

Recognition of Hindi (Arabic) Handwritten Numerals

Rawan I. Zaghoul, Dojanah M.K. Bader Enas and F. AlRawashdeh

Al-Balqa'a Applied University, Amman College, Amman, Jordan

Received 2012-03-26, Revised 2012-07-26; Accepted 2012-07-28

ABSTRACT

Recognition of handwritten numerals has been one of the most challenging topics in image processing. This is due to its contributions in the automation process in several applications. The aim of this study was to build a classifier that can easily recognize offline handwritten Arabic numerals to support those applications that are deal with Hindi (Arabic) numerals. A new algorithm for Hindi (Arabic) Numeral Recognition is proposed. The proposed algorithm was developed using MATLAB and tested with a large sample of handwritten numeral datasets for different writers in different ages. Pattern recognition techniques are used to identify Hindi (Arabic) handwritten numerals. After testing, high recognition rates were achieved, their ranges from 95% for some numerals and up to 99% for others. The proposed algorithm used a powerful set of features which proved to be effective in the recognition of Hindi (Arabic) numerals.

Keywords: Recognition, Binary Image, Hindi Numeral, Arabic Numeral, Optical Recognition System (Ocr), Projection, Classifier

1. INTRODUCTION

The development of Optical Character Recognition system OCR is considered one of the most important fields of research areas in pattern recognition. OCR allows a machine to automatically recognize characters through an optical mechanism. In other words, it is electronic translation for the images of handwritten numerals into computer textual format.

Recently, the recognition of handwritten numerals becomes an intensive area of research; in order to increase the functionality of OCR system. Numeral recognition systems can be utilized in several applications such as: check verification in banks, office automation, postal address reading and communication technology.

There are several approaches that deal with numerals/characters recognition problem, each approach depends on a set of features to be extracted and the ways of extracting them.

Handwritten numerals recognition is a hard task due to the restricted shape variations (In size, shape, slant and the writing style) and the different kinds of noise that break the strokes in numbers or change their topology. That's why we can see that handwriting varies when a person writes the same character twice. One can expect enormous dissimilarity among people. **Figure 1** shows a sample of standard and handwritten Hindi (Arabic) numerals.

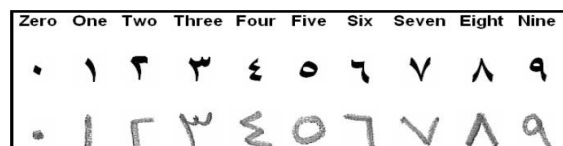


Fig. 1. Handwritten Hindi (Arabic) numerals versus standard of Hindi (Arabic) numerals

This study describes an off-line recognition technique for Arabic handwritten numerals by extracting features from numeric images to provide efficient and reliable results. The most important aspect of handwriting recognition scheme is the selection of a powerful set of features, which is reasonably invariant and robust with respect to the shape and slant variations that are caused by various writing styles.

2. MATERIALS AND METHODS

The recognition of handwritten text or number is a hard task because it depends on the writer and its accuracy. Thus, clear and accurate writing will help the OCR system to achieve very high recognition rates.

Hindi (Arabic) numerals are used by Arabs and Latin-based languages. Where, the term, Hindi (Arabic) numerals refer to the Indian numerals that are used in Arabic writing.

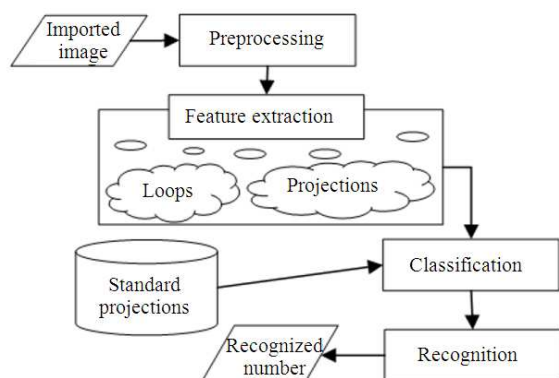


Fig. 2. Numeral recognition model

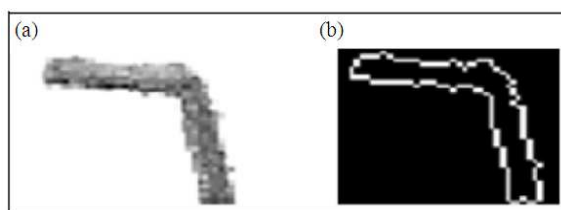


Fig. 3. (a) Imported image, (b) the preprocessed image

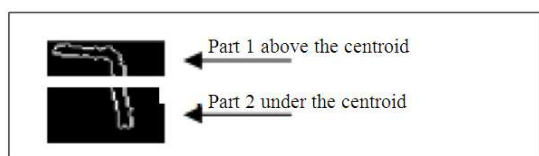


Fig. 4. Numeral segmentation according to the centroid

Sub image	Projection

Fig. 5. Analysis and Projections for Numbers 6 and 7 respectively

2.1. General Outline of the Proposed Approach

As depicted in Fig. 2, the proposed model is composed of four steps: importing numeral image, preprocessing, 2 extracting features, classification and finally, recognizing the imported numeral.

One	١		Two	٢	
Three	٣		Four	٤	
Six	٦		Seven	٧	
Eight	٨				

Fig. 6. Standard Hindi (Arabic) Numerals and their Projections

2.2. Image Preprocessing:

This step starts by applying the preprocessing techniques to the imported image, the preprocessing step include a set of operations: binarization, removing noise, edge detection. Figure 3 shows an example for imported numeral image and its preprocessing result.

2.3. Feature Extraction

Checking the existence of loops: Several techniques can be applied to detect loop in an image. In this study, the technique that is proposed by Kim *et al.* (2009) was applied.

Finding the Centroid of the image: the centroid is a useful feature that is used to describe the central weight of objects in image. Centroid is calculated by computing: $\{\text{Mean}(X), \text{Mean}(Y)\}$, where X and Y are the pixel's coordinates of the Numeral image.

Image Segmentation: is the process of partitioning a digital image into multiple parts or sub images. Segmentation is used to simplify and/or change the representation of an image into something that is more meaningful to analyze. More precisely, in this study, we suggest to divide the numeral image into two parts (sub images) according to its centroid value. An example of partitioning is shown in Fig. 4.

Horizontal projections: Another feature is suggested in this study, for each sub image the horizontal projection (projection on x-axis) is determined. Figure 5 shows examples of numeral projections.

2.4. Image Classification

In this step, the resulted features (loops or projections) are used to recognize the numeral. This is achieved by comparing the resulted features with the features of standard Hindi (Arabic) numerals as shown in Fig. 6 and 7 respectively.

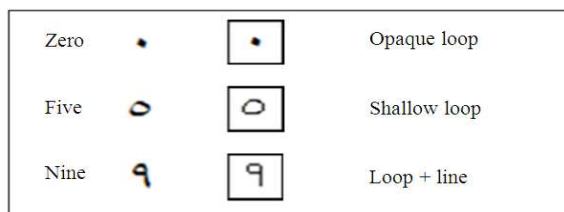


Fig. 7. The features of the standard Hindi (Arabic) numerals that contains loops

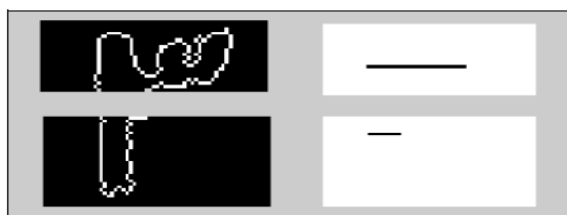


Fig. 8. Wrong Detection for Number “three”



Fig. 9. Correct detection of number three

2.5. Proposed Algorithm

Our proposed method is composed of the set of operations that are discussed earlier. The details of the proposed algorithm are illustrated in the following steps:

- Read the image of the handwritten numeral
- Convert the image into black and white (binarization)
- Apply the thinning and edge detection techniques over the image
- Check for the existence of loops. If so, then the detected number is one of the following {"five", "zero", "nine"}
- If the number is a filled loop then it is zero, while, if it contains a shallow loop then it is either "five or nine". So, to distinguish between them, "Nine" contains a line and a shallow loop, while "Five" is only a shallow loop as explained in **Fig. 5**
- If no loops exist. Compute the centroid for the image.
- Divide the image according to the centroid point.
- Find the projection for the generated images
- Finally, to recognize the number correctly, compare your projection results with the standard set of projections that are shown in **Fig. 6**

Table 1. Detection rates for characters without secondary parts

Number	Recognition rate
Zero	98%
One	98%
Two	98%
Three	95%
Four	97%
Five	98%
Six	98%
Seven	99%
Eight	99%
Nine	98%

Actually, by applying these steps good results were achieved, but sometimes an error may occur in detecting number “three”. It sometimes detected as “two” this is according to the position of the centroid point, as shown in **Fig. 8**.

Thus, to increase the robustness of the system, we propose that, if the projection’s result of your numeric image is the same as of the number “two”, try to insure that the number is correctly detected. So, re-apply the steps 6 and 7 for the upper sub image only. So, if it results in projections like those in **Fig. 9**, then it is “three” but if it doesn’t then it is truly detected as “two”.

3. RESULTS

The experiments were applied over a collection of Hindi (Arabic) handwritten numerals which collected from a large number of people in different ages; to test the proposed model. All the experiments are implemented under Matlab environment.

The results of the proposed method were highly accurate; it reaches high recognition rates for several samples as shown in **Table 1**.

4. DISCUSSION

In this study we describe a new approach to off-line, handwritten numeral recognition. There are a lot of problems for recognition due to writing habits and instruments; we suggest a recognition method which is able to account for a variety of distortions due to eccentric handwriting.

Various methods have been proposed and high recognition rates are reported, for the recognition of English handwritten digits (Berkes, 2005; Liu *et al.*, 2004; Kussul and Baidyk, 2004; Tang, 2006). In recent years, many researchers have addressed the recognition of Arabic text, including Arabic numerals (Al-Omari and Al-Jarrah, 2004; Bouzlama, 1999; Salourn, 2001; Salah *et al.*, 2002; Alma’adeed *et al.*, 2004; Touj *et al.*, 2005).

Alfonse *et al.* (2010), presented a hybrid classifier for segmenting Arabic numerals. The classifier is built using both of the Multilayer neural networks and the decision trees. They reach accuracy about 83% (Alfonse *et al.*, 2010).

Mahmoud and Awaida (2009), suggested a technique for automatic off-line handwritten Arabic (Indian) numerals recognition, by using Support Vector Machines and Hidden Markov Models. They achieved average recognition rates about 99.83% and 99.00% using, the Support Vector Machines and Hidden Markov Model classifiers respectively (Mahmoud and Awaida, 2009).

Mahmoud and Abu-Amara (2010a; 2010b) proposed a technique for the recognition of off-line handwritten Arabic numerals using Radon and Fourier Transforms. They reach high recognition rates around 98% (Mahmoud and Abu-Amara, 2010a; 2010b).

5. CONCLUSION

In this study a robust algorithm for offline Hindi (Arabic) numerals recognition is proposed. Its robustness comes from the set of extracted features. In summary, the proposed model starts by extracting a set of features like: detecting the loops or dividing the numeral image according to its centroid point position, finally classify the number according to the shape of the horizontal projection, or the existing of loops. The experimental results of this model show high accuracy and recognition rates around 98% among all numerals.

6. REFERENCES

- Alfonse, M., M. Almosry and M.S. Barakat, 2010. Eastern Arabic handwritten numerals recognition. *Int. J. Comput. Elect. Eng.*, 2: 1793-8163.
- Alma'adeed, S., C. Higgins and D. Elliman, 2004. Off-line recognition of handwritten arabic words using multiple hidden markov models. *Knowl. Based Syst.*, 17: 75-79. DOI: 10.1016/j.knosys.2004.03.002
- Al-Omari, F.A. and O. Al-Jarrah, 2004. Handwritten Indian numerals recognition system using probabilistic neural networks. *Adv. Eng. Inform.*, 18: 9-16. DOI: 10.1016/j.aei.2004.02.001
- Berkes, P., 2005. Handwritten digit recognition with nonlinear fisher discriminant analysis. *Proceedings of the 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications, (ICANN' 05)*, ACM Press, Heidelberg, pp: 285-287.
- Bousslama, F., 1999. Structural and fuzzy techniques in the recognition of online Arabic characters. *Int. J. Recog. Artif. Intell.*, 13: 1027-1040.
- Kim, J., Y. Han and H. Hahn, 2009. Character segmentation method for a license plate with topological transform. *World Acad. Sci. Eng. Technol.*, 56: 39-42.
- Kussul, E. and T. Baidyk, 2004. Improved method of handwritten digit recognition tested on mnist database. *Image Vis. Comput.*, 22: 971-981.
- Liu, C.L., K. Nakashima, H. Sako and H. Fujisawa, 2004. Handwritten digit recognition: Investigation of normalization and feature extraction techniques. *Patt. Recog.*, 37: 265-279. DOI: 10.1016/S0031-3203(03)00224-3
- Mahmoud, S.A. and M.H. Abu-Amara, 2010a. The use of radon transform in handwritten Arabic (Indian) numerals recognition. *WSEAS Trans. Comput.*, 9: 252-267.
- Mahmoud, S.A. and M.H. Abu-Amara, 2010b. Recognition of handwritten Arabic (Indian) numerals using Radon-Fourier-based features. *Proceedings of the 9th WSEAS International Conference on Signal Processing, Robotics and Automation, (ISPRA' 10)*, ACM Press, USA., pp: 158-163.
- Mahmoud, S.A. and S.M. Awaida, 2009. Recognition of off-line handwritten Arabic (Indian) numerals using multi-scale features and support vector machines vs. Hidden Markov Models. *Arabian J. Sci. Eng.*, 34: 429-444.
- Salah, A.A., E. Alpaydin and L. Akarun, 2002. A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Trans. Patt. Anal. Mach. Intell.*, 24: 420-425. DOI: 10.1109/34.990146
- Salour, S., 2001. Arabic hand-written text recognition. *Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications, Jun. 25-29*, IEEE Xplore Press, Beirut, pp: 106-109. DOI: 10.1109/AICCSA.2001.933959
- Tang, Q., 2006. Two-dimensional penalized signal regression for hand written digit recognition. *Master's thesis, Louisiana State University.*
- Touj, S., N. Amara and H. Amiri, 2005. Arabic handwritten words recognition based on a planar hidden Markov model. *Int. Arab J. Inform. Technol.*, 2: 318-325.