HANDWRITTEN CHARACTER CLASSIFICATION USING THE HOTSPOT FEATURE EXTRACTION TECHNIQUE

Olarik Surinta, Lambert Schomaker and Marco Wiering
Department of Artificial Intelligence, University of Groningen, Nijenborgh 9, Groningen, The Netherlands
o.surinta@rug.nl, {mwiering, l.schomaker}@ai.rug.nl

Keywords: Handwritten Character Recognition, Feature Extraction, *k*-Nearest Neighbors, Classification

Abstract: Feature extraction techniques can be important in character recognition, because they can enhance the efficacy of recognition in comparison to featureless or pixel-based approaches. This study aims to investigate the novel feature extraction technique called the hotspot technique in order to use it for representing handwritten characters and digits. In the hotspot technique, the distance values between the closest black pixels and the hotspots in each direction are used as representation for a character. The hotspot technique is applied to three data sets including Thai handwritten characters (65 classes), Bangla numeric (10 classes), and MNIST (10 classes). The hotspot technique consists of two parameters including the number of hotspots and the number of chain code directions. The data sets are then classified by the *k*-Nearest Neighbors algorithm using the Euclidean distance as function for computing distances between data points. In this study, the classification rates obtained from the hotspot, mark direction, and direction of chain code techniques are compared. The results revealed that the hotspot technique provides the largest average classification rates.

1 INTRODUCTION

Feature extraction can play an important role in handwriting recognition. It is used for generating suitable feature vectors, and using them as representation of handwritten characters. The difference between handwritten characters and printed characters lies in the diversity of characters. In printed characters, the structural pattern of characters is always the same, therefore the main challenges are coping with different fonts, or scanning qualities. However, in handwritten characters, the pattern of characters is different, even for those of the same writer.

The main objective for using feature extraction is to reduce the data dimensionality by extracting the most important features from character images (Lauer et al., 2007). When the feature vector dimensionality is smaller, a set of features can be useful for representing the characteristics of characters. Moreover, feature extraction can play a significant factor for obtaining high accuracies in character recognition systems (Trier et al., 1996), especially if there is not a lot of training data available.

The present study aims to investigate a novel feature extractor for handwritten characters and other types of characters such as handwritten digits from different scripts. The main aim of this paper is to propose a fast and easy to use feature extraction method that obtains a good performance. This study focuses on isolated characters. Three data sets including MNIST, Bangla numeric, and Thai were used to test our novel proposed feature extraction technique. The hotspot technique is used to determine the distance along a particular direction between the hotspots and the first black pixel of the object. The hotspots are distributed at fixed locations over the character images. This technique extracts some important information from the character images and is fairly robust to translation and rotation variances. The important parameters of this technique are the number of hotspots and the number of chain code directions.

Related work. Sanossian and Evans (1998) proposed a scanning technique for English characters. They used 64×64 pixels of binary images. The feature values are calculated by scanning through the image in horizontal, vertical, and inner (inside the character) directions of character images. Ferdinando (2001) used a vertical and two horizontal directions for digits. The feature vectors from this technique are the positions of crossing points between each line. In addition, an interesting approach is a direction technique consisting of 4 feature windows, and 4 neigh-

bor marks. Firstly, Kawtrakul and Waewsawangwong (2000) used 4 feature windows including horizontal, vertical, left diagonal, and right diagonal directions. This technique found the contour of Thai character images. Subsequently, every feature window (3×3 pixels) is used to slide through all cells of character images. The feature vectors of this technique represent the number of perfectly matching windows in the feature window with the part of the character windows. Secondly, Pal et al. (2008) presented 4 direction neighbors. These directions are used to count a matching number of directions from contour images. Rajashekararadhya and Vanaja Ranjan (2009) suggested the application of a feature extraction algorithm for Kannada script. In this study, the zone and projection distance metric technique was proposed. The distance values from four different directions including vertical downward direction, vertical upward direction, horizontal right direction, and horizontal left direction were calculated.

Contributions of this paper. We propose a novel method for feature extraction which is suitable for isolated handwritten characters. We have used three different handwritten data sets from different scripts to compare our novel feature extraction method to two state-of-the-art techniques. The results show that our novel method significantly outperforms the other 2 methods on 2 data sets containing handwritten digits. Only on the Thai data set containing 65 classes, one other technique achieves higher recognition accuracies. The average recognition rate over the three data sets is also highest for our novel technique, which demonstrates its effectiveness.

2 DATA COLLECTION AND PRE-PROCESSING

The data sets used in the present study include Thai, Bangla numeric and MNIST (LeCun and Cortes, 1998). Figure 1 shows some examples of handwritten characters. Each data set consists of isolated characters. MNIST consists of 60,000 training examples and 10,000 test examples. MNIST is a handwritten numeric data set that has been widely used as benchmark for comparing feature extraction techniques (Lauer et al., 2007). In the present study, 10,000 records (10 classes) of the MNIST data set were used. For the Bangla numeric data set, 9,595 records (10 classes) are used. The Thai data set used in the present study includes 65 classes consisting of consonants, vowels and tones. There are 5,900 Thai examples in this data set. The Thai data set was collected from characters written by writers aged from 20-23 years old. Among this group of data, there are characters written by 7 female writers and 3 male writers.

Figure 1: Some examples of character images used in the present study. (a) Thai data set. (b) Bangla numeric data set. (c) MNIST data set.

Pre-processing starts off with cropping the exceeding parts of scanned images. The exceeding area is the non-character area so that there are only character pixels in the images. These images are then transformed into binary images. Consequently, the images are scaled to 40×40 pixels. Finally, the thinning technique is used to make the images absolutely thin and ready for feature extraction and classification.

3 HOTSPOT FEATURE EXTRACTION TECHNIQUE

Feature extraction can play a significant factor for increasing the efficacy of recognition systems (Trier et al., 1996). It is a process that extracts the important information from the character images and transforms them into vector data. When a feature extraction technique is applied, the dimensionality of the resulting feature vector is smaller in comparison with that of raw data (Lauer et al., 2007). Since the smaller feature vectors are afterwards used in a classification algorithm, with little training data they may suffer less from overfitting than pixel-based methods.

The hotspot technique is our novel method useful for representing the character. The distance between black pixels and the hotspots in each direction is used to describe the whole object. In this technique, the size of the hotspot was defined as $N \times N$. For example, the size of the hotspot can be 3×3 (Figure 2). The distance between black pixels and the hotspots from the first to the last hotspot is calculated. The direction of the hotspots is defined by the chain code directions (Figure 3). The hotspot feature vector P_s is defined as (Equation (1));

$$P_s = \{(x_s, y_s), \{d_i\}, \{D_{si}\}\} \quad (1)$$

Where (x_s, y_s) is the coordinate of the hotspot, $d_i \varepsilon \{0, 1, 2, ..., 7\}$ when chain code direction is considered as 8-directional codes, and D_{si} is the distance

between the hotspot and the first black pixel of the object found in the direction d_i. It is noted that, if there is no object pixel found then the distance D_{si} is set to d_{max}. The distance is measured by using the following equation (Equation (2));

$$D_{si} = \begin{cases} \sqrt{(x_s - x_i)^2 + (y_s - y_i)^2} & if\ (x_i, y_i)\ exists, \\ d_{max} & else \end{cases} \quad (2)$$

Where (x_i, y_i) is the coordinate of the closest black pixel of the object in the specified direction. As feature vector we only consider 4 or 8 values of every hotspot and then concatenate it into some specific order to create a feature vector. The complete notation of feature vectors can be defined as $f = \{D_{11}, \ldots, D_{1K}, \ldots, D_{L1}, \ldots, D_{LK}\}$, where L is the number of hotspots and K is the number of directions. The feature vector size depends on the number of hotspots and the number of directions.

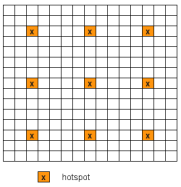

Figure 2: An example to illustrate the location and distribution of the hotspots.

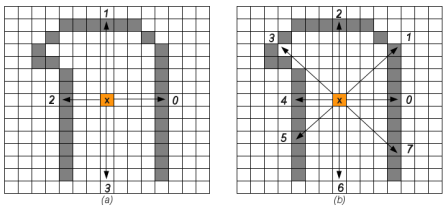

Figure 3: The chain code directions for identifying the distance, starting from the hotspot until the object in each direction was found, (a) 4 directions chain code and (b) 8 directions chain code.

There are two parameters that influence this method including 1) number of hotspots and 2) number of chain code directions. The preliminary results demonstrated that the best setting uses 25 hotspots and 4 directions, so that the hotspot technique provides 100 features.

4 EXPERIMENTAL RESULTS

We presented a novel method for feature extraction, called the hotspot technique, which is compared in this section to other methods including mark direction and the direction of chain code technique.

The mark direction technique is also known as the direction feature (Blumenstein et al., 2003) and suitable for tracking the directions of the character image. The mark size for the mark direction technique is 3×3 pixels (Kawtrakul and Waewsawangwong, 2000). The mark directions consist of horizontal mark, vertical mark, left-diagonal mark, and right-diagonal mark (Blumenstein et al., 2003). The number of features obtained from mark direction was 64 features.

The direction of chain code technique is an efficient technique in handwriting recognition (Bhowmik et al., 2007). We applied this technique to the present study according to the methods described by Pal et al. (2008), although we adapted their technique to deal more efficiently with our data sets by identifying the starting point of the direction in each block.

The feature vectors obtained from these techniques are classified by the k-Nearest Neighbors (k-NN) method. The outcome of the classification process is the classification rate. All different extractors were applied to three data sets including Thai, Bangla numeric, and MNIST. These three data sets were treated with the same methods so that the character image's size for all data sets is determined as 40×40 pixels.

The data sets were divided into 10 subsets (90% training set and 10% test set). We randomly divided the data into a test and training set 10 different times. The value of k of the k-Nearest Neighbor classifier was optimized for each method and dataset.

Table 1: Comparison of data classification efficacy of feature extraction techniques by using kNN.

Data set	Feature extraction technique		
	Hotspot	Mark direction	Direction of chain code
Thai	83.3 $\sigma = 0.5$	**88.0** $\sigma = 0.6$	71.3 $\sigma = 0.7$
MNIST	**89.9** $\sigma = 0.3$	85.1 $\sigma = 0.3$	83.5 $\sigma = 0.2$
Bangla numeric	**90.1** $\sigma = 0.4$	87.6 $\sigma = 0.4$	82.7 $\sigma = 0.4$

The size of the feature vectors obtained from hotspot, mark direction, and direction of chain code technique were 100, 64, and 128 dimensions, respectively, For the hotspot technique we used $d_{max} = 20$ for the two data sets containing digits, and $d_{max} = 0$ for the Thai data set, which worked slightly better for an unknown reason than $d_{max} = 20$.. Table 1 shows the comparison of classification efficacy of the different feature extraction techniques. It is found that the best feature extraction technique for classification is hotspot, followed by mark direction and direction of

chain code, respectively. The average classification rate obtained from hotspot, mark direction, and direction of chain code are 87.8%, 86.9%, and 79.2%, respectively. Our new technique significantly outperforms the other feature extraction method on the two data sets containing digits. The mark direction technique outperforms our method on the Thai data set. The direction of chain code technique obtains the worst performance by far. This technique is more complicated and involves several subtleties which requires adapting it to different data sets. Much better results for MNIST have been reported in literature (above 99% accuracy), but in those studies more training patterns were used (60,000 compared to 10,000 in our study). This dataset has a very large number of examples and few classes, which makes pixel-based methods more effective. However, we believe that by more fine-tuning, using more examples and better classifiers, and combining multiple feature extraction methods, we are able to obtain similar performances.

5 CONCLUSIONS

The present study proposed a new technique for feature extraction, named the hotspot technique. In this technique, the distance values between the closest black pixels and the hotspots in each direction are used as representation for a character. There are two key parameters to be taken into account; 1) number of hotspots and 2) number of chain code directions. The hotspot technique was applied to numeric data sets including MNIST and Bangla numeric, and Thai characters.

For the two data sets with few classes, namely the handwritten digit data sets, Bangla and MNIST, the novel hotspot technique significantly outperforms the other methods. However, the mark direction technique outperforms the hotspot technique on the Thai data set that has much more classes (65). Maybe the hotspot technique needs more examples for this data set, possibly because it is less robust to variances in the handwritten characters than the mark direction technique. Still, our results on data sets of multiple scripts show that the hotspot technique achieves the highest average recognition rate.

In future work, we want to compare different feature extraction techniques, among those the ones described in this paper, to pixel-based methods. Several neural network architectures have obtained very high recognition rates on the MNIST data set, and we are interested in finding the utility of feature extraction methods compared to the use of strong classifiers that immediately work on pixel representations. Furthermore, keypoint methods have not deserved a lot of attention in handwriting recognition, and we want to explore the use of adaptive keypoints to be more translation invariant and also use generative models to maximize the probability of generating the data.

ACKNOWLEDGMENTS

We are sincerely grateful to Dr. Tapan K. Bhowmik for providing the Bangla numeric data used in the present study. We thank Jean Paul van Oosten for useful remarks on a preliminary version of this paper.

REFERENCES

Bhowmik, T. K., Parui, S., Kar, M., and Roy, U. (2007). HMM parameter estimation with genetic algorithm for handwritten word recognition. In Ghosh, A., De, R., and Pal, S., editors, *Pattern Recognition and Machine Intelligence*, volume 4815 of *Lecture Notes in Computer Science*, pages 536–544. Springer Berlin / Heidelberg.

Blumenstein, M., Verma, B., and Basli, H. (2003). A novel feature extraction technique for the recognition of segmented handwritten characters. *Document Analysis and Recognition, International Conference on*, 1:137.

Ferdinando, H. (2001). Handwriting Digit Recognition With Fuzzy Logic. *Jurnal Teknik Elektro*, 1(1).

Kawtrakul, A. and Waewsawangwong, P. (2000). Multifeature extraction for printed thai character recognition. In *Natural Language Processing, 2000. SNLP 2000. 4th International Conference on*.

Lauer, F., Suen, C. Y., and Bloch, G. (2007). A trainable feature extractor for handwritten digit recognition. *Pattern Recognition*, 40(6):1816–1824.

LeCun, Y. and Cortes, C. (1998). The MNIST database of handwritten digits.

Pal, U., Sharma, N., Wakabayashi, T., and Kimura, F. (2008). Handwritten character recognition of popular south indian scripts. In Doermann, D. and Jaeger, S., editors, *Arabic and Chinese Handwriting Recognition*, volume 4768 of *Lecture Notes in Computer Science*, pages 251–264. Springer Berlin / Heidelberg.

Rajashekararadhya, S. and Ranjan, P. (2009). Zone based feature extraction algorithm for handwritten numeral recognition of kannada script. In *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pages 525–528.

Sanossian, H. and Evans, D. (1998). Efficient feature extraction technique for english characters. *International Journal of Computer Mathematics*, 66(3-4):257–265.

Trier, Ø. D., Jain, A. K., and Taxt, T. (1996). Feature extraction methods for character recognition-a survey. *Pattern Recognition*, 29(4):641–662.