2009 10th International Conference on Document Analysis and Recognition

Using Multiple Frame Integration for the Text Recognition of Video

Jian Yi, Yuxin Peng*, and Jianguo Xiao
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
{yijian, pengyuxin, xjg}@icst.pku.edu.cn

Abstract

This paper proposes a new approach for the multiple frame integration of video, whose novelty mainly lies in three phases: Firstly, in the text-block group (TBG) identification, we identify the blocks with the same text by considering jointly the location, edge distribution and contrast of the text block. Then, in the TBG filtering, to avoid the bad effects of the blurred text on the result of integration, we measure the clarity of the text using the proposed text-intensity map, and select the blocks with the clear text for the integration. Finally, in the TBG integration, we integrate the text blocks by using the average and minimum integrations in the text and background of the image respectively, which can obtain the clean background and clear text with high contrast for the effective text recognition. The experimental results show our method can improve the performance of text recognition in video significantly.

1. Introduction

Nowadays, video has become one of the most popular media types delivered through the internet, broadcast and wireless network, which leads to an increasing need for the video content analysis and retrieval. Among these techniques, text recognition is very useful because it can provide the text information about the content of video, and support the query-by-text video retrieval, which is a convenient and important way preferred by most of users.

Conventional methods [1-2] for the text recognition of video mainly focus on recognizing the text in each single frame independently. To utilize the redundant text information existing in the multiple video frames, some researchers have proposed the methods based on the multiple frame integration (MFI) [4-9], which helps to obtain the cleaner background, higher contrast and clearer text. Generally, there are two key phases in the existing MFI methods: the *text-block group (TBG) identification* and the *TBG integration*, where a text-block group is a set of text blocks with the same text.

In the *TBG identification*, some approaches [6][7] identify the text-block groups by using the image matching techniques, but they have the relatively high computational complexity. To achieve high efficiency, other methods in [9] regard the text blocks at the same location in the continuous frames to be the same text. However, their accuracy is relatively low, because they make errors when blocks with different texts are at the same location in the video. These methods [6][7][9] fail to achieve good performance in the accuracy and efficiency simultaneously.

In the *TBG integration*, the methods in [7][9] obtain the clean background and the clear text by using the average integration. However, these approaches may cause the low contrast in some cases. By using the minimum integration, [4][5][8] can enhance the contrast between the text and background, but they could be easily affected by the noises. These methods [4-5][7-9], generally, can not obtain the clean background and clear text with high contrast by using the average integration or minimum integration independently, which are very useful for improving the result of text recognition in the video.

In addition, another problem of the existing MFI methods [4-9] is that they do not consider the bad effects of the blurred text on the integration result, which is caused by the low-quality video frames. In [4-5][7-8], they deal with each text block equally, including the blocks with the blurred text. In [9], they select the blocks with high contrast for the integration, but the selected blocks can also contain the blurred and unclear text. These blurred texts can make the result of integration unclear and hard to recognize.

Based on the above analysis and considerations, we propose a novel approach for the MFI. Different from the conventional approaches [4-9] which mainly consist of two phases, our method contains three phases: *TBG identification*, *TBG filtering*, and *TBG integration*. The *TBG filtering* is proposed to filter the

* Corresponding author.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.58

blurred text that has bad effects on the integration. The main contributions of our method are as follows:

- *TBG Identification: we identify the blocks with the same text by considering jointly the location, edge distribution, and contrast of the text block, which achieves the high performance in both the accuracy and efficiency.*
- *TBG filtering: to filter the blurred text that can bring bad effects to the result of integration, we measure the clarity of the text based on the proposed text-intensity map and select the blocks with the clear text for the integration.*
- *TBG integration: we integrate the text blocks by using the average and minimum integrations in the text and background of the image, which can get the clean background and clear text with high contrast for the effective text recognition.*

2. Framework of our approach

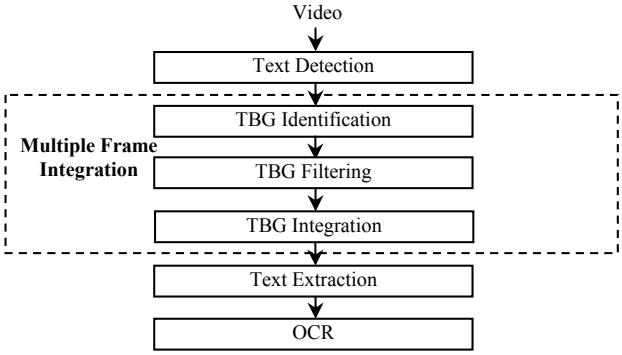

Fig. 1. **Framework of our approach**

Fig. 1 illustrates the framework of our approach for the text recognition of video, which is mainly composed of four parts: text detection, MFI, text extraction and OCR. In this paper, we mainly focus on the part of MFI, which can obtain the cleaner background, higher contrast, and clearer text for the effective text recognition. As shown in Fig. 1, our method for the MFI mainly consists of three phases: Firstly, in the *TBG identification* phase, we consider jointly three simple but effective characters of the text block: location, edge distribution and contrast, and regard the blocks with these three similar characters in the continuous frames to be the same text. In this way, we can avoid the errors in the method [9] makes when blocks with different texts are at the same location, and the computational complexity is much lower than the image matching techniques used in [6-7]. Then, in the *TBG filtering* phase, we propose the text-intensity map to measure the clarity of the text in the image. In general, the clearer text has the higher intensity in the text-intensity map than the blurred text. To avoid the bad effects of the blurred text on the result of integration, which is not considered in the conventional MFI methods [4-9], we filter the blocks with the lower text clarity, and select the blocks with the higher text clarity for the integration. Finally, in the *TBG integration*, we divide the pixels in the text blocks into the background and the text, and utilize the average and minimum integrations in the text and background of the image respectively. In this way, we can obtain the integration result containing the clean background, and clear text with high contrast for the effective text recognition. Details of each phase in our method are described as follows.

3. Multiple frame integration

3.1. TBG identification

As discussed in above, in the *TBG identification*, we consider jointly three simple but effective characters of the text block: location, edge distribution, and contrast. Text blocks in the continuous video frames are indentified to be the same text, only if they have these three similar characters. Firstly, because the same text existing in the multiple video frames generally keeps being at the same location, the blocks with the same text in different frames should be at the same location in the video, and have the large overlap that contains the text. Secondly, the edge map of text block image mainly contains the edges of the text. As a result, if two blocks contain the same text, their edge maps are expected to have the similar edge distributions. Thirdly, the contrast in the text block image is mainly determined by the difference between the text and background, so the blocks with the same text should have the similar contrasts in the image.

Let t_a and t_b denote two text blocks in the continuous video frames, Fig. 2 describes the details of our method. In step 2, $Overlap(t_a,t_b)$ is the overlap of t_a and t_b, and r_1 is the constant between 0 and 1. *SimilarLoc* is set to be true, only if $Overlap(t_a,t_b)$ is larger than the product of r_1 and the minimum area between t_a and t_b. In step 3, E_a and E_b are the edge maps of t_a and t_b, p is a pixel in $Overlap(t_a,t_b)$, and r_2 is the constant between 0 and 1. $NoneZero(E_a,E_b)$ is the set of pixels whose intensities are non-zero in both E_a and E_b, which can be used to measure the consistency of the edge distributions of t_a and t_b. *SimilarEdgeDis* is set to be true, only if $NoneZero(E_a,E_b)$ is larger than the product of r_2 and $Overlap(t_a,t_b)$, which means t_a and t_b have the

consistent edge distributions. In step 4, p is a pixel in $Overlap(t_a,t_b)$, D_{MAX} is a threshold. $EdgeIDiff(t_a,t_b)$ is the sum of the intensity differences of the pixels between E_a and E_b. Since the edge map E_a and E_b can represent the contrasts in t_a and t_b, $EdgeIDiff(t_a,t_b)$ can be used to describe the contrast difference between t_a and t_b. $SimilarCon$ is set to be true, only if $EdgeIDiff(t_a,t_b)$ is smaller than the product of D_{MAX} and $Overlap(t_a,t_b)$, which means the difference between the contrasts of t_a and t_b is small.

1. $SimilarLoc = SimilarEdgeDis = SimilarCon =$ False
2. If $Overlap(t_a,t_b) > r_1 \times Min(area(t_a), area(t_b))$
   $SimilarLoc =$ True
3. $NoneZero(E_a,E_b) = \{p \mid E_a(p) > 0 \,\&\, E_b(p) > 0\}$
   If $NoneZero(E_a,E_b) > r_2 \times Overlap(t_a,t_b)$
   $SimilarEdgeDis =$ True
4. $EdgeIDiff(t_a,t_b) = Sum(|E_a(p) - E_b(p)|)$
   If $EdgeIDiff(t_a,t_b) < D_{MAX} \times Overlap(t_a,t_b)$
   $SimilarCon =$ True
5. If $SimilarLoc$ & $SimilarEdgeDis$ & $SimilarCon$
   t_a and t_b are identified to be the same text
   Else
   t_a and t_b are identified to be the different texts

Fig. 2. **Text-block group identification**

3.2. TBG filtering

To filter the blurred text which brings bad affects to the result of integration, we measure the clarity of the text in image using the text-intensity map, and select the blocks with clear text for the integration. In our method, the text-intensity map is detected by using four text-intensity detectors on the image. Fig. 3 shows these text-intensity detectors, which correspond to the text strokes in the horizontal, vertical, left diagonal and right diagonal directions respectively. In general, the clear text has higher intensity than the blurred text in the text-intensity map, and we can use the intensity of the text in the text-intensity map to measure the clarity of the text in the image. Fig. 4 shows the text-intensity maps of two text block images with the same text. Fig. 4 (b) is the text-intensity map of the image with blurred text, and Fig. 4 (d) is the text-intensity map of the image with clear text. Obviously, the pixels of the clear text in Fig. 4 (d) have higher intensities than the corresponding pixels of the blurred text in Fig. 4 (b). Based on the above defined text-intensity map, our algorithm for the *TBG filtering* mainly consists of four steps. Firstly, we generate the text-intensity map $TIMap_i$ for each text block t_i. Secondly, we divide $TIMap_i$ into the text $TIMap_i^{text}$ and the background $TIMap_i^{back}$. The intensity of $TIMap_i^{text}$ can be used to measure the clarity of the text in image. Thirdly, we use the average intensity of $TIMap_i^{text}$ to be the clarity of the text in the image. And finally, we select the blocks with the highest text clarity for the integration, and filter other blocks.

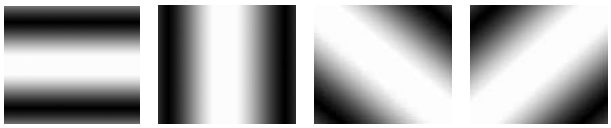

Fig. 3. **(a) Horizontal text-intensity detector. (b) Vertical text-intensity detector. (c) Left diagonal text-intensity detector. (d) Right diagonal text-intensity detector.**

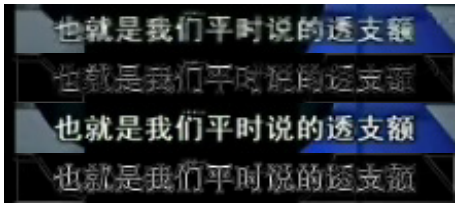

Fig. 4. **(a) Text block image with the blurred text. (b) Text-intensity map of (a). (c) Text block image with the clear text. (d) Text-intensity map of (c).**

1. For each t_i, $(1 \le i \le M)$
   $TIMap_i = Max(TInt_i^H, TInt_i^V, TInt_i^{LD}, TInt_i^{RD})$
2. $t_{AVG} = AVG(t_1,t_2,.....,t_M)$
   For each $TIMap_i$
   $TIMap_i^{text} = \{p \mid t_{AVG}(p) > H_{otsu}\}$
   $TIMap_i^{back} = \{p \mid t_{AVG}(p) \le H_{otsu}\}$
3. For each t_i, $(1 \le i \le M)$
   $TextClarity_i = \sum_{p \in TIMap_i^{text}} TIMap_i(p) / |TIMap_i^{text}|$
4. Select the M' blocks with highest text clarity for the integration, and filter the blocks with the low text clarity

Fig. 5. **Text-block group filtering**

Let $t_1,t_2,...,t_M$ to be M blocks with the same text, our aim is to select M' blocks with the relatively clear text from $t_1,t_2,...,t_M$. Fig. 5 shows the details of our algorithm. In step 1, $TInt_i^H$, $TInt_i^V$, $TInt_i^{LD}$ and $TInt_i^{RD}$ are the text intensities on the horizontal, vertical, left and right diagonal directions, detected by the four text-

intensity detectors shown in Fig. 3. In step 2, t_{AVG} is the average of $t_1, t_2, ..., t_M$, H_{otsu} is the local threshold calculated in t_{AVG} by the OTSU method [10], p is a pixel in $TIMap_i$, and $t_{AVG}(p)$ is the corresponding intensity in t_{AVG}. In t_{AVG}, the pixels with higher intensities than H_{otsu} are regarded as the text, because the text is generally white with the high intensity. In step 3, $TextClarity_i$ denotes the clarity of the text in t_i. The larger value $TextClarity_i$ is, the clearer text t_i has.

3.3. TBG integration

Existing methods in [4-5][7-9] fail to obtain the clean background, and clear text with high contrast, which are extremely useful for improving the result of text recognition in video. In this paper, to complement this deficiency, we divide the text blocks into the text and the background parts by using the local OTSU threshold [10]. Then, in the text part, we integrate the text blocks by utilizing the average integration, which can avoid the bad effects of the noises and get the clear text. In the background part, we integrate the text blocks using the minimum integration, which can enhance the contrast between text and background, and get the clean background.

$$t_{\text{int}}(p) = \begin{cases} \min\{t'_i(p)\}\ 1 \le i \le M' & p \in t_{back} \\ \sum_{1 \le i \le M'} t'_i(p)/M' & p \in t_{text} \end{cases} \quad (1)$$

$$t_{text} = \{p \mid t_{AVG}(p) > H_{otsu}\},\ t_{back} = \{p \mid t_{AVG}(p) \le H_{otsu}\} \quad (2)$$

Let $t'_1, t'_2, ..., t'_{M'}$ to be the M' blocks with clear text selected from $t_1, t_2, ..., t_M$ using the method in section 3.2. Our method for the integration can be described by Eqa (1) and (2), where t_{int} is the result of integration, and p is a pixel in the text block. t_{AVG} is the average of $t_1, t_2, ..., t_M$, and H_{otsu} is the local threshold calculated in t_{AVG} using the OTSU method [10]. The pixels with the higher intensities are regarded as the text, because the text in the video is usually white and has higher intensity. t_{text} and t_{back} are the text and background of the image respectively.

4. Experimental results

To evaluate the performance of the proposed approach, we set up an experimental database consisting of 10 web videos, which are collected from several famous Chinese websites, such as CCTV, Sohu, Xinhuanet, and Chinanews. We use the web videos here since they generally contain the complex background, low contrast and blurred Chinese texts, text recognition of these videos is relatively difficult, and the performances of different methods for the MFI can be well evaluated by using these videos. We manually label the text lines and count the Chinese characters in these videos. In total, there are 1809 different text lines and 11312 Chinese characters in these videos.

As discussed in above, our approach for the video text recognition consists of four parts: text detection, MFI, text extraction and OCR. In the experiments, we focus on the performance comparison of MFI, so we compare different methods for the MFI, and adopt the same approaches for the other three parts. In this way, we can evaluate and compare the performances of the different approaches for the MFI by the results of text recognition. Three metrics are adopted for the evaluation, including *Recall*, *Precision*, and *Repeat*. A high value of *Recall* indicates the superior ability to recognize the relevant characters, while a high value of *Precision* indicates the high recognition rate with correct characters. *Repeat* is utilized because a text may last for a long time, and may be detected and recognized repeatedly. The performances of the approaches for the video text recognition are mainly determined by *Recall* and *Precision* other than *Repeat*, because recognizing the correct characters is far more important than recognizing the characters repeatedly. These metrics are defined as the follows:

$$\text{Re}call = \frac{CN_{correct}}{CN_{truth}},\ \text{Pr}ecision = \frac{CN_{allcorrect}}{CN_{all}},\ Repeat = \frac{CN_{repeat}}{CN_{all}}$$

where $CN_{allcorrect} = CN_{correct} + CN_{repeat}$, $CN_{allcorrect}$ is the number of the correctly recognized characters with the repeated characters, $CN_{correct}$ is the number of the correctly recognized characters without the repeated characters, and CN_{repeat} is the number of the correctly and repeatedly recognized characters. CN_{truth} is the number of the characters in the ground truth, and CN_{all} is the number of the recognized characters.

In our experiments, we employ the text detection and extraction methods in [3], and use the Founder Ruisi OCR software for the recognition, which is applied widely to document recognition field. To show the effectiveness of our approach for the MFI, four methods are implemented for comparison: I. Multiple frame integration in [9]. II. Multiple frame integration with the *TBG identification* proposed in this paper, and *TBG integration* in [9], which employs the average integration on the text blocks with high contrast. III. Multiple frame integration with the *TBG identification* proposed in this paper, and *TBG integration* in [4], which employs the minimum integration on the text blocks with the same text. IV. Multiple frame

integration with the proposed *TBG identification*, *TBG filtering*, and *TBG integration*. Videos in the database are processed by the different text recognition approaches using the above different methods for the MFI, and the same methods for the other three parts.

Table 1. **Experimental results**

	I	II	III	IV
Recall	47.55%	54.82%	53.34%	57.43%
Precision	54.12%	59.50%	59.28%	60.43%
Repeat	6.88%	7.59%	8.36%	8.01%

(a)
(b)
(c)

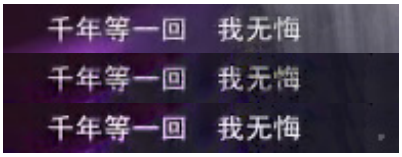

Fig. 6. **(a) Result of average integration used in [9]. (b) Result of minimum integration used in [4]. (c) Result of integration by our method.**

Table 1 shows the results of our experiments. Our three methods (methods II, III and IV) outperform the method in [9] (method I) in terms of *Recall* and *Precision*. This is because our method for the *TBG identification* can identify the text-block groups more accurately than the method in [9]. Comparing methods II, III with IV, method IV achieves the better performance than methods II and III, which has two main reasons: on one hand, in the *TBG filtering*, we measure the clarity of the text based on the text-intensity map, and select the blocks with the clear text for the integration, which can avoid the bad effects of the blurred text and get the clear text; on the other hand, in the *TBG integration*, we utilize the average and minimum integrations in the text and background of the image respectively, which can obtain the clean background and clear text with high contrast for the effective recognition. Fig. 6 shows some results of methods II, III and IV. Fig. 6 (a) and (b) are the results of method II and III respectively, while Fig. 6 (c) is the result of method IV. Comparing Fig. 6 (a) with (c), the contrast in the result of method IV is higher than that in the result of method II, and is better for the text recognition. Comparing Fig. 6 (b) with (c), the text result of method IV is much clearer, and the text recognition result of this image can be better.

5. Conclusion

In this paper, we propose a new approach for the MFI. The novelty of our method mainly lies in three phases. Firstly, in the *TBG identification*, to achieve the good performance in both the accuracy and efficiency, we identify the text-block groups by considering the location, the edge distribution, and the contrast of the text block jointly. Secondly, in the *TBG filtering*, we filter the blurred text that brings bad effects to the result of integration by using the text-intensity map. Finally, in the *TBG integration*, we employ the average and minimum integrations in the text and background of the image respectively, which can obtain clean background, and clear text with high contrast for effective text recognition. Experimental results have shown our method is effective to improve the text recognition performance in the video.

6. Acknowledgements

The work described in this paper was fully supported by the National Natural Science Foundation of China under Grant No. 60873154 and 60503062, the Beijing Natural Science Foundation of China under Grant No. 4082015, and the Program for New Century Excellent Talents in University under Grant No. NCET-06-0009.

7. References

[1] Q. X. Ye, Q. M. Huang, W. Gao and D. B. Zhao, "Fast and Robust Text Detection in Images and Video Frames", *Image and Vision Computing*, vol. 23, pp. 565-576, 2005.
[2] D. T. Chen, J. M. Odobez and H. Bourlard, "Text detection and recognition in images and video frames", *Pattern Recognition*, vol. 37, pp. 595-608, 2004.
[3] M. R. Lyu, J. Song and M. Cai, "A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction", *IEEE Transactions on CSVT*, vol. 15, no. 2, pp. 243-255, 2005.
[4] R. Lienhart and A. Wernicke, "Localizing and Segmenting Text in Images and Videos", *IEEE Transactions on CSVT*, vol. 12, no. 4, pp. 256-268, 2002.
[5] T. Sato, T. Kanade, E. Hughes, M.Smith and S. Satoh, "Video OCR: indexing digital news libraries by recognition of superimposed captions", *Multimedia Systems*, vol. 7, no. 5, pp. 385-385, 1999.
[6] H. P. Li, D. Doermann and O. Kia, "Automatic Text Detection and Tracking in Digital Video", *IEEE Transactions on IP*, vol. 9, no. 1, pp. 147-156, 2000.
[7] H. P. Li and D. Doermann, "Text Enhancement in Digital Video Using Multiple Frame Integration", *ACM Multimedia*, pp. 19-22, 1999.
[8] R. R. Wang, W. J. Jin and L. Wu, "A Novel Video Caption Detection Approach Using Multi-Frame Integration", *International Conference on Pattern Recognition*, 2004.
[9] X. S. Hua, P. Yin and H. J. Zhang, "Efficient Video Text Recognition Using Multiple Frame Integration", *International Conference on Image Processing*, 2002.
[10] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms", *IEEE Transactions on Systems, man and Cybernet*, January, 1979.