2009 10th International Conference on Document Analysis and Recognition

Separate Chinese Character and English Character by Cascade Classifier and Feature Selection

Yuanping Zhu[1], Jun Sun[1], Akihiro Minagawa[2], Yoshinobu Hotta[2], Satoshi Naoi[1]
1 Fujitsu R&D Center Co., Ltd, Beijing, China
2 Fujitsu Laboratories Ltd, Kawasaki, Japan
{yuanping.zhu,sunjun}@cn.fujitsu.com; {minagawa.a,y.hotta,naoi.satoshi}@jp.fujitsu.com

Abstract

The separation of Chinese character and English character is helpful for OCR technique. In this paper, a multi-level cascade classifier combined with feature selection is constructed to identify Chinese character and English character based on individual character. Most of samples are identified by the first node classifier, the remained low classification confidence samples are fed to the next node classifiers to get the final result. For the motivation of utilizing feature complementarity, each node classifier is trained on low classification confidence samples of its previous node classifier with independent feature selection. Furthermore, a confidence bias is utilized to improve the classifier generalization. The experiment results validate the effectiveness of this classifier.

1. Introduction

Chinese characters and English characters have distinctive properties. It makes trouble for OCR on Chinese-English mixed documents. A common strategy for both two types of characters is not suitable. Identifying Chinese type and English type character and utilizing different strategies or methods to two types is helpful to improve OCR performance.

This task is to divide the whole character space into two sub-spaces: Chinese character and English character(including digits). Treating it as a two-class classification problem, the task is to find effective features and construct a powerful classifier. But this is a special two-class problem. First, the diversity in each class is great. Each class is a character set. The diversity among the samples of each class is not only from different characters but also from different character types such as font, size and so on. Second, a serious data overlapping problem exists between two character sets. Some simple Chinese characters are similar with English characters. Therefore, separating them is not easy. Both the effective features and powerful classifier are important for this task.

Generally, this technique belongs to language identification. Most of language identification researches are based on document[1-2], text line[3-6] or word[7-8] level features. They have the assumption of uniformity in the text entities. Language identification based on individual character is seldom discussed. Zheng[9] used fisher classifier and SVM classifier to identify Chinese and English characters. Liu[10] used a fuzzy construction process(FCP) method[11] to classify characters into Chinese, English and Japanese. FCP is a fuzzy c-means(FCM) clustering based prototype learning method. Actually, it is a multi-template based classification method. Many prototypes are produced. Sanguansat[12] combined off-line and on-line features to distinguish on-line Thai and English characters. Spatial information features, such as looped part, cavities, are off-line features. Temporal information features, such as start-to-end point distance feature, are taken as on-line features.

In this paper, a cascade classifier with feature selection is constructed to identify an individual character as Chinese or English character. Its advantage is approximate accuracy to strong classifier by combining simple classifiers. The first classifier identifies most of characters. If the classification confidence of the first classifier is low, the target character will be fed to the next node classifiers one by one until it gets the final result. For the motivation of utilizing feature complementarity, each node classifier is trained on low classification confidence samples of its previous node classifier with independent feature selection, which makes them not share same feature space. And, a confidence bias is applied to improve classifier generalization. The experiment results validate the effectiveness of these measures in this task.

The organization of this paper is as follow: section 2 describes the motivation of the proposed method.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.164

Section 3 discusses the feature extraction. Cascade classifier construction is described in Section 4. Section 5 gives the experiment results and analysis. Then, it shows conclusions in section 6.

2. The proposed method

In this research, English character set including digits has 62 characters (26 capital letters, 26 low case letters and 10 digits). Chinese character set includes all 3755 characters of GBK Level 1. Most of Chinese characters are great different from English characters, like first 10 Chinese characters in GBK level 1 shown in Fig.1(a). But some Chinese characters have simple structures similar to English characters. The second row of Fig. 1 shows some of them.

Those similar characters make two classes be overlapped in feature space. Essentially, this is a non-linearly separable problem. Cascade classifier has shown effectiveness and efficiency in many tasks. Cascade classifier can deal with non-linearly separable problem while maintaining the simplicity of linear classifier[13]. Because different characters are sensitive to different features, the complementarity exists among the features, and the whole feature set does not always bring gain in classification. To utilize the feature complementarity, a rejection-based cascade classifier with feature selection is adopted. Let node classifiers in cascade classifier select different features. Put the sample into the header node classifier. If cannot get high classification confidence, it is regarded to be insensitive to this node classifier's features. Then, it is rejected and transferred into the next node classifier which is trained on the low confidence samples of its previous classifier. By feature selection and training sample selection, the next node classifier is expected to have stronger discrimination power on these samples than its previous classifier. Going through the cascade classifier, the sample will be identified by one node classifier or rejected until the last classifier identifies it. Fig. 2 shows the workflow of this cascade classifier.

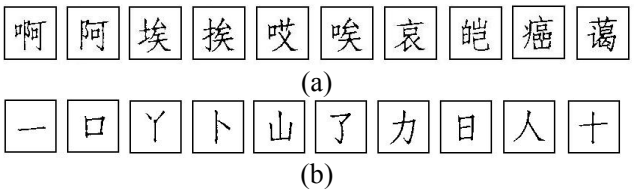


Fig. 1. Chinese character samples. (a) First 10 Chinese characters in GBK level 1; (b) 10 simple Chinese characters.

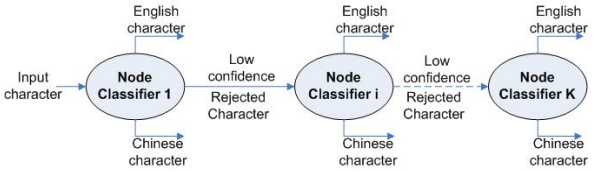


Fig. 2. Workflow of the cascade classifier

3. Feature Extraction

The first step for classifier construction is to extract effective features. By observing the difference between Chinese and English characters, several distinctive properties are useful:

1) Chinese characters have similar width and height while many English characters' width is smaller than height.

That means the Chinese characters' aspect ratio approximate to 1. Although many capital letters have similar width and height, many lower case letters and numbers' aspect ratios are smaller than 1.

2) Chinese characters are more complex in topology.

Almost all English characters have only one connected component while most of Chinese characters have more. At the same time, Chinese characters have more strokes and much more complex structure. Fox example, usually, Chinese characters are composed of several parts and most kinds of parts have more than one stroke. Furthermore, there are many kinds of spatial relationship among strokes and parts. All of above make Chinese character a more complex pattern than English character. This is the most distinctive difference between Chinese characters and English characters.

3) Chinese characters have much more horizontal and vertical strokes than cursive strokes [9].

Almost all lower case letters and numbers are composed of cursive strokes. In contrary, about 75% of the strokes of Chinese characters are horizontal and vertical. So this is also important information to discriminate Chinese and English characters.

According to these knowledge, aspect ratio, connected component number, stroke contour pixel density, stroke density histogram [9], normalized run length histogram, stroke density derivative accumulation and projection profile derivative accumulation are selected as features. Because the font, size and stroke width of samples vary much, normalization is considered during feature extraction.

1) Stroke contour pixel density

Because English characters have fewer strokes, as a whole, the density of stroke pixel is lower. To remove the effect of font and stroke width, the stroke contour pixel, as shown in Fig.3(a), is used instead of the

stroke pixel. And, the contour pixels of one-pixel-width stroke should be counted by twice.

2) Projection profile derivative accumulation

Because of the complex structure in Chinese character, the diversity of projection is larger, especially in the local regions. As a whole, the accumulation of this diversity is not small for Chinese characters. For a character image $I(M,N)$, let $P_h(j)/M$ and $P_v(i)/N$ denote normalized horizontal and vertical projection profile. The projection profile derivatives indicate the projection change in local regions or the complexity in structure. Then, the horizontal and vertical projection profile derivative accumulation C_{ph}, C_{pv} are defined as:

$$C_{ph} = \frac{1}{M}\sum_{j=1}^{N-1} |P_h(j+1) - P_h(j)| \quad (1)$$

$$C_{pv} = \frac{1}{N}\sum_{i=1}^{M-1} |P_v(i+1) - P_v(i)| \quad (2)$$

3) Stroke density histogram

Chinese characters have more strokes. In other word, the stroke density is higher. Fig. 3(b) shows the stroke density of a Chinese character. The scan line penetrates the character five times, so the vertical stroke density of this line is five. By this way, we can get the histogram of the horizontal and vertical stroke density respectively. The histogram is divided into several equal bins as the feature, such as five bins. This feature is not stable for font size. To remove this negative effect, horizontal and vertical stroke density histogram should be normalized by character width and height respectively.

(a) (b)

Fig. 3. Stroke property. (a)stroke contour; (b) stroke density

4) Stroke density derivative accumulation

The stroke density diversity can also refer the change in local regions or the complexity in structure.

Let $P_{sh}(j)$ and $P_{sv}(i)$ denote the horizontal and vertical stroke density projection profile. Similar to projection profile derivative accumulation, the horizontal and vertical derivative accumulation C_{sh}, C_{sv} are defined as:

$$C_{sh} = \sum_{j=1}^{N-1} |P_{sh}(j+1) - P_{sh}(j)| \quad (3)$$

$$C_{sv} = \sum_{i=1}^{M-1} |P_{sv}(i+1) - P_{sv}(i)| \quad (4)$$

5) Normalized run length histogram

Horizontal or vertical strokes have more long runs than cursive strokes. Chinese characters always have more horizontal and vertical strokes. So their run length histograms are different. Divide the histogram into five bins equally and take the number of run lengths in each bin as the feature.

Considering the character size diversity, horizontal and vertical histogram should be normalized by character height and character width respectively. And to remove the stroke width's influence, the stroke width normalization is applied in run length histogram. Zheng[9] used a morphological method to normalize stroke width on character image. This paper normalize feature vector directly. Firstly, horizontal and vertical stroke widths are calculated by average run length. Then, stroke width normalization is performed on horizontal strokes and vertical independently. The horizontal stroke pixels and vertical stroke pixels are discriminated by their run length on two directions. Fig. 4 shows an example that the horizontal and vertical stroke pixels are divided. Finally, all runs are normalized to a fixed width, such as 3, according to its stroke types.

Fig. 4. The horizontal and vertical stroke pixels.

4. Cascade classifier construction

In this section, the cascade classifier is constructed. Feature selection is combined in this cascade classifier. As mentioned before, feature selection can find suitable feature sub-set for each stage in cascade classifier. Firstly, the first classifier is training on whole sample set with feature selection. The samples with confidence lower than a confidence threshold are put into training set of the next classifier. The feature selection is also applied on the second classifier. Repeat this procedure until no accuracy gain from new added node classifiers, and then cascade classifier is constructed.

The classification confidence can be computed based on the sample's distance or membership grade to two classes according to the specific classifier. As to the prototype classifier, let d1 and d2 be a sample's distances to two class centriods, then the confidence is defined as:

$$Conf = 100^* |d1 - d2| / (d1 + d2); \quad (5)$$

Where, the range of $Conf$ is [0,100].

Confidence bias is utilized to improve generalization of node classifier. Because the samples around the confidence threshold maybe be randomly put into high confidence sample set or low confidence sample set, an overlap area between high and low confidence sample sets is help to enhance the generalization. A confidence bias is utilized to make this overlap area. By this way, the samples with confidence lower than threshold plus bias are put into training set of next node classifier in training stage. While in classification stage, only sample with confidence lower than threshold are rejected and transferred to the next node classifier. In short, more samples are involved in training. The bias b is set to 5 in this paper.

Let $Cc(k)$ denotes a k-level cascade classifier including node classifiers $Cn(i), i = 1,...,k$, its training accuracy is $A(Cc(k))$. As to i-th node classifier $Cn(i)$, let $Th(i)$ be its confidence threshold and $S(i)$ be its training set. $Conf(i, j)$ is the confidence of sample j in i-th node classifier. Then, the cascade classifier training algorithm is as follows:

Step 1. Train first node classifier $Cn(i), i = 1$ on whole training set.

Step 2. Construct training set for next node classifier $Cn(i), i = i + 1$,

If $Conf(i-1, j) < Th(i-1) + b$, move sample j from $S(i-1)$ to $S(i)$

Step 3. Train node classifier $Cn(i)$ on $S(i)$.

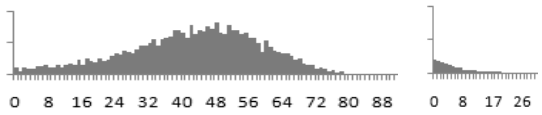
Step 4. Construct and evaluate cascade classifier $Cc(i)$

If accuracy decreases, $A(Cc(i)) < A(Cc(i-1))$, stop the training and final cascade classifier is $Cc(i-1)$.

Else return to Step 2, continue the training procedure.

In node classifier training, a uniform threshold for all node classifiers is not appropriate. Threshold search is performed to maximize $A(Cc(i))$. From the confidence histogram in Fig. 5, {5,10,15,20,25} is selected as search space for the simplicity.

In practice, the major contribution of accuracy increasement is from first several node classifiers. We can set a maximal level number, 5.



(a) (b)

Fig. 5. Confidence histogram of first node classifier. (a) Correct sample confidence histogram; (b) wrong recognition sample confidence histogram;

5. Experiments

To evaluate the classifier performance and validate the effectiveness of the proposed method, several experiments are performed. Both English character set and Chinese character set have 20k samples in this experiment. All these samples are split into two parts equally, one for training, one for test.

Firstly, a comparison among confidence thresholds and biases in cascade classifiers is shown in Table 1 and Fig. 6. FSS(Forward Sequential Selection) is selected as the feature selection method in following experiments. In Table 1 and 2, ER, CR and AR denote English, Chinese and average recognition correct rate (%) respectively. Trn and Tst are training and test stage. L is the classifier level number with the best accuracy. In column Th, U denotes unfixed confidence threshold. The wave-like accuracy curves of fixed threshold classifiers mean they are easy to get into local optimum in classifier construction. Unfixed confidence threshold can avoid this problem. About the effect of confidence bias, generally, it is unable to increase training accuracy comparing without bias. But in the most of times, classifiers with bias show better in test set. That means the utilization of bias is help to improve the generalization of the classifier. It is also be found that most of discrimination power is released by first 4 node classifiers, especially first 2 node classifiers. Far from obtaining accuracy gain, more node classifiers will decrease the performance.

Table 2 gives the comparison among different classification methods. The results of single prototype classifier, Bayesian classifier, SVM classifier, cascade classifier(2-level) without feature selection, single classifier with feature selection, and optimal cascade classifier are given. The last classifier in Table 1 is taken as optimal cascade classifier. The LibSVM[14] tool is used in SVM method and RBF kernel and its default settings are chosen. As the similar results in [9,10], SVM has the best correct rate. However, the gap of the proposed classifier is very small, especially on test set. The advantage of the proposed method is its approximate accuracy to strong classifier by simple base method as prototype classifier. An interesting phenomenon is that the cascade classifier without feature selection achieves a little worse accuracy even than single prototype classifier. The possible reason is that after first classifier, the remained samples are similar in original feature space so that it is hard to separate them by original features. That also proves the significance of feature complementarity and feature selection in cascade classifier.

The above experiments are run on the computer with

Table 1. Result comparison on confidence threshold and bias

Th	b	L	ER		CR		AR	
			Trn	Tst	Trn	Tst	Trn	Tst
10	0	2	98.37	98.94	99.5	99.21	98.93	99.08
15	0	4	99.14	99.34	99.17	98.96	99.16	99.15
20	0	4	98.63	98.98	99.36	98.79	98.99	98.88
U	0	5	99.33	99.37	99.25	98.99	99.29	99.18
10	5	4	98.99	99.42	99.29	98.93	99.14	99.18
15	5	4	98.68	99.07	99.41	98.98	99.05	99.03
20	5	2	99.05	99.27	99.23	98.85	99.14	99.06
U	5	4	99.25	99.42	99.29	99.07	99.27	99.25

Table 2. Classifier accuracy comparison

Classifier	ER		CR		AR	
	Trn	Tst	Trn	Tst	Trn	Tst
Prototype classifier	99.03	99.37	92.87	91.6	95.95	95.49
Bayesian classifier	99.19	99.76	95.36	94.27	97.27	97.01
SVM	99.72	99.66	99.58	98.92	99.65	99.29
Prototype classifier with FSS	98.49	99.3	98.1	96.48	98.3	97.89
Cascade classifier Without FSS (L=2)	95.28	95.83	95.6	94.6	95.44	95.22
Optimal cascade classifier	99.25	99.42	99.29	99.07	99.27	99.25

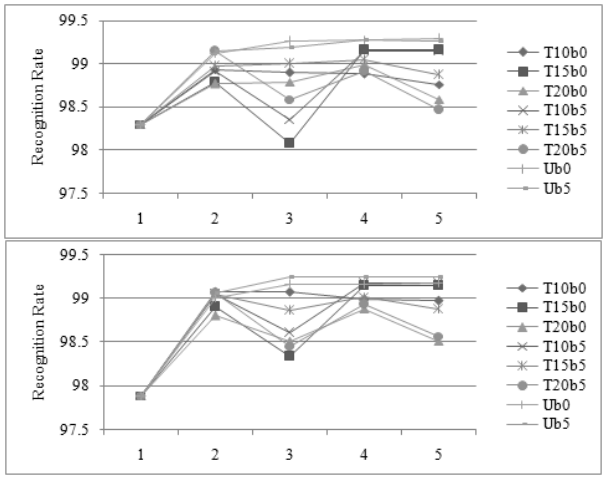


Fig. 6. Multi-level cascade classifier accuracy curves on training and test set.

3.0G CPU and 4G memory. The classification speed is about 2.2ms per character including feature extraction.

6. Conclusions

In this paper, a cascade classifier with feature selection is constructed for Chinese and English character separation based on individual character. The experiment proves that the cascade classifier utilizing the complementarity among features is effective in solving the character set separation problem. Additional confidence bias in node classifier training sample selection is help to enhance the generalization of cascade classifier. As a whole, the advantage of the proposed method is its approximate accuracy to strong classifier by simple base method as prototype classifier. Although this paper only discussed the separation of Chinese and English character, it is easy to extend this classifier to other language identification problems by replacing the features.

7. References

[1] G. Zhu, X. Yu, Y. Li, David Doermann. Unconstrained language identification using a shape codebook. In Proc. of 11th Int'l Conf. on Frontiers in Handwriting Recognition, pp.13-18, 2008

[2] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns. Automatic script identification from document images Using Cluster-based templates. IEEE Trans. Pattern Analysis and Machine Intelligence, 19(2):176–181, 1997.

[3] A. Spitz, Determination of the script and language content of document images. IEEE Trans. Pattern Analysis and Machine Intelligence, 19(3): 235-245, 1997

[4] J. Ding, L. Lam, and C. Suen. Classification of oriental and european scripts by using characteristic features. In Proc. of 4th Int'l Conf. on Document Analysis and Recognition, Vol.2, pp.1023–1027, 1997.

[5] S. Lu and C. Tan. Script and language identification in noisy and degraded document images. IEEE Trans. Pattern Analysis and Machine Intelligence, 30(2):14–24, 2008.

[6] A. Elgammal, M. Ismail, Techniques for language identification for hybrid Arabic-English document images, In Proc. of 6th Int'l Conf. on Document Analysis and Recognition, pp.1100-1104, 2001

[7] H. Ma and D. Doermann. Word level script identification on scanned document images. In Proc. of Document Recognition and Retrieval(SPIE), pp.124–135, 2004.

[8] S. Jaeger, H. Ma, and D. Doermann. Identifying script on wordlevel with informational confidence. In Proc. of 8th Int'l Conf. on Document Analysis and Recognition, vol.1, pp. 416–420, 2005.

[9] Y. Zheng, C. Liu, X. Ding. Single character type identification. In Proc. of the SPIE Document Recognition and Retrieval IX. Vol.4670, pp.49-56, 2002.

[10] Y. Liu, C. Lin, F. Chang. Language identification of character images using machine learning techniques. In Proc. of 8th Int'l Conf. on Document Analysis and Recognition, Vol.2, pp.630-634, 2005.

[11] F. Chang, C. Chou, C. Lin, and C. Chen, A Prototype Classification Method and Its Application to Handwritten Character Recognition, In Proc. of IEEE Conf. on System, Man, and Cybernetics, pp. 4738-4743, 2004.

[12] P. Sanguansat, P. Yanwit, et al, Language-based hand-printed character Recognition: a novel method using spatial and temporal informative features, In Proc. of IEEE 13th Workshop on Neural Networks for Signal Processing, pp. 527-536, 2003.

[13] M. Elad, Y. Hel-Or, and R. Keshet, Pattern Detection Using a Maximal Rejection Classifier. Pattern Recognition Letters, Vol.23, pp.1459–1471, 2002.

[14] C. Chang and C. Lin. LIBSVM: A Library for Support Vector Machines, 2001. [online] Available: http://www/csie.ntu.edu.tw/~cjlin/livsvm.