Technical Report 2003-01-01
Internetworking and Media Communications Research Laboratories
Department of Computer Science, Kent State University
http://medianet.kent.edu/technicalreports.html

Contour Approximation can Lead to Faster Object Based Transcoding with Higher Perceptual Quality

Javed I. Khan and Oleg Komogortsev
Media Communications and Networking Research Laboratory
Department of Math & Computer Science, Kent State University
233 MSB, Kent, OH 44242
javed@kent.edu

1 Abstract

Object aware video rate transcoding can significantly improvement the perceptual quality of relatively low bit-rate video. However, precise object detection in arbitrary video scene is computationally extremely challenging. This paper present an interesting experimentation with live eye-gaze-tracker which suggests that object detection particularly for perceptual encoding may not have to be precise. Indeed, intelligent approximation can not only reduce the complexity of the detection process, but also result in improved perceptual quality and yield very fast transcoding algorithms.

Key words: transcoding, perceptual encoding.

2 Introduction

Video rate transcoding is increasingly gaining importance in recent years. The asymmetry in the Internet capacity- particularly at the egress networks is growing dramatically. The emerging digital video standards such as DTV or HDTV will bring an enormous flux of high quality video content. However, the relatively differential of bandwidth at network edges and the advent of small devices (such as Personal Digital Assistant) seems to indicate that in near future the Internet applications have to deal with increased bandwidth asymmetry. Consequently, there will be increased need higher video transcoding ratio. Most of the current video transcoding techniques are based on frame wide requantization [5,7,9,10,11,12]. Unfortunately, frame-wide requantization very fast degenerate the perceptual quality of video. Research in first stage coding has already shown that object based encoding can play an increasingly important tool for creating perceptually pleasant video at lower rates [1,2,4,6]. MPEG-4 has been proposed to transport object-based coded video stream. However, conversion of a regular video to an object based stream is computationally challenging. [3,4] because of the high computational complexity of object detection. Thus live stream transcoding is still very difficult. It seems the first generation MPEG-4 systems will see most of its application in computer model generated synthetic video (where objects are already given), or for small format known content video [3] (such as head and shoulder video).

2.1 Related Work

Among the latest methods employed for object detection in video, Ngo et. al. [8] described object detection based on motion and color features using histogram analysis. This technique could process less than 2 frames in one second. Unfortunately, many of the other techniques presented such as [15] did not provide evaluation of time performance. However, it depends on even more involved image processing methods such as active contour model which spends considerable effort to determine the shape boundary, and is thus likely to be slower. More recently some compressed domain techniques have been suggested such as by Wang et al [14]. This system achieved about 0.5 sec/frame for the CIF size on a Pentium III 450 MHz. It should be noted that the **stream transcoding** scenario has some notable difference from first stage video encoding. First of all, in transcoding object detection has to be performed extremely fast at the rate of the stream. Secondly, the transcoder receives an encoded video stream. An unprocessed input stream seldom contains pixels level information rather contains highly structured coded and refined information such as motion vectors, thus access to pixel level information means severe computational overhead. While, in first stage encoding the opposite is true. A stream transcoder thus must take advantage of the scenario for being effective.

2.2 Computational Complexity

It seems, except from the compressed domain techniques, most of the object detection techniques tried in video have been derived from image level algorithms

and targeted towards scene comprehension type applications. Current approaches, even when it uses compressed domain fields, mixes pixel level data in analysis for boundary approximation. Thus most are too slow to meet the need of live perceptual stream transcoding. It seems a major source of computational burden originates from contour estimation. While precise contour detection has been considered as a critical part of scene interpretation and image understanding research is it possible they have less importance in perceptual encoding? To answer the question we have recently performed a live eye-gaze study to observe the perceptual effectiveness of several fast algorithms for perceptual coding incorporating contour approximation of various degrees.

2.3 Perceptual Quality

Over the years eye-gaze research has shed important light about visual perception. The retinal distribution of photoreceptor cells is highly non-uniform. Correspondingly, only about 2 degree in our about 140 degrees vision span has sharp vision [1,2]. Scientists have identified several intricate types of eye movements such as drift, saccades, fixation, smooth pursuit eye-movement, involuntary saccades. The quality perception is known to be highly correlated with the image quality around the fixations. Visual sensitivity reduces exponentially with eccentricity from the fovea.

In this paper we will therefore present a direct comparison of these techniques based on live eye-gaze. The study suggests indeed approximation not only significantly reduces the complexity of the object detection process, but also results in improved perceptual quality. The schemes can provide dramatic speed improvement in MPEG-2/MPEG-2 or MPEG-2/ MPEG-4 object based perceptual transcoding.

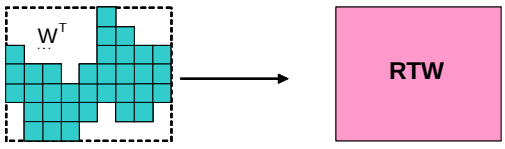

Fig.1 Rectilinear W^{TW} approximation

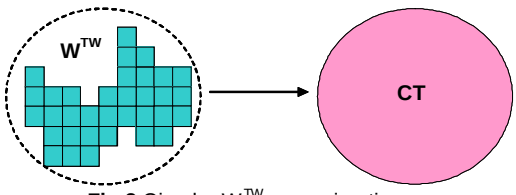

Fig.2 Circular W^{TW} approximation

3 Contour Approximation Appraoch

We started with a fast base BOF algorithm [13]. The algorithm itself is not the focus of this paper. However, we will describe it here briefly. This is a strictly compressed domain method which tracks object with only motion vector analysis. It estimates the projection of various scene objects, background motion, and camera motions on the motion vectors in the coded stream. Thus by observing the motion vectors it then detects and tracks video objects. It is important to note that motion vectors are not simply codified information about block's motion, as it may appear at first glance. Rather these also contain highly refined color, texture and shape information. However, what it cannot contain is precise shape information beyond block boundaries. Using Kalman filter prediction this fast algorithm automatically detects, and tracks the region covered by the scene objects for subsequent perceptual encoding. A detail of the algorithm is given in [13]. To study the impact of contour approximation we incorporate three contour approximation techniques with this algorithm. To measure the impact of the approximations on perceptual efficiency, we then play the live video to a subject and directly observe the eye gaze fixations on the video frames.

3.1 Approximated Tracking Window

The BOF algorithm for each object provides an object window called W^{TW}. This window tracks the moving blocks corresponding to a moving object in the scene. We consider three versions of the algorithm. The first is the raw object window and the other two approximations are explained below.

3.2 Rectilinear Approximation

The first is the rectangular approximation. All the points in W^{TW} are sorted and the window W^{RTW} is constructed using the min max corners. W^{RTW}= {(x_{min}, y_{max}), (x_{min}, y_{min}), (x_{max},y_{max}), (x_{max}, y_{min})}, where x_{max} = max {x_i}, x_{min}=min(x_i), y_{max}=max{y_i}, y_{max}=max{y_i}, where: all $x_i, y_i \in W^{TW}$.

3.3 Circular Approximation

The other form of approximation of W^{TW} is circular approximation shown in Figure 3.4.2. This approximation is built in the following way. Let (x_i, y_i) is the macroblock center. We determine:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_i)^2}$$

Let

$$D_{km} = \max\{D_{ij}\}$$

Where, $MB_k[x_k, y_k]$ and $MB_m[x_m, y_m]$ are the ones which have the biggest distance from each other. Let there distance is $D_{km} = \sqrt{(x_k - x_m)^2 + (y_k - y_m)^2}$, then W^{CTW} is defined as a circle with radius R=0.5D, and center at $(x_c = \frac{x_k + x_m}{2}, y_c = \frac{y_k + y_m}{2})$. CTW gaze containment varies from 70% to 85%. As we will see in most cases, it performs much better than TW gazes containment.

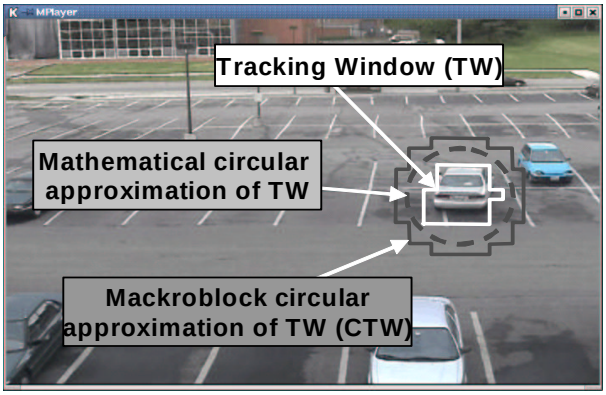

Fig. 3. Video 1. Frame number 233. Circular tracking window build explanation scheme.

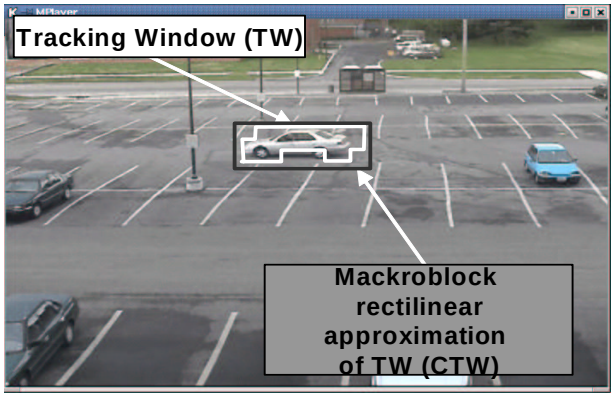

Fig. 4. Video 1. Frame number 1343. Circular tracking window build explanation scheme.

4 Experiment

4.1 Setup

We have implemented the system with integrated Applied Science Laboratories High speed Eye tracker Model 501. The eye position video capturing camera worked at the rate of 120 samples per second. For this experiment we defined fixation when the eye does not move more than 1 degree in 100msec. We modified the Percept Media Transcoder in a way that it can generate Reflex Window, track the object and use this both methods for creating Perceptual Object Window. All are videos were 720x480 and were captured with Sony TRV20 digital camera at high resolution with more than 500 lines at a frame rate of 30 frames per second. Number of frames per GOP is 15. Number of "B" frames between any give two "P" frames is two. The video was projected on the wall of the dark room. The projected physical dimensions of the image are width 60 inches, height 50 inches, and the distance between subject eyes and the surface of the screen is about 100-120 inches.

4.2 Sample Shots

First we will share some actual example. Fig-3 and Fig.4 show some actual test shots from the live system. These figures show sample frames 233 and 1343. The video has original encoding at 10 Mbps. Fig-3 and Fig.4 show the rectangular and circular approximation correspondingly. After approximation frames are perceptually encoded based on the corresponding approximation method. We reduced bit-rate about 10m times to 1 Mbps. In the rate reduction full resolution was maintained at the approximated window macro-blocks. The MPEG-2 TM-5 rate control was used to determine the quantization of the remaining blocks. The actual perceptually video samples as per all three window models, including the originals can be

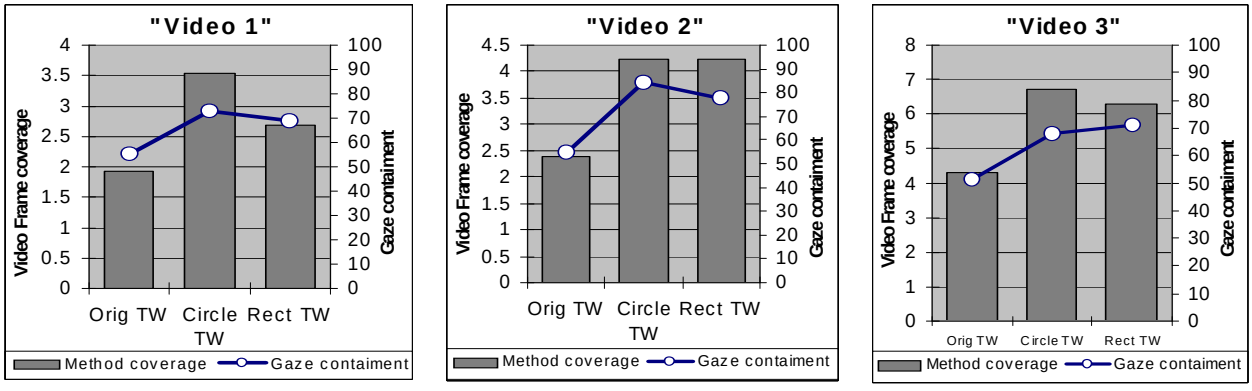

Fig. 5. Gaze containment and video frame coverage by different TW approximation methods and video samples.

obtained for direct visual appreciation from [16].

4.3 Video Sets

For the system we have tested a large number of video scenes. Perceptual encoding is highly dependent on the video content. Therefore, we avoid providing any 'average' performance. Rather we use case videos each carefully selected to offer a special challenge to the system. In this paper we present three cases. (a) "Video 1" contains car driving in a parking lot. The object speed is smooth and continuous. (b) "Video 2" has two radio controlled toy cars moving at different speeds with rapid unpredictable movements. In this video we asked subject to concentrate on just one car. (c) "Video 3" has two relatively close up toy cars offering much larger area of focus. Cars move in different directions inconsistently. Subject is asked to concentrate only on one car.

4.4 Containment Efficiency

The ideal perceptual encoding requires all the fixations to be contained within the object window. Ideally, if all gazes are within the window then it is possible to design optimum perceptual encoder. Thus, we defined the quantity *gaze containment* as the fraction of gazes successfully contained within the window:

$$\xi = \frac{\left|S^{w}(t)\right|}{\left|S(t)\right|} \qquad \ldots(5.1)$$

Where, S(t) is the entire sample set and $S^{W}(t) \subseteq S(t)$ be the sample subset contained within the object tracking window W(t).

The right y-axes of Fig 5(a), (b) and (c) show the results of gaze containment for W^{TW} W^{CTW} and W^{RTW} for the three videos. Compared to the strict object boundary based TW, both of these approximations increases the containment significantly. For Video 1, containment increases from 49% to about 70% for rectilinear approximation and to about 89% for circular approximation. Same tendency we can see for Video 2 and Video 3. It is interesting to note that the base BOF algorithm itself uses block approximation. Thus a true object contour is expected to contain less than 50% of the eye-gazes!

4.5 Coverage Efficiency

With larger visual windows more gazes can be contained, however, there will not be any perceptual redundancy to extract. Therefore, we were also curious to see how tight was the windows. We define a second performance parameter called "*perceptual coverage*".

$$\chi(t) = \frac{\left|\Delta(W(t) \cup F(t))\right|}{\left|\Delta(F(t))\right|} \qquad \ldots.(5.2)$$

Where, F(t) is the size of the total viewing frame, and W(t) is the method provided perceptually encoded window (delta for area or volume).

The bars plotted on the left y-axes of Fig-5(a), (b) and (c) show the coverage factor for all the cases. It can be noted that the approximated window increases very slightly--1.8% to 2.7% of the frame for Video 1, or 4.2% to 6.8% for the most difficult video 3.

It seems that circular approximation slightly outperforms rectilinear one in number of gazes contained and the increase in the area size for the circular approximation is negligible comparing to the total frame size.

5 Conclusions & Current Work

Many of contemporary object-based perceptual coding techniques are based on the implicit assumption that the object area inside the object contour should be coded with higher resolution. The assumption is perhaps imperfect. It is highly likely that mental process of visual perception drives eye to scan areas beyond particularly the area slightly outside. The high emphasis on object boundary has lead to many involved schemes for video object detection. It seems the efforts spent in exact contour extraction, is perhaps counter productive in perceptual coding. The proposed approximations can lower cost and improve performance effective. Without significant increase of the perceptual coverage we were able to achieve around 90% gaze containment in the proposed simple approximations. Clearly, other approximations techniques to suit the video coding constraint can also be designed.

This research has been supported by the DARPA Research Grant F30602-99-1-0515.

6 References:

[1] Z. Wang, and A. C. Bovik, "Embedded foveation image coding", IEEE Trans. Image Proc., Oct. 2001

[2] Z. Wang, Ligang Lu, and Alan C. Bovik, "Rate scalable video coding using a foveation-based human visual system model", ICASSP 2001.

[3] Aizawa, K., H. Harashima, & T. Saito, Model-based Image Coding for a person's Face, Image Commun, v.1, no.2, 1989, pp 139-152.

[4] Hotter, M., & R. Thoma, Image Segmentation based on Object-Oriented Mapping Parameter Estimation, Sinal Process., v. 15, 1998, pp.315-334.

[5] Keesman, Gertjan; Hellinghuizen, Robert; Hoeksema, Fokke; Heideman, Geert, Transcoding of MPEG bitstreams Signal Processing: Image Communication, Volume: 8, Issue: 6, pp. 481-500, September 1996.

[6] Javed I. Khan, Q. Gu, Network Aware Symbiotic Video Transcoding for Instream Rate Adaptation on Interactive Transport Control, IEEE Int. Symp. on Network Computing and Applications, IEEE NCA' 2001, Oct, 8-10, 2001, Cambridge, MA, pp.201-213.

[7] Youn, J, M.T. Sun, and J. Xin, "Video Transcoder Architectures for Bit Rate Scaling of H.263 Bit Streams," ACM Multimedia 1999', Nov.,

1999. pp243-250.

[8] Ngo, Chong-Wah, Ting-Chuen Pong and Hong-Jiang Zhang, "On clustering and retrieval of video shots", ACM Multimedia 2001, Oct., 2001. pp51-60.

[9] U. Chong and S. P. Kim, Wavelet Trancoding of block DCT-based images through block transform domain processing, SPIE Vol. 2825, 1996, pp901-908.

[10] Niklas Björk and Charilaos Christopoulos, Video transcoding for universal multimedia access; Proceedings on ACM multimedia 2000 workshops, 2000, Pages 75 - 79

[11] J. Youn, M.T. Sun, and C.W. Lin, "Motion Vector Refinement for High Performance Transcoding," IEEE, Transactions on Multimedia, Vol. 1, No. 1, pp.30-40, March 1999.

[12] P. Assuncao and M. Ghanbari, "A frequency-domain video transcoder for dynamic bit rate reduction of MPEG-2 bit streams," Trans. On Circuits Syst. Video Technol., vol. 8, no. 8, pp. 953-967, 1998.

[13] Khan Javed I. Zhong Guo, & W. Oh, Motion based Object Tracking in MPEG-2 Stream for Perceptual Region Discriminating Rate Transcoding, Proceedings of the ACM Multimedia, 2001, October 2001, Ottawa, Canada, pp572-576.

[14] Wang R., H.J. Zhang and Y.Q. Zhang. A confidence measure based moving object extraction system built for compressed domain. 2000.

[15] Kuehne, Gerald, Stephan Richter and Mark Beier, "Motion-based segmentation and contour-based classification of veideo objects", ACM Multimedia 2001, Oct., 2001. pp41-50.

[16] O. Komogortsev, & Javed I. khan, TR2002-06-01 Perceptually Encoded Video Set from Dynamic Reflex Windowing, Technical Report, Kent State Univrsity, Jun 2002, [http://medianet.kent.edu/techreports/TR2002-06-01/TR2002-06-01-videoset-KK.htm.