A Novel Italic Detection and Rectification Method for Chinese Advertising Images

Jie Liu
High-Tech Innovation Center
Institute of Automation, Chinese Academy of Sciences
Beijing, P.R. China
jliu@hitic.ia.ac.cn

Heping Li, Shuwu Zhang, Wei Liang
High-Tech Innovation Center
Institute of Automation, Chinese Academy of Sciences
Beijing, P.R. China
{hpli, swzhang, wliang}@hitic.ia.ac.cn

***Abstract*—The italic detection and slant rectification is a key step of optical character recognition (OCR). In this paper, a novel method is proposed to detect and rectify italic characters in Chinese advertising images. Based on observations on structures of many characters, the centroid angle is proposed and a statistical study on it is presented. According to the statistical results, the centroid angle of a Chinese character approximately obeys a Gaussian distribution with its slant angle. Moreover, a Markov Random Field (MRF) model, considering the font-face similarity of neighboring characters and the strong correlation between the centroid angle and the slant angle of a character, is then presented to estimate the slant angle of a character. The italic characters can be detected and rectified by the estimated angle. The experimental results demonstrate the proposed method is effective and applicable.**

Keywords- italic detection; slant rectification; centroid angle; Markov Random Field

I. INTRODUCTION

There are many italic texts in Chinese advertising images. It is noted that the characters of italic style in a line may be overlapped in the vertical projection. The overlapping characters will make character segmentation difficult. A document with italic-face characters may decrease the recognition accuracy even in the document analysis systems commercially available. Therefore, the majority of recent optical character recognition (OCR) system contains a preprocessing stage dealing with italic characters rectification.

Figure 1. Examples extracted from Chinese advertising images.

Currently, there have been several studies concerned on italic character detection and rectification [1-7]. Fan *et al.* [1] extract structural information from virtual stroke embedded in the characters to classify them in three types. The normal and italic characters are can be distinguished depending of their type using either gradient information, curvatures of strokes, and then the exact shear angle of italic character is calculated to rectify the italic character. Nicchiotti *et al.* and Xia *et al.* [2-3] propose a method to discriminate italic and normal style by analyzing the weighted projection profile histogram of an appropriate character area. And then the slant angle is estimated by comparing the features of different sheared patterns [3]. Lee *et al.* [4] use the slant angle of vertical strokes to detect italic characters. Zhang *et al.* [5] resort to statistical analysis of stroke patterns on a wavelet decomposed word images to detect italic characters. Chaudhuri *et al.* [6] assume that there is always a black line going from the bottom of the character to the top of it, they search for the angle that this line makes with the base line to define the character as italic type. Ma *et al.* [7] propose a method that automatically selects features under the assumption that OCR results are available, and then a Gaussian mixture model is constructed to create clusters of characters that are classified between styles. Word styles are determined using a weighted majority vote.

So far, most contributions give more attention to English document. These methods do not perfectly deal with italic characters in Chinese advertising images. There are left-italic characters, mixed normal/italic characters, characters with varying size and characters arranged in irregular direction in Chinese advertising images. Some examples are shown in Fig. 1. These bring challenge to italic detection. The purpose of this paper is to find a fast and efficient approach of detecting and rectifying italic characters in Chinese advertising images.

In this paper, the centroid angle is proposed and a statistical study on it is presented. The statistical results reveal that the centroid angle of Chinese character obeys approximately a Gaussian distribution with real slant angle mean. According to this fact, the centroid angle of a Chinese character can be regarded as the latent slant angle of it plus additive Gaussian noise. Moreover, Markov Random Field (MRF), considering the font-face similarity of neighboring characters and the strong correlation between the centroid angle and the slant angle of a character, is presented to model the problem of estimating the slant angle. The exact slant

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.146

angle of a character is estimated by minimize the total graph energy. Therefore, italic characters can be detected and rectified by the estimated angle.

This paper is organized as follows. The centroid angle of character is proposed and the results of statistical study on it are demonstrated in section 2. Section 3 discusses the approach about estimation of slant angle of a character based on MRF. Experimental results are reported in section 4, and followed by some conclusions in section 5.

II. RESULTS OF STATISTICAL STUDY ON THE CENTROID ANGLE OF CHARACTER

Based on our observations, the structures of many Chinese characters are symmetric. According to this fact, the centroid angle is proposed. It reflects the real slant angle of character. Therefore, the real slant angle of character can be easily estimated by the centroid angle using statistical inference method.

A. The Centroid Angle

As shown in Fig. 2, the centroid angle θ is given by the angle included by the vertical line and the straight line between the centroids of upper-half and lower-half part of a character.

The centroid angle θ_v is defined as

$$\theta_v = \tan^{-1}\left(\frac{X_{centroid\_low} - X_{centroid\_up}}{Y_{centroid\_low} - Y_{centroid\_up}}\right). \quad (1)$$

where $(X_{centroid\_low}, Y_{centroid\_low})$ and $(X_{centroid\_up}, Y_{centroid\_up})$ denote the coordinates of the centroids of lower-half and upper-half parts of a character image respectively.

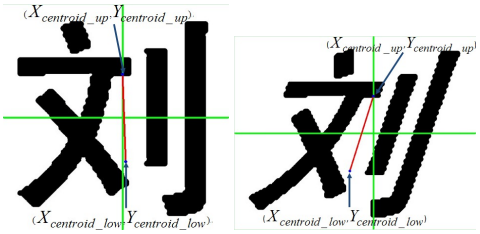

Figure 2. Example illustriting the centroid angles of normal and italic Chinese characters.

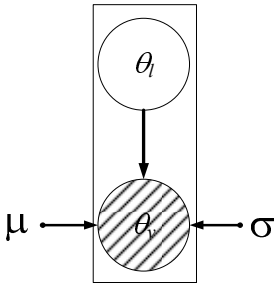

Figure 3. The generative model of the centroid angle.

Given an image $f(i,j)$ $1 \le i \le N$, $1 \le j \le M$, the centroid $(X_{centroid}, Y_{centroid})$ of $f(i,j)$ is calculated as

$$X_{centroid} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} i \times f(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M} f(i,j)}, Y_{centroid} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} j \times f(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M} f(i,j)} \quad (2)$$

B. The Statistical Results

To study on the centroid angle of Chinese character, we check almost regular and italic common Chinese characters. More details can be referred to in section 4. The statistical results are summarized as follows:

1) The centroid angle of regular/italic Chinese character approximately obeys a Gaussian distribution with real slant angle mean.

2) In 70% cases, the centroid angle of normal/italic Chinese characters arranges in the interval $[\, mean-8, mean+8 \,]$, where mean denotes the real slant angle of character.

From a probabilistic perspective, the centroid angle of a Chinese character can be demonstrated by the slant angle of it plus additive a Gaussian noise with zero mean, so that

$$\theta_v = \theta_l + \varepsilon. \quad (3)$$

where θ_l is a real slant angle of a character and ε is a zero-mean Gaussian-distributed noise with the standard deviation σ. The generative model is illustrated in Fig. 3.

Therefore, the conditional distribution and exception of centroid angle θ_v given a real slant angle θ_l are

$$p(\theta_v \mid \theta_l) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta_l)}{2\sigma^2}}. \quad (4)$$

$$E(\theta_v \mid \theta_l) = \theta_l. \quad (5)$$

where σ is the standard deviation of Gaussian distribution, it is set to 8 according to the statistical results.

Based on the formula (5), the slant angles of Chinese characters can be estimated by the average value of the centroid angles of characters belonging to the same class when sufficient characters exist in a document.

III. ESTIMATION OF SLANT ANGLE

In most Chinese advertising images, there are not sufficient characters to estimate slant angle. Besides, Arabic number and English character scattered in document also bring challenge to estimation of slant angle since the centroid angles of non-Chinese characters are not reliable feature to inference the slant angles. To overcome these difficulties,

MRF model is presented to estimate slant angle by considering of both the Gaussian property of centroid angle and the font-face similarity of neighboring character.

A. Esetimating Slant Angle based on MRF

MRF [8] is an undirected graphical model to estimate probability distribution conditioned on observations.

As discussed in section 2, there is a strong correlation between the centroid angle and the slant angle of a character. Besides, the font faces of neighboring characters usually are same. Based on these facts, the problem of estimating slant angles of characters in Chinese advertising images can be modeled by MRF, as shown in Fig. 4, in which slant angles are latent variables, and centroid angles are visible variables corresponding to slant angles plus a Gaussian noise.

According to the definition of MRF, we can formulate estimation of slant angle into slant angle labeling problem: given the centroid angle set $V = \{\theta_{v1}, \theta_{v2}, ...\}$, the objective is to find the best latent slant angle label $L = \{\theta_{l1}, \theta_{l2}, ...\}$ to minimize the total graph energy J.

The total energy function for the model is defined as

$$J = \beta \sum_{\{i,j\}} | \theta_{li} - \theta_{lj} | + \eta \sum_{i} | \theta_{vi} - \theta_{li} | . \qquad (6)$$

where i and j are indices of neighboring characters, θ_{li} and θ_{vi} denote the slant angle and the centroid angle of character i respectively, β and η are combination coefficients. The first term in the right side of formula (6) guarantees the similarity of neighboring characters, and the second term guarantees the strong correlation between the centroid angle and the slant angle of a character.

B. Algorithm Description

As shown in Fig. 5, the characters are first located by Liu's work [10].

For slant angle estimation of the MRF model, we use iterated conditional modes (ICM) [9] to minimize the total graph energy J. It starts with estimations of the slant angles, and then, for each character, it chooses an angle in a set of possible slant angles giving the largest decrease of the energy function. This process is repeated until convergence.

The set of possible slant angles $L = \{\theta_{l1}, \theta_{l2}, ...\}$ has to be first initialized, which could be done by simply setting $\theta_{li} = \theta_{vi}$ for all i. However, considering the effectiveness and efficiency, θ_l should be assigned to more reliable value. Based on the form (5), the real slant angles can be estimated by the average value of the centroid angles of characters belonging to the same face. Therefore, the centroid angles of characters are first clustered, and then the average values of every cluster are calculated as the possible latent slant angles. If any two centroid angles θ_{vi} and θ_{vj} satisfy the following condition (7), they are clustered into same class.

$$| \theta_{vi} - \theta_{vj} | < \varepsilon . \qquad (7)$$

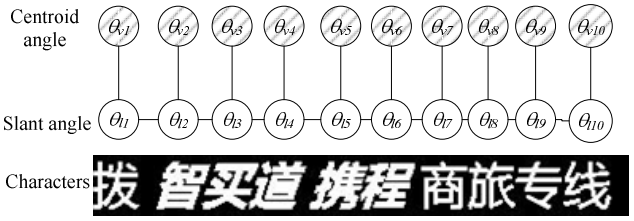

Figure 4. A example of MRF model for estimating slant angle.

Figure 5. Example of characters location by Liu's work.

where i and j denote the indices of any two characters, ε is a threshold which can be set to 5 according to the statistical results.

The overall process can be demonstrated as follow steps:

Step 1: Locating characters, and then initializing the latent slant angles $L = \{\theta_{l1}, \theta_{l2}, ...\}$ and the centroid angles $V = \{\theta_{v1}, \theta_{v2}, ...\}$;

Step 2: Keeping all other variables fixed and setting θ_{li} to whichever state has the lower energy by formula (8), and then repeating the update for every character;

$$\theta_{li}^{(k+1)} \leftarrow \arg\min_{\theta_{li} \in L} J(\theta_{li}^{(k)}) . \qquad (8)$$

Step 3: Finishing the process when formula (9) is satisfied, otherwise, moving to Step 2.

$$\theta_{li}^{(k+1)} = \theta_{li}^{(k)}, \forall l_i . \qquad (9)$$

In our experiment, the process iterates only one time to improve the efficiency while the effectiveness can be guaranteed.

A character will be identified as italic style and be rectified according to θ_{li} which minimize the energy if the following condition is satisfied

$$| \theta_{li} | > \lambda, \forall i . \qquad (10)$$

where λ is a threshold , it can be set to 8 since the absolute values of most italic characters' slant angles are greater than 8.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Statistical Study on The Centroid Angle of Character

In this section, we present a statistical study on the centroid angles of 3755 categories of the daily-used Chinese characters.

Considering the usual printed Chinese characters in Chinese advertising images with font families as Song font,

Black font, Round font, large number of samples are constructed in order to study on the statistical properties of the centroid angle of Chinese character. The number of samples is 3755×3×2, where 3755 denotes the number of categories of the daily-used Chinese characters, 3 is the number of usual font families, 2 denotes normal and italic font faces.

Fig. 6 shows the statistical distributions of the centroid angles of normal and italic faces Chinese characters with different font families. Table I gives the corresponding statistical properties.

We also study on the centroid angle of non-Chinese character. Table II gives the corresponding statistical properties.

Based on these statistical analyses, we can draw conclusions:

1) The centroid angle of regular/italic Chinese character approximately obeys a Gaussian distribution with real slant angle mean.

2) In 70% cases, the centroid angle of normal/italic Chinese characters arranges in the interval $[\, mean-8, mean+8\,]$, where mean denotes the real slant angle of character.

3) The centroid angle of non-Chinese character is not reliable feature to estimate the real slant angle, especially for lowercase English letter.

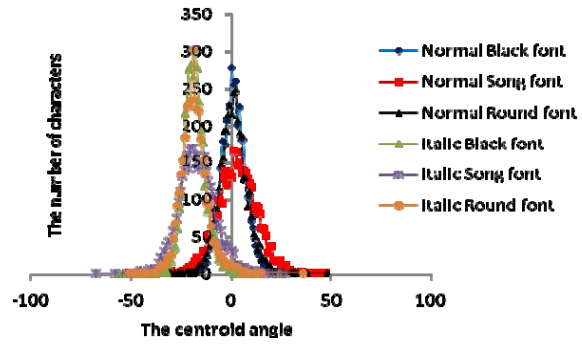

Figure 6. Distribution of the centroid angle of Chinese character.

TABLE I. THE STATISTICAL PROPERTY OF DISTRIBUTION OF CHINESE CHARACTER CENTROID ANGLE

	Mean	*The standard deviation*
Normal Black font	1.2	6.2
Normal Song font	2.8	10.2
Normal Round font	1.0	7.0
Italic Black font	-18.3	5.6
Italic Song font	-16.8	9.4
Italic Round font	-18.5	6.7

TABLE II. THE STATISTICAL PROPERTY OF CENTROID ANGLE OF NON-CHINESE CHARACTER

	Mean	*The standard deviation*
Normal Arabic number	0.9	8.3
Normal Uppercase English letter	1.7	9.9
Normal Lowercase English letter	-5.1	27.8
Italic Arabic number	-16.7	7.3
Italic Uppercase English letter	-18.1	8.7
Italic Lowercase English letter	-27.2	27.8

TABLE III. SUCCESS RATE OF ITALIC CHARACTERS DETECTION

	Normal	*Italic*	*Mixed Normal /Italic*
Our method	95.86%	95.28%	93.26%
Xia's method[3]	96.59%	91.12	
Lee's method[4]	85.00%	86.22	

B. *Evaluation of Characters Detection*

To evaluate our method for italic characters identification, we extract 100 normal text lines, 100 italic text lines and 100 mixed normal/italic text lines from Chinese advertising images downloaded from: http://www.sohu.com and http://www.sina.com.cn. To evaluate the proposed method, we compared our method with Xia'x method [3] and Lee's method [4]. The comparison is shown in table III.

Our method outperforms Xia's approach in italic and mixed normal/italic cases since Xia's method cannot deal with left-italic characters and mixed normal/italic characters. The performance of Lee's method is poor for Chinese advertising images mainly since it cannot determine effectively the styles of the characters with a few or no vertical strokes and the characters with vary size arranged in a text line.

It is noted that our method can deal with left-italic and mixed normal/italic characters. Our method also can detect the italic characters in mixed Chinese/English/Arabic texts since there are only a few non-Chinese characters scattered in Chinese advertising images and their slant angles can be exactly estimated by our approach based on MRF, which considers the font-face similarity of neighboring characters. Most detection errors of our approach are resulted from non-Chinese characters in the majority in texts.

V. CONCLUSION

In this paper, we propose the centroid angle of character and present a statistical study on it. Furthermore, the font-face similarity of neighboring characters, in addition to the statistical property of the centroid angle of character, is used

to construct a MRF model to estimate the slant angle of character. Experimental results have demonstrated that our method is meaningful for texts in Chinese advertising images.

For further improving the performance of our approach, we need to identify the non-Chinese characters and estimate their angle by more reliable feature.

ACKNOWLEDGMENT

This work has been supported by the National Key Technology R&D Program of China under Grant No. 2009BAH48B 02, 2009BAH43B04, 2011BAH16B01 and 2011BAH16B02. The authors thank the anonymous reviewers for valuable comments.

REFERENCES

[1] K.C. Fan, C.H. Huang and T.C. Chuang, "Italic detection and rectification," In Proc. Conf. Image Processing (ICIP 05), Sept. 2005, pp. 530-533.

[2] G. Nicchiotti and C. Scagliola, "Generalised projections: a tool for cursive handwriting normalisation," In Proc. Conf. Document Analysis and Recognition (ICDAR 99), Sept. 1999, pp. 729-732.

[3] Y. Xia, C.H. Wang and R.W. Dai, "Segmentation of Mixed Chinese/English Document Including Scattered Italic Characters," In Proc. Conf. Computer Processing of Oriental Languages (ICCPOL 06), Dec. 2006, pp. 13-21

[4] H.J. Lee and J.T Huang, "Performance Improvement Techniques for Chinese Character Recognition," In Proc. Conf. Document Analysis and Recognition (ICDAR 05), Aug. 2005, pp. 710-714.

[5] L. Zhang, Y. Lu and C.L. Tan, "Italic font recognition using stroke pattern analysis on wavelet decomposed word images," In Proc. Conf. Pattern Recognition (ICPR 04), Aug. 2004, pp. 835-838.

[6] B.B. Chaudhuri and U. Garain, "Automatic detection of italic, bold and all-capital words in document images," In Proc. Conf. Pattern Recognition (ICPR 98), Aug. 1998, pp. 610-612.

[7] H.F. Ma and D. Doermann, "Adaptive word style classification using a Gaussian mixture model," In Proc. Conf. Pattern Recognition (ICPR 04), Aug. 2004, pp. 606-609.

[8] S. Z. Li, Markov Random FieldModeling in Image Analysis. Springer, 2001.

[9] J. Besag, "On the Statistical Analysis of Dirty Pictures," J. Royal Statistical Soc., Series B, vol. 48, no. 3, pp. 259- 302, 1986.

[10] J. Liu, S.W. Zhang, H.P. Li and W. Liang, "A Chinese Character Localization Method based on Intergrating Structure and CC-Clustering for Advertising Images," In Proc. Conf. Document Analysis and Recognition (ICDAR 11), Sept. 2011.