A
synopsis
of
research work undertaken for the award of Ph.D.
on

AUTHENTICATION OF DOCUMENT IMAGES

By

PARVEEN KUMAR

(Enrollment No. 15911006)

Under the guidance of

Dr. AMBALIKA SHARMA

Department of Electrical Engineering

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE - 247 667 (INDIA)
NOVEMBER, 2019

1 Introduction

The intense development and extensive evolution of multimedia technologies is a current trend, and progressively multimedia data are used to provide authentication of document images. This implies robustness of a digital image against any security breaches. Moreover, it also renders the service of providing authorization and verification of users for distinct type of services worldwide [1–4]. With the concurrent, continuous and progressive improvisations in modern steganalysis techniques, the security of the information stored in digital images is at potential danger [5, 6].

This report gives information about Writer Identification (WI) related problems [7, 8]. The goal of the research is to design authentication systems for predicting writer of a given sample image and answer certain questions regarding handwriting style, representation and performance of the models created by reducing the parameters involved for the same, along with minimizing human intervention in the process. The approaches used in this report raise certain questions in computer vision, for example, whether handwriting style of an individual can be characterized using various algorithms and what features must be used to represent the model and how can they be combined in these models. The proposed models were evaluated on different datasets, the computer algorithms being unaware of what was written on these datasets sample images. These methods have potential to make it feasible for practical applications like in forensic science, banking system, human identification *etc*.

The term WI is used to identify a writer from handwritten images. In most of the languages, various types of symbols and signs have been employed for communication purposes in the present work. Generally, writer identification research is conducted using handwritten document in different languages such as English, Arabic, Devanagari, *etc* [9]. Handwriting-based WI is an active research area in pattern recognition and machine learning. There are many intermediate steps to identify the writers from handwritten documents, which are as follows;

- Design and development of distinct writer recognition methods and algorithms.
- Different feature extraction techniques and methods.
- Identify appropriate characteristics from the handwriting feature set.
- Evaluate handwritten image-based writer identification performance.

Features extraction and segmentation of an image plays significant role in many pattern recognition applications such as handwriting recognition and writer identification. Some of the feature extraction techniques that are used for WI are Global Wavelet-based features [10], Pattern-based features [11], Contour-based Orientation and Curvature-based features [12], Edge Structure Code (ESC) Distribution feature [13], Grid Micro-structure features [14], Curvature-free features [15], *etc*. Feature extraction is used to resize the vector dimension of the feature. If the feature vector is too large to process, it requires to be reduced to a set of features, called the feature vector, in order to improve the model's efficiency. The extracted and selected features have appropriate information of the input data. Further, segmentation is the method of splitting a digital image into several sections (pixels) in image processing [16]. The fundamental objective of segmentation is to simplify or alter the image as required by the user. It is used as a pre-processing technique for many applications of pattern recognition. Several approaches have been developed for authentication of documents images by researchers during last few decades for enhancing authenticity [1, 3]. WI, Signature Verification [17, 18] and Presence of seal in the documents are some of the approaches to provide the authentication of document images [19, 20].

In the era of smart computation, artificial intelligence and machine learning play an important role to simplify the human lives by developing smart devices and systems. However, intrusion and falsification cannot be avoided. To ensure the identification, several recognition systems have already been developed, commercialized and functional at peak. Writer identification is one of the methods to identify the document writer. Statistics show that machine learning methods help in predicting the writer in a better way compared to humans. The advent of deep learning revolutionized the learning and improved the performance of systems exponentially. Though, deep learning computation is expensive, it outperforms the traditional methods. In the present research work, the existing features have been used and new feature extraction techniques have been developed, and these features are used to learn the model based on machine and deep learning to classify the document writer. The designed and developed proposed research scheme is illustrated in Figure 1.

The contributions of the proposed work are as follows:

i. A new model, Histogram Weight Transformation (HWT), is proposed for Writer Identification and Verification (WIV) that provides an authentication of handwritten document images. The model targets the drawbacks of traditional data analysis and

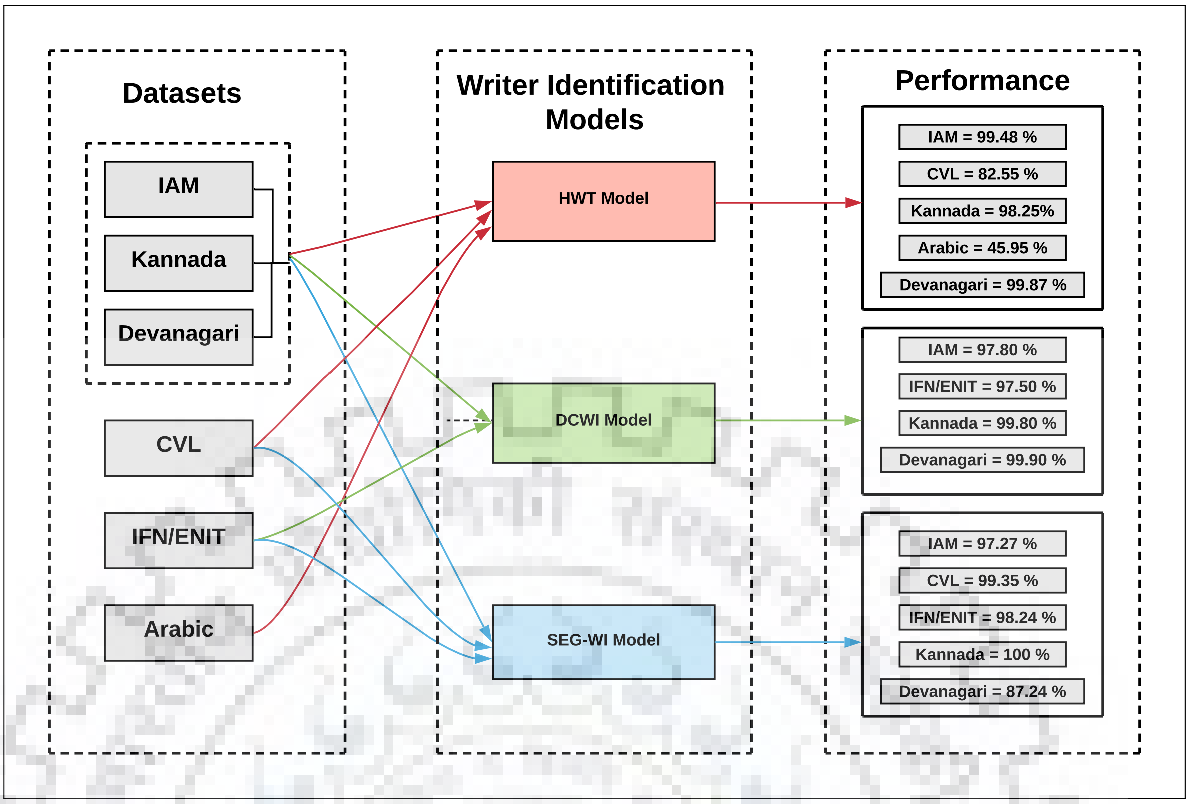

Figure 1: Designed and Developed Proposed Research Scheme.

Histogram Symbolic Representation (HSR) approach. The majority vote technique is adapted to identify the writer of a document having multiple text-lines.

ii. Next, a novel approach of feature extraction based on Distribution Descriptive Curve (DDC) and Cellular Automata (CA) has been presented and utilized in a new developed model to obtain high accuracy compared with recent techniques. Eventually, an efficient model, DCWI, for writer identification has been presented based on DDC and CA.

iii. Furthermore, a Segmentation-free Writer Identification (SEG-WI) model based on CNN is proposed to identify the writer. The region selection mechanism is also developed to improve the overall performance of the model. The lexical similarity between two documents of different writers makes the training difficult, therefore, a new training strategy is suggested to train the model.

2 Motivation

Documents found in a crime scene may contain a pool of unseen evidence, that can be identified by utilizing a scope of techniques and specific equipments. Document forensic experts analyze the authenticity of such documents. A document examiner answers about the reliability and authenticity of a document by utilizing scientific processes. Main aim of the invention of authentication was to provide the robustness of a digital image against any security breach plus it renders the service of providing authorization of users for distinct type of services worldwide. However, in last few decades, researchers have paid significant attention for the development of authentication techniques. Apparently, the evolution in technology such as advent of machine and deep learning techniques helps a lot for the development of new and advance methods and models for the authentication of document images in a more precise way.

3 Research Gaps and Objectives

In present scenario, researchers have worked extensively in the field of authentication of document images, such as documents analysis, writer identification, signature verification, and seal present in the documents. Still a lot of scope exists as several problems are encountered. The following research gaps have been identified from the literature survey carried out on authentication of document images.

i. Document level writer identification has been performed by considering only raw data. However, different features can be extracted to enhance the state-of-the-art techniques. It has been observed that, there is no publicly Indian script dataset available for authentication of the document images. Hence, the work can be carried out on Indian document images.

ii. For document analysis and authentication, signature verification and writer identification play a vital role. However, it has been observed that very few studies have been carried out for document authentication using WI approaches. Therefore, the work can be enhanced in future using different methods.

iii. The fusion of writer identification, signature verification and deep learning concepts and techniques can be applied in order to identify the individual.

On the basis of the above research gaps, the following research objectives have been addressed in this research work. The objectives are as follows:

i. To improve the classification performance in standard datasets using improved kernel, features and classifier combination. By focusing on these three measures *i.e.* better kernel, robust features and combination of multiple classifiers, classification error rate can be minimized. The new feature extraction technique will be introduced to enhance the performance of the models.

ii. The Indian script dataset will be prepared for the authentication of document images and this dataset will be used for writer identification.

iii. Writer identification, signature verification, and deep learning techniques will be applied for document authentication to identify the individual, and to achieve the high authentication rate.

4 Thesis Organization

Thesis is structured in the following manner. Chapter 1 presents the basic concepts of authentication of document images. Chapter 2 discusses a review of the state-of-the-art techniques in authentication and writer identification of document images. Various types of authentication and WI schemes such as biometric authentication, text-dependent and text-independent writer identification, on-line and off-line writer identification, etc. are reported in the literature review. Chapter 3 illustrates the proposed Histogram Weight Transformation (HWT) model based writer identification and verification. Chapter 4 discusses the proposed writer identification based on Distribution Descriptive Curve and Cellular Automata. Chapter 5 describes another proposed method, the Segmentation-free Writer Identification (SEG-WI) based on convolutional neural network and region selection mechanism. Chapter 6 concludes the study and suggests the future scope of the work.

5 Research Work

Handwriting based authentication of an individual is one of the most sought-after biometric security trademarks. Now in the technology era, a person's handwriting-based authentication is continuously used for legitimate, in addition, general authority, official and formal purposes. There is tremendous research done in the handwriting-based Writer Identification (WI) domain, and it is over three decades old. Authentication is essential in terms of digital image security and identification. Writer identification and signature verification are efficient techniques that are designed to provide image authentication. WI is an authentication approach and a well-known research area in the field of pattern recognition and machine learning.

It is outstanding that people naturally use some body attributes, for example, face, step or voice to perceive one another. The utilization of natural physical or behavioral characteristic for identification of persons is known as biometrics. Biometrics has a long custom in forensic and scientific research. Understood instances of natural physical biometrics that are utilized in criminological research are properties of fingerprints, DNA, faces, hands and irises. Instances of behavioral biometrics are signatures, handwriting, writer identification, speech, keystroke intervals and gait. Handwriting is generally used for communication of text. Nonetheless, this does not imply that handwriting is any less helpful as a biometric in criminology and legal sciences. In scientific and forensic WI, handwriting segments are compared for WI of a so-called questioned document. These questioned documents can be fraud letters, compromising letters, threatening letters, misrepresentation letters or suicide notes (*i.e.*, whether the note was written by the alleged suicider or not). Traditionally, human forensic experts perform the prediction of the writer. However, in recent times, WI systems are frequently used to help the forensic expert in recognizing the writer.

This report addresses the issues of Writer Identification and Verification (WIV) using scanned handwritten text images. WIV of handwritten text using image-based techniques is an interesting image processing and pattern recognition technique that has direct application in historic document analysis and forensic fields. There are many approaches of document image authentication like standard cryptography, digital signatures, watermarking, *etc.* However, in the case of WI, a scanned handwritten text document is used to identify the writer. First preprocessing methods such as noise reduction are used to enhance image quality of the handwritten text images. Then, different features are obtained by applying the extraction techniques such as connected components that are used to decrease the size of the vector space. Finally, by applying the classification methods, the writer is anticipated. The various statistical measurements are used to evaluate the system's general efficiency. The results obtained are compared using various similarity measurements.

In this report, Histogram Weight Transformation (HWT) model is proposed for Writer Identification and Verification (WIV) to provide an authentication of handwritten document images. In this work, HWT model targets some drawbacks of traditional data analysis and

Histogram Symbolic Representation (HSR) approach. The HWT model transforms features into histogram weight features. The feature extraction techniques are used in such a way that it improves the performance of the model. The feature vector of extracted features is computed for each text-line and used to develop a histogram transformation model. A weight vector is generated by applying standard transformation over feature vector. These weight vectors for each class are concatenated and a global weight vector is generated for classification and verification. The HWT model classifies each text-line in its respective writer class. The majority vote technique is adapted to identify the writer of a document having multiple text-lines. The proposed model is shown in Figure 2. The experiments are carried out on different datasets for different languages, such as IAM and CVL for English, ICFHR 2012 for Arabic, Kannada and Devanagari (Hindi) for Indic. The proposed model, HWT, improves the performance in terms of accuracy as compared with the state-of-the-art.

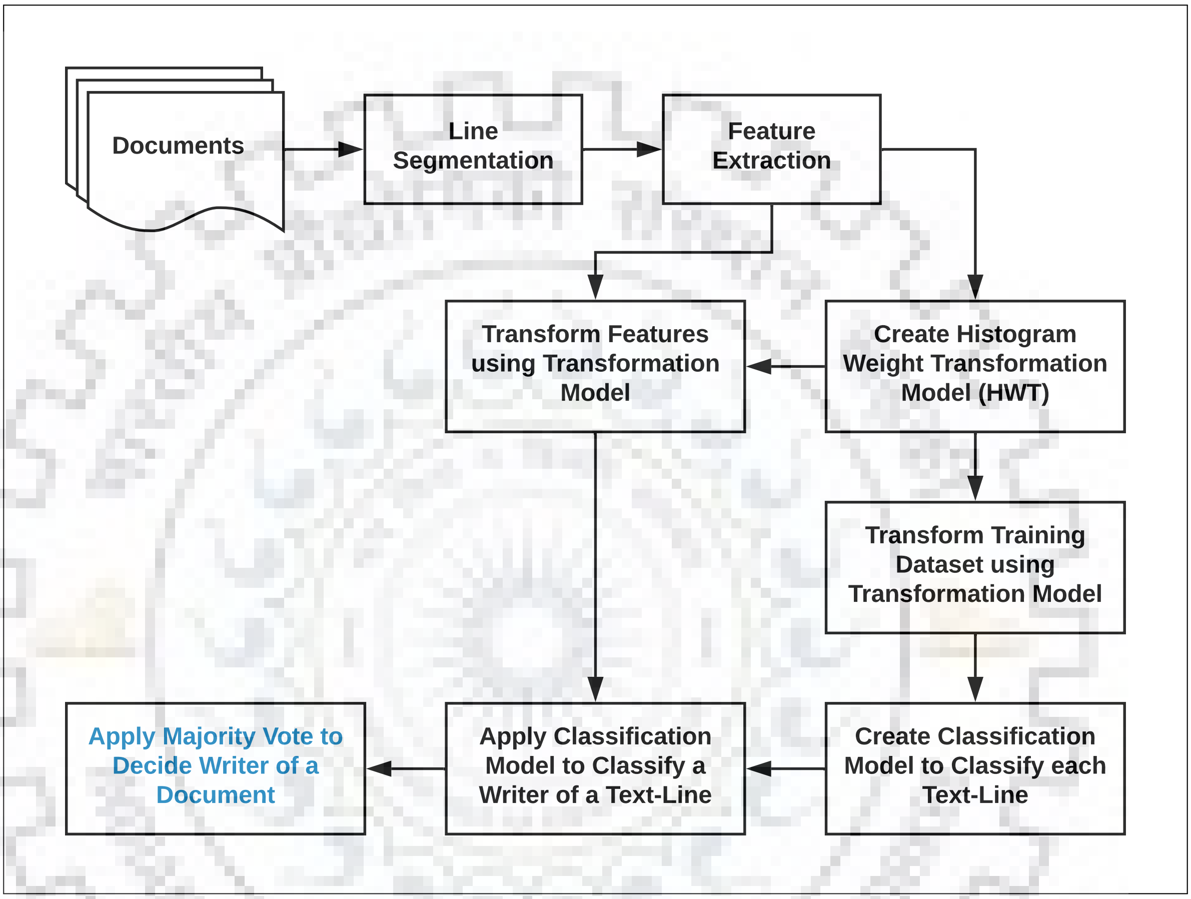
Documents
Line Segmentation
Feature Extraction
Transform Features using Transformation Model
Create Histogram Weight Transformation Model (HWT)
Transform Training Dataset using Transformation Model
Apply Majority Vote to Decide Writer of a Document
Apply Classification Model to Classify a Writer of a Text-Line
Create Classification Model to Classify each Text-Line

Figure 2: Writer Identification Modeling Framework (HWT).

In the domain of pattern recognition, the extraction of discriminative features of different writers has become very challenging. In order to address this concern, unique feature extraction techniques are further proposed based on Distribution Descriptive Curve (DDC)

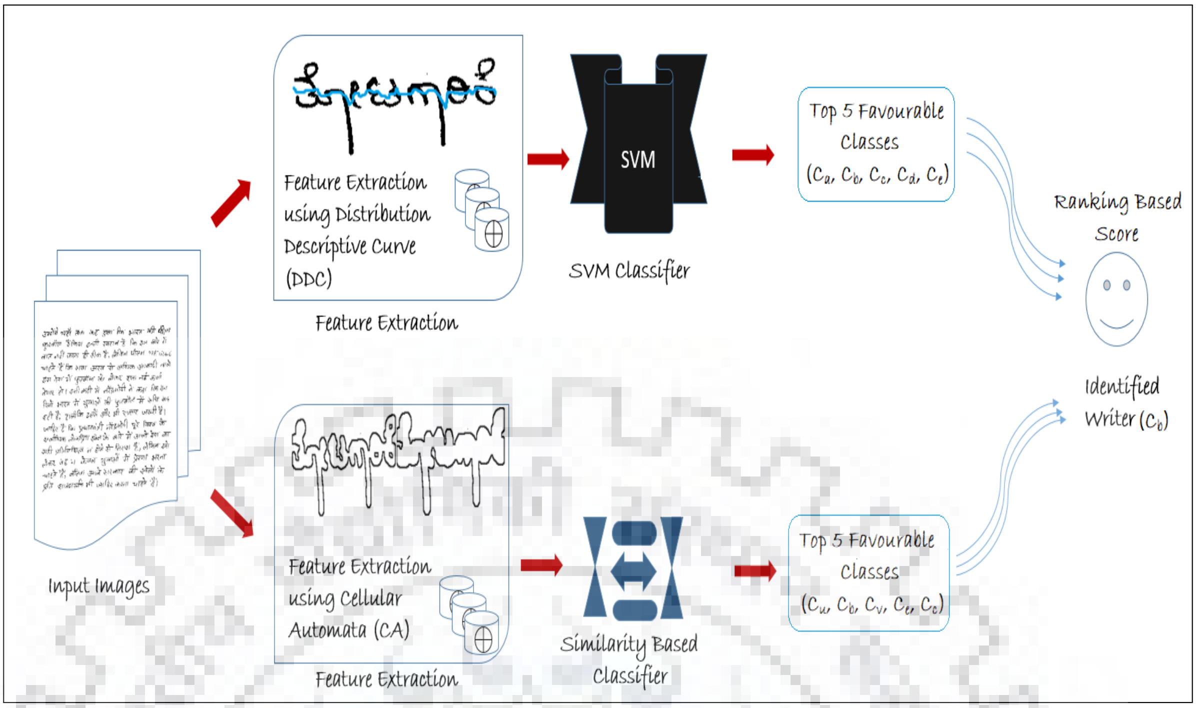
Feature Extraction using Distribution Descriptive Curve (DDC)
Feature Extraction
SVM
SVM Classifier
Top 5 Favourable Classes (C_a, C_b, C_c, C_d, C_e)
Ranking Based Score
Identified Writer (C_b)
Input Images
Feature Extraction using Cellular Automata (CA)
Feature Extraction
Similarity Based Classifier
Top 5 Favourable Classes (C_u, C_b, C_v, C_e, C_c)

Figure 3: Proposed DCWI Model for Writer Identification.

and Cellular Automata (CA). The DDC technique utilizes the idea of pixel distribution of handwritten text images to generate a unique curve as a feature vector. The generated feature vector is then fed to a Support Vector Machine (SVM) as an input to identify the writer. Simultaneously, in a parallel mode, the initial handwritten text images are processed repeatedly with CA to generate another set of feature vectors. This new set of generated feature vectors are fed to a Similarity-Based Classifier (SBC) as an input. The writer is predicted on the basis of the similarity of the features. The results from both of the approaches (DDC + SVM and CA + SBC) are merged to improve the performance of the model. This proposed model, DCWI, has better writer identification capabilities compared with recent techniques [7, 8]. Eventually, WI is accomplished using the ranking-based score scheme. The complete procedure of the proposed model is illustrated in Figure 3. The proposed model is evaluated on different datasets, *e.g.*, IAM for English, IFN/ENIT for Arabic, and Kannada and Devanagari (Hindi) for the Indic script. The results show that the proposed model has better performance compared to recent techniques.

Handwriting recognition is one of the desired aspects of document understanding and analysis. It deals with the writing style of the document and learns the features which differentiate the writers. As a further extension of the work, Segmentation-free Writer Identification (SEG-WI) model is proposed, based on Convolution Neural Network (CNN) and a weakly

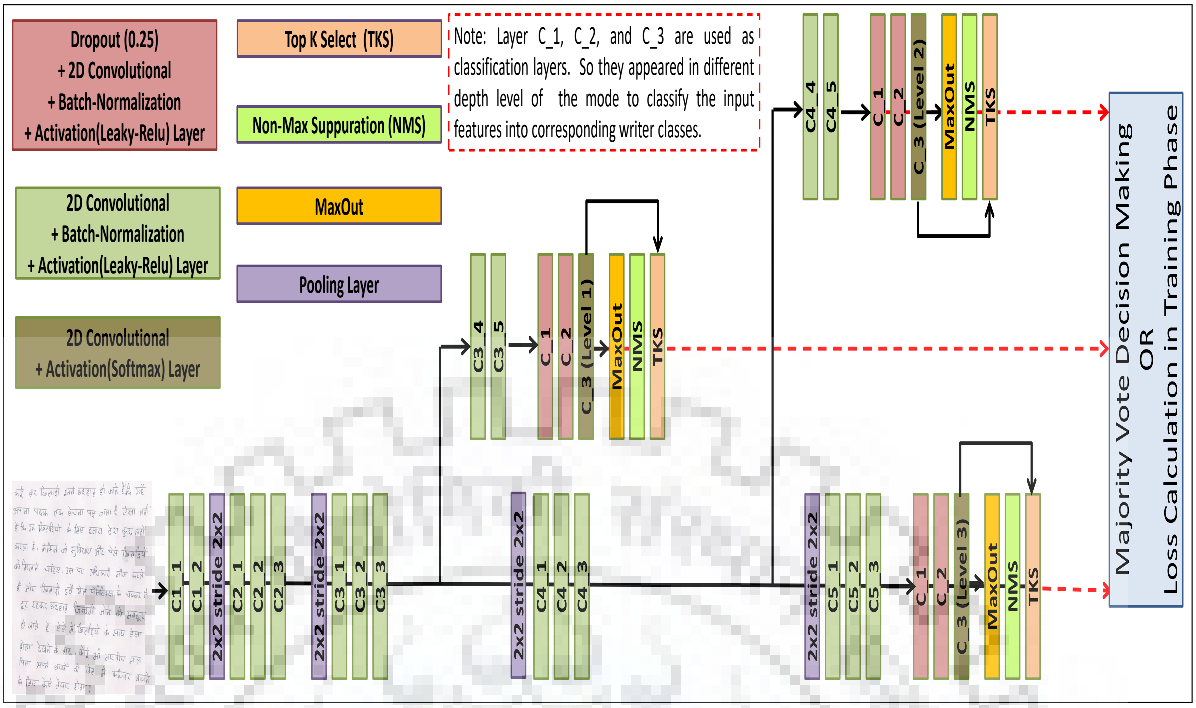
Dropout (0.25) + 2D Convolutional + Batch-Normalization + Activation(Leaky-Relu) Layer
Top K Select (TKS)
Non-Max Suppuration (NMS)
Note: Layer C_1, C_2, and C_3 are used as classification layers. So they appeared in different depth level of the mode to classify the input features into corresponding writer classes.
2D Convolutional + Batch-Normalization + Activation(Leaky-Relu) Layer
MaxOut
Pooling Layer
2D Convolutional + Activation(Softmax) Layer
C1_1
C1_2
2x2 stride 2x2
C2_1
C2_2
C2_3
2x2 stride 2x2
C3_1
C3_2
C3_3
C3_4
C3_5
C_1
C_2
C_3 (Level 1)
MaxOut
NMS
TKS
2x2 stride 2x2
C4_1
C4_2
C4_3
C4_4
C4_5
C_1
C_2
C_3 (Level 2)
MaxOut
NMS
TKS
2x2 stride 2x2
C5_1
C5_2
C5_3
C_1
C_2
C_3 (Level 3)
MaxOut
NMS
TKS
Majority Vote Decision Making OR Loss Calculation in Training Phase

Figure 4: Description of the Proposed Model (SEG-WI, CNN based).

supervised region selection mechanism. The model, SEG-WI, takes an unsegmented text document and produces the writer-ID with the region probability map. The probability vectors at each cell location in the input document constitute a region probability map. To achieve the best performance, the SEG-WI model selects top 10% to 50% cell regions for decision making and a voting mechanism among the selected regions is used to identify the writer. The proposed model is depicted in Figure 4. The model is evaluated on different datasets such as IAM, and CVL for English, IFN/ENIT for Arabic, Kannada and Devanagari for Indic and outperforms compared with the recent techniques [15, 21]. Moreover, a comparative analysis of the proposed model with and without the region selection is performed to validate the effect of region selection mechanism and the results show that the region selection mechanism improves the performance in terms of accuracy of the model.

6 Conclusions

This work presents a step towards efficient, effective and robust writer identification techniques. The off-line writer identification techniques have been proposed based on handwritten text images. The proposed techniques work efficiently on different types of languages and datasets. First, a Histogram Weight Transformation (HWT) model based writer identification and verification technique using weight vectors with SVM to utilize the inner-class information is presented. The efficient discriminative features of proposed transformed weights are analyzed and used for different datasets such as IAM, CVL, ICFHR, Kannada, and Devanagari for writer identification. A comparative evaluation of HWT model on different datasets of different languages has been performed. The results show a substantial gain in the performance of proposed model by using this transformation model. These comparative results illustrate the effectiveness of HWT model in comparison with recent techniques.

Next, DDC and CA based model was designed and developed for writer identification. The DDC and CA are used to extract the features from handwritten text images, and then the writer is predicted using SVM and SBC classifiers. The proposed model, DCWI, merges the results from both the SVM and SBC classifiers. Finally, using the ranking-based score scheme, writer identification is achieved. A comparative evaluation and analysis of DCWI using different datasets for different languages was carried out. The results obtained by DCWI show that there is significant improvement in performance compared with existing state-of-the-art techniques.

Extending further, a segmentation-free deep convolution neural network model for writer identification has also been presented in this research work. The region selection mechanism and training schemes are also introduced to improve overall performance of the model. A comparative evaluation of SEG-WI model is performed for different datasets of different languages. The comparative results show the robustness of SEG-WI model against current techniques.

In essence, an attempt has been made to successfully evolve writer identification and authentication techniques that would be of great help in present scenario of complex and unrecognizable human identification.

Author’s Publications

Refereed Journals

1. **Parveen Kumar** and Ambalika Sharma. “DCWI: Distribution descriptive curve and Cellular automata based Writer Identification”. Expert Systems with Applications (Elsevier) 128 (2019): 187-200. (Published)

2. **Parveen Kumar** and Ambalika Sharma. “Histogram weight transformation model based writer identification and verification”. (The Computer Journal, Oxford University Press) (Under Review).

3. **Parveen Kumar** and Ambalika Sharma. “SEG-WI: SEGmentation-free Writer Identification based on convolutional neural network and region selection mechanism”. (Computers and Electrical Engineering, Journal - Elsevier) (Under Review).

Conferences

4. **Parveen Kumar**, Manu Gupta, Mayank Gupta and Ambalika Sharma. ”Profession Identification using Handwritten Text Images.” International Conference on Computer Vision & Image Processing (CVIP-2019, 2019) (Presented).

5. **Parveen Kumar**, Mohd Haroon Ansari and Ambalika Sharma. ”MBC-CA: Multithreshold Binary Conversion based salt-and-pepper noise removal using Cellular Automata.” International Conference on Computer Vision & Image Processing (CVIP-2019, 2019) (Presented).

References

[1] A. Haouzia and R. Noumeir, “Methods for image authentication: a survey,” *Multimedia tools and applications*, vol. 39, no. 1, pp. 1–46, 2008.

[2] F. Monrose and A. D. Rubin, “Keystroke dynamics as a biometric for authentication,” *Future Generation computer systems*, vol. 16, no. 4, pp. 351–359, 2000.

[3] A. Mitra, P. Kumar, and C. Ardil, “Automatic authentication of handwritten documents via low density pixel measurements,” *International Journal of Computational Inteligence*, vol. 2, no. 4, pp. 219–223, 2005.

[4] J. Kodovskỳ, T. Pevnỳ, and J. Fridrich, “Modern steganalysis can detect yass,” in *Media Forensics and Security II*, vol. 7541. International Society for Optics and Photonics, 2010, p. 754102.

[5] H. Ge, M. Huang, and Q. Wang, “Steganography and steganalysis based on digital image,” in *2011 4th International Congress on Image and Signal Processing*, vol. 1. IEEE, 2011, pp. 252–255.

[6] N. F. Johnson and S. Jajodia, “Steganalysis: The investigation of hidden information,” in *1998 IEEE Information Technology Conference, Information Environment for the Future (Cat. No. 98EX228)*. IEEE, 1998, pp. 113–116.

[7] Y. Hannad, I. Siddiqi, and M. E. Y. El Kettani, “Writer identification using texture descriptors of handwritten fragments,” *Expert Systems with Applications*, vol. 47, pp. 14–22, 2016.

[8] A. Chahi, Y. Ruichek, R. Touahni *et al.*, “Block wise local binary count for off-line text-independent writer identification,” *Expert Systems with Applications*, vol. 93, pp. 1–14, 2018.

[9] P. Kumar and A. Sharma, “Dcwi: Distribution descriptive curve and cellular automata based writer identification,” *Expert Systems with Applications*, 2019.

[10] Z. He, X. You, and Y. Y. Tang, “Writer identification using global wavelet-based features,” *Neurocomputing*, vol. 71, no. 10-12, pp. 1832–1841, 2008.

[11] B. Helli and M. E. Moghaddam, “A text-independent persian writer identification based on feature relation graph (FRG),” *Pattern Recognition*, vol. 43, no. 6, pp. 2199–2209, 2010.

[12] I. Siddiqi and N. Vincent, “Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features,” *Pattern Recognition*, vol. 43, no. 11, pp. 3853–3865, 2010.

[13] J. Wen, B. Fang, J. Chen, Y. Tang, and H. Chen, “Fragmented edge structure coding for chinese writer identification,” *Neurocomputing*, vol. 86, pp. 45–51, 2012.

[14] X. Li and X. Ding, “Writer identification of chinese handwriting using grid microstructure feature,” in *International Conference on Biometrics*. Springer, 2009, pp. 1230–1239.

[15] S. He and L. Schomaker, “Writer identification using curvature-free features,” *Pattern Recognition*, vol. 63, pp. 451–464, 2017.

[16] E. Cermeño, S. Mallor, and J. A. Sigüenza, “Offline handwriting segmentation for writer identification,” in *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*. IEEE, 2014, pp. 13–17.

[17] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger, “Signature detection and matching for document image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2015–2031, 2008.

[18] R. Mandal, P. P. Roy, and U. Pal, “Signature segmentation from machine printed documents using conditional random field,” in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 1170–1174.

[19] P. Forczmański, “Stamp detection in scanned documents,” *Annales Universitatis Mariae Curie-Sklodowska, sectio AI–Informatica*, vol. 10, no. 1, 2010.

[20] B. Micenková, J. van Beusekom, and F. Shafait, “Stamp verification for automated document authentication,” in *Computational Forensics*. Springer, 2012, pp. 117–129.

[21] M. Kumar, M. Jindal, R. Sharma, and S. R. Jindal, “A novel framework for writer identification based on pre-segmented gurmukhi characters,” *Sādhanā*, vol. 43, no. 12, p. 197, 2018.