International Conference on Computer Applications 2016 [ICCA 2016]

ISBN	978-81-929866-5-4		**VOL**	05
Website	icca.co.in		**eMail**	icca@asdf.res.in
Received	14 – March– 2016		**Accepted**	02 - April – 2016
Article ID	ICCA001		**eAID**	ICCA.2016.001

Recognition of Isolated Handwritten Arabic and Urdu Numerals along with their Variants

Md Sohail Siddique[1], **Ayatullah Faruk Mollah**[2]
[1,2]Department of Computer Science and Engineering, Aliah University, New Town, Kolkata, India

Abstract- *Arabic script is one of the most widely used scripts in the world. Besides 340 millions native speakers, more than 200 million nonnative speakers are there in the world. Many languages such as Urdu, Persian, etc. adopted this script. Therefore, recognition of optical handwritten Arabic characters has received significant attention from the past decade. Unlike numerals of any other scripts, some Arabic numerals have more than one representation. For instance, Arabic '4', '5' and '7' have two variations each. Considering these variants as the same class may deviate the recognition performance. The present work proposed a mechanism to deal with such variations for improved classification. At the first phase, the recognition problem is considered as a 13 class classification problem instead of 10. Then, in the second phase, the classes are reorganized and post-processed for improved classification. A comparative study has also been included with the conventional approach that considers the variants as the same class. Experimental results reflect the efficiency of the proposed technique.*

I. INTDRODUCTION

Optical character recognition (OCR) of handwritten text is an active area of research. Unlike printed text, handwritten text recognition involves a number of additional challenges mainly due to varying handwriting styles and slanted as well as cursive nature of such texts. Although, significant works have been carried out for handwritten Roman numerals, recognition of handwritten Arabic numerals is yet an unsolved problem 1. Arabic script is one of the most widely used scripts in the world. Besides 340 millions native speakers, more than 200 million nonnative speakers are there in the world. About fifty languages such as Urdu, Persian, etc. have adopted this script 2.

	0	1	2	3	4	5	6	7	8	9
Classical Arabic	٠	١	٢	٣	۴	۵	٦	۷	٨	٩
Modern Standard Arabic	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Urdu	٠	١	٢	٣	۴	۵	٦	۷	٨	٩

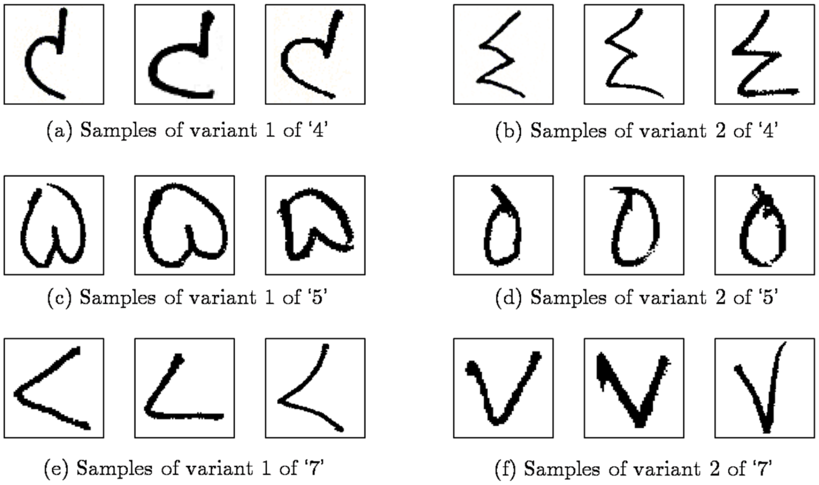
Figure 1. Numerals and their variants of Arabic script

There are two variants of Arabic language viz. Classical Arabic and Modern Standard Arabic. Variations in numerals among these

Cite this article as: Md. Sohail Siddique, Ayatullah Faruk Mollah. "Recognition of Isolated Handwritten Arabic and Urdu Numerals along with their Variants". *International Conference on Computer Applications 2016*: 01-05. Print.

language-variants are shown in Figure 1. It may be noted that some numerals have more than one variants that must be taken care of during recognition. Although, Arabic is written from right to left, Arabic numerals are written from left to right. It is also evident that Urdu numerals are identical to Classical Arabic numerals.

Figure 1 shows that numerals '4', '5' and '7' have significant variations. Some handwritten samples of these letters are also shown in Figure 2.



(a) Samples of variant 1 of '4'

(b) Samples of variant 2 of '4'

(c) Samples of variant 1 of '5'

(d) Samples of variant 2 of '5'

(e) Samples of variant 1 of '7'

(f) Samples of variant 2 of '7'

Figure 2. Variants of same handwritten numerals

So, a complete OCR system for Arabic numerals needs to consider these variants for recognition. Existing works on Arabic numerals recognition focuses on Modern Standard Arabic Numerals 345678. In this paper, an OCR system for Arabic numerals along with their variants is presented.

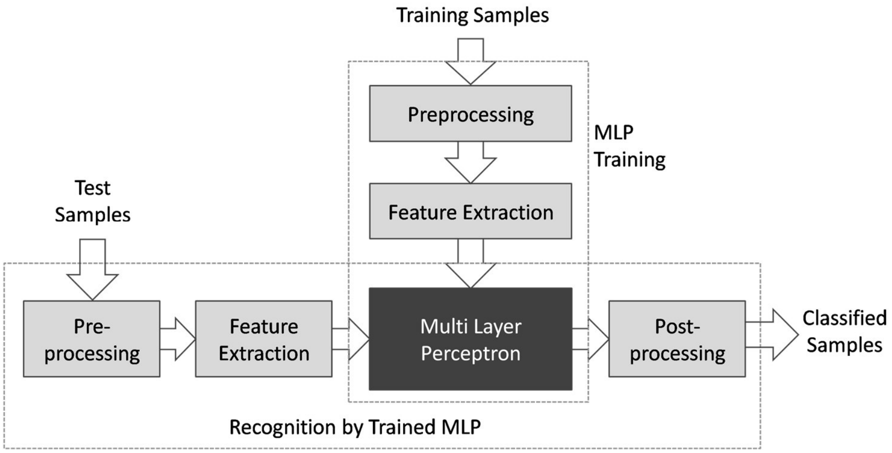


Figure 3. Block diagram of the MLP based OCR system

II. Present Work

The block diagram of the present system is shown in Figure 3. At first, training samples are preprocessed and the features extracted from them are used to train a multi-layer perceptron (MLP). Then, the preprocessed recall samples are tested using the trained MLP by feeding the same features used during training.

A. Preprocessing

At the preprocessing phase, the segmented training samples are binarized using Otsu's 9 global thresholding technique. Then, boundary box is detected and the sample is normalized to a standard size. In this work, 48x48 pixels have been taken as the normalized size. In Figure 4, sample views at different stages for a sample pattern are shown.

Cite this article as: Md. Sohail Siddique, Ayatullah Faruk Mollah. "Recognition of Isolated Handwritten Arabic and Urdu Numerals along with their Variants". *International Conference on Computer Applications 2016*: 01-05. Print.

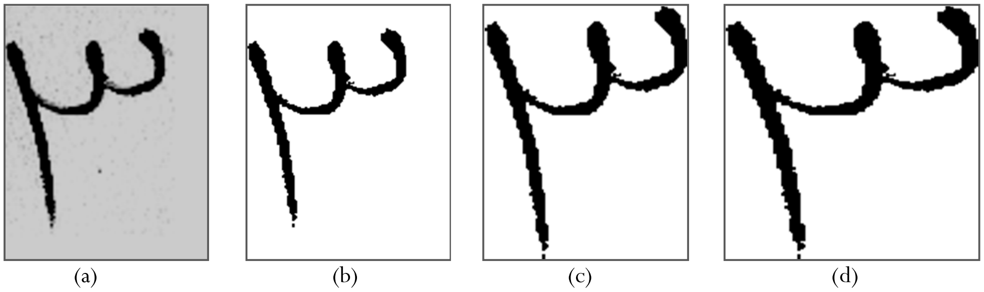


Figure 4. Samples at different preprocessing stages, (a) segmented input sample (Arabic '3'), (b) binarized view, (c) bounded sample, (d) normalized view (48x48 pixels)

B. Feature Extraction

Features are extracted from all samples for the purpose of training and testing. In this work, we have implemented octant centroid (16), longest run (20) and shadow features (24) reported by Basu et. al.10. Octant feature set is prepared by taking the position of centroid of all octants. Longest run feature set computes the longest run length at various direction and shadow features calculate the lengths of shadowed projection from beams of light sent from various directions.

C. Architecture of MLP

In the present work, a feed forward back propagation multi-layer perceptron is used having a three layer architecture. The first layer is the input layer in which the number of neuron is equal to the number of input features. The second layer is the hidden layer in which the number of neuron is heuristically determined and the last layer is the output layer where the number of classes is taken as the number of neuron. The architecture is shown in Figure 5.

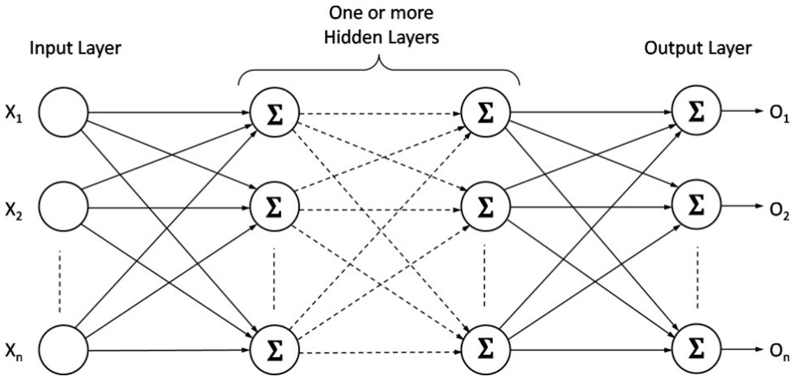


Figure 5. Architecture of the MLP based classifier

D. Classification of Numerals

As discussed in Section I that the variants of some numerals would be considered, we have designed the classification problem in two ways. At first, we have considered all variants of the same class together and in this way we have got 10 classes (i.e. 0-9) as shown in Figure 6. Then, each variant is considered as a separate class and in this way 13 classes are found as shown in Figure 7.

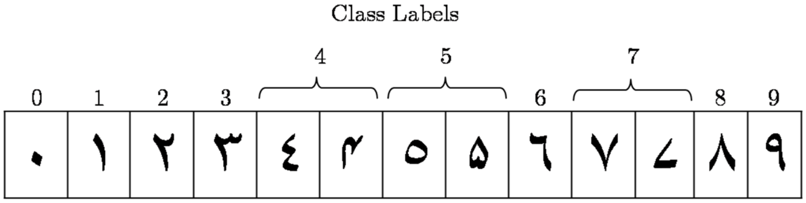


Figure 6. Classification problem with conventional 10 classes

Cite this article as: Md. Sohail Siddique, Ayatullah Faruk Mollah. "Recognition of Isolated Handwritten Arabic and Urdu Numerals along with their Variants". *International Conference on Computer Applications 2016*: 01-05. Print.

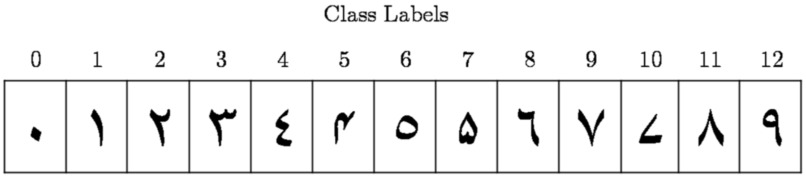


Figure 7. Classification problem with proposed 13 classes

III. Experimental Results

To validate the performance of the present system, handwritten numerals data have been collected at first. People of different ages, backgrounds and literacy who are familiar with Arabic and Urdu numerals have given their isolated handwritten numerals in a given format shown in Figure 8. Then, these sheets are segmented and individual numerals' images are stored into the database. This database contains a total of 3900 samples (300 samples for each of the 13 classes). For the present experiment, training and test samples are taken in the ratio of 2:1. So, 2600 random samples are chosen for training and the remaining 1300 samples are taken for testing.

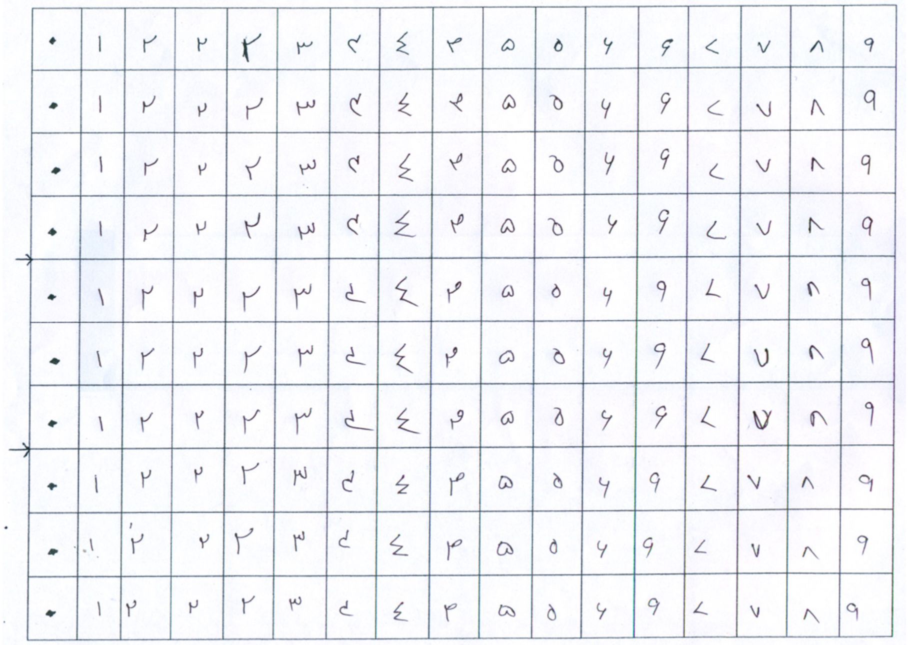


Figure 8. Sample sheet for data collection

In the first phase, the MLP classifier is run for 10 classes as shown in Figure 6. Numerals having two variants each have 600 samples that are divided into 400 training samples and 200 test samples. So, Class 4 {٤, ۴}, Class 5 {٥, ۵} and Class 7{ ٧, ۷} will have two variants each.

Table I Recognition performance with varying number of classes for different feature sets

Sl	Name of Feature	Number of Features	Recognition Accuracy with 10 classes	Recognition Accuracy with 13 classes
1	Set 1 {Octant}	16	90.92	91.46
2	Set 2 {Longest Run}	20	89.31	92.31
3	Set 3 {Shadow}	24	93.62	94.15
4	Set 4 {Octant, Longest Run}	36	93.15	95.46
5	Set 5{Octant , Shadow}	40	94.69	94.85
6	Set 6{Longest Run, Shadow}	44	94.85	95.15
7	Set 7{Octant, Longest Run, Shadow}	60	95.31	95.22

Cite this article as: Md. Sohail Siddique, Ayatullah Faruk Mollah. "Recognition of Isolated Handwritten Arabic and Urdu Numerals along with their Variants". *International Conference on Computer Applications 2016*: 01-05. Print.

In the second phase, the classifier is run for initially 13 classes and then the variants are merged in order to make them belong to the respective true classes. Recognition accuracy obtained from both the approaches for three feature sets and their combinations are shown in Table I and Figure 9. It may be noted from Table I that recognition accuracy with the latter approach is significantly higher compared to the former one except for the Set 7 where the accuracy with 13 class approach is slightly less. However, for six sets out of seven, 13 class approach yields better classification performance.

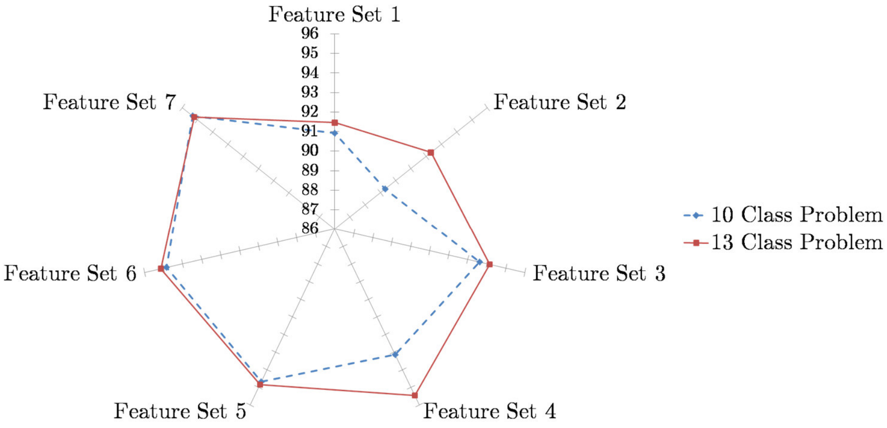


Figure 9. Comparative chart for recognition accuracy obtained with various feature sets for the conventional approach and the present approach

IV. Conclusion

This paper presents an approach for taking the variants of handwritten Arabic and Urdu numerals into consideration while having high recognition accuracy. Instead of three fold validation, training and test samples are randomly chosen for ensuring generality of the system. Reasonably good accuracy i.e. more than 95% is obtained for feature sets having more than 30 features. However, the accuracy can be improved by feature optimization and post-processing. Designing an optimized feature set and stronger post-processing are left for future work.

References

1. L.M. Lorigo, V. Govindaraju, "Offline Arabic handwriting recognition: a survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.28, no.5, pp.712-724, May 2006.
2. http://www.omniglot.com/writing/arabic.htm
3. Yousef Al-Ohali, Mohamed Cheriet, Ching Suen, “Databases for recognition of handwritten Arabic cheques”, Pattern Recognition, vol. 36, no. 1, pp. 111-121, January 2003.
4. Mahmoud, Sabri. "Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models." Signal Processing 88.4 (2008): 844-857.
5. Lawal, Isah, Radwan E. Abdel-Aal, and Sabri Mahmoud. "Recognition of Handwritten Arabic (Indian) Numerals Using Freeman's Chain Codes and Abductive Network Classifiers", 20th IEEE International Conference on Pattern Recognition (ICPR), 2010.
6. Parvez, Mohammad Tanvir, and Sabri A. Mahmoud. "Arabic handwriting recognition using structural and syntactic pattern attributes." Pattern Recognition 46.1 (2013): 141-154.
7. Zaghloul, Rawan I., Dojanah MK Bader Enas, and F. AlRawashdeh. "Recognition of Hindi (Arabic) Handwritten Numerals." American Journal of Engineering and Applied Sciences 5.2 (2012).
8. Ghaleb, Mohamed H., Loay E. George, and Faisel G. Mohammed. "Printed and Handwritten Hindi/Arabic Numeral Recognition Using Centralized Moments", International Journal of Scientific & Engineering Research, vol. 5, no. 3, pp. 140-144, March-2014.
9. Otsu, Nobuyuki. "A threshold selection method from gray-level histograms." Automatica 11.285-296 (1975): 23-27.
10. S.Basu, N.Das, R.Sarkar, M.Kundu, M.Nasipuri, D.K.Basu, “Handwritten ‘Bangla’ Alphabe recognition using an MLP based classifier”, NCCPB-2005, Bangladesh, pp.285-291.

Cite this article as: Md. Sohail Siddique, Ayatullah Faruk Mollah. “Recognition of Isolated Handwritten Arabic and Urdu Numerals along with their Variants”. *International Conference on Computer Applications 2016*: 01-05. Print.