Scene Text Extraction with Edge Constraint and Text Collinearity

SeongHun Lee*, Min Su Cho*, Kyomin Jung‡ and Jin Hyung Kim*
Dept of Computer Science, KAIST, Korea
{*leesh, mscho, jkim*}*@ai.kaist.ac.kr*, *kyomin@kaist.edu*‡

Abstract

In this paper, we propose a framework for isolating text regions from natural scene images. The main algorithm has two functions: it generates text region candidates, and it verifies of the label of the candidates (text or non-text). The text region candidates are generated through a modified K-means clustering algorithm, which references texture features, edge information and color information. The candidate labels are then verified in a global sense by the Markov Random Field model where collinearity weight is added as long as most texts are aligned. The proposed method achieves reasonable accuracy for text extraction from moderately difficult examples from the ICDAR 2003 database.

1. Introduction

The purpose of scene text extraction is to separate text regions from camera-captured images before the extracted text is put through a character recognition process. There are many challenging issues related to separating text from the background and maintaining a text region as a single component. For example, the images usually have non-uniform illumination due to uncontrolled lighting conditions and the presence of shadows. In addition, it is common for the content and background in outdoor images to have complex layouts. Such complications make extracting text from scene images a persistent challenge.

For robust isolation of text region in natural scene images, we propose an scene text extraction algorithm seamlessly combining texture features, color and edge information, and text collinearity. The proposed algorithm employs two steps: generation of text region candidates and verification of the label (text or non-text) of the region.

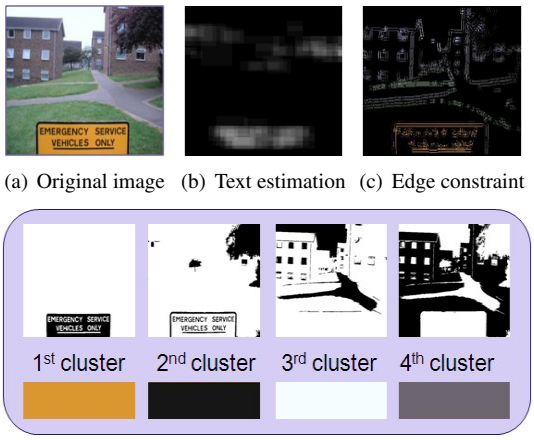
2. Generation of text region candidates

Text regions are predicted by a bottom-up process of image segmentation. Image segmentation is built based on two assumptions about text regions: homogeneity of text color and distinctiveness between text and background regions. Since the pixel colors are similar within a given text region, they can be grouped separately from the background based on their color value. Likewise, since edges are formed in the boundary between the text and background regions, an edge constraint is utilized to force the grouping of text colors and background colors near boundaries into different clusters.

A K-means clustering algorithm [3] is used to find the most dominant K colors from the image and assign each pixel in the image into one of the K colors. The dominant colors are determined by the frequencies of color values in the image. Since the portion of the text region in the image is generally relatively small, bare K-means clustering based on the color distribution often yields inadequate segmentation results.

The modified K-means clustering algorithm finds dominant colors from possible text regions.

By changing the weight of color frequencies that may contain text regions, the text color can be chosen from among the dominant colors. We generate a text saliency map that contains the likelihood that various regions in an image are text regions and this likelihood is used as a weight on the color's pixel count. A text localizer is trained to give a confidence score of how likely it is that each pixel belongs to a text region. The text localizer makes determinations based on the combination of three texture features of multi-scale segments: mean difference, standard deviation, and histogram of gradient [2]. The text localizer is applied to all subregions of the whole image at multiple scales to capture various font sizes, and estimates text areas with the confidence score on a scale of 0 to 1 (Figure 1(b)). Weighted frequency of a color cluster is calculated by multiplying the likelihood of text region and the count of pixels for the color cluster. Since pixels in the text areas has a high confidence score and most of backgrounds have a low confidence score, text color could be chosen as dominant color based on the weighted color frequencies.



(a) Original image (b) Text estimation (c) Edge constraint

(d) Clustering result

Figure 1. text region candidates

Edge constraint is also utilized to force text colored pixels and background color pixels into different clusters in the proposed K-means clustering algorithm. Lists of two-color pairs in edge constraints are obtained from the normal vectors of edge contour pixels [4] (Figure 1(c)). To satisfy the edge constraint, one color is assigned to the cluster of the closest centroid and the other is assigned to the cluster of the other centroid. When a cluster contains both colors of the constrained instance, one color of the instance is moved into the other cluster where its centroid is nearest to the current cluster centroid. By doing so, the constraint is satisfied. Constraint K-means clustering is regarded as finding the solution minimizing the constrained vector quantization error (CVQE) [1]. The formula for CVQE is given below:

$$CVQE_j = \frac{1}{2}\{\sum_{j=1}^{K}\sum_{s\in C_j}\omega_i D_{HCL}(\bar{c}_j, s) + \sum_{(s_i,s_j)\in EC} D_{HCL}(\bar{c}_y(x_i), \bar{c}^*_{ij})\Delta(y(s_i), y(s_j))\},$$

$$D_{HCL}(p,q) = 0.1(l_p - l_q)^2 + \{0.2 + (h_p - h_q)\} * A_{CH}\{c_p^2 + c_q^2 - 2c_p c_q \cos(h_p - h_q)\}, \quad (1)$$

We adopt a color distance measure called hue, chroma, and luminance (HCL) distance (D_{HCL}) to express color difference in HCL color space [5]. HCL distance is more suitable for use with scene text images because it emphasizes hue difference, and hue is an indicator less likely to be affected by changes in illuminations than luminance or RGB color distance approaches.

Like the K-means algorithm, the constraint K-means algorithm is iterative, alternating between the allocation step and the centroid update step. By iteratively updating centroids of clusters to minimize CVQE, K dominant colors are obtained. The initial K value is set as five for handling complex image configurations, but when the distance between two centroids of clusters is less than a given threshold, they are combined during the iteration. Determining optimal K value is still an open problem [8]. Each pixel in the image is assigned into one of the K clusters, and the pixels connected in all 8 directions are grouped into a single component (Figure 1(d)).

3. Verification of text candidate components

The extracted components are determined as text or non-text regions in the verification step. Since characters have common shape characteristics such as aspect ratio and constant stroke thickness which are distinctive from those of background regions, the geometric shape of a single component is considered for verification measurement. In addition, characters in a text line usually have a similar font and color, so spatial relationships among neighboring components are also important factors to determine the label of these components. The Markov Random Field (MRF) model provides a convenient way to model spatial relationships among components as an undirected graphical model [7]. Given the component set $Y = (y_1, y_2, ...)$ and component label $X = (x_1, x_2, ...)$, the joint probability of a pair-wise MRF model can be written as Eq. 2. The probability of a configuration of the components in the MRF model can be calculated by the Belief Propagation [9].

$$P(X,Y) = \frac{1}{Z}\Pi_i \phi_i(y_i|x_i)\Pi_{i,j\in C}\psi_{ij}(x_i,x_j) \quad (2)$$

$\phi_i(y_i|x_i)$ is the one-node potential that represents the probability of the component being one of two classes (text or non-text). In addition, $\psi_{ij}(x_i, x_j)$ is the pair-wise potential which represents the probability of a given set of two neighboring components of being one of three classes (both text, both non-text, one text or one non-text). The features of the components such as their locations, sizes, and shapes for the one-node potential and two-node potential are the same as those used in the our previous system [6].

Even though the pair-wise MRF model shows good performance in many low-level vision applications, it has a limited ability to detect multi-part objects because it only captures two-node relationships. In other words, the traditional pair-wise MRF model cannot capture a unique spatial relationship where more than two characters are aligned on a straight line or a smooth curve. To overcome the limitations of the pair-wise MRF model, we redefined the pair-wise potential by multiplying the co-linearity weight. The colinearity weight between the two components i and j is defined as the function score of θ_{ij} (Eq. 3). θ_{ij} is the angle of a vector connecting the center points of two components i and j. θ_{hi} and θ_{jk} are obtained from the neighboring components of i and j having the least angular distance from θ_{ij}. Compared to the higher-order MRF model [10], our proposed method considers the geometric relationships between up to four components while maintaining low computational complexity of the lower-order MRF framework.

$$\omega_{ij}(y_i, y_j, \theta_{ij}) = \quad (3)$$
$$\frac{1}{2}\{exp(-\frac{(||\theta_{ij} - \theta_{hi}||)^2}{\sigma_\theta^2}) + exp(-\frac{(||\theta_{ij} - \theta_{jk}||)^2}{\sigma_\theta^2})\},$$
$$\theta_{hi} = \arg\min_{\theta_{hi}} ||\theta_{ij} - \theta_{hi}||, h \in n_i, h \neq j$$
$$\theta_{jk} = \arg\min_{\theta_{jk}} ||\theta_{ij} - \theta_{jk}||, k \in n_j, k \neq i$$

Collinearity weight becomes high when more than three components are aligned. By using the collinearity weighting scheme in the Belief Propagation approach, the influence from aligned neighbors is stronger than that from non-aligned neighbors in determining the label of a component. Figure 2(b) shows the collinearity weight among all the components, where the aligned components have higher weights than the non-aligned. In Figure 2(c), most text components remain (blue) and the non-text components are removed (red).

(a) original (b) co-linearity weight (c) final

Figure 2. Ex. Component verification

4. Experimental Results

We collected over 3000 images from various environments (signboards, book covers and English and Korean characters) and manually segmented text regions for the ground truth images[2].

We compared our method with two others applied in the International Conference on Document Analysis and Recognition (ICDAR) 2005 Text Location Competition (Table 1). For fair comparison, we adopted the 2005 ICDAR performance evaluation criteria of defining precision rate, recall rate, and standard f measure based on area matching ratio. As shown in Figure 3, the text areas were well detected in most cases, even those with complex backgrounds. Non-text areas were eliminated effectively, although there were some errors. However, we found that the proposed method cannot handle strong reflection effects because color information in the affected areas is missing. The performance is also dependent on the segmentation result of the K-means clustering method. For future work, a feedback mechanism from the text verification step is needed to handle the segmentation errors.

Table 1. Text detection result

	Precision	Recall	f
1st ICDAR' 05	0.62	0.67	0.62
2nd ICDAR' 05	0.60	0.60	0.58
Proposed method	0.69	0.60	0.64

Figure 3. Text extraction results

5. Conclusion

In this paper, a robust scene text extraction algorithm utilizing edge constraints and text collinearity is proposed. Color, edge and texture information from images are utilized to generate text region candidates. Then, the candidate labels are verified in a global sense on an MRF model, in which collinearity weight is added to consider geometric relationships between text components. Our study shows that the proposed method extracts text regions reasonably well while eliminating most non-text regions for the ICDAR 2003 competition database.

Acknowledgements

This work was supported by the KOSEF grant funded by the Korea government 2009-0078943, BK21 and NAP of Korea Research Council of Fundamental Science & Technology

References

[1] M. Bilenko, S. Basu, and R. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*, 2004.

[2] S. Hanif, L. Prevost, and P. Negri. A cascade detector for text detection in natural scene images. In *ICPR*, pages 1–4, 2008.

[3] J. Hartigan and M. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[4] T. Kasar and R. A. Ganesan. COCOCLUST: Contour-based Color Clustering for Robust Binarization of Colored Text. In *CBDAR, IEEE*, 2009.

[5] E. Kim, S. Lee, and J. Kim. Scene Text Extraction using Focus of Mobile Camera. In *ICDAR*, pages 166–170. IEEE, 2009.

[6] S. Lee, J. Seok, K. Min, and J. Kim. Scene Text Extraction using Image Intensity and Color Information. In *CJK Joint Workshop on PR*, 2009.

[7] S. Li. Markov random field models in computer vision. *LNCS*, 801:361–370, 1994.

[8] J. Park and S. Park. Detection of Text Region and Segmentation from Natural Scene Images. *Advances in Visual Computing*, pages 666–671, 2005.

[9] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

[10] D. Zhang and S. Chang. Learning to detect scene text using a higher-order MRF with belief propagation. In *CVPR, IEEE*. Citeseer, 2004.

[2]KAIST scene text database is available at http://ai.kaist.ac.kr/home/DB/SceneText