Chinese Characters Recognition Based on HALCON*

Guangrun Zheng, Kaicheng Li, Lei Yuan
National Engineering Research Center of Train Control System, Beijing Jiaotong University, Beijing 100044, China
11120329@bjtu.edu.cn

Abstract - This paper presents a fast Chinese character recognition method based on HALCON image processing software. After character image pre-processing, dynamic threshold segmentation method combined with region morphology is used to segment characters, and then construct the feature vector. At last, improved artificial neural network classifier is applied to classify and identify characters. Experimental results show that this method can accelerate the speed Chinese characters recognition system, improve the Chinese characters recognition rate and has strong practicality and feasibility**.**

Index Terms - Chinese Characters Recognition; Dynamic Threshold Segmentation; Regional Morphology; Artificial Neural Network

1. Introduction

Optical character recognition technology (OCR) is an important branch of pattern recognition, which combines the knowledge of the various aspects, such as computer graphics, digital signal processing and artificial intelligence, is an integrated technology having a wide range of applications in the computer and its related fields [1].

Commonly used Chinese character segmentation methods include the projection method, template matching and domain connecting method. Projection method and the traditional domain connecting method is not a better solution of the Chinese characters disconnected problem, the template matching is too dependent on a priori knowledge when producing templates [2]. To solve these problems, dynamic threshold segmentation method combined with region morphology and connected region segmentation method is proposed in this paper, a better way to solve the problem of Chinese characters disconnected.

After segmenting Chinese characters and extracting characters features, improved artificial neural network classifier is used to classify and identify characters [3]. The artificial neural network classifier is fast at classification, a good choice if the training can be applied offline and thus is not time critical.

Image processing software HALCON is comprehensive standard software for machine vision developed by the German MVTec company. It is a set of image processing library with more than 1600 operators for Blob Analysis, Morphology, Pattern Recognition, Matching, Measuring, Identification, and 3D vision, to name just a few [4]. The interactive programming environment can be used for rapid development of machine vision applications, or new operators can be added to integrate their visual function. Introduced with HALCON's powerful computational analysis capability, Chinese characters recognition is greatly simplified, and produces good results.

2. Progress of Chinese Characters Recognition

As shown in Fig. 1, the character recognition process consists of training stage and recognition process. In the learning process, known as the offline process, comprises the training of the font, i.e., regions that represent Chinese characters (in the following just called 'characters') are extracted and stored together with the corresponding character names in training files, making it easy to find the errors that occurred during the training on the one hand and the other one hand, you can reuse the contained information for the case that you want to apply a similar application in the future. Now the training files are used to train the font.

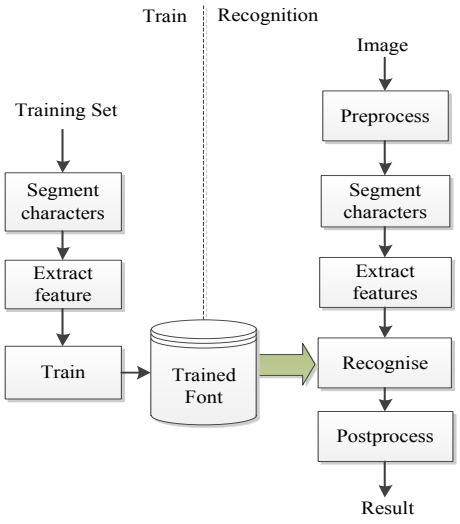

Fig. 1 Characters Recognition Progress

In the recognition process, known as the online process, the regions of unknown characters are extracted from image input, most suitably by the same method that was used within the offline training process, and classified, i.e., the characters are read.

3. Recognition Program Design and Implementation

In the following, we illustrate the proceeding for Chinese OCR classification with characters '人', '机', '控', '目', '完', '全', '视', '觉'as example. Note that the number of classes as

* This research was partially supported by the project of National High-tech R&D Program of China (2012AA112801)

well as the number of training samples is very small as the example is used only to demonstrate the proceeding and how to design and implement characters recognition with HALCON. Typically, a larger number of classes are trained and a lot of samples and probably a different set of features are needed to get a robust classification. Figure 7.1 shows the training images set for the characters ‘人’, ‘机’, ‘控’, ‘目’, ‘完’, ‘全’, ‘视’, ‘觉’ .

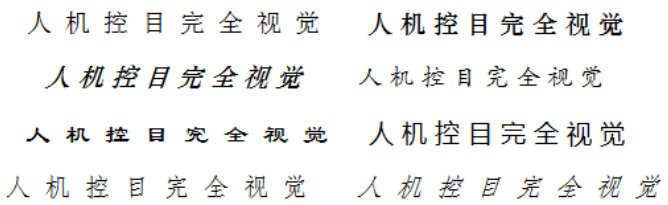

Fig. 2 Training Images Set

A. *Characters Segmentation*

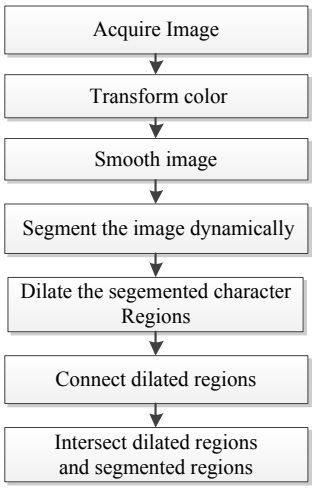

Fig. 3 Characters Segmentation Progress

Fig. 3 shows the proceeding for segment characters. Within HALCON application, the task of image acquisition is thus reduced to a few lines of code, Reading images from files is even simpler: It consists of a single call to the operator read_image. Then, use the operator rgb1_to_gray to transforms the RGB image into a gray image ,which has gray values ranges of 0 - 255 (see Fig. 4 (a)).

Before segment the characters, operator mean_image should be called to smooth the gray image by averaging. The smoothed image (see Fig. 4 (b)) is used as an input object to operator dyn_threshold, which segment an image using a local threshold. The characters can be segmented easily with the dynamic thresholding operation,which can suppress noise in the segmentation better(see Fig. 4 (c)).

If we compute the connected components of the dynamice thresholded region in Fig. 4 (c), we can see that the characters and their strokes are separate components (see Fig. 4 (d)). To solve this problem, we first need to connect the strokes with their characters. This can be achieved using operator dilation_circle of region morphology operations to dilate the characters(see Fig. 4 (e)). With this,the correct connected components are obtained by using operator connection (see Fig. 4 (f)). Unfortunately, they have the wrong shape because of dilation. To correct them, we should use operator intersection to intersect the regions with the dynamic segmented regions . Fig. 4 (g) shows that,we have obtained the correct regions of each characters.

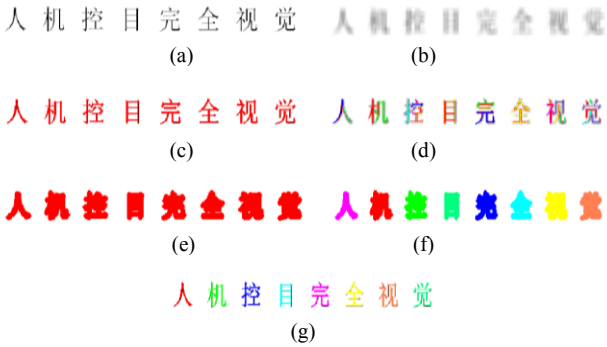

Fig. 4 Results of Each Step of Characters Segmentation Progress

B. *Train OCR*

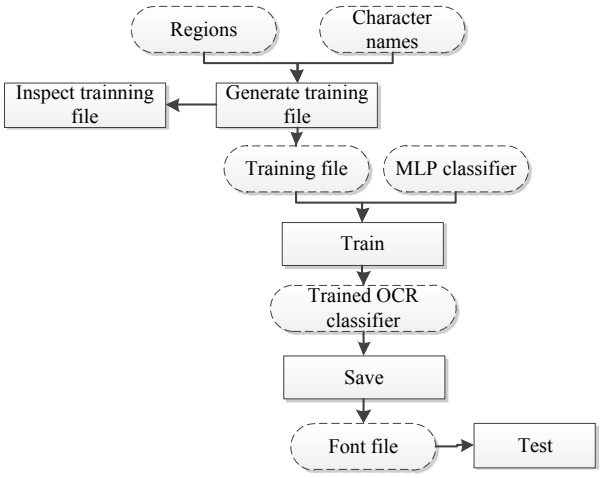

Fig. 5 Training OCR

Fig. 5 shows an overview on training OCR. Firstly, to each of the single characters extracted from sample image a name already known must be assigned. This can be done either by typing it in. Secondly, the regions together with their names are written into training files. The most convenient operator to do this is append_ocr_trainf. Before applying the training, we had better check the correctness of the training files by using the operator read_ocr_trainf combined with visualization operators. Thirdly, create a neural network (multi-layer

perceptron or MLP) classifier using create_ocr_class_mlp. Fourthly, the training is applied using trainf_ocr_class_mlp. After the training, we typically save the classifier to disk for later use by write_ocr_class_mlp.

C. *Characters Recognition*

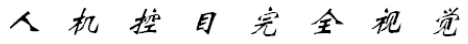

Fig. 6 Image to Recognize

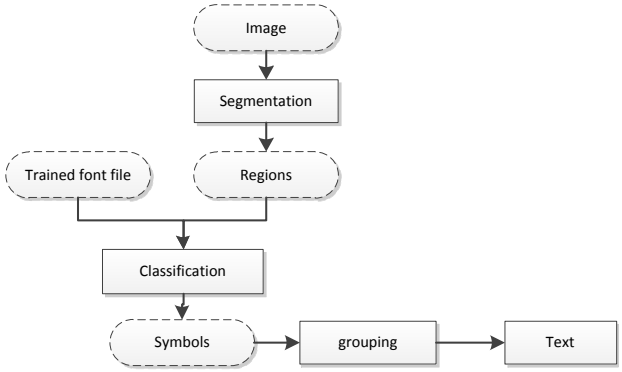

Fig. 7 Recognition Process

Fig. 6 shows the image to recognize using the font file trained above while Fig. 5 shows an overview on the recognition process. Firstly, the characters must be extracted using method that returns the characters in a form similar to the ones used for training. After reading the classifier (font file) from file using operator read_ocr_class_mlp, the classifier can be used for recognition. The segmented charecters are past to the reading operators do_ocr_multi_class_mlp or operators do_ocr_single_class_mlp for classification.If operators do_ocr_multi_class_mlp is used, for each region the corresponding name and the confidence are returned. While if operators do_ocr_single_class_mlp is used, not only to return the characters with the highest confidence but also others with lower confidences, which is sometimes necessarily when the characters are easily mistaken. As a final step it might be necessary to group characters to words. This can be realized with the region processing operators.

As shown in Fig. 8, the recognition results is displayed.

Fig. 8 Results and Display

4. Conclusion

Chinese characters recognition is a very active field for research and development in OCR, and has become one of the most successful applications of automatic pattern recognition [5]. Chinese character recognition consists of two main tasks: segmentation and classification. Character segmentation is a main pre-processing step. In this paper dynamic threshold segmentation method combined with region morphology method is put forward to solve the disconnected problem in Chinese characters segmentation. Improved artificial neural network classifier in HALCON is applied in classification. From the results shown above, these proposed methods are actually experimentally proved to be fast and efficient. But How to identify Chinese characters having noise interference and distortion, such as hand writing with HALCON should be our further study task.

5. Acknowledgment

The research was supported by the project of National High-tech R&D Program of China (2012AA112801), the Research Funds for the Central Universities (2012JBM024), and National Engineering Research Center of Train Control System.

6. References

[1] Kejun Wang, Weixing Feng, *Chinese printed document identification technology* ,Beijing: Science Press ,2010, pp.1-3.

[2] R.G.CASEY, E. LECOLINET, "A survey of methods and strategies in character segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 690-706, July 1996.

[3] Naiping Hu, Li Wang, "The design of developed neural networks arithmetic in identifying the character," Automation& Instrumentation, vol. 2, pp. 12-13, 16, February 2002.

[4] Jianhang Huang, "Implementation of Character Recognition in Annular Region Based on HALCON," Modern Computer, vol. 7, pp. 58-60, July 2010.

[5] Weiguang Liu, *Image Fusion and Recognition*, Beijing: Publish House of Electronics Industry, 2007, pp.196-198.