IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ∺ Peer-reviewed & Refereed journal ∺ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15841

Historical Handwritten Document Age Identification using DCT-based Frequency Analysis

Dr. Pushpalata Gonasagi

Associate Professor, Department of Computer Science, Govt. First Grade College Mahagaon Cross, Kalaburagi, Karnataka, India.

Abstract: Historical handwritten documents hold invaluable information about the development of language, literature, governance, education and cultural heritage. The automatic age estimation of such documents has become a major research question in digital humanities, archival science and forensic document analysis. In this paper, we propose a frequency-domain approach for age estimation of historical handwritten documents based on the Discrete Cosine Transform (DCT). The proposed framework starts with the preprocessing to minimize the scanning noise and normalize the document images. Each document image is divided into non-overlapping blocks and the 2D DCT is applied to derive frequency coefficients which represent the structural features of the handwriting and paper texture. To create discriminative feature vectors, we employ the statistical descriptors of certain DCT coefficients. SVM classifiers are used to categorize these features. Using the publicly accessible Medieval Paleographic Scale (MPS) dataset, which comprises handwritten manuscripts from several historical eras, we assess the suggested method. The findings of the experiment demonstrate that frequency features based on DCT may reliably classify ancient manuscripts and capture their age-related characteristics.

Keywords: Historical document, DCT, MPS, Document age estimation, SVM.

I. INTRODUCTION

Historical handwritten documents are valuable cultural assets that record the linguistic, social, administrative, and educational development of civilizations over many centuries. Large-scale digitization projects carried out by libraries, museums, and archives have made these documents accessible for digital preservation and scholarly research. Digital document collections enable applications such as handwriting recognition, document retrieval, writer identification, manuscript authentication, and historical document dating. Determining the approximate age of historical manuscripts is an important task because many historical documents do not explicitly record their creation date. Accurate dating assists historians in establishing document authenticity, identifying the period of origin, studying the evolution of handwriting styles, and organizing archival collections. Traditionally, paleographers estimate document age through manual examination of handwriting characteristics, writing materials, and document appearance. However, manual analysis is time-consuming and depends heavily on expert knowledge. Image processing and machine learning techniques provide an efficient alternative for automatic document age identification. Earlier methods have primarily relied on spatial-domain texture descriptors. Although these techniques effectively describe local texture information, they may not sufficiently represent the global frequency characteristics associated with handwriting evolution and document degradation. The DCT provides a compact frequency-domain representation of image content by transforming spatial pixel intensities into cosine frequency coefficients. Most structural information is concentrated in low-frequency components, whereas handwriting details, stroke boundaries, and document texture are represented by middle- and high-frequency coefficients. Variations in writing style, ink distribution, and manuscript texture across historical periods produce measurable changes in these frequency components, making DCT a suitable feature extraction technique for historical document age estimation.

II. RELATED WORK

Several research studies have been carried out for identifying the age of historical and printed documents. A brief review of some significant contributions in the area of document age estimation is presented below.

Sheng He et al. [1] introduced a document dating approach using scale-invariant features on the MPS dataset. Their method generated stroke elements by analyzing the primary strokes of handwritten characters and applied an Evaluation

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁘ Peer-reviewed & Refereed journal ⁘ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15841

of Self-Organizing (EOS) map to study the evolution of writing patterns over time. The proposed system achieved an accuracy of 85.1% using the K-Nearest Neighbor (K-NN) classifier. Sheng He et al. [2] investigated historical document dating by analyzing stroke fragments and local contour information. The extracted stroke and contour features were designed to be rotation and scale invariant. These features were represented using a Bag-of-Words model for classification. Their approach achieved a mean absolute error of 14.9, demonstrating effective performance in historical document age estimation. Halder et al. [3] proposed a technique for estimating the age of ink in printed documents. The method extracted color-based features, including average intensity, pixel profiles, and kurtosis values, from Life magazine cover images published between 1930 and 1970. A neural network classifier was trained to estimate the age of unknown document samples, achieving an accuracy of 74.5%. Raghunandan et al. [4] presented a method for handwritten document classification based on foreground and background information to distinguish original and forged documents. Their approach utilized Fourier coefficients to analyze contrast variations and classify documents into old and new categories. The method obtained accuracies of 78.5% for new documents and 77.5% for old documents. Barboza et al. [5] proposed a document age estimation method based on background information extracted from RGB color components of scanned document images. The study considered birth certificates, wedding certificates, and other documents from the period between 1913 and 1952, with a one-year age interval. Gaussian distribution models were applied to normalized color components for estimating document age. Sheng He et al. [6] introduced a Polar Stroke Descriptor (PSD)-based method for historical document dating. The PSD represents local stroke characteristics using rotation-invariant features and captures stroke length information in different directions. The extracted features were classified using the K-NN classifier, achieving an accuracy of 60.7%. Fernando et al. [7] developed a model for dating historical color photographs using color component derivatives and angle-based features. The extracted features were classified using linear and nonlinear SVM classifiers, resulting in an accuracy of 85.5%. P Gonasagi et al. [8] proposed a historical handwritten document age identification method using Local Binary Pattern (LBP) and Local Phase Quantization (LPQ) features on the publicly available MPS dataset. The extracted texture features were classified using K-NN and SVM classifiers. The proposed method achieved classification accuracies of 90.7% with K-NN and 92.8% with SVM, demonstrating the effectiveness of texture-based descriptors for historical document age classification. Gonasagi et al. [9] proposed a printed document age identification method using Histogram of Oriented Gradients (HOG) and LPQ features extracted from Life magazine cover pages published between 1930 and 1970. The extracted features were classified using a SVM classifier, achieving an overall classification accuracy of 86.2%.

The above studies demonstrate that various feature extraction techniques, including stroke-based features, color information, Fourier analysis, and texture descriptors, have been explored for historical document age identification. However, challenges remain due to variations in document degradation, writing styles, and environmental conditions. Therefore, frequency-domain approaches such as DCT-based feature extraction can provide an effective alternative for capturing age-related characteristics of historical documents.

III. PROPOSED METHOD

This work proposes a DCT-based framework for classifying historical handwritten documents into different historical periods using statistical frequency features and supervised machine learning classifiers. The major contributions of this work include: Frequency-domain representation of historical manuscripts using Discrete Cosine Transform. Extraction of statistical descriptors from DCT coefficients for historical document characterization. Classification using Support Vector Machine classifiers. Evaluation on a publicly available historical handwritten document dataset. The methodology consists of four major stages: historical document dataset preparation, image preprocessing, DCT-based feature extraction, and classification. Initially, historical document images are collected and organized into different classes based on their respective time periods.

A. Dataset

The proposed method is evaluated using the MPS dataset, which contains handwritten manuscript images collected from medieval European archives [10]. The dataset includes documents written between 1300 CE and 1550 CE, covering eleven historical periods separated by approximately twenty-five years. The documents exhibit considerable variation in handwriting style, character formation, ink density, paper quality, scanning conditions, and natural aging effects. These variations make the dataset suitable for evaluating document age identification algorithms. Sample images of MPS are shown in the Fig.1.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⩕ Peer-reviewed & Refereed journal ⩕ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15841

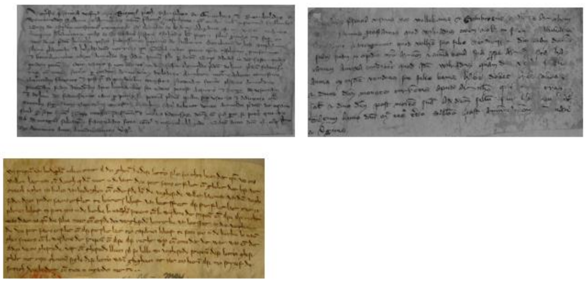

Fig. 1 Sample MPS document images

B. Preprocessing

Historical manuscript images often contain various types of degradation and noise introduced by aging, ink fading, paper deterioration, and the digitization process. Therefore, a preprocessing stage is performed before feature extraction to improve image quality and eliminate unwanted artifacts. Initially, the input color document image is converted into a grayscale image to reduce computational complexity while preserving essential handwriting information. A median filter is then applied to remove impulse noise without affecting the handwritten strokes. To enhance the visibility of the text, histogram equalization is performed to improve image contrast. Subsequently, morphological operations are used to remove small isolated background noise and other unwanted objects. Finally, image intensity normalization is carried out to minimize illumination variations and produce a consistent appearance across all document images. After preprocessing, each cleaned document image is segmented into text blocks of 512 × 512 pixels, ensuring that only regions containing complete handwritten text are selected for DCT-based feature extraction.

C. DCT-Based Feature Extraction

After preprocessing, each normalized text block is divided into non-overlapping blocks of 8 × 8 pixels. A two-dimensional Discrete Cosine Transform (DCT) is applied to each block to transform the spatial representation of the handwritten document into the frequency domain. In the DCT representation, the low-frequency coefficients capture the overall intensity and background information of the document, the middle-frequency coefficients represent important handwriting strokes, paper texture, and structural characteristics, while the high-frequency coefficients mainly contain fine details and noise. The resulting DCT coefficients are arranged using a zigzag scanning technique, which orders the coefficients from low to high frequency and facilitates the selection of the most informative components. Statistical measures, including mean, variance, standard deviation, energy, entropy, skewness, and kurtosis, are then computed from the selected DCT coefficients. These statistical descriptors effectively summarize the frequency characteristics associated with handwriting style, texture, and document degradation. Finally, all the extracted statistical features are concatenated to form a compact feature vector that represents each historical manuscript and serves as the input for the classification stage. The statistical features extracted from each text block are concatenated to form the final feature vector as: feature vector = [mean, variance, standard devaition, energy, entropy, skewness, kurtosis]

D Classification

The extracted feature vectors are classified using supervised learning algorithms. SVM constructs optimal separating hyperplanes between historical document classes. Model performance is evaluated using 10-fold cross-validation, ensuring that every document is used once for testing and nine times for training. Average classification accuracy is reported across all folds.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of the proposed DCT-based historical handwritten document age identification framework is evaluated using the MPS dataset. The dataset consists of historical handwritten document images belonging to different periods between 1300 CE and 1550 CE. The documents are categorized into eleven classes based on their approximate period of origin. Each class represents a specific historical era with an interval of approximately 25 years. The experiments are conducted using the DCT-based statistical features extracted from the preprocessed document text blocks. The extracted feature vectors consist of frequency-domain information representing handwriting variations, paper texture, ink distribution, and degradation patterns. The effectiveness of the proposed method is evaluated using two machine learning classifiers: SVM. A 10-fold cross-validation strategy is adopted to evaluate the classification performance and generalization ability of the proposed framework. In each iteration, the dataset is divided into ten equal subsets, where nine subsets are used for training and the remaining subset is used for testing. This process is repeated ten times, and the average classification accuracy is considered as the final performance measure.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 15, Issue 8, August 2026
DOI: 10.17148/IJARCCE.2026.15841

A Classification Performance

The DCT features effectively capture the frequency characteristics of historical handwritten documents. These features represent changes in handwriting style, ink degradation, and paper aging across different historical periods. The extracted features provide good discrimination between document classes. The SVM classifier uses these features to construct optimal decision boundaries and accurately classify the documents into their respective historical periods. The classification results obtained SVM classifiers are presented in Table 1.

TABLE 1: CLASSIFICATION ACCURACY OF DCT-BASED FEATURES

Classifier	Accuracy (%)
SVM	95.8%

B Discussion

The experimental results demonstrate that DCT-based frequency features are effective for identifying the age of historical handwritten documents. The transformation of document images into the frequency domain provides valuable information about the structural characteristics of manuscripts. The low-frequency DCT coefficients capture general intensity variations and background characteristics of the historical documents, whereas the middle-frequency coefficients represent handwriting strokes, character structures, and texture variations. These characteristics change gradually over different historical periods due to differences in writing practices, ink composition, paper quality, and environmental conditions. The statistical measures extracted from DCT coefficients further improve the representation capability of the features. Mean and variance describe the distribution of frequency components, while energy and entropy capture texture complexity and randomness. Skewness and kurtosis provide additional information about the shape of frequency distributions, which helps distinguish documents from different eras.

C Comparison with Existing Methods

The proposed DCT-based method is compared with existing texture-based approaches for historical document age identification. The results demonstrate that DCT effectively captures frequency-domain characteristics related to handwriting style, paper texture, and document degradation, resulting in competitive classification performance. Unlike spatial-domain descriptors, DCT provides a compact and computationally efficient feature representation while reducing the influence of local noise, making it well suited for historical manuscript classification. Table 2 presents the comparative analysis.

TABLE 2 COMPARISON WITH EXISTING APPROACHES

Method	Classifier	Accura%cy (%)
Gonasagi, et al [8]	SVM	92.8
Proposed method	SVM	95.8%

V. CONCLUSION

This paper proposed a DCT-based frequency-domain approach for automatic age identification of historical handwritten documents. The DCT coefficients effectively capture variations in handwriting patterns, paper texture, and document degradation. Statistical features extracted from DCT coefficients are used with SVM classifiers for document classification. Experimental results on the MPS dataset demonstrate that the proposed method is effective for historical document age identification. The approach is computationally efficient and can support digital preservation and archival analysis. Future work will focus on combining DCT with other feature extraction techniques and deep learning models to improve classification performance.

REFERENCES

[1]. Sheng. He, P. Samara, J. Burgers, and L. Schomaker, "A multiple-label guided clustering algorithm for historical document dating based on writing evolution," Pattern Recognition, vol. 51, pp. 280–297, 2016.
[2]. Sheng. He, P. Samara, J. Burgers, and L. Schomaker, "Historical manuscript dating based on temporal pattern codebooks," Computer Vision and Image Understanding, vol. 152, pp. 167–175, 2016.
[3]. B. Halder, U. Garain, and D. Dutta Majumder, "Color feature-based approach for determining ink age in printed documents," in Proceedings of the 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010, pp. 3212–3215.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁘ Peer-reviewed & Refereed journal ⁘ Vol. 15, Issue 8, August 2026
DOI: 10.17148/IJARCCE.2026.15841

[4]. K. S. Raghunandan, P. Shivakumara, B. J. Navya, G. Pooja, N. Prakash, G. H. Kumar, U. Pal, and T. Lu, "Fourier coefficients for fraud handwritten document classification through age analysis," in Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, 2016, pp. 25–30.
[5]. R. da Silva Barboza, R. Dueire Lins, and D. Marinho de Jesus, "A color-based model to determine the age of documents for forensic purposes," in Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 2013, pp. 1350–1354.
[6]. Sheng. He, P. Samara, J. Burgers, and L. Schomaker, "Historical manuscript dating using Polar Stroke Descriptor," Pattern Recognition Letters, vol. 100, pp. 64–70, 2017.
[7]. B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling image time with content and metadata for historical color photograph dating," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3930–3938.
[8]. P. Gonasagi, S. S. Rumma, and M. Hangarge, "Classification of Historical Documents Based on LBP and LPQ Techniques," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 3, pp. 1533–1538, 2020.
[9]. Gonasagi, Pushpalata, Shivanand Rumma, and Mallikarjun Hangarge. "Texture Analysis Technique for Document Age Identification." In Innovations in Computer Science and Engineering: Proceedings of the Ninth ICICSE, 2021, pp. 331-337. Singapore: Springer Singapore, 2022.
[10]. https://zenodo.org/record/1194357#.XV2E5eMzbIV.

