Localize and Segment Scene Text

Xiaoqian Liu[1] and Weiqiang Wang[1,2]

[1] School of Computer and Control Engineering,
University of Chinese Academy of Sciences, Beijing, China
[2] Key Lab of Intell. Info. Process.,
Institute of Computing Technology, CAS, Beijing China

Abstract. In this paper, we present a scene text extraction approach which can realize text localization and segmentation simultaneously. Two popular paradigms (machine learning method and rule-based method) are combined to achieve competitive performance. For a given image, a sliding window is used to detect scene text. The texture feature Local Binary Pattern is extracted to represent the content of each window, and an unbalanced SVM classifier is designed to identify candidate text regions. Then, candidate text windows are further verified using color contrast and binarized by an adaptive local thresholding computation to get candidate text connected components. Further, non-text ones among them are removed utilizing some empirical rules. Finally, text connected components are linked into text lines according to their spatial relationships and appearance similarities. The evaluation results on two challenging and standard datasets ICDAR 2003 and ICDAR 2011 demonstrate that the proposed approach can effectively detect and segment scene text with different sizes, fonts, colors and arrangement directions.

Keywords: LBP, SVM, rule-based method, text detection, text extraction.

1 Introduction

With the advent of the information age, more and more intelligent devices flood into the life of humans. Semantic understanding is an indispensable part of realizing various intelligent applications on the devices, so automatically detecting and recognizing text in images and videos as a key technique has become a hot research topic which attracts a lot of researchers' attention. In the last decade, text extraction technology has been exploited in many applications, such as intelligent traffic management system [1], postal letter sorting [2], helpful devices for blind and visually impaired people [16], and automatic caption extraction system [4]. Graphic text in images and videos has two forms, artificial text and scene text. The artificial text is manually overlaid on images or video frames and generally owns uniform color, size, position and arrangement direction, e.g., captions. Scene text refers to those captured by cameras as a part of natural scenes. Compared with artificial text, the colors, fonts, sizes and locations of scene text are diverse and they are also vulnerable to the influence of illumination variance

B. Huet et al. (Eds.): PCM 2013, LNCS 8294, pp. 59–68, 2013.

and shooting angles. So, figuring out an effective scene text extraction approach is still a challenging problem.

Most text extraction algorithms can be decomposed into two parts, text detection and text segmentation, which aim to localize text regions and identify text pixels respectively. Overall, the existing mainstream methods for text extraction can be broadly classified into two categories, the rule-based method and machine learning method. The rule-based methods are relatively simple and intuitive in system construction, and they generally utilize some distinct characteristics summarized from colors, edges or structures of characters to filter out non-text areas or pixels based on some empirical rules. For instance, Li *et al.* [5] first extract connected components by the Maximally Stable Extremal Region(MSER) algorithm and then filter out non-text components based on some geometric rules. Finally, some false positives are eliminated according to stroke width of text. Chen *et al.* [6] first exploit a multi-scale Laplacian of Gaussian (LoG) edge detector to obtain the edge set, and then combine both edge and color information to localize text regions. The rule-based methods usually involve many thresholds and too much dependence on the choice of thresholds is prone to reduce the robustness and objectivity of the system. To overcome this disadvantage, some text extraction approaches integrate the machine learning techniques, like SVM, Neural Networks, into their system. In [7], a cascade with four strong classifiers are trained by an AdaBoost machine learning algorithm to classify text candidate regions. Thillou *et al.* [8] exploit an unsupervised clustering algorithm to separate text pixels from background by combining color feature with spatial information. Lukas *et al.* [9] use the SVM to classify the extracted Maximally Stable Extremal Regions (MSERs) into character and non-character regions. In [10], text detection and recognition are realized based on a scalable unsupervised feature learning method.

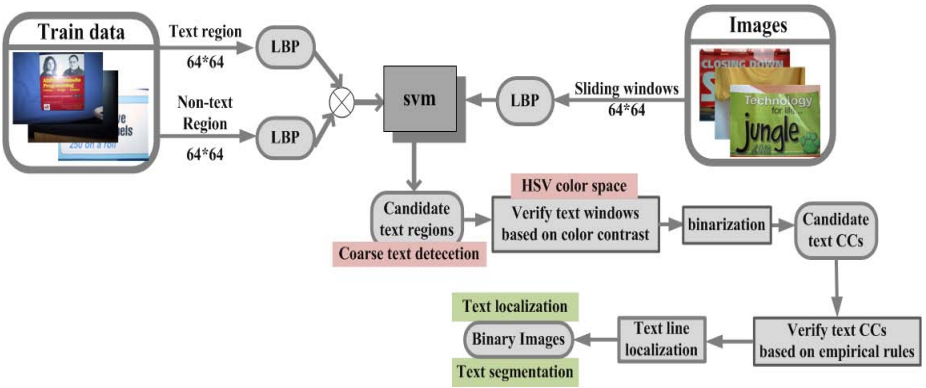

Fig. 1. The framework of the proposed approach

In this paper, we present a scene text extraction approach combining the advantages of the rule-based algorithm and the machine learning algorithm together. First, an unbalanced SVM classifier is designed to efficiently detect the existence of scene text at each position of sliding windows, so that candidate text regions can be coarsely localized. Next, the color contrast information is exploited to further verify and identify text windows, and the binarization computation based on adaptive local thresholds are conducted on them to obtain candidate text connected components. Then, non-text connected components are removed out based on some empirical rules. Finally, text lines are obtained by linking the text components with similar appearance and close spatial distance together.

The rest of this paper is organized as follows. Section 2 presents the proposed method. The experimental results are reported in Section 3. Section 4 concludes the paper.

2 Methodology

The framework of the proposed approach is shown in Figure 1. For a given image, a sliding window is used to detect scene text. At each window position, the Local Binary Pattern (LBP) feature is extracted and exploited by an SVM classifier to coarsely localize candidate text regions. Next, candidate text windows are further verified based on color contrast and binarized by an adaptive local thresholding computation to get candidate text connected components. Further, non-text ones among them are removed utilizing some empirical rules. Finally, text connected components are linked into text lines according to their spatial relationships and appearance similarities. The more details are described in the corresponding subsection respectively.

2.1 Localizing Candidate Text Regions Using SVM

For a given image, a m by n window slides over it from top to bottom and from left to right with a certain step, and at each window position an SVM classifier is exploited to judge whether scene text exists in it. In our system, the size of the sliding window is 64 by 64 pixels and the sliding step is 32 pixels. In each window, the texture feature of Local Binary Patterns (LBP) is extracted, which has been acknowledged to be effective and has a good invariance to lightness. The basic representation of LBP is utilized in our approach based on the consideration of computational efficiency. Concretely, for each pixel in the window, the grayscale values of its eight neighboring pixels are compared with its value. The neighboring pixel is marked as 1 if it is greater than the current pixel; otherwise, it is marked as 0. Thus, an 8 bit unsigned number is formed through concatenating a sequence of 1s and 0s corresponding to the eight neighboring pixels as the LBP code of the current pixel. Further, for the current window, the statistical histogram of these LBP codes is computed to describe its content as the feature representation $\mathbf{x}$.

To train the classification model, we randomly sample 10140 positive instances and 100000 negative instances from the training images of ICDAR2003. Figure 2 gives some examples of positive instances and negative instances in our training datasets. Each training instance has the size of 64×64 pixels and different instances are permitted to be partially overlapped. The SVM is chosen as the classification model, and the linear kernel is used due to its low computation complexity. Formally, the classification function $y = sign(\mathbf{w}^T\mathbf{x}+b)$. If $\mathbf{w}^T\mathbf{x}+b > 0$, the current window is classified as text window with $y = 1$; otherwise, it is a non-text window with $y = 0$. In our system, the SVM is exploited to quickly filter out most non-text regions. At the same time, we also expect that the obtained SVM classifier has a recall as high as possible, so as to guarantee that as many text windows as possible are remained. To realize this, after we have obtained the classifier $y = sign(\mathbf{w}^T\mathbf{x} + b)$, the bias b is adjusted. Apparently a lower bias b corresponds to a higher recall and a lower accuracy. Our experiments show that through the classifier, lots of non-text region can be removed. To avoid filtering out text sub-images in this step, we adjust the parameter of classifier to ensure that the recall is equal to 1. Exploiting the SVM classifier obtained, each window is classified, and all the windows which are classified as text windows are merged together to form candidate text regions. We assume that scene text consist of more than one word, so if the text sub-image is isolated, it will be taken as non-text one and removed according to the constraint of size.

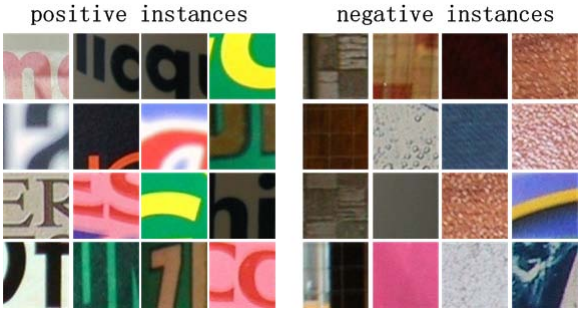

Fig. 2. The examples of positive instances (text windows) and negative instances (non-text windows)

2.2 Extracting Text Connected Components

The scene text on the billboards and signs usually owns strong color contrast with background so that they can be easily captured and recognized by human eyes. The color characteristics of scene text can be exploited to extract text connected components in the following computation. For those candidate text regions identified by the SVM classifier, the windows containing scene text are

further verified, and the binarization computation is conducted on them to segment out the text connected component through an adaptive local thresholding computation. For an image I containing k candidate text regions, each text region $R_i, (i = 1, 2, \ldots, k)$ are formed by a group of candidate text windows R_i^j, where j denotes the j^{th} windows in R_i. Here we should note that these text windows R_i^j are just those labeled by the SVM classifier. Then, the system calculates the mean $m_{i,j}$ and variance $\sigma_{i,j}^2$ of the hue values of pixels in R_j^i in the HSV color space. If the following condition

$$\sigma_{i,j} > \theta \tag{1}$$

is satisfied, the candidate text window R_i^j is binarized with threshold $m_{i,j}$, where threshold $\theta = 0.3$ in our system. Through the above computation, a set of candidate text connected components can be obtained.

Further, five rules are established by us and utilized to identify text connected components. They are summarized as follows.

- The text connected components can not border on the boundaries of images.
- The area N_c of text connected components, i.e., the number of pixels forming the connected component, should not be too small. In our system, $N_c \geq 10$.
- The height h_c and width w_c of the minimum enclosing rectangle (MER) of a text connected component should not be too small. In our system, both h_c and w_c should exceed 8 pixels.
- The aspect ratio $R_a = h_c/w_c$ of a text connected component should satisfy a certain ratio according to the structure of text strokes. In our system, $R_a \in [0.5, 2]$.
- If the occupation ratio $R_o = N_c/(h_c * w_c)$ is more than 0.8 or less than 0.1, the corresponding connected components is removed out as background.

The above rules are designed based on the geometric characteristics of characters, and the specific choice of thresholds is based on trial and error. Through these simple rules, non-text connected components can be effectively removed.

2.3 Text-Line Localization and Text Segmentation

Finally, these isolated characters segmented are linked into text lines according to their spatial relationships and appearance similarities. The spatial location of each text connected component c_i is represented by its centroid. We used the Euclidean distance d_{ij} to measure the proximity of two connected components c_i and c_j. Let w_i, w_j and h_i, h_j denote the width and height of MERs of c_i and c_j respectively. If the $d_{ij} \leq (w_i + w_j)$, c_i and c_j will be considered to be close enough. Appearance similarity requires the MERs of two linked connected components own similar shapes, i.e.,

$$|w_i - w_j| \leq \beta * \max\{w_i, w_j\} \tag{2}$$

$$|h_i - h_j| \leq \beta * \max\{h_i, h_j\} \tag{3}$$

where $\beta = 0.2$ in our system. Although the arrangement directions of text may be arbitrary, we assume that the characters in the same word or sentence should be approximately in a straight line. In our implementation, we combine the connected components whose centroids are approximately collinear. Only when the connected components satisfy all of the three constraints above, they are combined into a text line and marked by its minimum enclosing rectangle with the angle consistent to the arrangement direction.

3 Experiments

To validate the proposed approach, our evaluation experiments use two challenging datasets ICDAR 2003 [14] and ICDAR2011 [16], which are the standard public datasets in the competition of scene text localization. They contain 251 and 491 images for testing respectively.

3.1 Results of Localizing Candidate Text Regions

In this group of experiments, we evaluate whether the unbalanced SVM classifier can effectively filter out non-text windows by using the simple LBP feature. A window is labeled as a text window if it contains some pixels belonging to text regions in the ground-truth. Through adjusting the bias b in the classification function, two performance curves of precision vs. recall can be obtained corresponding to ICDAR 2003 and ICDAR 2011 datasets respectively, as shown in Fig. 3. The recall is the ratio of the number of text windows correctly classified to the total number of text windows in the ground-truth, and the precision is the ratio of the number of windows correctly classified to the total number of windows. It can be seen that when the recall approaches 100%, the precision is still above 55%, which means that most non-text windows are filtered out due to the computation of this step. Since the classification model is established based on the images of ICDAR 2003, the result on ICDAR 2011 is slightly lower. Whatever, it shows that the SVM classifier based on the LBP can effectively filter out most of non-text regions.

Table 1. Performance comparison of text localization on ICDAR 2003 dataset

Method	P	R	F
Our method	**0.72**	**0.69**	**0.705**
Epshtein*et al.* [11]	0.73	0.60	0.66
Lukas *et al.* [12]	0.72	0.62	0.67
Becker [13]	0.62	0.67	0.64
Ashida [14]	0.55	0.46	0.50

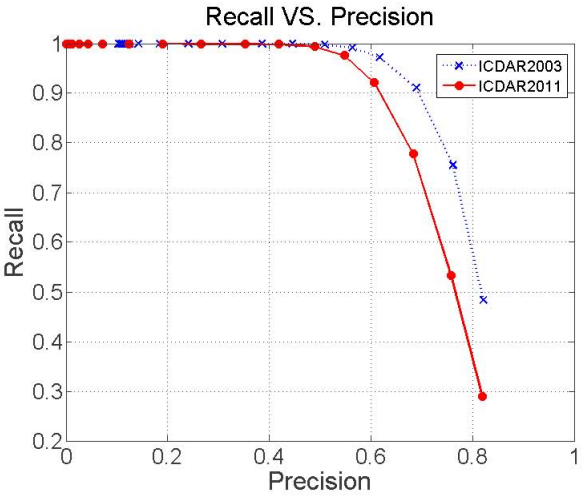

Fig. 3. The results of text sub-images classification. The blue dashed line and the solid red line describe the results on ICDAR 2003 and ICDAR 2011 respectively.

Table 2. Performance comparison of text localization methods on ICDAR 2011 dataset

Method	P	R	F
Kim* [16]	0.83	0.625	0.713
Our method	**0.70**	**0.68**	**0.69**
Lukas *et al.* [15]	0.647	0.731	0.687

Fig. 4. The results of text localization and extraction on ICDAR 2003 dataset

Fig. 5. The results of text localization and extraction on ICDAR 2011 dataset

3.2 Results of Localizing and Segmenting Scene Text

Three metrics are adopted to evaluate the proposed approach, i.e., precision P, recall R and F-measure F. P is the ratio of area of the correctly localized text regions N_r to area of the whole detected regions N_d by our system, where N_r can be calculated by taking an operation "And" between the area of text regions localized and the text regions as groundtruth. R is the ratio of area of N_r to area of the text regions in ground-truth. The area of a region is represented by the number of pixels in it. $F = \frac{2 \times P \times R}{P+R}$ is a composite indicator combining P and R. The localization results on ICDAR 2003 and ICDAR 2011 are shown in Table 1 and Table 2 respectively. For ICDAR 2003 dataset, the localization performance of our approach is the best according to R. Although the corresponding precision P is slightly lower compared with the best one, the F-measure is still the best compared with other state-of-the art methods, where [13] and [14] are the champions of ICDAR 2005 competition and ICDAR 2003 competition respectively. The performance of our approach is also competitive on the latest dataset ICDAR 2011 and they are comparable with the winner of ICDAR 2011 Robust Reading competition marked by asterisk in Table 2.

In addition to localizing scene text regions, text segmentation can be realized simultaneously in our approach. The examples of results of text localization and extraction on both datasets are shown in Fig. 4 and Fig. 5 respectively. The red rectangles enclosing the text regions own the same inclination angles with the arrangement direction of text lines.

4 Conclusion

In this paper, we present a scene text extraction approach which can realize text localization and segmentation simultaneously. Two popular paradigms (machine learning method and rule-based method) are combined to achieve competitive

performance. The evaluation experiments on two challenging public datasets demonstrate the proposed approach can effectively handle with scene text with different colors, sizes, fonts and arrangement directions.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 61232013, No. 61271434, No. 61175115.

References

1. Park, S.H., Kim, K.I., Jung, K., Kim, H.J.: Locating car license plates using neural networks. Electronics Letters 35(17), 1475–1477 (1999)
2. Liu, C.-L., Koga, M., Fujisawa, H.: Lexicon-driven segmentation and recognition of handwritten character strings for japanese address reading. IEEE Transactions on Pattern Analisis and Machine Intelligence 24(11), 1425–1437 (2012)
3. Ezaki, N., Bulacu, M., Schomaker, L.: Text detection from natural scene images: towards a system for visually impaired persons. In: Proc. of 17th International Conference on Pattern Recognition, Cambridge, England, UK, August 23-26 (2004)
4. Liu, Q., Jung, C., Kim, S., Moon, Y., Kim, J.: Stroke filter for text localization in video images. In: Proc. of IEEE International Conference on Image Processing, Atlanta, GA, USA, October 8-11, pp. 1473–1476 (2006)
5. Li, Y., Lu, H.: Scene text detection via stroke width. In: Proc. of 21st International Conference on Pattern Recognition, Tsukuba, Japan, November 11-15, pp. 681–684 (2012)
6. Chen, X., Yang, J., Zhang, J., Waibel, A.: Automatic detection and recognition of signs from natural scenes. IEEE Transactions on Image Processing 13(1), 87–99 (2004)
7. Chen, X., Yuille, A.L.: Detecting and reading text in natural scenes. In: Proc. of Computer Vision and Pattern Recognition, Washington, DC, USA, June 27-July 2, pp. 366–373 (2004)
8. Mancas-Thillou, C., Gosselin, B.: Spatial and color spaces combination for natural scene text extraction. In: Proc. of the 13th International Conference on Image Proceedings, Atlanta, GA, October 8-11, pp. 985–988 (2006)
9. Neumann, L., Matas, J.: A method for text localization and recognition in real-world images. In: Proc. of Asian Conference on Computer Vision, New Zealand, November 8-12, pp. 30–35 (2010)
10. Carpenter, B., Case, C., Satheesh, S., Suresh, B., Wang, T., Wu, D.J., Ng, A.Y.: Text detection and character recognition in scene images with unsupervised feature learning. In: Proc. of the 11th International Conference on Document Analysis and Recognition, Beijing, China, September 18-21, pp. 440–445 (2011)
11. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: Proc. of the 23rd IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, June 13-18, pp. 2963–2970 (2010)
12. Neumann, L., Matas, J.: Text localization in real-world images using efficiently pruned exhaustive search. In: Proc. of the 11th International Conference on Document Analysis and Recognition, Beijing, China, September 18-21, pp. 687–691 (2011)
13. Lucas, S.M.: Icdar 2005 text locating competition results. In: Proc. of the 8th International Conference on Document Analysis and Recognition, Seoul, Korea, August 29-September 1, pp. 80–84 (2005)

14. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R.: ICDAR 2003 robust reading competitions. In: Proc. of the 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, UK, August 3-6, 2003, pp. 682–687 (2003)
15. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: Proc. of the 25th IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, June 16-21, pp. 3538–3545 (2012)
16. Shahab, A., Shafait, F., Dengel, A.: ICDAR 2011 robust reading competition challenge 2: Reading text in scene images. In: Proc. of the 11th International Conference on Document Analysis and Recognition, pp. 1491–1496 (2011)