THAM 15.3

REGION-BASED OBJECT TRACKING FOR MULTIPOINT VIDEO CONFERENCING USING WAVELET TRANSFORM

Kai-Tat Fung, Ngai-Fong Law and Wan-Chi Siu
Centre for Multimedia Signal Processing,
Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

ABSTRACT

The main features of our proposed wavelet-based video coder include: 1) a user-specified region of interest selection as to which the region can be changed by the user at any I-frame; 2) a dynamic region tracking technique by which the video is tracked and updated according to motion activity and 3) an adaptive bit allocation that allows the user to specify the relative quality between the foreground and the background. This architecture guarantees a high video quality in the region of interest while reducing the overall bit rate and the computation time. Experimental results confirm that the approach produces a good video quality even under low bit rates.

1. INTRODUCTION

With the advance of video compression and networking technologies, multipoint video conferencing becomes popular in the consumer market [1]. Most video conferencing systems use DCT-based encoders. A good performance can be achieved with a large bandwidth [2]. However, under low bit rates, the DCT-based encoder exhibits visually annoying blocking artifacts. Recently, wavelets have been used in internet applications. The major advantage of using a wavelet is its high quality and the absence of blocking artifacts when compared to the conventional video encoder [3-4]. Although a wavelet-based coder can achieve a good quality, its computational speed is an area of concern. A way to speed up the computation is to explore the fact that the various regions in an image are not of equal importance. This concept has been adopted in dynamic bit allocation and frame-skipping technique [5].

In this paper, a new region-based video coder architecture is proposed to achieve a good video quality with a low complexity. The proposed video coder is based on the adaptive region-based updating technique by which the video is updated according to the motion activity. This architecture allows a high quality video in the region of interest while reducing the overall bit rate and computation time. Since the user might be in fast motion when active, a simple and fast object tracking technique is proposed to locate the region of interest. This approach produces a good video quality even under low bit rates.

2. THE PROPOSED ARCHITECTURE

Figure 1 shows the system architecture for the proposed video coder in multipoint video conferencing. Since the video encoder is wavelet-based, blocking artifacts are avoided. Our proposed region-based video encoder has two major features:

1. selection of the region of interest; and
2. adaptive bit allocation for the foreground (region of interest) and the background.

The purpose of region selection is to identify the region of interest in an image, e.g. the speaker's face in video conferencing. This region is updated automatically by tracking the object's motion. The wavelet-based coder is then applied separately to the foreground and background. Because the size of the region of interest is small, the computation time could be reduced significantly. Adaptive bit allocation is then performed. It makes sure that the video quality of the foreground is always better than that of the background. This is particularly important for unstable networks or low bit rate applications.

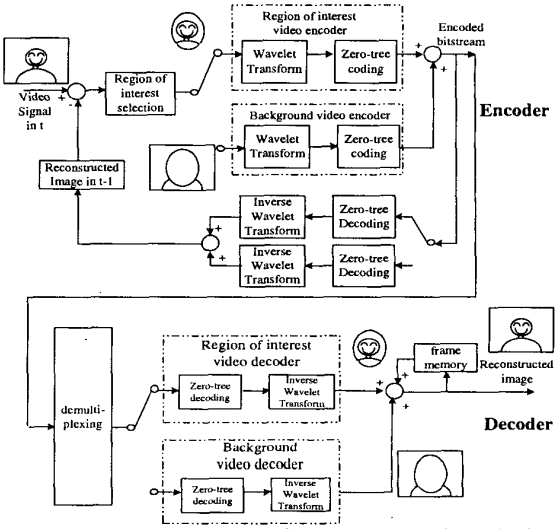

Figure 1: The system architecture for the proposed video coder in multipoint video conferencing.

2.1 Region of interest selection

Our proposed system allows the user to specify the region of interest so as to define the foreground and the background initially. If the user does not specify the region, the center region will be used (see Figure 2). In any I-frame, the user can change the region of interest which makes the video conferencing interactive. In subsequent frames, the intelligent video conferencing system calculates the difference between the current and the previous frames in order to reduce the temporal redundancy. This difference frame contains information about the motion of the user. In most video conferencing applications, the background is stationary and the difference frame shows only the changed region. Therefore, instead of coding the whole difference frame, only the region that contains large motions needs to be encoded. This reduces the overall bit rate and the encoding time while maintaining a good video quality even under low bit rates. We further propose in this paper to make an analysis of the histogram of the difference frame, due to its simplicity and invariance to rotation and translation. The matching criterion is defined as,

$$\sum_{i=0}^{255} [H_{defined}(i) - H_{neighbour}(i)] \qquad (1)$$

0-7803-6622-0/01 $10.00 © 2001 IEEE

where $H_{defined}(i)$ and $H_{neighbour}(i)$ represent respectively the histogram information in the defined and the neighboring regions. A search is carried out only around the center checking point as shown in Figure 2a. By using the histogram [6], instead of the minimum absolute difference as the matching criterion, a fast region tracking with high accuracy can be achieved.

The best match is obtained for a search range when the matching criterion defined in eqn.1 is the minimum. If the minimum is found at the center, the procedure stops. Otherwise, further search is conducted around the point where the minimum has just been found. The procedure continues until the winning point becomes a center point of the checking block or when the checking block hits the boundary of the predefined search range. In a practical situation, the interested region can be easily tracked as shown in Figure 2b. In summary, the proposed intelligent video conferencing system allows the user to select a region of interest and track the region using a fast algorithm.

Figure 2: The proposed searching technique, (a) the region of interest defined by the user and (b) the tracked object in the subsequent frame.

2.2 Adaptive Bit Allocation

Once the region of interest is defined, different video quality for the foreground and the background can be obtained by applying the wavelet transform and the zero-tree coding separately to the region of interest and the background. Therefore, different number of bits can be allocated in different regions so that a good video quality for the foreground can always be guaranteed. Also, the small region of interest can greatly reduce the processing time in the wavelet-based coder. The percentage of bits allocated to the region of interest is specified by the user. Thus the user can control the video quality dynamically.

3. EXPERIMENTAL RESULTS

The proposed encoder for the multipoint video conferencing system is tested for the 64kbit/sec case. Figure 3 shows a comparison between our proposed system and the conventional DCT-based system. The overall performance is shown in Figure 4. Although the background has a lower PSNR as compared to the conventional approach, the foreground has a much higher PSNR. Moreover, the subjective performance is much better than the conventional approach and the blocking artifacts are avoided. In fact, the subjective superiority is even more profound at low bit rates. In this case, the blocking artifacts associated with the DCT-based encoders are severe. However, by using our proposed algorithm, the quality in the foreground can still be maintained. Besides, the proposed video conferencing system has an improvement factor of about 2 to 3 times as compared to the case when the wavelet transform is applied to the whole image.

Figure 3: Reconstructed frames from (a) our proposed and (b) the DCT-based encoders.

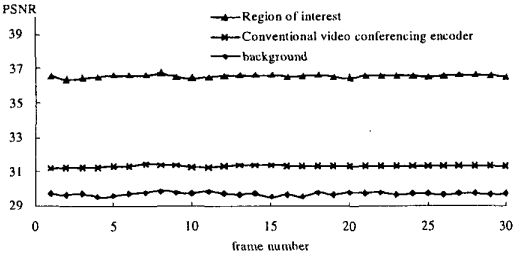

Figure 4: A comparison of the PSNR in different frames between our proposed encoder and the DCT-based encoder.

4. CONCLUSIONS

A region-based video encoder for multipoint video conferencing is proposed. The proposed architecture consists of region of interest selection, wavelet-based encoders for the foreground and the background, motion-based region updating steps and adaptive bit allocation strategy. To make the system interactive, the user can specify the region of interest at any I-frame as well as the relative quality between the foreground and the background. Experimental results confirm that our proposed method produces a good video quality even under low bit rates for real-time video conferencing applications.

5. ACKNOWLEDGEMENTS

This work is supported by the Centre for Multimedia Signal Processing, the HK Polytechnic University. K.T. Fung acknowledges the research studentships provided by the University and N.F. Law is grateful for the support she receives from the University under its research fellowship scheme.

6. REFERENCES

[1] H.T Chen, P.C. Wu, Y.K Lai and L.G. Chen, 'A multimedia video conference system: using region base hybrid coding', IEEE Trans. on Consumer Electronics, Vol. 42, No.3, 1996, pp.781 -786.

[2] Y.H. Chan and W.C. Siu, 'General approach for the realization of DCT/IDCT using convolutions', Signal Processing, Vol. 37, No.3, 1994, pp.357-364.

[3] N.F. Law and W.C. Siu, 'Successive Structural Analysis Using Wavelet Transform for Blocking Artifacts Suppression', revised version submitted to Signal Processing, Jan 2001.

[4] N.F. Law and W.C. Siu, 'Progressive Image Coding based on Visually Important Features', Vol. II, ICIP, 1999, Japan, pp.362-366.

[5] Kai-Tat Fung, Yui-Lam Chan and Wan-Chi Siu, 'Low-Complexity and High Quality Frame-Skipping Transcoder', paper accepted, to be published on proceedings, ISCAS'2001.

[6] M.J. Swain and D.H. Ballard, 'Color Indexing', International Journal of Computer Vision, Vol. 7, No.1, 1991, pp.11-32.