Proceedings 2001 International Conference on Image Processing, 2001, Vol. 2, pp. 977-980
Thessaloniki, Greece, 7-10 Oct., 2001

CONTENT-BASED SELECTIVE ENHANCEMENT FOR STREAMING VIDEO

M. van der Schaar and Y.-T. Lin

Philips Research USA, Briarcliff Manor, NY 10510, USA
(email: Mihaela.vanderschaar@philips.com, Yun-Ting.Lin@philips.com)

ABSTRACT

Video transmission over bandwidth-varying networks is becoming increasingly important due to emerging applications such as streaming of video over the Internet. The fundamental obstacle in designing such systems resides in the varying characteristics of the Internet (i.e. bandwidth variations and packet-loss patterns). In MPEG-4, a new scalability scheme, called Fine-Granular-Scalability (FGS), was recently standardized, which is able to adapt to bandwidth variations in real-time while using the same pre-encoded stream. This paper presents a novel technique that enables the FGS coding scheme to perform content-based enhancement and prioritized transmission of specific regions. The proposed selective enhancement mechanism gives the FGS framework the flexibility to perform differentiated bit-allocation to enhance specific objects, such that an improved visual image quality can be obtained at various bit-rates. In our system, we employed a novel real-time face detection algorithm, that in conjunction with the proposed content-based selective enhancement method consistently leads to better subjective visual quality of streaming video under various transmission bit-rates. The FGS selective enhancement method presented in this paper has been adopted in the MPEG-4 standard.

1. INTRODUCTION

Video transmission over bandwidth-varying networks is becoming increasingly important due to emerging applications such as streaming of video over the Internet [1]. The fundamental obstacle in designing such systems resides in the varying characteristics of the Internet (i.e. bandwidth variations and packet-loss patterns). Recently, several scalable coding methods have been successfully proposed for video transmission through heterogeneous networks [1][2][3]. One of these techniques is the MPEG-4 Fine-Granular Scalability (FGS) scheme [4][5], that is able to adapt in real-time (i.e. at transmission time) to the Internet bandwidth variations while using the same pre-encoded stream. To ensure a good image quality for the FGS streams at all various bit-rates, a differentiated bit-allocation needs to be employed that allows the enhancement of specific objects within a sequence.

Selectively enhancing a particular region or object in an FGS stream differs from the well-known bit-rate allocation problem encountered in the rate-control of non-scalable video compression schemes, like non-scalable MPEG-4 or H.263 video codecs. While the non-scalable codecs target a specific, pre-determined transmission bit-rate, in the FGS case, the transmission bit-rate is unknown at encoding time. Thus, the adopted rate-control needs to have a good performance over a range of bit-rates. This is because the Internet channel is extremely volatile over space and time in terms of communication capacity and quality, and the bandwidth available to the various clients accessing the (same) content can vary significantly [1].

This paper presents a *Selective Enhancement (SE)* technique that enables the FGS coding scheme to perform content-based enhancement and prioritized transmission of selective regions. The proposed SE mechanism is essential in ensuring that the FGS framework provides a high visual image quality across a large range of transmission bit-rates, since it gives the flexibility to perform differentiated bit-allocation to enhance specific objects. To ensure improved subjective quality, SE needs to be applied in conjunction with a robust real-time segmentation mechanism that accurately identifies the visually important regions within a sequence.

However, defining the visually important content (objects) within a sequence is a challenging task. Segmentation algorithms that exploit both the temporal and spatial coherence information to handle the issue of foreground/background separation have a good performance, but at the expense of a relatively high complexity. Consequently, more efficient content extraction methods target specific classes of objects.

A good example of a visually important object is the human face. Hence, in our system, we combined the proposed SE method with a novel real-time face-detection algorithm to improve the subjective image quality of human faces within the MPEG-4 FGS framework. The proposed system can be successfully employed to obtain improved subjective quality in various systems, ranging from video-conferencing to video streaming of movies and news programs, wherein faces represent the visually most important objects. It is also important to notice that both the proposed SE mechanism and the real-time face-detection algorithm can be extended to obtain subjective quality improvement in other *embedded* video coding schemes, e.g. 3-D wavelet codecs [3][7][8].

The paper is organized as follows. In Section 2, the MPEG-4 Fine-Granular Scalability coding technique is briefly presented. Subsequently, a method for selectively enhancing specific regions within an FGS video sequence is presented in Section 3. Section 4 describes a real-time face detection algorithm that can be easily incorporated within an FGS encoder for improved visual quality. Section 5 evaluates the simulation results of the proposed content-based selective enhancement algorithm both subjectively and objectively, and in Section 6, the conclusions are drawn.

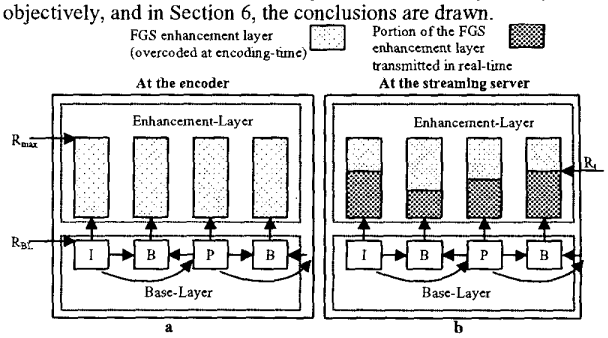

Fig. 1. The FGS structure at the encoder and streaming server for a typical Internet streaming application.

2. INTERNET VIDEO STREAMING WITH FGS

The scalability structure of the FGS framework is portrayed in Fig. 1a. In addition to the base-layer, which is coded with an MPEG-4 compliant non-scalable coder, FGS consists of a single intra-coded (i.e. without motion-compensation) enhancement-layer coded in a progressive (fine granular) manner. Under this framework, the scalable video content can be compressed over any desired bit-rate range $[R_{min}, R_{max}]$, where R_{min} and R_{max} are the minimum and maximum bandwidth available over the network at all times.

The base-layer is coded with bit-rate R_{BL}, chosen so that the available bandwidth (over the time-varying network) is higher than R_{BL} at all times ($R_{BL} \leq R_{min}$). Subsequently, the enhancement-layer is over-coded at encoding-time using a bit-rate ($R_{max}-R_{BL}$), as

0-7803-6725-1/01/$10.00 ©2001 IEEE

shown in Fig. 1a. The enhancement-layer is progressively coded using embedded compression techniques [5].

At the streaming server, the enhancement-layer improves upon the base-layer video, by fully utilizing the bandwidth Rt available at transmission-time (see Fig. 1b). Then at the decoder side, the base-layer and the received portion of the enhancement-layer data are decompressed. A more detailed description of the FGS technique can be found in [4] and [9].

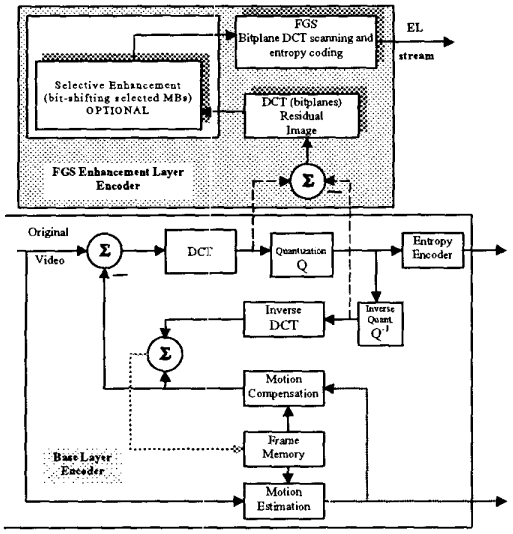

Fig. 2: FGS encoder with Selective Enhancement.

The progressive codec used in MPEG-4 for compressing the FGS residual signal is based on a bitplane DCT coding scheme [10]. Since the DCT transform is used in both the base and the enhancement layers, the FGS residual signal can be directly computed in the DCT-domain (see Fig. 2). Each DCT FGS-residual frame consists of N bitplanes:

$$N = \left\lfloor \log_2 \left(\max_{i,j,k} \left[|c(i,j,k)| \right] \right) \right\rfloor + 1$$

with $c(i;j;k)$ being the DCT coefficient i of block j within macroblock k of the residual frame. After determining N, the FGS enhancement-layer encoder scans the residual signal on a block-by-block basis using the traditional zigzag scanning method starting from the most significant bitplane BP(1) to the least significant bitplane BP(N). Run-length codes are used for (lossless) entropy-coding of the zeros and ones in each 8x8 bitplane block. Subsequently, the run-length codes are VLC coded to constitute the FGS compressed bitstream.

At the receiver side, the VLD re-generates the DCT residual bitplanes starting from the most significant bitplane to the least significant. Depending on the transmission bit-rate, the decoder may not receive all the bitplanes. The resulting DCT residual is then inverse-transformed to generate the SNR residual pixels. These residual pixels are then added to the base-layer decoder output to generate the final enhanced scalable video.

It is worth noticing that the "basic" FGS coding scheme previously described did not contain any mechanism allowing for the adaptive quantization and prioritized transmission of specific regions of a sequence. However, these tools are necessary for obtaining good visual image quality at all transmission bit-rates.

3. SELECTIVE ENHANCEMENT

In standards like MPEG-2, MPEG-4 and H.26L, adaptive quantization (AQ) is used to improve the visual quality of transform-coded video, by controlling the quantization factor on a macroblock-basis. Performing AQ on bitplane signals has to be achieved through a different set of techniques. The AQ notion for the FGS bitplane signal was first introduced in [11]. FGS-based AQ is achieved through *bitplane shifting* of selected macroblocks within an FGS enhancement-layer frame. The bitplane shifting is equivalent to a multiplication by a power-of-two. For example, in order to emphasize a particular macroblock k, all coefficients within this macroblock can be multiplied by a factor $2^{se(k)}$. Therefore, the new value $c'(i;j;k)$ of a coefficient i of block j (within macroblock k) is:

$$c'(i,j,k) = 2^{se(k)} \cdot c(i,j,k)$$

where $c(i,j,k)$ is the original value of the coefficient. This is equivalent to *up-shifting* the set of coefficients of macroblock k [$c(i;j;k)$, i=1, 2, ... 64, j=1,...4] by $se(k)$ bitplanes relative to the non-enhanced macroblocks coefficients (see Fig. 3). Hence, after AQ, the number of bitplanes N' needed for representing the FGS frame increases to:

$$N' = \max_k \left\{ se(k) + \left\lfloor \log_2 \left(\max_{i,j} \left[\left| c(i,j,k) \right| \right] \right) \right\rfloor + 1 \right\} > N$$

This adaptive-quantization tool is referred to as *Selective Enhancement* (SE), since through this approach selected macroblocks within a frame can be enhanced relatively to others.

The allowed SE shifting factors range was chosen to be 5, since every increment of the shifting factor corresponds to a factor-of-two decrease in the equivalent quantization step-size. Thus, even a moderate shifting factor range suffices to provide the encoder the flexibility necessary for maneuvering local control of quantization.

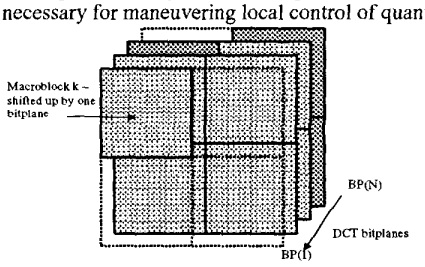

Fig. 3: Example of SE, where a "selected" macroblock k is emphasized by up-shifting its coefficients with one bit-plane (se(k)=1).

The FGS encoder with SE described above is illustrated in Fig. 2. At the encoder side, bitplane shifting due to selective-enhancement is performed on the residual FGS signal prior to the scanning and entropy coding of the bitplanes. Bitplane de-shifting is performed at the decoder side after entropy decoding and prior to the computation of the inverse DCT of the FGS residual signal.

It is important to emphasize that SE is a relative operation. Only a limited number of macroblocks should be selected for enhancement in order to obtain any perceived improvement in quality at low bit-rates. Furthermore, the rate-distortion performance of an FGS coder that uses SE may actually degrade due to the overhead introduced by the SE shifting factors, i.e. se, and the higher number of bit-planes to be coded, i.e. N'. However, the aim of FGS SE is not to improve the rate-distortion performance, but the visual quality of the resulting video.

In addition to performing SE at the macroblock-level, bitplane shifting of selected DCT coefficients could lead to further visual quality improvement for the FGS codec [12].

To obtain good visual quality at all transmission bit-rates, the proposed SE algorithm should be combined with a robust segmentation mechanism that identifies the visually important regions of a sequence. However, since it is embedded in the FGS encoder, the segmentation algorithm needs to be performed in real-time and have a low-complexity.

4. FACE DETECTION

The problem of identifying visually important objects/segments in a video is very challenging. For example, the compressed-domain information alone (i.e. motion-vectors and texture), which can be easily extracted in MPEG-based codecs, often leads to inaccurate detection of visually important areas within the scene. Alternatively, simple assumptions, such as considering the middle of the screen as being the most important area, often fail if the important object is outside this predefined area, or has a prominent motion. More sophisticated segmentation algorithms, that exploit the spatio-temporal coherence and perceptual grouping, can lead to more semantically meaningful and generic object segmentation results, but often at the expense of a relatively high complexity [13]. To achieve a meaningful segmentation in real-time, some of the best methods for selecting visually important objects target specific classes of objects.

A good example of a visually important object is the human face. In applications like video-conferencing and video-mail, as well as video streaming of movies, sit-coms, and news programs, human faces usually represent the most important visual objects within a scene. While the detection and tracking of faces have been active topics in computer vision research for many decades, only in recent years has robust real-time performance become feasible.

Here, we describe a real-time face detection and tracking algorithm that has been integrated with the previously described FGS selective enhancement method to provide improved image quality. The algorithm includes the following steps.

- *Color classification:* non-skin color areas are first identified and pruned.
- *Structure matching:* the facial luminance structure is examined only on the remaining skin-color areas.
- *Tracking:* skin color adaptation is used to continuously track the target face in order to accommodate skin-tone variations due to ethnicity or varying lighting conditions.

In recent years, skin color has become an indispensable cue for human face extraction mainly for two reasons: (1) skin color classification is much faster than luminance-based object detection algorithms; (2) human skin tones form a highly condensed cluster in certain color spaces [14], which can be effectively modeled and classified. To model the skin color distribution in facial regions, we use the θ value in the Yrθ color space, where θ is the angle to the positive U-axis in the UV space, with $U, V \in [-128,127]$, $\theta \in [0,2\pi)$, and $r=\text{sqrt}(U^2+V^2)$. Similarly to the hue, it can be shown that θ is insensitive to luminance or shading. From analyzing a large number of facial data, the color distribution of human skin-tone can be readily modeled as a function of θ by a Gaussian distribution: $p_s(\theta(x,y); \theta_s) \propto exp[(\theta - \theta_s)^2/(2\sigma_s^2)]$, where (x,y) is the pixel location on a 2D image, and θ_s is the reference skin-color (initially set from training data), representing the mean value of skin-color pixels. It is updated each time a face is detected or tracked. σ_s^2 is the variance of the skin color.

A simple thresholding on the summation of p_s over a candidate block, which is defined as a rectangular region on a reduced resolution version of the image, determines if the color of the candidate block resembles that of a face. (Since human faces can have different sizes in an image, multiple downscaled versions of the image are searched.) Non skin-tone blocks are discarded automatically without further examination of their grayscale structures. This pruning step is crucial for reducing the overall computation cost since in a typical video scene, only a small portion of the scene corresponds to skin-color areas, assuming normal lighting conditions.

Although using color alone provides a fast mechanism to extract human faces, it tends to extract non-face skin-tone objects as well. This leads to false acceptance that potentially will harm not only the coding efficiency, but also the overall perceived picture quality, since a wrongly detected area (i.e. non-face area) is enhanced at the expense of the face area. To detect a frontal-view face, a structure distance to a generic face represented by eigenface [15] is computed for each skin-tone candidate. This structure distance is compared against a preset threshold to determine if a candidate is reliably classified as a face.

A successful detection initiates a tracking event. To accommodate various head poses, the following two strategies are applied. (1) The similarity in color plays a more important role than that in the grayscale structure, since color is less sensitive to possible view changes of the face. (2) The threshold on structure distance is larger during tracking. Specifically, if the structure distance becomes too large, the candidate with the closest color to the reference skin-tone θ_s is selected. Also, to reduce the number of skin-tones candidates, color adaptation is used by updating the reference skin-tone each time the face is tracked reliably.

The face detection and tracking algorithm has been tested in several live environments. The detection rate is above 95%. In the successful cases, the subject's face is detected whenever he/she presents a frontal view to the system. Afterwards, within a tolerable view change (up to 45-degree up-and-down or sideways, and up to 30-degree rotation), the system can continuously track the detected face at a speed of more than 15 frames/s (running on a *Philips* TriMedia TM-1000 processor). The algorithm can output for each frame either a rectangular face bounding box or a post-processed pixel-resolution face mask by clustering face-color pixels in the neighborhood defined by the detected bounding box. The face mask outputted frame by frame serves as the input for FGS SE.

5. EVALUATION OF FGS SELECTIVE ENHANCEMENT

In this section, we evaluate the visual quality improvement that can be obtained by combining the previously described selective enhancement method and the face-detection segmentation algorithm proposed in Section 4. An example of such a segmentation mask at pixel level is portrayed in Fig. 4A for the 72nd frame of the MPEG-4 sequence "Foreman". Since the selective enhancement algorithm in the FGS framework is performed on a macroblock-basis, we need to determine the shifting factors *se* for each macroblock *k*. Converting the binary mask (which separates the image into foreground and background) generated by the face-detection algorithm into a collection of shifting factors can be done with a very simple technique:

If (Number_of_foreground_pixels in MB(k)) > Th, then se(k)=S; else se(k)=0.

The previously described algorithm can be easily adapted to allow for different levels of enhancement if the segmentation mechanism provides various levels of importance for the objects/ regions.

In our experiments, the threshold for enhancing a macroblock *Th* has been set to one 8x8 block (i.e. 64 samples) belonging to the foreground FG and a shifting factor S for the foreground equal to 3. Using these algorithm settings and the binary mask of Fig. 4A, the proposed selective enhancement algorithm has been employed on the "Foreman" sequence and the results obtained are presented in

Fig.4C for a base-layer bit-rate of 100kbit/s and an enhancement-layer of 150kbit/s (i.e. Rt=250kbit/s). As a reference, Fig. 4B shows the results obtained without selective enhancement at the same bit-rate.

From the images shown, it can be clearly seen that the type of FGS AQ does "selectively" enhance the visual quality of the chosen macroblocks. This obviously results in some quality degradation to the "unselected" regions. For example, notice the logo in the top left part of the image portrayed in Fig. 4C and the wall in the background of the "foreman" sequence. Although the face has been enhanced due to SE, the logo and the wall have been degraded.

To quantify in an objective manner the improvement in image quality obtained by selectively enhancing a specific region, we computed the PSNR values for the foreground (FG) and background (BG) of the "Foreman"-sequence (coded at CIF-resolution with 10 frames/s) at several transmission bit-rates and R_{BL}=100kbit/s. The results are given in Table 1. It is important to mention that the results have been obtained by truncating the same pre-encoded bit-stream at various bit-rates.

Table 1. PSNR results for "Foreman"-sequence with and without Selective enhancement (SE).

Transmission bit-rate	PSNR no SE	PSNR BG with SE	PSNR FG with SE
200kbit/s	33.48	31.48	34.99
300kbit/s	34.50	31.50	36.04
400kbit/s	35.97	33.11	36.55

At this point it is important to notice that since the selective enhancement does take place within the enhancement-layer only, even if the face-detection algorithm fails for a particular frame, this will have no or little visual impact on the sequence. This is because there is no motion-compensation within the FGS enhancement-layer and thus, if the visually important object is not enhanced in a certain frame, the human-visual system will be able to filter away the decrease in resolution (quality) for that particular frame.

6. CONCLUSIONS

In this paper, a novel technique is presented that enables the FGS coding scheme currently adopted by the MPEG-4 standard for Internet video streaming, to perform content-based enhancement of selected regions. The proposed technique has been successfully combined with a simple, yet effective face-detection mechanism to allow for selective enhancement of human faces.

The presented system has applications for both real-time (e.g. video-conferencing) and off-line (e.g. video on demand) Internet video streaming. Furthermore, it is important to notice that alternative approaches can be employed for detecting the visually important parts within an image (e.g. MPEG-7 descriptors can be used for this purpose). These content detection mechanisms can then be used to provide segmentation masks that can be easily plugged in the proposed FGS selective enhancement module.

In our future research, segmentation techniques aimed at detecting visually important content (objects) within a sequence, e.g. faces, will be used beyond rate-control. For instance, in Internet video streaming applications, better error protection strategies can be employed for the visually important parts of a scene. An example is the use of more FEC packets or packet-retransmission requests for the visually important regions (e.g. faces).

REFERENCES

[1] J. Lu, "Signal Processing for Internet Video Streaming: A Review", *Proc. SPIE*, vol. 2974, p. 246-259, Jan. 2000.

[2] B. Girod et al, "Internet Video Streaming", Multimedia Communications, Springer, p. 547-556, 1999.

[3] W. Tan, A. Zakhor, "Real-Time Internet Video Using Error Resilient Scalable Compression and TCP-Friendly Transport Protocol", *IEEE Trans. on Multimedia*, vol. 1, no. 2, June 1999.

[4] H. Radha et al, "Scalable Internet Video Using MPEG-4," Signal Processing: *Image Communication*, Sept. 1999.

[5] Study of ISO/IEC 14496-2:1999/FPDAM4, N3670, La Baule, France, USA, Oct. 2000.

[6] T. Ebrahimi, C. Horne, "MPEG-4 natural video coding – An overview", Signal Processing: *Image Communication*, 2000.

[7] Y.Q. Zhang, S. Zafar, "Motion-compensated wavelet transform coding for color video compression," *IEEE Trans. on Circ. and Sys. for Video Technology*, vol. 2, no. 3, Sept. 1992.

[8] S.J. Choi, J. W. Woods, "Motion-Compensated 3-D Subband Coding of Video," *IEEE Trans. on Image Proc.*, Febr. 1999.

[9] H. Radha, M. van der Schaar, Y. Chen, "The MPEG-4 Fine-Grained Scalable Video Coding method for Multimedia Streaming over IP", *IEEE Trans. on Multimedia*, March 2001.

[10] W. Li et al, "Experiment Result on Fine Granular Scalability", m4473, Contrib. to 46th MPEG Meeting, March '99.

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

[12] W. Li, "Frequency Weighting for FGS", m5589, Contrib. to 50th MPEG Meeting, Dec. 1999.

[13] F. Dufaux, F. Moscheni, A. Lippman, "Spatio-temporal segmentation based on motion and static segmentation", Proc. ICIP 1995.

[14] J. Yang, W. Lu, A. Waibel, "Skin-Color Modeling and Adaptation", Technical Report, CMU-CS-97-146, 1997.

[15] B. Moghaddam, A. Pentland, "Probabilistic Visual Learning for Object Detection", Proc. ICCV, Cambridge, MA, 1995.

ACKNOWLEDGEMENTS

The authors would like to thank Yingwei Chen and Hayder Radha from Philips Research USA for their contributions to the standardization of the Selective Enhancement tool.

Fig.4. The impact of FGS selective enhancement: (A) pixel-level face-detection mask and FGS decoded image at 250kbits/s; (B) without selective enhancement; and (C) with selective enhancement based on the face-detection mask.