FMOE-MR: Content-Driven Multi-Resolution MPEG-4 Fine Grained Scalable Layered Video Encoding

S. Chattopadhyay, X. Luo, S. M. Bhandarkar, K. Li
Department of Computer Science, the University of Georgia
415 Boyd Graduate Studies Research Center, Athens, GA 30602-7404, USA

ABSTRACT

The MPEG-4 Fine Grained Scalability (FGS) profile aims at scalable layered video encoding, in order to ensure efficient video streaming in networks with fluctuating bandwidths. In this paper, we propose a novel technique, termed as FMOE-MR, which delivers significantly improved rate distortion performance compared to existing MPEG-4 Base Layer encoding techniques. The video frames are re-encoded at high resolution at semantically and visually important regions of the video (termed as Features, Motion and Objects) that are defined using a mask (FMO-Mask) and at low resolution in the remaining regions. The multiple-resolution re-rendering step is implemented such that further MPEG-4 compression leads to low bit rate Base Layer video encoding. The Features, Motion and Objects Encoded-Multi-Resolution (FMOE-MR) scheme is an integrated approach that requires only encoder-side modifications, and is transparent to the decoder. Further, since the FMOE-MR scheme incorporates "*smart*" video preprocessing, it requires no change in existing MPEG-4 codecs. As a result, it is straightforward to use the proposed FMOE-MR scheme with any existing MPEG codec, thus allowing great flexibility in implementation. In this paper, we have described, and implemented, unsupervised and semi-supervised algorithms to create the FMO-Mask from a given video sequence, using state-of-the-art computer vision algorithms.

Keywords: scalable video encoding, MPEG-4 FGS, multi-resolution, content based

1. INTRODUCTION

Streaming video across the Internet has become one of the most important means of distributed information sharing. Since the Internet bandwidth availability is dynamic, it is essential to dynamically adapt the bit rate of streaming video, in order to ensure uninterrupted video streaming. The MPEG-4 standard uses a layered video encoding scheme, termed as Fine Grained Scalability (FGS) profile [1] [2] [3], to achieve adaptive video streaming. MPEG-4 FGS encodes the video into a Base Layer and an Enhancement Layer. The Base Layer bit rate is the minimum bit rate at which the video can be streamed. However, this bit rate may not be sufficient for low bandwidth networks which cannot support even the low bit rate of the Base Layer. Theoretically, the Base Layer may be encoded at even lower bit rates to allow streaming to these low bandwidth networks; however, this inevitably leads to such a reduction in video quality that the visual information is almost useless. Hence, the rate distortion performance of standard MPEG-4 Base Layer encoding for FGS calls for improvement, in order to allow even lower bit rates at a reasonable video quality.

One way of improving the rate distortion performance of MPEG-4 encoding scheme for the low bit rate Base Layer is by filtering out semantically and visually "*uninteresting*" information from the video frames. This can be achieved by re-encoding each frame as a multi-resolution frame, with the visually and semantically "*interesting*" information at high resolution, and the rest in low resolution. The multi-resolution scheme is implemented such that it fits within the scheme of the standard MPEG-4 compression pipeline, consisting of quantization and variable length encoding of the DCT space of the video frames.

In this paper, we have described and implemented a *mask*-based multi-resolution (MR) step in the standard MPEG-4 FGS Base Layer encoding pipeline, which can achieve acceptable video quality in visually important regions of the video at very low overall bit rates. Each frame of the FGS Base Layer video is re-encoded, using the proposed Features, Motion and Objects Encoded-Multi-Resolution (FMOE-MR) scheme, such that the regions defined by the *mask* are at high resolution, whereas the remaining regions in the frame are at low resolution. The multi-resolution scheme is implemented such that when the MPEG-4 video encoding pipeline converts the color-space of frames to their corresponding DCT space, the DCT coefficients require a very low number of bits for encoding. This leads to low bit

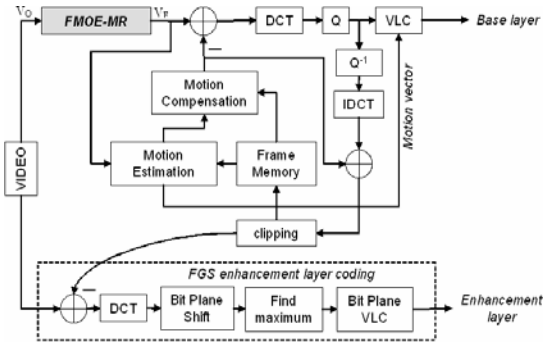


Figure 1: The MPEG-4 FGS scalable video encoding pipeline, with the proposed FMOE-MR step added. (grey box).

rate Base Layer MPEG-4 video encoding scheme. We demonstrate unsupervised and semi-supervised methods to create *Features, Motion and Objects* (*FMO)-masks* based on the presence of features, motions and objects detected in the video sequence. The *FMO-Mask* essentially defines the importance of a pixel in the image by assigning it a weight that lies between 0 and 1. The *FMO-Mask* is obtained computationally via content analysis of the video sequences using appropriate computer vision algorithms. The combination of the *FMO-Mask* and Base Layer selective video enhancement using multi-resolution (MR) techniques is what we term as the proposed FMOE-MR scheme.

It must be noted that there exist technologies [4] to selectively enhance the quality of spatial regions in the video frame while streaming within a constrained bandwidth. MPEG-4 selective enhancement [5] is employed in the Enhancement Layer of MPEG-4 FGS in order to stream higher quality video within selected image regions. However, MPEG-4 selective enhancement does not provide quality improvement for the Base Layer. Improving video quality of the Base Layer is essential, because for low network bandwidths since the Base Layer is often the only layer which can be streamed. In order to improve the quality of the Base Layer, MPEG-4 FGS uses adaptive quantization (FGS-AQ) [5]. FGS-AQ quantizes each 8×8 DCT block differently based on its relevance in improving the overall video quality.

The proposed FMOE-MR technique has several advantages over the existing MPEG-4 FGS-AQ based Base Layer encoding; (a) FMOE-MR results in significantly better rate distortion performance compared to FGS-AQ, by using a pixel-level multi-resolution video frame representation (b) FMOE-MR is transparent to the decoder; FGS-AQ, on the other hand, requires additional parameters and components for the decoder to decode each frame (c) FMOE-MR requires no additional changes to the existing MPEG-4 codecs, thus making the overall scheme very portable.

The rest of the paper is organized as follows: In Section 2, an overview of the existing technologies for MPEG-4-based layered coding of video (FGS), and FGS-AQ, is provided. Section 3 describes the proposed FMOE-MR scheme in detail. Detailed quantitative and qualitative comparisons of the proposed FMOE-MR Base Layer video encoding scheme with the existing MPEG-4 FGS-AQ scheme are in Section 4. Finally, the conclusions and potential future work are presented in Section 5.

2. BACKGROUND

MPEG-4 fine grained scalability profile, FGS, separates the video frames into two layers, which are referred to as the Base Layer and the Enhancement Layer (**Figure 1**, minus the shaded box). The Base Layer is encoded at the minimum bit rate available to the video streaming network. The Enhancement Layer is obtained by encoding the difference between the original DCT coefficients and the coarsely quantized Base Layer coefficients in a bit-plane manner [3] [4]. The Enhancement Layer can be truncated at any bit position and can provide fine granularity of the reconstructed video quality which is proportional to the number of bits actually decoded.

The Base Layer, since it is encoded at the minimum bit rate, is often the most significant layer that can be streamed across to the client when the bandwidth drops down to a certain minimum value. Thus it is necessary that the Base Layer retain the highest quality in the semantically and visually important regions of the video for a specified bit rate. MPEG-4 uses adaptive quantization [4] [5] in its Fine Grained Scalability Base Layer encoding (FGS-AQ) scheme to assign more bits to the DCT coefficients of the blocks that correspond to the desired regions that need to be enhanced. FGS-AQ is achieved via a quantization matrix that defines different quantization step sizes for the different transform coefficients within a block (prior to performing entropy coding on these coefficients). These adaptive quantization tools have been employed successfully in the MPEG-2 and MPEG-4 (base-layer) standards [6]. However, the aim of FGS-AQ is not to improve the rate distortion performance, but rather to improve the visual quality of the resulting video. As a result, the rate distortion performance of an FGS encoder that uses FGS-AQ may actually degrade due to the overhead entailed in the transmission of the FGS-AQ parameters.

One of the main aims of the proposed FMOE-MR technique is to actually obtain better rate distortion performance for the Base Layer encoding compared to the FGS-AQ technique. A crucial by product of using the proposed FMOE-MR

scheme is its transparency to the existing MPEG-4 codecs; thus not requiring any additional codecs for decoding multi-resolution video frames. In the next section, we describe the proposed FMOE-MR technique in detail.

3. OUR APPROACH: FMOE-MR

The proposed FMOE-MR scheme is based on the fundamental observation that applying a low pass filter in the color space of an image is equivalent to DCT coefficient truncation in the corresponding DCT space of the image [10]. The FMOE-MR scheme is a two step process:

(i) *Creating the FMO-Mask*: Features, Motion and Objects (FMOs) are detected in the video sequence using state-of-the-art computer vision algorithms. A corresponding mask (FMO-Mask) is created to mark the regions corresponding to the presence of the FMOs. The mask has floating point values between (and inclusive of) 0 and 1, where 0 represents a completely uninteresting region and 1 represents a vital region for visual and semantic understanding of the image.

(ii) *Multi-Resolution (MR) re-encoding*: The original frame is re-encoded as a multi-resolution representation, guided by the FMO-mask such that regions corresponding to mask values near 1 are at high resolution compared to regions corresponding to mask values near 0.

3.1 Creating the FMO-Mask

The FMO-Mask is essentially a combination of one or more of the following three individual masks; the *Feature*-Mask (F-Mask), *Motion*-Mask (M-Mask) and the *Object*-Mask (O-Mask).

3.1.1 Feature mask (F-mask)

The F-Mask captures the low-level spatial features of the video frame. Edges are one of the most important low-level features of an image (or video frame), because human perception tends to detect edges first for objective recognition of the scene or object.

Edges can be detected automatically from a given image. There are many ways to perform edge detection. However, the majority of different methods may be grouped into two categories: gradient-based and Laplacian-based. The gradient-based methods detect the edges by looking for the maximum and minimum in the first derivative of the image intensity (color) function. The Laplacian-based methods, on the other hand, search for zero crossings in the second derivative of the image intensity (color) function to find edges. We have used a gradient-based edge detection algorithm, known as the Canny edge detector [11], to find edges in a video frame.

Once the edges in the video frame are determined by using the Canny edge detector, the F-mask is created by assigning the value of 1 to regions in and around the edges, and the value of 0 elsewhere. Note that the mask is actually a weighting matrix, and as such each pixel may be assigning any values between (and inclusive of), 0 and 1.

3.1.2 Motion mask (M-mask)

Motion within a video sequence constitutes a very important visual phenomenon. The human eye tends to follow the moving objects to note their activities. Therefore, in situations which demand reduction in quality of the video, the regions with motion in the video can be rendered at high resolution and the rest of the video at low resolution. Detection of motion in video sequences is summarized in two major steps:

(i) *Background Subtraction*: The background of a video sequence is either the stationary backdrop, or backdrops which change as a result of camera motion such as panning, translation etc. Background subtraction is required in order to extract foreground objects which are moving relative to the camera, or had been moving recently. Background subtraction is done typically by creating background models first [12]. Then, the video sequences are compared with the background model to detect regions which are not part of background. These regions are classified as foreground object.

(ii) *Foreground Object Tracking*: Once the foreground objects are detected, they are tracked over the sequence. Note that a simpler modification would be to just detect the foreground in each frame, as the mask is dependent on the foreground objects only. However, this approach fails when the moving object stops temporarily; in such a case, just the foreground becomes a part of the background. Tracking, on the other hand, will still detect the still object as part of the foreground.

Motion tracking is itself an extremely well-researched area, and is clearly beyond the scope of this paper. We have implemented a novel tracking algorithm based on optical flow-based multi-scale elastic matching [13]. The algorithm

can detect and track multiple objects moving in a video sequence. The tracked objects constitute an important part of the M-mask.

3.1.3 Object mask (O-mask)

All the foreground objects in a video sequence may not be equally important visually or semantically. For example, in a video sequence containing a news reader, with a rotating background logo, the face of the news reader is more important than the moving logo (which too is a part of the foreground). In this case, the face is deemed an object of interest amongst the foreground regions.

Face recognition and tracking is typically done using feature points that recognize a face, and tracking the feature points. A detailed description of the face recognition and tracking algorithm is again beyond the scope of this paper. Faces in a video sequence can be detected using the algorithm described in [16] and tracked using the algorithm described in [13] to create an O-mask based on human faces automatically.

3.3.4 Mask Combinations

The three masks mentioned above are implemented such that each pixel value represents a weight in the range [0, 1]. The higher the weight, the more significant the pixel in the overall visual and semantic content of the image frame. A combination of these masks may be more appropriate for a particular application. The three masks, F-Mask, M-Mask and O-Mask, can be combined as follows:

Original **FMO-Mask**

Figure 2: An *optimistic combination* of F-Mask, M-Mask and O-Mask to form the FMO-mask.

(i) *Optimistic Combination*: An arithmetic MAX operation at each pixel would assign the maximum of the three mask values at each pixel location. This is an optimistic combination, as this would mean that a non-interesting pixel would not affect an interesting pixel.

(ii) *Pessimistic Combination*: An arithmetic MIN operation on each pixel would set the minimum mask of the three mask values at each pixel location. This is a pessimistic combination, as an uninteresting pixel would render the combined value to be that of the uninteresting pixel.

(iii) *Arithmetic Combination*: Each pixel of the combined FMO-mask is the arithmetic mean of the three mask values at each pixel location. This gives the maximum weight to pixels where all the masks values are 1, and the minimum weight at pixels where all the values are 0.

Since the M-Mask requires identification of objects in the video scene, one may argue that a natural method of encoding would be to use MPEG-4 object-based coding [14][15]. However, MPEG-4 FGS does not support object based coding. In addition, object-based coding requires additional components in both the encoder and the decoder end. FMOE-MR, on the other hand, does not require any modifications either in the encoding side, or at the decoding side.

The visual performance of FMOE-MR can be improved by using a proper combination of the above mentioned masks. The combination depends on the application on hand. For example, in order to encode a surveillance video, combinations of motion mask (to render moving objects at good resolution) and face mask (to recognize faces of the people moving around) might be required. For generic movies, where contents are not known, a combination of motion mask and edge masks can be used, as these two features are generally the most robust recognition mechanisms for human perception. An example of a combination of all the three masks to create an FMO-Mask, using *optimistic combination*, is given in **Figure 2**. A detailed study of the effect of mask combinations is present in Section 4.4. It must be noted that the FMO-Mask is used only at the encoding side, and, as such, is transparent to the decoding side.

3.2 Multi-resolution (MR) Re-encoding using FMO-Mask

Once the FMO-Mask is created, it remains now to re-encode the original frame in a multi-resolution manner, guided by the FMO-Mask. The goal of the MR step is to take the original video frame, V_O, re-encode it at multiple resolutions using the FMO-Mask, and produce the final video frame, V_F. The final video frame, V_F, is in turn fed to the MPEG-4 FGS Base Layer encoding pipeline (**Figure 1**) to obtain the Base Layer for FGS.

The simplest method of creating the MR video frame V_F is to render a weighted combination of two video frames of different resolutions. The original video frame, V_O, is used to render two video frames, V_H and V_L, such that V_H is a high resolution rendering and V_L is a low resolution rendering of the same video frame. We assume that a Gaussian filter, denoted as $G(\sigma)$, with parameter σ representing the standard deviation, is used as a representative low pass filter. V_L can

be obtained by convolving V_O with a Gaussian filter $G(\sigma_L)$; similarly, V_H can be obtained by convolving V_O with a Gaussian filter $G(\sigma_H)$. Keeping $\sigma_L > \sigma_H$ ensures that V_L is more smoothed compared to V_H; in other words, V_L is rendered at a lower resolution compared to V_H. In order to combine the two resolutions, the mask weight matrix, **W** (matrix version of the FMO-Mask), is created which describes, in terms of normalized weights (lying between 0 and 1), the regions in a frame which need to be rendered at good resolution. An intermediate video frame, V_I, is created from the two video frames, V_H and V_L, and the weight matrix **W**, given by

$$V_I = (\mathbf{I} - \mathbf{W})V_L + \mathbf{W}V_H \qquad (1)$$

where **I** is the matrix with all entries as 1. The intermediate video frame, V_I, is a multi-resolution frame re-encoding of the original video frame, V_O. However, V_I encounters abrupt changes in resolution, which is not pleasing to the eye. Another smoothing operation is performed on the intermediate frame V_I with a Gaussian filter $G(\sigma_I)$, to yield the final frame V_F as a multi-resolution version of V_I. The detailed description of the FMOE-MR step is depicted pictorially in **Figure 3**. Note that the final video frame V_F is fed to the standard MPEG-4 Base Layer encoding pipeline, as shown in **Figure 1**.

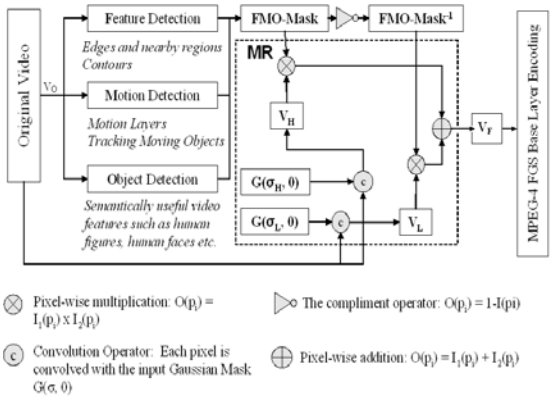


Figure 3: The proposed FMOE-MR enhancement step. The steps in the dotted box create the Multi-Resolution (MR) rendering of the original video frame.

The Base Layer video quality and encoded bit rate, after FMOE-MR and MPEG-4 video encoding, depend on the MR-Parameters σ_L, σ_H, σ_I and the FMO-Mask, depicted by the matrix **W**. The parameters σ_L, σ_H, and σ_I are bounded scalar quantities which control the bit rate; the weight matrix **W** controls the quality of the encoded video frame (it also controls, to a some extent, the encoded bit rate). Detailed discussions on the analysis, evaluation and performance of FMOE-MR are presented in the next section.

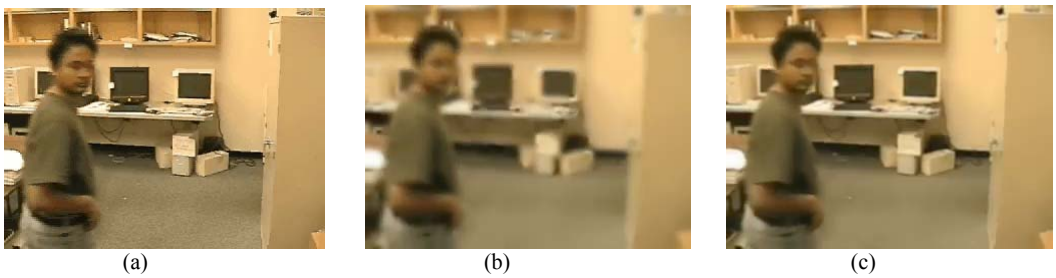
4. ANALYSIS, EVALUATION AND DISCUSSION

In this section, we first describe our evaluation methodology for the performance of the FMOE-MR scheme, followed by an objective comparison between the FMOE-MR and FGS-AQ schemes. Next, we analyze the rate distortion performance of FMOE-MR as a function of the three MR-Parameters σ_L, σ_H and σ_I. Finally, we analyze the rate distortion performance of FMOE-MR with respect to the FMO-Mask.

4.1 Evaluation Methodology

We have implemented the F-Mask using the Canny edge detector [11], and the M-Mask/O-Mask using optical flow based multi-scale elastic matching algorithm given in [13]. In order to have an objective quantification of video quality, we have used the measure of PSNR. Similarly, the bit rate is computed by measuring the size of the video file after FMOE-MR followed by MPEG-4 compression. A set of different videos have been used to compute the bit-rates and average PSNR per frame. The videos have been obtained under various background conditions (stationary, moving), various lighting conditions (moderately lighted, well lighted), various levels of motion complexity (single moving person, multiple moving persons), and various frame rates. Due to space constraints, the reported results in the section are based on the following four representative videos:

- *Video 1*: A 16 second video of a single person walking in a well lighted room. Frame rate: 30 fps, Frame Size: 320 × 240 pixels.
- *Video 2*: A 30 second panning view across a room in poor light (non-stationary background). Frame rate: 30 fps, Frame Size: 176 × 144 pixels.
- *Video 3*: Another panning video sequence of 30 seconds, this time at Frame Rate: 15 fps, Frame Size: 176 × 144 pixels.
- *Video 4*: A 40 second video of two people moving together in a well lighted room. Frame rate: 30 fps, Frame Size: 320 × 240 pixels.



(a) Original (5.12 Mbps)

(b) MPEG-4 Adaptive Quantization PSNR=22.77 dB (0.17 Mbps)

(c) Proposed Multi-Resolution PSNR = 26.5 dB (0.17 Mbps)

Figure 4: Comparison of video quality of *Video 1*, using the proposed F-MR (only edge mask is used) Base Layer encoding; compared with MPEG-4 Adaptive Quantization (FGS-AQ) Base Layer video encoding technique; (a) The original 320 X 240 frame; original AVI video is encoded at 5.12 Mbps (b) Video frame after MPEG-4 Adaptive Quantization for target bit rate around 0.17 Mbps; PSNR = 22.77 dB (c) Multi-Resolution Video frame; bit rate = 0.17 Mbps; PSNR = 26.5 dB.

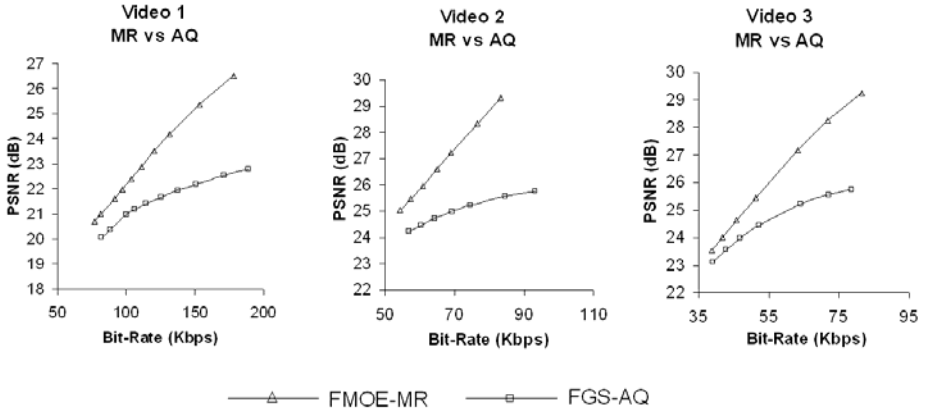


Figure 5: Rate-Distortion comparisons for the three video sequences, *Video 1* (30 FPS, 16 seconds), *Video 2* (30 FPS, 30 seconds) and *Video 3* (15 FPS, 30 seconds).

4.2 FMOE-MR vs MPEG-4 FGS-AQ

In order to compare the proposed FMOE-MR technique with the existing MPEG FGS-AQ technique, we have compared the quality of video sequences (using PSNR) for a given target bit rate. We have used a Gaussian kernel for the low pass filter. **Figure 4** shows a frame of *Video 1*, encoded using FGS-AQ and FGS-MR, at the same bit rate (0.17 Mbps), using an edge mask as the F-mask. The visual quality of the video frame using F-MR (PSNR = 26.5 dB) is significantly better than that obtained by FGS-AQ (PSNR = 22.77 dB). We observed empirically that assigning $\sigma_L = 15$, $\sigma_H = 3$ in **equation (1)**, in order to obtain an F-MR representation of the videos, produces the best results for this video. In order to provide objective evidence of the superior video quality resulting from FMOE-MR compared to FGS-AQ, **Figure 5** shows plots of PSNR as a function of the target bit rate for *Video 1*, *Video 2* and *Video 3*, by fixing $\sigma_L = 15$, $\sigma_H = 3$ and varying σ_I from 3 to 25 in **equation (1)**. All the videos have been encoded using an F-Mask. Note that the PSNR values for FGS-MR are much higher than those of the FGS-AQ technique for the entire range of bit rates.

The results above show that the FMOE-MR technique yields a higher quality video for the same bit rate, compared to the FGS-AQ technique. This can be attributed to the fact that pixel-level enhancement can be performed in FMOE-MR, whereas FGS-AQ does block-based enhancements in the DCT space. Thus, in order to enhance even a single pixel in a block, FGS-AQ enhances the whole block; although this improves the overall PSNR slightly, the overall bit rate increases. This is not the case for FMOE-MR, since FMOE-MR enhances a single pixel. The DCT coefficients in the corresponding block in the case of FMOE-MR may be enhanced less drastically.

Thus, for the same target bit rate of the Base layer for FGS, the resulting video quality (measured in terms of PSNR) is significantly better by using FMOE-MR adaptation, compared to standard MPEG-4 FGS-AQ adaptation.

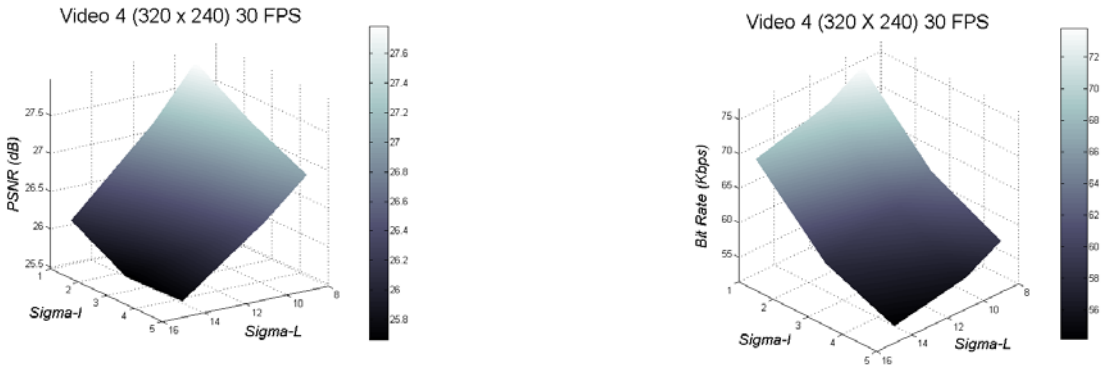


Figure 6: Rate distortion 3D plots of FMOE-MR vs σ_L σ_H and σ_I. $\sigma_H = 3$ in all cases.

4.3 FMOE-MR vs MR-Parameters (σ_L, σ_H, σ_I)

In order to analyze the rate distortion performance of FMOE-MR as a function of the three MR-Parameters, σ_L, σ_H and σ_L, we have generated 3D plots of PSNR and bit rate as functions of the MR-Parameters. Since the plots for all the four videos are similar, we show the plots for only *Video 4* in **Figure 6**. This figure shows the dependence of PSNR and bit rate on the two MR-Parameters, σ_L and σ_I, by keeping $\sigma_H = 3$ (fixing σ_H to an empirically chosen value of 3 renders the foreground at a consistently good resolution). The two persons in *Video 4* have been tracked using the algorithm described in [13], to create the motion mask (M-Mask), for the video sequence.

The bit rate varies more with σ_I compared to σ_L. This is because σ_I smoothes the video frame uniformly overall, thus effectively truncating all the DCT coefficients of the frame. The dependence of PSNR on the values of σ_L than σ_I are not apparent, except for the fact that PSNR increases as either decreases. From the analysis, it is clear that if the bit rate is the primary concern, the MR-parameter σ_I should be used to control the bit rate. The visual quality of the video frame depends on the proper combination of σ_L and σ_H. We have empirically found that $\sigma_H = 3$ renders the "*interesting*" regions defined by the mask at good visual quality, and also reduces the overall bit rate to some extent.

The optimal values of the MR-parameters can be computed by optimizing a suitable figure of merit function. We have devised a figure of merit function δ, which is the ratio of the visual quality (Q) to the obtained compression ratio (C) for a given video sequence. The metric δ is designed such that the higher its value, the better the rate distortion performance. We have defined δ as

$$\delta = Q/C \qquad (2)$$

where $Q = 2^{PSNR(\sigma L,\, \sigma H,\, \sigma I)/10}$ and C = Compression Ratio (FMOE-MR + MPEG-4 encoded versus MPEG-4-only encoded Base Layer). The expression Q is obtained from the standard PSNR equation: PSNR = 10log(Q), where Q is the ratio of the mean square error to the number of pixels in the frame. The metric δ is maximized for the video of best visual quality with the least bit rate requirement. Thus, the higher the metric δ, the more *efficient* the video encoding scheme. The values of σ_L, σ_H, σ_I which maximize δ can potentially be found by exhaustively enumerating all possible values of σ_L, σ_H, σ_I, and computing δ for each combination. The MR-Parameters σ_L, σ_H, σ_I are essentially the size of the Gaussian convolution masks, which take odd integral values, and are bounded from above. Hence, enumerating all possible combinations of σ_L, σ_H, σ_I is not as daunting as it seems, and is certainly computationally not challenging for bounded, small values of the MR-Parameters.

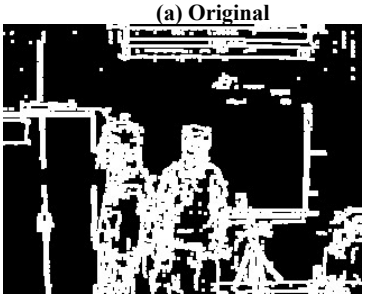
4.4 FMOE-MR vs FMO-Mask

The FMO-Mask plays a significant role in the rate distortion performance of FMOE-MR. In order to compare the effect of the FMO-Mask on performance of FMOE-MR for *Video 4* (*Video 4* is chosen arbitrarily), we fix values of the MR parameters empirically as follows: $\sigma_L = 21$, $\sigma_H = 3$ and $\sigma_I = 1$. We have used three kinds of masks: F-mask (edge mask, implemented using the Canny edge operator [11]), M-mask (object tracking mask, implemented using the algorithm described in [13]), and an *optimistic combination* of the two masks (FM-Mask).

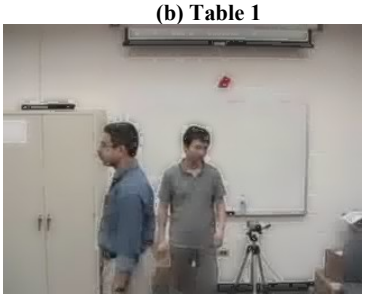
(a) Original

Mask Type	Compression	PSNR	δ
F-Mask	82%	28.79	1.05
M-Mask	60%	24.94	1.25
FM-Mask	80%	29.04	1.10

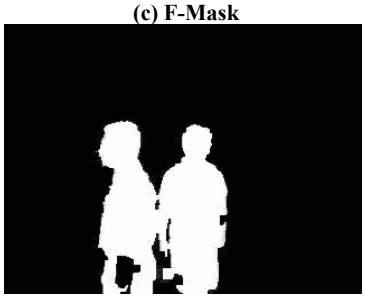
(b) Table 1



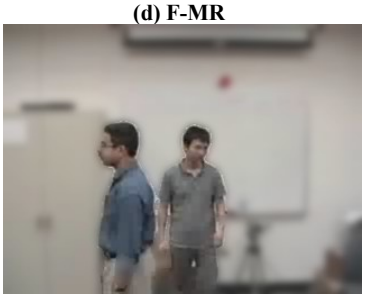
(c) F-Mask



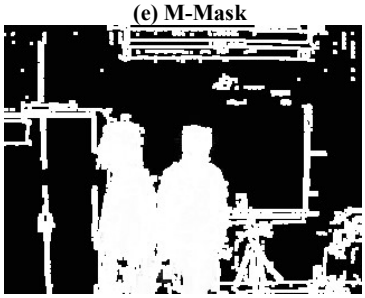
(d) F-MR



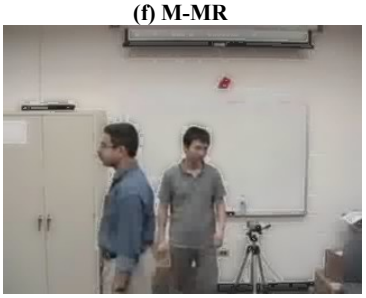
(e) M-Mask



(f) M-MR



(g) FM-Mask



(h) FM-MR

Figure 7: Frame 1070 from *Video 4* (a) The original frame (b) The Table of results showing compression ratio, PSNR and evaluation metric (discussed in text) (c) The F-Mask, obtained by edge detection using Canny edge detector, and padding added (d) Multi-Resolution (MR) Frame, re-rendered aided by F-Mask (e) Motion mask using Motion tracking [13] (f) MR-Frame aided by M-Mask (g) *Optimistically combined* FM-Mask (h) MR-Frame aided by FM-Mask.

Figure 7 shows the multi-resolution re-rendering of the original frame based on the three types of masks mentioned above. The F-Mask has high resolution areas spread over the entire frame. An interesting observation is that the effect of the M-mask is synonymous to the focusing of the human eye, which focuses on an object of interest by blurring the other background objects. The *optimistically combined* FM-Mask delivers the best visual quality, as expected, but at a price of only a modest compression gain. **Figure 7(b)** shows the compression ratio (FMOE-MR versus standard MPEG-4), the

corresponding PSNR obtained by using the various mask types, and the corresponding value of δ (using **equation (2)**) in a tabular format.

The values of δ reveal that the resulting compression from the M-Mask is the best in terms of rate distortion performance. A brief analysis will reveal why this is the case. Although the M-Mask seems to yield the worst overall PSNR, the bit rate obtained is significantly less too, compared to that obtained by the other masks. The bit rate is significantly lower for the M-Mask because the white regions (i.e. semantically and visually important regions) in the frame are grouped together in the mask. The grouping results in large DCT coefficients in only the DCT blocks in which the white portions of the mask are present. In the multi-resolution frame resulting from the F-Mask and FM-Mask, on the other hand, the relevant regions (or white portions of the mask) are spatially distributed throughout the entire frame. This makes FMOE-MR less effective, because after using the proposed FMOE-MR scheme, the MPEG-4 DCT-based compression is faced with large DCT coefficients for almost all the DCT blocks in the frame.

4.5 Encoding time of FMOE-MR

The time taken T_{base} to encode a given raw video as an MPEG-4 FGS Base Layer is given by

$$T_{base} = T_f + T_m \qquad (3)$$

where T_f is the time taken to re-encode the given raw video frames in multi-resolution format, and T_m is the time required to encode the re-encoded raw video using the standard MPEG-4 video compression pipeline. Since standard MPEG-4 video encoding can be done in near real time, we limit our discussion to the complexity of T_f, given by:

$$T_f = T_{mask} + T_{\sigma L} + T_{\sigma H} + T_{merge} \qquad (4)$$

where T_{mask} is the time taken to create the mask, $T_{\sigma L}$ is the time taken to create the low resolution image, $T_{\sigma H}$ is the time taken to create the high resolution image, and T_{merge} is the time to merge the two different images. For an edge mask, the time to create the mask is $O(nmN)$, where the raw video has N 3-channel frames, each frame of size $n \times m$ pixels. Similarly, in the case of motion masks, optical flow analysis for each pair of successive frames is computationally the most complex aspect of the computation and takes $O(nm)$ time; thus the overall time is $O(nmN)$ for all the N frames. In general, the computational complexity for T_{mask} is $O(nmN)$. $T_{\sigma L}$ depends on the time taken for convolution of the image of size $n \times m$ with a Gaussian mask of constant size; thus $T_{\sigma L}$ is $O(nmN)$ for all the N frames. Similarly, $T_{\sigma H}$ is also $O(nmN)$. Since the merging is done on a pixel-by-pixel basis, T_{merge} is $O(nmN)$ for all the N frames. The video frame size of $n \times m$ pixels is $O(n^2)$ since n and m are typically scaled to a fixed ratio of 4:3. Thus, from equation (4), we get $T_f = O(n^2N)$, where n is the width (or height) of each video frame, and N is the total number of frames in the video.

Theoretical analysis apart, the proposed FMOE-MR scheme, when implemented using a simple mask such as an edge mask, is capable of real-time performance for all of the four video examples mentioned above. For more complex frames that entail motion estimation using optical flow, near real-time performance can be achieved.

5. CONCLUSION

A novel multi-resolution Base Layer encoding technique for MPEG-4 fine grained scalability (FMOE-MR) video encoding has been described and implemented. Results show that the rate distortion performance of the proposed FMOE-MR technique is significantly better than that of the existing MPEG-4 adaptive quantization technique (FGS-AQ) for FGS Base Layer encoding. FMOE-MR entails "*smart*" preprocessing of the video prior to MPEG-4 encoding for creation of the Base Layer for FGS; as a result, existing codecs for creating FGS video can be easily used in the proposed scheme. In addition, FMOE-MR is transparent to the decoder; FGS-AQ, on the other hand, requires special AQ parameters, and components, at the decoder end to reconstruct the video.

FMOE-MR is a mask based technique; the effectiveness of the MR video depends on the creation of the FMO-Mask. The FMO-Mask is designed to highlight features, motion and objects in the video sequence. We have proposed unsupervised and semi-supervised algorithms for effective creation of the FMO-Mask from any given video sequence.

FMOE-MR is more than just a tool; it is a whole new approach to intelligent, content-based scalable video encoding. Since FMOE-MR is inherently parametric, a potential future research endeavor will be to compute the MR-Parameters automatically from the given video. In addition, many new types of masks may be used, such as application-specific object masks. We are working on real-time applications where the masks can be created in real time in order to facilitate applications such as video conferencing in the presence of dynamically changing quality constraints.

REFERENCES

[1] *Coding of Audio-Visual Objects*, Part-2 Visual, Amendment 4: Streaming Video Profile, ISO/IEC 14 496-2/FPDAM4, July 2000.

[2] Li, W. *Overview of Fine Granularity Scalability in MPEG-4 Video Standard*, IEEE Trans. Circuits and Systems for Video Technology, vol. 11, no. 3, Mar. 2001, pp. 301-317.

[3] Radha, H., van der Schaar, M. and Chen, Y., *The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP*, IEEE Trans. Multimedia, vol. 3, pp. 53–68, Mar. 2001.

[4] Richardson, Iain E. G. *H.264 and MPEG-4 Video Compression: Video Coding for Next Generation Multimedia*, Wiley, 2004.

[5] Van der Schaar, M. and Lin, Y.-T., *Content-based selective enhancement for streaming video*, in Proc. IEEE Intl. Conference on Image Processing, vol. 2, 2001, pp. 977–980.

[6] Van der Schaar, M., Chen, Y., and Radha, H., *Adaptive Quantization Modes for Fine-Granular Scalability*, Contrib. to 48th MPEG Meeting, m4938, July 1999.

[7] Davies, E., *Machine Vision: Theory, Algorithms and Practicalities*, Academic Press, 1990, pp 42 - 44.

[8] Geusebroek, J.-M., Smeulders, A. W. M. and van de Weijer, J., *Fast Anisotropic Gauss Filtering*, IEEE Trans. Circuits and Systems for Video Technology, vol. 11, no. 3, March 2001.

[9] *Syntax of fine granularity scalability for video coding*, ISO/IECJTC1/SC29/WG11,MPEG99/M4792, July 1999.

[10] Gonzalez, R.C. and Woods, R.E., *Digital Image Processing,* Addison-Wesley, Reading, MA, 1992.

[11] Canny, J., *A computational approach to edge detection*", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, pp. 679--698, 1986.

[12] Luo, X. and Bhandarkar, S.M., *Robust Background Updating for Real-time Surveillance and Monitoring*, Proc. International Conference on Image Analysis and Recognition, Toronto, Canada, September, 2005. pp.1226-1233.

[13] Luo, X. and Bhandarkar, S.M., *Tracking of Multiple Objects Using Optical Flow Based Multi-scale Elastic Matching*, Workshop on Dynamical Vision at the International Conference on Computer Vision, Beijing, China, October, 2005.

[14] Bertini, M., Cucchiara, R., Del Bimbo, A. and Prati, A., *Object-based and Event-based Semantic Video Adaptation*, in Proceedings of IAPR International Conference on Pattern Recognition (ICPR 2004), vol. 4, Cambridge, UK, pp. 987-990, Aug, 23-26, 2004

[15] Wang, H., Schuster, G. M. and Katsaggelos, A. K., *Rate distortion optimal bit allocation scheme for object-based video coding*, IEEE Trans. Circuits and Systems for Video Technology, vol. 15, no. 9, pp. 1113-1123, Sept. 2005.

[16] Zarit, B. D., Super, B.J. and Quek, F.K.H., *Comparison of Five Color Models in Skin Pixel Classification*, Intl. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Washington DC, USA, September 1999, pp.58-63.