Fast Text Caption Localization on Video Using Visual Rhythm

Seong Soo Chun[1], Hyeokman Kim[2], Jung-Rim Kim[1],
Sangwook Oh[1], and Sanghoon Sull[1]

[1] School of Electrical Engineering, Korea University, Seoul Korea
{sschun,jrkim,osu,sull}@mpeg.korea.ac.kr
[2] School of Computer Science, Kookmin University, Seoul Korea
hmkim@cs.kookmin.ac.kr

Abstract. In this paper, a fast DCT-based algorithm is proposed to efficiently locate text captions embedded on specific areas in a video sequence through visual rhythm, which can be fast constructed by sampling certain portions of a DC image sequence and temporally accumulating the samples along time. Our proposed approach is based on the observations that the text captions carrying important information suitable for indexing often appear on specific areas on video frames, from where sampling strategies are derived for a visual rhythm. Our method then uses a combination of contrast and temporal coherence information on the visual rhythm to detect text frames such that each detected text frame represents consecutive frames containing identical text strings, thus significantly reducing the amount of text frames needed to be examined for text localization from a video sequence. It then utilizes several important properties of text caption to locate the text caption from the detected frames.

1 Introduction

With rapid advances in digital technology, the amount of multimedia information available continues to grow. As multimedia contents become readily available, archiving, searching, indexing and locating desired content in large volumes of multimedia, containing images and video in addition to the textual information, will become even more difficult. One important source of information that can be obtained from image and video is the text contained therein. The video can be easily indexed if access to this textual information content is available. They provide clear semantics of video, and are extremely useful in deducing the contents of video.

A large number of methods have been extensively studied in recent years to detect text in uncompressed images and video. Ohya et al. [1] perform character extraction by local thresholding and detect character candidate regions by evaluating gray level difference between adjacent regions. Hauptmann and Smith [2] use the spatial context of text and high contrast of text regions in scene images to merge large numbers of horizontal and vertical edges in spatial proximity to detect text. Shim et al. [3] use a generalized region labeling algorithm to find homogeneous regions for text. Wu et al.

S.-K. Chang, Z. Chen, and S.-Y. Lee (Eds.): VISUAL 2002, LNCS 2314, pp. 259–268, 2002.

[4] use texture analysis to detect and segment texts as regions of distinctive texture using pyramid technique for handling text fonts of different sizes. Lienhart [5] provide split and merge algorithm based on characteristics of artificial text to segment text. Li et al. [6] used wavelet analysis and employed a multi-frame coherence approach to cluster edges into rectangular shape. Sato et al. [7] adopted a multi-frame integration technique to separate static text form moving background.

A few methods have been also proposed to detect text regions in compressed domain. Yeo and Liu [8] propose a method for the detection of text caption events in video by modified scene change detection which cannot handle captions that gradually enters or disappears from frames. Zhong et al. [9] examined the horizontal variations of AC values in DCT to locate text frames and examined the vertical intensity variation within the text regions to extract the final text frames. Zhang and Chua [10] derived a binarized gradient energy representation directly from DCT coefficients which are subject to constraints on text properties and temporal coherence to locate text. However, none of them exploits the temporal coherence of text useful for reducing processing time by not applying all steps (detection, localization, and OCR) to every frames, which results in duplicates of the same text string in the database.

The main contribution of this paper is to develop an efficient and fast compressed DCT domain method to locate text captions on specific areas in digital video through a visual rhythm [12], an abstraction of video that is constructed by sampling certain group of pixels of each frame and by temporally accumulating the samples along time. Our method uses a combination of contrast and temporal coherence information on the visual rhythm to detect text frames such that each detected text frame represents consecutive frames containing identical text strings, thus significantly reducing the amount of text frames needed to be examined for text localization from a video sequence. It then utilizes several important properties of text caption to locate text caption from the detected frames. The visual rhythm constructed for text localization also serves as a visual feature to efficiently detect scene changes.

This paper is organized as follow: Section 2, gives a brief description of visual rhythm. Section 3 describes the proposed text frame detection, and text caption localization algorithm. Section 4 describes experimental results. In Section 5, we give concluding remarks.

2 Related Work

2.1 Visual Rhythm

For the design of an efficient real-time text caption detector, we resort on using a portion of the original video. This partial video must retain most, if not all, text caption information. We claim that a visual rhythm, defined below, satisfies this requirement. Let $f_{DC}(x,y,t)$ be the pixel value at location (x,y) of an arbitrary *W*x*H DC image* [11] which consists of the DC coefficients of the original frame *t*. Using the sequences of DC images of a video called the *DC sequence*, we define a *visual rhythm*, *VR*, of the video *V* as follows:

$$VR = \{ f_{VR}(z,t) \} = \{ f_{DC}(x(z), y(z), t) \}, \tag{1}$$

where $x(z)$ and $y(z)$ are one-dimensional functions of the independent variable z. Thus, the visual rhythm is a two dimensional image where the vertical z axis consists of a certain group of pixels from each DC image and the samples are accumulated along time in the horizontal t axis. That is, the visual rhythm is a two dimensional image consisting of pixels sampled from a three-dimensional data (DC sequence). The visual rhythm is also an important visual feature that can be utilized to detect scene changes [12].

The sampling strategy, $x(z)$ and $y(z)$, must be carefully chosen for a visual rhythm to retain text caption information. We define $x(z)$, $y(z)$ as follows :

$$(x(z), y(z)) = \begin{cases} \left(\frac{W}{H} z, z \right), & 0 \le z < H \\ \left(2W - \frac{W}{H} z, z - H \right), & H \le z < 2H \\ \left(\frac{W}{2}, z - 2H \right), & 2H \le z < 3H \end{cases}, \tag{2}$$

where W, H are the width and the height of a DC sequence respectively. Figure 1 illustrates the sampling strategy of the DC sequence for the construction of visual rhythm. The diagonal pixels of a frame from bottom-left most corner to top-right most corner are sampled when $0<z<H$, the diagonal pixels of a frame from bottom-right most corner to top-left most corner are sampled when $H<z<2H$, and vertical pixels of a frame from upper-middle to bottom-middle are sampled when $2H<z<3H$

This is partially, if not entirely, due to empirical observations that text caption carrying important information and herewith suitable for indexing is often embedded such that portions of text caption appear on these regions. However the sampling strategy can be flexibly set for specific video materials where the approximate locations of text caption to be located are known a priority. Figure 2 illustrates a vertical line of the visual rhythm of the DC sequence of Figure 1 constructed from equation (1) and (2).

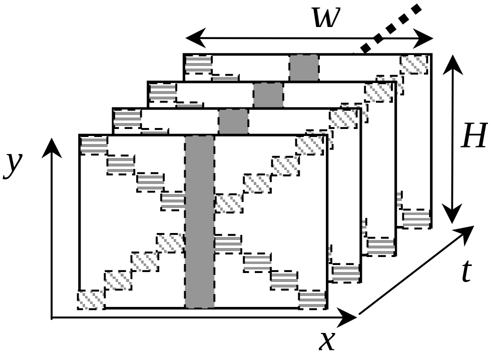


Fig. 1. Representation for regions of text appearance

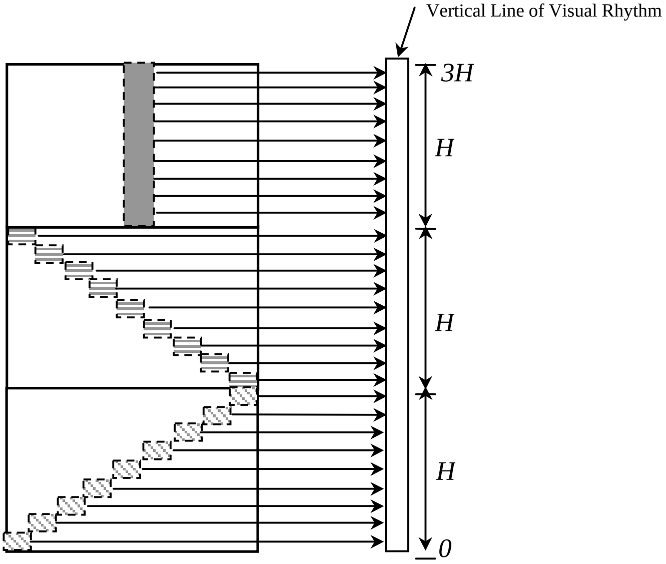


Fig. 2. The vertical line of a visual rhythm obtained by sampling the pixels of a DC sequence

2.2 Fast Generation of Visual Rhythm

Many compression schemes use the discrete cosine transform (DCT) for intra-frame encoding. Thus, the construction of a visual rhythm is possible without the inverse DCT. We simply extract the DC coefficients of each frame. As for the P- and B-frames of MPEG, algorithms for determining the DC images from inter-frame compressed P- and B-frames of MPEG-1 [11] and MPEG-2 [13] have already been developed. Therefore, it is possible to generate a visual rhythm fast, at least for the DCT-based compression schemes, such as Motion JPEG and MPEG videos.

3 Proposed Strategy

3.1 Text Frame Detection

A text frame is defined as a video frame that contains one or more text captions. Since a text caption usually appears in a number of consecutive frames, we propose an algorithm, which detects a representative text frame from the consecutive frames containing identical text strings to avoid unnecessary text caption localization for identical text strings.

The text frame detection algorithm detects text frames based on the following characteristics of text caption within video:

- Characters in a single text caption are mostly uniform in color.
- Text caption contrasts with their background.
- Text caption remains in a scene for a number of consecutive frames.

On the visual rhythm obtained through a DC sequence, the pixels corresponding to text caption manifest themselves as long horizontal line with high contrast with their background. Hence, horizontal lines on the visual rhythm with high contrast with their background are mostly due to text string, and they give us clues of where and when each text string appears within the video. The pixel value of the horizontal line on the visual rhythm also gives us clue on the pixel value of the text caption in DC image, allowing us for a simple algorithm for text caption localization within the frame.

To detect potential text frames, any horizontal edge detection method can be used on the visual rhythm. In our experiment we used Prewitt edge operator with convolution kernels

$$\begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

on the visual rhythm to obtain $VR_{edge}(z,t)$ as follows:

$$VR_{edge}(z,t) = \sum_{i=-1}^{1}\sum_{j=-1}^{1} w_{i,j}\, f_{VR}(z+j, t+i)\,. \tag{3}$$

To obtain *text lines* which we define as horizontal lines with high contrast with their background on a visual rhythm possibly formed due to text caption, the edges with $VR_{edge}(z,t)$ value greater than a threshold value τ (we set τ as τ=150 in our experiment) and uniform value $f_{VR}(z,t)$ are connected in the horizontal direction. Text lines lasting shorter than a specific amount of time is not considered, since text usually remains in the scene for a number of consecutive frames. Through observations on various types of video materials, shortest captions appear to be active for at least two seconds, which translates into a text line with frame length of 60 if the video is digitized at 30 frames per second. Thus the text lines with length less than 2 seconds can be eliminated. The resulting set of text lines appear in the form:

$$LINE_k, [z_k, t_k^{start}, t_k^{end}],\, k = 1,\ldots, N_{LINE}, \tag{4}$$

where $[z_k, t_k^{start}, t_k^{end}]$ denotes the *Z* coordinate, the beginning and the end frame of the occurrence of text line $LINE_k$ on the visual rhythm, respectively. The text lines are ordered by the increasing starting frame number,

$$t_1^{start} \le t_2^{start} \le \ldots \le t_{N_{LINE}}^{start}\,. \tag{5}$$

Figure 5(b) shows the binarized representation of text lines possibly formed by text caption from the visual rhythm in Figure 5(a). The frames not in between the temporal duration of $LINE_k$, do not contain any text caption and are omitted for further consideration as text frame candidates.

Once the frames without text have been excluded as text frame candidates, it is highly probable that the remaining frames of a video contain text caption within them.

However, it would be very inefficient to perform the text caption localization repeatedly for the same text caption remaining on the screen over multiple frames.

Since each text line possibly represents a single text caption, we only need to access a single frame to extract its corresponding text. Therefore, the number of text frames to be examined for text caption localization can be minimized by obtaining a maximum cardinality collection of disjoint intervals of text lines through the following algorithm:

```
$n \leftarrow 0$
$SET = \bigcup_{1<k<N_{LINE}} \{k : k \in N\}$
$WHILE\,(SET \neq \phi)\{$
    $e = \min(t_{k:}^{end} : k \in SET);$
    $A = \{\forall k \mid t_k^{start} < e\};$
    $F_n = (\max(t_k^{start} : k \in A) + e)/2;$
    $n{+}{+};$
    $SET \leftarrow SET - A;$
$\}$
```

Fig. 3. Pseudo-code to find minimal number of text frames

where F_j is the j^{th} frame to be accessed for text caption localization as the final output of the text frame detection stage, where $j < n$.

3.2 Text Caption Localization

The text caption localization stage spatially localizes text caption within a frame. Let $f_{DC}(x,y,t)$ be the pixel value at (x,y) of the DC image of frame t. From the visual rhythm obtained by the sampling strategy given by Equation (2), we can observe that $LINE_k$ is possibly formed due to a portion of a character located on $(x,y)=(x(z_k), y(z_k))$ in frames between t_k^{start} and t_k^{end} with the pixel values $f_{VR}(z_k, t)$ where $t_k^{start} < t < t_k^{end}$.

Furthermore, if a portion of a character is located on $(x,y) = (x(z_k), \mathrm{y}(z_k))$ within a DC image it can be assumed that portions of characters belonging to the same text caption to appear along $y=y(z_k)$ because text caption are usually horizontally aligned. Therefore, the text line information obtained from the text frame detection stage can be used to approximate the location of text within the frame, and enable us to provide an algorithm to focus on specific area of the frame.

From each of the detected frames F_j, we verify whether $LINE_k$, $t_k^{start} < F_j < t_k^{end}$ is formed by portions of text string located along $y=y(z_k)$.

For the text line $LINE_k$, we first cluster the pixels with pixel value $f_{VR}(z_k, t)$ from the pixels of horizontal scanline $y=y(z_k)$ using a 4-connected clustering algorithm to form text candidate regions in frame F_j where $t_k^{start} < t\,,F_j < t_k^{end}$. From each of the clustered regions, the top-most coordinate is computed and collected in an alignment histogram H_T, where the bin corresponds to the row number of the DC image as illustrated in Figure 4. The H_B is computed in the same way using the bottom-most coordinates of each region. We declare the existence of an upper boundary B_T of text caption if at least 50% of the elements in H_T are contained within three or fewer adjacent histogram bins. The lower boundary B_B is computed in the same way using H_B. The height of the localized text caption can thus be obtained. To find the width of the caption text, the regions with width longer than 1.5*height are firstly discarded. From the final set of regions, the following criterion is used to merge regions corresponding to characters to obtain the width of the text caption:

- Two regions, A and B, are merged if gap between A and B is less than 3 times the height.

We can thus verify whether $LINE_K$ is formed due to text caption and if so localize text caption, which appears along the duration of $LINE_K$, which does not have to be verified again.

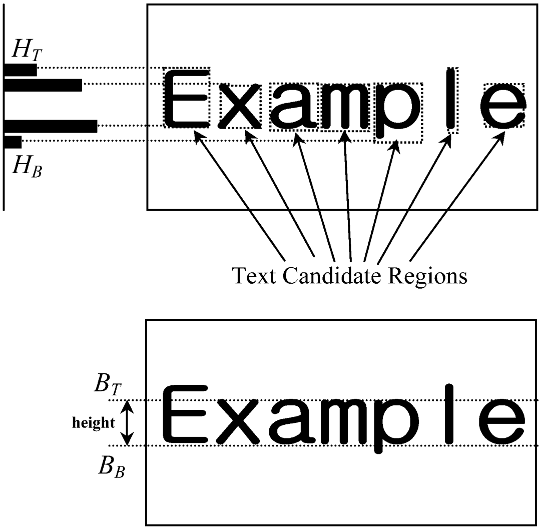


Fig. 4. Computation of the upper and lower boundary B_T and B_B of text caption through alignment histogram H_T and H_B.

Since several text lines can be formed due to the same text caption, the whole text localization process is omitted when $LINE_k$ and its corresponding horizontal scanline y= y(z_k) intersects with any text caption localized by any of the previous text line. Figure 6 shows an example of the localized text caption. The usefulness of this text caption localization stage is that it is inexpensive and fast, robustly supplying bounding boxes around text caption along with their temporal information.

4 Experimental Results

4.1 Environment of the Experiment

To evaluate the performance of the proposed method, we have tested it on various types of MPEG video clips consisting of 1) a news broadcast clip (14m 52s), which covered a variety of events including outdoor and newsroom news programs, and weather forecast, 2) a sports clip of golf lesson (37m 21s), and baseball (22m 4 s), 3) a commercial clip (7m 35 s), which contains various embedded captions and credits.

4.2 Performance Evaluation of Proposed Algorithm

Table I shows the results of the proposed algorithm. The second row of Table I gives the total number of text caption present in each category. The next row is the count of the correctly identified text caption. The total number of false positives and false negative is stated in the next two rows. Finally the recall and precision in each case is stated.

Our proposed text caption localization has an overall average recall of about 80% and a precision of 86%.

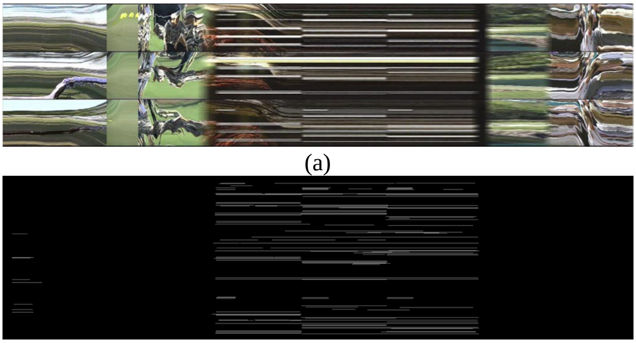
4.3 Computational Time of Proposed Algorithm

The processing speed of the proposed caption localization method is fast since it only works on few of the pixels sampled from the entire video in compressed domain compared to the conventional approaches operated by using the entire pixels of a video. Table 2 shows the processing time of each stage using Pentium III-500Mhz. It took approximately 7 minutes to produce the visual rhythm of the video clips corresponding to a total length of approximately 1 hour and 20 minutes. From the visual rhythm of the video clips, it took about 22 seconds to detect potential text frame subject to text caption localization as the final result of the text frame detection stage.

From the detected text frames, it took approximately a total of 2 minutes to locate text caption from the detected frames. Thus the whole process time took about 9 minutes.

5 Conclusions

The proposed algorithm on localizing text caption proved to be very fast by using text caption characteristics on visual rhythm. Moving text caption and text caption embedded on locations other than the assumed locations, resulted in a rather low average recall rate of 80% since our proposed algorithm locates only static text caption located on assumed locations. It took 9 minutes to localize text caption and its temporal duration, for a 1 hour and 22 minutes worth of video clip. This includes the construction time of visual rhythm, which can be used to detect scene changes for video indexing with very little processing. The proposed method also reduces the time for OCR since identical text caption appearing along consecutive frames are not fed into OCR repeatedly.


(a)

(b)

Fig. 5. Characteristics of text caption on visual rhythm: (a) Visual rhythm of video material (b) Text lines representing text caption

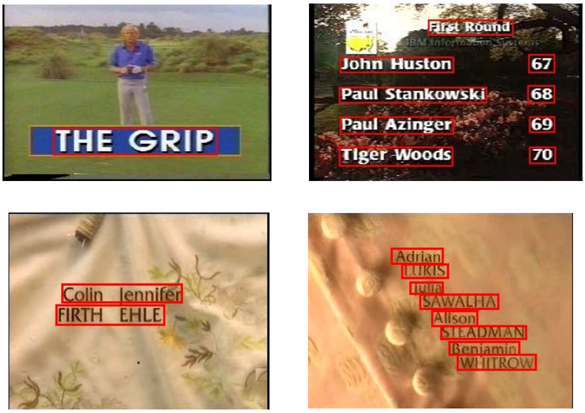


Fig. 6. Results of text caption localization

Table 1. Recall and precision for text caption localization

Video Type	News	Sports	Commercials
Distinct text caption	55	302	37
True Pos	44	241	30
False Pos	8	40	4
False Neg	11	61	7
Recall(%)	80.0	79.8	81.1
Precision(%)	84.6	85.8	88.2

Table 2. Execution time of visual rhythm construction, text frame detection and text caption localization.

Video Type	News	Sports	Commercials	Total
Duration	14m 52s	59m 25s	7m 35s	1h 21m 52s
Visual Rhythm	1m 16s	5m 6s	38s	7m
Detection Time	3s	17.21s	1.42s	21.63s
Localization Time	16s	1m 12s	8s	1m 36s
Total Processing Time	1m 35s	6m 35.21s	47.42s	8m 57.63s

References

1. Ohya, J., Shio, A., Akamatsu, S.: Recognizing Characters in Scene Image. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16. (1994) 214-220
2. Haupmann, A., Smith, M.: Text, Speech, and Vision for Video Segmentation: The Informedia Project. AAAI Symposium on Computational Models for Integrating Language and Vision, (1995)
3. Shim, J., Dorai, C., Bolle, R.: Automatic Text Extraction from Video for Content-Based Annotation and Retrieval. IEEE International Conference on Pattern Recognition, Vol. 1. (1998) 618-620
4. Wu, V., Manmatha, R., Riseman, E.: Finding Text in Images. Proceedings of the 2nd ACM International conference on Digital Libraries (1997) 3-12
5. Lienhart, R.: Automatic Text Recognition for Video Indexing. Proceedings of ACM Multimedia (1996) 11-20
6. Li, H., Doermann, D., Kia, O.: Automatic Text Detection and Tracking in Digital Video. IEEE Transactions on Image Processing, Vol. 9. (2000) 147-156
7. Sato, T., Kanade, T., Hughes, E., Smith, M., Satoh, S.: Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption. ACM Multimedia Systems, Vol. 7 (1998) 385-394
8. Yeo, B.L., Liu, B.: Visual Content Highlighting Via Automatic Extraction of Embedded Captions on MPEG Compressed Video. IS&T/SPIE/IS&T Symposium on Electronic Imaging: Digital Video Compression, (1996)
9. Zhong, Y., Karu, K. Jain, A.: Automatic Caption Localization in Compressed Video. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22. (2000) 385-392
10. Zhang, Y., Chua, T.: Detection of Text Captions in Compressed Domain Video. Proceedings of Multimedia Information Retrieval ACM Multimedia. (2000) 201-204
11. Yeo, B.L., Liu, B.: Rapid Scene Analysis on Compressed Video. IEEE Transactions on Circuit and Systems for Video Technology, Vol. 5. (1995) 533-544
12. Kim, H., Lee, J., Song, S.M.: An Efficient Graphical Shot Verifier Incorporating Visual Rhythm. Proceedings of IEEE International Conference on Multimedia Computing and Systems. (1999) 827-834
13. Song, J., Yeo, B.L.: Spatially Reduced Image Extraction from MPEG-2 Video: Fast Algorithms and Application. Proceedings of SPIE Storage and Retrieval for Image and Video Database VI, Vol. 3312. (1998) 92-107