A Novel Video Image Text Detection Method

Lin Zhou[1], Xijian Ping[1], Haolin Gao[1] and Sen Xu[2]
[1]Zhengzhou Information Science and Technology Institute
Zhengzhou, Henan 450002 - P. R. China
[2]Yancheng Institute of Technology
Yancheng, Jiangsu, 224000 - P. R. China
[e-mail: {zhoulin8382, brunda}@163.com, pingxijian@yahoo.com.cn, xusen@hrbeu.edu.cn]
*Corresponding author: Lin Zhou

Received November 14, 2011; revised February 6, 2012; accepted March 5, 2012; published April 25, 2012

Abstract

A novel and universal method of video image text detection is proposed. A coarse-to-fine text detection method is implemented. Firstly, the spectral clustering (SC) method is adopted to coarsely detect text regions based on the stationary wavelet transform (SWT). In order to make full use of the information, multi-parameters kernel function which combining the features similarity information and spatial adjacency information is employed in the SC method. Secondly, 28 dimension classifying features are proposed and support vector machine (SVM) is implemented to classify text regions with non-text regions. Experimental results on video images show the encouraging performance of the proposed algorithm and classifying features.

Keywords: Text detection, SC algorithm, multi-parameters kernel function, SVM

This work was supported by the National Natural Science Foundation of China under grant No. 60970142 and No. 60903221, and Talent Introduction Special Foundation of Yancheng Institute of Technology under grant No. XKR2011019.

http://dx.doi.org/10.3837/tiis.2012.04.011

1. Introduction

At present, multimedia information in the Internet increases tremendously, especially for digital video. It is a pressing task to develop effective methods to manage and retrieve these multimedia resources by their content. Text, which carries high-level semantic information, is a kind of important object that is useful for this task. Caption text in news videos usually annotates information on where, when and who of the happening events. Sub-title in sport videos often annotates information of score, athlete and highlight. Compared with other image features, text is embedded into videos by human, which can directly reveal the video content in a certain point of view without requiring complex computation. Therefore, it has inspired a lot of research on text detection and recognition in videos [1]. The goal of text detection is to find image regions containing only text that can be directly highlighted to the user or fed into an optical character recognition (OCR) module for recognition. It is an essential step of text recognition.

Text detection can be applied to many fields such as video coding based on content, robot vision system and video annotation. Although there are plenty of literatures[2][3][4] on video annotation, they rely on visual contents. As texts in video images provide highly condensed and intuitionistic information about the content of video, text detection can well complement the existing video annotation methods. Video images text detection is similar to foreground extraction of the visual monitoring system[5][6][7] because they all extract the foreground from the whole image, the difference is that the foreground in v video images text detection are caption texts while in foreground extraction of the visual monitoring system are person, car and so on. However, due to complex backgrounds or various fonts, colors and sizes, text detection from video is a difficult and challenging task.

Text detection methods can be approximately divided into three kinds: region based [8], edge based [9][10] and texture based [11][12]. Region based methods use the properties of the color or gray-scale in a text region, and then group small components into successive larger components until all regions are identified in a video image [13]. This method can detect text quickly. However, when the background is complex, it may fail. Edge based method utilize abundant and various edge features of text to judge whether a region is text or non-text. As there are usually many strokes in text region, and the color and lightness of text present a striking contrast to the background, so the text region generally possess ample edge feature. This method is simple and effective, however, it is not robust when the size of text is large or the color of text and background is similar. Texture based methods regard text as a special texture. It usually divide the whole image into some blocks and extract texture features of each block, then use neural network (NN) or SVM to classify each block as text or non-text. The problem of texture-based methods is large computational complexity in the texture classification stage and it may confuse when text-like regions appear [14].

Considering the existing problems, this paper hence proposes a novel and universal method of video text detection and obtains satisfying results. The wavelet coefficients energy feature is obtained by decomposing images using SWT [15]. It doesn't need downsampling and the size of output images are the same as the input one. Then SC method [16]][17][18] is adopted for detecting text in video images. Compared to the traditional clustering algorithms such as k-means or single linkage [19], SC has many fundamental advantages. Results obtained by SC often outperform the traditional approaches, and it can be implemented simplely and be solved efficiently by standard linear algebra methods [20]. After getting text candidates,

density-based region growing method [21] is used to connect text pixels into text regions, then 28 dimension classifying features are proposed and SVM is implemented to classify text regions with non-text regions.

The remainder of the paper is organized as follows. In section 2, coarse-to-fine text detection is introduced. Experimental results and analysis are shown in section 3. At last, conclusion is drawn in section 4.

2. Text Detection

Text detection is to locate all kinds of text appearing in the video image. Here, we divide text detection into two steps: coarse detection and fine detection. The first step is to detect text regions precisely as far as possible. And the second step is to reject falsely located ones.

2.1 Text Coarse Detection

SWT shown in **Fig. 1** is adopted for coarsely detecting text, where Q and G are the lowpass and highpass filter respectively. SWT is similar to Discrete Wavelet Transform (DWT) in that the high and low pass filters are applied to decompose the input image at each level. However, the filters in SWT don't need to subsample at each level. SWT gets better approximation than the DWT as it is redundant, linear and shift invariant, and the output images are the same size as the input one. It is greatly useful in representing the image features.

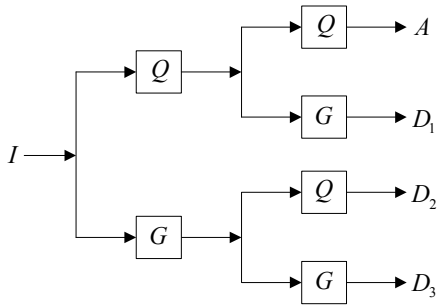


Fig. 1. Stationary wavelet transform

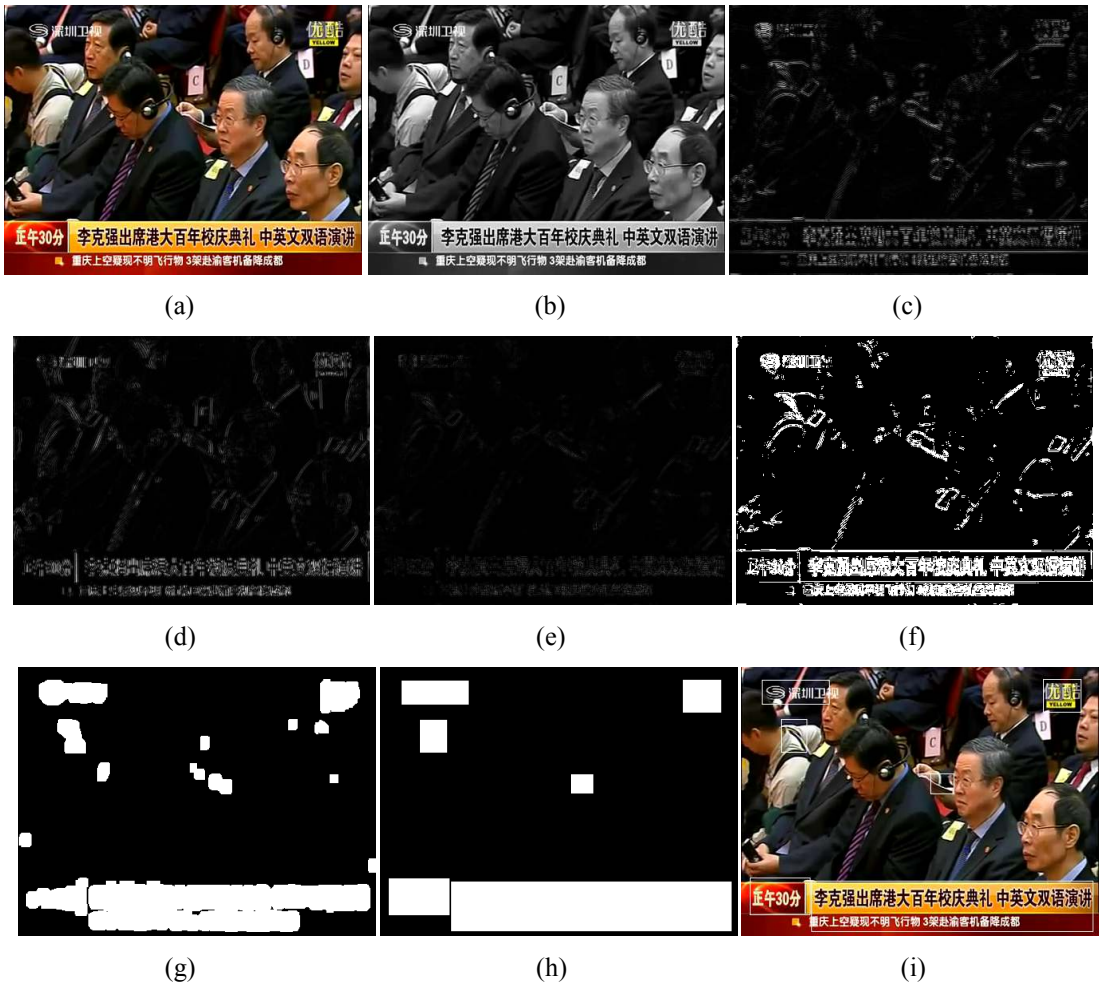
As the color image and gray image of size 448×336 shown in **Fig. 2 (a)-(b)**, three high frequency subband images LH, HL and HH obtained by SWT are shown in **Fig. 2 (c)-(e)** respectively. From these images we can see that large wavelet coefficients near or on the edge are thick white color compared to its background. So we select energy feature to reflect this change, the feature vector at pixel (i, j) is defined as

$$E(i,j)=[|D_1(i,j)|,|D_2(i,j)|,|D_3(i,j)|] \tag{1}$$

where, $D_k(i,j), k=1,2,3$ is the coefficient of pixel (i, j) in LH, HL and HH subband images respectively. As $E(i, j)$ is composed of three detail subbands, it can reflect the change of gray in direction of horizontal, vertical and diagonal.

After getting feature vector, the SC method is applied to classify image pixels into two clusters: text candidates and background. The idea of the SC algorithm originates from spectral graph partitioning theory [22]. It considers clustering as a problem of multi-way

partitioning to undirected graph. Given a partitioning criterion, for example, normalized cut criterion proposed in [23], the SC method optimizes the criterion and makes the points in the same cluster have high similarity while in different clusters have low similarity [24].



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 2. Intermediate steps in text location: (a) Color image, (b) Gray image, (c) Horizontal subband image(LH), (d) Vertical subband image(HL), (e) Diagonal subband image(HH), (f) Candidate text pixels, (g) Density-based region growing image, (h) Visible and (i) Text coarse detection result

As belonging to pairwise grouping methods, the SC method requires to compare all data points for composing the affinity matrix. For instance, a given image of 256×256 pixels has 65536 points, so the size of the generated affinity matrix will be 65536×65536. This requires great cost of computation and storage. The Nyström method [25][26] is a better solution. It randomly chooses a small subset of pixels to do spectral grouping, and then extrapolates to the full set of pixels in the image. This method can substantially reduce the computational complexity in spectral clustering.

In the original SC algorithm, the affinity matrix presenting the similarity information between datas is calculated as

$$S(i,j) = e^{-\frac{\|E(x_i)-E(x_j)\|^2}{2\sigma_v^2}} \tag{2}$$

where $E(x_i)$ and $E(x_j)$ denote feature vector of pixel x_i and x_j respectively, $S(i,j)$ is the similarity of the two samples, σ_v is the scaling parameter which represents the weightiness of features similarity. In this paper, we select multi-parameters kernel function to calculate the affinity matrix

$$\begin{aligned} S(i,j) &= e^{-\frac{\|E(x_i)-E(x_j)\|^2}{2\sigma_v^2}} \times e^{-\frac{\|P(x_i)-P(x_j)\|^2}{2\sigma_c^2}} \\ &= e^{-\frac{\|E(x_i)-E(x_j)\|^2}{2\sigma_v^2}-\frac{\|P(x_i)-P(x_j)\|^2}{2\sigma_c^2}} \end{aligned} \tag{3}$$

where $P(x_i)$ and $P(x_j)$ denote the position of pixel x_i and x_j respectively, and σ_c is the scaling parameter which represents the weightiness of spatial adjacency. As the model combining the features similarity information and spatial adjacency information, it can get better clustering results. The procedure of the multi-parameters SC algorithm can be summarized in as follows:

Input: the total number of pixels in image N, pixel set $X = \{x_1, x_2, \cdots, x_n\}$, feature vector set $\Pi = \{E(x_1), E(x_2), \cdots, E(x_N)\} \in R^{N\times 3}$, the number of clusters K.

(1) Construct the affinity matrix $S \in R^{N\times N}$ defined by

$$S(i,j) = \exp(-\|E(x_i) - E(x_j)\|^2 / 2\sigma_v^2 - \|P(x_i) - P(x_j)\|^2 / 2\sigma_c^2).$$

(2) Apply Nyström method to approximate the eigenvalues and eigenvectors of affinity matrix S.

(3) Select the eigenvectors correspond with the K largest eigenvectors to form matrix $Y \in R^{N\times K}$.

(4) Normalize each row in Y to have unit length.

(5) Set $g_i \in R^K$ as the column vector of Y's row, use k-means or other algorithm to cluster $G = \{g_i \mid i = 1, \cdots, N\}$ in to K clusters $C_1, \cdots, C_K$.

Output: pixel clusters $D_1, \cdots, D_K$, where $D_i = \{x_j \mid g_j \in C_i, x_j \in X\}, 1 \le i \le K$.

The outputs of the multi-parameters SC method are two clusters: text candidates and background. Usually, the number of text pixels is less than background pixels in video images. So a cluster is classified as text if its number of data is less than another cluster. The sample output of the SC algorithm is shown in **Fig. 2-(f)**. After getting candidate pixels, we adopt density-based region growing method to connect text pixels into text regions as shown in **Fig. 2-(g)**. Then the horizontal and vertical boundaries of the text lines are found by projection profile analysis and the bounding boxes are fixed. A located text block is eliminated if its width or height is less than 10 pixels or area is less than 255 pixels. The located text blocks are

filled with the white color to make visible in **Fig. 2-(h)**. Then the detected text blocks in color image are located as shown in **Fig. 2-(i)**.

2.2 Text Fine Detection

From **Fig. 2-(i)** we can see that there are also some falsely detected non-text regions. In order to reduce false alarms, the following features are extracted for classifying.

(1) *Edge image statistical features*: Sobel operator is adopted for calculating the edge image. In video images, text usually dispose horizontally or vertically and the words are mostly composed by horizontal, vertical and diagonal strokes. So the edges in 0^o, 45^o, 90^o and 135^o direction can present the edge characteristic of text commendably. The following 4 features in each direction are calculated:

$$m = \sum_{i=0}^{L-1} z_i h(z_i) \tag{4}$$

$$\sigma^2(z) = \sum_{i=0}^{L-1} (z_i - m)^2 h(z_i) \tag{5}$$

$$U = \sum_{i=0}^{L-1} h^2(z_i) \tag{6}$$

$$e = -\sum_{i=0}^{L-1} h(z_i) \log_2 h(z_i) \tag{7}$$

where z_i and $h(z_i)$ denote gray level and gray histogram of the edge image respectively. There are totally 4×4 features.

(2)*Wavelet moment features*[21]: The intensity variance and spatial distribution of text and non-text are different. The mean and central moment features of wavelet coefficients are adopted to reflect these differences. The mean, second order and third order central moments are calculated as

$$m = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} |W(i,j)| \tag{8}$$

$$u_2 = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (|W(i,j)| - m)^2 \tag{9}$$

$$u_3 = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (|W(i,j)| - m)^3 \tag{10}$$

where M and N are the width and height of image block respectively , and $W(i,j)$ is the SWT coefficient. These 3 features are calculated in three high frequency subbands .There are totally 3×3 features.

(3)*Cross count features*: **Fig. 3** shows the difference between binarized text and non-text images. There are frequent alternations of white and black pixels in binarized text images. However, this phenomenon does not occur in binarized non-text images.

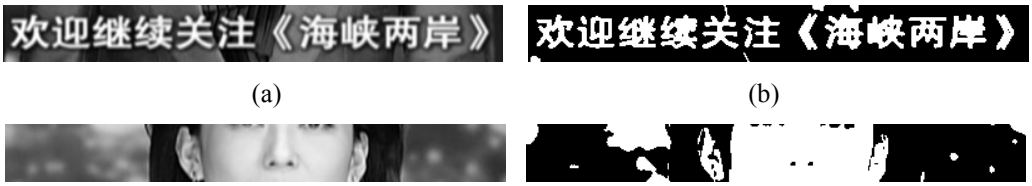


Fig. 3. Text and non-text images and their binarized images

Cross count denotes the alternate frequency of white and black pixels in a certain direction. Horizontal cross count (HCC), vertical cross count (VCC) and diagonal cross count (DCC) are calculated as

$$CCH = \frac{1}{M \times N} \sum_{j=0}^{N-1} \sum_{i=0}^{M-2} \left(p(i,j) \oplus p(i+1,j) \right) \tag{11}$$

$$CCV = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-2} \left(p(i,j) \oplus p(i,j+1) \right) \tag{12}$$

$$CCD = \frac{1}{M \times N} \sum_{i=0}^{M-2} \sum_{j=0}^{N-2} \left(p(i,j) \oplus p(i+1,j+1) \right) \tag{13}$$

where $\oplus$ denotes exclusive OR, and $p(i,j)$ is the pixel value in the binarized image. There are 3 features.

After getting the 28-dimension feature vector, SVM is adopted as the classifier. It is a new universal learning machine based on the statistical learning theory [27]. It works by projecting the data into a higher-dimensional space and finding the optimal linear separator among different classes. SVM shows better generalization performance than traditional techniques, such as neural networks, in pattern classification. In experiments, the Radial Basis Functions (RBFs) is used in SVM and the parameters are obtained by the method of grid-search.

3. Experiments and Analysis

3.1 Experimental Setup

3.1.1 Video Image Databases

As there is no standard dataset available in literature, our own dataset is created. The videos including movies, news and sports are downloaded from several popular webs such as YouKu and CCTV. 400 frame images containing 1023 actual text blocks are extracted to make the test set. 28 dimension features are calculated and SVM is used to classify text blocks with non-text blocks.

3.1.2 Evaluation Criteria

Recall, false alarm and accuracy rate are used to evaluate the performance of the methods. They are calculated respectively as

$$\text{Recall rate} = \frac{\text{Number of truly detected text block}}{\text{Number of manually count actual text block}} \times 100\% \tag{14}$$

$$\text{Falsely alarm rate} = \frac{\text{Number of falsey detected text block}}{\text{Number of (truly detected text block+falsey detected text block)}} \times 100\% \tag{15}$$

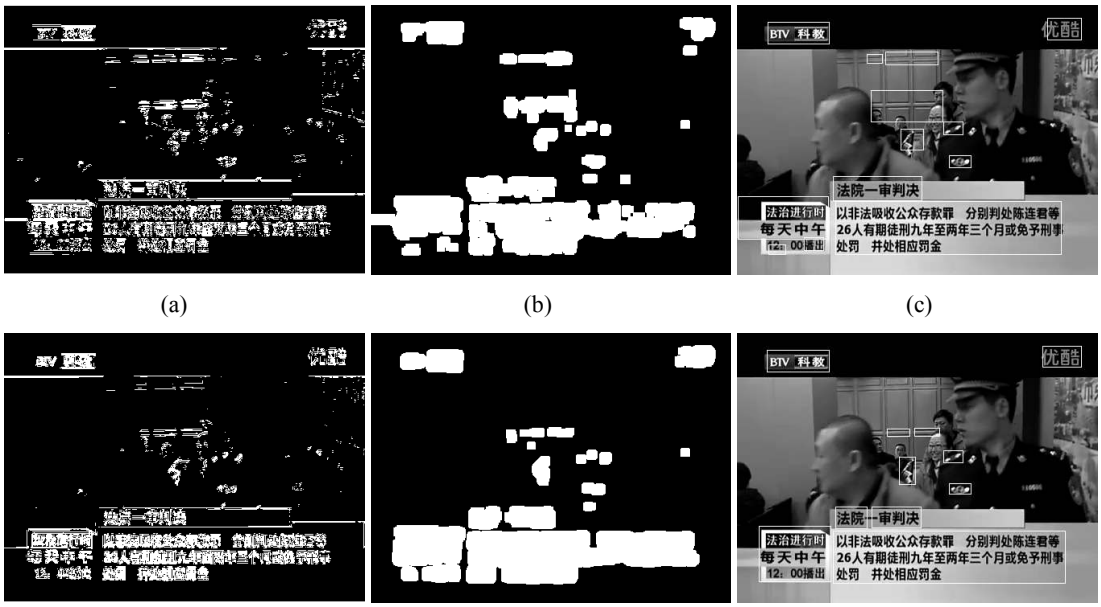
$$\text{Accurate rate} = \frac{\text{Number of (truly detected text block+truly detected non-text block)}}{\text{Number of (text block+non-text block)}} \times 100\% \tag{16}$$

3.2 Coarse Text Detection Experiments

The aim of coarse text detection is to locate text regions and eliminate background regions as far as possible. In this paper, we adopted multi-parameters SC method to fulfill the task.

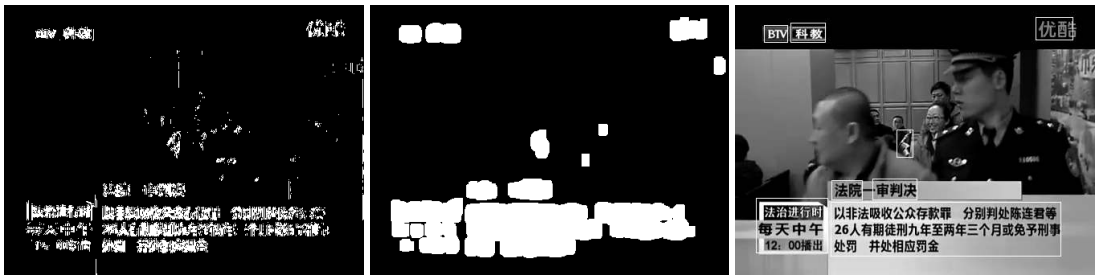
3.2.1 Comparison to Other Clustering Methods

As clustering algorithm is usually employed in coarse text detection, two other clustering methods are implemented for comparison: (1) k-means; (2) the original SC [20]. In the three methods, the same feature vector as formula (1) is adopted. Experiment on an example video image is shown in **Fig. 4**. The image is a news video frame. The candidate text pixels and text regions detected by k-means method and the original SC method are shown in **Fig. 4 (a)-(b)** and **Fig. 4 (d)-(e)**. The text pixels are correctly classified. However, there are numerous background pixels classified falsely. However, from **Fig. 4 (g)-(i)** we can see that there are less number of falsely detected background pixels.



(a) (b) (c)

(d) (e) (f)



(g) (h) (i)

Fig. 4. Coarse text location results of a video image: (a) Candidate text pixels detected by k-means method, (b) Density-based region growing image, (c) Candidate text regions detected by k-means method, (d) Candidate text pixels detected by the original SC method, (e) Density-based region growing image, (f) Candidate text regions detected by the original SC method, (g) Candidate text pixels detected by the multi-parameters SC method, (h) Density-based region growing image and (i) Candidate text regions detected by the multi-parameters SC method

Table 1 shows the performance of the three algorithms in the test dataset. k-means method gets the worst performance. As the multi-parameters kernel function which combining the similarity information features and spatial vicinity is adopted in the proposed method, it gets better performance than the original SC method as shown in **Table 1**.

Table 1. Performance comparison of three clustering methods

Methods	Recall (%)	False Alarm (%)
Text detection based on k-means	95.5	18.3
Text detection based on Original SC	96.8	11.3
Proposed coarse text detection method	97.1	9.2

3.2.2 Comparison to non-clustering methods

In order to validate the performance of the proposed coarse text detection method, two other non-clustering algorithms are also implemented for comparison: (1) adaptive threshold method based on wavelet energy [21]; (2) multi-scale edge-based text detection algorithm [28]. The performance of the three algorithms is shown in **Table 2**. From it we can see that the recall rate of the proposed method is higher than the two clustering method, meanwhile, the false alarm is lower than the other two.

Table 2. Performance comparison of three methods

Methods	Recall (%)	False Alarm (%)
Adaptive threshold method based on wavelet energy [21]	93.6	16.4
Multi-scale edge-based text detection algorithm [28]	94.9	11.7
Proposed coarse text detection method	97.1	9.2

3.3 Text Detection Experiments

In this paper, text detection is divided to two steps. Firstly, the multi-parameters SC method is adopted to coarsely detect text regions based on SWT. Secondly, 28 proposed features and SVM are employed to classify text regions with non-text regions. In order to show the strength of the proposed text detection method, three existing methods [11][28][29] are implemented for comparison. Method [11] is based on wavelet features; method [28] is based on multiscale edge information; method [29] is based on gradient information. The performance of the four algorithms is shown in **Table 3.** It shows that the proposed method obtains the best text detection performance, wavelet features based method [11] and edge information based method [28] get the secondary results and gradient information based method [29] performances worst.

Table 3. Performance comparison of four text detection methods

Methods	Recall (%)	False Alarm (%)
Wavelet features based method [11]	95.4	5.2
Edge information based method [28]	94.6	7.9
Gradient information based method [29]	73.2	12.5
Proposed text detection method	96.2	3.1

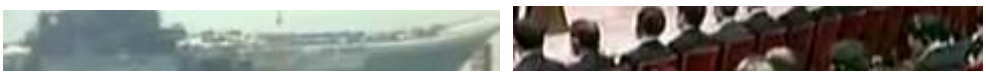
3.4 Comparison to Prior Classifying Features

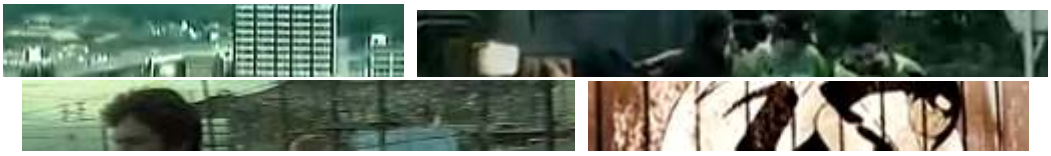
For the purpose of demonstrating the superiority of the proposed features, wavelet coefficients kurtosis features [30] and stroke second-order central moments features [31] are selected for comparison. All these features are measured by SVM. **Fig. 5** shows the text and non-text samples for training. **Table 4** shows performance of these three class classifying features. It is easily noticed that the proposed features obtains the best results.

Table 4. Performance of classifying features

Features	Accuracy (%)
Wavelet coefficients kurtosis features [30]	95.0
Stroke second-order central moments features [31]	92.1
Proposed features	97.9

(a)





(b)

Fig. 5. Train samples: (a) Text blocks and (b) Non-text blocks

4. Conclusion

This paper proposes a novel method of video image text detection. The method consists of two steps: coarse detection and fine detection. In the first step, SWT is adopted to calculate feature vector and the multi-parameters SC algorithm is implemented to locate text regions. Then in the second step, 28 features are proposed and SVM is employed for text regions refinement. It is illustrated that the proposed algorithm and classifying features work well in experiments.

References

[1] Q.X. Ye, Q. M. Huang, W. Gao, D. B. Zhao, "Fast and robust text detection in images and video frames," *Image and Vision Computing*, vol.23, no.6, pp.565-576, Jun.2005.Article (CrossRefLink)

[2] M. Wang, X. S. Hua, R. Hong, J. H. Tang, G. J. Qi and Y. Song, "Unified video annotation via multigraph learning," *IEEE Transaction on Circuits and Systems for Video Technology*, vol.19, no.5, pp.733-746, May.2009. Article(CrossRefLink)

[3] M. Wang, X. S. Hua, J. H. Tang and R. Hong, "Beyond distance measurement: constructing neighborhood similarity for video annotation," *IEEE Transaction on Multimedia*, vol.11, no.3, pp.465-476, Apr. 2009. Article(CrossRefLink)

[4] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proc. of.* the *ACM Multimedia*, pp.660–667, Oct.2004. Article(CrossRefLink)

[5] H. S. Lee, S. J. Hong and E. Kim, "Probabilistic background subtraction in a video-based recognition system," *KSII Transaction on Internet and Information Systems*, vol.5, no.4, pp.782-804, Apr. 2011. Article(CrossRefLink)

[6] N. M. Oliver, B. Rosario and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.22, no.8, pp.831-843, Aug. 2000. Article(CrossRefLink)

[7] J. C. Nascimento and J. S. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Transactions on Multimedia*, vol.8, no.4, pp.761-774, Aug.2006. Article(CrossRefLink)

[8] S.C. Pei and Y.T. Chuang, "Automatic text detection using multi-layer color quantization in complex color images," in *Proc. of the IEEE International Conference on Multimedia and Expo*, pp.619-622, Jun.2004. Article(CrossRefLink)

[9] P. Shivakumara, T. Q. Phan and C. L. Tan, "Video text detection based on filters and edge features," in *Proc. of the IEEE International Conference on Multimedia and Expo*, pp.514-517, Jun.2009. Article(CrossRefLink)

[10] C. M. Liu, C. H. Wang and R. W. Dai, "Text detection in images based on unsupervised classification of edge-based features," in *Proc. of the 8th International Conference on Document Analysis and Recognition*, pp.610-614, Aug.2005. Article(CrossRefLink)

[11] P. Shivakumara, T. Q. Phan and C. L. Tan, "A robust wavelet transform based technique for video text detection," in *Proc. of the 10th International Conference on Document Analysis and Recognition*, pp.1285-1289, Jul. 2009. Article(CrossRefLink)

[12] K. I. Kim, K. Jung and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuous adaptive mean shift algorithm," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, vol.25, no.12, pp.1631-1639, Dec.2003. Article (CrossRefLink)

[13] Z. Ji, J. Wang and Y. T. Su, "Text detection in video frames using hybrid features," in *Proc. of the 8th International Conference on Machine Learning and Cybernetics*, pp.318-342, Jul.2009. Article(CrossRefLink)

[14] Y. Song, A. N. Liu, L Pang, S. X. Lin, Y. D. Zhang and S. Tang, "A novel image text extraction method based on k-means clustering," in *Proc. of the 7th IEEE/ACIS International Conference on Computer and Information Science*, pp.318-342, Jul.2009. Article(CrossRefLink)

[15] R. Guo, Y. H. Peng and H. L Wan, "Palmprint feature extraction and recognition based on stationary wavelet transform," *Computer Engineering and Applications*, vol.42, no.17, pp.62-65, Jul.2006. Article(CrossRefLink)

[16] K. Ersahin, I. G. Cumming and R. K. Ward, "Segmentation and classification of polarimetric SAR data using spectral graph partitioning," *IEEE Transaction on Geoscience and Remote Sensing*, vol.48, no.1, pp.164-174, Jan.2010. Article(CrossRefLink)

[17] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol.32, no.2, pp.335-347, Feb.2010. Article(CrossRefLink)

[18] T. Sakai and A. Imiya, "Randomized algorithm of spectral clustering and image/video segmentation using a minority of pixels", in *Proc. of the 12th IEEE International Conference of Computer Vision Workshops*, pp.468-475, Sep.2009. Article(CrossRefLink)

[19] N. Iam-on, T. Boongoen and S. Garrett, "Refining pairwise similarity matrix for cluster ensemble problem with cluster relations", in *Proc. of the 7th IEEE International Conference on Discovery Science*, vol.5255, pp.222-233, 2008. Article(CrossRefLink)

[20] U. V. Luxburg, "A tutorial on spectral clustering," *Journal of Statistics and Computing*, vol.17, no.4, pp.395-416, Dec.2007. Article(CrossRefLink)

[21] Q. X. Ye and Q. M. Huang, "A new text detection algorithm in images/video frames," *Lecture Notes in Computer Science*, vol.3332, pp.858-865, 2004. Article(CrossRefLink)

[22] L. C. Jiao, X. R. Zhang, B. Hou, S. Wang and F. Liu, "*Intelligent SAR image processing and interpretation*," Science Press, 2008.

[23] J. B. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.22, no.8, pp.888-905, 2000. Article(CrossRefLink)

[24] C. Yang, X. R. Zhang and L. C Jiao, "Self-tuning semi-supervised spectral clustering," in *Proc. of the IEEE International Conference on Computation-al Intelligence and Security*, pp.1-5, Dec.2008. Article(CrossRefLink)

[25] C. Fowlkes, S. Belongie, F. Chung, L. Malik, "Spectral grouping using the Nyström method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 214-225, February, 2004. Article(CrossRefLink)

[26] X. R. Zhang, L. C. Jiao, F. Liu, L. F. Bo and M. G. Gong, "Spectral clustering ensemble applied to SAR image segmentation," *IEEE Transaction on Geoscience and Remote Sensing*, vol.46, no.7, pp. pp.2126-2136, Jul.2008. Article(CrossRefLink)

[27] V. Vladimir, "*The nature of statistical learning theory*," Springer-Verlag, 1995.

[28] X. Q. Liu and J. Samarabandu, "Multiscale edge-based text extraction from complex images," in *Proc. of the IEEE International Conference on Multimedia and Expo*, pp.1721-7124, Jul.2006. Article(CrossRefLink)

[29] E. K. Wong and M. Y. Chen, "A new robust algorithm for video text extraction", *Pattern Recognition*, vol.36, pp.1397-1406, 2003. Article(CrossRefLink)

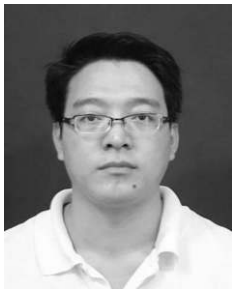
[30] T. X. Zhao, G. M. Sun, C. Zhang and D. M. Chen, "Study on video text processing," in *Proc. of the IEEE International Symposium on Industrial Electronics*, pp.1215-1218, Jun.2008. Article(CrossRefLink)

[31] Z. Wang and Z. Q. Wei, "A comparative study of feature selection for SVM in video text detection," in *Proc. of the 2th International Symposium on Intelligence and Design*, pp.552-556, Dec.2009. Article(CrossRefLink)

Lin Zhou received his M.S. degrees in Signal and Information Processing from Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2005, and is currently pursuing the Ph.D. degree in Zhengzhou Information Science and Technology Institute. His research interests include image processing and pattern recognition.

Xijian Ping received his M.S. degrees in Signal and Information Processing from Beijing University of Aeronautics and Astronautics, Beijing, China, in 1982. He is currently a professor with Department of Information Science, Zhengzhou Information Science and Technology Institute, where he is also the supervisor of Ph.D. candidates from 1998. His research interests include image processing, pattern recognition and information hiding.

Haolin Gao received his M.S. degrees in Signal and Information Processing from Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2005, and is currently pursuing the Ph.D. degree in Zhengzhou Information Science and Technology Institute. His research interests include video processing and pattern recognition.



Sen Xu received his M.S. and Ph.D. degree in Computer Science from Harbin Engineering University, Harbin, China, in 2004 and 2008, respectively. He is currently an associate professor in Yancheng Institute of Technology. His research interests include machine learning and pattern recognition.