Text Processing in Video Frames with Complex Background

Shuicai Shi [1, 2], Tao Cheng [1, 2], Shibin Xiao [1, 2], Xueqiang Lv[1, 2]

[1] Chinese Information Processing Research Center, Beijing Information Science and Technology University,100101 Beijing, China
[2] Beijing TRS Information Technology Co., Ltd, 100101 Beijing, China

cheng.tao@trs.com.cn; shi.shuicai@trs.com.cn, xiao.shibin@trs.com.cn, lv.xueqiang@trs.com.cn

ABSTRACT: Information deficiency is a huge problem when researching on video indexing and retrieval. On the other hand, text in video frames implies lots of semantics inherently, and can provide supplemental but important information for video data processing. In this paper, we present a fast and robust approach for text detection, localization, extraction, and reorganization in video frames with complex background. Here, block change rate (BCR for short) is imported to realize text detection and localization, smoothness model is used to narrow the scope of the text stroke, element image division in Lab color space is implied in binary text extraction, and Langue model is imported to evaluate the text extraction results. Experiments based on a large amount of video frames from different sources show that this approach is robust, effective and compatible for variety videos with complex background.

KEYWORDS: block change rate; smoothness model; element image division; langue model

I. INTRODUCTION

Information in video data is mainly concealed in a sequence of video frames, and the content of video frames can be divided into two main categories: perceptual content and semantic content [1]. Perceptual content includes attributes such as color, intensity, shape, texture, and their temporal changes, whereas semantic content means objects, events, and their relations.

Text in video frames is form of semantic content. It is very useful for describing the contents of video data, it can be easily extracted compared to other semantic contents, and it enables applications such as keyword-based video retrieval, automatic video logging, and text-based image indexing.

Text processing in video frames is divided into the following sub-problems: (I) detection, (II) localization, (III) tracking, (IV) extraction, and (V) recognition (OCR). The architecture of this processing is shown in Fig. 1.

Text detection is to determine the presence of text in given frames. Text localization is the process of determining the location of text in the given frame. We use BCR to realize text detection and text localization at the same time. Text tracking is to determine the life cycle of text area. Text extraction is to distinguish between background pixel and foreground pixel, and then output a binary image that can be fed into optical character recognition (OCR) engine to recognize. We use element image division in Lab color space to resolve the sub-problem of text extraction in this paper. In the text recognition (OCR) module, we apply langue model to justify the result of text extraction module.

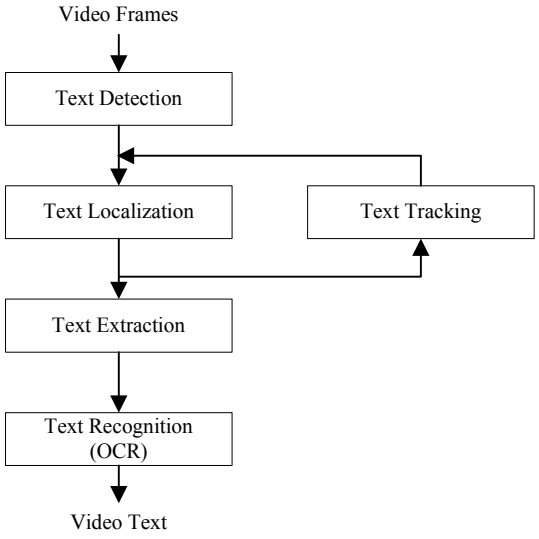

Figure 1. Architecture of text Processing.

Although text processing includes 5 sub-problems, the main researches are focused on the text localization module and text extraction module.

The methods for text localization can be concluded to the followings: using edge feature [2-6], using wavelet energy feature [7], using neural network [8], using DCT coefficient [9], and others.

Dozens of text extraction algorithms are presented in the past years such as SVM (Support Vector Machine) based method [3, 4, 7], local thresholds [2, 5], clustering the image pixels by color [10], using Hough transform technique [11], using kurtosis feature [8], via corners-set matching scheme [12], and others.

978-0-7695-3600-2/09 $25.00 © 2009 IEEE
DOI 10.1109/IFITA.2009.128

In order to demonstrate thinking of this paper clearly, we list the symbols which will appear in the following chapters firstly.

TABLE I. CONSTANT SYMBOL DEFINITIONS

Symbol description	Symbol name
Width of video frame	$\$width$
Height of video frame	$\$height$
Width of text area	$\$textWidth$
Height of text area	$\$textHeight$
Time of video	$\$time$
Number of blocks in line	$\$blockNum = 30$
Size of block	$\$blockSize = \frac{\$width}{\$blockNum}$

TABLE II. FUNCTION SYMBOL DEFINITIONS

Symbol description	Symbol definition
RGB pixel in frame	$Rgb(x, y) = (r, g, b)$
Lab pixel in frame	$Lab(x, y) = (l, a, b)$
Neighbors of a pixel (distance lower than d)	$Neighbor(x_0, y_0, d) = \{(x, y) \mid (x - x_0)^2 + (y - y_0)^2 \le d^2\}$
Frame image	$F(t) = \mid Rgb(x, y) \mid_{\$width \times \$height}$
One line in frame	$L(t, h) = \{Rgb(x, y)\}_{y=h}$
Location of text area	$Loc = (x_1, x_2, y_1, y_2)$
Content of a text area	$T(Loc, t) = (Loc, F(t))$
Difference of 2 pixels in RGB color space	$DifRGB(Rgb1, Rgb2) = (Rgb1.r - Rgb2.r)^2 + (Rgb1.g - Rgb2.g)^2 + (Rgb1.b - Rgb2.b)^2$
Difference of 2 pixels in Lab color space	$DifLab(Lab1, Lab2) = \sqrt{(Lab1.l - Lab2.l)^2 + (Lab1.a - Lab2.a)^2 + (Lab1.b - Lab2.b)^2}$

Value ranges of some parameters in TABLE II are:

$$\begin{aligned} &0 \le x < \$width \\ &0 \le y, h < \$height \\ &0 \le t < \$time \end{aligned} \tag{1}$$

II. TEXT DETECTION AND LOCALIZATION

This chapter describes our method of text detection and localization. For one frame in given time, we expect to get information of the existence of text in this frame, and if texts exist, we also expect to obtain the locations of these texts.

Through analyzing the character of text area and none text area, we find that texture of text area is sharper, and that of none text area is smoother. In other words, color value in text area change more violent and color value in none text area change more tender.

We import the concept BCR in this chapter. For each line $L(t,h)$ in frame $F(t)$, we divide $L(t,h)$ into $\$blockSize$ (30 in our experiment) blocks. Then each block can be expressed as fellows:

$$B(t,i,h) = \{RGB(x,h)\}_{i*\$blockSize \le x < (i+1)*\$blockSize} \tag{2}$$

In (2), symbol i ranges from 0 to $\$blockSize - 1$. For each block $B(t,i,h)$, the following is the formula to compute the corresponding BCR:

$$BCR(t,i,h) = \sum_{j=i*\$blockSize}^{(i+1)*\$blockSize} (DifRGB(Rgb(j,h), Rgb(j+1,h)) + DifRGB(Rgb(j,h), Rgb(j+1,h))/3) \tag{3}$$

Equation (3) defines the change rate of a block considered both horizontal change and vertical change. However, the value of BCR computed by (3) will change with the value $\$blockSize$. So it muse be normalized to a certain range.

$$BCR(t,i,h) = \frac{BCR(t,i,h)}{\alpha * \$blockSize} \tag{4}$$

In (4), symbol α is an experiment parameter. Huge amounts of experiments to validate that 3200 is the suitable value for parameter α.

The effects of BCR method can be validated by create "BCR Picture". "BCR Picture" is consist of binary lines whose existence is depended on $Signal(BCR)$.

$$Signal(BCR) = \begin{cases} 1 & if\ (BCR(t,i,h) \ge 1) \\ 0 & other \end{cases} \tag{5}$$

In "BCR Picture", if $Signal(BCR)$ equals 1, we draw the block $B(t,i,h)$ as black color (a black line in fact), else we draw it as white color (nothing in the picture). We can see clearly that if a text exists somewhere in a video frame, most of $Signal(BCR)$ in corresponding location will equal 1, and correspond black lines in "BCR Picture".

Fig. 2 and Fig. 3 are 2 examples of video frames and their corresponding "BCR Picture". Fig. 4 and Fig. 5 are the localization results of Fig. 2(I) and Fig. 3(II).

(I) (II)

Figure 2. Video frame1 (I) its BCR picture (II)

(I) (II)

Figure 3. Video frame2 (I) its BCR picture (II)

Figure 4. Localization results for Fig. 2(I)

Figure 5. Localization results for Fig. 3(I)

III. TEXT EXTRACTION

In this chapter, we use Lab color space to compute the difference of 2 pixels. Lab color space is more perceptually uniform than RGB. Perceptually uniform means that a change of the same amount in a color value should produce a change of about the same visual importance. Lab color space is a color-opponent space with dimension L for lightness and a&b for the color-opponent dimension. In our experiments, the "L" of Lab ranges from 0 to 10, the "a" of Lab ranges from 0 to 18, and the "b" of Lab ranges from 0 to 19. Therefore, there are 11*19*20=4180 different colors in our experiments.

In this chapter, we will take the first text area in Fig. 4 (that is "ayumi hamasaki") for example to introduce our approach. The output of text extraction is a sequence of binary images that can be directly processed by OCR software.

A. *Prepare work for text extraction*

When finishing locating the text area, we track it to obtain its life cycle. Suppose that at time $t1$, we detected a text and located it at the location Loc. For each frame at time ti, we compute the difference between $T(Loc,t1)$ and $T(Loc,ti)$. Suppose that at time tn, the difference is beyond some definite threshold, then the life cycle of the text if from time $t1$ to tn.

Then we superpose the content of text area in Loc from time $t1$ to tn. Experiments have validated that this operation can enhance the stroke of the text and weaken the background effectively. Fig. 6(I) is the text image at time $t1$, and Fig. 6(II) is the text image superposed the content of text area from time $t1$ to tn. We can clearly see that the background of Fig. 6(II) is smoother than that of Fig. 6(I).

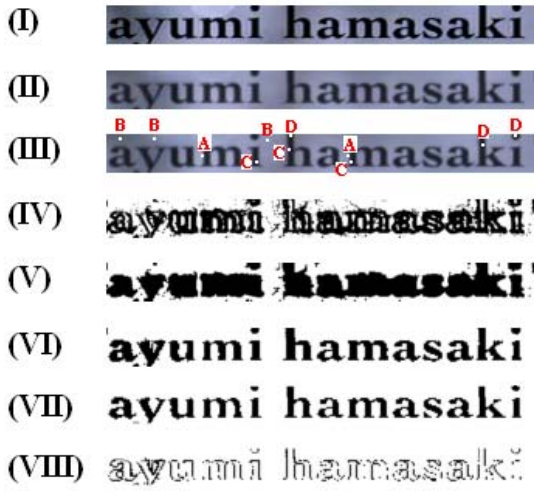

Figure 6. Progress of text extraction

B. *Smoothness model*

The aim of using smoothness model is to narrow the scope of the text stroke.

First, we conclude the pixels in Fig. 6(III) into 4 kinds.

"A" kind pixels are background pixels whose locations are close to the edges of text stroke, so the smoothness of them is higher relatively.

TABLE III. 4 KINDS OF PIXELS

Kind	Background or Foreground	Smoothness
A	Background	high
B	Background	low
C	Foreground	high
D	Foreground	low

"B" kind pixels' locations are far from the edges of text stroke and the smoothness of them are lower.

"C" kind pixels are foreground pixels. They are close to the edge of text stroke and on the text stroke. Therefore, the smoothness of them is higher.

"D" kind pixels are located in the center of the text stroke, and the number of these pixels is limited.

In most text areas, the number of these pixels is more than that of others. Then the aim of using smoothness model is to select these pixels, and mark them as background. In order to realize the aim, we import the concept of $Smooth(x, y)$ to quantify the smoothness.

$$Smooth(x,y) = \sum_{i=0}^{n} DifLab(Lab(x,y), Lab(x_i, y_i)) \quad (6)$$

In (6), symbol n is the number of elements in set $Neighbor(x, y, d)$, symbol i is range from 0 to n, and (x_i, y_i) belongs to $Neighbor(x, y, d)$, $d = \sqrt{10}$.

Fig. 6(IV) is created according to the following steps. Firstly, we define the average $Smooth(x, y)$ of all pixels in text area as threshold. If $Smooth(x, y)$ is greater than the threshold, we draw the pixel (x, y) as black color, else we draw it as white color (nothing in the picture).

We can see in Fig. 6(IV) that through computing the smoothness, "A" kind pixels or "B" kind pixels are marked as black color. Through observation of the character of "D" kind pixels, we find that these pixels are in holes mostly. Then we fill these holes and generate the Fig. 6(V). Then, the scope of the text stroke is narrowed, and distributed mainly in the area of black pixels in Fig. 6(V).

However, there are some "B" kind pixels in Fig. 6(V) yet. We cluster the color values of pixels which are marked as white color in Fig. 6(V) in Lab color space and get some clusters of color value. We chose the biggest 5 clusters to express the colors of background. We subtract the pixels belong to these clusters from pixels marked as black color in Fig. 6(V). Finally, we obtain the result as Fig. 6(VI) shows. We call pixels marked as black colors in fig. 6(VI) "candidate foreground pixels" (CFP for short). They are the narrowed scope of the text stroke.

C. Element image division

The foreground pixels are limited CFP according to the last section and we will generate a sequence of element image in this section. The element image is created according to the following steps.

Firstly, we cluster the color values of CFP in Lab color space. Then, we chose the biggest 5 clusters to express the colors of foreground. For each cluster, we can generate an element image.

Finally, for each pixel in the text area, if this pixel belongs to CFP and belongs to the cluster at the same time, we consider it as foreground pixel. Else, we consider it as background pixel.

Fig. 6(VII) and Fig. 6(VIII) are 2 element images of fig. 6(II). The following figure shows the element images of other text areas.

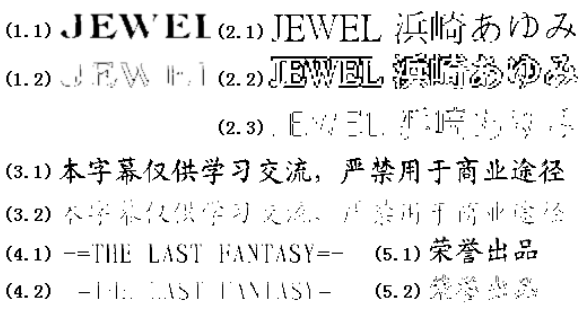

Figure 7. Other result of text extraction

IV. TEXT RECOGNITION

It is very difficult to identify the best element image generated in the last chapter by knowledge of image processing. We resort to language model. TABLE IV is the element images and their corresponding OCR results.

TABLE IV. ELEMENT IMAGES AND THEIR OCR RESULTS

Image	OCR result	probability
Fig.6(VII)	avumi hamasaki	13.58
Fig.6(VIII)	龊婀圈讯 ¨ i Jh 暑 ¨ H_墨 s 函 j。《i	7.03
Fig.7(1.1)	JWEEL	12.67
Fig.7(1.2)	．。瓦鸸	9.86
Fig.7(2.1)	JWEEL 浜崎弱国 A	10.91
Fig.7(2.2)	凰丽厦皿群崂为渗黔	8.56
Fig.7(2.3)	E-1'㈣寄一 ij■	6.43
Fig.7(3.1)	本字幕仅供学习交流，严禁用于商业途径	14.68
Fig.7(3.2)	表字幕仅供学习交流，严禁用亏商业途径	13.95
Fig.7(4.1)	一=rHE LAST FAN TASY=一	12.39
Fig.7(4.2)	一川．．＼)I I、＼I＼)I 一	6.37
Fig.7(5.1)	荣誉出品	14.23
Fig.7(5.2)	，之，、…．矗	8.52

Using a simple language model, we can easily compute the sentence probability of each OCR result. In this paper, we use letter appearance probability (Statistic 438,023 letters by Dewey. G), Chinese character appearance

probability and high frequency words table to compute the sentence probability of each OCR result.

The content of the 3rd column in TABLE IV is the sentence probability of each OCR result. We trend to choose the OCR result with greater sentence probability as the final result as TABLE V shows.

TABLE V. FINAL RESULT OF TEXT PROCESSING

Index	OCR result	Precision
1	avumi hamasaki	92.31%
2	JWEEL	100%
3	JWEEL 浜崎弱国 A	70%
4	本字幕仅供学习交流，严禁用于商业途径	100%
5	一=rHE LAST FAN TASY=一	94.44%
6	荣誉出品	100%

V. CONCLUSIONS

Text processing in video frames involves detection, localization, extraction, and recognition of the text from a given video data. This paper proposes an efficient approach to handle existing difficulties in text processing. The approach has been implemented and tested with a variety of video sources, and the result of experiments is promising.

ACKNOWLEDGMENT

The research work is supported by 863 Key Program of China (2006AA010105); National Natural Science Foundation of China (60772081， 60872133); Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PXM2007_014224_044677,PXM2007_014224_044676); Scientific Research Common Program of Beijing Municipal Commission of Education (KM200710772010).

REFERENCES

[1] K. Jung, K.I. Kim, A.K. Jain, Text information extraction in images and video: a survey, Pattern Recognition, 37(2004) 977-997.

[2] Min Cai, Jiqiang Song, and Michael R. Lyu, A new approach for video text detection, Image Processing, 1(2002) 117-120.

[3] Chengjun Zhu, Yuanxin Ouyang, Lei Gao, et al., An automatic video text detection, localization and extraction approach, Proc. of Inter. Conf. on Signal-Image Techno. & Internet, 2006 166-175.

[4] Datong Chen, Hervé Bourlard, and Jean-Philippe Thiran, Text identification in complex background using SVM, Computer Vision and Pattern Recognition, 2(2001) 621-626.

[5] Lyu, M. R. Jiqiang Song, Min Cai, A comprehensive method for multilingual video text detection, localization, and extraction, Circuits and Systems for Video Technology, 15(2005) 243- 255.

[6] Chong-Wah Ngo and Chi-Kwong Chan, Video text detection and segmentation for optical character recognition, Multimedia Systems, 10(2005) 261-272.

[7] Qixiang Ye, Qingming Huang, Wen Gao, et al., Fast and robust text detection in images and video frames, Image and Vision Computing 23(2005) 565-576.

[8] Tianxue Zhao, Guangmin Sun, Cheng Zhang, et al., Study on video text processing, Industrial Electronics, 2008 1215-1218.

[9] Xueming Qian and Guizhong Liu, Text Detection, Localization and segmentation in compressed videos, Acoustics, Speech and Signal Processing, 2(2006) 14-19.

[10] Julinda Gllavata and Bernd Freisleben, Adaptive fuzzy text segmentation in images with complex backgrounds using color and texture, Computer Analysis of Images and Patterns, 10(2005) 756-765.

[11] Bouaziz, B., Mahdi, W., Ardabilain, M.,et al., A new approach for texture features extraction: Application for text localization in video images, Multimedia and Expo, 2006 1737-1740.

[12] Rui Ma, Wei Hu, Qiao Huang, et al., Robust text stroke extraction from video, Multimedia and Expo, 2007 1391-1394.