Proceedings, 3rd IEEE International Conference on Image Processing (ICIP-96), Lausanne, Switzerland, September 1996.

SPATIO-TEMPORAL MODEL-ASSISTED COMPATIBLE CODING FOR LOW AND VERY LOW BITRATE VIDEOTELEPHONY

Jae-Beom Lee and Alexandros Eleftheriadis

Department of Electrical Engineering
and Image Technology for New Media Center
Columbia University, New York, NY 10027, USA
{jbl,eleft}@ctr.columbia.edu

ABSTRACT

We introduce the concept of *Spatio-Temporal Model-Assisted Compatible* (STMAC) coding, a technique to selectively encode areas of different importance to the human eye in terms of space and time in moving images. For this, we use the fact that human "eye contact" and "lip synchronization" are very important in person-to-person communication. Several areas including the eyes and lips need different types of quality, since different areas have different perceptual significance to human observers. For example, for the eyes "high resolution" is needed for clear communication, while for the lips "frequent refresh" is needed. The approach provides a better rate-distortion tradeoff than conventional image coding technologies based on MPEG-1, MPEG-2, H.261, as well as H.263, since STMAC coding is applied on top of an encoder, taking full advantage of its core design. The decoder does not need to be changed in any way although the encoder's rate control unit is slightly modified. This characteristic leads to the name "compatible" in the proposed concept. Experimental results are given using ITU-T H.263, addressing very low bit rate compression (13-17Kbps).

1. INTRODUCTION

Recent research presented a way to integrate techniques from computer vision to low bit rate coding systems for video telephony applications in the form of Model-Assisted Compatible(MAC) coding [3, 1, 2]. The focus was to locate and track the faces and selected facial features of persons in typical head-and-shoulders video sequences, and to exploit the location information in a "classical" video coding system. The motivation was to enable the system to selectively encode various image areas and to produce perceptually pleasing coded images where faces are sharper. Since the approach only affects the bit allocation performed at the encoder, no change is needed in the decoder. Consequently, the technique is applicable to wide range of coding techniques (including H.261 and H.263), with full compatibility with existing decoders [4, 5].

We propose a method which achieves temporally selective encoding as well as spatially selective one. Previous model-assisted coding was successful because high resolution "eye contact" is very important in person-to-person communication. In this paper we also use the fact that "lip synchronization" is very important in communication. With the same face model as the previous MAC coding, we assign different budgets of bits or frames to areas in different spatial and/or temporal locations. We refer to this approach as "Spatio-Temporal Model-Assisted Compatible (STMAC) coding". As with MAC, the proposed coding technique doesn't necessitate any modification on the decoder, and hence can be used in a compatible way. We present the application of the concept to a motion- compensated block-based transform codec, and particularly H.263, and present comparative results with baseline H.263. The technique is also applicable to other codecs of the same genre (e.g., MPEG-1 and MPEG-2), but it is most appropriate for videoconferencing applications at very low bit rates.

2. STMAC CODING

We define STMAC coding as a technique which selectively encodes areas of different importance to the human eye in terms of space and time in moving images. The basic idea of STMAC coding is to use both spatial and temporal scalabilities in a frame simultaneously. In this paper, we use 4 different areas, that is, eye area, lip area, face area, and background area. High resolution is needed in the eye area, but it doesn't need high temporal frequency. Conversely, medium resolution is sufficient in the lip area, but a high temporal frequency is desired for good lip synchronization. The facial area needs to change at least gradually, to avoid "shearing." Since the background area doesn't generally have important information, it can be degraded as much as possible to save bits for more important regions. This area-selective operation makes the technique especially suitable for very low bitrate coding purposes. The MAC coding operation is depicted in Fig. 1. As we mentioned, a relatively fine step size is used for "eye" area. A medium step size is allocated to "face" and "lip" areas. And, a very rough step size is given to "background" area. Note that this is concerned with only "spatial scalability." With this MAC coding, we add a "temporal scalability" to the moving image coding as in Fig. 2. This combined coding with Fig. 1 and Fig. 2 is defined as STMAC coding.

The first task in order to apply STMAC coding is to discriminate these four areas from a given frame. In this

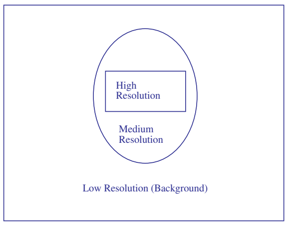

Figure 1: MAC coding: spatial scalability coding.

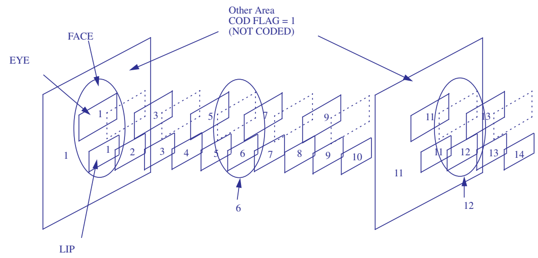

Figure 2: Proposed temporal scalability coding for STMAC coding.

paper, we use a manually-chosen fixed face model as in Fig. 1. For a better solution, we could follow the same "automatic tracking and detection" procedure of conventional MAC coding. The specific procedures are given in [3, 2].

The second task is to assign a proper set of "quantization parameters (QP)" and "frame number per second (fps)." Note that quantization step size (QS) is equal to twice the quantization parameter. The eye QS should be finer than that of other areas, and the face QS finer than that of the background. In addition, the lip area fps must be higher than that of other areas, and the face area fps higher than that of the background. Under this constraint we need to apply a rate control mechanism. In the next section, we propose "Template Adaptive Rate (TAR)" control mechanism for this STMAC coding. For convenience, we define Template, in this paper, as a set of QP and fps corresponding to the face model. If we take the order of "eye", "lip", "face", and "background" by 1, 2, 3, and 4, then we can represent the spatial quantization parameters as $S = (s1, s2, s3, s4)$, the temporal freqency fps as $T = (t1, t2, t3, t4)$ as in Fig. 4. We set the average quantizer set by $S = (5, 9, 7, 31)$, $T = (15, 30, 6, 3)$ based on subjective experiments. Note that eye area spatial QS is quite fine and lip area temporal fps is high enough compared with that of regular compression case.

Now, we need a basic functionality to suspend a macroblock data transmission of the selected regions. Most of MC-DCT encoder/decoder pairs such as H.261, H.263, MPEG-1, and MPEG-2, etc. have "not coded" mode where the decoder just copies the macroblock from the previous reference frame at the exactly same position. The "not coded" mode can be used very easily simply by setting the so-called COD flag to 1 in H.263 [5]. This is the basic function for selecting different temporal scalabilities in the different image areas. Fig. 2 shows the STMAC coding concept and the basic functionality in the H.263 videotelephony example.

The third task is to accomodate the "motion" of the model in STMAC coding. The reason for this is that the motion cannot be described until the next refresh occurs for each model area. The worst case can be easily imagined when we consider the situation where the person we are communicating with moves abruptly. The person will not actually there, but the shoulders (because they are included in the background) are still there until the background is refreshed, although other areas are already updated. To overcome this, we can use a global motion vector of a model as a reference, in addition to using the shoulder area. One easy way to take a representative global motion vector is to choose the relative motion of the "lip area" (or the eye area). The reason why we use the lip (eye) area as a reference is that it has the highest (relatively high) temporal refresh frequency. In this paper, in every frame the lip area is refreshed. That is, if some motion occures in the image, we compute the motion vectors of other model areas based on the relative lip area motion, and put it into other model area macroblocks artificially as in Fig. 3. As we expect, this is not the perfect solution, but it is quite good engineering solution to get around the difficulty. Because the refresh time is very short (in this paper, it is 1/3 second) for the worst case-background image movement, such a solution is generally acceptible to the human observer. Note that we do the model detection (in this paper, manually chosen-model resetting) every starting time of background refreshment. This means we think the model as "rigid" between background refreshement times. Experiment says that this is quite good assumption for "video telephony" images, although it highly depends on the characteristics of input image sequence. In our test sequence (Akiyo) the model is fixed throughout its duration.

3. TEMPLATE ADDAPTIVE RATE CONTROL

For STMAC coding, we cannot use traditional rate control mechanisms. Conventional rate control schemes are based on examination of the current occupancy of the transmission buffer: if the occupancy is large, the encoder uses a large step size to quantize the DCT coefficients. Conversely, if the occupancy is small, the encoder uses a finer step size for quantization. The number of quantizers is always one so that conceptually there is no difficulty. In the STMAC coding, the difficulties for the rate control come from the fact that we already have face model with various scalabilities in terms of space and time. For example,

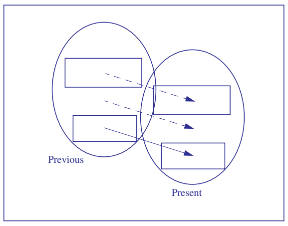

Figure 3: Relative motion vector as a global motion for human face model.

we have 4 different step sizes for spatial model: eye area, lip area, face area, and backgound area. In addition, we also have 4 different frame rates for temporal model: for the same areas such as eye, lip, face, and background. For the STMAC coding rate control, we should control these 8 different spatial scalabilities (QS) and temporal scalabilities (fps), given a target bitrate. In this paper, we propose a method called "Template-Adaptive Rate Control", which uses a fixed number of templates to control the bitrate based on a given face model. The template is a set of quantization parameters and temporal resolutions for the different areas of interest. Each template is composed of 8 parameters; for eye area, lip area, face area, and background area, each needs both quantization parameters and temporal frame rate. The precise template configuration has been determined after experimentation. Table 1 shows the templates that we obtained in our experiments. In this paper, we use the same quantization parameters as in "TMN4" for deciding "template index TI". In "TMN4", the quantization parameter is given by $QP = 16$ in the beginning. After the first picture, the buffer content is set to:

$$\frac{R}{f_{target}} + 3 \cdot \frac{R}{FR} \tag{1}$$

and

$$B_{i-1} = \bar{B}$$

Where f_{target} and $\bar{B}$ are calculated at the starting point of each frame:

$$f_{target} = 10 - \frac{Q\bar{P}_{i-1}}{4} \tag{2}$$

$$\bar{B} = \frac{R}{f_{target}}$$

For the following pictures the quantization parameter is updated at the beginning of each new macroblock line. The new quantization parameter is given by:

$$QP_{new} = Q\bar{P}_{i-1}\left(1 + \frac{\triangle_1 B}{2\bar{B}} + \frac{12\triangle_2 B}{R}\right) \tag{3}$$

$$\triangle_1 B = B_{i-1} - \bar{B}$$

$$\triangle_2 B = B_{i,mb} - \frac{mb}{MB}\bar{B}$$

Where,

$Q\bar{P}_{i-1}$: the mean quantizer parameter for the previous picture.

B_{i-1} : the number of bits spent for the previous picture.

$\bar{B}$: the target number of bits per picture.

mb : present macroblock number.

MB : number of macroblocks in a picture.

$B_{i,mb}$: the number of bits spent until now for the picture.

R : bitrate.

FR : frame rate of the source material.

Note that quantization parameters shoule be bounded by 31. From this quantization parameter, we compute the "Template Index" TI_i using:

$$TI_i = \left\lfloor \frac{PQ_i}{4} \right\rfloor \tag{4}$$

The buffer content is updated after each complete picture in the following way:

buffer_content = buffer_content + $B_{i,99}$;
while(buffer_content $> 3 \cdot \frac{R}{FR}$) {
buffer_content = buffer_content- $\frac{R}{FR}$;
frame_incr++;
}

Since the range of the quantization parameter is limited by 31, a given TI is also bounded by 8. By corresponding TI to Table 1, we can get the template quantizer set.

To make a Template Table, it is very important to note that in H.263 case "difference quantization parameter" is used for macroblocks. We must take one of DQUANT values 2,1,0,-1,-2 so that we should consider the QS difference of each area not to be exceeded more than 3. But for the background, we want to assign a QP of 31 by default. (if not, we would transmit the backgound in as much quality as other areas.) In order not to break the rule of low quality background transmission, at background refresh time, we copy the eye, lip, face areas from the previous frame. It is important to note that in the first few frames the quality of the coded images is not good, because there is no previous frame from which to copy the other model areas. DQUANT is the most restrictive syntactic component in H.263 for our "multi-scalability" coding concept. Fortunately, in other popular encoders such as H.261, MPEG-1, and MPEG-2, this does not happen, because we can take full advantage of quantization step sizes from 1 to 31 for each macroblock.

4. SIMULATION RESULTS AND CONCLUSIONS

We provide simulation results comparing the STMAC scheme with conventional H.263 compression. Note that our compression efficiency depends on the base codec efficiency. For example, if we have a very efficient H.263 encoder in which

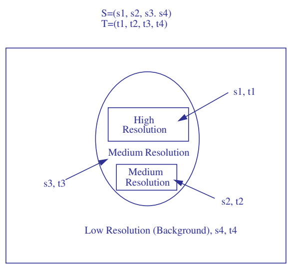

Figure 4: Template for STMAC coding.

Figure 5: Conventional H.263 at 17Kbps.

all functionalities are switched on such as Advanced Prediction Mode, PB frame Mode, etc., the STMAC encoder will be correspondingly more efficient. If the specific H.263 encoder at hand is not efficient, then the quality for the specified average bit rate will not be as good. Consequently, of importance is the relative performance compared to the base codec. Most interesting are experiments where the relative bitrates are compared for the same perceptual quality. Our experiments show that we can make similar quality video sequences around face area using STMAC coding at 13Kbps, and a conventional H.263 encoder operating at 17Kbps. The temporal resolution for H.263 was 9 fps, while the maximum temporal resolution for the STMAC encoder was 10 fps. The results on a single frame are shown in Fig. 5 and Fig. 6. for comparison. We should note that in STMAC coding the eye area is of higher resolution and the lip area is of higher temporal rate than those of the conventional H.263, although background area is degraded enough to see the worse blocking effect.

In conclusion, for very low bitrate image coding area selectivity is inevitable (object-based coding is area-selective as well). Once an appropriate model is selected and matched to the input image, we assign various spatial and temporal scalabilities in each of the areas: we assign more bits to eye area by deducting bits from the budget of the background, and we decimate different number of frames temporally according to the perceptual importance of each area. The STMAC approach gives us the benefits of high resolution eye rendition which is important for maintaining eye contact, and very sharp lip synchronization due to the high temporal frequency with which the specific area is encoded.

Figure 6: STMAC H.263 at 13Kbps.

Template Index (TI)	S	T
1	(2,6,4,31)	(15,30,6,3)
2	(3,7,5,31)	(15,30,6,3)
3	(4,8,6,31)	(15,30,6,3)
4	(5,9,7,31)	(15,30,6,3)
5	(6,10,8,31)	(15,30,5,2)
6	(7,11,9,31)	(15,30,5,2)
7	(8,12,10,31)	(15,30,5,2)
8	(9,13,11,31)	(15,30,5,2)

Table 1: Template index vs. template mapping table.

5. REFERENCES

[1] A. Eleftheriadis and A. Jacquin. Model-assisted coding of video teleconferencing sequences at low bit rates. *Proc. ISCAS '94*, May-June 1994.

[2] A. Eleftheriadis and A. Jacquin. Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit rates. *Image Communication journal*, 7(3):231–248, September 1995.

[3] A. Eleftheriadis and A. Jacquin. Automatic face location detection for model-assisted rate control in h.261 compatible coding of video. *Image Communication Journal, Special Issue on Coding Techniques for Very Low Bit rate Video*, 7(4-6):435–455, November 1995.

[4] ITU. Draft revision of recommendation H.261: video codec for audiovisual services at 64kbps. *Signal Processing:Image Communication*, 2(2):221–239, August 1990.

[5] ITU. Draft ITU-T recommendation H.263: video coding for low bitrate communication. *Expert's Group on Very Low Bitrate Video Telephony Draft*, July 1995.