Triangle Shape Feature based on Selected Centroid for Arabic Subword Handwriting

Nur Atikah Arbain[1], Mohd Sanusi Azmi[2], Azah Kamilah Muda[3], Amirul Ramzani Radzid[4]
Faculty of Information and Communication Technology,
University Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia

Azrina Tahir[5]
Department of Information & Communication Technology,
Politeknik Ungku Omar,
Jalan Raja Musa Mahadi,
31400 Ipoh, Perak, Malaysia

***Abstract*—Features are normally modelled based on color, texture and shape. However, some features may have different constraints based on types, styles and pattern of an image. The Arabic subword handwriting, for example, cannot be recognized by color and not suitable to be characterized based on texture. Therefore, features based on shape are suitable to be used for recognizing Arabic subword handwriting since each of the character has various characteristics such as diacritics, thinning and strokes. These characteristics can contribute to particular a shape that is unique and can represent Arabic subword handwriting. Currently, geometry shape such as triangle has been adopted to extract useful features based on triangle properties without implicating any triangle form. In order to increase classification accuracy, these properties have been categorized into several zones where the number of features produced is directly proportional to the number of zones. Nevertheless, shape representation does not implicate any triangle properties such as ratio of side, angle and gradient. By using shape representation, it helps in reducing the number of features. Thus, this paper presents feature based on triangle shape that can represent the identity of Arabic subword handwriting. The method based on triangle shape identifies three main coordinates of triangle formed based on selected centroids. The AHDB dataset is used as a testing data. The Support Vector Machine (SVM) and Random Forest (RF), respectively were used to measure the accuracy of the proposed method using triangle shape as a feature. The accuracy results have shown better outcome with 77.65% (SVM) and 76.43% (RF), which prove the feature based on triangle shape is applicable for Arabic subword handwriting recognition.**

Keywords*—*Arabic subword; feature extraction; random forest; support vector machine; triangle geometry

I. Introduction

Subword handwriting is one of the popular handwriting studies that have been actively explored for many years due to challenges in identifying the styles, patterns, and signatures of subword handwriting. The recognition of the handwritten words is based on the recognition of segmented characters from subword. Images of handwritten documents will be processed from imaging from pages to lines, and words to subwords.

Due to the challenging task in subword handwriting, there have been intensive responses and encouragement by numerous researchers to develop or improvise the existing recognition methods and systems. Based on [1], work in Arabic character recognition is limited. However, Arabic handwritten character recognition systems have achieved much improvement over the years. Since many languages such as Farsi, Curds, and Urdu used Arabic characters in their writing, it makes tasks more challenging due to different words used, strength and sequential order of the writing.

Over the past decades, a lot of handwritten subwords databases [1]–[3], which contain images were developed. Image processing is required to process the images, for example, to convert the image into binary pixels. Image processing is one of the vital elements that are widely used in a research area within engineering and computer science disciplines. Arabic handwritten documents currently exist in a big number of resources in physical and web form, providing a challenge for the word recognition process. The features extraction process will play an important role before the classification process. The problem with segmentation to the single characters is that the characters may be overlapped and some of the characters share the same shape, for example, characters: "ب" ,"ت" and "ث".

The features for recognizing Arabic subword have been introduced, which led to in-depth studies due to variation of style, pattern, characteristic and type of Arabic subword handwriting. Thus, the generated features must have hallmarks that can differentiate them from another subword. Two groups namely analytical and holistic are used as a recognition method in handwritten text. In the analytical group, a word segmented into components such as character or subword, a feature is extracted from each other, and a general vector is obtained from each word. Besides, the character segmentation is needed, and the errors may occur in the recognition step. In the holistic group, a feature vector is extracted from the whole word image without any need to image segmentation.

In this study, a holistic group is applied in producing novel features for Arabic subword handwriting. The feature based on triangle shape is proposed using three main coordinates of triangle formed based on selected centroids. This paper is organized as follows. In Section II, the related work is discussed. The proposed method is discussed in Section III. Next, the experiment and evaluation of study is presented in Section IV. Finally, Section V concludes this paper.

II. Related Works

A huge number of pre-modern texts have been scanned as subword images to remain against aging. Nevertheless, the ability of researchers to handle with the images were limited due to the difficulty to handle certain tasks such as query search [4]. According to [4], it is important to provide researcher with algorithms for automatic transliteration and transcription of scanned images, which would extract the textual content of the image and reproduce into an editable text file.

With the advanced technology, many approaches and methods have been introduced for Arabic handwritten text recognition. The feature learning framework had been proposed by [5] using a Bag-of-Feature (BoF) paradigm for Arabic handwritten text recognition. Besides that, scale invariant feature transform (SIFT) descriptor was used by [6] to represent the object in detail to reduce the computation cost. However, [6] has stated that the complexity of testing image cannot be too high when performing object recognition and image retrieval on big data. This is because a vector with 128 dimensions represents one feature point and an image will have several feature points. Thus, more time is needed to compare the feature points individually.

Therefore, it is important in selecting suitable features applicable to represent the image. The structural approach is applied in generating triangle shape feature based on selected centroid. Then, the holistic approach is applied where the whole subword image is used without segmenting each character from subword image.

The holistic approach has been applied by [7] in producing features for AHDB subword images using Discrete Cosine Transform (DCT) and histogram of oriented gradient (HOG). The features are produced based on whole subword image without any word segmentation. An array of the best 50 DCT coefficients and 324 of HOG features are produced as the parameters of the features for subword images. Besides, a study in [5] also has used holistic approach in producing features based on Bag-of-Feature (BoF) paradigm. The BoF framework is exploited by [5] as to learn robust feature representations for Arabic handwriting recognition. Several approaches have been implemented as there are few stages in the framework that will use different approaches. The Harris detector and dense sampling have been applied for selecting representative image regions. Then, Principle Component analysis (PCA) is applied to reduce Scale-Invariant Feature Transform (SIFT) descriptors to 64-D vectors. In a study by [8], a holistic group approach also has been applied in generating novel features for recognition of Persian/Arabic handwritten words. The generated feature is proposed based on a geometric attribute of components forming the word. The number, angle, location and size of a line are the parameters that represent the features in [8].

The geometry features have been adopted in object recognition, which is especially used for identifying the style and pattern of writing, font, authors, and number of authors, place of writing and originality of the documents. Apart from that, these features also have been extensively used for recognizing the type of writing and calligraphy in existing documents especially for ancient manuscripts [9]. The geometry features can be produced based on geometry shapes such as polygons including triangles, squares and pentagons. These polygons have respective properties that can be used in object recognition.

Most of the properties, for example, triangle properties have been used by researchers to produce proposed features for image classification [9], [10]. The properties are extracted after the polygon is formed. The geometry method also has been broadly used in various domains such as face recognition [11]–[13], fingerprint recognition [14]–[16], vehicle detection [17], intrusion [18] and digit recognition [9], [10]. Each of the domain has a special form that uses an indicator to determine the corner points of the formed geometry shape.

In face recognition, eyes and nose are face elements that are used as indicator to determine the points on the face. The minutiae, ridges and valleys were used as indicator in fingerprint recognition. In vehicle detection, flat road assumption has been used as indicator to search for vehicles that are located on the ground. Besides that, geometry method was also used in recognizing digit recognition and calligraphy [9], [10], [19]. A local foreground image was applied to construct triangle points based on the size of image. The author of [9] proposes new features based on triangle properties. The triangle is formed based on three triangle points of corners A, B and C. The determination of the three triangle points of corners plays a big role in triangle formation. Any fault in determining the exact coordinates of triangle points can affect the triangle formation. The midpoint of triangle is important to determine the position of triangle's point of A and B.

However, the current algorithm to extract features from face and fingerprint recognition respectively cannot be implemented in recognizing subword images. The limitation of elements used such as eyes, mouth, nasal tip, ear hole and corner of mouth in face recognition as well as minutiae in fingerprint recognition cannot be applied due to the aforementioned non-elements that exists in subword images. Thus, elements from both face and fingerprint recognition cannot be used as new feature parameter based on triangle geometry for subword images. Nevertheless, the current algorithm using triangle geometry in digit recognition is possible to be applied on subword image. However, there are constraints where every feature must be produced for all the 33 zones, which eventually lead to the increasing number of features into 297. The algorithm has increased training time in feature extraction process concomitantly with big data image used. Thus, a research on feature based on triangle shape is needed to be extended in order to facilitate subword image.

III. Proposed Method

A. Pre-processing

Before proposed features are produced, binarization process is performed in pre-processing stage for selecting adequate threshold of grey level for extracting objects from image background. Thus, Otsu thresholding method [20] is applied to convert subword image into binary form. The binarization process will transform image into binary form where '0' represents foreground of image while '1' represents background of image as shown in Fig. 1.

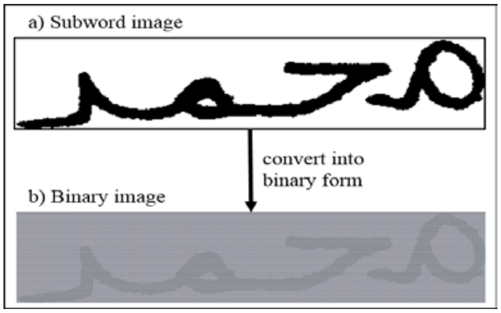


Fig. 1. Subword Image is Converted into Binary Form.

B. Feature Extraction

1) Zoning method: In this stage, the zoning method is applied to divided image into several zones, which contains useful information that can be extracted as the features. The zoning method is known as one of handwriting recognition method where handwriting image will be divided into several zones that provide regional information according to feature needs. There are four types of zoning method applied namely Cartesian plane zone, horizontal zone, vertical zone and 45-degree zone. These zoning methods also have been used in digit recognition [10], [19]. TABLE I shows the summary of zoning method information while Fig. 2 illustrates the image output from Cartesian plane zone method. Based on Fig. 2, binary image is divided into five zones including main image using Cartesian plane zone method. The binary image is measured based on height and width of the image. The height and width of binary image is obtained based on the number of binary pixels including '0' and '1'.

TABLE I. SUMMARY OF ZONING METHOD

Zoning method	**Number of zones**
Cartesian Plane Zone	5 including main image
Horizontal Zone	6
Vertical Zone	14
45-Degree Zone	8
Total	**33**

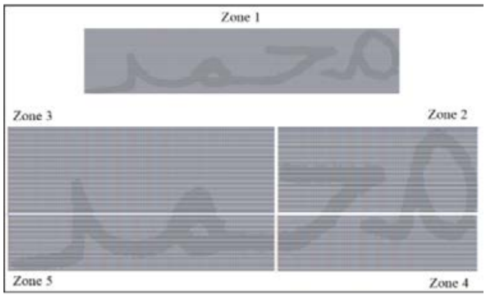


Fig. 2. Output Image after using Cartesian Plane Zone Method.

2) Geometry Method: After applying zoning method, the features can be extracted from each of zones using geometry method. In this study, triangle geometry method has been applied to generate features where the triangle shape is formed inside divided zones. There are 33 triangle shapes formed based on a total number of zones for all types of zoning method (refer to TABLE I). The features are generated based on triangle shape where the three main coordinates of triangle are formed based on selected centroids. There are six types of possible centroids that form the triangle shape as shown in TABLE III. The algorithm to obtain three main triangle coordinates is shown in Fig. 3.

```
Input: binary image of zone
Output: triangle shape points (A, B, C)
Begin
    •  Read image I from dataset
    •  $N$ ← total number of pixels at $x$-axis
    •  Get point $C$ (centroid)
    •  h ← centroid height of zone,
       w ← centroid width of zone
    •  Get point $A$.
       Find $Ax = Cx \; until \; Ax <= N - 1$
    •  Get point $B$. Find $Bx = 0 \; until \; Bx <= Cx$

End
```

Fig. 3. Algorithm for Triangle Shape Coordinates.

After identifying three main triangle coordinates based on selected centroids, the coordinates are used to extract the features based on triangle shape. The number of features based on triangle shape produced 99 features (3 features × 33 zones). The description of triangle shape features based on three main triangle coordinates from selected centroids is shown in TABLE II.

TABLE II. DESCRIPTION OF TRIANGLE SHAPE FEATURES

No	**Triangle shape features**	**Formula**
1	Length of side a	$a = \sqrt{b^2 + c^2 - 2bc.cosA^{\circ}}$
2	Length of side b	$b = \sqrt{a^2 + c^2 - 2ac.cosB^{\circ}}$
3	Length of side c	$c = \sqrt{a^2 + b^2 - 2ab.cosC^{\circ}}$

I. EXPERIMENT AND EVALUATION

In this study, Arabic subword handwriting from AHDB database is used. This dataset contains more than 2000 images of Arabic words and texts written by a hundred different writers where 70% data is used as training data while 30% is used as testing data. As to evaluate the data, Support Vector Machine (SVM) and Random Forest (RF) are applied to measure the data based on accuracy. As known, the SVM is one of most popular approach that has been used in measuring classification accuracy for handwriting recognition. Thus, the libSVM is required to gain the highest cross-validation (CV) accuracy for each of the SVM parameter. The Gaussian kernel is applied to search the best grid point of cost and gamma with highest cross-validation. Then, the best value of cost and gamma are used to train the dataset using SVM. With the best value of cost and gamma, a good accuracy is achieved accordingly to the dataset nature and characteristic. The cost and gamma value for proposed method in [9] and our proposed method respectively is shown in TABLE IV.

TABLE III. DESCRIPTION OF TRIANGLE SHAPE BASED ON SELECTED CENTROID

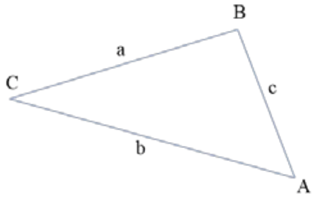
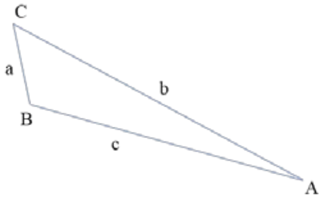
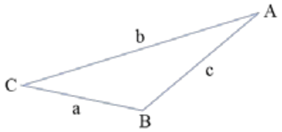
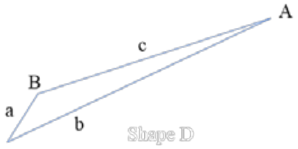
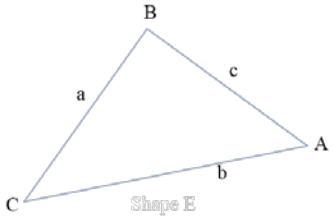
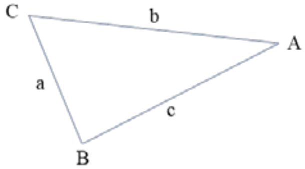
Shape types	**Rules for centroid**	**Triangle output**
A	$yA \geq yC \geq yB$	
B	$yA \geq yB \geq yC$	
C	$yA \leq yC \leq yA$	
D	$yA \leq yB \leq yC$	
E	$yA \geq yB \leq yC$	
F	$yA \geq yC \leq yB$	

TABLE IV. COST AND GAMMA RESULTS USING LIBSVM FUNCTION

Proposed Method	**Cost (c)**	**Gamma (γ)**
M. S. Azmi (2013) [9]	32.0	0.001953125
Our proposed method	32.0	0.03125

TABLE V. CLASSIFICATION ACCURACY RESULTS BASED SVM

Proposed Method	**Number of features**	**Accuracy (%)**
M. S. Azmi (2013) [9]	297	76.122
Our proposed method	99	77.653

The results of accuracy based on SVM classifier are compared between prior method [9] and the proposed method. Based on TABLE V, the accuracy result for the proposed method has shown better outcome by obtaining 77.653% compared to proposed method by proposed method of [9] which obtained only 76.122%. The results based on SVM showed that the proposed method has achieved target to apply minimum number of features by using triangle shape feature. Number of features is possible to be reduce from 297 to 99 by using different approaches of triangle shape features types. Furthermore, triangle shape feature can differentiate triangle shape from another triangle shape types.

Besides that, the accuracy results are also compared using other classifier based on different features used on AHDB dataset. Based on TABLE VI, the accuracy result based on random forest has shown good result for our proposed method by increasing about 7% compared to the proposed method by prior method [21]. It has shown that the proposed features using triangle shape is efficient and applicable to be used in Arabic handwritten text recognition. However, the handwriting styles, pattern and types may influence in producing the features, which made recognizing the Arabic handwriting text more challenging.

TABLE VI. CLASSIFICATION ACCURACY RESULTS BASED RF

Proposed Method	**Features**	**Accuracy (%)**
J. Salem (2017) [21]	MOMENTS	68.750
Our proposed method	Triangle shape	76.417

II. CONCLUSION

This paper presents a feature based on triangle shape that formed three main coordinates using selected centroids. The proposed feature based on triangle shape has been proven applicable to be used as a feature for recognizing Arabic subword handwriting. The results based on SVM and RF have shown good result for the proposed method compared to prior methods. The further research can be extended where other geometry shapes can be applied as a feature.

ACKNOWLEDGMENT

The authors would like to express their appreciation to Universiti Teknikal Malaysia Melaka for providing a scholarship of UTeM Zamalah Scheme. Besides, thank you to Faculty of Information Technology and Communication for providing excellent research facilities.

REFERENCES

[1] S. Al-Ma’adeed, D. Elliman, and C. Higgins, “A database for Arabic handwritten text recognition research,” Int. Arab J. Inf. Technol., vol. 1, no. 1, pp. 117–121, 2004.

[2] J. J. Hull, “A Database for Handwritten Text Recognition Research,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 16, no. 5, pp. 550–554, 1994.

[3] H. C. Fernando, N. D. Kodikara, and S. Hewavitharana, “A database for handwriting recognition research in Sinhala language,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2003–Janua, no. Icdar, pp. 1262–1264, 2003.

[4] Y. Chherawala, R. Wisnovsky, and M. Cheriet, “TSV-LR: topological signature vector-based lexicon reduction for fast recognition of pre-modern Arabic subwords,” Proc. 2011 Work. Hist. Doc. Imaging Process., pp. 6–13, 2011.

[5] M. O. Assayony and S. A. Mahmoud, “An Enhanced Bag-of-Features Framework for Arabic Handwritten Sub-words and Digits Recognition,” J. Pattern Recognit. Intell. Syst., vol. 4, no. 1, pp. 27–38, 2016.

[6] C. Wu, C. Te Chiu, and Y. S. Hsu, “Object recognition using bag of words with kernels for big data,” Dig. Tech. Pap. - IEEE Int. Conf. Consum. Electron., pp. 89–90, 2014.

[7] M. S. Kadhm and A. K. A. Hassan, “Arabic handwriting text recognition based on efficient segmentation, DCT and HOG features,” Int. J. Multimed. Ubiquitous Eng., vol. 11, no. 10, pp. 83–92, 2016.

[8] R. Tavoli, M. Keyvanpour, and S. Mozaffari, “Statistical geometric components of straight lines (SGCSL) feature extraction method for offline Arabic/Persian handwritten words recognition,” IET Image Process., vol. 12, no. 9, pp. 1606–1616, 2018.

[9] M. S. Azmi, “Fitur Baharu Dari Kombinasi Geometri Segitiga Dan Pengezonan Untuk Paleografi Jawi Digital,” Doctoral dissertation, Universiti Kebangsaan Malaysia, 2013.

[10] M. S. Azmi, N. A. Arbain, A. K. Muda, Z. Abal Abas, and Z. Muslim, “Data Normalization for Triangle Features by Adapting Triangle Nature for better Classification,” 2015 IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol., pp. 1–4, 2015.

[11] M. M. M. Tin, M. M. Sein, and H. Township, “Multi Triangle Based Automatic Face Recognition System By,” I2MTC 2009 - Int. Instrum. Meas. Technol. Conf., no. May, pp. 5–7, 2009.

[12] J. Zheng, Y. Gao, and M.-Z. Zhang, “Fingerprint Matching Algorithm Based on Similar Vector Triangle,” in Image and Signal Processing, 2009. CISP ’09. 2nd International Congress, 2009, pp. 1–6.

[13] Z. Zhang, S. Wang, and A. I. Morphing, “Multi-feature facial synthesis based on triangle coordinate system,” Proc. 2nd Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2012, no. 2, pp. 141–145, 2012.

[14] M. Ghazvini, H. Sufikarimi, and K. Mohammadi, “Fingerprint matching using genetic algorithm and triangle descriptors,” 19th Iran. Conf. Electr. Eng., pp. 1–6, 2011.

[15] A. Gago-Alonso, J. Hernández-Palancar, E. Rodríguez-Reina, and A. Muñoz-Briseño, “Indexing and retrieving in fingerprint databases under structural distortions,” Expert Syst. Appl., vol. 40, no. 8, pp. 2858–2871, 2013.

[16] W. Yang, J. Hu, S. Wang, and J. Yang, “Cancelable Fingerprint Templates with Delaunay Triangle-Based Local Structures,” Cybersp. Saf. Secur., pp. 81–91, 2013.

[17] A. Haselhoff and A. Kummert, “A Vehicle Detection System Based on Haar and Triangle Features,” in Intelligent Vehicles Sysmposium, 2009, pp. 261–266.

[18] P. Tang, R. Jiang, and M. Zhao, “Feature selection and design olintrusion detection system based on k-means and triangle area support vector machine,” Second Int. Conf. Futur. Networks ICFN’10, pp. 144–148, 2010.

[19] N. A. Arbain, M. S. Azmi, S. S. S. Ahmad, I. E. A. Jalil, M. Z. Masud, and M. A. Lateh, “Detection on Straight Line Problem in Triangle Geometry Features for Digit Recognition,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 1019–1025, 2016.

[20] P. Smith, D. B. Reid, C. Environment, L. Palo, P. Alto, and P. L. Smith, “A Threshold Selection Method from Gray-Level Histograms,” IEEE Trans. Syst. Man. Cybern., vol. 9, no. 1, pp. 62–66, 1979.

[21] J. R. A. Salem, “Segmentation Methods of Arabic Handwriting Using Neighbourhood Information For Voronoi Diagrams,” Doctoral dissertation, Universiti Kebangsaan Malaysia, 2017.