

ENCODING AND TRANSCODING MULTIPLE VIDEO OBJECTS WITH VARIABLE TEMPORAL RESOLUTION

Anthony Vetro and Huifang Sun

MERL - Mitsubishi Electric Research Laboratories
Murray Hill, NJ

ABSTRACT

A key advantage of object-based coding schemes is that the quality of each video object may be varied. In this paper, we describe the problem associated with varying the temporal quality of multiple video objects, also referred to as the composition problem. We propose a solution to this problem that is based on changes in the shape boundaries over time. Finally, we discuss how this shape feature, or shape hint, can be used to vary the temporal resolution of multiple video objects. The use of the proposed shape hint may be considered within an encoding or transcoding application. This hint has also been adopted into the emerging MPEG-7 standard.

1. INTRODUCTION

With the standardization of MPEG-4, arbitrary-shaped objects may be encoded and decoded as separate video object planes (VOPs) [1]. At the receiver, these video objects are composed to form compound objects or scenes.

The encoding of multiple objects has been studied in [2]. In that paper, the encoding of multiple objects was constrained to the same temporal resolution, i.e., a *fixed* temporal rate. The main reason for this constraint was that choosing different temporal rates for each VOP, i.e., *variable* temporal rates, could lead to problems with composition at the decoder. Whether there is actually a problem or not is dependent on several things, including the definition of the VOPs (or segmentation) and their motion with respect to each other. Although the MPEG-4 standard states that VOP's may be decoded independent of one another, the encoder must still be aware of some joint influences at the encoder. Among these joint influences are the composition of objects, which is dictated by the temporal rates chosen for the set of objects; another joint influence is the shared buffer, which is impacted by (but also impacts) the allocation of bits to each object. This paper is mostly dedicated to the ability to choose variable temporal rates, but it also touches on the bit allocation problem under these conditions. Further details on bit allocation for multiple videos with different frame rates may be found in [3].

Besides encoding, the solution that we propose is also applicable to transcoding applications, where one may be interested in reducing the overall bit-rate of the objects in a scene. The solution is based on a set of measures that indicate the percentage of shape boundary change over time. In fact, this measure (or hint) has been adopted as part of the emerging MPEG-7 standard [4].

Contact Information: MERL - Mitsubishi Electric Research Laboratories, 571 Central Ave, Suite 115, Murray Hill, NJ 07974 {avetro,hsun}@merl.com

In general, the MPEG-7 standard aims to standardize a set of Descriptors (D's) and Descriptions Schemes (DS's) that can be used to describe multimedia content for a variety of applications. One such application that it plans to support is Media Conversion or Transcoding. In the current draft of the standard, there exists a MediaTranscodingHint DS, which includes a variety of attributes that assist the transcoder. The set of attributes reduce complexity and/or improve the quality of a transcoder, as well as enable certain functionalities (such as variable temporal resolution). In this paper, we emphasize the use of one particular transcoding hint, the shape hint, but also discuss how other hints may be applied.

The rest of the paper is organized as follows. In the next section, we briefly discuss the composition problem. In Section 3, we present the proposed shape hint that is used to overcome this problem. In section 4, several uses of the shape hint are presented, with corresponding simulation results shown in section 5. Finally, some conclusions are drawn in section 6.

2. COMPOSITION PROBLEM

To motivate our reasons for considering a shape hint, the composition problem is discussed. The foreman sequence is shown in the Figure 1. This sequence has two objects, a foreground and background. The left image shows the decoded and composited sequence for the case when both objects are encoded at 30Hz. The right image shows the decoded and composited sequence when the foreground is encoded at 30Hz and the background at 15Hz.

When these two objects are encoded at the same temporal resolution, there is no problem with object composition during image reconstruction in the receiver, i.e., all pixels in the reconstructed scene are defined. However, a problem occurs when the objects are encoded at different temporal resolutions. When the shape of the object has movement in the scene and these object are being encoded at different rates, it causes undefined pixels (or holes) in the reconstructed frames. These holes are due to the movement of one object, without the updating of adjacent or overlapping objects. The holes are uncovered areas of the scene that cannot be associated with either object and for which no pixels are defined. The holes disappear when the objects are resynchronized, i.e., background and foreground are coded at the same time instant.

3. SHAPE HINT EXTRACTION

We now shift our discussion to measures that define the shape change or distortion of an object through time. The normalized measure is referred to as the shape hint in the MPEG-7 standard.



Figure 1: Illustration of composition problem for encoding multiple objects at different temporal rates. (Left) foreground and background both coded at 30Hz, no composition problem. (Right) foreground coded at 30Hz, background coded at 15Hz, composition problem is shown.

We also provide several examples of the shape hint for various test sequences.

3.1. Shape Distortion Measures

Below, two shape distortion measures that have been proposed in [5] for key frame extraction are reviewed. The distortion is measured from one frame to the next, however such shape metrics can also relate to the scene at various other levels, e.g., at the segment level over some defined period of time.

The first difference measure that is considered is the well-known Hamming distance, which measures the number of different pixels between two shapes. As defined in [5], the Hamming distance, d , is given by:

$$d = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \|\alpha_1(m, n) - \alpha_2(m, n)\| \quad (1)$$

where $\alpha_1(m, n)$ and $\alpha_2(m, n)$ are corresponding segmentation planes at different time instances. The second shape difference measure is the Hausdorff distance, which is defined as the maxmin function between two sets of pixel:

$$h(A, B) = \max \{ \min \{ d(a, b) \} \} \quad (2)$$

where a and b are pixels of the sets A and B of two video objects respectively, and $d(a, b)$ is the Euclidean distance between these points. The above metric indicates the maximum distance of the pixels in set A to the nearest pixel in set B . Because this metric is not symmetric, i.e., $h(A, B)$ may not be equal to $h(B, A)$, a more general definition is given by:

$$H(A, B) = \max \{ h(A, B), h(B, A) \} \quad (3)$$

It should be noted that the above measures are taken between corresponding shapes at different time instants. For a difference between time instants, small changes indicate a potential for greater variations in the temporal resolution for each object, whereas larger changes indicate smaller variations. If the duration of time between instances is made larger and the differences remain small, then this also indicates a potential for savings through variations in the temporal quality.

In addition, the above measures must be normalized in the range $[0, 1]$ over the video segment being considered. For the Hamming distance, we may normalize the distortion by the number of pixels in the object. A value of zero would indicate no

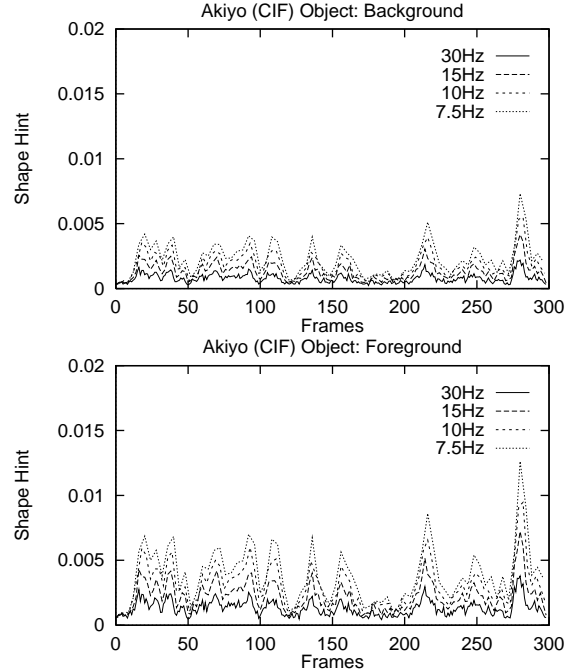


Figure 2: Shape Hints for Akiyo.

change, while a value of one would indicate that the object is moving very fast. If the Hausdorff distance is used, we may normalize with respect to the maximum width or height of the object's bounding box or of the frame. Due to the non-symmetric nature of the Hausdorff distance, it may also be useful in to distinguish differences between a growing shape boundary and a reducing shape boundary, i.e., a shape boundary may grow as the result of a zoom in and reduce as the result of a zoom out. With regard to the composition problem, shape boundaries that grow may be treated differently than those that reduce. Regardless of the measure used, it is sufficient to average the sequence of shape hints to yield a single shape hint value for the VOP over a given period of time. Actually, the shape hints themselves can be used to provide a temporal segmentation of the VOP's so that the averaged shape hint is more reliable for the specified duration of time.

We should note that these difference measures are most accurate when computed in the pixel-domain, however, as in [5], approximated data from the compressed-domain can also be used. The pixel-domain data are readily available in the encoder, but for the transcoder, it may not be computationally feasible (or of interest) to decode the shape data. Rather, the macroblock coding modes (e.g., transparent, opaque, border) could be used. In this case, a macroblock-level shape boundary could be formed and the distortion between shape can be computed.

3.2. Examples of Extracted Shape Hints

The shape hints are extracted for each object in a number of sequences at 30Hz, 15Hz, 10Hz and 7.5Hz. For each of the sequences, we interpret how the shape hint mat be used by a transcoder. In many cases, the final outcome and transcoder decision is based on additional information.

Shape Hints for Akiyo (Fig. 2): The shape hints for this se-

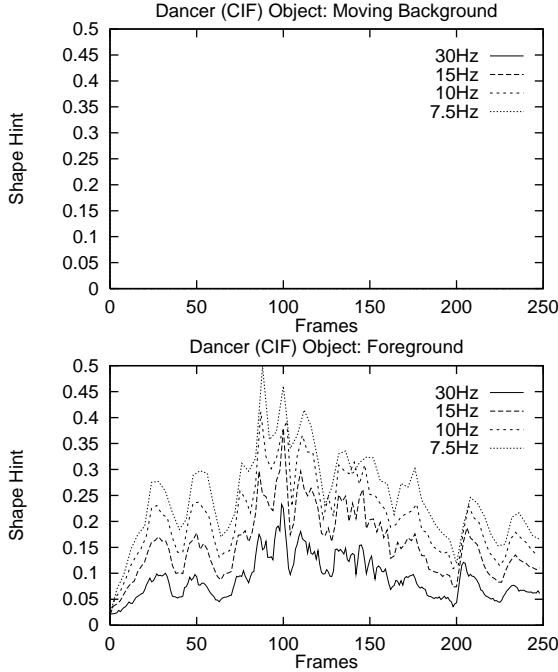


Figure 3: Shape Hints for Dancer.

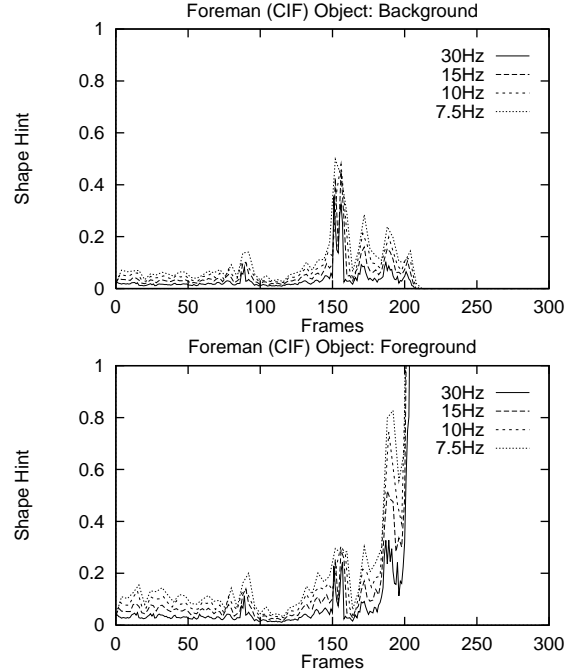


Figure 4: Shape Hints for Foreman.

quence indicate very low movement of objects. This would be a very good candidate for variable temporal reduction. Depending on other content characteristics, an encoder may reduce the resolution of either object.

Shape Hints for Dancer (Fig. 3): The shape hint for the background indicates no change (it is a full rectangular frame), while the foreground object is experiencing large movement in shape. Is it possible to have variable resolutions with this scene, since the full background is coded and it will never cause holes. However, due to high degree of shape movement in the foreground object, one may decide to encode this object at the full temporal rate.

Shape Hint for Foreman (Fig. 4): The movement of the background and foreground agree with each other and show large movement for the first 200 frames. After that, the shape hint indicates that the foreground no longer exists and the background eventually fills the entire frame.

4. ACHIEVING VARIABLE TEMPORAL RESOLUTION

In the previous section, the extraction of the shape hint was discussed. Here, we describe how it may be used for video object encoding or transcoding. In Fig. 5, we illustrate the steps that would be taken to compute variable temporal rates for each object. First, the shape hints for each VOP are extracted according to the methods discussed in the previous section. The collection of hints are passed into an analyzer whose function is to determine if composition problems will occur at the decoder if variable temporal rates are used. If variable temporal rates are allowed, then with the assistance of additional transcoding hints or content characteristics, the temporal rates for each object can be computed. However, this does not always imply that they will be employed. On the other hand, if variable temporal rates are not allowed, the encoder or transcoder may still adjust the outgoing temporal rate equally for

all objects. These decisions would be based on the outgoing bit-rate and content characteristics.

To make the decision if variable temporal rates should be allowed, a simple algorithm is proposed. First, it is important to determine the existence of any still objects, which are either full-rectangular objects that consume the entire frame or arbitrarily shaped objects that do not move. Such objects are easy to find since the shape hint should be exactly equal to zero (see background of dancer sequence, Fig. 3). Since it takes at least two moving objects to cause composition problems, we can immediately allow variable rates if the number of still objects is less than the total number of objects minus two. If this condition is not satisfied, we must then determine if the movement of the non-still objects is tolerable or not. To do so, each shape hint is compared to an empirically determined threshold. If the threshold is exceeded by any one hint, then we may conclude that too much distortion in the composed scene will result. Consequently, the temporal rate should be fixed.

Given that the temporal rate may vary, additional information about the objects is needed to determine the actual rates that each VOP is assigned. The method that we describe is based on an importance measure that was first proposed in [6] to identify the key objects in the scene. This measure uses two other transcoding hints that have been specified in the MPEG-7 standard, the difficulty and motion intensity hints. The difficulty of an object, B , is defined by its bit usage or complexity, while the motion intensity, σ , is defined as the variance of motion vector magnitudes. The importance of each object is given by,

$$\eta = (\sigma + c) \cdot \frac{B}{\gamma} \quad (4)$$

where $c > 0$ is a constant to allow zero motion objects to be decided based on bits and γ is a normalizing factor that accounts for object

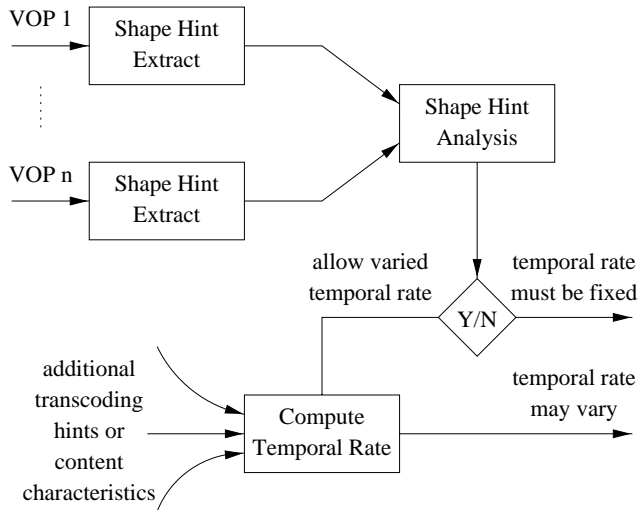


Figure 5: Block diagram to illustrate use of shape hint and how variable temporal resolution could be computed.

size. Larger values of η indicate objects of greater significance. Since the values of σ and B are specified in the range $[0,1]$, we know that the η is in the range $[0,1+c]$. If we consider a fixed set of temporal rates, we may then partition the importance scale and assign the temporal rate for each object according to its importance.

The above algorithm for computing the temporal rates only illustrates how one may exploit the shape hints to achieve some savings in bit rate. However, it does not directly address the problem of bit allocation to meet a target rate. To do so, one must consider more sophisticated algorithms that examine the ratio of incoming to outgoing bit-rate and the impact of reducing the temporal rate of each object. The impact of temporal rate reduction must also be considered jointly with the spatial quality, i.e., quantizer selection. Such spatio-temporal trade-offs are a topic of further research and are beyond the scope of this paper.

5. SIMULATION RESULTS

Given that the shape hint will identify scenes (or segments) that can take advantage of variable temporal resolution, we provide results to demonstrate the gain. In the previous subsection, the Akiyo sequence has been identified as a scene that can be encoded or transcoded with variable temporal resolution, without creating disturbing artifacts due to composition. To examine this further, we generate the R-D curves for two cases. The first case is our reference and simply encodes both foreground and background objects at 30Hz with a constant QP. This is referred to as the fixed case. As a comparison, we maintain the temporal resolution of the foreground at 30Hz, but reduce the temporal resolution of the background to 7.5Hz. This case is referred to as the variable case.

A comparison of the fixed and variable R-D curves is shown in Fig. 6. The set of QP's that were used to generate the curves were $Q = 9, 15, 21$ and 27 . The finest QP was 9 since reducing the temporal resolution for high bandwidth simulations is not of much interest. It should be emphasized that the PSNR for all simulations are computed with respect to the original 30Hz sequence. The curves in this figure show that a 25% reduction in bit rate can

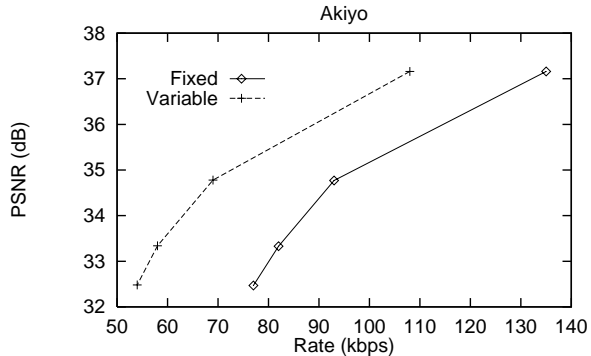


Figure 6: R-D curves comparing Akiyo sequence coded for fixed and variable temporal resolution. With fixed temporal resolution, both foreground and background are coded at 30Hz. With variable temporal resolution the foreground is still coded at 30Hz, but the background is coded at 7.5Hz.

be achieved by reducing the resolution of the background only, with no loss of quality. As indicated by the shape hints for this sequence, the movement among shape boundaries is very small. Consequently, any holes in the reconstructed image are negligible and do not impact the PSNR values.

6. CONCLUSIONS

In summary, we have described the composition problem for encoding and transcoding multiple video objects with variable temporal resolutions. A method to overcome this problem using a measure that computes the change in shape boundary over time for each VOP has been proposed. The extraction and use of this shape hint has been described and simulations with variable temporal resolution show that a 25% reduction in bit-rate can be achieved without loss of quality. Algorithms that provide bit allocation to VOP's with varying temporal rate to meet target bit rates are currently under investigation.

7. REFERENCES

- [1] ISO/IEC 14496-2:1999 "Information technology – coding of audio/visual objects," Part 2: Visual.
- [2] A. Vetro, H. Sun, and Y. Wang, "MPEG-4 rate control for multiple video objects," *IEEE Trans. Circuits and Syst. Video Technol.*, Feb. 1999.
- [3] C.W. Hung and D.W. Lin, "Towards jointly optimal rate allocation for multiple videos with possibly different frame rates," *Proc. IEEE Int'l Symp. Circuits and Systems*, Geneva, Switzerland, May 2000.
- [4] MDS Group, "MPEG-7 Multimedia Descriptions Schemes XM (v4.0)," ISO/IEC N3464, Beijing, China, July 2000.
- [5] B. Erol and F. Kossentini, "Automatic key video object plane selection using shape information in the MPEG-4 compressed-domain," *IEEE Trans. Multimedia*, June 2000.
- [6] A. Vetro, A. Divakaran, H. Sun and T. Poon, "Adaptive transcoding system based on MPEG-7 meta-data," *Proc. IEEE Pacific-Rim Conf. on Multimedia*, Sydney, Australia, Dec. 2000.