A Selection Algorithm of Training Set Based on Similar Classification

Xiaowen Liang
Department of Computer, Communication University of China, Beijing, China
Email: 158887513@qq.com

Wei Gong, Wenlong Fu
Department of Computer, Communication University of China, Beijing, China
Email: {gongwei@cuc.edu.cn, fwl2000@163.com}

Jing Qi
Beijing, China
Email:183135570@qq.com

***Abstract*—License Plate Recognition (LPR) combines computer vision technology and pattern recognition technology and plays an important role in Freeway Toll System, Urban Road Monitoring System and the Intelligent Parking Lot Management System. Therefore, it has attracted an ever increasing number of scholars from home and abroad. Despite many years of unremitting effort which has resulted in breakthrough achievements, it remains unsatisfactory in meeting real world application requirements. LPR primarily employs pattern recognition and digital image process technology. This paper is focused on the study of pattern recognition. The segmented characters are trained utilizing the BP neural network. Selecting the ideal training set from the usually large sample set we have is the first step to train a good network which has a high recognition rate. At present, training sets are randomly selected, which affects the accuracy of recognition as well as its speed. Thus, selecting the best training sets is of uttermost importance. In this paper, Similarity Comparison Sampling method is proposed to improve the training results.**

***Index Terms*—training set, selection algorithm, neural network, character recognition**

I. Introduction

License Plate Recognition is a specific application of the character pattern recognition technology in the traffic field, which includes a serial of steps ranging from image capture, license plate locating to character segmentation and character recognition. After initial stage of the last century, this technology has become mature and has found broad applications in the public security, parking, security and other fields.

However, the domestic research is still in a fledging period, and existing LPR systems have exposed many a flaw such as unstable all-weather recognition rate resulted from changing conditions and lighting in the image acquisition of the vehicle. License plate may be stained, faded and tilted and motion may lead to blurry and fuzzy image. All this may contribute to difficulties in implementation of location algorithm and recognition algorithm. Currently, there is no algorithm that covers all circumstances. Individual algorithm has to be developed to deal with specific situations.

Some of the more or less established LPR methods are based on either neural network, or support vector machine, or key point template match, or character recognition by contour features, etc. Either way, a large data set is needed. Different sample of the same character may vary. Thus it is worthwhile to select training sample to represent the character from many façade. The first step to train a good neural network with satisfactory results starts from selecting the best training sample.

II. Selection Method of Training Set

Current pattern recognition algorithm depends mostly on the training set and its size and distribution have a great impact on the performance of the final classifier. When the feature dimensions of training sample exceeds the size of sample set, the LPR systems based on BP neural network have an inclination of easy over-fitting, i.e. high recognition rate on the training set with low recognition on the test set. This is a classic example of small sample learning problem, which is caused by the following reasons:

- The presence of noise in the training data;
- Lack of representation in the training data ;
- Insufficient training data.

In the process of selecting categorical training set, balanced volume of data is required in each category. Too many training set usually takes up too much memory and slows the training process down.

Commonly used methods to select training set are:

Bagging, also known as self-aggregation, is a technology that repeatedly samples from uniformly distributed data set. Allow each self-aggregated sample set to be the same size as the original data and train a base classifier for each sample set. After training k number of

doi:10.4304/jsw.9.3.726-731

classifiers, the test sample is assigned to the class of highest number of vote. For the noise data, bagging is less susceptible to the effects of over-fitting.

Lifting, an iterative process, is used to self-adaptively change the distribution of the sample so that the base classifier is focused on the hard-to-distinguish samples[1].

III IMAGE PREPROCESSING

Ideally, the character will have a high recognition rate if the plate in which located is clear, untitled and noiseless. Unfortunately, It might make image blur, skew and defect due to weather, light intensity, taking angle and other causes, thus, if the plate is not preprocessed, the recognition rate of the character looks bad. In order to improve the recognition rate of the character and adapt to all kinds of terrible cases, the plate have to be made a optimize processing. Image Processing plays a critical role as the first step of character recognition system, typically, it includes the following steps: 24-bit color images will be converted to 8-bit gray image, and process with gray stretch, binarization, slant correction and eliminating noise. These steps will be described in detail in the following paragraphs.

A. Color Image Transform into Gray Image

In general, Obtained the license plate is a color images based on RGB model, three matrices have be used to save the image in the computer, the three color components R, G, B are stored in the three matrices, of which each element in the range [0, 255]. The process of a color image will be converted to gray image is three-dimensional color space mapped into one-dimensional color space by a linear mapping T (R, G, B), Its form is as follows:

$$T(R, G, B) \rightarrow Y \tag{1}$$

$$Y = aR + bG + cB \tag{2}$$

Wherein a, b, c are constants.
In this paper, the conversion of the form:

$$Y = 0.299R + 0.58G + 0.114B \tag{3}$$

The results are shown as follow:

Figure 1. The color plate image

Figure 2. The gray plate image

B. Image Enhancement

Image enhancement aims to highlight the target object, weaken or eliminate some of the interference information, enhance the contrast between target and background, so that subsequent operations, improve the accuracy and the success rate of subsequent operations.

As the distribution of image's gray is often uneven, especially when the gray values are concentrated in a certain region, the contrast of the region is relatively small, so that the details of the region is not prominent, thus the distribution of the image's gray needs to be expanded scope to highlight the contrast in the region. In this paper, the method called gray stretch[2-3] is considered to process image. Piecewise linear transformation is used to work on gray values. Its form is shown as follow:

$$f(x) = \begin{cases} \dfrac{y_1}{x_1} x (x < x_1) \\ \dfrac{y_2 - y_1}{x_2 - x_1}(x - x_1) + y_1 (x_1 \le x \le x_2) \\ \dfrac{255 - y_2}{255 - x_2}(x - x_2) + y_2 (x_2 \le x) \end{cases} \tag{4}$$

The results are shown as follow:

Figure 3.The gray plate image

Figure 4.The plate image by gray stretch

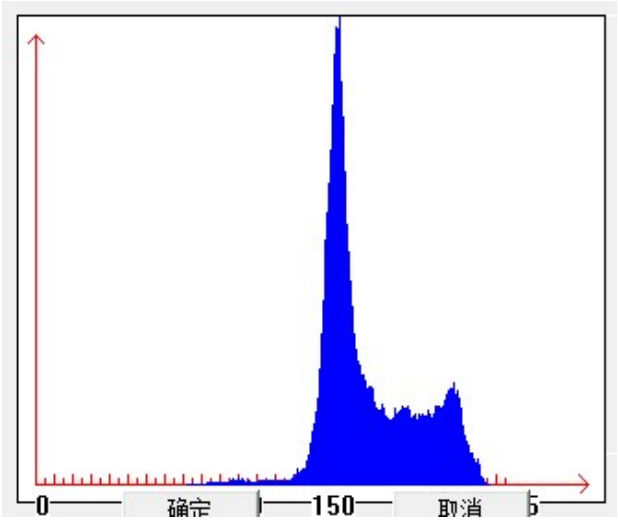

Figure 5. Gray histogram before gray stretch

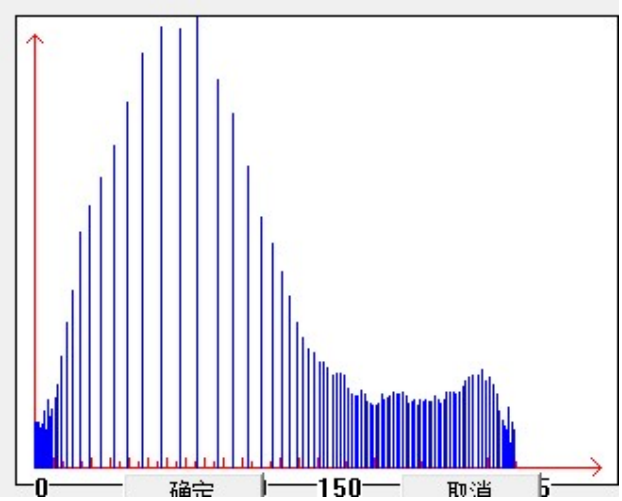

Figure 6. Gray histogram after gray stretch

C. Image Binarization

What is image binarization? Gray image will be transformed into another form which every pixel only contains two values. Generally the value of every pixel is set to 0(black) or 255(white). Its essence is to segment image by giving the corresponding threshold, in other word, according to the given threshold distinguish object to be extracted with the background in the picture. So the appropriate threshold is selected crucially.

In this paper, adaptive threshold binarization methods[4], adaptive mean binarization methods, and adaptive Gaussian binarization methods are used.

The result is shown as follow:

Figure 7.The image after binarization

D. Remove Discrete Noise

Sometimes, images also contain some discrete noises after binarization. In this paper, regional connectivity method[5-6] is used to remove these noises, they are consecutive points which the white pixel is less than or greater than a threshold.

The result is shown as follow:

Figure 8. The image after removing discrete noise

E. Slant Correction

When the plate is inclined, maybe the incomplete characters or slanted characters are segmented if try to segment character without any processing, which will bring to great difficulty on subsequent recognition, and thus the plate should make slant detection and correction. Firstly, the plate is rotated by an angle $\Delta\theta$. Secondly, make horizontal projection for each image rotated and record the height of horizontal projection, if total changed angle is $\theta(10 <= \theta <= 10)$, when the height of the horizontal projection is minimum, take the θ as tilt angle. Finally, make use of the tilt angle to rotate the plate image. The resulting image is the plate image which is revised.

The result is shown as follow:

Figure 9. The image after slant correction

IV. The Process of the LPR System

A. Grabbing Image

Generally, images are took by a camera, because of weather or light and so on, the image which are collected maybe take some discrete noises.

B. Image Processing

Make use of kinds of image processing technology to improve quality of images.

C. License Plate Location

The purpose of license plate location is to determine the position of the plate from a car image, thereby facilitating the subsequent character segmentation and recognition.

D. Character Segmentation

Character segmentation requires the ability to accurately locate character boundary, then all the characters in the license are extracted.

E. Character Recognition

Character recognition is the most important step in the whole process, and it requires the ability to accurately recognize the plate numbers. BP neural network is used to recognize characters.

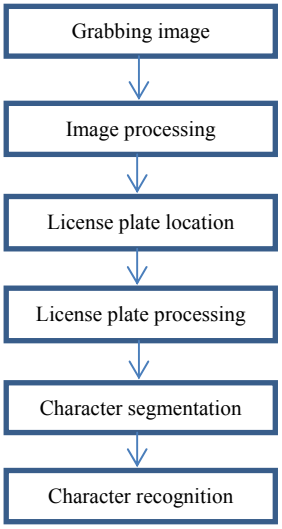

Figure 10. The process of the LPR

V. BP Neural Network

At present, Applications of BP neural network have been successfully carried out in the fields of various fields due to the fact that it has such advantages as self-study, self-organization, fault tolerance, highly nonlinear, highly robust, associational memory function, high-speed parallel processing and distribution storage information, etc. It is also capable of achieving the goal of pattern information processing based on pattern recognition theory, which cannot be completed on the calculation theoretical level[7]. Thus, using neural network is a new way to open up the development of pattern recognition. Its network mode includes supervised learning network, unsupervised learning work, self-supervised learning work and hybrid learning network.

BP neural network is a multi-layered network made up of three or more layers with each layer being composed of many neurons. BP neural network is trained by supervised learning. After the learning mode is provided to the network, its activating value of neurons will propagate from the input layer through the intermediate layers to the output layer. Then the network response which corresponds to the input mode will be output in the output layer.

Finally, according to the principle of reducing the errors between expected output and the actual output to correct the each connection weights value from the output layer, through the intermediate layer, and finally return to the input layer. The correction process starts from the input layer and is carried out layer by layer, known as the error back propagation algorithm. With the ongoing of the error back propagation training, the correct rate of the network response to the input mode continues to rise.

The learning steps of the BP neural network can be illustrated as follows:

- Initialize network. Set the parameters of error "e", learning rate "η" and maximum learning number M;
- Provide training samples to network;
- Run forward propagation and calculate the output mode of each layer;
- Run reverse propagation and calculate the errors of each layer and correct connection weight values;
- Train the network with cycle memory;
- Distinguish learn results.

In this paper, the selected parameters when training the neural network were as follow:

- A three-layer neural network is selected. In principle, the three-layer neural is able to cope with all the problems.
- There are 512 neurons in the input layer. 512 features were extracted from character image.
- There are 6 neurons in the output layer. 6 binary can represent all license plate characters.
- There are 48 neurons in the hidden layer.

$$h = \sqrt{m * (n + 1) + 1} \quad (5)$$

The above is an empirical formula of number of neurons in the hidden layer, m is the number of neurons in the input layer, n is the number of neurons in the output layer.

- Learning rate η is 0.5.
- The maximum number of learning M is 1500.
- Activation function is selected as follow:

$$f(x) = \frac{1}{1 + e^{-x}} . \quad (6)$$

- The initial weight values of the layers are random number between (-1, 1).

The BP neural network algorithm flowchart is shown as Fig. 11.

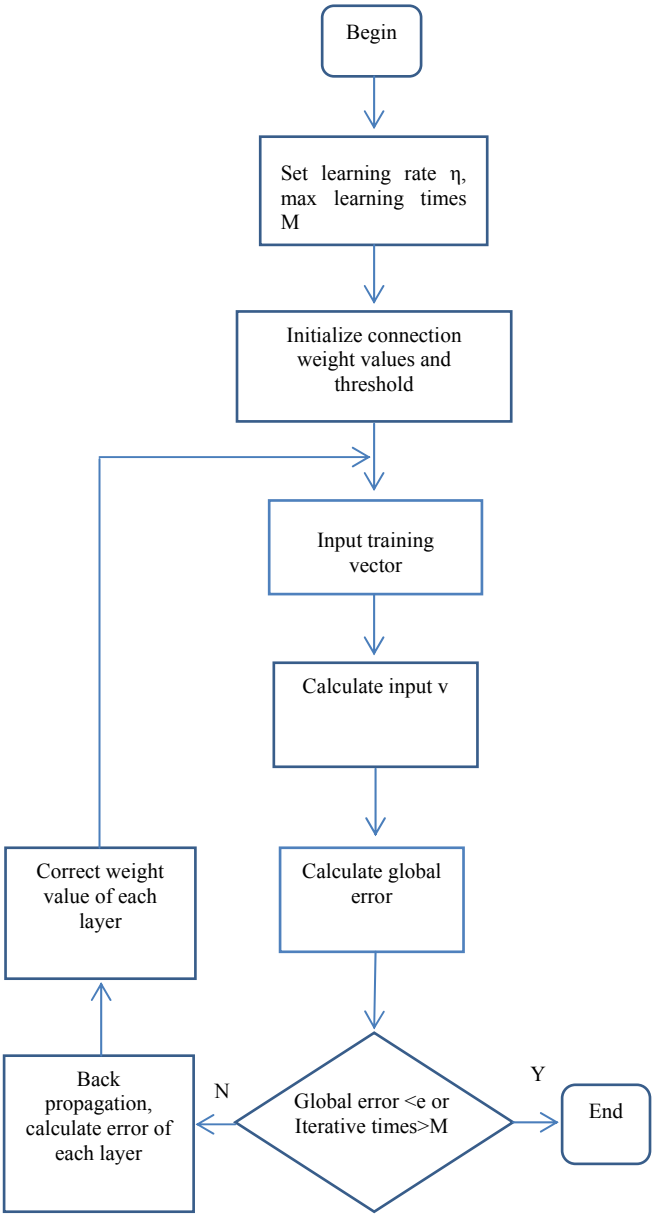

Figure 11. BP neural network algorithm flowchart

VI. Experimental Results and Analysis

The 8569 samples were selected in the trial, the character set included the figures "0, 1, 2, 3, 4, 5, 6, 7, 8,

9", the letters "A, B, C, D, E, F, G, H, J, K, L, M, N, P, Q, R, S, T, U, V, W, X, Y, Z", the Chinese characters "bei, yue, gan, hei, ji, jin, jing, liao, lu, meng, su, yu, gua". The sample size of each numeric character is shown in Table 1. The sample size of each letter character is shown in Table 2. The sample size of each Chinese character is shown in Table 3.

TABLE 1

SAMPLE SIZE OF THE NUMERIC CHARACTERS

Figure	0	1	2	3	4	5	6	7	8	9
Sample size	717	230	574	520	94	718	835	518	766	716

TABLE 2

SAMPLE SIZE OF THE LETTER CHARACTERS

Letter	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	T	U	V	W	X	Y	Z
Sample size	683	115	140	109	87	334	57	214	79	142	78	103	95	146	30	80	22	8	48	21	18	14	30	30

TABLE3

SAMPLE SIZE OF THE CHINESE CHARACTERS

Chinese character	bei	yue	gan	hei	ji	jin	jing	liao	lu	meng	su	yu	gua
Sample size	4	15	9	4	2	2	2	19	31	51	1	7	4

Generally, Image is classified through extraction of low level features such as grey scale, color, texture, shape, and location. This image classification method has a common drawback of requiring large amount of data and high complexity of computation, yet it yields more accurate classification. SVM is a new and very effective classification, which is able to avoid "curse of dimensionality" and over-fitting, and converged solution is a global solution, thus it is very suitable for applying to classify images[8-10]. SVM is used as classifier in this paper. The classification results are shown in Table 4.

TABLE 4

NUMBER OF CATEGORIES OF INDIVIDUAL CHARACTER

Character	0	5	D	B	M	jin	bei
Number of categories	138	122	28	33	13	4	2

As there are so many characters involved, not all of them can be listed. Only a few representative characters are selected to be classified. The number of samples in each category is not the same after classification. For example, the character "0" has two similar samples in one category at the least and 174 similar samples in another category at the most; the character "B" has two similar samples in one category at the least and 28 similar samples in another category at the most. Certainly, the number of categories is also related to the number samples of each character. Examples of similar samples in the same class are as follows.

Figure 12. Similar samples

Training sets are randomly selected from each character set before classification and proportionally selected from each category of the character set after classification.

Because the character samples in each category are similar, these samples possess the ability to cover every style of the character. Thus the training sets selected are representative. BP neural network is used to train different training sets. The results of which are shown in Table 5.

Table 5 shows the method of selecting the training set can indeed improve the recognition rate of the character, with over 28.5% increase as compared to that of the pre-classification. However, this improvement is not enough for practical application. New improvement in extracting feature algorithm and BP neural network algorithm is needed to further increase recognition rate.

VII. CONCLUSION

In this paper, the training set selection algorithm had made improvement. Experiments show that this method has a good ability to cover the sample set and avoid over-fitting. It demonstrates high recognition rate on the training set as well as cheerful recognition performance on the test set.

TABLE 5
RECOGNITION RATE BEFORE AND AFTER CLASSIFICATION

	Sample set	Training set	Test set	Number of correct	Recognition rate
Before classification	8569	5990	2579	1635	63.4%
After classification	8569	4339	4230	3890	91.9%

Training set selection algorithm has been a basic point and a research focus in the license plate recognition system, so it is worth spend time and effort to study it. License Plate Recognition is an integrated system, we must build a foundation which the training set to be selected for making future research easier and more meaningful.

ACKNOWLEDGMENT

We would like to thank the National Key Science &Technology Pillar Program of China (The key technology research and system of stage design and dress rehearsal, 2012BAH38F05; Study on the key technology research of the aggregation, marketing, production and broadcasting of online music resources, 2013BAH66F02) and the communication university of China engineering program(3132013XNG138).

REFERENCES

[1] M.Faisal Zaman, and Hideo Hirose,. "Classification Performance of Bagging and Boosting Type Ensemble Methods with Small Training Sets." New Generation Computing, vol.29, 2011, pp. 277-292.
[2]. Shuai Fang,Rong Deng,Yang Cao,Chunlong Fang "Effective Single Underwater Image Enhancement by Fusion". Journal of Computers.vol.8(4), 2013, pp. 904-911.
[3] Hao Xu, Tiesheng Fan. "A Segment Linear Function—Based on Image Enhancement Method".Journal of Liaoning University(Natural Science Edition), vol.33, 2006, pp. 362-364.
[4] N. Otsu. "A Threshold Selection Method from Gray-Level Histograms". IEEE Transactions on Systems, Man, and Cybernetics., vol.9(1), 1979, pp.62-66.
[5] M. Dillencourt, H. Samet, M. Tamminen. "A general approach to connected component labeling for arbitrary image representations". Journal of the ACM,39(2), pp253-280, April 1992.
[6] R. Gonzales, R. Woods. Digital Image Processing. pp.42-45, Addison-Wesley, 1992.
[7] Wang Xuewu and Tan Dejian, "Application and Development Trend of Neural Network", Computer Engineering and Application. 2003, pp. 98-100.
[8] Wang Chenfei and Xiao Shibin, "Research the Image Classification Based on the SVM", Computer & Digital Engineering. vol.34, 2006, pp. 74-76.
[9] Lei Ding, Fei Yu, Sheng Peng, Chen Xu. "A Classification Algorithm for Network Traffic based on Improved Support Vector Machine", Journal of Computers. vol.8(4), 2013, pp. 1090-1096.
[10] Deyuan Zhang, Bingquan Liu, Chengjie Sun Xiaolong Wang, "Learning the Classifier Combination for Image Classification", Journal of Computers, vol.6(8), 2011, pp. 1756-1763.

Xiaowen Liang bachelor degree achieved in 2009, master degree will achieve in 2014, current research interest is data mining.

Wei Gong associate professor, participated in a number of Chinese national natural science fund project, published many papers.

Wenlong Fu lecturer, mainly research on communication information security, large-data processing

Jing Qi computer engineer.