2012 International Conference on Image and Information Processing (ICIIP 2012)
IPCSIT vol. 46 (2012) © (2012) IACSIT Press, Singapore

Farsi/Arabic Digit Classification Using Triangle Based Model Features with Ranking Measures

Mohd Sanusi Azmi [1 +], Mohamad Faidzul Nasrudin [2] and Khairuddin Omar [3], Khadijah Wan Mohd Ghazali [4] and Azah Kamilah Muda [5]

[1, 4, 5] Faculty of Information Communication and Technology, Universiti Teknikal Malaysia Melaka, Malaysia

[2, 3] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

Abstract. The HODA digit dataset is a huge digit dataset for Farsi and Arabic characters. With the huge collection of ≈80000 images, Supervised Machine Learning (SML) has been the more preferable technique for classification of the dataset among researchers, as well as Multi-layer Perceptrons, Radial Basic Function and Support Vector Machine (SVM). In this paper, we propose the use of Unsupervised Machine Learning (UML) with ranking measures that are widely used in information retrieval. The UML algorithm used is the Euclidean Distance Method (EDM) whereas the information retrieval measures used for the classification are Majority Voting (MV) and Mean Average Precision (MAP). Experiments have been conducted using ≈ 60000 images for training and 20000 images for testing for both the UML and SML techniques. The features used are from Triangle Based Model which has been proposed in our previous works. Result from the tests prove that the features from Triangle Model techniques with the UML and MAP techniques give better result compared to SML with Multi-layer Perceptrons and UML with MV.

Keywords: HODA digit dataset, triangle based model, features extraction, ranking measures, mean average precision.

1. Introduction

A very comprehensive digit dataset for Farsi/Arabic characters is introduced in [1]. The dataset was named as HODA Farsi Digit Dataset. The dataset has huge number of images for training and testing. The total images for training is ≈60000 and testing is 20000 [1–4].

The dataset was introduced in 2007 [1], since then many researchers have been using SML techniques as its classification. SMLs that have been explored by past researchers are Multi-layer Perceptrons [2], [3], [5] Radial Basic Function [3] and Support Vector Machine [4–6]. Our studies so far have not found any published research on HODA dataset that uses UML as its classification method. In this paper, we propose classification of the HODA digit dataset using a UML algorithm named Euclidean Distance Method (EDM) with two information retrieval measures that determines accuracy by ranking, namely the mean average precision (MAP) [7] and majority voting [8]. The reasons why EDM is chosen has been detailed in [6]. The features used for the UML and SML in this paper are those that had been proposed in [6].

The features we introduced in [6] were based on geometrical features. The features were extracted from triangles that are based on the important coordinates defined in [9]. The features, named as Triangle Based Model are explained in Table. 2 . In this paper, we implement the four quadrants of the proposed features in [6]. The features from the implementation of four quadrants were used in the UML with MAP, UML with MV and SML using MLP.

[+] Corresponding author. Tel.: + 60196264558.
E-mail address: sanusiazmi@gmail.com.

In this paper, sections below are divided into Pre-processing, Experimental Result, Conclusion and Future Improvement.

2. Pre-Processing

In this part, the images from the dataset were extracted to the Triangle Based Model features. The processes took place in stages as follows:

a. Data Collection
b. Features Extraction
c. Experimental Setup

2.1. Data collection

The dataset used by this research were based on the features extracted from the standard Farsi/Arabic digit dataset HODA as introduced by [1]. The example of HODA dataset is shown in Fig. 1. There are 10 classes of digit in the dataset. The total number of images in the training is 56790 images, with 20000 of them for testing. The testing images consist of ten classes of 2000 images each. The distribution of training / testing is shown in Table. 1.

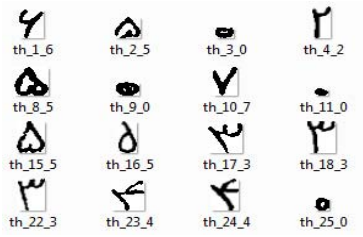

Fig. 1: HODA Digit Dataset

Table. 1: Distribution of classes

class	Model/Training	Testing
hoda0	5712	2000
hoda1	5665	2000
hoda2	5666	2000
hoda3	5635	2000
hoda4	5670	2000
hoda5	5692	2000
hoda6	5715	2000
hoda7	5664	2000
hoda8	5670	2000
hoda9	5701	2000
Total	**56790**	**20000**

2.2. Features Extraction

The features used in this research are shown in Table. 2. The features were extracted from the main image named as "*Main Triangle*" and four quadrants as illustrated in Fig. **2**. The quadrants were named as "*A*" on the upper right, "*B*" on the upper left, "*C*" on the lower right and "*D*" on the lower left. The "*Main Triangle*" and the four quadrants illustrated in the figure were used to calculate features based on the three important points for each of them. The three important points are defined in [6], [10].

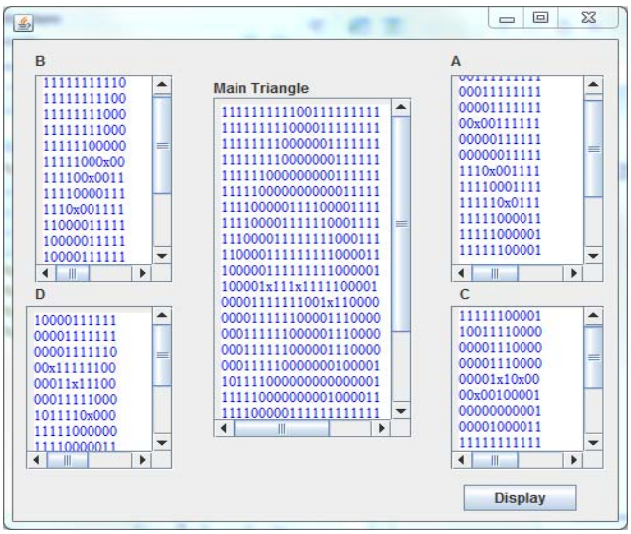

Fig. 2: Segregation of Isolated Characters to Four Quadrants

The three important points for each of the quadrants and "Main Triangle" are labelled as "*x*" as shown in Fig. **2** above. The "*x*"s are the coordinates of triangles used to extract features. The features are shown in Table. 2 below.

Table. 2: Features from the Triangle as in [6]

No.	Feature Name	Description
1	*c/a*	Ratio of side *c* to *a*
2	*a/b*	Ratio of side a to b
3	*b/c*	Ratio side b to c
4	*A*	Angle of A
5	*B*	Angle of B
6	*C*	Angle of C
7	*GraBA*	Gradient of B and A
8	*GraBC*	Gradient of B and C
9	*GraCA*	Gradient of C and A

Fig. **2** shows the five zones of the image. Using the nine features on each zone, 45 features were selected from each image. The proposed features were obtained by using sides of triangles, calculated using the Pythagorean Theorem. Fig. **3** and the list of formula below show how the features were extracted from the triangle.

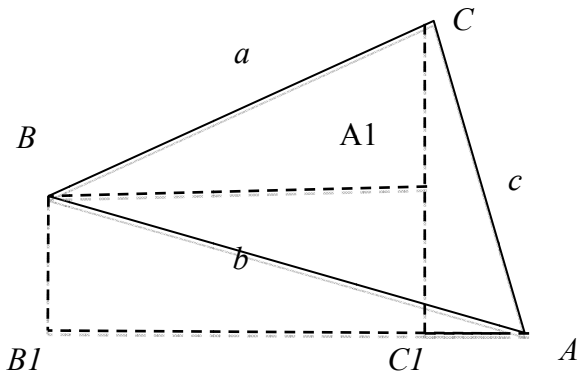

Fig. 3: Calculation of the triangle's sides

$$a^2 + b^2 = c^2 \tag{1}$$

$$\therefore$$

$$a = \sqrt{((A1(y) - C(y))^2) + ((A1(x) - B1(x))^2)} \tag{2}$$

$$b = \sqrt{((B1(y) - B(y))^2) + ((A(x) - B1(x))^2)} \tag{3}$$

$$c = \sqrt{((C1(y) - C(y))^2) + ((A(x) - C1(x))^2)} \quad (4)$$

After sides *a*, *b* and *c* were calculated, the ratios were defined by dividing the sides as in Table. 2. The features of angles of A, B and C were calculated using the formula below.

$$A = arccos \frac{b^2+c^2-a^2}{2bc} \quad (5)$$

$$B = arccos \frac{a^2+c^2-b^2}{2ac} \quad (6)$$

$$C = arccos \frac{a^2+b^2-c^2}{2ab} \quad (7)$$

The gradients features were computed based on coordinates *A*, *B* and *C*. The formulas below show how the gradients were calculated.

$$GraBC = \frac{B(y)-C(y)}{B(x)-C(x)} \quad (8)$$

$$GraBA = \frac{B(y)-A(y)}{B(x)-A(x)} \quad (9)$$

$$GraCA = \frac{B(y)-C(y)}{B(x)-C(x)} \quad (10)$$

2.3. Experimental Setup

The experiment was conducted using Unsupervised Machine Learning (UML) and Supervised Machine Learning (SML). The amount of data for training and testing was based on [1], that is ≈60000 data for training and 20000 data for testing for both SML and UML.

For the UML, the 20000 data for testing were clustered into ten classes; each represents a Farsi/Arabic digit. Then, the testing for the UML was conducted with MV and MAP. The criteria used in the MV and MAP were based on the "*Top 5*", "*Top 10*" and "*Top 20*". The SML was conducted using Neural Network with learning rate of 0.4. Result from the confusion matrix in Fig. **4** was used in order to get the percentage of each class. The results for the testing on SML and UML are shown in Experimental Result.

3. Experimental Result

For the first experiment, features extracted from Triangle Based Model were tested using UML with MV, followed by UML with MAP. The testing was done in three phases. The first phase tested "Top 5", the second tested "Top 10" and the last phase tested "Top 20". The result is shown in Table. 3 below.

Table. 3:Result for UMLwith MV and UML with MAP

Type	**UML-MV**			**UML-MAP**		
Rank	**Top 5**	**Top 10**	**Top 20**	**Top 5**	**Top10**	**Top20**
Hoda0	89.62	88.365	86.705	93.728	92.459	90.759
Hoda1	95.65	95.07	94.72	96.924	96.511	95.972
Hoda2	83.83	82.56	81.5425	88.075	86.834	85.257
Hoda3	66.85	64.675	62.2925	76.671	73.861	70.374
Hoda4	49.9	47.71	44.6275	62.275	59.812	55.812
Hoda5	89.19	88.43	87.545	92.077	91.541	90.624
Hoda6	80.34	78.56	76.2375	86.454	84.550	82.241
Hoda7	91.5	90.47	89.9225	93.726	92.976	92.026
Hoda8	95.14	94.455	93.7025	96.757	96.276	95.611
Hoda9	85.87	84.895	83.695	90.018	88.982	87.604
Average	**82.789**	**81.519**	**80.099**	**87.671**	**86.380**	**84.628**

Based on the result in Table. 3 above, UML with MV gives good result in the Top 5 test, outperforming the Top 10 and top 20 tests. The result shows that, Hoda8 and Hoda1 can be classified up to 95%. However, for Hoda4, the accuracy is only 49.9%. The UML with MAP test gives better result for all classes compared to the UML with MV in Table. 3. The top five gives ≈87.67%, outperforming the Top 10 and top 20 tests. The results from Table. 3 were evaluated using Multi-layer Perceptrons with learning rate of 0.4.

Table. 4 below shows result for SML using Neural Network. The confusion matrix for the Table. 4 is as shown in Fig. 4.

Table. 4: Result for SML with MLP

Class	SML-MLP (Learning Rate 0.4)
Hoda0	92.45
Hoda1	96.45
Hoda2	88.20
Hoda3	70.35
Hoda4	73.1
Hoda5	88.35
Hoda6	86.05
Hoda7	90.6
Hoda8	96.75
Hoda9	88.2
Average	**87.05**

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j   <-- classified as
 1849   48    3    2   45   36   10    3    2    2 |    a = hoda0
   37 1929    7    1   14    0    8    0    0    4 |    b = hoda1
    4   12 1764   66   93    0   27   25    0    9 |    c = hoda2
    6    3  128 1407  434    3   13    2    2    2 |    d = hoda3
   61   14   38  291 1462   31   22   31   14   36 |    e = hoda4
   79    6    4    5  103 1767   18    7    7    4 |    f = hoda5
   32   19   47   10   41   14 1721   31    3   82 |    g = hoda6
   10   10   36    5   54    9   64 1812    0    0 |    h = hoda7
    3    6    0    0   16    2    7    0 1935   31 |    i = hoda8
   24   50   13    1   28    9   99    0   12 1764 |    j = hoda9
```

Fig. 4: Confusion Matrix

The evaluation results for the test that have been conducted are as shown in Table. 5 below. The top five from UML from both tests were evaluated with the result from SML with MLP.

Table. 5: Comparison Result between UML and SML

	UML		SML
Class\Top	**Top 5 Majority Voting (MV)**	**Top 5 Mean Average Precision (MAP)**	**SML-MLP (Learning Rate 0.4)**
Hoda0	89.62	93.728	92.45
Hoda1	95.65	96.924	96.45
Hoda2	83.83	88.075	88.20
Hoda3	66.85	76.671	70.35
Hoda4	49.9	62.275	73.1
Hoda5	89.19	92.077	88.35
Hoda6	80.34	86.454	86.05
Hoda7	91.5	93.726	90.6
Hoda8	95.14	96.757	96.75
Hoda9	85.87	90.018	88.2
Average	**82.789**	**87.670**	**87.05**

The implementation of four zones using the proposed features in this research gives significant result improvements to HODA digit dataset. The result can be evaluated with the result in [6]. The UML with MAP for top 5 gives best result.

4. Conclusion

UML with MAP based on top 5 outperforms SML with MLP. This can be benchmarked with the previous researches as the experiments used the same features and same size of training and testing data. There are rooms of improvement for the accuracy of the classification of HODA digit dataset by expanding features into more zones that discriminate classes in the dataset better. The implementation of four zones using the proposed features in this research gives significant result improvements to HODA digit dataset compared to the [6]. As conclusion, the UML with MAP can be used in classification for the huge dataset like HODA.

5. Future Improvement

There are rooms of improvement for the accuracy of the classification of HODA digit dataset. First, the classification using UML with MAP need to be conducted several times using random data by maintaining the number of size for training and testing. The average of the accuracy from results taken from random data will increase the level of confidence. Second, the number of zones should be increased to boost the accuracy of classification. Finally, the class hoda4 Table. 5 need to be studied thoroughly in order to minimize chances of incorrect classification.

6. Acknowledgment

The authors would like to thank the Government of Malaysia for sponsoring the research, Universiti Teknikal Malaysia Melaka for allowing study leave for the main researcher, and the Pattern Recognition Group, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia for providing excellent research faculties and facilities.

7. References

[1] H. Khosravi and E. Kabir, Introducing a very large dataset of handwritten Farsi digits and a study on their varieties, *Pattern Recognition Letters*, vol. 28, pp. 1133-1141, 2007.

[2] M. M. Javidi, R. Ebrahimpour, and F. Sharifizadeh, Persian handwritten digits recognition: A divide and conquer approach based on mixture of MLP experts, *International Journal of the Physical Sciences*, vol. 6, no. 30, pp. 7007-7015, Nov. 2011.

[3] R. Ebrahimpour, A. Esmkhani, and S. Faridi, Farsi handwritten digit recognition based on mixture of RBF experts, *IEICE* Electronics *Express*, vol. 7, no. 14, pp. 1014-1019, 2010.

[4] M. Hamidi and A. Borji, Invariance analysis of modified C2 features: case study—handwritten digit recognition, *Machine Vision and Applications*, vol. 21, no. 6, pp. 969-979, Aug. 2009.

[5] H. Parvin, H. Alinejad-rokny, and S. Parvin, Divide and Conquer Classification 1 1, *Science*, vol. 5, no. 12, pp. 2446-2452, 2011.

[6] M. S. Azmi and K. Omar, "Arabic Calligraphy Classification using Triangle Model for Digital Jawi Paleography Analysis," in *11th International Conference on Hybrid Intelligent Systems*, 2011, pp. 704-708.

[7] M. F. Nasrudin, K. Omar, C.-Y. Liong, and M. S. Zakaria, Object Signature Features Selection for Handwritten Jawi Recognition, *Distributed Computing and Artificial Intelligence, The International Symposium on Distributed Computing and Artificial Intelligence (DCAI'10)*, vol. 79, pp. 689-698, 2010.

[8] D. Jang and C. D. Yoo, "Music information retrieval using novel features and a weighted voting method," *Information Retrieval*, no. ISlE, pp. 1341-1346, 2009.

[9] M. S. Azmi, K. Omar, M. Faidzul, N. Khadijah, and W. Mohd, "Arabic Calligraphy Identification for Digital Jawi Paleography using Triangle Blocks," in *2011 International Conference on Electrical Engineering and Informatics*, 2011, no. July, pp. 1714-1718.