Research Journal of Recent Sciences _____ ISSN 2277-2502
Vol. **3(11),** 98-102, November **(2014)** *Res.J.Recent Sci.*

Offline Urdu Numeral Recognition Using Non-Negative Matrix Factorization

Shahab Uddin, Muhammad Sarim, Abdul Basit Shaikh and Sheikh Kashif Raffat
Department of Computer Science, Federal Urdu University of Arts, Sciences and Technology, Karachi, PAKISTAN

Available online at: www.isca.in, www.isca.me
Received 1st December 2013, revised 17th March 2014, accepted 25th July 2014

Abstract

By the rapid change and advancement in technology a need for processing and preserving many texts had been felt. These texts are either in hard copies or in handwritten form. Hand-written numerals, written in various languages and scripts, are an integral part of these texts. Several efforts have been made to recognize numerals and a variety of Optical Character Recognition (OCR) systems have been successfully implemented and marketed. Urdu numerals, as opposed to English numerals, are different due to their style and format of writing. Various methods have been proposed but majority of them only address computer typed numerals in different forms and sizes. Therefore we need to develop new and enhance existing handwritten Urdu numerals recognition systems due to their wide scale use and application in many fields. This research addresses the problem of handwritten offline numerals. A novel approach of Non-negative Matrix Factorization (NMF) for Urdu handwritten character recognition has been proposed in this research.

Keywords: Urdu, handwritten, numeral, recognition, optical character recognition, offline, NMF, OCR.

Introduction

For the purpose of recognition, numerals can be classified into two classes. i.e. On-line and off-line. The word on-line suggests that the writing and recognition are carried out simultaneously. While in the case of off-line recognition, a digital image is presented to the system. Moreover, we use off-line recognition for printed and handwritten numerals recognition. On-line recognition has an added advantage of time coordinate that is not available in case of off-line recognition[1]. For a specific font type, printed numerals have only one style whereas styles and sizes vary in case of handwritten numerals for the same writer at different instances and between different writers. Furthermore, if we compare Urdu numerals to English numerals we find that Urdu numerals are written from right to left whereas English numerals are written from left to right. Handwritten numerals may look similar but they are different. It is also hard for the recognition system to spot the dissimilarity. Additionally, the length and width of the numerals can also be different. Moreover, same numerals can be written differently in various forms. In an automatic recognition system, the selection of feature extraction method might be the most important step for achieving high recognition accuracy.

In multivariate analysis and linear algebra, Non-negative matrix factorization (NMF) is a matrix decomposition and dimension reduction technique based on low-rank approximation which makes use of a range of algorithms. Two such algorithms are based on multiplicative update rule and alternating least-squares algorithm. NMF reduces the number of features along with the constraint that the features will have to be nonnegative. If A is a matrix then NMF of A gives two factor matrix (not unique) namely W and H. Where W may be called weight matrix and H may be called basis matrix. Mathematically,

$$NMF(A) \rightarrow W \times H \quad (1)$$

Where A is $m \times n$ matrix and j is a positive integer such that

$$min(m,n) > j \quad (2)$$

Thus NMF returns a non-negative matrices W of size $m \times j$ and H of size $j \times n$ which are approximate non-negative factors of A that minimize the root mean square residual D where:

$$\mathrm{D} = \sqrt{\frac{|A - W \times H|}{\mathrm{N} \times \mathrm{M}}} \quad (3)$$

Where j columns of W show transformations of matrix's data while the j rows of H show the coefficients of the linear combinations of n data in A which had resulted in the form of transformed data in W. Since $j <$ rank (A) thus the product $W \times H$ gives condensed approximation of A.

NMF is similar to Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) but the constraint with NMF is that it has non-negative factors. NMF had been used by many researchers in many fields for both the feature extraction and classification.

Related Work: Sagheer et al.[1] prepared the Urdu datasets of isolated digits, numeral strings, alphabets, dates, isolated letters and special characters. Normalized images of 36 x 36 pixels were used to extract 32-direction gradient maps. The image is then divided into 9 x 9 blocks by extracting 4 x 4 feature values from each block. Gaussian filtering was used to down sample the directions and blocks to get the feature of 400 dimensions. Support Vector Machine (SVM) with the Radial Basis kernel Function (RBF) was used as a classifier.

Das et al.[2] exploited the 72 shadow features and 16 centroid features of normalized 32 x 32 pixel bounded boxed images. The image is divided into octants. Considering the normalized lengths (i.e. the length of each projection divided by the length of maximum possible length of projection on respective side) of projections on three sides of each such octant, 24 shadow features are extracted from each window of the digit image. Then three such overlapping windows are considered that make the number of shadow features equal to 72. Furthermore, the x and y coordinates of centroids of images in the octant window are added to the feature set. Finally, multi-layer perception was used as a classifier.

Razzak et al.[3] proposed fuzzy rule base, Hidden Markov Model (HMM) and Hybrid approaches for Urdu and Arabic numerals in unconstrained environment. For Persian numerals, Harifi and Aghagolzadeh[4] used asymmetrical segmentation patterns and shadow coding feature extraction and multi-layer perceptron (MLP) for classification. For Kannada, Tamil, Telugu and Malayalam, the Indian scripts based languages, Rajashekararadhya and Ranjan[5] presented centroid (zone and image based) distance for the extraction of features. Stuti Asthana et al.[6] conducted numeral recognition of Urdu, English, Tamil, Devnagri and Telugu scripts through multilayer feed-forward back-propagation algorithm.

Shuwair Sardar and Abdul Wahab[7] had developed a system which was tested on 1050 individual urdu characters and liguatures and tested it on both online and offline characters. They have claimed an accuracy overall accuracy of 97.12%, 97.09% accuracy in extracting the lines of text and 98.86% accuracy in primary and secondary stroke extraction. Pal and Sarkar[8] took advantage of water-reservoir features and binary-tree classifier for classification of 3050 characters and numerals and achieved 97.8% accuracy. According to Akram and Hussain[9], converted segments of text to a word sequence. Here, space, colon, semi colon etc were assumed to be word separators. As it is the case with several scripts whiles other scripts do not have clear cut boundaries for the words. In the latter case, linguistic knowledge, lexicon and machine-learning[4] based approaches can be utilized for recognition.

Alaei et al.[10] utilized IFHCDB[9], an isolated Farsi and Arabic handwritten character data set, for character recognition. The data set consists a total of 52020 handwritten character samples out of which 36682 samples were considered for training and the remaining 15338 characters were used for testing. Various feature such as line-fitting information, intersection/junction/ endpoint, under-sampled bitmap, shadow, directional chain code and gradient were used. Moreover, SVM, Nearest Neighbour (NN) and k-Nearest Neighbours (k-NN) were utilized for the purpose of classification. Finally, results from the combinations of all of the above mentioned classifiers and feature sets were calculated. In conclusion, the gradient features in combination with SVM as a classifier showed the best results of 96.91% accuracy in recognition rate.

Mozaffari et al.[11] proposed a new method for isolated handwritten Farsi/Arabic characters and numerals recognition using fractal codes. Fractal codes represent affine transformations. Each fractal code contained six parameters, such as corresponding domain coordinates for each range block, brightness offset and an affine transformation, which were used as inputs for a multilayer perceptron neural network for learning and identifying an input. This method was robust to scale and frame size changes. Farsi characters (32 in number) were categorized to eight different classes. Each class comprised of structurally similar characters. According to experimental results, classification rates of 91.37% and 87.26% were obtained for digits and characters respectively.

Mowlaei et al.[12] used discrete Wavelet transform to produce Wavelet coefficients. These coefficients were used for classification. Haar wavelet was used for feature extraction. The features so extracted were given to Neural Network as an input. Eight classes of structurally similar characters were created. Recognition of 92.33% and 91.81% was attained for handwritten Farsi characters and numerals respectively. Mozaffari et al.[13] compared the method of fractal codes and wavelet transform. Though the wavelet transform proved to be 25 times faster than the fractal codes but there wasn't any much difference in the recognition rates. Husain et al.[14] exploited various structural features of ligatures. Approximately 50000 words were extracted from these ligatures. The recognition rate of base ligatures was 93% and 98% for base ligatures, secondary strokes respectively.

Data Collection : First of all a two-page form was developed to get the input of Urdu numerals from a variety of people. A specimen of the form can be seen in figure-1 and figure-2. As it can be seen that the first page of the form was divided in to three portion and three rectangular boxes were drawn. The first page had some Urdu pre-printed words to take some personal information and guide the user to input the numerals. While the second page contained just two boxes.

The first and the medium size box showed both the English and their equivalent Urdu numerals to serve as an illustration and guide for the person filling the form. This was very essential in our case since it has been observed that most of the Urdu, Sindhi, Arabic and Persian and some other languages' numerals are almost the same and the remaining digits are similar in shapes and thus causing the confusion for bi-lingual or tri-lingual people at several occasions to distinguish between respective languages' numerals. The second box from top was for the input of National Identity Number to get some random numerals but few people filled the small squares within it as it was either not applicable to them due to their age or they didn't felt secure to disclose such kind of information publicly.

The third and the bottom most boxes were divided into ten columns and each column had a printed Urdu numeral typed within its top most boxes. Each column contained five (05)

empty boxes beneath it to take five inputs from our writers. The two boxes of second page were equal in size and were similar to the third box of first page with the exception that no numerals were written on the top of both of the boxes. The writers were allowed to freely input their own numerals. Basically the data was collected from three groups of people based on their gender, age, education and mother tongue. Input was taken from approximately 800 people related to the above mentioned three categories. After that the forms were scanned using a high resolution scanner to get png images of the 2-page forms. Nearly 1600 pages were scanned to serve as an input for off-line Urdu numeral recognition process.

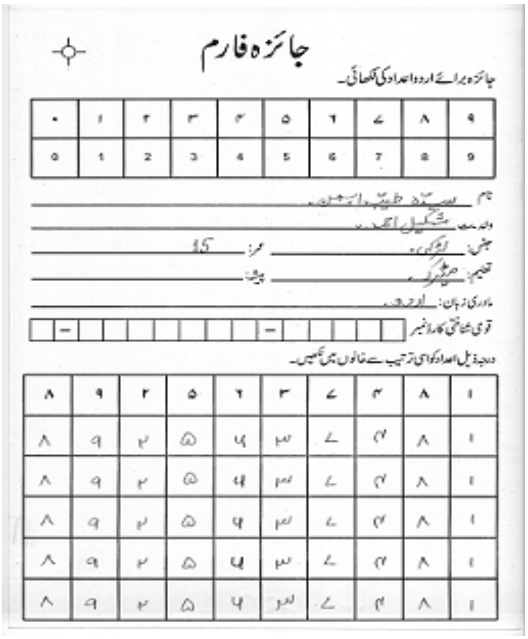

Figure-1
Form Page 1

Pre-Processing: For our purposes proper pre-processing of the forms was the most crucial and time taking process. The numerals from the first page of the form were used for training whereas the second page numerals were used for testing. It involved following steps: Removing noise - small sized and easily removable, locating the rectangular boxes, removing noise - large sized and often confused with the numerals, identifying the numerals and isolating each numerals image from the forms, padding the numerals with appropriate margins to preserve their orientations, normalizing the numerals Images to 175 x 175 image size, identifying, naming the first page numerals only and saving the numerals.

Data Preparation: After the pre-processing, we had two data sets of images. The first dataset was from the first page's numerals images each of size 175 x 175 pixels. Each such image had already been properly named according to its numeral. i.e. each image's naming convention is designed in such a way that it bears the numeral number in its first place. Then its position starting from the top-left to the bottom-right of each box. Then the box number and finally the form number to which each image belonged to. Each naming part separated by a dash, as shown in table-1. This naming scheme was necessary since it had been used to measure the performance of our method.

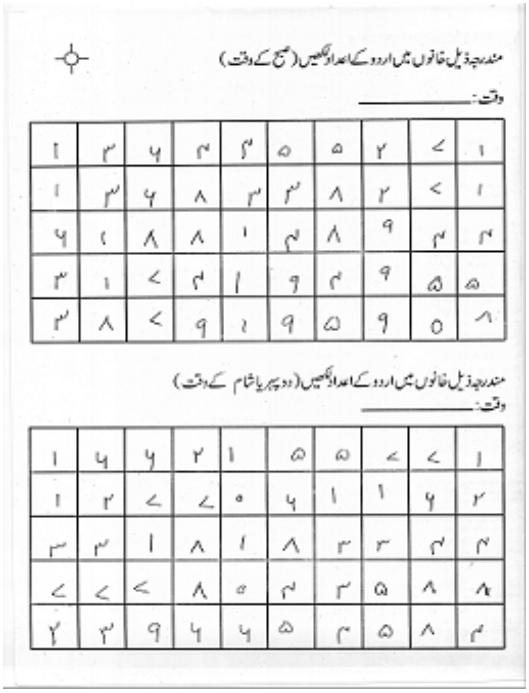

Figure-2
Form Page 2

Table-1
Naming System of Numerals Saved from First Page

7	39	1	1258
Numeral Number	Position	Box Number	Form Number

The second dataset was from the second page's numerals images each of size 175 x 175 pixels also. No such naming convention had been used for this dataset as in the case of the first page rather an arbitrary naming convention was used. Training set was prepared from the first dataset and Testing set from the second data set. Each image of size 175x175 pixels was vectored and converted into column image. Thus each image amounted to 30625 values in a column. Let us say that N and M images were taken from the training and testing sets respectively. As a result, we had the training and testing matrices of sizes 30625 x N and 30625 x M respectively. We had used the label of A for training set and B for testing set.

Feature Extraction: Now the technique of NMF is to be applied on matrix A to reduce the matrix dimension. As a result, we get two matrices namelyAw and Ah. Whereas Aw stands for the weight matrix while Ah stands for the basis matrix of A after its decomposition.

$$NMF(A) \rightarrow Aw \times Ah \quad (4)$$

Where j columns of Aw show transformations of matrix A's data while the j rows of Ah show the coefficients of the linear combinations of n data in A which have resulted in the form of transformed data in Aw.

The rank of approximation to be obtained is termed by j. It also specifies the desired level of decomposition and the number of non-negative factors. This value is adjusted according to the required level of accuracy and satisfaction. The higher the value of j, the more close the low rank approximation. But this value does not exceed or even gets closer to the rank of A If j's value is set very close to the rank of A then NMF is of no use here. In this case, the purpose of such decomposition dies. j has to be less than the rank of A to provide the condensed approximation of A and the product $Aw \times Ah$ provides such condensation.

We had applied the function of NMF onto the matrix A to extract Aw, the weight matrix and h , the basis matrix using multiplicative update rule and alternating least-squares algorithm. A large number of experiments were conducted using different sample sizes and different j values for the proper approximation of the matrix A and finding the most suitable's value. It was found by trial and error that the most suitable value for j is 25 and it should be used for further investigations.

After calculating the low approximation weight matrix Aw and basis Ah matrix of the matrix A, we had used the weight matrix Aw to extract the basis matrix Bh of matrix B by dividing the matrix B by the matrix Aw. This is done through left division of B by Aw such that:

$$Bh = B \backslash Aw \quad (5)$$

Results obtained at $j = 25$ are summarized in the table-2

Table-2
Recognition Rates of Different Sized Datasets

Testing Data Set Size	Training Data Set Size			
	*	100	300	500
	100	67%	83%	86%
	300	73.67%	80.67%	84%
	500	73.8%	80.4%	85.8%

We first trained our algorithm on 100 and tested the results on 100, 300 and 500 images respective. Then the same procedure was repeated for 300 and 500 images which produced the table-2.

Classification: We had used L_2-Norm on the decomposed matrices Ah and Bh to classify the images. Those converted, reduced and approximated images in Bh which had the minimum difference with images in Ah were matched and classified according to their naming convention used in the table.1. This table mentioned the naming system for saving the training images in A matrix. i.e. the matrix A (and thus the matrix Ah) only contained the pre-classified images according to the numerals to which those images belonged to.

Figure-3 showed the results, in each of these figures the first image is from the database in the matrix. While the second adjoining figure besides it is the matched result from the second matrix B. The variation covered by the algorithm can be evidently seen from these figures.

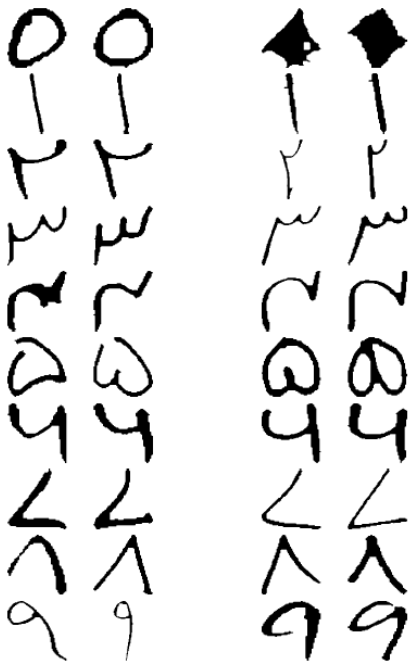

Figure-3
Results

Conclusion

Several techniques[15] including neuro-cognitive and probabilistic pattern recognition techniques[16] have been used for handwritten numeral recognition up till now but NMF is used for the first time for Urdu handwritten characters. Most of the prevalent techniques show good recognition rate but these techniques are not computationally efficient. NMF, on the other hand, allows us to do efficient computations due to its ability to reduce the matrix to a lower dimension approximation. Recognition rate of around 86% has been achieved through this technique.

References

1. Sagheer M.W., He C.L., Nobile N. and Suen C.Y., Holistic urdu handwritten word recognition using support vector machine, *Int. Conf. on Pattern Recognition*

(ICPR), 1900 –1903 **(2010)**

2. Das N, Mollah A.F., Saha S. and Haque S.S., Handwritten arabic numeral recognition using a multi layer perceptron, *National Conf. on Recent Trends in Inf. Sys.,* 200-203 **(2006)**

3. Razzak M.I., Hussain S.A., Sher M. and Khan Z.S., Combining offline and online preprocessing for online urdu character recognition, *Int. MultiConf. of Engineers and Computer Scientists*, **1**, 18–20 **(2009)**

4. Harifi A. and Aghagolzadeh A., A new pattern for handwritten persian/arabic digit recognition, *Int. Conf. on Info. Tech. (ICIT2004)*, Istanbul, Turkey, **(2004)**

5. Rajashekararadhya S.V. and Ranjan P.V., Efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south indian scripts, *J. of Theoretical and Applied Info. Tech.*, **4(12)**, 1171–1181 **(2008)**

6. Asthana S., Haneef F. and Bhujade R.K. , Handwritten multiscript numeral recognition using artificial neural networks, *Int. J. of Soft Comput. & Engin.*, **1(1)**, 1–5 **(2011)**

7. Sardar S. and Wahab A., Optical character recognition system for Urdu, *Int. Conf. Info. and Emerg. Tech. (ICIET)*, 14-16 **(2010)**

8. Pal U. and Sarkar A., Recognition of Printed Urdu Script, *7th Int. Conf. on Doc. Anal. and Recog. (ICDAR)*, **2**, **(2003)**

9. Akram M. and Hussain S., Word segmentation for urdu ocr system, *8th Workshop on Asian Language Resources*, Beijing, China, 88-94 **(2010)**

10. Alaei A., Pal U. and Nagabhushan P., A comparative study of persian/arabic handwritten character recognition, *Int. Conf. on Frontiers in Hand-writing Recognition*, **(2012)**

11. Mozaffari S., Faez K. and Kanan H.R., Recognition of isolated handwritten farsi/arabic alphanumeric using fractal codes, *Image Analysis and Interpre-tation, 6th IEEE Southwest Symposium*, 104-108 **(2004)**

12. Mowlaei A., Faez K. and Haghighat A.T., Feature extraction with wavelet trans-form for recognition of isolated handwritten farsi/arabic characters and numerals, *14th Int. Conf. Digital Signal Processing*, **2**, 923-926 **(2002)**

13. Mozaffari S., Faez K. and Kanan H.R., Feature comparison between fractal codes and wavelet transform in handwritten alphanumeric recognition using svm classifier, *17th Int. Conf. Pattern Recognition*, **2**, 331-334 **(2004)**

14. Husain S.A., Sajjad A. and Anwar F., Online urdu character recognition system, *Conf. on Machine Vision Applications (IAPR MVA)*, Tokyo, Japan, 16-18 **(2007)**

15. Sharif M., Shah J.H. Mohsin S. and Raza M., Sub-holistic hidden markov model for face recognition, *Res. J. of Rec. Sci.*, **2(5),** 10–14 **(2013)**

16. KhanY.D., Ahmad F. and Khan S.A., A Survey on use of Neuro-Cognitive and Probablistic Paradigms in Patterm Recognition, *Res. J. of Recent Sci.*, **2(4)**, 74-79 **(2013)**