TEXT DETECTION, LOCALIZATION AND SEGMENTATION IN COMPRESSED VIDEOS*

Xueming QIAN, *Guizhong LIU*

School of Electronics and Information Engineering, Xi'an Jiaotong University, 710049, China

ABSTRACT

Video text information plays an important role in semantic-based video analysis, indexing and retrieval. Video texts are closely related to the content of a video. Text-based video analysis, browsing and retrieval are usually carried out in the following for steps: video text detection, localization, segmentation and recognition. Videos are commonly stored in compressed formats where MPEG coding techniques are adopted. In this paper, a DCT coefficient based multilingual video text detection and localization scheme for compressed videos is proposed. Candidate text blocks are detected in terms of block texture constraint. An adaptive method for the horizontal and vertical aligned text lines determination is then designed according to the run length of the horizontal and vertical block numbers. The remaining block regions are further verified by local block texture constraints. And the text block region can be localized by virtue of the horizontal and vertical block texture projections. Finally, a foreground and background integrated (FBI) video text segmentation approach is adopted in this paper to eliminate the complex background in text regions. The final experimental results show the effectiveness of our methods.

1. INTRODUCTION

Videos constitute a kind of popular recreation resources. They are usually integrated with audio, image, graph, text and so on. Video texts can provide more intuitive information to video clips with rich semantic information.

In general, there are two types of texts in videos, namely, the scene texts and the artificial texts. The scene texts appear naturally in scenes which are captured by cameras. Artificial texts are mechanically added to video frames to supplement the visual and audio contents [1]. Since artificial text is purposefully added, it is usually more structured and closely related to the subject than a scene text. In some kinds of videos such as the sports, however, artificial texts can be used to semantic based video retrieval, indexing and browsing. As addressed in [2], text-based video retrieval, indexing and abstracting are more reliable than audio and image based methods due to the fact that many existing commercial optical character recognition (OCR) systems are far more robust than the speech analysis techniques and visual object analysis systems.

A video text detection approach presented in [3] is based on background complexities. Video frames are classified into four types according to the edge densities. Edges of sparse non-text regions are gradually removed by repeated shifting and smoothing operators.

Lyu *et al*. [2] proposed a multi-resolution based multilingual video text detection method. Multi-resolution text regions can be detected from the edge maps down-sampled from that of the original images, which are produced by virtue of four directional Sobel operators. The final text detection results are the integration from different channels. Moreover, they also make full use of the temporal redundancy to eliminate the falsely detected text regions.

In [4] and [5] video texts are detected using DCT coefficients and intra-coded type numbers in B and P frames in compressed videos.

Zhong et al. [6] used horizontal DCT texture intensity to detect the candidate text blocks, and then refine those blocks with vertical DCT energy constraint. This method is efficient to detect characters with higher horizontal and vertical texture, such as English text. However, it is not efficient for the Asian characters, which have fertile diagonal texture. And the coarsely detected text blocks are not verified and refined to get accurate text box.

In this paper, video text detection and localization algorithms based on block DCT textures are proposed to checking the selected I frames. Firstly, text block regions can be detected using DCT texture information from selected I frames. Secondly, accurate text boxes are localized from vertical and horizontal block texture projections. Then, a foreground and background integrated (FBI) video text segmentation is used to reduce the influence of complex background. Finally, the extracted text characters are processed by a commercial OCR.

The rest of this paper is organized as follows. In Section 2, our video text detection and localization methods based on DCT texture are proposed. A foreground and

* This work is partially supported by China National Natural Science Foundations (No.60272072, No.60572045), Ministry of Education of China Trans-Century Elitists Project (TCEP, 2002), and Ministry of Education of China 'the Tenth Five Years Plan"-"211" project at Xi'an Jiaotong University, and Microsoft Research Asia.

background integrated text segmentation approach is presented in Section 3. Experimental results and their discussions are given in Section 4. Conclusions are finally drawn in Section 5.

2. DCT TEXTURE BASED TEXT DETECTION AND LOCALIZATION

In order to make our DCT texture based video text detection method both efficient for Chinese and English texts, three horizontal AC coefficients, three vertical AC coefficients and one diagonal AC coefficient of a 8×8 block are selected in this paper to capture the horizontal, vertical and diagonal textures of a block.

2.1. Text Blocks Region Detection

The technique of the 8×8 DCT transform is used in MPEG coding standards. The DCT coefficients AC_{uv} of an 8×8 image block $f(x,y)$ are expressed as.

$$AC_{uv}=\frac{1}{8}C_uC_v\sum_{x=0}^{7}\sum_{y=0}^{7}f(x,y)\cos\frac{(2x+1)\pi u}{16}\cos\frac{(2y+1)\pi v}{16} \quad (1)$$

where u and v denote the horizontal and vertical coordinates $(u,v=0,1,\cdots,7)$ and $C_u,C_v=\begin{cases}\frac{1}{\sqrt{2}} & u,v=0\\ 0 & others\end{cases}$. Let $\{AC_{01}(i,j),\cdots,AC_{03}(i,j),AC_{10}(i,j),\cdots,AC_{30}(i,j),AC_{11}(i,j)\}$ denotes the set of the selected AC coefficients for a block indexed by (i,j). Correspondingly, the DCT texture of block (i,j), denoted by $T_{AC}(i,j)$, is defined by

$$T_{AC}(i,j)=\sum_{u=1}^{3}\left|AC_{u0}(i,j)\right|+\sum_{v=1}^{3}\left|AC_{0v}(i,j)\right|+\left|AC_{11}(i,j)\right| \quad (2)$$

As textures of texts are region-related, a region based block texture filter is used to suppress the textures in the non-text regions. Let MT_{AC} denotes the matrix of the filtered block texture. Then the texture map $MAP(i,j)$ is specified by Equations (3) and (4).

$$MAP(i,j)=\begin{cases}1 & \text{if } MT_{AC}(i,j)>\max\{\alpha\times\overline{MT_{AC}},MT_{Th}\}\\ 0 & \text{else}\end{cases} \quad (3)$$

$$\overline{MT_{AC}}=\frac{1}{M_b\times N_b}\sum_{i=0}^{M_b-1}\sum_{j=0}^{N_b-1}MT_{AC}(i,j) \quad (4)$$

where MT_{Th} is a minimum texture constraint, which is used to diminish false detection in the smooth regions; here we set $MT_{Th}=280$ and $\alpha=1.5$ experimentally. M_b and N_b are the block numbers in the horizontal and vertical directions.

Text block regions and texture rich non-text block regions consisting of the candidate blocks are extracted by the simple texture constraint described above. As artificial video texts are usually aligned horizontally or vertically, an adaptive horizontal and vertical aligned text adjustment is implemented based on the run length of the "white" ($MAP(i,j)=1$) blocks in horizontal and vertical directions. Let $RN\_Hor(i,j)$ and $RN\_Ver(i,j)$ denote the horizontal and vertical run length of block (i,j) respectively. $RN\_Hor(i,j)$ is calculated as the maximum length of horizontally continuous "white" blocks passing (i,j) and $RN\_Ver(i,j)$ is calculated as the maximum length of vertically continuous "white" blocks passing (i,j). By virtue of $RN\_Hor(i,j)$ and $RN\_Ver(i,j)$, the horizontal texture map denoted as $HMAP(i,j)$ and vertical texture map denoted as $VMAP(i,j)$ are defined by.

$$HMAP(i,j)=\begin{cases}\text{if } RN\_Hor(i,j)>RN\_Ver(i,j)\\ 1 \quad \text{and } RN\_Hor(i,j)>RN_{\text{th}}\\ \quad \text{and } MAP(i,j)=1 \text{ and } RN\_Ver(i,j)>1\\ 0 \quad \text{else}\end{cases} \quad (5)$$

$$VMAP(i,j)=\begin{cases}\text{if } RN\_Ver(i,j)>RN\_Hor(i,j)\\ 1 \quad \text{and } RN\_Ver(i,j)>RN_{\text{th}}\\ \quad \text{and } MAP(i,j)=1 \text{ and } RN\_Hor(i,j)>1\\ 0 \quad \text{else}\end{cases} \quad (6)$$

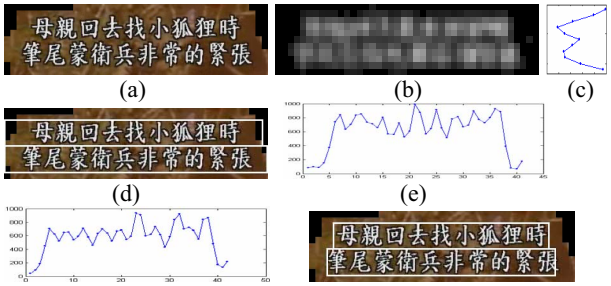
where RN_{th} is the run length threshold, in this paper RN_{th} is set to be 5 by experimentation. Equations (5) and (6) show that block (i,j) belongs to horizontal or vertical aligned text regions it must satisfying region-related characteristic. Thus, the video texts detection involves the horizontal and vertical texts detection. And the final text detection results are an integration of that of horizontal and vertical. A close operator is used to bridge the gaps in horizontal directions. And a dilation operator with a structure element of size 3×3 is implemented to draw some text blocks back, because in some circumstances only a small fraction of the 8×8 blocks are occupied by text, where the texture intensities of the blocks are comparatively low. From above operations, the candidate text block regions in horizontal and vertical directions are detected effectively as shown in Fig.1. In the following sub-section we will refine the text block regions by DCT texture based projections.

2.2. Text Localization

Since the DCT textures of text blocks are higher than those of non-text blocks, we could refine the text block regions by the horizontal and vertical texture projections. Different from the text localization method used in [7], where the accuracy of the text line localization is of pixels, our text line localization method is based on blocks DCT texture projections, which only has detection accuracy of blocks. However, this has little influence to the text segmentation, because our foreground and background integrated text segmentation method is efficient to reduce the influence of complex background. Fig.2 shows an example of text line localization, from which we found that text lines are gradually refined by horizontal and vertical projections.

(a) (b) (c)

Fig.1. Original image (a) and the corresponding horizontal (b), vertical regions (c) detected by DCT coefficients-based method.



(a) (b) (c) (d) (e) (f) (g)

Fig.2. Text line localization. (a) bounding boxes of coarsely detected block regions in pixel domain; (b) texture images of (a); (c) horizontal texture projection profiles of (b); (d) horizontal projection refinement; (e) vertical texture projection profiles of the first text lines; (f) vertical texture projection profiles of the second text lines; (g) final localization results.

3. FOREGROUND AND BACKGROUND INTEGRATED TEXT SEGMENTATION

Different from the previous multi-frame integrated text segmentation methods, which only use the redundant foreground information to reduce the influences of complex background for video text segmentation [2], a foreground and background integrated video text segmentation approach is adopted in this paper.

Let $BKG^l_{\text{app-1}}$, $BKG^l_{\text{dis+1}}$, TXT^l_{app} and TXT^l_{dis} denote the two background regions and the appearing and disappearing text regions of the *l*-th text line respectively. Still backgrounds in the text regions are effectively removed by the difference images, which are expressed as follows

$$DBT^l_{\text{app}} = \left| BKG^l_{\text{app-1}} - TXT^l_{\text{app}} \right| \tag{7}$$

$$DBT^l_{\text{dis}} = \left| BKG^l_{\text{dis+1}} - TXT^l_{\text{dis}} \right| \tag{8}$$

The global thresholds calculated by Otsu method [8], are used to convert the different images DBT^l_{app} and DBT^l_{dis} into the binary ones represented by $BDBT^l_{\text{app}}$ and $BDBT^l_{\text{dis}}$. So a minimum image $MBDBT^l$ can erase most of falsely detected moving background regions, which is expressed as

$$MBDBT^l(x,y) = \begin{cases} 1 & \text{if } BDBT^l_{\text{app}}(x,y) = 1 \\ & \text{and } BDBT^l_{\text{dis}}(x,y) = 1 \\ 0 & \text{others} \end{cases} \tag{9}$$

Generally speaking, $MBDBT^l$ contains ground truth pixels of the text regions. So these pixels can be selected as text seeds. A special filling method is used to find text regions from the minimum binary image, denoted as $MINOtsu^l(x,y)$.

$$MINOtsu^l(x,y) = \begin{cases} 1 & \text{if } BTXT^l_{\text{app}}(x,y) = 1 \\ & \text{and } BTXT^l_{\text{dis}}(x,y) = 1 \\ 0 & \text{others} \end{cases} \tag{10}$$

where $BTXT^l_{\text{app}}(x,y)$ and $BTXT^l_{\text{dis}}(x,y)$ represent respectively the binary images of TXT^l_{app} and TXT^l_{dis} extracted using Otsu method. Starting from the text seeds in $MBDBT^l(x,y)$, their connect regions are found in $MINOtsu^l(x,y)$. If the regions filled from those text seed are reached boundaries of text line, then the filled connected regions are rejected as background.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In order to evaluate the performance of the proposed text detection, localization and tracking and segmentation results, about 320 minutes videos containing Chinese and English texts are used for thresholds selection. Those videos are converted into MPEG-2 coded videos, with different resolutions. Totally, about 100 minutes of videos named *Wild Australasia* (denoted as *wild*) with resolution 320×592and *Foxes of the Kalahari* (denoted as *Foxes*) with resolution 432×576 are used to test the performance of our algorithms.

We use the recall (N_R) and precision (Np) to quantitatively evaluate the performance of our text detection and localization and recognition results [2,3].

4.1. Performance Evaluation for Text Detection and Localization

Fig.3 shows the experimental results of our DCT texture based video text detection and localization. In Fig.3, video texts with different font-sizes, colors and layouts are successfully detected. Both English and Chinese texts embedded in different backgrounds are correctly located.

TABLE I shows the performances of the text detection and localization before and after tracking for the test video sequences. Totally, 6471 I frames in the videos are checked, which contain 3905 text lines. There are 3724 text lines correctly localized with 236 false alarmed ones. The corresponding recall and precision are 95.36% and 94.04% respectively. (where N_F ,N_C and N_G stand for the numbers of falsely, correctly detected and ground truth text lines respectively.)

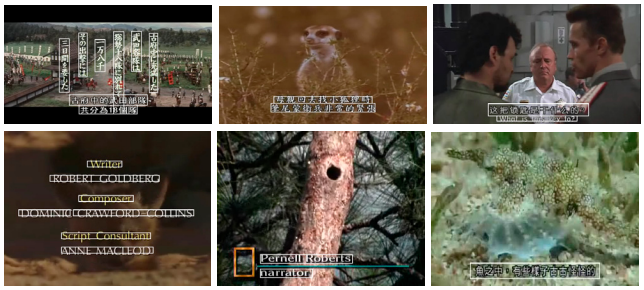
4.2. Performance Evaluation for Text Segmentation and Recognition

We compare our video text segmentation approach with the methods proposed by Ngo *et al.* [3], Lyu *et al.* [2] and Otsu

et al. [8]. Fig.4 shows the segmentation results of 2 text lines by those methods, from which we found that our method robust against complex background. In order to evaluate the text segmentation performances of different algorithms, the extracted characters are input into a commercial OCR package [9] for characters recognition.

We test the text segmentation methods with two datasets; one is English characters consisting of 64 text images, and the other is Chinese characters containing 158 text lines. Totally, there are 1346 Chinese characters and 1484 English characters in the tested text images.

The performances of English characters and Chinese characters for different text extraction approaches are listed in TABLE II and TABLE III. It is found that our approach achieves the best recall and precision for both English and Chinese test characters, compared with the Ngo, Lyu and Otsu methods.



(a) (b) (c)

Fig.3.Video text detection and localization results.

(a) (b)

Fig.4. Comparison of four video text segmentation methods

TABLE I. Performance of Text Detection and Localization

Test video	Checked I Frame	N_G	Performance			
			N_C	N_F	N_R(%)	N_P(%)
Foxes	3406	2652	2583	129	97.40	95.24
Wild	3065	1253	1141	107	91.06	91.43
Total	6471	3905	3724	236	95.36	94.04

TABLE II. OCR Results For English Characters

method	Correct	Missed	False	Recall	Precision
Original	1268	216	210	85.44	85.79
Otsu	1168	316	225	78.71	83.83
Ngo	1175	309	230	79.18	83.63
Lyu	1209	275	203	81.47	85.62
Ours	1395	89	83	94.00	94.38

TABLE III. OCR Results For Chinese Characters

method	Correct	Missed	False	Recall	Precision
Original	1244	102	108	92.42	92.08
Otus	1097	249	171	81.50	86.79
Ngo	1036	310	288	76.97	78.25
Lyu	1186	160	211	88.11	84.90
Ours	1306	40	46	97.03	96.60

5. CONCLUSION

This paper proposes multilingual video text detection and localization methods based on DCT texture of compressed videos. Seven DCT coefficients of a 8×8 block are selected to represent the texture of the block, which can capture horizontal, vertical and diagonal texture information which can be used for multilingual video text detection including Chinese and English. Text detection is separated into horizontally and vertically aligned text detection by virtue of the run length of the candidate text blocks. A block DCT texture projection based text localization method is proposed to get more accurate box of each text line. A foreground and background integrated video text segmentation method is proposed which can effectively reduce the influence of complex background.

6. REFERENCES

[1] H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 147–156, Jan. 2000.

[2] M.R. Lyu,, J.-Q. Song , M. Cai , "A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction ," *IEEE Trans. Circuits and Systems for Video Technology*, vol.15, No.2, pp.243-255, Feb. 2005.

[3] C.-W. Ngo, C.-K. Chan, "Video text detection and segmentation for optical character recognition," *Multimedia Systems*, vol.10, No.3,pp.261-272,Mar,2005.

[4] U. Gargi, S. Antani, and R. Kasturi, "Indexing text events in digital video databases," in Proc. 14th Int. Conf. Pattern Recognit., vol. 1, 1998, pp.916-918.

[5] Y.-K. Lim, S.-H. Choi, and S.-W. Lee, "Text extraction in MPEG compressed video for content-based indexing, " *in Proc. Int. Conf. on Pattern Recognit.*, vol. 4, 2000, pp. 409–412.

[6] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic Caption Localization in Compressed Video," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.22, No.4, pp.385-392, Apl. 2000.

[7] Rainer Lienhart, and Axel Wernicke, "Localizing and Segmenting Text in Images and Videos," *IEEE Trans. Circuits and Systems for Video Technology*, vol.12, No.4, pp.256-267, Apr. 2002.

[8] N. Otsu, "A threshold selection method from gray-level histograms, "*IEEE Trans. Syst., Man, Cybernet.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[9] Available: http://www.cnau.net/SoftView/SoftView_302.html.