Improving Scene Text Detection by Scale-adaptive Segmentation and Weighted CRF Verification

Yi-Feng Pan, Yuanping Zhu, Jun Sun, Satoshi Naoi
Fujitsu Research & Development Center CO., LTD.,
No. 56 Dong Si Huan Zhong Rd, Beijing, 100025, P. R. China
{*yfpan,yuanping.zhu,sunjun,naoi*}*@cn.fujitsu.cn*

***Abstract*—This paper presents a hybrid method for detecting and localizing texts in natural scene images by stroke segmentation, verification and grouping. To improve system performance, novelties on two aspects are proposed: 1) a scale-adaptive segmentation method is designed for extracting stroke candidates, and 2) a CRF model with pair-wise weight by local line fitting is designed for stroke verification. Moreover, color-based text region estimation is used to guide segmentation and verification more accurately. Experimental results on ICDAR 2005 competition dataset show that the proposed approach can detect and localize scene texts with high accuracy, even under noisy and complex backgrounds.**

***Keywords*-Text detection; Stroke segmentation; Stroke verification; Conditional random field (CRF)**

I. Introduction

With the fast development of text-based image content analysis applications, e.g. signboard translator and mobile text recognizer [1], scene text detection has received increasing attentions. In the past decades, related techniques have been studied intensively and many methods have been proposed and achieved promising results [2], [3]. However, it remains a challenge due to variations of texts' size, font, alignment and degraded images with light reflection, cluttered background and noise.

The existing text detection methods can be categorized into region-based ones, stroke-based ones and combination ones. Region-based methods retrieve local text patches by a scanning window, from where texture features are extracted and fed into machine learning-based classifiers for removing noises. Then local text patches are grouped together into specific character structures (words or lines) according to their spatial relationships. These methods are robust to noise and image degrading but not suitable for characters with irregular fonts and alignments.

For stroke-based methods, text stroke candidates are extracted by segmentation, verified by feature extraction and classification, and grouped together by clustering. These methods are easy to implement on specific applications because of their intuition and simplicity. However, complex backgrounds often make text strokes hard to segment and verify.

Recently, some combination methods [4], [5] which utilize both regional and stroke information are proposed for text detection task. These methods first estimate values of text's confidence and scale as assistant information by local textural analysis. Then stroke segmentation, verification and grouping, guided by assistant information, are sequentially proceeded for detecting and localizing texts. Although the combination of textual and stroke information can improve detection performance, existing methods are still difficult to accurately segment strokes and remove noises, especially for the images with illumination changes and complex backgrounds.

To improve text detection performance, this paper presents a combinational scene text detection method with novelties on two aspects: 1) a scale-adaptive segmentation algorithm is designed for stroke candidates extraction, and 2) a CRF model with pair-wise weight by local line fitting is designed for stroke verification. Moreover, color-based text region estimation is adopted to guide stroke segmentation and verification accurately. Experimental results on ICDAR 2005 competition dataset show that the proposed approach can detect and localize scene texts with high accuracy, even under noisy and complex background.

II. System Overview

The proposed method contains four major stages: 1) text region estimation for text confidence and scale information retrieval, 2) scale-adaptive segmentation for stroke candidate extraction, 3) weighted CRF based stroke verification for noise removal and 4) stroke grouping for text line (word) generation. Note that all stages are proceeded sequentially and the first stage: text region estimation, is used for guiding the following stages: stroke segmentation and verification more accurately. The flowchart of the proposed method is given in Fig. 1 and detailed description will be given in the following sections.

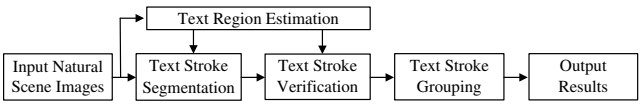

Figure 1. Flowchart of the proposed method.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.158

III. Text Region Estimation

Text's confidence and scale information have been proved helpful for accurate stroke segmentation and verification [4], [5]. Generally, image pyramid is first generated to capture texts with different sizes. Then texture features, such as gray-level statistics and histogram of oriented gradient (HOG), are extracted from local patches and fed into a machine learning classifier for removing noises.

In this paper, to avoid the compression loss from color to gray-level, color HOG features are directly extracted from local patches. It distinguishes the original gray-level one [6] from that pixel gradients are decomposed in RGB channels separately and merged into the target histogram together. For multi-channel merging on each histogram bin, squares of decomposed RGB values are summarized since it can strengthen edge-like responses in any one of three channels. A HOG value corresponding to the bin of orientation i on the pixel x is calculated as

$$GBin_i(x) = \sqrt{\left[\frac{\sum_{j\in\{R,G,B\}} grad_i^j(x)}{3}\right]^2}, \quad (1)$$

where $grad_i^j(x)$ is x's gradient decomposition value on the orientation i and color channel j.

Due to the high efficiency, Waldboost [7] is selected to estimate how likely the local patch contains textural information and it outputs a confidence value. Then, all the pixel confidence and scale values at different pyramid layers are projected back to the original image for calculating the final text confidence and scale maps. These information will be used for guiding subsequent stroke segmentation and verification. Fig. 2 shows some confidence maps generated by gray-level and color HOG features, where the color from blue to red corresponds to the confidence value from low to high. Note that the confidence maps generated by color HOG are less sensitive to illumination changes than gray-level ones.

IV. Text Stroke Segmentation

Stroke segmentation is crucial for system performance and several methods have been proposed for handling this task. Local binarization, as a well-known algorithm in document analysis domain, has been used in scene text segmentation [5]. However, it shows poor performance in the cases where more than two dominant colors exist around texts. Another widely used methods are based on K-means algorithm by clustering all pixels into fixed-size color seeds. Lee et al. [4] proposed a modified K-means algorithm with edge constraint for extracting stroke candidates. Although stroke segmentation performance could be improved, the number of color seeds still needs to be manually selected, which makes segmentation not adaptive.

To address these problems, we propose a scale-adaptive segmentation approach for extracting stroke candidates. A

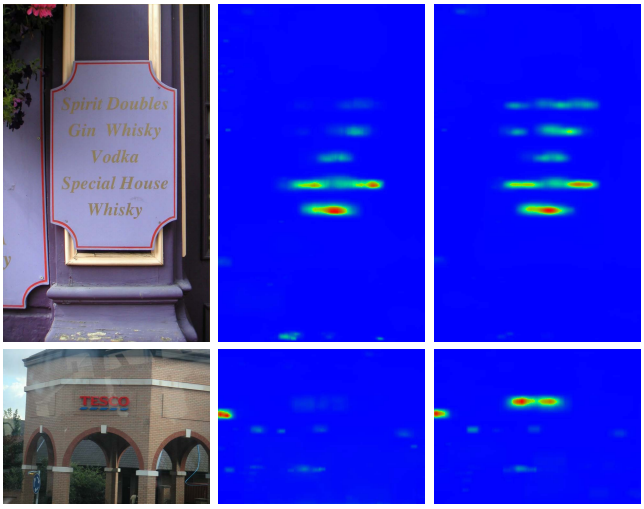

Figure 2. Examples of text confidence maps. From left to right: original images, confidence map generated by gray-level HOG and color HOG, respectively.

graph-based segmentation algorithm [8] is selected since it does not need to fix color number and could preserve details in low-variability image regions while ignoring details in high-variability regions. First, gradient magnitudes of all pixels are computed and normalized by standard deviation in RGB channels. Then, four neighboring pixels over the whole image are linked by edges to initialize a region adjacent graph, where one pixel corresponds to one region. All linking edges are built up as a list and sorted based on their gradient magnitudes in ascending order. From the top to the bottom of the edge list, regions r_i and r_j for each edge can be merged together recursively if they satisfy the following merging rule:

$$Dif(r_i, r_j) < MInt(r_i, r_j), \quad (2)$$

where Dif is the weight value (gradient magnitude) between two merged regions, and the minimal region difference $MInt$ is defined as

$$MInt(r_i, r_j) = \min[Int(r_i)+\tau(r_i), Int(r_j)+\tau(r_j)], \quad (3)$$

where Int is set as the biggest weight value of all inner edges of the region. $\tau(r) = k/|r|$ is a regularization term, where $|r|$ is the region pixel number and k is a constant balancing parameter.

However, manually choosing k's value makes it difficult to extract strokes with different sizes. To solve this problem, we propose a scale-adaptive computation of τ by adding a region scale term whose value is estimated from the text scale map (introduced in section III). The new computation formula of τ is defined as

$$\tau(r) = \frac{k \cdot sc(r)}{|r|}, \quad (4)$$

where $sc(r)$ is the scale value of the region r by averaging all region pixels' scale values. Examples in Fig. 3 show that the proposed scale-adaptive segmentation algorithm can keep stroke shape more regular than the original one.

Figure 3. Examples of text stroke segmentation. From left to right: original images, stroke segmentation with the original computation of τ [8], stroke segmentation with the proposed computation of τ.

V. Text Stroke Verification

For stroke noise removal, some probabilistic graphical model based methods [9], [4], [5] have recently been proposed and they achieved better performance than traditional individual classifiers. The most important advantage of these methods is that neighboring stroke relationships can be considered together with individual stroke properties and formulated into a unified framework, which can be optimized overall.

The method in [9] combined single stroke properties with three-order stroke relationships into a Markov random filed (MRF) model for stroke verification and it achieved better results than single classifiers. The method in [4] differs from the former one in a way that it adopted pair-wise stroke relationships embedded with collinearity weights, which gives a more reliable representation of the stroke relationship. In the method of [5], MRF is replaced by conditional random field (CRF), which is more flexible since it can model stroke posterior probabilistic directly and optimize parameters by supervised training.

In this paper, CRF model is selected for modeling single stroke properties and pair-wise stroke relationships simultaneously. A major improvement compared with the existing methods is that the weight value of pair-wise relationship is estimated by local line fitting. First, any two stroke candidates whose spatial distance is less than a threshold are linked together to generate stroke neighboring graph. The threshold value is proportional to the minimal scale value of strokes x_1 and x_2 as $k \times \min[sc(x_1), sc(x_2)]$, where k is set as 32 empirically.

For each image, given all stoke candidates $X = (x_1, ..., x_n)$ with labels $Y = (y_1, ..., y_n)$. Then (X, Y) is a conditional random field when the probability of Y conditioned on X obeys the Markovian property. The stroke posterior probability can be approximated by

$$P(Y|X) = \frac{1}{Z} \prod_i \Bigg[\psi_u(x_i, y_i, \lambda_u) \cdot \prod_{j \in \mathcal{N}_i} [\lambda_{ij} \cdot \psi_b(x_i, x_j, y_i, y_j, \lambda_b)] \Bigg], \quad (5)$$

where $\psi_u(\cdot, \lambda_u)$ and $\psi_b(\cdot, \lambda_b)$ denote two-class ("text" and "non-text") unary and three-class (both "text", both "non-text" and different labels) binary potential functions, respectively. $\mathcal{N}_i$ denotes i's neighborhood set and λ_{ij} is a coefficient measuring the relationship between i and j.

In the method [5], λ_{ij} is simply normalized by the size of $\mathcal{N}_i$ as $\lambda_{ij} = \frac{1}{|\mathcal{N}_i|}$ which means that all members of $\mathcal{N}_i$ have a same influence on i. It is conflicted with the fact that, neighboring strokes positioning the same text line of i, should be more meaningful to i than others. Based on this observation, method [4] computed λ_{ij} by collinearity based on angle difference as

$$\lambda_{ij} = \frac{1}{2}\left[exp(-\frac{||\theta_{ij} - \theta_{hi}||)^2}{\sigma_\theta^2}) + exp(-\frac{||\theta_{ij} - \theta_{jk}||^2}{\sigma_\theta^2}) \right], \quad (6)$$

in which two angle difference terms: $\theta_{ij} - \theta_{hi}$ and $\theta_{ij} - \theta_{jk}$ are used to measure the collinearity between i and its neighborhood j. Although different weights can be assigned to different neighborhoods by collinearity, this method is still not accurate enough especially when there are many close text lines and so strokes from different lines have confusing influence to each other.

To be more reliable for reflecting the linearity nature of text line, the pair-wise weight is computed by local line fitting in our method. All i's neighborhoods are collected to fit a local text line ℓ_i by weighted least square regression, where the weight $\omega_{j \to i}$ indicates to what extent the neighboring stroke relates to the target stroke. Its value is calculated as

$$\omega_{j \to i} = \omega_{conf}(j) \cdot \omega_{sc}(i, j) \cdot \omega_{dist}(i, j), \quad (7)$$

where $w_{conf}(j)$ denotes the confidence value of j, and $w_{sc}(i, j)$ and $w_{dist}(i, j)$ denote scale and spatial similarity between i and j, respectively. Note that the former two terms are calculated from the text confidence and scale maps. With the fitted text line ℓ_i, the pair-wise weight can be calculated from point-to-line distance:

$$\lambda'_{ij} = exp\left[-\frac{||dist(j, \ell_i)||^2}{\sigma_{\ell_i}^2} \right], \quad (8)$$

where $dist(j, \ell_i)$ is the spatial distance from j to ℓ_i and σ_{ℓ_i} is a normalization factor chosen empirically. Finally, the

weight value is normalized over all i's neighborhoods:

$$\lambda_{ij} = \frac{\lambda'_{ij}}{\sum_{k \in \mathcal{N}_i} \lambda'_{ik}}. \quad (9)$$

The proposed CRF model is trained based on the method of [5]. Multi-layer perceptron (MLP) is selected as unary and binary potential functions and several texture, shape and geometric features are selected as defined in [5]. All parameters of potential functions (λ_u, λ_b) are optimized by the LOG-likelihood of Margin (LOGM) criterion [10] using stochastic gradient descent algorithm. Belief propagation algorithm [11] is chosen for stroke labeling due to its efficient performance. In order to alleviate the computational burden, obvious noises are first filtered out by unary feature thresholds before verification by CRF models.

An example of stroke verification by the proposed CRF model is given in Fig. 4. Note that for the target stroke (red point in sub-figure (c)), its neighboring strokes close to the fitted local line have greater pair-wise weights with bigger sizes in the image. This makes the performance of the proposed weighted CRF model better than previous ones.

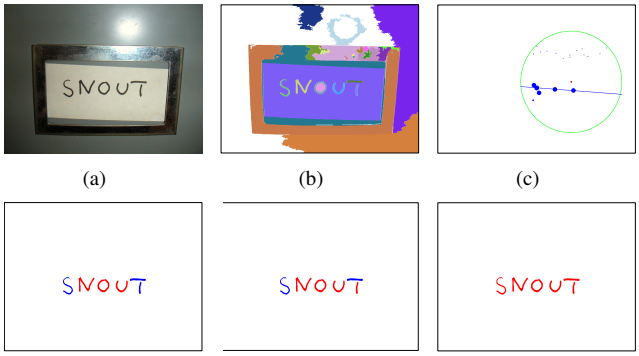

Figure 4. Examples of stroke verification. (a) original image. (b) stroke candidates after pre-filter. (c) pair-wise weights on the stroke "−" of the character "T" (the bigger of the point size, the greater of the weight value) by local line fitting. (d, e, f) verification results of the MLP, the original equal-weighted CRF, and the proposed weighted CRF (estimated strokes and noises are marked in red and blue, respectively).

VI. Text Stroke Grouping

For convenient character recognition and fair comparison with other methods, it is needed to group strokes into text lines or words. Under the assumption that text lines are always straight, we adopt the method of [5] to cluster strokes together followed by partitioning them into lines or words.

In implementation, strokes are first linked together into an overall chain by the Kruskal algorithm which is a standard minimal spanning tree method. Euclidean distance is selected to measure how close of two strokes. Then the overall stroke chain is partitioned into lines by cutting off between-line edges which are confirmed if satisfying one of the following rules:

- The spatial distance between strokes x_1 and x_2 linked by the target edge exceeds a scale related threshold: $dist(x_1, x_2) > k_1 \times min[sc(x_1), sc(x_2)]$.
- The height of the text line (ℓ) containing the target edge exceeds a scale related threshold: $height(\ell) > k_2 \times sc_{avg}(\ell)$. The line height is defined as the sum of distances from the top stroke to the line and the bottom stroke to the line. $sc_{avg}(\ell)$ denotes the average scale value of line strokes. k_1 and k_2 are all fixed to 32 in our experiments.

For word partition, a similar process is implemented and between-word edges are confirmed if the ratio between the stroke bounding box distance of the target edge and the average stroke bounding box distance of separated words exceeds 2. Finally, spatial close text words are grouped together to avoid over-splitting and words with too small size are removed as noises. Some text word localization examples are given in Fig. 5.

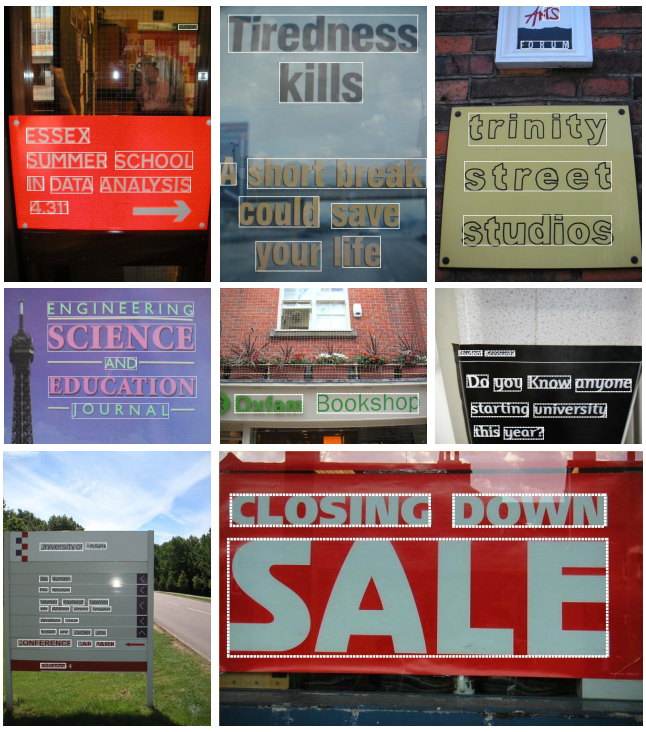

Figure 5. Examples of text word localization.

VII. Experimental Results

To evaluate the performance of the proposed method, we have done experiments on ICDAR 2005 competition dataset [12] which includes 258 training images and 251 test images with English and Arabic number texts.

For training the Waldboost and CRF models, samples were manually collected from the training images. The window size for text region estimation was fixed to 16×16

and 4-orientation color HOG was used to extract local patch features. Image pyramid interval step was set to 1.2 for capturing texts with different sizes.

To evaluate the proposed CRF model, we compared stroke verification performance with four different classifiers: 1) MLP, only MLP based unary potential function is used, 2) EW-CRF, the CRF model with equal pair-wise weights [5], 3) CW-CRF, the CRF model with the collinear pair-wise weights [4], and 4) LW-CRF, the CRF model with the proposed pair-wise weights by local line fitting. Results in Fig. 6 show that 1) CRF models have better performance than unary classifier since information of neighboring stroke relationships are considered, and 2) the proposed local line fitting based pair-wise weight has better performance than others due to it captures the linearity nature of text line.

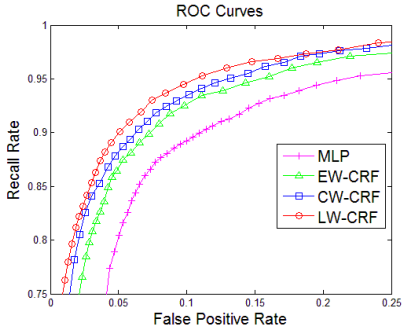

Figure 6. Stroke verification results.

To evaluate the proposed text localization method, we adopted the performance evaluation criterion as the setting of ICDAR 2005 competition [12]. It defines the precision (p) and recall (r) based on area matching ratio. Their harmonic mean: $f\text{-}measure = [(2r)^{-1} + (2p)^{-1}]^{-1}$ is used to evaluate the overall performance. The proposed method is compared with the winner method of ICDAR 2005 competition and Lee et al.'s method [4] using the same dataset and evaluation criteria. Note that the result of [5] has not been compared since it uses the evaluating dataset of ICDAR 2005 competition which is different from others. As shown in Tab. I, the precision and recall of the proposed method is comparable with the other two methods and the best overall performance is achieved.

Table I
TEXT LOCALIZATION RESULTS OF DIFFERENT METHODS.

	Precision (%)	Recall (%)	$f\text{-}measure$
1st ICDAR'05 [12]	62	67	64
Lee et al. method [4]	69	60	64
The proposed method	68	67	67

VIII. CONCLUSION

In this paper, a hybrid method is proposed for detecting and localizing texts in natural scene images. Scale-adaptive segmentation is designed for stroke candidate extraction and weighted CRF based on local line fitting is designed for stroke verification. Furthermore, color-based text region estimation is used for guiding stroke segmentation and verification more accurately. Experimental results show that the proposed method is less sensitive to illumination change, text alignment and background noise and achieves comparative results with the state-of-the-art methods. In the future, following scene character recognition module needs to be developed for a complete text information extraction system.

ACKNOWLEDGMENT

The authors are grateful to the anonymous reviewers for valuable comments.

REFERENCES

[1] J. Liang, D. Doermann, and H.-P. Li, "Camera-based analysis of text and documents: a survey," *Int. J. Document Analysis and Recognition*, vol. 7, no. 2-3, pp. 84–104, 2005.

[2] K. Jung, K. Kim, and A. Jain, "Text information extraction in images and video: A survey," *Pattern Recognition*, vol. 37, no. 5, pp. 977–997, 2004.

[3] J. Zhang and R. Kasturi, "Extraction of text objects in video documents: Recent progress," in *Proc. 8th IAPR Workshop on DAS*, 2008, pp. 1–13.

[4] S. Lee, M. S. Cho, K. Jung, and J. H. Kim, "Scene text extraction with edge constraint and text collinearity," in *Proc. 20th ICPR*, 2010, pp. 3983–3986.

[5] Y.-F. Pan, X. Hou, and C.-L. Liu, "A hybrid approach to detect and localize texts in natural scene images," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 800 –813, 2011.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. CVPR*, 2005, pp. 886–893.

[7] J. Sochman and J. Matas, "Waldboost - learning for time constrained sequential detection," in *Proc. IEEE Conf. CVPR*, 2005, pp. 150–156.

[8] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[9] D.-Q. Zhang and S.-F. Chang, "Learning to detect scene text using a higher-order MRF with belief propagation," in *Proc. IEEE Conf. CVPRW*, 2004, pp. 101–108.

[10] X.-B. Jin, C.-L. Liu, and X. Hou, "Regularized margin-based conditional log-likelihood loss for prototype learning," *Pattern Recognition*, vol. 43, no. 7, pp. 2428–2438, 2010.

[11] Y. Weiss and W. T. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 736–744, 2001.

[12] S. M. Lucas, "ICDAR 2005 text locating competition results," in *Proc. 8th ICDAR*, 2005, pp. 80–84.