MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Object-Based Coding for Long-Term Archive of Surveillance Video

Anthony Vetro Tetsuji Haga Kazuhiko Sumi
Huifang Sun

TR-2003-98 July 2003

Abstract

This paper describes video coding and segmentation techniques that can be used to achieve significant increase in storage capacity. Specifically, we examine the possibility to use object-based coding for efficient long-term archiving of surveillance video. We consider surveillance systems with many camera sources in which we are required to store several months of video data for each source, thus storage capacity is a major concern. The paper considers several automatic segmentation algorithms. With each algorithm, we will analyze the shape coding overhead and implication on overall storage requirements, as well as the effect each algorithm has on the reconstructed quality of frames. Additionally, this paper reviews techniques to dynamically control the temporal rate of objects in the scene and perform bit allocation. Experimental results show that up to 90% savings in storage can be achieved with the proposed method compared to frame-based video coding techniques. The cost for this savings is that the accuracy of the background is compromised; however, we feel that this is satisfactory for the application under consideration.

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2003
201 Broadway, Cambridge, Massachusetts 02139

Proc. IEEE Int’l Conference on Multimedia and Expo, July 2003.

OBJECT-BASED CODING FOR LONG-TERM ARCHIVE OF SURVEILLANCE VIDEO

Anthony Vetro[*], *Tetsuji Haga*[**], *Kazuhiko Sumi*[**], *Huifang Sun*[*]

[*] MERL – Mitsubishi Electric Research Laboratories, Murray Hill, NJ, USA
[**] Mitsubishi Electric Corporation, Advanced Technology R&D Center, Tsukaguchi, Japan

ABSTRACT

This paper describes video coding and segmentation techniques that can be used to achieve significant increase in storage capacity. Specifically, we examine the possibility to use object-based coding for efficient long-term archiving of surveillance video. We consider surveillance systems with many camera sources in which we are required to store several months of video data for each source, thus storage capacity is a major concern. The paper considers several automatic segmentation algorithms. With each algorithm, we will analyze the shape coding overhead and implication on overall storage requirements, as well as the effect each algorithm has on the reconstructed quality of frames. Additionally, this paper reviews techniques to dynamically control the temporal rate of objects in the scene and perform bit allocation. Experimental results show that up to 90% savings in storage can be achieved with the proposed method compared to frame-based video coding techniques. The cost for this savings is that the accuracy of the background is compromised; however, we feel that this is satisfactory for the application under consideration.

1. INTRODUCTION

Most video compression schemes operate on frames of the video sequence, where hybrid motion-compensated DCT coding is employed. The compression efficiency is typical measured by the quality that can be achieved for a given bit-rate, or conversely, the bit-rate that can be achieved at a fixed quality. To achieve a fixed quality, a constant quantization parameter can be used to compress every frame. Quality is usually gauged with exact metrics such as the mean-squared error (MSE) of pixel differences.

Object-based coding offers the flexibility to compress individual objects within the scene. As with frame-based coding, motion-compensated DCT coding is employed, but done so only for data contained within an object boundary, which implies that the object boundary must also be coded. MPEG-4 has defined tools for this purpose [1]. The flexibility that object-based coding provides is that the quality of each object need not be the same and also the temporal rate of each object can be different.

In prior studies of object-based coding [2], quality was measured by traditional frame-based metrics, such as MSE. The main drawback in measuring the quality this way is that it varying the temporal rate of objects in a scene would be discouraged due to the high overall penalty that would be incurred for objects coded at a lower temporal rate. For example, if the background of a scene that contains small local motions is coded with a lower temporal rate compared to the foreground, it is likely that pixel differences measured in the background will result in an overall decrease in measured quality, even if there is no perceptual difference, i.e., subjective quality is maintained. To realize the potential of object-based coding and fully utilize the flexibility that it offers, we must consider alternate means of judging the coding performance.

In a recent study on using background modeling and texture replacement and mapping for content-based coding [3], alternative metrics, such as the weighted signal-to-noise ratio and noise quality measure [4], were used to evaluate the effectiveness of the proposed techniques. The authors also rely on subjective quality assessment. This paper supports the notion shared by this paper that traditional distortion metrics are not suitable to evaluate the effectiveness of applying object-based coding and manipulation techniques.

This paper presents a surveillance application system that utilizes object-based coding techniques to achieve efficient storage of video content. In such a system, certain inaccuracies in the reconstructed scene can be tolerated, however subjective quality and semantics of the scene must be strictly maintained. As shown in Figure 1, the target of our system is for the long-term archiving of surveillance video, where several months of video content from multiple cameras would need to be stored.

The rest of this paper is organized as follows. In the next section, an overview of a pixel-based and a block-based segmentation algorithm is given. The impact of the accuracy of these segmentation algorithms on the reconstructed quality is discussed. In section 3, techniques for object-based coding in the target system are covered. In section 4, experimental results are provided and concluding remarks are given in section 5.

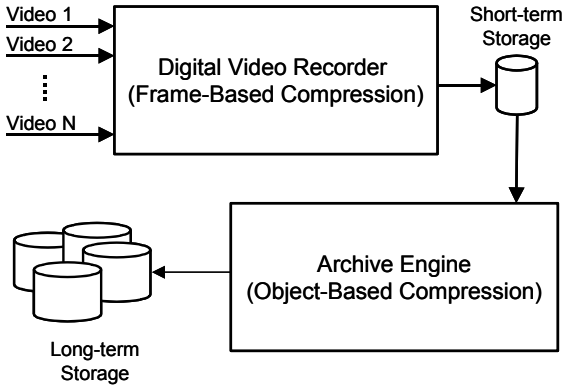

Figure 1 Application system for surveillance system employing object-based coding for long-term archive of video contents.

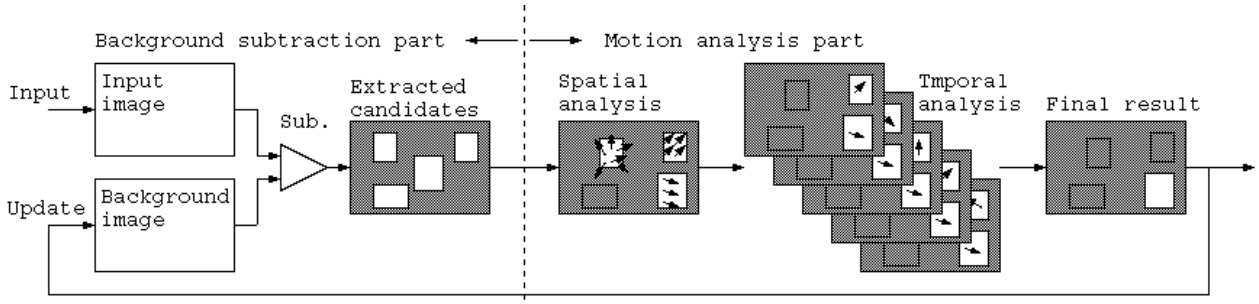

Figure 2 Flow of automatic object segmentation algorithm.

2. OBJECT SEGMENTATION

This work considers three automatic object segmentation algorithms. All of them consist of two series of processes as shown in Figure 2. The former part is the background subtraction and the latter part is the motion vector analysis. In surveillance video scenes, most of the picture typically belongs to the background, where the pixel value is easily estimated with the use of a statistical background model.

Recent works can deal robustly with even non-static background. However, the existing background models are not perfect as they are often faced with the trade-off of sensitivity between false positive and false negative. Our policy is to avoid miss detection in the background subtraction stage by controlling the false detection in the motion vector analysis stage.

2.1 Background Subtraction

The following three background models are considered in this work: a Mixtured Gaussian Model [5], a Non-parametric Background Model [6], and a Normalized Correlation Model [7]. These models are briefly described below.

With the Mixtured Gaussian Model, each pixel value of background model is represented by summation of 3 to 5 weighted Gaussians. Each weight is calculated according to the frequency with which the input data matches the Gaussian. All the means and the standard deviations are continuously updated in the exponential averaging manner. An input pixel whose value does not match within the any of 2.5 standard deviations of k Gaussians, is regarded to be a part of the foreground. This model is later referred to in the experimental results as *Pixel1*.

In the Non-parametric Background Model, each pixel value of background model is represented by the probability density function Pr(X). As Pr(X) indicates the probability with which the pixel of value X belongs to the background, extraction is done in one global threshold value TH in all over the image. Pr(X) is shaped as the average of N Gaussians estimated by pixel values and those deviations between consecutive values observed during the last N frames. This model is later referred to as *Pixel2*.

For the Normalized Correlation Model, an input picture frame is divided into small blocks with W by W pixels and the pixel data of each blocks is rescanned into one dimensional vector data with length of W^2. In the similar way, the reference image, which is the median of the training images, is split into small blocks and the pixel data is rescanned into one dimensional vector data. The background model is represented by block-wise mean and standard deviation calculated from normalized correlation result between the vector data from training image and that from the reference image. A block of input image is regarded to be a part of the foreground if the normalized correlation between its vector data and that from the corresponding block of the reference image does not match within the T standard deviations. The size of block W is 8 and the threshold T is empirically decided. This model is later referred to as *Block*.

The Mixtured Gaussian Model and the Non-parametric Background Model are robust to the bimodal background like swaying trees, The Normalized Correlation Model, to the lighting change, respectively.

2.2 Motion Vector Analysis

This part is common to all the three methods. To remove the surviving false positive errors, we focused on the motion of the detected region. The typical error factors in surveillance video scenes are lighting change, swaying branches or leaves, diffused reflection on the surface of water, etc. Most of these error factors have confused and discontinuous motion, whereas moving objects lumped together have uniform and continuous motion. We can extract them through analyzing the uniformity and the continuity of local motion inside of the changed region.

The algorithm of motion vector analysis is as follows. First, we place small blocks in the changed region given in the background subtraction process. Then, calculate the block correlation between adjacent frames. Next, we accumulate the correlation results to check the motion uniformity. The region whose accumulated correlation map has no distinct peak is removed as a background with confused motion or motionless light change. Finally, we extract the average motion in the remaining region and track them at a certain period of time, and again accumulate the cumulative correlation map to check the motion continuity. The region whose doubly accumulated correlation map has no distinct peak is removed as a background with random or repetitive motion.

3. OBJECT CODING

One of the major advantages of object-based coding is that each object can vary in its temporal quality. Varying the temporal rate for each object is expected to yield modest gains under certain conditions using traditional PSNR metrics, and extremely high gains if only subjective quality is being assessed. In this paper,

we depart from traditional PSNR metrics and mainly rely on subjective evaluation.

With traditional metrics in mind, models to estimate the expected rate and distortion of coding video with a dynamic temporal rate are required; such models have been presented in [8]. Simulation results have been shown for frame-based video coding, and for sequences with low to moderate motion, slight gains in PSNR were observed. Encoding and transcoding with variable temporal resolution on an object-basis was investigated in [9]. At this time, the major focus was still on minimizing exact pixel differences. One of the major problems encountered with object-based coding is the composition problem, i.e., the full background is not available and coding fast moving foreground objects with a different temporal rate than the background will cause holes in the composed scene. Fortunately, in the surveillance system under consideration in this paper, obtaining the full background without any foreground objects present is not a problem.

In this paper, a single background image is compressed using frame-based coding, and the sequence of segmented foreground objects are compressed using object-based coding. Both background and foreground are coded at a constant quality using fixed quantization parameters, and the background image is simply repeated for each reconstructed frame. In an actual system, it is expected that a preset criteria will be used to judge when the background image should be refreshed, if needed. For dynamic skipping of objects, the models and algorithms described in [8] may be used.

Performing the object-based compression in this way gives rise to differences in the background pixels, especially for outdoor scenes that have trees and objects that sway due to wind conditions or are subject to other weather conditions. It is critical, however, to have the correct segmentation of foreground objects. Assuming that one does, moving elements in the background for this application are simply noise and wasteful bits are spent in coding them. Using a still image background can be seen as a noise removal in which the semantics of the scene are still maintained, and improved coding efficiency can be achieved.

4. EXPERIMENTAL RESULTS

For the purpose of this paper, the test sequences used are relatively short in duration and an adaptive update of the background is not required. As mentioned in the previous section, we compress a single background image for each test sequence using MPEG-4 frame-based coding, and the sequences of segmented foreground objects using MPEG-4 object-based coding. Both background and foreground are coded at a constant quality using fixed quantization parameters (Q=10). In the object-based reconstructed frames, the still background image is simply repeated.

Sample reconstructed frames of the gunfestA sequence using frame-based and object-based coding are shown in Figure 3. In this test sequence, wind is blowing the striped curtain, tree branches and hanging ornaments; all of which are coded and accurately represented by the frame-based coding result. However, for the object-based coding result, these moving background elements are not recorded and only the moving foreground object is coded at each time instant. Semantically, however, these sample frames are equivalent.

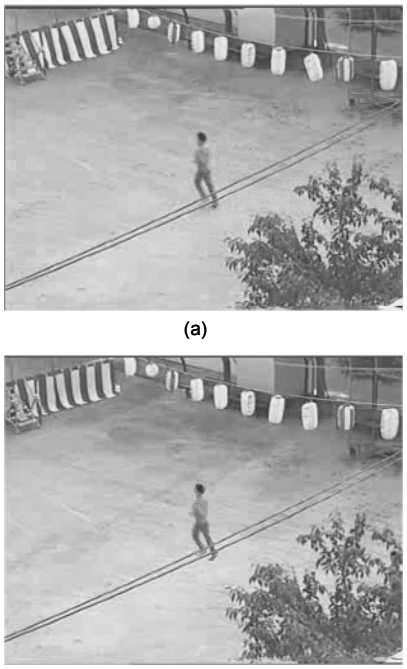

Figure 3 Sample reconstructed frame of gunfestA sequence. (a) Frame-based reconstruction, (b) Object-based reconstruction.

Given that this type of distortion is tolerable, we perform tests over a wider range of test sequences to understand the potential bit-savings of object-based coding compared to frame-based coding. Table 1 summarizes the storage requirements for 8 test sequences of varying duration and content. These sequences have been coded with the frame-based and the object-based coding. For the object-based results, four different types of segmentation map generation techniques are applied, including a manual segmentation and the 3 techniques described earlier in section 2. It is clear from this table that regardless of the segmentation used, object-based coding provides favorable savings in the bits required to store the compressed video sequences. The percentages of bit-saving over frame-based coding are provided in Table 2. While the amount of bit-savings has little dependence on segmentation algorithm, we can clearly see that it does however vary as a function of the sequence from as much as 62% with gunfestC to as much as 91% with gunfestA. On average, the bit-savings is approximately 78%. These differences are mainly due to the amount of background movement contained within the scene. It should be noted that gunfestC contains very little motion in the background, but savings are still realized. This indicates that even relatively small changes in the pixel intensities (which could even occur with indoor scenes subject to varying lighting conditions) create potential for bit-savings.

With regards to quality, the segmentation accuracy has the biggest impact on the perception of the reconstruction quality. The most significant problem occurs when the segmentation algorithm misses pixels that are part of the object, however it is also somewhat problematic when the algorithm detects background pixels as being part of the foreground. Using the manual segmentation results, such artifacts are not observed and

Table 1 Comparison of storage requirements in KB for frame-based and object-based coding. Background image for object-based coding results are included (gunfest: 9KB, san_Rain: 4KB, san_Tree: 4KB).

Sequence	gunfestA	gunfestB	gunfestC	gunfestD	gunfestE	gunfestF	san_Rain	san_Tree
Frame-Based	178	173	35	116	65	111	78	43
Object-Based (Seg: Manual)	17	18	13	18	15	19	22	10
Object-Based (Seg: Pixel1)	17	17	13	19	13	17	25	11
Object-Based (Seg: Pixel2)	17	17	13	19	13	17	31	18
Object-Based (Seg: Block)	13	17	13	16	14	18	25	12

Table 2 Percentage of bit-savings with object-based coding. Observe that savings are more dependent on the sequence characteristics than the particular segmentation used. Maximum savings with gunfestA: 91.01% (average over all segmentations). Minimum savings with gunfestC: 62.86% (average over all segmentations). Overall savings: 78.57% (average over all sequences and segmentations).

Sequence	gunfestA	gunfestB	gunfestC	gunfestD	gunfestE	gunfestF	san_Rain	san_Tree
Object-Based (Seg: Manual)	90.45%	89.60%	62.86%	84.48%	76.92%	82.88%	71.79%	76.74%
Object-Based (Seg: Pixel1)	90.45%	90.17%	62.86%	83.62%	80.00%	84.68%	67.95%	74.42%
Object-Based (Seg: Pixel2)	90.45%	90.17%	62.86%	83.62%	80.00%	84.68%	60.26%	58.14%
Object-Based (Seg: Block)	92.70%	90.17%	62.86%	86.21%	78.46%	83.78%	67.95%	72.09%

the reconstructed frames are quite acceptable. In our experiments, we have found that the non-parametric background model (pixel2) provides the most accurate segmentation results in terms of the reconstruction quality that it produces.

5. CONCLUDING REMARKS

This paper presented a surveillance system that utilizes object-based coding techniques for long-term archiving of video contents. In this system, we diverge from traditional MSE-like measures of quality and focus on maintaining the subjective quality and semantics of the original video. Several segmentation algorithms to identify the objects in the scene were described. In terms of detecting the object-of-interest in the scene, different performance is obtained by each algorithm. It was observed that the performance of the segmentation algorithm has the most significant impact on the reconstructed quality of the scene. Regardless of which segmentation algorithm was used, significant savings of up to 90% in the bits needed to store the coded contents was achieved. Future work will be concentrated on improving the accuracy of the automatic segmentation algorithms. Furthermore, with such an object-based system, the analysis of individual objects is possible, and indexing and annotation of the scheme may be done at this level.

REFERENCES

[1] ISO/IEC 14496-2:2001, "Coding of Audio-Visual Objects – Part 2: Visual," 2nd Edition, 2001.

[2] A. Vetro, H. Sun and Y. Wang, "MPEG-4 rate control for coding multiple video objects," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 186-199, Feb. 1999.

[3] A. Dumitras and B.G. Haskell, "A background modeling method by texture replacement and mapping with application to content–based movie coding," *IEEE Int'l Conf. Image Processing*, Rochester, NY, USA, Sept. 2002.

[4] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Processing*, vol. 9, no. 4, pp. 636-650, Apr. 2000.

[5] C. Stauffer and W.E.L Grimson, "Adaptive background mixture models for real-time tracking", *Proc. Computer Vision Pattern Recognition*, Fort Colins, CO, June 1999.

[6] A. Elgammal, D. Harwood, and L.S. Davis, "Nonpara-metric background model for background subtraction," *Proc. 6th European Conf. Computer Vision*, 2000.

[7] T. Matsuyama, T. Ohya, and H. Habe, "Background subtraction for non-stationary scenes", *Proc. 4th Asian Conference on Computer Vision*, pp.662-667, 2000.

[8] A. Vetro, H. Sun and Y Wang, "Rate-Distortion optimized video coding with frameskip," *IEEE Int'l Conf. Image Processing*, Thessaloniki, Greece, Sept. 2001.

[9] A. Vetro and H. Sun, "Encoding and transcoding of multiple video objects with variable temporal resolution," *Proc. IEEE Int'l Symp. Circuits and Systems*, Sydney, Australia, May 2001.