A Robust Video Text Extraction and Recognition Approach using OCR Feedback Information

Guangyu Gao[1], He Zhang[2] and Hongting Chen[3]

[1]School of Software, Beijing Institute of Technology, Beijing, China.
[2] Operations Office (Beijing), People's Bank of China, Beijing, China
[3]School of Computer, Beijing University of Posts and Telecom. Beijing, China
Email : guangyu.ryan@gmail.com,zhanghe86@126.com, hongting.chen@yahoo.com

Abstract. Video text is very important semantic information, which brings precise and meaningful clues for video indexing and retrieval. However, most previous approaches did video text extraction and recognition separately, while the main difficulty of extraction and recognition with complex background wasn't handled very well. In this paper, these difficulty is investigated by combining text extraction and recognition together as well as using OCR feedback information. The following features are highlighted in our approach: (i) an efficient character image segmentation method is proposed in consideration of most prior knowledge. (ii) text extraction are implemented both on text-row and segmented single character images, since text-row based extraction maintains the color consistency of characters and backgrounds while single character has simpler background. After that, the best binary image is chosen for recognition with OCR feedback. (iii) The K-means algorithm is used for extraction which ensures that the best extraction result is involved, which is the binary image with clear classification of text strokes and background. Finally, extensive experiments and empirical evaluations on several video text images are conducted to demonstrate the satisfying performance of the proposed approach.

Keywords: text extraction, text recognition, character segmentation, *K*-means, OCR feedback.

1 Introduction

According to the official statistic-report of *YouTube* over 6 billion hours of video are watched each month and about 300h of video are uploaded every minute [1]. Thus, it has become a crucial and challenging task to retrieve videos in those large dataset. Video text is one of the most important high-level semantic features for this. Generally, there are two types of video text: the superimposed text (added during the editing process) and the scene text (existing in real scene). Moreover, there are three steps involved before text recognition, i.e., detection, localization and extraction. Here in video text extraction, text pixels remain after the background pixels in the text rows are removed.

*** et al. (Eds.): PCM 2015, LNCS 9309, pp. **-**, 2015.

For video text recognition, the Optical Character Recognition (OCR) is always used to deal with the image pre-processed by text extraction.

Comparing to scene text, the superimposed text offers concise and direct description of the video content. For example, subtitles in sport video highlight the information of scores and players, and captions in movie can be used to summarize the core description of the story [2, 3]. Therefore, in this paper, we mainly discussed the video text extraction combined with recognition for superimposed texts. Actually, both text extraction and recognition problems are handled giving the detected video text image. Compared with previous studies, our main contributions include:

- In extraction, an efficient and accurate character segmentation method is proposed, in which the peak-to-valley eigenvalues is defined to evaluate character width.
- Extraction on row-text images and segmented character images are combined in order to remain both of their advantages. Meanwhile, the best extracted binary image is adaptively chosen for recognition.
- The K-means algorithm with different "K"s is used for text extraction to ensures the best extraction is obtained and chosen by OCR feedback information.

2 Related Work

Recently, video annotation has become a very attractive functionality in video analysis and understanding. i.e., Kim [4] applied manual video tagging to generate text-based metadata for context-based video retrieval. Bhute et al. [5] have given a review of text based approach for indexing and retrieval of image and video. However, manual metadata annotation is both time consuming and incapable of preventing new errors. Meanwhile, although automated speech recognition is applied to provided text scripts of spoken language, poor recording quality and noises affect the achieved performance beyond further usefulness [6]. Therefore, video text recognition based video annotation is still efficient and precise for large scale videos.

As a crucial process before recognition, text extraction needs to separate the pixels of a localized text region into categories of text and background. Video text extraction is always difficult because of complex background, unknown text character color, and various stroke widths [7]. Most existing methods can be classified into:

1) *threshold-based*. Otsu method [8] was the widely used threshold-based text extraction method due to its simplicity and efficiency. Leedham et al. [9] proposed the automatic selection or combination of appropriate algorithms. Besides, Ngo et al. [10] used the adaptive thresholding with four sets of different operations for text extraction. In [11], a novel extraction framework with an improved color-based thresholding is proposed. However, thresholding-based methods did not work well facing complex background, since a strict threshold of text and the background does not always exist.

2) *statistic model-based*. The statistic model-based methods deemed that features of text pixels obey some probability distribution. Gao et al. [12] expressed all the same color regions with a Gaussian kernel function and then to determine the class for each region. Chen et al. [13] used Markov Random Field to determine which Gaussian term each pixel belongs to, and consequently, to segment the text from background. Fu et al.

[14] proposed to extract multilingual texts in images, through discriminating characters from non-characters based on the Gaussian mixture modeling of neighbor characters. Meanwhile, Roy et al. [15] presented a statistical model based scheme for automatic extraction of text components from digital images. Statistic model-based methods can handle complex background well since it took consideration of multi-peak distribution of text color, but they always need establish different models for different images, and also owing to simply using color information, it is hard to determine model functions.

3) *connected components-based.* These methods considered that character strokes are always preserved connectivity but it is not true for background. Lienhart et al. [16] adapted image segmentation, to cluster the foreground pixels with unsupervised color clustering for text extraction. Lyu et al. [7] proposed a synthetic method with local adaptive threshold and connected components analysis. Li et al. [18] proposed a two-threshold method using stroke edge filter, which can effectively identify stroke edges in subjective evaluation. In addition, the authors of [19] proposed a video text extraction scheme using Key Text Points (KTPs). Liu et al. [20] proposed a novel multi-oriented CC based video text extraction algorithm specifically for Chinese text.

However, none of these works have perfectly solved the problem of complex background especially in videos. Meanwhile, while the goal of text extraction is to get a good result for recognition, it is not reasonable to discuss the performance of text extraction separately. The feedback of OCR will be helpful for selecting and evaluating the extraction results. Besides, while most previous works performed text extraction in a whole text row, Sharma et al. [21] proposed to do character segmentation from detected text before binarization recognition, and to achieve better accuracy even for scene text. Meanwhile, our previous work [22] also got a satisfactory performance by separating the text row into individual characters for text extraction.

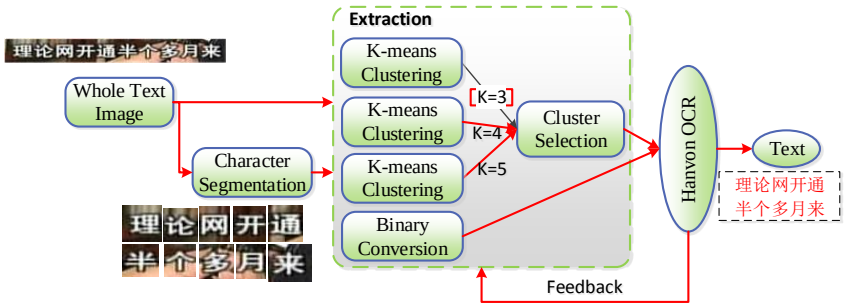

Fig. 1. Framework Diagram of the Proposed Approach

Thus, in this paper, text extraction are applied on both whole text and segmented characters considering their own advantages. Meanwhile, we propose a novel way to optimally configure K since the best K can't be easily obtained in the K-means algorithm. Finally, the extraction result with the best *K* is used to recognize text accurately. The whole framework is shown in Fig.1.

The rest of the paper is organized as follows. In Section 3, Video text segmentation will be discussed. While in Section 4, the text extraction in both segmented character and whole text row will be introduced. Meanwhile, the OCR recognition feedback

based best scheme choosing is illustrated in Section 5. The experimental results are shown in Section 6. Finally, conclusion will be included in Section 7.

3 Video text Segmentation

An assumption is made that the video text rows is obtained with state-of-the-art video text detection and localization method [23]. Characters always have strong spatial frequency variations compared with un-character in a text image. Edge features correspond to spatial frequency which synthesizes vertical and horizontal gradient together, where the vertical gradient is the most useful one. Because most of the video texts are horizontal texts (it can be rotated to horizontal one for the vertical text.), and the vertical gradient can strongly discriminate characters from un-character regions or gaps. Besides, horizontal gradient may also introduce noises to segmentation process. Thus, here, only vertical gradient is considered for segmentation.

The gradient projection is performed on the gradient map to get the projection map, namely gradient projection array. Meanwhile, this array is smoothed by the mean filter. Character Divider Lines (DLs) has to be corresponded to the troughs in the gradient projection array. Thus, in the smoothed gradient array, all troughs form the initial DL set, $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_N\}$. And then, the Peak-to-Valley Eigen values (PVE) is defined for each DL ℓ_i as follows.

$$V(i) = \frac{C(i-1)+C(i+1)-2\times T(i)}{T(i)} \tag{1}$$

where $V(i)$ is the PVE of ℓ_i, $C(i)$ and $T(i)$ refers to the i_{th} crest and trough respectively. In fact DLs who's PVEs are less than the average of all PVEs are removed. The remains are represented by,

$$L = \{l | V(i) \geq \sum_j \frac{V(j)}{N}\} \tag{2}$$

In the end, character DLs are reserved. They divide characters from each other with the following steps:

- DLs are sorted with their PVEs and the biggest M ones are remained. Generally, given that the characters are near-square word with the same height and width, we set $M = 1.5 \times W/H$, where *W* and *H* refer to the width and height of the text image. The reason of using weight of 1.5 is that characters are only near-square, not really square, especially for English characters. Therefore, it make sure that enough DLs are existed, and also it ensure not to abandon any character DLs.
- Then, we aim to estimate the single Character Width (CW). The gap between adjacent *DL*s is named as *DL* interval. Generally, characters always have the same width, so we sorted DL intervals with descending order as $W = \{\omega_1, \omega_2, \ldots, \omega_{M+1}\}$, and then the *CW* is calculated as,

$$CW = \frac{\sum_{\omega \in W`} \omega}{||W`||}, \quad W` = \{\omega_k | |\omega_k - \omega_{k-1} < T, \omega_k - \Omega_{K+1} < T \tag{3}$$

In fact, DL interval must be big enough compared with CW. Here, while a DL interval is less than 0.2CW, the right DL of this interval is removed as noise.

- For Chinese characters with left-right structure, there could be several false DLs located between character components (including two components with 0.5CW or components with 0.3CW and 0.7CW). Specifically, for the first case, if the width of two adjacent DL intervals are both greater than 0.4CW and less than 0.6CW, and also the PVE of the middle DL is smaller than that of both sides' DLs, we merge this two DL intervals. For the second case, if the width of any DL intervals is smaller than 0.3CW, we check that if its left/right DL interval is smaller than 0.9CW (normal width with 0.1CW offset), we merge these two DL intervals. Otherwise, these three DL intervals are merged together.
- We estimate CW again with the above steps corresponding to the new numbers of DLs. The width of DL intervals between these new DLs may be greater than 1.5CW. Therefore, a second segmentation which is similar to the above steps, is conducted in DL intervals whose width are larger than 1.5CW.

4 Text Extraction

Algorithm 1. Best Extraction Scheme Selection Mechanism.

Input: Text image I and clusters number K.
Output: Character binary images B_1 and B_2.
1: Classify pixels into initial classes based on mod operation of pixel index and K.
2: Select the average gray values of RGB channels in each class as initial centers.
3: **while** Not convergence or iteration smaller than maximum **do**
4: Calculate the distance between each pixel and the K cluster centers.
5: Classify pixel into cluster center with the smallest distance.
6: Refresh the cluster centers with average value calculation.
7: **for** Each cluster do
8: Calculate mean square deviation *cd* and the mean *cm* as in [22].
9: Choose the two binary images B_1 and B_2 with the two smallest *cd*s.
10: Output character binary images B_1 and B_2

Now, we get the final character DLs which accurately segment a whole video text row into several single characters, as shown in Fig. 2. We perform the K-means clustering to retrieve several candidate binary images as shown in Algorithm 1. K is set to 4 in our previous work since a text image is composed of *text characters*, *contrast contours around characters*, *background* and *some noise points*. However, it will introduce false extraction results when the noisy points or the contrast contours around characters are not so obvious, for instance, some character points will forcibly be classified into noises or contours. Actually, we find that the binary conversion or clustering with $K=3$ and *5* have more satisfactory performance sometime. Consequently, both $K=5$, $K=4$ and $K=3$ based K-means clustering (as shown in Algorithm 1) as well as the binary conversion (namely, $K=2$) are both applied to each text image.

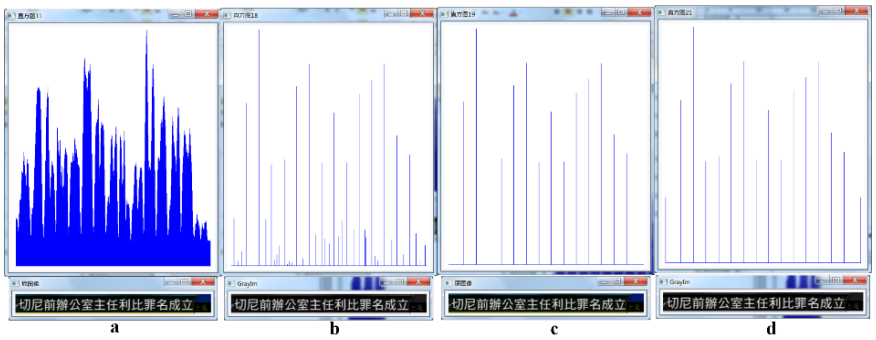

Fig. 2. Segmentation histogram of Text Segmentation. (a) Gradient Projection map, b Initial DLs, (c) First Segmentation Results, and (d) Final Character Divider Lines.

For all the clustering results and binary conversion, dam point labeling and inward filling [22] are used firstly. Then, variable *cd* [22] is used to select the binary images refer to characters for recognition. Considering the clustering bias, binary images corresponding to the two smallest *cds* are chose as candidate images.

So far, for each whole text or segmented single character image, two binary images have been got for a specific K. In fact, there are 13 binary images for recognition, including each 2 binary images for *K=3, K=4, K=5* on both whole text and segmented single characters, and also one for directly binary conversion.

5 Best Extraction Schemes Choosing

While 13 binary images are generated by different *Ks*, the confidence value is also given to each binary image. Then the binary image with the biggest confidence will be the best one for recognition. Specifically, as characters in the same row have nearly the same width, an assumption is made that the character width in each text row follows Gaussian distributions. And the confidence values are evaluated as follows. Let a text as character string $\mathcal{C} \to c_i C$, where c_i denotes a character, and C means a string. As shown in Section 3, we denote the initial DL intervals as $W = \{\omega_1, \omega_2, \ldots, \omega_{M+1}\}$. Then, Gaussian distributions of the character width named C_ω is defined as,

$$C_\omega \sim P_{C_\omega}(x) \sim N(\mu, \sigma^2) \tag{4}$$

$$\mu = \bar{\omega} = \frac{1}{M+1}\sum_{i=1}^{M+1} \omega_i, \ \sigma^2 = \frac{1}{M+1}\sum_{i=1}^{M+1}(\omega_i - \bar{\omega})^2 \tag{5}$$

While a binary text image is input into *Hanvon OCR*, the feedback information: the total recognized characters number N and the dubious characters count D, are used. That is to say, the confidence value S of each binary image is calculated as,

$$S = P_{C_\omega}(N)e^{-D/N} = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(N-\mu)^2}{2\sigma^2} - D/N} \tag{6}$$

The binary image with the biggest confidence S is chose as the best extraction result. Consequently, the corresponded best extraction binary image is input into OCR to successfully recognize the text.

Table 1. The Video Set Used in Our Experiments.

Name	Duration (Seconds)	# Text Row
CCTV TV News	1503	208
Liaoning TV News	538	67
Petroleum TV News	311	46
Hunan TV News	476	70
Phoenix TV News	593	75

6 Experimental Results

To assess the performance of our approach, a large set of videos with different resolutions and characteristics is selected whose details are shown in Table 1.

6.1 Performance of Text Segementation

The text segmentation performance is evaluated with: *Recall*=DT/(DT+LS) and *Precision*=DT/(DT+FA), where DT, LS, FA are the number of correct detects, loss detects and false alarms, respectively. The combination of recall and precision is also evaluated: $RP=Recall \times Precision$, as shown in Table 2. Some segmentation examples can be seen in Fig. 3. Actually, the segmentation performance in video of *Phoenix News* are relatively weaker than the others. Because there are more English words and numbers combined with Chinese characters, causing the inconsistent character width.

Table 2. Text segmentation performance in test videos.

Test Videos	Recall (%)	Precision (%)	PR (%)
CCTV TV News	98.3	100	98.3
Liaoning TV News	97.9	98.1	96
Petroleum TV News	99	100	99
Hunan TV News	98.6	99.5	98.1
Phoenix TV News	95.3	96.4	91.8

6.2 Performance of Text Extraction

In order to evaluate the performance of the proposed approach, our approach is compared with: the Otsu method [8], the Lyu's method [7] and the Li's method [19]. The Otsu method is a simple but very classic solution, while Lyu method is robust to various background complexities and text appearances, and also broadly used for comparison in recent works. Meanwhile, the Li method is one of the most recent approach for text extraction. The Character Error Rates (CER) [7] is adopted to evaluate

the extraction performance by *Hanvon OCR*. For the total of 4849 characters in 466 text rows, the comparisons on CERs of the four methods are shown in Table 3.

From Table 3 and Fig. 4, we can see that our approach get more satisfactory results than the other three methods. The main advantage of our approach is that not any single result generated by different way (using whole text or segmented characters and different value of *K*) is adopted as outputs alone, but we applied all these ways on the text image and chose the best one in term of the OCR feedback.

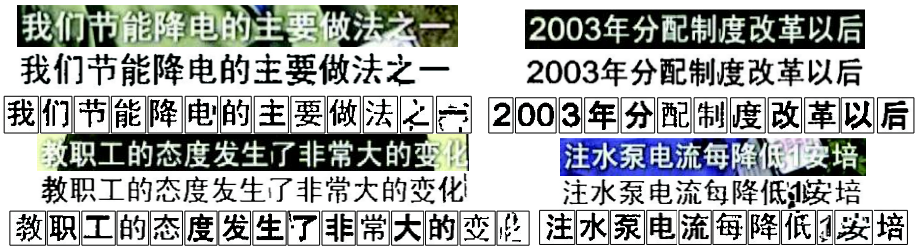

Fig. 3. Text Segmentation Results. The colorful image is the original one, the first binary image refer to the whole row, while the following binary images means the segmented characters.

Table 3. CERs Evaluation of Four Text Extraction Methods.

	Chinese CER	English CER
Our Method	0.122	0.073
Li's Method [19]	0.187	0.151
Lyu's Method [7]	0.201	0.160
Otsu's Method [8]	0.690	0.411

6.3 Recognition Performance with Best Scheme Choosing

Without considering of the recognition performance, the text extraction evaluation is incomplete. Furthermore, the purpose of extraction is to recognize the text. Therefore, to assess the recognition performance of different methods, we also adopt the definition of *Recall* and *Precision*, where *DT* is the number of correctly recognized characters, *LS* is the number of characters loss to be recognized, and *FA* is the number of false alarm in recognized characters by *Hanvon OCR* toolkit.

First, the recognition performance that applying the Hanvon OCR on the extracted texts of different methods, including the proposed method without using OCR feedback (*Without Feedback*), *Otsu method* [8], *Lyu method* [7], *Li method* [19] as well as our approach with OCR feedback (*With Feedback*), is shown in Table 4. Specifically, in *Without Feedback*, K=4 is used directly for clustering, and the single character based extraction is adopted. From Table 4 we can see that without OCR feedback, the recognition performance is similar to those robust works of *Li's method* and *Lyu's method*. Besides, with unpredictable complex background, none of the method can extract perfect text for general text recognition.

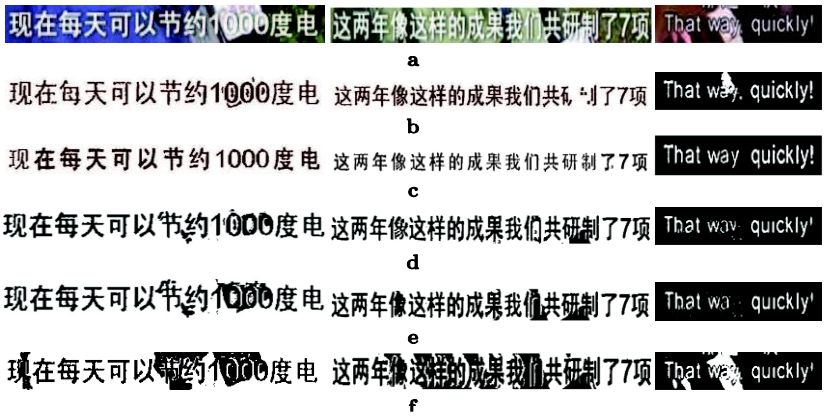

Fig. 4. Comparison of text extraction results: (a) the original text; (b) our whole row results; (c) our single character merged results; (d) Li's results; (e) Lyu's results; (f) Otsu's results.

Table 4. Comparison of Text Recognition Performance.

Methods	Recall (%)	Precision (%)	RP (%)
With Feedback	98.3	85.4	83.9
Without Feedback	88.7	75.0	66.5
Li's Method [19]	89.8	74.1	66.5
Lyu's Method [7]	87.1	75.2	65.5
Otsu's Method [8]	75.3	72.2	54.4

The recognition performance of our approach with OCR feedback is better than most of the previous methods. Our recognition recall even achieved 0.983, which means almost all of the characters are successfully recognized. However, the recognition precision is not so satisfactory, and the reasons include: the frame resolution is low, several noises are introduced in the detection and extraction steps, and also the OCR's recognizing capability is limited.

In addition, the best scheme with the most satisfactory recognition results does not correspond to any fixed configuration of the conditions, including the value of K and the extraction results using whole text or single characters as well as the binary image. Although the clustering based method is proved to be robust for text extraction, the binary image can sometimds generate more accurate recognition results. In other words, the proposed approach that determine the best extraction scheme with consideration of the OCR feedback is reasonable for satisfactory performance.

7 Conclusions

Most previous text extraction methods focus on extraction itself, but not the results of recognition. In order to make video text extraction and recognition more robust and general for different video conditions, we propose a robust text extraction and recognition approach using OCR feedback information. Specifically, we applied the K-means based color clustering on both whole text row, and segmented single characters.

Firstly, we proposed an efficient character segmentation method. Then, we calculated the confidence value with the OCR feedback information. After that, the best extraction scheme is adaptively determined. Finally, we compared our approach with several typical methods, and the result shows that our approach is able to extract and recognize almost all the characters in the test videos efficiently and accurately.

References

1. https://www.youtube.com/yt/press/statistics.html.
2. D. Zhang, and S. Chang. Event detection in basketball video using superimposed caption recognition. In Proc. Of the ACM MM, 2002, pp. 315-318.
3. D. Zhang, R Rajendran, and S. Chang. General and domain-specific techniques for detecting and recognizing superimposed text in video. In Proc. of ICIP, pp. I-593-I-596.
4. H. H. Kim. Toward video semantic search based on a structured folksonomy. Journal of the American Society for Society for Infor. Science and Technology, 2011, 62(3): 478-419.
5. A. N. Bhute, B.B. Meshram. Text Based Approach For Indexing And Retrieval Of Image And Video: A Review. Advances in Vision Computing, 2014, 1(1): 27-38.
6. V. Mitra, H. Franco, M. Graciarena, and D. Vergyri. Medium-duration modulation cepstral feature for robust speech recognition. In Proc. of ICASSP, 2014, pp. 1749-1753.
7. M. R. Lyu, J. Song, and M. Cai. A Comprehensive Method for Multilingual Video text Detection, Localization, and Extraction. IEEE Trans. on CSVT, 2005, 15(2): 243-255.
8. N. Otsu. A threshold selection method from gray-level histograms. IEEE Trans. on Circuits and Systems for Video Technology, 1979, 9(1): 62-66.
9. G. Leedham, C. Yan, K. Takru, J.H.N. Tan, and L. Mian. Comparison of Some Thresholding Algorithms for Text/Background Segmentation in Difficult Document Images. In Proc. of ICDAR, 2003, pp. 859-864.
10. C.W. Ngo and C.K. Chan. Video text detection and segmentation for optical character recognition. Multimedia Systems, 2005, 10(3): 261-272.
11. W. Kim and C. Kim. A New Approach for Overlay Text Detection and Extraction From Complex Video Scene. IEEE Trans. on Image Processing, 2009, 18(2): 401-411.
12. J. Gao and J. Yang. An adaptive algorithm for text detection from natural scenes. In Proc. of CVPR, 2001, pp. II-84-II-89.
13. D. Chen, J.M. Olobez, and H. Bourlard. Text Segmentation and Recognition in Complex Background Based on Markov Random Field. In Proc. of ICPR, 2002, pp. 227-230.
14. H. Fu, X. Liu, Y. Jia, and H. Deng. Gaussian Mixture Modeling of Neighbor Characters for Multilingual Text Extraction in Images. In Proc. of ICIP, 2006, pp. 3321-3324.
15. A. Roy, S.K. Parui, and U. Roy. A Pair-copula Based Scheme for Text Extraction from Digital Images. In Proc. of ICDA, 2013, pp. 892-896.
16. R. Lienhart and A. Wernicke. Localizing and segmenting text in images and videos. IEEE Trans. on Circuits and Systems for Video Technology, 2002, 12(4): 256-268.
17. Y. Song, A. Liu, L. Pang, S. Lin, Y. Zhang, and S. Tang. A Novel Image Text Extraction Method Based on K-means Clustering. In Proc. of ICIS, 2008, pp. 185-190.
18. X. Li, W. Wang, Q. Huang, W. Gao, and L. Qing. A Hybrid Text Segmentation Approach. In Proc. of ICME, 2009, pp. 510-513.
19. Z. Li, G. Liu, X. Qian, D. Guo, and H. Jiang. Effective and efficient video text extraction using key text points. IET Image Processing , 2011, 5(8): 671–683.
20. Y. Liu, Y. Song, Y. Zhang, Q. Meng. A Novel Multi-Oriented Chinese Text Extraction Approach from Videos. In Proc. of ICDAR, 2013, pp. 1355-1359.
21. N. Sharma, P. Shivakumara, U. Pal, M. Blumenstein, and C. L. Tan. A New Gradient based Character Segmentation Method for Video text Recognition. ICDAR, 2011, pp. 126-130.
22. X. Huang, H. Ma, and H. Zhang. A New Video text Extraction Approach. In Proc. of ICME 2009, 2009, pp. 650-653.
23. P. Shivakumara, T. Q. Phan, and C.L. Tan. A Laplacian Approach to Multi-Oriented Text Detection in Video. IEEE Trans. on PRMI, 2011, 33(2): 412-419.
24. X. Huang, H. Ma, H. Yuan. A Novel Video text Detection and Localization Approach. In Proc. of PCM, 2008, pp. 525-534.