ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Offline Writer Recognition for Kurdish Handwritten Text Document Based on Proposed Codebook

Twana Latif Mohammed[1]*, Ahmed Abdullah Ahmed[2]

[1]Department of Information Technology, Technical College of Informatics, Sulaimani Polytechnic University, Sulaymaniyah, Kurdistan Region, Iraq, [2]Department of Software Engineering, Faculty of Engineering and Computer Science, Qaiwan International University (QIU)/Raparin, Sulaymaniyah, Kurdistan Region, Iraq

ABSTRACT

Handwritten text recognition has been an ongoing attractive task to research in the field of document analysis and recognition with applications in handwriting forensics, paleography, document examination, and handwriting recognition. In the present research, an automatic method of writer recognition is presented using digitized images of unconstrained texts. Despite the increasing efforts by prior literature on the different methods used for the same purpose, such methods performance, particularly their accuracy, has not been promising, leaving plenty of room for improvements. This method made use of codebook-based writer characterization, with each writing sample represented by a group of computed features from a primary and secondary codebook. The writings were then represented through the computation of the probability of codebook patterns occurrence, and the probability distribution was employed for each writer's characterization. Writer identification process involved comparing two writings through the computation of the distances between their respective probability distribution. The study carried out experiments to determine the performance of the implemented method in light of rates of identification with the help of standard datasets, namely, KRDOH and IAM, the former being the most current and largest Kurdish handwritten datasets with 1076 writers, and the latter being a dataset containing 650 writers. The outcome of the experiments was promising with a rate of identification of 94.3%, with the proposed method outperforming the state-of-the-art methods by 2–3%.

Index Terms: Writer Identification, Feature Extraction, Text Independent, Codebooks, Feature Combination

1. INTRODUCTION

An individual is distinguished from another through the distinct identities that he possesses. Such identities may be physical or behavioral and they can be employed to identify individuals in a scientific field called biometrics. Biometrics plays a key role in researches dedicated to forensic science, where forensic experts make use of physical or behavioral biometrics to recognize and identify individuals. Physical biometrics includes DNA as illustrated in the study by Holland and Parsons [1], fingerprints as presented by Abu-Faraj *et al.* [2], ear prints [3], [4], irises [5], and soft and hard tissues as illustrated in the study by Zewail *et al.* [6]. Examples of behavioral biometrics are speech [7], [8] and gait [9], [10]. This also includes keystroke intervals as in Delac and Grgic [11], and signatures, and handwriting. In this thesis, the researcher focuses and examines handwriting.

Whether handwriting is alphabetical or pictographic based, it has been employed as a significant communication means

Access this article online

DOI:10.21928/uhdjst.v5n1y2021.pp21-27	E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Twana Latif Mohammed, Technical College of Informatics, Sulaimani Polytechnic University, Sulaymaniyah, Kurdistan Region, Iraq. E-mail: twana.latif@spu.edu.iq

Received: 19-10-2020 Accepted: 27-03-2021 Published: 31-03-2021

from the beginning of time and has made certain evolutions. According to Huber and Headrick [12], writing styles are developed based on local culture, geographical location, historical background, and temporal situations. The earlier handwriting record keeping came from China, dated around 2000 years ago, following the invention of the first inks and papers. During that time, early handwritings and the following handwriting were written in some standard writing models. In general, writers have a tendency not to follow standard writing models, and thus, handwritings show deviation. Individual writer characteristics are invaluable in distinguishing one writer from another.

2. LITERATURE REVIEW

This section presents a comprehensive review of the techniques developed for writer identification on offline writing samples, and this is one of the main topics addressed in the research. The section presents an extensive review of offline handwritten datasets, outlines the existing methods on text-dependent writer identification methods while the subsection discusses the significant contributions to the text-independent writer recognition domain which is also the primary focus of this research. Finally, the last part of this section conducts a comparative analysis among the methods with detailed presentations of their performances.

A recent study [13] uses a simplified and rapid methodology by avoiding character appearances and making a distinction from approaches similar to those in traditional codebook-based methods. Here, researchers present new descriptors that are derived from different scales of geometrical interest points. This is achieved by documenting the geometric associations between parts of the script, such as strokes, loops, endings, and junctions. These descriptors are easier to use, more effective, provide better results on unseen datasets, and reduce processing time dramatically over existing methods. In addition to these benefits, this method has a drawback in terms of the amount of data needed to build a model with consistent results.

Some researchers in Al-Maadeed *et al.* [14] showed the identification of various writers using their proposed collection of curvature, direction, and tortuosity-based geometrical features. They also suggested improvement of edge-based directional features using a filled moving window instead of an edge moving window alongside chain code-based features using a fourth-order chain code list for enhancing its recognizing ability. This method was tested in the handwriting databases of IAM and QUWI. In addition, the authors [15] proposed a new method called (DLS_CNN) for writer recognition so, in this research used the combination between neural network (NN) with line segmentation. On the other hand in Chahi *et al.* [16], the authors proposed a new algorithm called LSTP but at the classification step base on NN achieved Hamming distance.

While great interest and substantial progress have been observed in the field of handwriting based biometrics and its applications [17], the identification of writers is based on relatively complex scripts, that is, Arabic [18] and Chinese [19] which remains a less investigated area [20], rare comparable between each of the scripts among researchers have resulted in vague results which are not proportional to its widespread use. Most of the researches in this area share the same purpose of determining a script's authorship through the acquisition of individual handwriting characteristics. In the current process, all document features are found and created, after which the feature vector distance is compared between the query and the library image. Nevertheless, this performance is considered far from being achieved, and it is computationally very costly, particularly a significant problem in document image analysis and retrieval is the search for the relevant document from large and complex document image repositories. Apart from issues of database size (scale), there is a problem of data heterogeneity. The smooth incorporation of such techniques with the current knowledge in forensic handwriting is still unknown. Such approaches are not obsolete and are still used today. Some researchers in the same field have used the principal component analysis method to extract the most important information. The data performed as testing and training on the grayscale's images like 12,500 images [21].

Ghiasi and Safabakhsh [22] generated feature vectors for each manuscript by taking advantage of the normalized and resampled contours of connected segments. Then, for writer recognition, they used these feature vectors to form a codebook. They also solve cursive handwriting, the method utilizes the occurrence histogram of the shapes in a codebook, connected complements can be too long and may have a wide range of shapes. To prevent complex patterns, the authors used small fragments of the connected components and implemented two effective methods for extracting code from contours. One of the techniques uses the actual pixel coordinates of contoured fragments, while the other utilizes linear piecewise approximation using segment angles and lengths and removes some of the unnecessary information. It helps to identify and group similar shapes. The shorter length of this code allows it to be applied faster and also helps the quicker generation of the codebook. The authors tested this code on two English databases and one Persian and found better performance than other contemporary techniques in 2013. It may be quicker to generate codebooks, but the computational times of this technique are long.

3. METHODOLOGY

In this study, the author brings forward a methodology characterizing the writer through the division of the text into smaller fragments, and potential clusters are searched within it. This is based on premise that writer recognition is related to the physical stroke's generation by the writer. Rather than dividing the writing in graphemes, the method divides them into small fragments that are enough to be utilized in writer recognition. Further detailed elaboration of the modules is provided in the next sections.

3.1. Binarization

Under this process, the digitized document images are scanned in the form of grayscale images. The research addresses handwritten scanned documents, and as such, there are two objects to the image, namely, handwriting and background, and in this research, the primary object of interest is handwriting. Therefore, handwriting is separated from the background through the use of binarization that is categorized into two classes. The study employs the popular Otsu's thresholding logarithm (known as bae benchmark) for the calculation of the threshold from the grayscale image. A grayscale original image along with its binarized version is demonstrated in Fig. 1.

3.2. Componentization of Writing

Before handwriting fragmentation, the applied method entails the division of handwriting into related components in what is known as componentization. This forms clusters of the entire related black pixels based on their connectivity. In an individual connected component, the pixels adjacencies are gauged through the use of 8-side connectivity, after which they are labeled with sequential numbers. For every pixel with the same label, a component obtained from the image is highlighted for their fragmentation. The connecting image components are depicted in Fig. 2.

3.3. Fragmentation of Components

This study brings forward a writer characterization method through a specific sample by examining small invariant fragments and exploiting the writing redundancy. The step entailing the division of handwriting into fragments is a significant one in the applied method, and it considers them along with the adjacent fragments connected to them. More specifically, the adjacent fragments are acquired through the writing division into windows after which, the main and adjacent fragments are categorized into individual codebooks for the ultimate writer characterization method.

3.4. Feature Extraction

It is the comparison among the fragments through pattern matching or by representing them with a set of features. Although pattern matching is a simple process, it calls for maintaining the fragment's pixel values, otherwise, the comparison outcome may lose its robustness to both noise and distortions. A comparison of features mitigates the representation space of the dimension but it is susceptible to distortions. Thus, the applied method represents each writing fragments (main and adjacent) using a set of features including vertical and horizontal projections, upper and lower profiles along with a group of familiar shape descriptors (i.e., elongation, solidity, rectangularity, orientation, and perimeter).

3.4.1. Horizontal and vertical projections

Projections provide the number of black pixels present in the fragmented image, within each row and column. More specifically, the horizontal projection is produced by determining the number of black pixels in every column of the image, while the vertical projection is produced by determining the number of black pixels in every role of the same.

3.4.2. Upper and lower profile

For the upper and lower profile, the former is described as the distance of the first black pixel from the top of every fragmented image, while the latter is the distance of the first black pixel from the bottom of every fragment. Both upper and lower fragment profiles are calculated by determining the column of the fragment and the distance between the upper black pixels to the lower one.

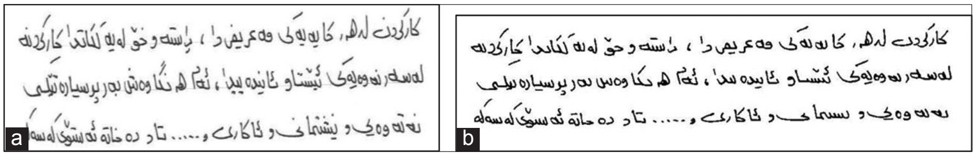


Fig. 1. Image binarization. (a) Grayscale handwriting image before binarization, (b) image after binarization.

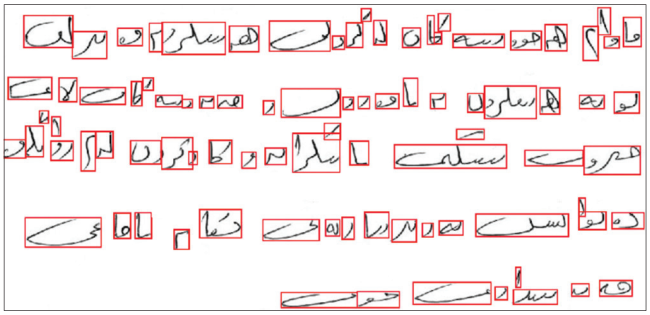


Fig. 2. Bounding box of connected components.

3.4.3. Orientation

The direction of a stroke (or its slope) in a fragmented image is calculated through its orientation feature, specifically by the angle between the X-axis and the major axis of an ellipse that approximates the fragment. It is evident from Fig. 3. 18b that an ellipse comprises a group of points that move around the black pixels of the fragmented stroke, whose sum constitutes the distance from two fixed points, namely, F1 and F2, and it remains constant.

3.4.4. Rectangularity

This feature refers to the ratio of the object area to the bounding box area, the latter of which is the smallest rectangle encapsulating the writing shape in a fragment. Rectangularity is mathematically defined as follows:

$$Rec\tan gularity = \frac{A_{FW}}{A_{BB}} \tag{1}$$

In the above equation, A_{FW} denotes the number of pixels in the fragmented window area, while A_{BB} denotes the bounding box area containing the stroke region.

3.4.5. Elongation

This feature refers to the ratio between the bounding box height and its width. A bounding box was obtained from a fragment enclosing a stroke. The stroke elongation is mathematically represented by the following equation:

$$Elongation = \frac{l_b}{s_b} \tag{2}$$

From the above equation, l_b represents the bounding box longer side and the S_b represents the bounding box shorter side.

3.4.6. Perimeter R

This feature represents the shape boundary's total length. More specifically, the boundary of the shape comprises a group of pixels in the boundary having a non-shape pixel as an adjacent pixel. Mathematically, the perimeter can be calculated by tracking the stroke's boundary pixel after which the steps are summed up.

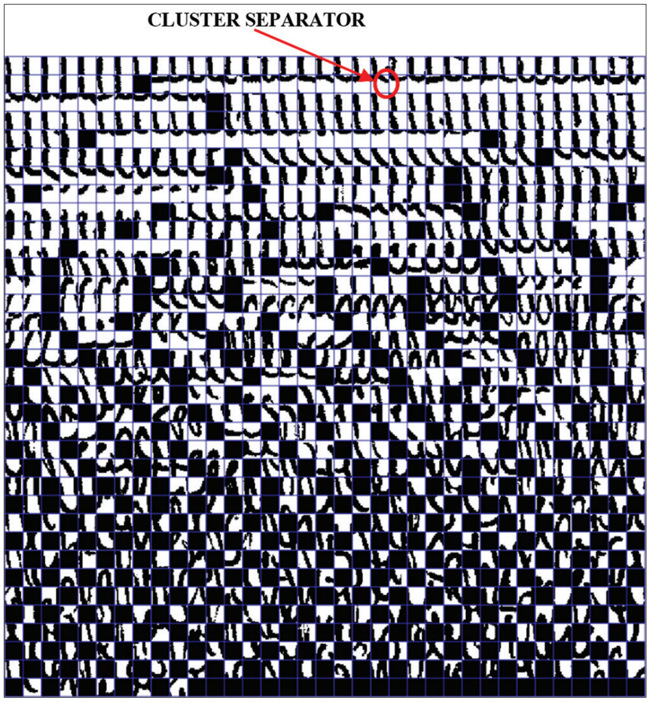


Fig. 3. Primary codebook obtained from the main fragmented windows on a writing sample (Sample: W0010Para2).

3.4.7. Solidity

This feature is useful in measuring the fragment's density and is calculated as the ratio between the fragment areas and is corresponding to convex. The solidity value ranges from 0 to 1 and a solidity value that is near to 0 depicts an irregular object, while that is near to 1 is a solid one. Solidity can be mathematically represented as:

$$Solidity = \frac{A_{FW}}{A_{CR}} \tag{3}$$

In the above equation, A_{FW} denotes the fragment area, while A_{CR} denotes the convex region area.

3.5. Clustering of Fragments

The present section proceeds to present the grouping of similar fragments, extracted through the use of main and adjacent windows, into clusters referred to as codebook. The features are used to make clusters, in that closely related fragmented patterns are clustered together to make a class. In each class, patterns are distinct from those in other classes, and in each cluster, every individual class contains a group of invariant writing patterns. The implemented method produces two distinct cluster sets, by matching the invariant patterns features. The entire main strokes extracted through the use of main windows are clustered to develop a primary cluster, while the entire adjacent windows fragments are clustered to develop a secondary cluster. Figs. 3 and 4 illustrate the primary and secondary codebooks generated from the main and adjacent fragmented windows.

4. RESULTS AND DISCUSSION

This section discusses the experimental evaluation of the applied technique. Many experiments were performed to evaluate the performance of the system and study the sensitivity of the performance to different parameters. Since two codebooks, primary and secondary, have been presented, writer recognition results on each of these codebooks are presented. All experiments are conducted on the latest, largest, and standard Kurdish handwritten documents database known as KRDOH [23]. Moreover, the implemented method is also benchmark against the best and up-to-date methods found in the literature of writer identification that has used IAM [24] data set. Sample forms from the database are shown in Fig. 5.

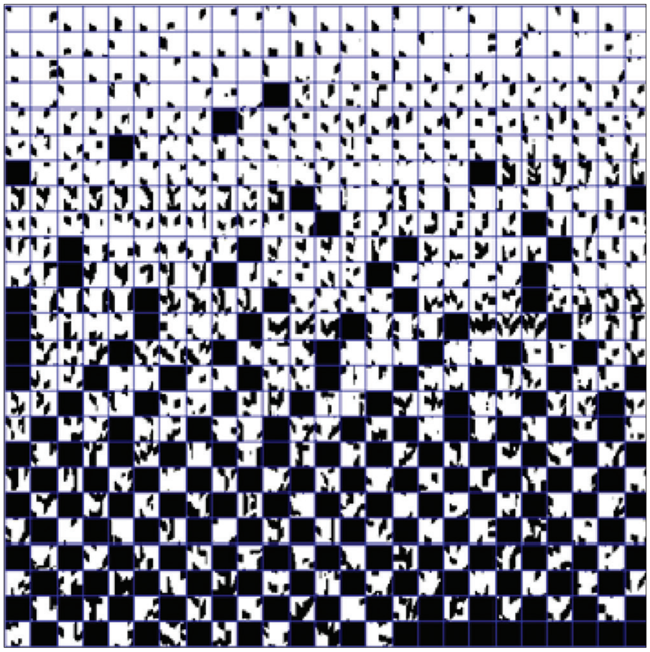
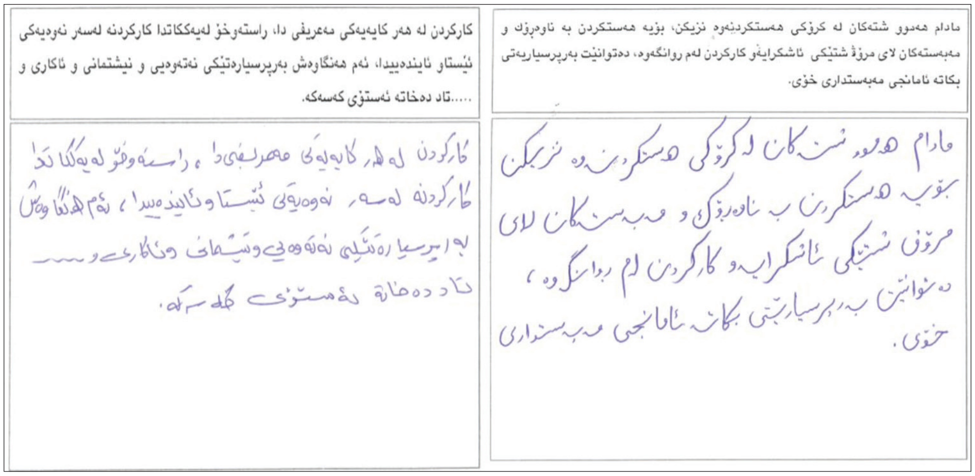


Fig. 4. Secondary codebook obtained from the adjacent fragmented windows on a writing sample (Sample: W0010Para2).

This study first evaluates the performance of primary and secondary codebook separately and then merges both codebooks. Initial experiments were conducted on 210 random writers from the KRDOH dataset. Table 1 summarizes the results of a primary codebook, secondary codebook, and merged codebooks. Using the primary codebook, an identification rate of 87.14% (Top-5: 91.03% and Top-10: 94.08%) is achieved with an EER of 5.92%. The secondary codebook achieves slightly better identification rate of 89.26% (Top-5: 92.14% and Top-10: 96.17%) with 3.83% EER. By merging the two codebooks, the overall identification rate is increased to 91.87% (Top-5: 93.3% and Top-10: 97.6%) and EER drops to 2.4%.

Later, Table 2 provides a performance comparison of the latest writer identification techniques. Oriented basic image features and the concept of graphemes codebook were employed by Durou *et al.*, 2019 [25], achieved 92% identification rate on the IAM dataset. Later (Nguyen *et al.*, 2019) [26], the author used a CNN-based method for text-independent writer identification on the



کارکردن لە هەر کایەیەکی مەعریفی دا، راستەوخۆ لەیەککاتدا کارکردنە لەسەر نەوەیەکی
ئێستاو ئایندەییدا، ئەم هەنگاوەش بەرپرسیارەتێکی نەتەوەیی و نیشتمانی و ئاکاری و
.....تاد دەخاتە ئەستۆی کەسەکە.

مادام هەموو شتەکان لە کرۆکی هەستکردنەوە نزیکن، بۆیە هەستکردن بە ناوەڕۆک و
مەبەستەکان لای مرۆڤ شتێکی ئاشکرایەو کارکردن لەم روانگەوە، دەتوانێت بەرپرسیاریەتی
بکاتە ئامانجی مەبەستداری خۆی.

Fig. 5. Examples of the scanned forms of KRDOH dataset.

TABLE 1: Applied method results on 210 writers from KRDOH dataset

Mission	Identification			Verification
Codebook	Top 1 (%)	Top 5 (%)	Top 10 (%)	EER (%)
Primary	87.14	91.03	94.08	5.92
Secondary	89.26	92.14	96.17	3.83
Merge	91.87	93.3	97.6	2.4

TABLE 2: Performance comparison of writer identification methods

Authors	Year	Dataset	Writers	Performance (%)
Durou *et al.*	2019	IAM	650	92
Nguyen *et al.*	2019	IAM	650	91.81
Proposed method	2020	IAM	650	94.37

same database of offline handwritten English text and achieved 90.12% identification rate. Using the same 650 sets of writers of the IAM dataset, the proposed study achieved an identification rate of 94.37% which is the best identification rate on this dataset so far using the writer-specific codebook technique.

5. CONCLUSION

This research primary aims to apply and test an automatic writer recognition method on offline Kurdish handwritten text. Such objective was achieved through the implementation of a new method addressing the issues of state-of-the-art methods and outperforming them. The approach involved the extraction of small writing fragments through the positioning of windows over the writing and clustering writing fragments that are similar, forming a codebook. In contrast to classical methods that produce a codebook of graphemes, the codebook of small writing fragments is script independent and is applicable to text of different languages. Moreover, the applied method extracts main strokes along with the adjacent strokes linked to the former and both fragments (main and adjacent) are separately clustered to generate the primary and the secondary codebooks. Following the generation of the codebooks, each writing sample is represented as a probability patterns distribution in them, after which two writings are compared by calculating the distance between their respective codebooks. Standard datasets (IAM and KRDOH) using both the codebooks (primary and secondary) and their integration reflect the optimal performance of the approach over the existing approaches.

REFERENCES

[1] M. M. Holland and T. J. Parsons. “Mitochondrial DNA sequence analysis validation and use for forensic casework”. *Forensic Sci. Rev.*, vol. 11, no. 1, pp. 21-50, 1999.

[2] Z. Abu-faraj, D. P. A. Atie, K. Chebaklo, S. Member and Z. E. Khoukaz. “*Fingerprint Identification Software for Forensic Applications*”. Electronics, Circuits and Systems, 2000. ICECS 2000. The 7th IEEE International Conferenceno. May, 2010.

[3] S. Black and T. J. U. Thompson. “*Body Modification*”. CRC Press, Boca Raton, 2007.

[4] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld. “Face recognition: A literature survey”. *ACM Computing Surveys*, vol. 35, no. 4, pp. 399-458, 2003.

[5] J. Daugman. “How iris recognition works”. *IEEE Journal*, vol. 14, no. 1, pp. 21-30, 2004.

[6] R. Zewail, A. Elsafi, M. Saeb and N. Hamdy. “*Soft and Hard Biometrics Fusion for Improved Identity Verification*”. The 2004 47th Midwest Symposium on Circuits and Systems, pp. 225-228, 2004.

[7] C. Champod and D. Meuwly. “The inference of identity in forensic speaker recognition”. *Speech Communication*, vol. 31, pp. 193-203, 2000.

[8] G. R. Joaquin and D. Ramos. Forensic automatic speaker classification in the coming paradigm shift. *In: Speaker Classification I. Springer, Berlin, Heidelberg*, pp. 205-217, 2007.

[9] J. K. Aggarwal and Q. Cai. “Human motion analysis: A review”. *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428-440, 1999.

[10] M. G. Grant, J. D. Shutler, M. S. Nixon and J. N. Carter. “*Analysis of a Human Extraction System for Deploying Gait Biometrics*”. 6th IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 46-50, 2004.

[11] K. Delac and M. Grgic. “*A Survey of Biometric Recognition Methods*”. 46th International Symposium Electronics in Marineno, pp. 16-18, 2004.

[12] R. A. Huber and A. M. Headrick. “*Handwriting identification: Facts and fundamentals*”. CRC Press, Boca Raton, Florida, 1999.

[13] A. Garz, M. Würsch and A. Fischer. “*Simple and Fast Geometrical Descriptors for Writer Identification*”. Society for Imaging Science and Technology, Springfield, Virginia, pp. 1-12, 2016.

[14] S. Al-Maadeed, A. Hassaine, A. Bouridane and M. A. Tahir. “Novel geometric features for off-line writer identification”. *Pattern Analysis and Applications*, vol. 19, no. 3, pp. 699-708, 2016.

[15] C. Shi-Ming and W. Yi-Song. “A robust off-line writer identification method”. *Renhe Test*, vol. 46, no. 1, pp. 108-116, 2020.

[16] A. Chahi, Y. Ruichek and R. Touahni. “Local gradient full-scale transform patterns based off-line text-independent writer identification”. *Applied Soft Computing*, vol. 2020, p. 106277, 2020.

[17] A. Forn, D. Albert and G. Josep. “CVC-MUSCIMA: A ground-truth of handwritten music score images for writer identification and staff removal”. *International Journal on Document Analysis and Recognition*, vol. 15, pp. 243-251, 2012.

[18] A. A. Ahmed and G. Sulong. “Arabic writer identification: A review of literature”. *Journal of Theoretical and Applied Information Technology*, vol. 69, no. 3, pp. 474-484.

[19] G. J. T. Rahim and M. S. M. Rahim. “Off-line text-independent writer recognition for chinese handwriting: A review”. *Jurnal Teknologi*, vol. 2, pp. 39-50, 2015.

[20] S. M. Awaida and S. A. Mahmoud. “State of the art in off-line writer identification of handwritten text and survey of writer identification of Arabic text”. *Educational Research Review*, vol. 7, no. 20, pp. 445-463, 2012.

[21] A. Junaidi, S. Trianingsih and M. Iqbal. “Writer identification of lampung handwritten documents based on selected characters”. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, vol. 6, no. 1, pp. 1-8.

[22] G. Ghiasi and R. Safabakhsh. “Offline text-independent writer identification using codebook and efficient code extraction methods”. *Image and Vision Computing*, vol. 31, no. 5, pp. 379-391, 2013.

[23] T. L. Mohammed, A. A. Ahmed and O. I. Al-Sanjary. “*KRDOH: Kurdish Offline Handwritten Text Database*”. In: 2019 IEEE 7th Conference on Systems, Process and Control (ICSPC), pp. 86-89, 2019.

[24] U. V. Marti and H. Bunke. “The IAM-database: An English sentence database for offline handwriting recognition”. *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39-46, 2002.

[25] A. Durou, I. Aref, S. Al-Maadeed, A. Bouridane and E. Benkhelifa. Writer identification approach based on bag of words with OBI features”. *Information Processing and Management*, vol. 56, no. 2, pp. 354-366, 2019.

[26] H. T. Nguyen, C. T. Nguyen, T. Ino, B. Indurkhya and M. Nakagawa. “Text-independent writer identification using convolutional neural network”. *Pattern Recognition Letters*, vol. 121, pp. 104-112, 2019.