Automatic Text Detection In Video Frames Based on Bootstrap Artificial Neural Network And CED

Yan Hao
Institute of Automation
Chinese Academy of Sciences
P.O.Box 2728-9Dep
100080, Beijing, P.R.China
hao.yan@mail.ia.ac.cn

Zhang Yi
Institute of Biophysics
Chinese Academy of Sciences
P.O.Box
100101, Beijing, P.R.China
mimicat0401@sina.com

Hou Zeng-guang
Institute of Automation
Chinese Academy of Sciences
P.O.Box 2728-9Dep
100080, Beijing, P.R.China
zengguang.hou@mai.lia.ac.cn

Tan Min

ABSTRACT

In this paper, one novel approach for text detection in video frames, which is based on bootstrap artificial neural network (BANN) and CED operator, is proposed. This method first uses a new color image edge operator (CED) to segment the image and achieve the elementary candidate text block. And then the neural network is introduced into the further classification of the text blocks and the non-text blocks in video frames. The idea of bootstrap is introduced into the training of the ANN, thus improving the effectiveness of the neural network greatly. Experiments results proved that this method is effective.

Key Words: text detection, video frame, bootstrap, artificial neural network, CED,

1. INTRODUCTION

With the development of the Internet and multimedia applications, there is an urgent demand for efficient and accurate content-based browsing and retrieving systems. Text embedded in video frames often carries the most important information, such as time, place, name or topics, etc. This information may do great help to video indexing and video content understanding. To extract text information from video frames, which is often referred as video OCR, the first essential step is to detect the text area in video frames.

Many methods have been introduced to detect and locate the text in video sequence. Most of the published methods for text detection can be classified into two categories. The first category is component-based methods. Text region are detected by analyzing the geometrical arrangement of edges or homogeneous color/grayscale components that belong to characters [1]. Smith detected text as horizontal rectangular structures of clustered sharp edges [2]. Combining using the features of color and size range, Lienhart identified text as connected components that have corresponding matching components in consecutive video frames [3]. The component-based methods can locate the text quickly but have difficulties when the text is embedded in complex background or touches other graphical objects [4]. The second category is texture-based methods. Jain has used various textures in text to separate text, graphics and halftone image regions in scanned grayscale document images [1][5][6]. Zhong further utilized the texture characteristics of text lines to extract text in grayscale images with complex backgrounds [1][7]. Zhong located candidate caption text regions directly in DCT compressed domain using the intensity variation information encoded in the DCT domain [1]. Those texture-based methods decrease the dependency on the text size, but they have difficulty in finding accurate boundaries of text areas. The two categories methods are limited to many special characters embedded in text of video frames, such as text size and the contrast between text

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Journal of WSCG, Vol.11, No.1., ISSN 1213-6972
WSCG'2003, February 3-7, 2003, Plzen, Czech Republic.

Copyright UNION Agency – Science Press

and background in video images. To detect the text efficiently, those methods usually defined a lot of rules that are largely dependent of the content of video. Because the video background is complex and moving/changing, traditional ways that tried to describe the contrast between text and video backgrounds have difficulty to detect text efficiently. So it is significant to synthesize both the traditional method using many locating rules and that based on statistical models for detecting and locating text in video frames.

In this paper, one new method based on bootstrap neural network and CED operator is proposed for text detection in video frames. Compared with the traditional edge operator, the CED (color edge detector) operates on the overall effect of three channels of Y.I.Q color space. Combining with morphological methods, the CED can locate not only gray images but also color images effectively. Artificial Neural Network (ANN) can embed the statistical features of one pattern into the structure and parameter of the ANN network. ANN has the special merit for the complex video objects. What is more important is that in this paper the idea of bootstrap, which is proposed by Sung to detect the face [8], is introduced into the training of ANN network, thus improving the effectiveness of the ANN greatly.

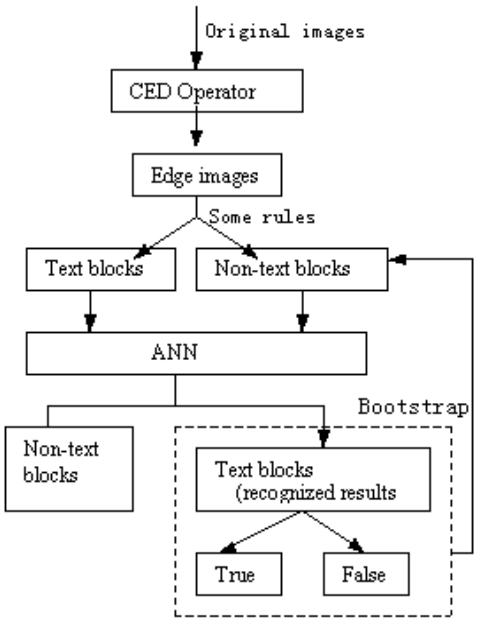


Figure 1. Flow chart of the proposed text detection algorithm

Figure 1 shows the flow chart of the proposed text location algorithm. Firstly, the CED is proposed to detect the edges of the original image and morphological methods are used to get the candidate blocks. Secondly, some rules are introduced to classify the blocks into text blocks and non-text blocks. Thirdly, the Gabor texture features are input as the train samples into the ANN to train the network. The bootstrap is introduced into this process. Those non-text blocks that are classified as text-blocks falsely are put in the non-text block training set of ANN as the new non-text blocks training samples. Finally, the ANN is used to classify the text blocks and non-text blocks after it is fully trained and then the detection result is achieved.

2. TEXT REGION DETECTION BASED ON CED

2.1 CED operator

High-level accuracy and ability for removing noises are the important requirements for the edge detection of color images, just as that of gray images. Here the traditional Roberts Operator is transformed into CED that makes use of the Y.I.Q color system. Considering that the Y, I and Q have different influences on video images, the different weight numbers are introduced to balance those influences. The CED operator is described as follows:

$$\text{CED}= \sqrt{\delta_1^2 + \delta_2^2} \tag{1}$$

Where δ_1 and δ_2 are defined as:

$$\delta_1 = Dis(i, j, i+1, j+1);$$
$$\delta_2 = Dis(i+1, j, i, j+1); \tag{2}$$

Where $Dis(i_1, j_1, i_2, j_2)$ is defined as the Eulerian distance between two pixels of the image in Y.I.Q color system, its definition is:

$$\begin{aligned} Dis(i_1,j_1,i_2,j_2) = \{ & \lambda_1 [I(i_1,j_1,y) - I(i_2,j_2,y)]^2 \\ & + \lambda_2 [I(i_1,j_1,i) - I(i_2,j_2,i)]^2 \\ & + \lambda_3 [I(i_1,j_1,q) - I(i_2,j_2,q]^2 \}^{1/2} \end{aligned} \tag{3}$$

2.2 Elementary Text Detection Based on CED

Post-processing is important for segmenting the text and the background in those images that have been processed by CED. Because the text lines in the video are usually horizontal, we must strengthen the image's horizontal edges. So the edge operator that has longitudinal character is used here to extract the edge of the image again after CED extracted it firstly. In this paper, the longitudinal *sobel* operator is used to extract edge after CED performed such operation. In this way, the binary image is achieved and the candidate text blocks can be located elementarily by morphological methods. The algorithm is described as follows:

(1) The original image I_1 in one of the video frames is processed by CED to get the grayscale edge image I_2。

(2) I_2 is processed by longitudinal *sobel* operator to get the binary edge image I_3。

(3) I_3 is processed by morphological methods to get the image I_4。Considering the horizontal features of texts in video images, we use the open operator to dilate I_3 in horizontal direction and then use the close operator to erode it in morphological direction.

After the processing described above is finished, some important rules are designed to locate some obvious text blocks and remove some obvious non-text blocks. Both the features of horizontal and longitudinal projection of image I_4 and the density features of it are considered to locate the text elementarily. The detailed rules are as follows:

(1) When both the horizontal projection P_h and the longitudinal projection P_v of one $m \times n$ block do not meet the inequality (4), this block is classified into non-text block set. To avoid the influence of text size on the algorithm, the pyramid method is used to extract the text in video images with different resolutions. That is, the images in different resolutions are classified respectively. And then the results got in different resolution are combined to get the final classification. Here if all of the block images in different resolution do not meet the inequality (4), those blocks are classified as non-text blocks.

$$P_h > \mu_1 \quad \text{and} \quad P_v > \mu_2 \tag{4}$$

where μ_1 and μ_2 are the low limit of horizontal and longitudinal projection respectively.

(2) When the density of $m \times n$ block are less than the threshold μ_3，the block is classified as non-text block。Where μ_3 is defined as the low limit of density determined.

(3) When the $m \times n$ block meet both $density > \mu_4$ and (4)，the block is classified as text block. Where μ_4 is defined as the low limit of density.

Then the elementary detection process is finished. And the rest of the candidate blocks except for those determined by the rules given above are to be processed by the neural network in the following section.

3. TEXT BLOCK CLASSIFICATION BASED ON BOOTSTRAP ANN (BANN)

After the image is processed in the way described above, the text blocks are located elementarily. The following task is to locate the text blocks more accurately and remove non-text blocks that are often classified as text blocks by the CED. Due to the complexity of the images in video frames, the BANN is used to further classify the text blocks and non-text blocks.

3.1 Artificial Neural Network (ANN)

In this paper, the Back Propagation (BP) ANN is adopted for classification. BP neural network is the most widely used neural network model. Its merit is that is has strong ability of nonlinear projection and flexible network structure. All of its network structure, the number of layers, the number of nerve units and study coefficients can be adjusted according to the specific cases. And to realize such models is easy and quick. The structure of the BP artificial neural network is described in figure 2:

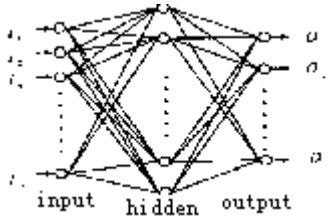


Figure 2. Structure of BP Neural Network

There are two output nodes of BP network in this paper, corresponding to the text block and non-text block respectively.

3.2 Feature Selection of Input Nodes of Back Propagation Neural Network

Because the text in video has the special texture, we adopt the texture characters of candidate blocks as the features to be recognized. Multichannel Gabor filter is a well-established method for texture analysis and has been demonstrated to have good performance in texture discrimination and segmentation [9]. In theory, any kind of texture analysis methods can be employed here. But experiments show that the Gabor filter has better performance [10] [11] [12], and therefore is used in this paper.

3.2.1 The Concept of Gabor Filter

In this paper, we use pairs of isotropic Gabor filters with quadrature phase relationship [10]. The models in spatial domain is as follows:

$$\begin{cases} h_e(x,y,f,\theta,\sigma) = g(x,y,\sigma)\times\cos[2\pi f(x\cos\theta + y\sin\theta)] \\ h_o(x,y,f,\theta,\sigma) = g(x,y,\sigma)\times\sin[2\pi f(x\cos\theta + y\sin\theta)] \end{cases} \quad (5)$$

where $h_e(x,y,f,\theta,\sigma)$ and $h_o(x,y,f,\theta,\sigma)$ responds to so-called even- and odd-symmetric Gabor filters respectively, and $g(x,y,\sigma)$ is an isotropic Gaussian function that is described as follows:

$$g(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\times\exp\left[-\frac{x^2+y^2}{2\sigma^2}\right] \quad (6)$$

f,θ,σ in (5) are three important parameters. They are spatial frequency, spatial orientation, and space constant of the Gabor envelope respectively. It is important to understand how to solve the problems in frequency domain for Gabor filter. So it is necessary to know the frequency responses of the Gabor filters that is described as follows:

$$\begin{cases} H_e(u,v) = \dfrac{[H_1[u,v] + H_2(u,v)]}{2} \\ H_o(u,v) = \dfrac{[H_1[u,v] - H_2(u,v)]}{2j} \end{cases} \quad (7)$$

where $j = \sqrt{-1}$, $H_1(u,v)$ and $H_2(u,v)$ are:

$$\begin{cases} H_1(u,v) = \exp\left\{-2\pi^2\sigma^2\left[(u-f\cos\theta)^2 + (v-f\sin\theta)^2\right]\right\} \\ H_2(u,v) = \exp\left\{-2\pi^2\sigma^2\left[(u+f\cos\theta)^2 + (v-f\sin\theta)^2\right]\right\} \end{cases} \quad (8)$$

3.2.2 Frequency Response

As described in Figure 3, the relationship between the input image p(x,y) and output image q(x,y) is :

$$\begin{cases} q(x,y) = \sqrt{q_e{}^2(x,y) + q_o{}^2(x,y)} \\ q_e(x,y) = h_e(x,y)\otimes p(x,y) \\ q_o(x,y) = h_o(x,y)\otimes p(x,y) \end{cases} \quad (9)$$

where $\otimes$ is defined as convolution. In practical application, we usually use the Fourier Transform to calculate the convolution. That is:

$$\begin{aligned} q_e(x,y) &= FFT^{-1}\left[P(u,v)\times H_e(u,v)\right] \\ q_o(x,y) &= FFT^{-1}\left[P(u,v)\times H_o(u,v)\right] \end{aligned} \quad (10)$$

where $P(u,v) = FFT\left[p(x,y)\right]$, which is the Fourier Transform of $p(x,y)$.

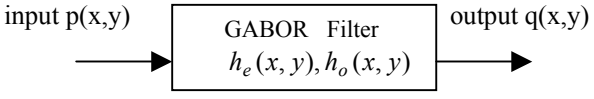


Figure 3. Frequency Response of Gabor Filter

3.2.3 Filter Design

Each pair of the Gabor filters $h_e(x,y)$, $h_o(x,y)$ are tuned to a specific band of spatial frequency and orientation, which respond to f and θ. How to select these parameters is an important problem. Tan presented that there is no need to uniformly cover the entire frequency plane so far as texture recognition is concerned [13]. He also pointed that since the Gabor filters are of central symmetry in the frequency domain, only half of the frequency plane is needed. So four values of orientation are selected: $\theta = 0^0, 45^0, 90^0, 135^0$. Zhu pointed that in order to achieve good results, for an image of size $N\times N$, central frequencies are chosen within $f < N/4$ [10]. In our experiments, the input image is tuned to the normal size 128×128. For each orientation θ, we select 2, 4, 8, 16, 32 as frequencies, getting a total of 20 Gabor channels ($4\times5 = 20$, 4 orientations and 5 central frequencies). The spatial constant γ is chosen as: $\gamma = 0.01$.

3.2.4 Features Extracted by Gabor Filters

In our experiments, the mean values ($\overline{q}$) and the Standard deviation (γ) of the channel output images are chosen to represent the features. The definition of them is

$$\begin{aligned} \overline{q} &= \frac{1}{N\times N}\sum_{x=1}^{N}\sum_{y=1}^{N} q(x,y) \\ \gamma &= \sqrt{\sum_{x=1}^{N}\sum_{y=1}^{N}\frac{[\,q(x,y) - \overline{q}\,]^2}{N\times N}} \end{aligned} \quad (11)$$

Thus, a total of $20\times2 = 40$ features are extracted from the input image. Figure 4 shows the flow chart of coarse feature extraction using Gabor Filters.

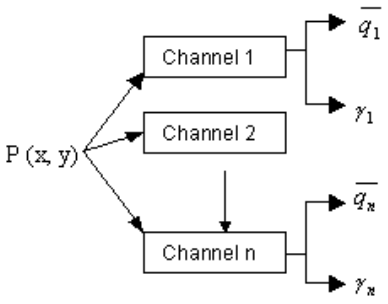


Figure 4. The feature extraction of Gabor Filter

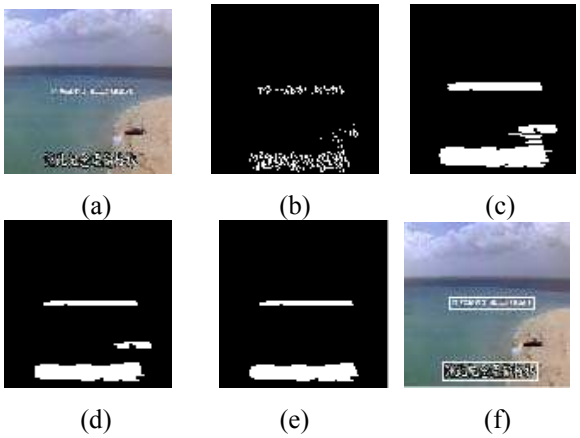
3.3 Bootstrap of BP Neural Network and Text Block Recognition

Just as those described in Figure 1, the blocks got by CED are first classified into text blocks and non-text blocks that are included into text block sample set and not-text block sample set for training the BP network respectively. The non-text block sample set is originally a very small set. Then the Gabor features of these blocks are input to train the BP network. During the training process, the bootstrap is introduced into our method. Bootstrap means that when the output of the BP network is text block that is in fact non-text block and classified falsely by BP network, this block is then included in the training sample set for non-text block. The process is iterated steadily until the non-text block samples are enough for training the network. Then a complete detection model is built up for text detection in video frames.

4.IMPLEMENTATION AND EXPERIMENTAL EVALUATION

4.1 Experimental results

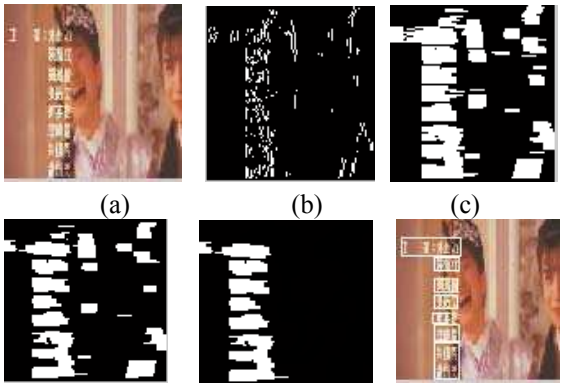
The experiments are performed following the algorithm presented in this paper. The experimental data are from the various videos of some movies. The total length of these videos is about 70 minutes. The testing data contain 205 video frames. Figur 5 and Figure 6 show the total process of text detection. In the images shown in each of them, (a) shows the original image I_1, (b) shows the edge image I_2 got by CED, (c) shows the binary image I_3 got after I_2 is processed by open morphological operator, (d) shows the binary image I_4 got after I_3 is processed by close morphological operator, (e) shows the image got by BANN, (f) shows the final detection results in the original video image. Figure 7 (a), (b), (c) and (d), (e), (f) are two other experiments respectively, in which the first one is the original image, the second one is the image processed by BANN, the last one is the detection result. From those images, we can see that although the background is complex, the detection of the text is accurate and effective.



(a) (b) (c)

(d) (e) (f)

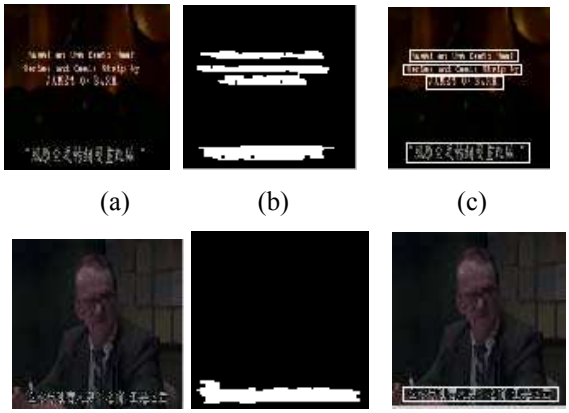
Figure 5. Experimental Results 1



(a) (b) (c)

(d) (e) (f)

Figure 6. Experimental Results 2



(a) (b) (c)

(d) (e) (f)

Figure 7. Experimental Results 2

4.2 Experimental Evaluation

The statistical experimental results are listed in Table 1.

Total_Frames	205
Total_Text_Blocks	964
Total_Missed_Text_Blocks	59
Total_False_Alarms	63
Detection_Rate	87.3%
False_Alarm_Rate	6.54%

Table 1. Statistical Detection Results

Where *False_Alarm_Rate* and *Detection_Rate* are defined respectively as follows:

$$False\_Alarm\_Rate = \frac{Total\ \_\ False\ \_\ Alarms}{Total\ \_\ Text\ \_\ Blocks})$$

$$Detection\_Rate = \frac{Total - Detected\ \_\ Text\ \_\ Blocks}{Total\ \_\ Text\ \_\ Blocks}$$

Total_Detected_Text_Blocks=(Total_Text_Blocks
-Total_Missed_Text Blocks
- Total False Alarms); (11)

From Table1, we can see that the method can detect and locate the text blocks efficiently. The detection rate is 87.3% and the false alarm rate is only 6.54%. However, we find it is difficult to recognize the small characters and may have false alarms in some blurred texts. Figure.8 shows some samples that have some false detection results. That is because the different texture features of the image have the different impact on the method in this paper. If the texture of one false text block is very similar with that of text block, the false alarms may occur when the CED segment the blocks falsely too.

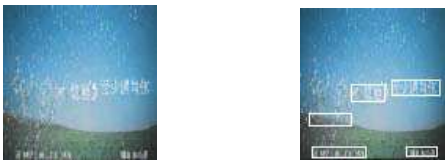


Figure 8. False Alarms in similar background

5. CONCLUSION AND FUTURE WORK

In this paper, a new text detection algorithm based on bootstrap neural network and CED operator is proposed. The detection rate is 87.3% in our experiments. Although the experimental results is satisfying, there are some future work to do:

(1) Improving the design of the classification rules. (2) Extract more effective features of text block and non-text blocks. (3) Enhance the speed of the algorithm to make it fit the video retrieval in large databases.

Reference

[1] Yu, Zhong; Hongjiang, Zhang; Jain, A.K, "Automatic caption localization in compressed video," IEEE Trans. On PAMI, Vol. 22, Issue 4, April, 2000, pp. 385-392.

[2] M.A. Smith and T. Kanade, "Video Skimming and Characterization through Language and Image understanding Techniques," technical report, Carnegie Mellon Univ. 1995

[3] R. Lienhart and F. Stuber, "Automatic Text Recognition in Digital Videos," Proc. Praktische Informatic IV, pp.68-131, 1996

[4] Jie Xi, Xian-Sheng Hua, Xiang-Rong Chen, Liu Wenyin, Hong-Jiang Zhang. A Video Text Detection and Recognition System. IEEE International Conference on Multimedia and Expo (ICME 2001), Waseda University, Tokyo, Japan, August 22-25, 2001.

[5] A. K. Jain and S. Bhatt acharjee, "Text Segmentation Using Gabor Filter for Automatic Document Processing," Machine Vision and Application, Vol. 5, No.3, pp. 169-184, 1992

[6] A. K. Jain and Y. Zhong, "Page Segmentation in Images and Video Frames," Pattern Recognition, Vol. 31, No. 12, pp. 2055-2-76, 1998

[7] Y. Zhong, K. Karu, and A.K.Jain, "Locating Text in Complex Color Images,"Pattern Recognition, Vol.28, No. 10, pp. 1523-1536, Oct.1995

[8] Sung K, Poggio T. Example-based learning for view based human face detection. IEEE Trans. on PAMI, 1998, Vol. 20, No. 1, 39-51

[9] M.R.Turner, "Texture Discrimination by Gabor Functions," Biological Cybernetics, Vol, 55, no.1, pp.55-73, Jan, 1990

[10] Yong Zhu, Tieniu Tan, Yunhong Wang, "Font Recognition Based on Global Texture Analysis", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.23 no.10, pp.1192-1200, Oct.2001.

[11] H.E.S. Said,K.D.Baker, And T.N.Tan, "Personal Identification Based on Handwriting," Proc.14th Int'l Conf.pattern Recognition, Assoc.for Pattern Recognitoion Int'l, pp.1761-1764,1998

[12] G.S.Peake and T.N.Tan,"Script and Language Identification from Document Images," Proc.BMVC'97,vol.2, pp.169-184, Sept, 1997

[13] T.N.Tan, "Texture Feature Extraction via Cortical Channel Modelling,"Proc.11th Int'l Conf.Pattern Recognition, Assoc. for Pattern Recognition Int'l, vol. III, pp. 607-610,1992

[14] W.Qi, et. Al. "Integrating Visual, Audio and Text Analysis for News Video," 7th IEEE Int. Conf. on Image Processing (ICIP 2000). Vancouver, British Columbia, Canada, 10-13 September 2000.