Classification and Feature Extraction for Document Forgery Images

Rishabh Singh, Garima Jaiswal, Aditi Jain, and Arun Shrama

Abstract Hyperspectral images (HSIs) are captured over continuous wavebands. It includes the object's spatial as well as spectral information. Different materials have different reflection and absorption properties, which allows obtaining the underlying sample's physical structure and chemical composition. Therefore, hyperspectral imaging technology gives comprehensive information about the sample. This paper uses a hyperspectral imaging approach to detect document forgery in the documents through ink mismatch detection. The proposed approach utilizes PCA to handle multiple dimensions in HSIs. It captures the spectral features which act as an input to a convolutional neural network for image classification. The method is applied to the UWA Writing Ink HSI (WIHSI) database. The results are compared with the state-of-art results that prove the proposed approach's potential.

Keywords Hyperspectral imaging · Spectral · PCA · Document forgery · Deep learning

1 Introduction

Hyperspectral Imaging is an advanced technique based on spectroscopy. It collects images over a wide and continuous range of wavelengths for the same spatial area. It captures the spatial and spectral information from the object under analysis. It divides

R. Singh · A. Jain
Netaji Subhas Institute of Technology, University of Delhi, New Delhi, India
e-mail: rishabhs.ec.16@nsit.net.in

A. Jain
e-mail: aditijain0802@gmail.com

G. Jaiswal (✉) · A. Shrama
Indira Gandhi Delhi Technical University for Women, Delhi, India
e-mail: garima121@gmail.com

A. Shrama
e-mail: arunsharma@igdtuw.ac.in

A. Tomar et al. (eds.), *Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication*, Lecture Notes in Electrical Engineering 915, https://doi.org/10.1007/978-981-19-2828-4_68

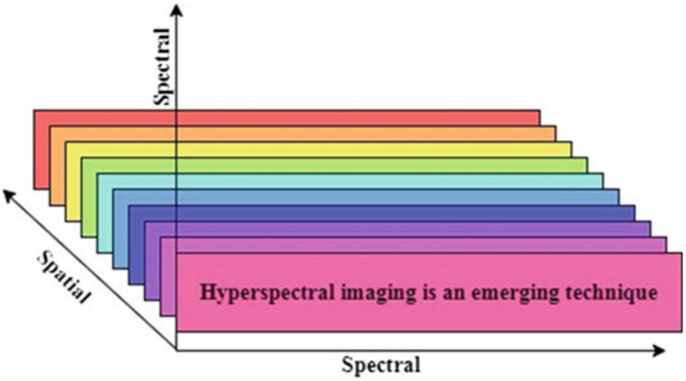

Fig. 1 Capturing Images at multiple bands

the spectrum into many more bands as illustrated in Fig. 1. It helps to determine the composition and assigns a unique spectral signature of the underlying material. It is being used in forensics, agriculture, archeology, security, medical diagnosis and surgery, remote sensing, and many more [1]. The unique spectral signature of each material in hyperspectral imaging (HSI) gives it tremendous power in identification and verification purposes [2].

Document forgery involves creating, falsifying, modifying the document with the wrong intention. In the present world of tools and techniques, detecting document forgery is a mysterious task that a naked eye cannot identify. In Document Analysis, the material's unique spectral signature helps us identify ink mismatches, the timeline of manuscripts, and recovering the degraded scripts. Cutting-edge technology like deep learning approaches has attained noticeable results in many domains [3].

In the present study, we utilized PCA for dimensionality reduction, and the extracted spectral features are fed to CNN to detect document forgery. The paper has been segregated into various sections. Section 2 illustrates the applications of hyperspectral imaging in various domains. Related work is elaborated in Sect. 3. The experimental setup and results are discussed in Sects. 4 and 5. Conclusion and Future work are depicted in Sect. 6.

2 Applications

Hyperspectral imaging is an emerging approach, gaining popularity for solving problems in many domains. This section highlights the application areas for hyperspectral imaging in the document analysis domain.

2.1 *Document Analysis*

Ink Mismatch Detection. The document forgery has always been of great concern and is frequently seen in fraudulent bank cheques, tampering with historical manuscripts and forensic evidence, and many more. If a document consists of more than one ink of similar color that may indicate a possible forgery. Currently, there are two approaches for identifying the ink mismatching-destructive and non-destructive. Non-destructive methods such as hyperspectral imaging are very useful in identifying the homogeneity of the ink.

Recovering Degraded Documents. The historical manuscripts, sometimes with time or due to some external factors, get degraded and unreadable. The unique spectral signature assigned to every element in an HSI can help to trace the text by applying image classification models [4].

Writer Identification. HSI is used to recognize the handwriting in a document. This approach helps to identify the writer or the owner of the document. It assists in identifying the modification and tampering done in the documents.

2.2 *Supply Chain Management*

Inventory Management. Supply chain management requires innovations and developments that may help it work in a more efficient and faster way in managing the inventory. The crucial role is to identify and also verify the articles. The problem usually comes in the verification of the article. It is a tedious task to separate articles manually and verify them. The customarily used cameras give us the RGB image, short of the subtle differences immune to naked eyes [5]. Here comes the beauty of HSIs, which allows capturing minute differences and the properties of the article's material. It allows us to exploit this and use it in varied applications.

One important application can be calculating the applicable charges on import and export according to the quality of the material. For example, there is a big difference in the import duty on varied fabrics. Still, they all may seem identical to a non-expert, or it usually becomes cumbersome to segregate the items manually. The materials appearing to be the same in the RGB camera should be charged differently according to the material's composition [6]. Here hyperspectral imaging can be used to facilitate knowledge of the properties of the material. Another practical application can be in the verification of the quality of the material. In businesses, there is a bulk sale and purchase of goods and ensuring the desired quality is their utmost priority. The use of HSI of the received articles will help to detect and match with the properties of the article promised.

Determine the quality of Food Products. The hyperspectral images of food products give essential spectral information about it. The commonly used colored image

(RGB) or seen by the naked eye only helps to determine the outer health [7]. While using hyperspectral images helps determine the complete health, i.e., inner and outer health [8]. It can be helpful in the management of the supply chain to separate unhealthy food like rotten vegetables or fruits. It can be used to effectively decide the product's cost based on their health.

3 Related Work

This paper focused on applying hyperspectral imaging for document forgery detection. We proposed to detect ink mismatch in HSIs to detect document forgery.

This section elaborates the related work for document forgery using machine learning and deep learning approaches in Table 1.

Table 1 Keypoints of the related work

S. No.	Paper ID	Approach	Limitations
1	Khan et al. [7]	JSBS for automatic mismatch detection for ink classification	Optimal band selection
2	Abbas et al. [8]	HySime for ink classification	HySime overestimates the number of inks
3	Khan et al. [9]	CNN with different architectures on spectral features	Spatial features unexplored
4	Khan et al. [10]	CNN with different architectures on spatial as well as spectral features	Uses supervised learning that requires prior knowledge
5	Islam et al. [11]	CNN on Spatio-Spectral features for writer identification	Unsupervised deep learning is not explored
6	Devassy et al. [12]	1-D-CNN, SAM and SID	Deeper neural networks unexplored
7	Qureshi et al. [13]	Literature Review of pattern recognition techniques	Handwriting classification unexplored
8	Devassy et al. [14]	Unsupervised clustering using t-SNE algorithm	Validation against other non-linear methods
9	Silva et al. [15]	PCA, MCR-ALS, and PLS-DA	Intersecting lines and identification of all the samples
10	Luo et al. [16]	Schwartz and P2P	Assumption on maximum number of inks mixed
11	Lian et al. [17]	Hyperspectral imager Nuance-Macro and software Nuance 1p46	Amount of ink applied and structure and surface of the paper substrate unexplored

4 Experimental Setup

This section elaborates the experimental setup for detecting document forgery using hyperspectral document images. We proposed a supervised neural network algorithm to detect document forgery in HSI that uses spectral information to classify the image's pixels (Fig. 2).

4.1 Database and Preprocessing

The WIHSI database [13] contains images of seven subjects. A single image in the database comprises five lines, all with same color (blue/black) but distinct ink. They are written in English by the subject. So, a total of 14 HSIs having 752 * 480 pixels, spanning across of 33 bands from 400 to 720 nm at a step of 10 nm, were captured [18]. The illumination in the images is non-uniform. Each hyperspectral image is exposed to preprocessing for further experimentation.

4.2 Preprocessing

We aim to process the data in such a way that it is ready for further experimentation. The first step is to separate each line to extract the background pixels via image thresholding. The global thresholding techniques fail to give satisfactory results as the image has non-uniform illumination. Sauvola's binarization method is used [9, 10] in this case as it threshold the image locally instead of globally. After this step, we have five hyperspectral data cubes, each containing an English phrase from every image along with their binary masks.

The objective of this work is to detect different inks in the same document with their unique spectral signature. To accomplish this, inks of the same subject were mixed in varying proportions [9, 10, 13]. No two different colored inks were mixed together, as it can be distinguished visually. Samples were generated using two, three, four, and five inks in equal and unequal proportions.

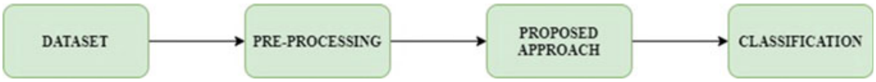

Fig. 2 Experimental setup

4.3 Proposed Approach

Principal Component Analysis is implemented before passing the image to the neural network to reduce the dimensions. With PCA, we can get rid of some of the features and map our dataset into a reduced subspace without losing essential information about the original dataset.

The objective is to extract the spatial features by preserving the spectral information of the HSI. The number of features to preserve has to be decided in this step. The number of principal components was varied from (3, 4, 5, 8, 9, 11, 13, 15, 17, 19, and 21). Its impact on accuracy was noted. It was concluded that with a count of 9 or above, the accuracy hardly changed. After the analysis, we selected 9 as the count.

Deep Learning is a cutting-edge technology that automatically captures features from a large dataset [19, 20]. A CNN consists of various layers: convolutional, activation, and pooling layers, followed by a connected layer that produces the output. The extracted spectral features were passed through a CNN model, as illustrated in Fig. 3 for classification.

5 Results

To analyze the proposed approach, we investigated the accuracy of ink mixing proportions for blue and black inks. The computed results are compared with the state-of-art approaches as illustrated in Table 2. The results depicted that black inks were challenging to identify compared to blue ink [19].

6 Future Prospects

The present study proposed a supervised deep learning-based method combined with PCA in HSIs for forgery detection. We evaluated the proposed approach by combining different inks in various proportions. The results clearly stated the effectiveness of the proposed approach. The present work may be extended using hybrid spectral and spatial features for forgery detection. Moreover, the supervised deep learning approach demands to know the count of the inks mixed in proportion in advance, which imposes a constraint for practical application. Unsupervised deep learning techniques may be examined to overcome this limitation.

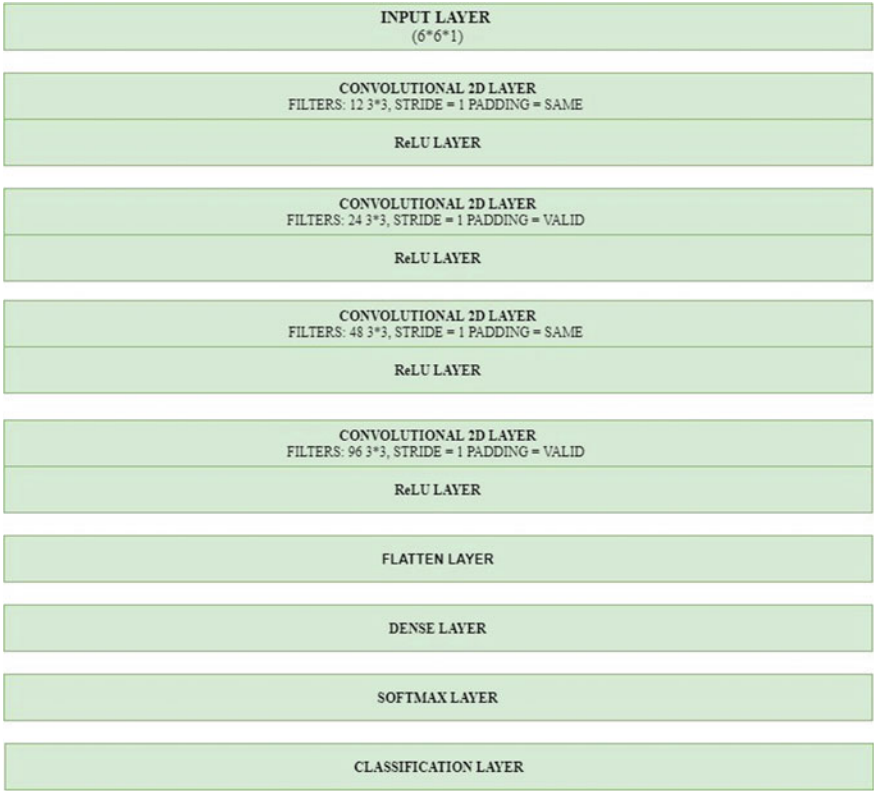

Fig. 3 CNN architecture

Table 2 Comparison of the proposed approach with state-of-art results

Citations	Approach	Average accuracy (%)		Maximum number of inks artificially mixed
		Blue inks	Black inks	
Proposed approach	Supervised deep learning	90.2	89.7	5
[13]	Feature selection	86.7	89.0	2
[14]	Unmixing	86.2	83.4	4
[21]	Unsupervised machine learning	89	82.3	2
[15]		86.7	81.9	2
[18]		85.6	81.4	2

References

1. Jaiswal G, Sharma A, Yadav SK (2021) Critical insights into modern hyperspectral image applications through deep learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1426
2. Jaiswal G, Sharma A, Yadav SK (2022) Deep feature extraction for document forgery detection with convolutional autoencoders. Comput Electr Eng 99(107770)
3. Jaiswal G, Sharma A, Yadav SK (2019) Analytical approach for predicting dropouts in higher education. Int J Inf Commun Technol Educ (IJICTE) 15(3):89–102
4. Jyothsnaa S, Gandhe A, Deshpande A, Bodas S (2010, November) Automated inventory management and security surveillance system using image processing techniques. In: TENCON 2010–2010 IEEE Region 10 Conference. IEEE, New York, pp 2318–2321
5. Siche R, Vejarano R, Aredo V, Velasquez L, Saldana E, Quevedo R (2016) Evaluation of food quality and safety with hyperspectral imaging (HSI). Food Eng Rev 8(3):306–322
6. Lorente D, Blasco J, Serrano AJ, Soria-Olivas E, Aleixos N, Gomez-Sanchis J (2013) Comparison of ROC feature selection method for the detection of decay in citrus fruit using hyperspectral images. Food Bioprocess Technol 6(12):3613–3619
7. Khan Z, Shafait F, Mian A (2015) Automatic ink mismatch detection for forensic document analysis. Pattern Recogn 48(11):3615–3626
8. Abbas A, Khurshid K, Shafait F (2017, November) Towards automated ink mismatch detection in hyperspectral document images. In: 2017 14th IAPR International conference on document analysis and recognition (ICDAR), vol 1. IEEE, New York, pp 1229–1236
9. Khan MJ, Yousaf A, Abbas A, Khurshid K (2018) Deep learning for automated forgery detection in hyperspectral document images. J Electron Imaging 27(5):053001
10. Khan MJ, Khurshid K, Shafait F (2019, September) A spatio-spectral hybrid convolutional architecture for hyperspectral document authentication. In: 2019 International conference on document analysis and recognition (ICDAR). IEEE, New York, pp 1097–1102
11. Islam AU, Khan MJ, Khurshid K, Shafait F (2019, December) Hyperspectral image analysis for writer identification using deep learning. In: 2019 Digital image computing: techniques and applications (DICTA). IEEE, New York, pp 1–7
12. Devassy BM, George S (2019) Ink classification using convolutional neural network. NISK J 12:1–16
13. Qureshi R, Uzair M, Khurshid K, Yan H (2019) Hyperspectral document image processing: applications, challenges and future prospects. Pattern Recogn 90:12–22
14. MelitDevassy B, George S, Nussbaum P (2020) Unsupervised clustering of hyperspectral paper data using t-SNE. J Imaging 6(5):29
15. Silva CS, Pimentel MF, Honorato RS, Pasquini C, Prats-Montalbán JM, Ferrer A (2014) Near infrared hyperspectral imaging for forensic analysis of document forgery. Analyst 139(20):5176–5184
16. Luo Z, Shafait F, Mian A (2015, August) Localized forgery detection in hyperspectral document images. In: 2015 13th International conference on document analysis and recognition (ICDAR). IEEE, New York, pp 496–500
17. Lian Y, Liang L, Li B (2017) Hyperspectral imaging technology for revealing the original handwritings covered by the same inks. J Forensic Sci Med 3(4):210
18. Khan Z, Shafait F, Mian A (2013, August) Hyperspectral imaging for ink mismatch detection. In: 2013 12th International conference on document analysis and recognition. IEEE, New York, pp 877–881
19. Jaiswal G, Sharma A, Yadav SK (2021) Efficient ink mismatch detection using supervised approach. In: Singh M, Tyagi V, Gupta PK, Flusser J, Ören T, Sonawane VR (eds) Advances in computing and data sciences. ICACDS 2021. Communications in computer and information science, vol 1440. Springer, Cham. https://doi.org/10.1007/978-3-030-81462-5_65

20. Tomar A et al (2020) Machine learning, advances in computing, renewable energy and communication, LNEE vol 768. Springer Nature, Berlin, 659 p. https://doi.org/10.1007/978-981-16-2354-7. ISBN 978-981-16-2354-7
21. Khan Z, Shafait F, Mian A (2013, August) Hyperspectral imaging for ink mismatch detection. In: 2013 12th International conference on document analysis and recognition. IEEE, pp 877-881