A Robust Approach for Recognition of Text Embedded in Natural Scenes

Jing Zhang[1], Xilin Chen[2], Andreas Hanneman[2], Jie Yang[2], Alex Waibel[2]
[1]*Mobile Technologies, LLC*
[2]*Interactive Systems Labs, School of Computer Science, Carnegie Mellon University*
jingzhang@computer.org, {xlchen, Hanneman, yang+, ahw}@cs.cmu.edu

Abstract

In this paper, we propose a robust approach for recognition of text embedded in natural scenes. Instead of using binary information as most other OCR systems do, we extract features from intensity of an image directly. We utilize a local intensity normalization method to effectively handle lighting variations. We then employ Gabor transform to obtain local features, and use LDA for selection and classification of features. The proposed method has been applied to a Chinese sign recognition task. The system can recognize a vocabulary of 3755 Level 1 Chinese characters in the Chinese national standard character set GB2312-80 with various print fonts. We tested the system on 1630 test characters in sign images captured from the natural scenes, and the recognition accuracy is 92.46%. We have already integrated the system into our automatic Chinese sign translation system.

1. Introduction

We encounter large amounts of information embedded in natural scenes in our daily lives. Signs are good examples of objects in natural environments that have high information content. A sign is an object that suggests the presence of a fact. It can be a displayed structure bearing letters or symbols, used to identify or advertise a place of business. Signs are everywhere in our lives. They make our lives easier when we are familiar with them, but they may pose problems or even danger when we are not. For example, a foreign tourist might not be able to understand a sign that specifies warnings or hazards. Automatic sign translation, in conjunction with spoken language translation, can help international tourists to overcome these barriers. A successful automatic sign translation system relies on three key technologies: sign detection, OCR (Optical Character Recognition) and machine translation.

OCR is one of the most successful areas in the pattern recognition field. For clearly segmented printed materials, state-of-the-art techniques offer virtually error-free OCR for several important alphabetic systems and their variants. However, error rates of OCR systems are still far from that of human readers in many applications such as video OCR and license plate OCR. The gap between the two is exacerbated when the quality of the image is compromised, e.g., input using a video camera. Video OCR, which is to recognize text from a video stream, was motivated by digital library application and visual information retrieval tasks. Many video images contain text contents. These texts can be part of the scene, or may come from computer-generated text, which is overlaid on the image (e.g., captions in broadcast news programs). The text, especially the subtitle in the video, provides useful information for video indexing. In a video OCR task, text in the foreground is usually uniformly distributed and its resolution can be enhanced by inter-frame information [7, 8, 11]. Compared with video OCR tasks, recognition of text embedded in natural scenes faces more challenges: text can vary in font, size, orientation, and position of text, be blurred by motion, be under shadow, be occluded by other objects, or be distorted by slant and tilt. The image from a camera under unconstrained environment can be very noisy.

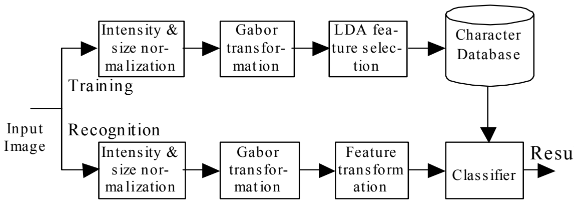


Figure 1. The flow chart of the system

In this paper, we propose a robust approach for recognition of text embedded in natural scenes. Figure 1 illustrates the procedure of the proposed approach. Text embedded in the natural scenes is captured by a camera. The method of character segmentation has been introduced before [2], and we will only focus on the recognition of the character in this paper. Instead of using only binary information as most other OCR systems, we extract features for recognition from intensity of an image directly [10]. The motivation for this is to avoid potentially information loss during the binarization process which is irreversible. We utilize a local intensity normalization method to effectively handle luminance variations of the captured character image. Gabor transforms are applied to obtain local features. We use LDA (Linear Discriminant Analysis) for selection and classification of features.

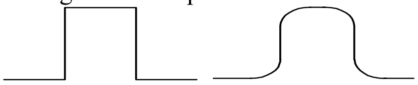
We have applied the proposed method to form a Chinese sign recognition system. The system can recognize a vocabulary of 3755 Level 1 Chinese characters with various print fonts in GB2312-80, which is the national Chinese character set standard. We tested the method on 1630 characters from sign images captured from natural scenes, and the recognition accuracy is 92.46%.

2. Features extraction

2.1. Image pre-processing

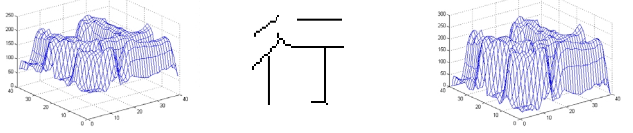
An imperfect text image is a major factor that reduces recognition accuracy when OCR is applied to a video image. Image preprocess, therefore, can enhance the recognition accuracy. We perform normalizations on both size and intensity of the text before feature extraction.

Lighting variations usually bring problems for most image-based recognition tasks. A common practice is to normalize the image's luminance using a histogram normalization method, which changes the luminance scope. The image intensity captured by a camera can be in a large dynamic scope, some with obvious contrast, some not, and some with high light while others may in dark. The goal of intensity normalization for character recognition is to make entire strobes within characters have the nearly same intensity distribution. A global normalization method cannot do it, so we propose to perform local normalization using the profile of a stroke as a priori. The ideal profile of a stroke is shown in figure 2(a), but it requires an unlimited bandwidth system to obtain such a response. We can only have a limited bandwidth system in practice and obtain a profile as shown in figure 2(b). We can normalize the intensity alone the orthogonal direction of a stroke using the stroke profile.



(a) (b)

Figure 2. Ideal (a) and practical (b) profile of a stroke



(a) Original intensity (b) Skeleton (c) Normalized intensity

Figure 3. An example of intensity normalization

Figure 3 shows an example of local intensity normalization for Chinese character "行". Figure 3(a) is the original intensity of the character, (b) is the character's skeleton, and (c) is the normalized intensity.

2.2. Gabor feature extraction

Gabor wavelet is a sinusoidal plane wave with particular frequency and orientation, modulated by a Gaussian envelope. It can characterize spatial frequency structure in the image while preserving information of spatial relations, and thus suitable for extracting orientation-dependent frequency contents of patterns. Gabor wavelet has been widely used in many applications, such as data compression [12], face recognition [13], and texture analysis [9], because of its good mathematic properties. Early applications of Gabor wavelet to OCR tasks were based on binary features [1, 3-5]. Yoshimura and his colleagues ever reported to extract features using Gabor from intensity of video images for character recognition and use LVQ (Linear Vector Quantization) for feature selection [14].

A complex-valued 2D Gabor function modulated by a Gaussian envelope is as follows:

$$G(x,y,k,\theta)=G_1(x,y)[\cos(R)-\exp\left(-\frac{\sigma^2}{2}\right)] \quad , \quad (1)$$

$$+iG_1(x,y)\sin(R)$$

where $G_1(x,y)=\frac{k^2}{\sigma^2}\exp[-\frac{k^2(x^2+y^2)}{2\sigma^2}]$, $R=kx\cos\theta+ky\sin\theta$, $k=\frac{2\pi}{\tau}$. The parameter σ is the deviation of the Gaussian envelope, τ and θ are the wavelength and orientation of Gabor function respectively. Frequency responses of Gabor filters in four orientations (0°, 45°, 90° and 135°) are illustrated in figure 4.

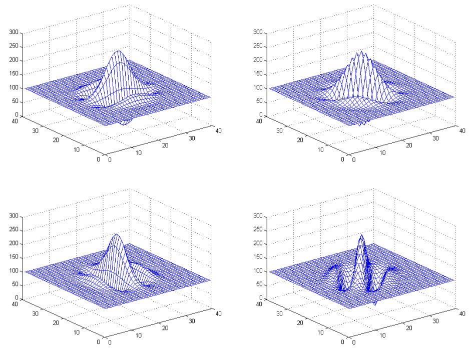


Figure 4. Four directions' Gabor filter

For a given pixel (x_1,y_1) with gray level $I(x_1,y_1)$ in an image, its Gabor feature can be regarded as the convolution:

$$J(x_1,y_1,k,\theta)=\int I(x_1-x,y_1-y)G(x,y,k,\theta)dxdy \ .(2)$$

Suppose that m frequencies and n orientations are used in extracting Gabor feature, we can have a vector of $m\cdot n$ complex coefficients for each position. We called this vector as a jet, which is used to represent the position's local features.

In our application, we divide a character into 7x7 grids as shown in figure 5, which results in $49 \cdot m \cdot n$ dimensions' feature vector for a character.

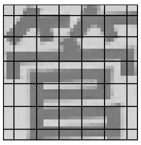


Figure 5. Grids for feature extraction

3. Features selection and classification

We need to reduce dimensions of feature vectors because they are computational expensive and not all of them are effective for recognition. LVQ and LDA are common tools for dimension reduction. We use LDA in this work because it can be used not only for dimension reduction, but also feature optimization. LDA is a method to find a transformation that can maximize the between-class scatter matrix S_b and minimize the within-class scatter matrix S_w simultaneously:

$$\arg_W \left\{ \max \frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}} \right\}, \tag{3}$$

The new space, **W**, is then the most discriminative space. Feature vector **x** in the original feature space is projected to this new space and have the new feature **y** using equation (4), which can be used for classification.

$$\mathbf{y} = \mathbf{W}^T \mathbf{x} \tag{4}$$

In the recognition process, the Gabor feature vector is extracted from the input character image and then is projected onto the LDA space. The classification is finally performed using KNN method.

4. Experimental Results

We have evaluated the proposed approach on a Chinese sign recognition task. The current training set includes all level 1 characters in Chinese national standard character set GB2312-80 (total 3755 different characters). We employed 6 kinds of different fonts SongTi, HeiTi, KaiTi, LiShu, YaoTi, and YouYuan, from the standard font lib to form the training set. Because the characters on signs were usually in bold style, bold characters were used and were printed on paper and were finally captured by a camera.

We also randomly selected 1630 character images from our sign library, which contains more than 8000 characters in more than 2000 sign images, to form the testing set. The different characters in the testing set roughly cover 1/5 of the level 1 Chinese characters, with variations in font, lighting condition, rotation, and even affine transform. Figure 6 is some examples from the testing set.

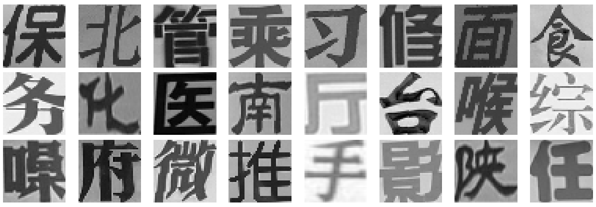


Figure 6. Some samples from testing set

Four orientations (0°, 45°, 90° and 135°) for gabor feature were used in our experiments because Chinese character's strokes were mainly drawn along the four directions. The key problem is to try to determine the best wavelength τ, which has the close relationship with the character's stroke width. Basically, the wavelength within the scope from half to twice of the average stroke width should produce good recognition results. Table 1 shows the character recognition rate for 6 different fonts of training set and recognition rate for character images in testing set when the average stroke width and one and a half of the width were assigned to the wavelength respectively. It can be seen that it has similar recognition rate using the two different wavelength's gabor features, not so good for testing set but quite well for training set especially for HeiTi, SongTi and YaoTi.

Table 1. Comparison of different wavelength for Gabor

	Accuracy	
	1.5 Average stroke width	Stroke width
Testing set	84.8816%	86.0717%
SongTi	99.6226%	99.4340%
HeiTi	99.8113%	99.8113%
KaiTi	83.3926%	76.9811%
LiShu	87.3558%	86.2264%
YaoTi	98.1132%	93.3962%
YouYuan	92.0755%	94.9057%

In a sign recognition task for natural scenes, font of character on different signs might be various, and the stroke width may also be diversified even for same font. So we combined the two wavelengths' gabor features to try to adapt to the various fonts and stroke width. The recognition rate is shown in the left column of the table 2.

Table 2. Comparison of different discrimination methods

	Accuracy	
	Vector angle	Euclidian Distance
Testing set	92.4606%	82.6057%
SongTi	99.7337%	99.7337%
HeiTi	99.8935%	99.7337%
KaiTi	93.3688%	85.0067%
LiShu	96.1651%	94.2477%
YaoTi	98.9880%	98.7750%
YouYuan	98.4021%	98.2157%

It has an evident improvement for the testing set compared with table 1 gotten by single wavelength's Gabor features. In the same time, the table 2 gives the comparison of 2 different discrimination methods, the vector angle in the left and Euclidian distance in the right. It shows that the vector angle produces the higher recognition accuracy than the other.

It is a quite satisfied result because the character images in the testing set are captured from natural scenes, with variations in font, lighting condition, rotation, and even affine transform. For further verifying the robustness of the proposed approach against noises, we would add the zero mean Gaussian noise to each character image in the testing set. For a given pixel whose intensity is $I(x,y)$ in the character image, we have

$$I'(x,y) = I(x,y) + n(x,y), \quad (5)$$

where $n(x,y) \sim N\left(0,(I(x,y)\cdot\gamma)^2\right)$.

Parameter γ represents the intensity of the noise. Figure 7 illustrates the impact on a Chinese character "随" from testing set when we added noise respectively $\gamma = 0.05$, $\gamma = 0.10$, $\gamma = 0.15$ and $\gamma = 0.20$.

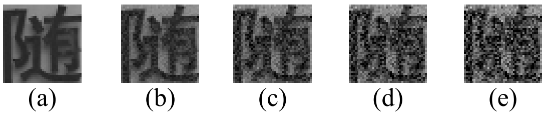


Figure 7. An example with different noises

(a) original image, (b) - (e) ones with Gaussian noise of 5%,10%,15%, and 20%

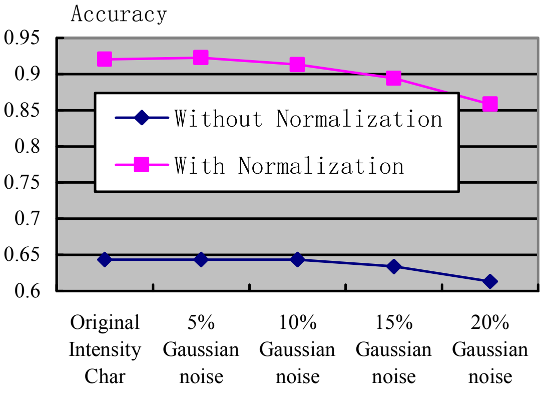


Figure 8. Improvement on accuracy from intensity normalization

The top curve of Figure 8 shows the recognition rate of 1630 characters in testing set when we added different intensity of Gaussian noise. It is only about 1% decrease of recognition rate when 10% noise is added, and about 6.5% decrease when 20% noise is added, compared with that of the original testing set. This Figure also illustrates the effective of local intensity normalization, which shows that the recognition accuracy is about 24% to 28% higher than that without normalization.

5. Conclusion

We have proposed a robust approach for recognition of text embedded in natural scenes. We propose to utilize local intensity normalization to effectively enhance the robustness. We use Gabor transform to extract features and LDA to select and reduce the dimensionality of the feature space. Experimental results have demonstrated its excellent recognition accuracy and robustness of the proposed method. We have already integrated the system into our automatic Chinese sign translation system.

References

[1] D. Deng, K. P. Chan, and Y. Yu, "Handwritten Chinese Character Recognition Using Spatial Gabor Filters and Self-Organizing Feature Maps", Proc. of ICIP, Vol. 3, pp. 940-944, 1994

[2] J. Gao, J. Yang, Y. Zhang, and A. Waibel, "Text Detection and Translation from Natural Scenes", Technical Report CMU-CS-01-139, Computer Science Department, Carnegie Mellon University, June, 2001.

[3] Y. Hamamoto, S. Uchimura, K. Masamizu, and S. Tomita, "Recognition of Handprinted Chinese Characters Using Gabor Features", Proc. of 3th ICDAR, Vol. 2, pp. 819-823, 1995

[4] Y. Hamamoto, S. Uchimura, M. Watanabe, T. Yasuda, and S. Tomita, "Recognition of Handwritten Numerals Using Gabor Features", Proc. of 13th ICPR, Vol. 3, pp. 250-253, 1996

[5] Q. Hou, Y. Ge, and Z. Feng, "High Performance Chinese OCR Based on Gabor Features, Discriminative Feature Extraction and Model Training", Proc. of ICASSP, Vol. 3, pp. 1517-1520, 2001.

[6] F. LeBourgeois, "Robust Multifont OCR System from Gray Level Image", Proc. of 4th ICDAR, Vol. 1, pp. 1-5, 1997

[7] H. Li, D. Doermann, and K. Omid, Automatic Text Detection and Tracking in Digital Video, IEEE Trans. on Image Processing, Vol. 9, No. 1, pp. 147-156, 2000

[8] R. Lienhart, Automatic Text Recognition for Video Indexing, Proceedings of ACM Multimedia 96, pp. 11-20, 1996.

[9] R. Mehrotra, K. R. Namuduri, N. Ranganathan, "Gabor filter-based edge detection", Pattern Recognition, Vol. 25, No. 12, pp. 1479-1494, 1992

[10] T. Pavlidis, "Recognition of Printed Text Under Realistic Conditions", Pattern Recognition Letters, Vol. 14, No. 4, pp. 317-326, 1993

[11] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, and S. Satoh, "Video OCR: indexing digital news libraries by recognition of superimposed captions", Multimedia Systems, Vol. 7, No. 5, pp. 385-95, 1999

[12] H. Szu, B. Telfer, and J. Garcia, "Wavelet transforms and neural networks for compression and recognition", Neural Networks, Vol.9, No. 4, pp. 695-708, 1996

[13] L. Wiskott, J. M. Fellous, N. Kruger, and C. Malsburg, "Face Recognition by Elastic Bunch Graph Match", IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 764-768, 1997

[14] H. Yoshimura, M. Etoh, K. Kondo, and N. Yokoya, "Gray-scale Character Recognition by Gabor Jets Projection", Proc. of 15th ICPR, Vol. 2, pp. 335-338, 2000.