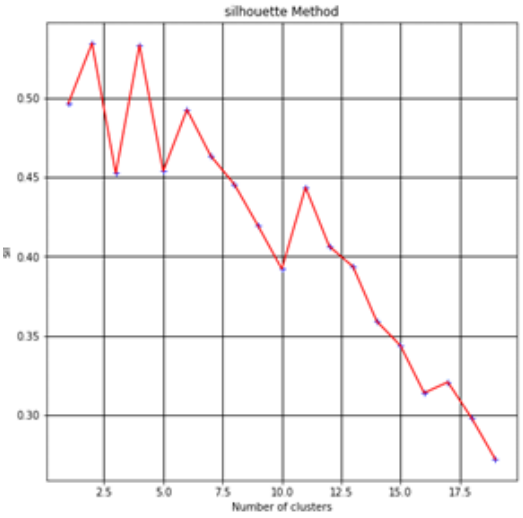

Fig. 4. Graphical Representation of Silhouette Method.

V. Conclusion

Detection of Ink mismatch is a key step in forgery detection. For distinguishing between visually similar inks in Multispectral documents images spectral information is very useful. We used K-Means Clustering technique with Elbow

as well as Silhouette method to get the optimal number of inks by distinguishing the spectral responses for ink mismatch detection. The proposed method gives better results in the questioned document.

REFERENCES

[1] 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.

[2] Brauns, E. B., Dyer, R. B. (2006). Fourier Transform Hyperspectral Visible Imaging and the Nondestructive Analysis of Potentially Fraudulent Documents. Applied Spectroscopy, 60(8), 833–840.

[3] 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.

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

[5] 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.

[6] F. Hollaus, M. Gau, and R. Sablatnig, "Enhancement of multispectral images of degraded documents by employing spatial information", 12th Int Conf on Document Analysis and Recognition (ICDAR), Aug 2013, pp. 145-149.

[7] 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.

[8] C. S. Silva, M. F. Pimentel, R. S. Honorato, C. Pasquini, J. M. PratsMontalban, and A. Ferrer, Near infrared hyperspectral imaging for forensic analysis of document forgery, Analyst, vol. 139, no. 20, pp.51765184, 2014

[9] 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.

[10] Ammad ul Islam, Muhammad Jaleed Khan, Khurram Khurshid, Faisal Shafait, Hyperspectral Image Analysis for Writer Identification using Deep Learning, Digital Image Computing: Techniques and Applications (DICTA), December 2019, Australia

[11] C. S. Silva, M. F. Pimentel, R. S. Honorato, C. Pasquini, J. M. PratsMontalban, and A. Ferrer, Near infrared hyperspectral imaging for forensic analysis of document forgery, Analyst, vol. 139, no. 20, pp. 51765184, 2014.

[12] E. H. Land, J. J. McCann et al., "Lightness and retinex theory," Journal of the Optical society of America, vol. 61, no. 1, pp. 1–11, 1971.

[13] Steinley, Douglas Brusco, Michael. (2011). Choosing the Number of Clusters in K-Means Clustering. Psychological methods. 16. 285-97. 10.1037/a0023346.

[14] Thorndike, R.L. Who belongs in the family?. Psychome- trika 18, 267–276

[15] Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65. Journal of Computational and Applied Mathematics. 20. 53-65. 10.1016/0377-0427(87)90125-7.

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

[17] Hedjam, Rachid Cheriet, Mohamed Kalacska, Margaret. (2014). Constrained Energy Maximization and Self-Referencing Method for Invisible Ink Detection from Multispectral Historical Document Images. Proceedings - International Conference on Pattern Recognition. 10.1109/ICPR.2014.522.