Caption Extraction in News Video Sequence using Frequency Characteristic

Younglae Bae, Byung Tae Chun, Seyoon Jeong

Image Processing Department, Computer & Software Technology Lab.,
ETRI(Electronics and Telecommunications Research Institute)
161 Kajong-Dong, Yusong-Gu, Taejon, 305-350, Korea
E-mail : yljb@etri.re.kr Fax : +82-42-860-4884

Abstract : Popular methods for extracting a text region in video images are in general based on analysis of a whole image such as merge and split method, and comparison of two frames. Thus, they take long computing time due to the use of a whole image. Therefore, this paper suggests the faster method of extracting a text region without processing a whole image.

The proposed method uses line sampling methods, FFT and neural networks in order to extract texts in real time. In general, text areas are found in the higher frequency domain, thus, can be characterized using FFT. The candidate text areas can be thus found by applying the higher frequency characteristics to neural network. Therefore, the final text area is extracted by verifying the candidate areas. Experimental results show a perfect candidate extraction rate and about 92% text extraction rate. The strength of the proposed algorithm is its simplicity, real-time processing by not processing the entire image, and fast skipping of the images that do not contain a text.

1. Introduction

Now days video image data is one of the most popular multimedia data [1][2][3]. This is because the price of storing media of video image data is getting cheaper. In future, the use of video image data will be also increased as the communication techniques are developed. One important source of information about video is the text contained therein. Text in TV programs such as news, travel guide and hobby programs, and in other source such as training and educational videos is a typical example. Such text is used to show the title of a program, the headline of a news story, the name of a person talking or a place being shown in the video, and statistical data or other information related to a topic.

Existing work on text recognition has focused primarily on optical character recognition in printed and hand-written documents since there exists a great demand in and market for document readers for office automation systems. These systems have attained a high degree of maturity [4]. Michael A. Smith and Takeo Kanade briefly describe a method[5] which concentrates on extracting regions from video frames that contain textual information. However, they deal with the preparation of the detected text for standard optical character recognition software. In particular, they do not try to determine the character's outline or to segment the individual characters. Another interesting approach to text recognition in scene images is that of Jun Ohya, Akio Shio, and Shigeru Akamatus[6]. Text in a scene image exists in a 3D space, so it can be rotated, tilted, or partially shadowed, and it can appear under uncontrolled illumination. In view of the many possible degrees of freedom of text characters, Ohya et al. restricted them to being almost upright, monochrome, and not connected in order to facilitate detection. We describe a text extraction method using FFT and neural network. The overview is shown in Fig. 1.

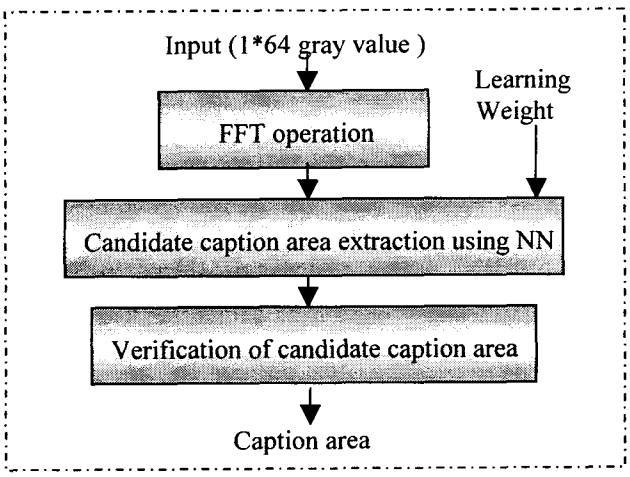
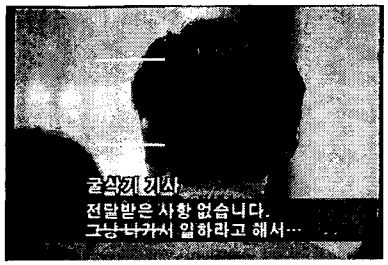


Fig. 1. Text extraction method using FFT and NN

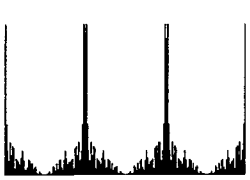
2. FFT characteristics of texts

As we known, there are various kinds of texts in digital video, whose properties differ according to the generation methods.

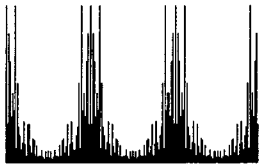
(a) News icon text



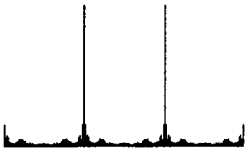
(b) context texts



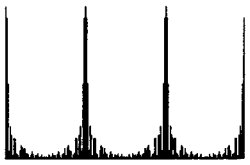
(c-1) FFT of (a) line-(1)



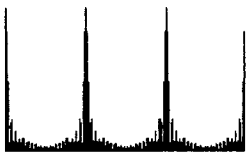
(c-2) FFT of (a) line-(2)



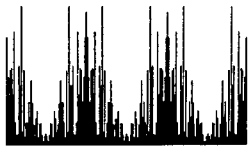
(c-3) FFT of (a) line-(3)



(d-1) FFT of (b) line-(1)



(d-2) FFT of (b) line-(2)



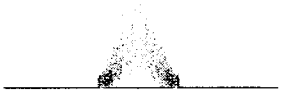
(d-3) FFT of (b) line-(3)

Fig. 2. Frequency characteristic of text and non-text

When some texts are generated by a video title machine, they can be categorized as news icon text, content text, on content text made by the producer Text areas usually tend to include characters with uniform density. Compared with the background area, a text area has the affinity of regularity at brightness and darkness. This affinity can be shown as a regular frequency characteristics in a frequency domain. This paper uses a fast Fourier transform for analyzing an image in terms of frequency domain. Fig.2 shows video images and their frequency characteristics by using FFT. When we perform FFT to Fig.2(a)(b), the frequency characteristic is shown in Fig.2(c-1),(c-2),(c-3),(d-1),(d-2),(d-3). It can be found in Fig. 2 (c-2)(d-3) that the frequency characteristics of a text region have stronger response at higher frequencies. As displayed in Fig. 2 (c-1),(c-3),(d-1),(d-2), non-text regions show stronger response at lower frequencies.

Therefore, in order to produce general frequency characteristics of non-text and text regions, this paper collects lots of frequency responses of different text and non-text regions, and applies them into a neural network. The general frequency characteristics of non-text and text regions are displayed in Fig. 3 where a non-text region has stronger response at lower frequencies and a text region at higher frequencies.

(a) the frequency characteristic of non-text



(b) the frequency characteristic of text

Fig. 3. The frequency characteristic

3. Automatic Text Extraction using Neural Networks

3.1 Extracting candidate areas for text area using Neural Networks

This section tells how to extract text regions by using FFT and neural network. Here, FFT is applied in terms of line sampling rather than of a whole image, shown in Fig. 5., in order to reduce the preprocessing time for text extraction and the search time. Also, FFT at each line is computed in terms of a overlapped-segment of $1*64(2^n$ n:=6) pixels with 32 overlapped-pixels. This is because the use of $128(2^n$ n:=7) is so large that it might cover non-text and text regions together. The results of FFT are used as the input of a neural network. The dimension of FFT computation is $64(2^n$: n = 6) pixels which are considered as representing a small area with text or not. When an input of FFT is 64, an input of representative frequency is 32 owing to Nyquist frequency theory.

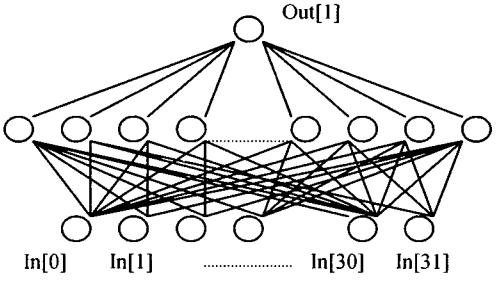


Fig. 4. Neural Network for learning

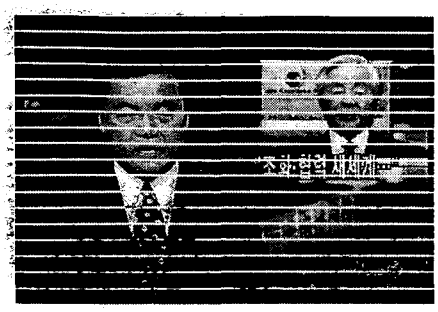


Fig. 5. Line sampling

So, the number of input neurons of a given neural network is 32, the number of hidden neurons is 40 and the number of output neuron is 1.

Fig 6. Extraction of candidate text area

Fig. 4 shows the topology of a constituted neural network model. This neural model is trained with collected data after FFT computation. To collect a set of training data, we use the line sampling method as shown in Fig.5. A given image is divided into several rows with constant range. The range value among rows is the same. And then the candidates of texts are set up. The candidate area is considered having a text if the output is higher than a given threshold. Fig. 6 shows candidates of extracted texts. Fig.7 shows the result of thresholding of extended candidates area of extracted text

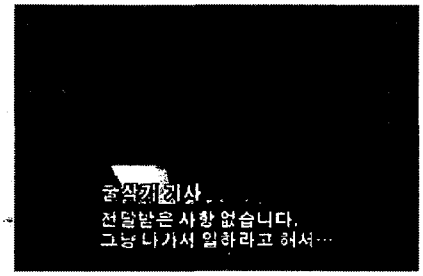
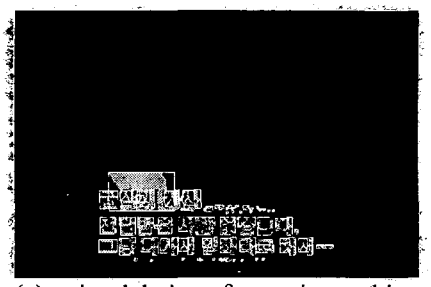


Fig. 7. Extraction of candidate text area

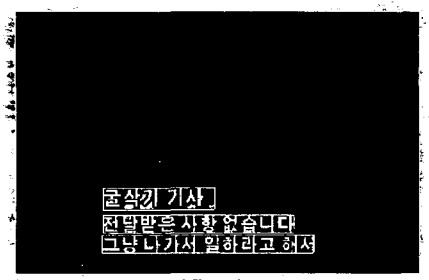
3.2 Verification of candidates of text area

The post processing method is in this section offered to decide whether a candidate has characters or not. Firstly, a noise-cancellation is applied before verifying the candidates of text regions. Labeling operation is conducted for candidated character regions and the coordinate of character composition is obtained as shown in Fig. 8 (a). Finally, the regions, which are too large or small for a character region, are removed. This preprocessing helps removing noise regions before verification step.

The candidate text regions are verified by two stages. In the first stage, the character regions are merged horizontally to form text lines. The size and other characteristics of the text lines are analyzed to remove non-text regions. If the fill factor is too low, the corresponding regions are discarded. Next, the width-to-height ratio of the blocks is calculated. If it exceeds limits, i.e. dose not lie between *min_ratio* and *max_ratio*, the corresponding regions are also discarded.



(a) noise deletion after region making



(b)results of verification text area

Fig. 8. Extraction of text area

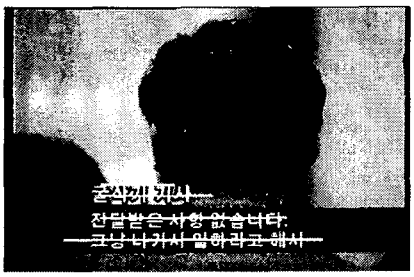
In the second stage, the text lines are segmented and the character sizes and the number of characters are examined. If certain conditions are satisfied, the text region is extracted. The extracted text area after verification is shown in Fig. 8(b).

4. Experiments and Results

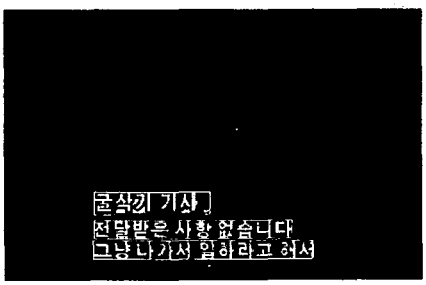
The experiments have been performed on a Pentium PC with 333 MHz CPU. The program is implemented in Visual C++ Ver.5.0. The MPEG1 data have been used for the experiments, and the data have been acquired with an RT5 image board. The video image is a news program called "KBS 9 : 00 news." The texts at images are coming out for 3 to 6 seconds. The image frames were extracted at 2 frames per second.

The number of training data is 854. The tolerant error was 0.0001. This experiment shows that the result of extracting candidate areas was 100% when this approach was applied to 3 scenes with 483 frames. Finally, 92% of extracting candidate areas was acquired when the verification was performed with the combination of various kinds of background and character colors.

In Fig. 9-(a), the extracted candidate areas are represented by lines. The results are shown at Fig9-(b) after applying the verification step.



(a) Extraction of candidates text area



(b)Extraction of text area

Fig. 9. Extraction of text area

5. Conclusion

This paper introduces a new method for extracting text areas in video images by FFT and neural network. Through a few experiments, the performance is approved. But the followings should be researched more in the future. This approach could be ineffective when applied to images with a little difference in color between text and background areas. In addition, the verification methods through learning by examples would be better than that by isolated and multiple areas.

Reference

[1] M. Flickner, H. Sawhney, W. Niblack, J. Hafner, D. Lee., D. Petkovic, "Query by Image and Video Content : The QBIC system," IEEE Computer, Sept., pp.23-32, 1995.

[2] HongJiang Zhang, Shuang Yeo Tan, Stephen W.Smoliar, Gong Yihong, "Automatic parsing and indexing of news video," Multimedia Systems Vol 2, pp.256-266, 1995

[3] Michael A. Smith and Takeo Kanade, "Video skimming for Quick Browsing based on Audio and Image characterization," Carnegie Mellon University, Technical Report CMU-CS-95-186, July 1995.

[4] Shunji Mori, Ching Y. Suen, Kazuhiko Yamamoto,"Historical review of OCR research and Development", Proceeding of the IEEE, Vol.80, No.7, pp.1029-1058, July 1992.

[5] Michael A. Smith and Takeo Kanade, "Video Skimming for Qucik Browsing based on Audio and Image characterization," Carnegie Mellon Unv. Technical Reprot CMU-CS-95-186, July 1995

[6] Jun Ohya, Akio Shio, and Shigeru Akamatsu, "Recogniton characters in scene images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No.2, pp.214-220, 1884