International Journal of Automation and Computing

15(4), August 2018, 454-461
DOI: 10.1007/s11633-015-0943-5

MFSR: Maximum Feature Score Region-based Captions Locating in News Video Images

Zhi-Heng Wang Chao Guo Hong-Min Liu Zhan-Qiang Huo

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003, China

Abstract: For news video images, caption recognizing is a useful and important step for content understanding. Caption locating is usually the first step of caption recognizing and this paper proposes a simple but effective caption locating algorithm called maximum feature score region (MFSR) based method, which mainly consists of two stages: In the first stage, up/down boundaries are attained by turning to edge map projection. Then, maximum feature score region is defined and left/right boundaries are achieved by utilizing MFSR. Experiments show that the proposed MFSR based method has superior and robust performance on news video images of different types.

Keywords: News video images, captions recognizing, captions locating, content understanding, maximum feature score region (MFSR).

1 Introduction

Text characters embedded in images carry a great deal of useful and important information, which is considered to be an important aspect of overall image understanding[1]. Generally, text in video images can be divided into two categories[2−5]: 1) Artificial text, which is also named as "superimposed text" and manually added; 2) Scene text, which is also named as "graphics text" and exists naturally in the video images. Artificial text and scene text also exist in news video images. In news videos, artificial text is manually added to the video in a post-processing step in order to convey additional information related to the news video, which can be further classified into caption text and subtitle text. As a brief of news video, caption is of great significance for content understanding or content-based news video retrieval[6−8]. Relative to caption text, subtitle text is less important for video content understanding[9]. Scene text is captured by a camera as part of scenes, such as the advertising boards, address boards of houses, landmarks of streets and so on, which are mostly random and usually do not contain useful content information[2]. Some examples of these three kinds of text in news video images are provided in Fig. 1, in which red boxes denote caption text, black boxes indicate the scene text and yellow boxes represent the subtitle text.

Caption recognizing is critical in news video processing for news content understanding, and locating is the first step for recognizing[10−12]. In recent years, many methods for text locating in images have been studied, most of which depend on text features such as color, edge or texture. Therefore, existing methods can be mainly classified into three categories: color-based, edge-based and texture-based methods.

(a) (b)

Fig. 1 Examples of three kinds of text in news video images

1) Color-based methods, known as connected component based methods, assume that characters in image share the uniform color. Through the clustering method, text is separated from the background as a special single color region. Shim et al.[13] discriminated text regions from the image based on the homogeneity of character pixels. When dealing with an image with the monochrome text, these methods would perform satisfactorily. But if the background is complex or the color of characters is polychrome, the result would be imperfect. On the other hand, because of the video compression, the text in video image often suffers from local color bleeding, which will also affect the results. To overcome the varieties of text color, Yan and Gao[14] proposed a new method. First, they separated the RGB image into different layers exploiting fuzzy c-means clustering algorithm. Then, generated bounding boxes around candidate text regions based on the connected component and eliminated some regions utilizing some heuristic rules such as the aspect ratio or the size of each bounding box,

Research Article
Manuscript received January 15, 2015; accepted March 14, 2015; published online January 11, 2016
Recommended by Associate Editor Victor Becerra

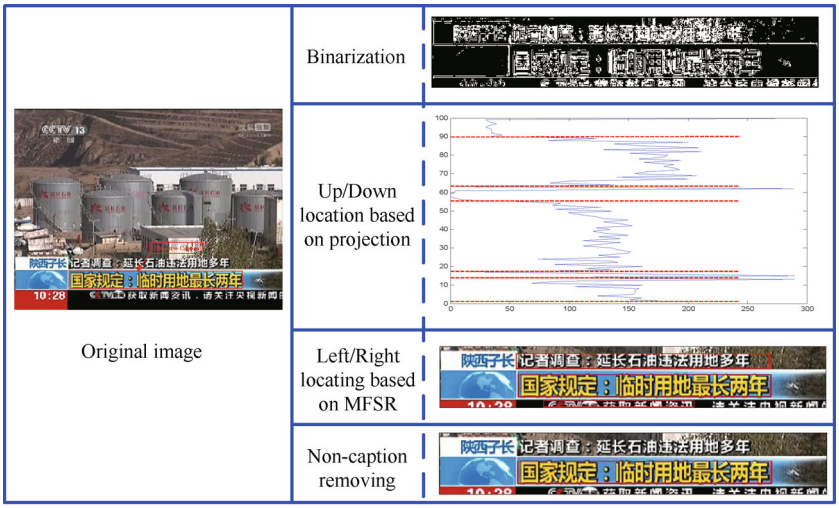


Fig. 2 Flowchart of the proposed MFSR

etc. Last, these candidate regions in each layer are merged. This method is robust to the text color in image, yet it is of high computational complexity.

2) Edge-based methods are also considered to be useful for text locating and have been used by many researchers. Edge-based methods focus on the high contrast between text and background. The usually utilized method is to apply an edge filter (e.g., a Sobel operator) to preprocess image and then determine text regions with high edge intensity and density[15]. Lyu et al.[8] located the text regions using the local threshold on a sobel-based edge map. Chen et al.[16] used the canny operator to preprocess images and obtain the text regions through the edge information. Anthimopoulos et al.[17] also utilized the canny operator to detect edges in an image. Then, based on the edge map, they determined the text regions exploiting the morphological operations and projection analysis. As opposed to other approaches, edge-based methods are simple and have a low time cost. But if the background also has too many strong edges, the experimental results will be not satisfactory.

3) Texture-based methods are always combined with the machine learning and they utilize the observation that text in image has distinct textural properties that distinguish text from the background. First, a method such as Fast Fourier Transform (FFT) or Wavelet Transform is applied to extract textural properties of each local region in an image. Then, these textural properties are fed into a classifier to estimate whether or not the local region contains text[18]. Ye et al.[19] suggested the use of wavelet decomposition to get the textural properties of different local regions followed by an adaptive threshold to distinguish between text and non-text area. Kim et al.[20] preprocessed every pixel using support vector machine (SVM) and then exploited a continuous adaptive mean shift algorithm to ascertain the text regions. Texture-based approaches have better robustness, but they have always high computational complexity since these methods always need a small sliding window to scan the entire image with a typical step of 1 or 2 pixels and determine whether or not the window contains text[21].

Although most existing methods can achieve good results under certain applications, most of them still have shortcomings. Color-based and edge-based methods can locate text regions quickly and have low computational complexity, but they are usually sensitive to the background and have weak robustness[22]. In addition, texture-based methods can reduce the interference of the background and have a strong robustness, but the executing speed is relatively slow and it is quite difficult to design reliable machine learning classifiers to classify small region as text or non-text[21].

Concluding the discussion on previous works of text locating and combining the characteristics of caption text in news video image, a new caption locating algorithm called maximum feature score region (MFSR) based method for news video image is developed in this paper. The remainder of this paper is organized as follows: Section 2 details the proposed MFSR method. Experimental results and analysis are presented in Section 3, and the conclusion is given in Section 4.

2 The proposed method

Before introducing the proposed MFSR, we have observed that captions in news video images of different channels usually are of the following characteristics: 1) Captions are mainly located in the bottom of the news video images (almost in the lower $\frac{1}{4}$ part) and have a certain distance from the image boundaries. 2) Captions are always aligned horizontally and arranged in one or two lines. 3) The color of caption text is uniform and obviously different from the background. 4) The size of caption text is larger than subtitle text. According to 1), it is acceptable that only the lower $\frac{1}{4}$ part of a news video image is processed in our method, which can simplify the problem and reduce the computational complexity.

As shown in Fig. 2, the proposed MFSR method for caption locating mainly consists of four steps: edge detec-

tion, up-down locating based on projection, left-right locating based on MFSR and non-caption removing. Here our method begins with the first edge detection step. Due to the existence of characters, caption regions in news video images usually have rich edge information, and position information of caption regions can be indicated in the edge map. In our method, the up and down positions of the objective caption are estimated based on horizontal projection of the edge map. The Sobel operator is utilized to compute the gradient map of the original image, and then the binary edge map is obtained by using a simple global threshold T, which is determined by computing the mean of the gradient magnitude of the whole region. In the following discussion, the edge pixel is expressed as foreground pixel and the non-edge pixel is expressed as background pixel.

2.1 Up/Down locating based on projection

Fig. 3 illustrates the principle of up/down locating based on projection. The number of edge points of each line is computed and a projection vector V, the dimension of which is equal to the height of the region, is constructed. Then, up/down locations can be estimated by using two thresholds T_1 and T_2, where $T_1 = \delta \times \sum V\ (i)/h$ refers to the minimum value, and T_2 (T_2=10 in our experiments) means the minimum continuous interval. Firstly, the vector V is converted to zero or one by using T_1 (one as text line and zero as non-text line), and then continuous intervals with one value larger than T_2 will be accepted as candidate intervals. In fact, usually more than one candidate intervals (up/down boundaries) $[U_i, D_i], i = 1, 2, \cdots, K$ will be found in this step.

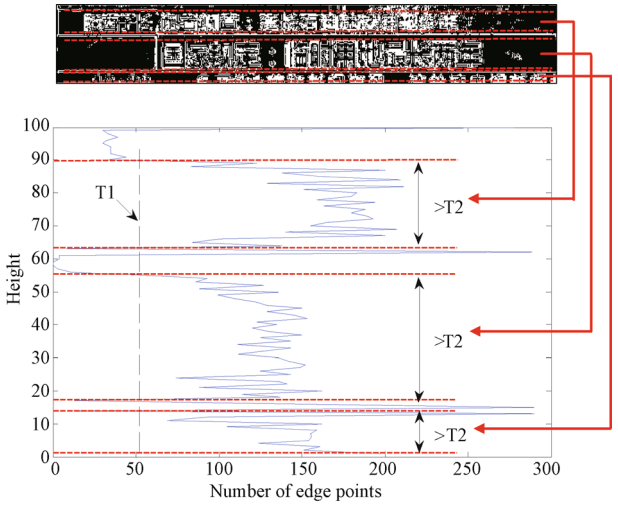


Fig. 3 Up/Down locating based on projection

2.2 Left/Right locating based on MFSR

Considering the region defined by up/down boundaries determined in the previous section, it is obvious that the characters are closely clustered with uniform inter-character distance as in Fig. 3, and determining left/right boundaries of the caption region means finding a rectangular region containing most of the foreground pixels with the smallest size.

How to obtain the smallest rectangular region containing most of the foreground pixels? The problem is converted to searching the maximum value of the following defined feature score, and the attained region is called MFSR:

$$FS(l,r) = \frac{B(l,r) \times P(l,r)}{(r-l+1)} \tag{1}$$

where $B(l,r)$, $P(l,r)$ are called beneficial item and penalty item, respectively, and l, r are variables representing the left and right boundaries of the objective region.

Specifically, beneficial item $B(l,r)$ and penalty item $P(l,r)$ are defined as follows: Our goal is finding a rectangular region containing most of the foreground pixels, and a simple strategy is defining the beneficial item as the total number of foreground pixels. $B(l,r)$ is expressed as follows:

$$B(l,r) = \left(\sum_{x \in G(l,r,U_i,D_i)} e(x) \right)^{\gamma} \tag{2}$$

where $G(l,r,U_i,D_i)$ refers the rectangular region with the boundaries as l,r,U_i,D_i and $e(x)$ is defined as

$$e(x) = \begin{cases} 1, & \text{if } x \text{ is a foreground pixel} \\ 0, & \text{if } x \text{ is a background pixel.} \end{cases} \tag{3}$$

To achieve robustness, a robustifying factor γ is introduced here, which is used to increase the proportion of beneficial item $B(l,r)$. In fact, the value of γ only needs to be greater than 1. In this paper, γ is set as 1.5.

The penalty item is expressed as

$$P(l,r) = P_L(l,r) \times P_R(l,r) \times P_M(l,r) \tag{4}$$

where

$$P_L(l,r) = \alpha - \frac{\sum_{x \in Line(l)} e(x)}{D_i - U_i + 1} \tag{5}$$

$$P_R(l,r) = \alpha - \frac{\sum_{x \in Line(r)} e(x)}{D_i - U_i + 1}. \tag{6}$$

First we discuss $P_L(l,r)$ and $P_R(l,r)$, which are called left penalty item and right penalty item respectively. The two summation terms in (5) and (6) refer to the number of foreground pixels on left and right boundary line, which mean that the left or right boundary should be located as a blank region without foreground pixels. In other words, it will be penalized if the left or right boundary is located on a character. In the above formula, α is a constant factor for ensuring non-negative value and α is set at 1.5 in this paper.

Now we discuss the third penalty item $P_M(l,r)$, which in fact provides constraints on the internal structure of the objective region. This penalty item is introduced mainly based on the following considerations: Characters in objective caption region are nearly uniformly distributed and thus foreground pixels can be considered to be continuous. Therefore, if one or more blank regions without any foreground pixels exist in a test region, it will be rejected as

a false one. In practice, this point can be achieved using the following steps: 1) A small window is defined with size as $(D_i\text{-}U_i\text{-}10) \times \varepsilon$, where ε is the width of the window; 2) The small window slides in the region defined by $[U_i,D_i]$ and the number of foreground pixels of the small window is stored; 3) Positions at which the small window contains no foreground pixels will be marked as blank position, as shown in Fig. 4; 4) The third penalty item $P_M(l,r)$ is expressed as follows (where N_B refers to the number of blank positions occurred in the test region):

$$P_M(l,r) = \begin{cases} 1, & N_B = 0 \\ 0, & N_B \geq 1. \end{cases} \tag{7}$$

Fig. 4 The blank positions

After determining the up/down positions of the objective caption region in the Section 2.1, a candidate region for caption can be easily attained by turning to the maximum feature score defined by the formula (1). For clearly understanding the details of MFSR, its pseudo-code is shown in algorithm 1.

Algorithm 1. MFSR
Require: Up/down boundaries $[U,D]$
Ensure: Region $[U,D,L,R]$
1) Initialization L=1; R=W; maxFS=0
2) for l=1:$W/2$
3) for r=l+10:W
4) compute feature score:
5) $FS(l,r)$=$B(l,r) \cdot P(l,r)/(r\text{-}l\text{+}1)$
6) if $FS(l,r) >$ maxFS
7) maxFS=$FS(l,r)$, L=l, R=r
8) **end if**
9) **end for**
10) **end for**
11) **return** $[U,D,L,R]$

2.3 Non-caption removing and multi-captions recognizing

As shown in Fig. 5, since that possibly more than one group of up/down boundaries can be found in Section 2.1, still multi-candidate caption regions can be outputted based on the MFSR, some of which are false ones as in Fig. 5 (a) while others may be due to multi-captions as in Fig. 5 (b). Therefore, further work should be done for distinguishing them.

(a) (b) (c) (d)

Fig. 5 Non-caption removing and multi-captions recognizing

In news video images, most caption texts usually have larger size than other characters, and it is acceptable that the gradient information of the caption region is richer than other regions. Here the gradient information is measured by the following formula (The two summation terms denote the sum of horizontal and vertical gradient value of each pixel in a candidate caption region, respectively):

$$GI = \sum_{X \in G} d_x(X) \times \sum_{X \in G} d_y(X). \tag{8}$$

The candidate region with the largest value of GI is directly regarded as the caption region. As for the rest candidate regions, they can be easily classified as multi-captions or non-caption by using the following facts: Heights of multi-caption regions are very approximative and left positions of multi-caption regions are not different from each other too much. Finally, multi-caption regions are retained and non-caption regions will be removed as in Figs. 5 (d) and 5 (c).

3 Experiments

To verify the effectiveness of the MFSR, the algorithm has been implemented in Matlab 2012a. 314 news video images with resolution 640×480 from different news channels are selected as the test-bed to evaluate the performance of the proposed algorithm. And 328 caption regions were existing in these test images.

3.1 Parameters selection

The parameters, such as the parameter T_2 to control the minimum height of the caption line, the parameter δ to control the value of T_1 and the parameter ε to control the width of the small window in Section 2.2, significantly affect the performance of the algorithm. Therefore, an important task before using the algorithm is to select appropriate parameters. To this end, we use the over-location rate (OR) and error-location rate (ER) to measure the effects of parameters on the identified results, i.e.,

$$OR = \frac{\text{Number of over} - \text{location caption regions}}{\text{Total number of caption regions}} \tag{9}$$

$$ER = \frac{\text{Number of error} - \text{location caption regions}}{\text{Total number of caption regions}}. \tag{10}$$

(a) (b) (c) (d)

Fig. 6 Over-location caption region and error-location caption region

Over-location caption region means that the caption region is not completely surrounded by the bounding box (Fig. 6 top). Error-location caption region means that the background is surrounded by the bounding box (Fig. 6 bottom). As shown in Fig. 6, the blue dotted bounding boxes

denote the ground truth caption regions and the red bounding boxes denote the experimental result with different parameters (The colorful figure can be seen in electric version). In our experiments, for the parameter T_2, we take T_2=10 to ensure that all caption regions are detected in news video image.

1) The parameters δ and T_1

The parameter T_1 is a benchmark, which is the minimum number of foreground pixels that text line should contain, which indirectly determines the positions of up/down boundaries. And the parameter δ is used to adjust the value of T_1, so determining the value of δ is an important task. If the value of δ is too small, error-location caption regions would appear in the experimental results, i.e., the background would be located as caption region as shown in Fig. 6 (c). If the value of δ is too large, over-location caption regions would appear in the experimental results, i.e., the caption region is not located completely as in Fig. 6 (a). To select the optimal parameter δ, many experiments are conducted with different δ for our 314 testing images with 328 caption regions. As shown in Fig. 7 (a), with the increase of δ the OR will increase after several zero points, whereas the ER will drop to zero. It is clear that the OR curve rises quickly and the ER curve declines slowly after δ=0.4. Moreover, when δ=0.4, the OR is equal to zero. Therefore, δ=0.4 is selected in our experiment.

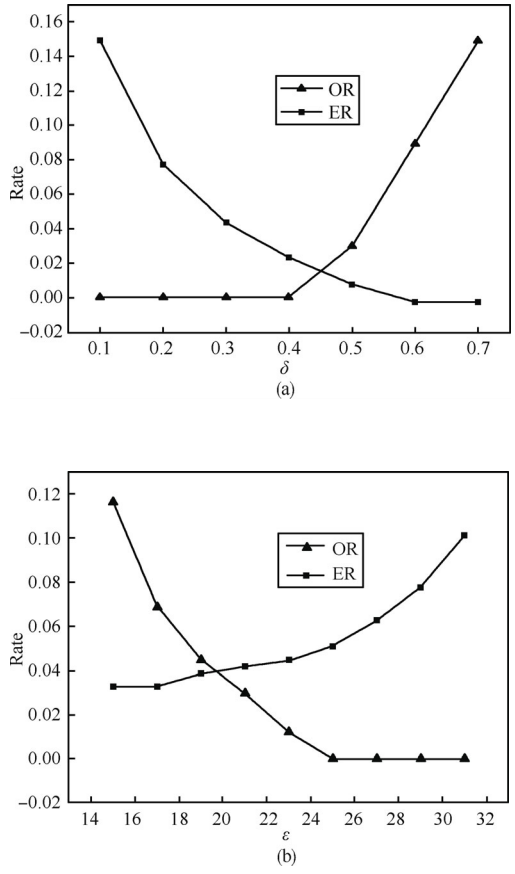


Fig. 7 Parameters determination: (a) OR and ER curves of different δ; (b) OR and ER curves of different ε.

2)The determination of ε

In reality, the caption regions usually have a certain distance from the image boundaries as shown in Fig. 6. To accelerate the execution speed, a small window defined in Section 2.2 is utilized to search the MFSR. The width of the small window is expressed as ε. Whereas punctuation may also occur in the caption region, such as the colon existing in Fig. 6, the inter-character distance will become larger. If the value of ε is too small, over-location caption regions would appear in the experimental results, i.e., the caption region is not located completely as in Fig. 6 (b). If the value of ε is too large, error-location caption regions would appear in the experimental results, i.e., the background would be located as caption region as shown in Fig. 6 (d). As shown in Fig. 7 (b), it is clear that the OR will drop to zero after ε=25 and the ER will rise quickly. To ensure that all caption regions can be detected with low ER, hence ε=25 is selected in our experiment.

3.2 Results and analysis

For conveniently viewing the experimental results, several experimental results of news video images from different channels are listed in Fig. 8, where red boxes denote the results. In addition, for a quantitative evaluation, the method mentioned in [19] is utilized to estimate whether or not an identified caption region is correct, i.e., if the intersection of the identified caption region (ICR) and the ground truth caption region (GCR) covers more than 75% of the ICR and 95% of the GCR, the identified caption region is correct; otherwise, it is incorrect. The GCRs in the test news video images are localized manually. In experiment, the identified results are classified into three kinds: correctly identified caption regions, incorrectly identified caption regions and mistakenly located non-caption regions. In fact, the incorrectly identified caption regions also contain the caption text, which just could not meet with the criteria mentioned in [19]. And the mistakenly located non-caption regions mean that some subtitle regions or background are located as caption regions. As shown in Fig. 9, white dotted bounding box refers to the correctly identified caption region, green dotted bounding box means the mistakenly located non-caption region and yellow dotted bounding box denotes the incorrectly identified caption region.

Here, based on the classification of experimental results, recall rate (RR) and false positive rate (FPR) are utilized to evaluate the algorithm quantitatively, i.e.,

$$RR = \frac{c}{c+m} \times 100\% \tag{11}$$

$$FPR = \frac{f}{c+m+f} \times 100\% \tag{12}$$

where c means the number of correctly identified caption regions, m refers to the number of incorrectly identified caption regions and f denotes the number of mistakenly located non-caption regions. The statistics of the experimental results is given in Table 1.

Fig. 10 shows some experimental results of our method and the method in [23], which is also based on edge detection. The left/right boundaries of caption region are attained by using the vertical projection method of [23], which may result in inaccurate location. However, our algorithm can overcome the shortcoming. As shown in Table 2, the RR of our method is higher than the method in [23], while the FPR is lower. Our proposed method processes a frame in 0.1503 s tested on a PC with two Intel Celeron 2.60 G CPUs. The processing time of our method is longer than the method in [23]. The main cause of the higher time is that the non-caption removing is done in our method. In general, our method is better.

3.3 Experiments on images with different languages

In reality, MFSR not only can be utilized to locate the captions in news video images, it can also be applied to locate the dialogue subtitles in movies or teleplays images. In our experiments, some movies and teleplays images with different languages, such as English, Korean and Japanese are selected from the internet to test the proposed algorithm. For conveniently viewing the results, some testing examples are listed in Fig. 11.

Table 1 The statistics of experimental results

Channel	Image number	GCR number	m	c	f	RR (%)	FPR (%)
CCTV13	108	108	0	108	2	100	1.82
CCTV4	47	47	0	47	0	100	0
CCTV1	23	33	0	33	0	100	0
News comprehensive	23	23	0	23	1	100	4.17
BTV	20	21	0	21	0	100	0
Shanghai TV	25	25	0	25	0	100	0
Other channels	68	71	2	69	5	97.18	6.58
Total	**314**	**328**	**2**	**326**	**8**	**99.39**	**2.38**

Fig. 8 Experimental results

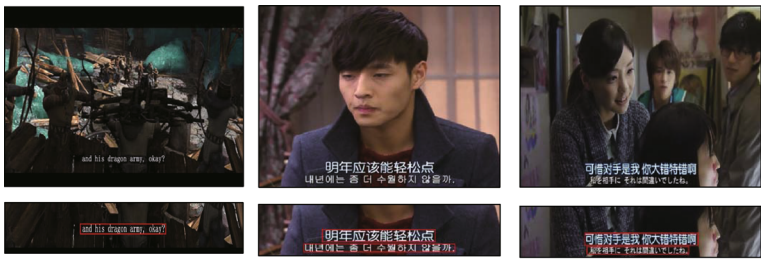
(a) (b)

Fig. 9 Classification of experimental result

Table 2 The statistics of experimental results

Method	RR (%)	FPR (%)	Speed (s/frame)
[23]	98.23	8.64	0.115 3
Our method	99.39	2.38	0.150 3

Fig. 10 The comparative results



(a) Enghish (b) Chinese and Korean (c) Chinese and Japanese

Fig. 11 Examples of different languages

4 Conclusions

In this paper, we present a simple but effective caption locating algorithm, which mainly consists of three modules. First, edge map projection is applied to determine the up/down boundaries of caption region. Then, left/right boundaries of caption region are determined based on maximum feature score region (MFSR). Finally, non-caption regions are removed based on the characteristics of caption text. The experimental results show that the proposed MFSR-based caption locating method could perform satisfactorily on news video images of different news channels. In addition, the proposed method can also be utilized to locate the dialogue subtitles in movies or teleplays images without language restriction.

Acknowledgement

This work was supported by National Natural Science Foundation of China (Nos. 61272394, 61201395 and 61472119), the program for Science & Technology Innovation Talents in Universities of Henan Province (No. 13HASTIT039), Henan Polytechnic University Innovative Research Team (No. T2014-3), and Henan Polytechnic University Fund for Distinguished Young Scholars (No. J2013-2).

References

[1] S. Y. Yan, X. X. Xu, Q. S. Liu. Robust text detection in natural scenes using text geometry and visual appearance. *International Journal of Automation and Computing*, vol. 11, no. 5, pp. 480–488, 2014.

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

[3] P. Shivakumara, T. Q. Phan, C. L. Tan. Video text detection based on filters and edge features. In *Proceedings of 2009 IEEE International Conference on Multimedia and Expo*, IEEE, New York, USA, pp. 514–517, 2009.

[4] Y. C. Wei, C. H. Lin. A robust video text detection approach using SVM. *Expert Systems with Applications*, vol. 39, no. 12, pp. 10832–10840, 2012.

[5] P. Shivakumara, W. H. Huang, T. Q. Phan, C. L. Tan. Accurate video text detection through classification of low andhigh contrast images. *Pattern Recognition*, vol. 43, no. 6, pp. 2165–2185, 2010.

[6] D. T. Chen, J. M. Odobez, H. Bourlard. Text detection andrecognition in images and video frames. *Pattern Recognition*, vol. 37, no. 3, pp. 595–608, 2004.

[7] N. Dimitrova, H. J. Zhang, B. Shahraray, I. Sezan, T.Huang, A. Zakhor. Applications of video-content analysisand retrieval. *IEEE Multimedia*, vol. 9, no. 3, pp. 43–55, 2002.

[8] M. R. Lyu, J. Q. Song, M. Cai. A comprehensive methodfor multilingual video text detection, localization, and extraction. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 243–255, 2005.

[9] D. T. Chen, J. M. Odobez, J. P. Thiran. A localization/verification scheme for finding text in images and video frames based on contrast independent features and machine learning methods. *Signal Processing: Image Communication*, vol. 19, no. 3, pp. 205–217, 2004.

[10] R. Liehart, A. Wernicke. Localizing and segmenting textin images and videos. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 4, pp. 256–268, 2002.

[11] C. Jung, Q. F. Liu, J. Kim. A new approach for text segmentation using a stroke filter. *Signal Processing*, vol. 88, no. 7, pp. 1907–1916, 2008.

[12] M. Cai, J. Q. Song, M. R. Lyu. A new approach for videotext detection. In *Proceedings of 2002 International Conference on Image Processing*, IEEE, Rochester, USA, pp. I–117–I–120, 2002.

[13] J. C. Shim, C. Dorai, R. Bolle. Automatic text extractionfrom video for content-based annotation and retrieval. In *Proceedings of the 14th International Conference on Pattern Recognition*, IEEE, Brisbane, Australia, pp. 618–620, 1998.

[14] J. Q. Yan, X. B. Gao. Detection and recognition of text superimposed in images base on layered method. *Neurocomputing*, vol. 134, pp. 3–14, 2014.

[15] J. Q. Yan, J. Li, X. B. Gao. Chinese text location undercomplex background using Gabor filter and SVM. *Neurocomputing*, vol. 74, no. 17, pp. 2998–3008, 2011.

[16] D. T. Chen, K. Shearer, H. Bourlard. Text enhancement with asymmetric filter for video OCR. In *Proceedings of the 11th International Conference on Image Analysis and-Processing*, IEEE, Palermo, Italy, pp. 192–197, 2001.

[17] M. Anthimopoulos, B. Gatos, I. Pratikakis. Multiresolution text detection in video frames. In *Proceedings of 2007International Conference on Computer Vision Theory and Applications*, Barcelona, Spain, pp. 161–166, 2007.

[18] C. Z. Shi, C. H. Wang, B. H. Xiao, Y. Zhang, S. Gao. Scene text detection using graph model built upon maximally stable extremal regions. *Pattern Recognition Letters*, vol. 34, no. 2, pp. 107–116, 2013.

[19] 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, 2005.

[20] K. I. Kim, K. Jung, J. H. Kim. Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25,no. 12, pp. 1631–1639, 2003.

[21] M. Anthimopoulos, B. Gatos, I. Pratikakis. A two-stage scheme for text detection in video images. *Image and Vision Computing*, vol. 28, no. 9, pp. 1413–1426, 2010.

[22] H. G. Zhang, K. L. Zhao, Y. Z. Song, J. Guo, Text extraction from natural scene image: A survey. *Neurocomputing*, vol. 122, pp. 310–323, 2013.

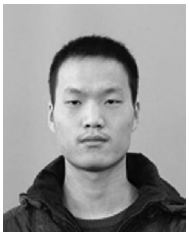
[23] H. Huang, P. Shi, L. W. Yang. A method of caption location and segmentation in news video. In *Proceedings of the 7th International Congress on Image and Signal Processing*, IEEE, Dalian, China, pp. 365–369, 2014.

Zhi-Heng Wang received the B. Sc. degree in mechatronic engineering from Beijing Institute of Technology, China in 2004, and the Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences, China in 2009. Currently, he is an associate professor at School of Computer Science and Technique, Henan Polytechnic University, China.

His research interests include computer vision, pattern recognition, and image processing.

E-mail: wzhenry@eyou.com

ORCID iD: 0000-0002-3241-0720



Chao Guo received the B. Sc. degree from Henan Polytechnic University, China in 2013. Currently, he is a master student at School of Computer Science and Technology, Henan Polytechnic University, China.

His research interests include image processing.

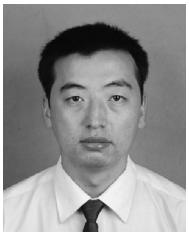
E-mail: xiaofuxing@126.com

Hong-Min Liu received the B. Sc. degree in electrical & information engineering from Xi′dian University, China in 2004, and her Ph. D. degree from the Institute of Electronics, Chinese Academy of Sciences, China in 2009. Currently, she works as an associate professor at School of Computer Science and Technique, Henan Polytechnic University, China.

Her research interests include image processing, especially on feature detection and matching.

E-mail: hongminliu@hpu.edu.cn (Corresponding author)

ORCID iD: 0000-0001-9834-4087



Zhan-Qiang Huo received the B. Sc. degree in mathematics and applied mathematics from the Hebei Normal University of Science & Technology, China in 2003. He received his M. Sc. degree in computer software and theory in 2006 and the Ph. D. degree in circuit and system in 2009 from Yanshan University, China. Currently, he is an associate professor in the college of computer science and technology at Henan Polytechnic University, China. He has published about 20 refereed journal and conference papers.

His research interests include computer software and theory, queuing systems and digital image processing.

E-mail: hzq@hpu.edu.cn