MODEL OF OBJECT-BASED CODING FOR SURVEILLANCE VIDEO

Yang Yu and David Doermann

Language and Media Processing Lab
Center for Automation Research
University of Maryland, College Park, MD 20742
email: {yyu,doermann}@umiacs.umd.edu

ABSTRACT

In this paper, we explore the model of potential savings of object-based coding for surveillance video. Moving foreground objects in stationary camera surveillance video are detected by a background subtraction technique [3] and encoded with MPEG-4 object-based coding. Experiment results show that compared with frame-based coding, object-based coding can achieve significant savings which are dependent on the video content. We further model the relationship of compression efficiency and the number and size of video objects using statistical learning method. Simulations show that the model is representative. The model can be used to predict the savings of object-based coding and select the coding methods for surveillance video.

1. INTRODUCTION

Today surveillance video systems are widely deployed to monitor activities at outdoor or indoor sites for detection of suspicious activity or site security. These systems produce a large quantity of video data which will be stored and then subjected to analysis or inspection. Efficient storage and access of such a large quantity of video data is an imminent problem to solve in current surveillance video systems. Most surveillance video systems employ frame-based coding such as MPEG-1, or MPEG-2. Recent video coding standards such as MPEG-4 [1] employ object-based coding. Object-based coding not only achieves higher coding efficiency but also enables video content access and interactivity.

In most surveillance video systems, stationary cameras are typically used. So background subtraction can be utilized to segment the moving foreground objects by building the model of the background and comparing the incoming new frames with the background model. As the background information is relatively unimportant, large distortion in the background can be tolerated and objected-based coding is employed to encode the foreground objects. In MPEG-4 object-based coding, hybrid motion-compensated DCT coding is still utilized to encode the texture of foreground objects. In addition shape information is needed to represent the binary mask of the encoded objects. With the change of the number and size of the video objects, the savings of object-based coding compared with frame-based coding where both foreground objects and background are encoded change. In this paper the potential of compression is studied when object-based video coding is employed in surveillance video. In [2] the storage saving of object-based coding was also noticed. However in this paper, for the first time, we derive the model of the compression efficiency with respect to the number and size of video objects using statistical learning method.

The paper is organized as follows. In section 2, MPEG-4 object-based coding is combined with background subtraction, and the potential of savings is explored. Section 3 proposes the scheme that gives the relationship model of the compression efficiency and the number and size of the foreground objects. Simulation results are given in section 4. And section 5 concludes the paper.

2. BACKGROUND SUBTRACTION AND OBJECT-BASED VIDEO CODING

Codebook-based background subtraction (CB-BGS) [3] is the background subtraction technique employed in this paper. CB-BGS adopts a quantization/clustering technique to construct a background model from long observation training sequences. For each pixel, it builds a codebook consisting of one or more codewords. Samples at each pixel are clustered into a set of codewords based on a color distortion metric together with a brightness ratio. Once the background model has been built, the foreground objects can be detected by comparing each pixel of the incoming frame with the codewords.

The output of CB-BGS is binary mask which is fed into MPEG-4 encoder and object-based video coding is applied to encode the foreground objects. As to the

background information, since it is unimportance and its large distortion does not influence the interpretation of the video content, only one frame of background is encoded by MPEG-4 simple profile which employs frame-based coding. The final encoded stream is the sum of the bit-streams of object-based coding and background information encoded by frame-based coding. Table 1 compares frame-based coding and object-based coding in terms of the number of bytes needed to encode three different surveillance video sequences. Here frame-based coding is implemented by MPEG-4 simple profile which employs the same technique in encoding the texture as object-based coding in MPEG-4. The quantization parameter QUANT is set to 10 in both cases. And we assume that the same quantization parameter gives approximately the same objective and subjective visual quality in foreground objects and distortion in background can be ignored.

Sequence Name	Frame-based Coding	Object-based Coding	Saving
Carleaving	188,943 Bytes	99,054 Bytes	47.6%
Exchange	311,606 Bytes	96,490 Bytes	69%
Wood	750,691 Bytes	91,025 Bytes	87.9%

Table 1. Comparison of storage needed for frame-based and object-based coding both with Quant = 10.

Table 1 shows that significant savings can be achieved if object-based coding is employed. Also for different sequences, the savings are different. This means that the saving depends on the video content which may have different objects with different number and size.

3. THE MODEL OF COMPRESSION EFFICIENCY AND VIDEO CONTENT

Section 2 shows that for a specific surveillance video sequence, the achievable savings of object-based coding compared with frame-based coding is decided by the video content. In object-based coding the encoded bit-stream contains both the texture and motion information for foreground objects and the shape information for the binary mask. In frame-based coding the coded-stream contains the texture and motion information for both foreground objects and background. So the saving should be related with the number and size of the foreground objects. If the number or the size of the foreground objects goes too large, too many bits may be wasted on shape information or other overhead information and the saving may become marginal and frame-based coding is preferred in terms of compression efficiency. So in the following we will study the relationship of coding efficiency with the number and size of objects.

It is well known that the bound of compression is given by rate distortion theory. However the distribution of video source is hard to obtain. Deriving analytical model of savings with respect to the number and size of objects is impractical. This paper proposes a scheme of deriving the model of compression efficiency of object-based coding using statistical learning specifically linear regression [4]. The scheme first derives the number of bits needed for a training sequence if object-based coding and frame-based coding are applied respectively. Then linear regression is used to obtain the model. The derived the model can be used to predict the possible saving of object-based coding and decide whether frame-based coding or object-based coding should be used in the target video sequence. The detail steps are as follows.

In the training phase, frame-based coding first is applied to a training sequence of the surveillance video. The number of bits needed for each frame k is $R_f(k)$. Then CB-BGS and object-based coding are applied to the same training sequence, and the number N_k and size S_k of the objects are obtained together with the bits $R_o(k)$ needed for each frame k. Here the size S_k of objects is just the number of pixels in the foreground objects. The coding efficiency Y_k of object-based coding is defined as.

$$Y_k = \frac{R_o(k)}{R_f(k)}$$

As compression process eliminates temporal and spatial redundancy, Y_k can be looked as a sequence of independent random variables. We model the relationship between Y_k and the number N_k and size S_k of the objects in frame k as a linear model, i.e.,

$$\begin{aligned} Y_k &= \beta_0 + \beta_1 N_k + \beta_2 S_k + \varepsilon_k \\ &= \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k \end{aligned}$$

where

$$\mathbf{x}_k = \begin{bmatrix} 1 & N_k & S_k \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 \end{bmatrix}^T$$

and ε_k is a random variable representing the error term in the model.

From the training sequence, we can get a sequence of training data $(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \cdots, (\mathbf{x}_N, Y_N)$. From the training data, the model is derived by minimizing the training error

$$Err_{train} = (\mathbf{Y} - \mathbf{x}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{x}\boldsymbol{\beta})$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_N \end{bmatrix}^T$$

and

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{bmatrix}^T$$

From linear regression theory [4]

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{Y}$$

minimizes the training error.

Once N_n and S_n for frame n have been obtained from CB-BGS, we can predict the coding efficiency of object-based coding for frame n as

$$\hat{Y}_n = \mathbf{x}_n \hat{\boldsymbol{\beta}} = \mathbf{x}_n (\mathbf{x}_n{}^T \mathbf{x}_n)^{-1} \mathbf{x}_n \mathbf{Y}$$

So the model can be used to decide whether object-based coding or frame-based coding should be applied to a specific sequence of video. During background subtraction, the average of the predicted saving $\bar{Y}$ can be obtained, i.e.,

$$\bar{Y} = \frac{1}{N} \sum_{k=1}^{N} \hat{Y}_k$$

If $\bar{Y} < 1$, object-based coding is employed otherwise frame-based coding is employed.

4. SIMULATION RESULTS

Simulations are performed on synthetic and real surveillance video sequences to test the proposed model. The synthetic surveillance video is generated as follows. First we obtain different foreground objects by applying CB-BGS to different real surveillance video sequences. Then the obtained foreground objects are added randomly to the existing background of a surveillance video sequence. The resulted sequence is the synthetic surveillance video sequence. Once the synthetic sequence is got, we encode the training sequence, using both MPEG-4 frame-based coding and MPEG-4 object-based coding. Then the model is derived as described in section 3. Finally the validity of the model is tested by use of the rest of the sequence or the test sequence.

p=2	**Model**	$\hat{Y} = 0.1698 + 0.0044N + 1.7S$
	Training error	0.0024
	Test error	0.0137
p=1	**Model**	$\hat{Y} = 0.1756 + 1.8810S$
	Training error	0.0024
	Test error	0.0192

Table 2. Summary of the regression models for the synthetic sequence

The synthetic sequence used here has totally 200 frames of size 360x240. The first 100 frames are used as the training sequence. Figure 1(a) shows the data obtained in the training period. Here the object size is normalized by the size of the frame 360x240. The model is derived as

$$\hat{Y} = 0.1698 + 0.0044N + 1.7S \qquad (1)$$

The plane in figure 1(a) denotes the derived model. Figure 1(a) also shows the data from the remaining 100 frames i.e., test sequence. Apparently the plane as shown in figure 1(a) fits the data quite well. The average training error is 0.024 and the average test error is 0.0137. From (1), we can see that the coefficient before N is rather small compared with the other two coefficients. So we also apply a model that skips the number of objects. The model that ignores the number of objects is

$$\hat{Y} = 0.1756 + 1.8810S$$

Figure 1(b) shows the training data and test data when the number of objects is ignored. The line in figure 1(b) is the derived model. Again the model fits the data quite well. The average training error is 0.0024 and the average test error is 0.0192 which increases a little compared with the model that takes account of the object number but is still very small. So the number of objects is a less important parameter than the size of the objects. This is because usually in surveillance video the total size of the objects increases with the number of objects and the two parameters are correlated. Table 2 summarizes the simulation results of the synthetic sequence 1.

p=2	**Model**	$\hat{Y} = 0.0386 + 0.0189N + 10.5511S$
	Training error	0.0043
	Test error	0.0062
p=1	**Model**	$\hat{Y} = 0.0560 + 11.2581S$
	Training error	0.0044
	Test error	0.0056

Table 3. Summary of the regression models for the video surveillance sequence "entrance"

Finally the simulation is performed on a real surveillance video sequence. The sequence was taken from the entrance of a building. The sequence has 1000 frames of size 320x240. Figure 2 demonstrates the training data, the test data and the regression model when the training sequence has 200 frames. Table 3 summarizes the models, the training error and test error in different cases. From these results, we can see that the derived model fits the data very well for real surveillance sequence.

5. CONCLUSION

This paper studies the potential of compression when object-based coding is used to encode surveillance video. CB-BGS is applied to the surveillance video and foreground objects are detected. Since the large distortion of background is tolerable, background can be encoded only once. Object-based coding is employed to encode the foreground objects. Experiment results show that object-based coding can achieve significant saving

compared with frame-based coding. We further propose a scheme that gives the model of the relationship of coding efficiency of object-based coding and video content specifically the number and size of objects. Simulation results show that the model matches the test data well. So the model can be used to predict the savings of object-based coding and make selection between object-based coding and frame-based coding.

6. ACKNOWLEGEMENT

The authors would like to acknowledge Kyungnam Kim and Kyongil Yoon of Department of Computer Science in University of Maryland for providing the CB-BGS program.

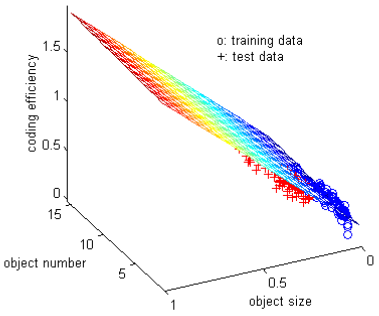
7. REFERENCES

[1] MPEG Video Group, "MPEG-4 Video Verification Model version 18.0," Doc. ISO/IEC JTC1/SC29/WG11 N3908, Jan. 2001.

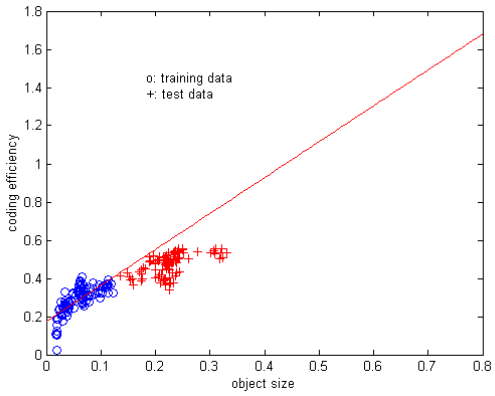
[2] Anthony Vetro, Tetsuji Haga, Kazujiko Sumi and Huifang Sun, "Object-based Coding for Long-term Archive of Surveillance Video," International Conference on ICME 2003, Vol. 2, 6-9 July 2003

[3] K.Kim, T.H.Chalidabhongse, D. Harwood and L. Davis, "Background Modeling by Codebook Construction," IEEE International Conference on Image Processing 2004 (accepted).

[4] Alvin C. Rencher, "Linear Models in Statistics," New York: Willey-Interscience 2000

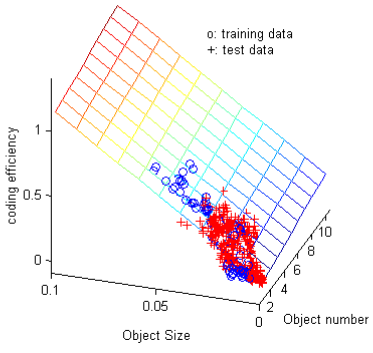


(a) $\hat{Y} = 0.1698 + 0.0044N + 1.7S$

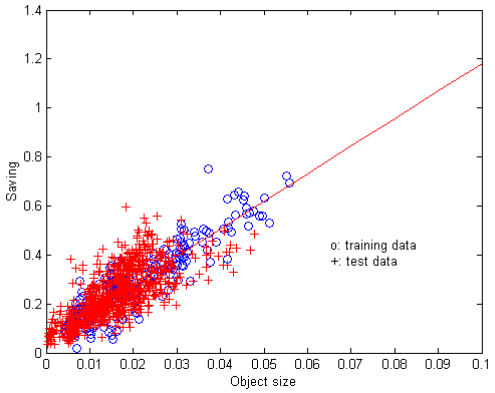


(b) $\hat{Y} = 0.1756 + 1.8810S$

Figure 1. The training data, the test data and the regression model (a) $\hat{Y} = 0.1698 + 0.0044N + 1.7S$ and (b) $\hat{Y} = 0.1756 + 1.8810S$ with 100 training frames for the synthetic sequence of 200 frames



(a) $\hat{Y} = 0.0386 + 0.0189N + 10.5511S$



(b) $\hat{Y} = 0.0560 + 11.2581S$

Figure 2. The training data, the test data and the regression model (a) $\hat{Y} = 0.0386 + 0.0189N + 10.5511S$ and (b) $\hat{Y} = 0.0560 + 11.2581S$ with 200 training frames for the surveillance sequence "entrance" of 1000 frames