Proceedings, Visual Communications and Image Processing Conference, San Jose, California, February 1997.

Motion adaptive model-assisted compatible coding with spatio-temporal scalability

Jae-Beom Lee and Alexandros Eleftheriadis

Department of Electrical Engineering and Image Technology for New Media Center
Columbia University, New York, NY 10027

{jbl,eleft}@itnm.columbia.edu

ABSTRACT

We introduce the concept of *Motion Adaptive Spatio-Temporal Model-Assisted Compatible* (MA-STMAC) coding, a technique to selectively encode areas of different importance to the human eye in terms of space and time in moving images with the consideration of object motion. Previous STMAC was proposed based on 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. The approach provides a better rate-distortion tradeoff than conventional image coding techniques based on MPEG-1, MPEG-2, H.261, as well as H.263. STMAC coding is applied on top of an encoder, taking full advantage of its core design. Model motion tracking in our previous STMAC approach was not automatic. The proposed MA-STMAC coding considers the motion of the human face within the STMAC concept using automatic area detection. Experimental results are given using ITU-T H.263, addressing very low bit rate compression.

Keyword: scalability coding, spatio-temporal scalability, automatic area detection, motion adaptivity, very low bitrate coding

1 INTRODUCTION

In practical very low bitrate image transmission systems like H.263, a key component is to make the image size small (around QCIF) and decrease the transmission image frame rate from 30 to 10. Since present techniques like MPEG-1 already give us 100 to 1 compression ratio, the obvious way for much more compression is just discarding input image frames and transmitting small size images. Since there are many applications of very low bitrate image transmission as in wireless, finding out improved techniques has become an important issue.

In very low bitrate videotelephony situations, state-of-art coding algorithms produce artifacts which are systematically present throughout the coded images; all the more as the image content in terms of motion and texture is high. These artifacts usually affect all areas of the image without discrimination. Viewers, however, will mostly find coding artifacts to be more noticeable in areas of particular interest to them. In particular, a user of a videotelephony or video teleconferencing system will typically focus his or her attention to the face(s) of the

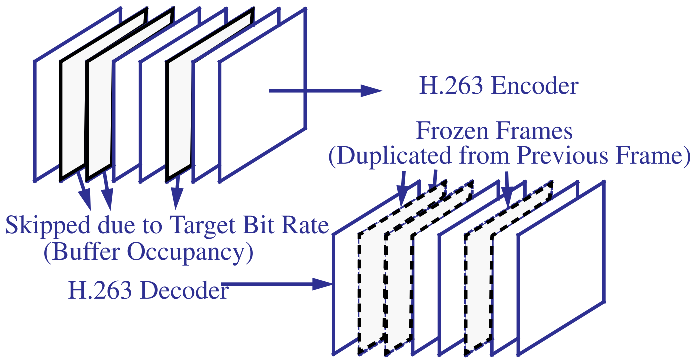

Figure 1: The image transmission on a very low bitrate channel.

person(s) on the screen, rather than to areas such as clothing or background. Besides, although fast motion is known to mask coding artifacts, the human visual system has the ability to lock on and track particular moving objects, such as a person's face. Communication between users of very low bitrate videotelephony or video teleconferencing will be intelligible and pleasing only when facial features are not plagued with an excessive amount of coding artifacts.[3]

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 have proposed a technique 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 that paper we also use the fact that "lip synchronization" is very important in communication.[6] In practical very low bitrate image transmission systems, there are two possibilities for lip synchronization loss. One possibility is from the fact that transmission frame rate is dropped from 30 to 10, which is accepted in prevailing practical very low bitrate video comunication systems such as H.263. The other possibility is from the fact that, as shown in Figure 1, some frames are discarded based on the occupancy of the transmission buffer. For either case in which frames are skipped in transmission, the decoder duplicates the non-transmitted images from the each previous frame. At this moment, lip synchronization is lost in the duplicants. In order not to distort lip synchronization, and using a similar face model as the previous MAC coding, we assigned different budgets of bits or frame rates to areas in different spatial and/or temporal locations with the consideration of model motion. Especially, to overcome the lip synchronization defect, we assigned the highest frame rate to the lip area. We refered to this approach as "Spatio-Temporal Model-Assisted Compatible (STMAC) coding". Therefore the various spatio-temporal scalabilities, based on different perceptual importance, are given to 4 basic different areas: eye, lip, face, and background. Note that the shoulders were included in the background area. For example, a relative fine step size was used for eye area, while medium step size was allocated to face and lip areas. Since the background area doesn't generally have important information, it can be

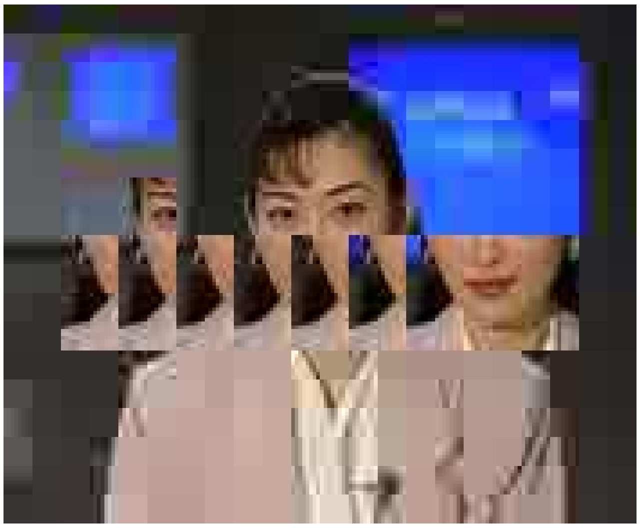

Figure 2: The encoding defect in conventional STMAC in a big motion image.

degraded as much as possible to save bits for more important regions. However, eye area doesn't need very high temporal frequency, while very high temporal frequency is needed in the lip area. This area-selective operation makes the technique especially suitable for very low bitrate coding purposes.[6]

In this paper, we also propose a motion adaptive technique which absorbs the motion of human face model with the consideration of automatic model area detection. In our previous work, object detection was performed manually. The spatio-temporal scalabilities concept presents difficulies in dealing with the motion of the pre-defined human model, since the model is considered as a "rigid" pattern. 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 be 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. Figure 2 shows the effect of extremely fast model motion. To overcome this, we propose a selective area refresh technique which encodes only particular areas inside the model. In this paper, a modified "automatic model area detection" is used for the model area discrimination. Since the "automatic model area detection" method has been developed in,[3,2] we adopt those techniques with little modification.

For controlling multi-scalabilities, we present a rate control scheme for this STMAC coding. We consider "Template Adaptive Rate (TAR) control" scheme, a rate control scheme for global bitrate trends. TAR allows the use of several pre-defined templates based on a given face model.

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 video telephony and videoconferencing applications at very low bit rates.

Section 2 provides the motivation for a Motion-Adaptive technique. STMAC consists of 3 tasks: task 1 is model area automatic tracking and detection, task 2 is motion adaptation of motion, and task 3 is assignment of a proper quantizer and frame rate. We point out that task 1 (for model area automatic tracking and detection) and task 3 (for motion adaptation of motion) are highly connected with each other, since motion area detection is a pre-requisite for motion adaptation. We also find that task 2 (quantizer and frame rate assignment) is directly related with the rate control scheme, since a target bitrate enforces the rate control scheme to assign a proper set

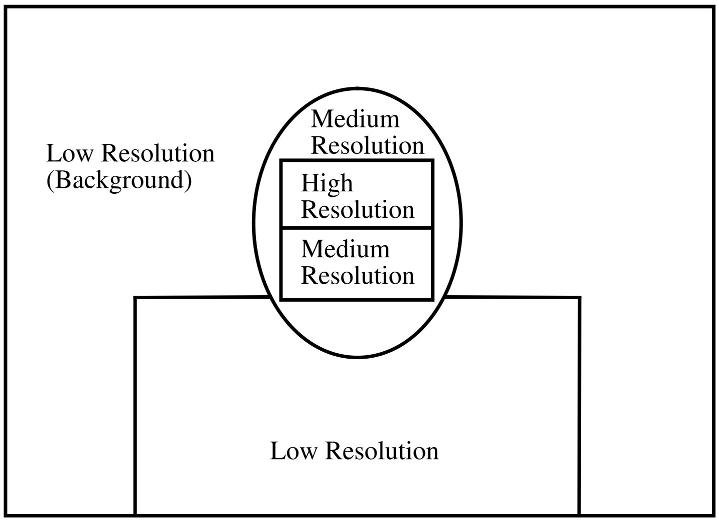

Figure 3: MAC coding: spatial scalability coding.

of quantizer and frame rate. Therefore, we discuss the automatic detection and tracking, and motion detection methods in Section 3. A rate control scheme for STMAC coding is proposed in Section 4.

2 MOTION ADAPTIVE STMAC CODING TECHNIQUE

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, lip, face, and background areas as in Figure 3, while shoulders area is used for motion adaptation, not for the scalabilities. 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. The MAC coding operation is depicted in Figure 3. As we mentioned, a relatively fine quantization 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 which includes shoulders 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 Figure 4. A most frequent refresh frame rate is given to the lip area for lip synchronization, while a very slow refresh frame rate is used for the background area. For the face area, a medium refresh frame rate is enough since face doesn't have any syncronization face feature for communications. Note that the shoulder area is included in the background area. This combined coding with Figure 3 and Figure 4 is defined as STMAC coding which considers various spatio-temporal scalabilities.

The first task in order to apply STMAC coding is to discriminate these four areas from a given frame. In this paper, for the model area matching, we follow the same "automatic tracking and detection" procedure of conventional MAC coding[3,1,2] which is composed of two sub-algorithms, automatic tracking of "head outline", and detection of "eye-nose-mouth region". We, however, need to extend the conventional algorithms, since it is necessarily used for extension of the MAC concept to that of STMAC. The specific procedures and the modified version are introduced in the next section.

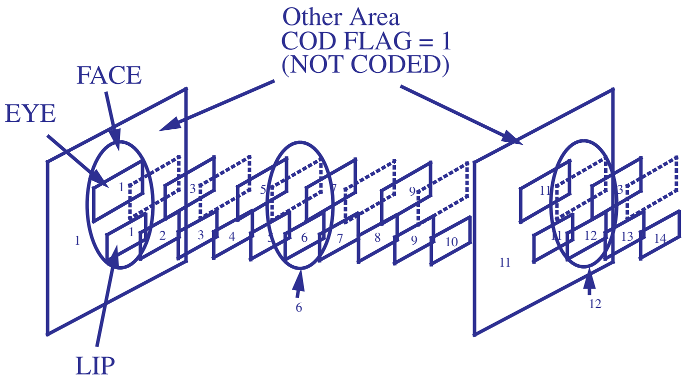

Figure 4: Proposed temporal scalability coding for STMAC coding.

The second task is to assign a proper set of "quantization parameters" (QP) and "frame rate" (fps). Note that quantization step size (QS) is equal to twice the quantization parameter. As we discussed, 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 to meet a target bitrate which affects QS and fps of each model areas. For convenience, we define a "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 $S = (S1, S2, S3, S4)$, and the temporal freqency fps as $T = (T1, T2, T3, T4)$. Note that the eye area spatial QS is quite fine and lip area temporal fps is high enough compared with that of regular compression case.

We now need as a basic functionality to suspend a macroblock's transmission for the selected regions. Most of MC-DCT encoder/decoder pairs such as H.261, H.263, MPEG-1, and MPEG-2, etc. have a "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. Figure 4 also shows the STMAC coding concept and the basic functionality in the H.263 video telephony example.

The third task is to accomodate the "motion" of the model in STMAC coding. We define Motion Adaptive STMAC coding as a technique which is to accomodate the motion of the model in previous STMAC coding. The reason for this is that the motion cannot be described until the next refresh occurs for each model area. To overcome this, we refresh and encode the moving area within the model when there is a big move. If there is not a big move, we use the regular STMAC coding. The decision of movement is made based on a global motion vector, which is calculated by the relative motion of ellipse centers in two consecutive frames. One potential way to take a representative global motion vector is to choose the relative motion of the lip area, since it has the highest temporal refresh frequency. Our experiment shows ellipse centers give better solutions, since automatic area detection introduces some noise to the exact location information of lips. However, the ellipse finding algorithm is much more robust and so we use the ellipse center as a motion decision criterion.

For the proposed Motion Adaptive STMAC coding, there is a restrictive syntactic component especially in H.263. Note that in H.263 case "difference quantization parameter (DQUANT) " 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

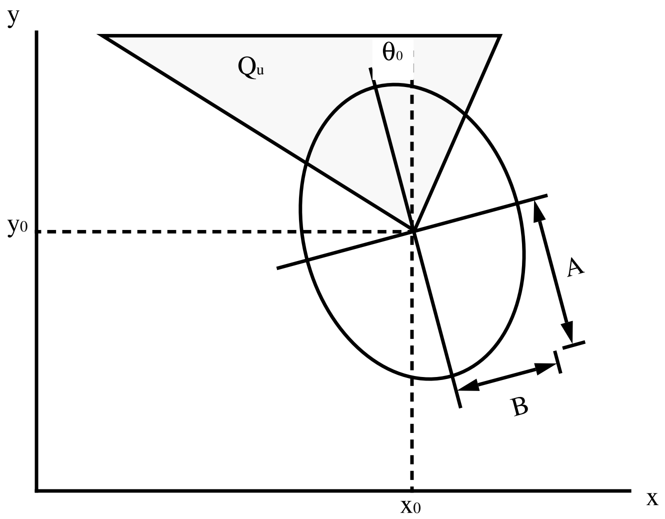

Figure 5: Elliptical face location model.

be exceeded more than 3. However, for the background we assign a QP of 31 by default. As easily expected, if not, we would transmit the background 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 eyes, 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 gradually upgraded from a bottom quality (QP = 31), because there is no previous frame from which to copy the other model areas. 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.

3 AUTOMATIC AREA DETECTION AND MOTION ADAPTATION FOR THE MODEL

The detection of head outlines as well as outlines of persons in still or moving images has been the object of active and recent research in computer vision.[3] Only very recently was it realized that such tasks, when performed totally automatically, could be helpful in low bitrate coding envirnments.[3] The task of detecting face locations in a sequence of images is facilitated by both the fact that people's head outlines are consistently roughly elliptical, even when the persons appear in a profile, and by the temporal correlation from frame to frame. The task of detecting the eye-nose-mouth region is facilitated by the axial symmetry inherently present in a human face for a person facing the camera, or looking slightly to the side, and appearing in projections in successive frames of a video signal. In this section, we use the same "automatic tracking and detection" procedure of conventional MAC coding in[3,2] with slight modification.

The model we adopt in order to represent the location of a face is simply that of an ellipse ε, as shown in Figure 5, characterized by a center (x_0, y_0), the lengths of its minor and major axes A, B, and a 'tilt' angle θ_0. Although the upper (hair) and lower (chin) areas in actual face outlines can have quite different curvatures, ellipses provide a good trade off between model accuracy and parametric simplicity. Moreover, due to the fact that this information is not actually used to regenerate the face outline, a small lack of model-fitting accuracy does not have any significant impact in the overall performance of the coding process.

Since an elliptical head outline can in some cases provide only a rough estimate of the face location, we have

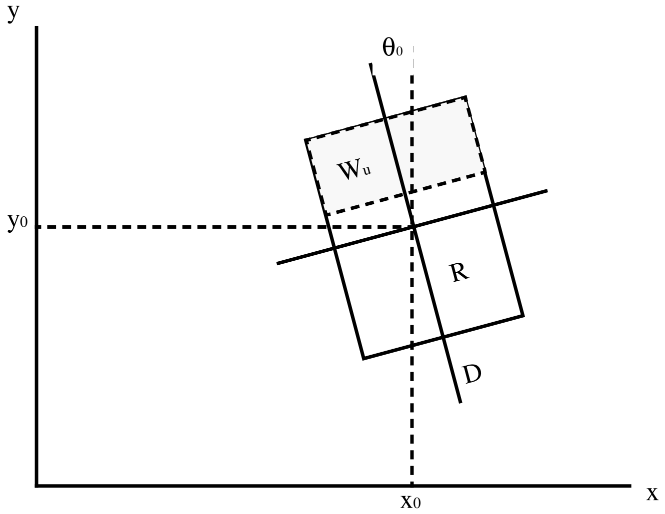

Figure 6: Rectangular window as eyes-nose-mouth location model.

chosen to refine this elliptical location model by identifying a rectangular region W inside the ellipse, which tightly captures the eyes, nose, and mouth of the person in the scene. This is depicted in Figure 6 with an extra degree of freedom which we introduce by allowing a slant of its vertical axis. This additional degree of freedom ensures that the detection will be robust in the (very frequent) case of slight head rotation. The upper third of window, denoted by W_u, is further identified to contain the eyes and eyebrows-two most reliably symmetric features in a human face. The window W is entirely characterized by a center (x_1, y_1), width w, height h, and slant angle θ_1.

The first step is getting a downsampled binary edge of input images which is used for head outline detection and feature traking as we proposed in.[3] Second step is applying the head outline detection algorithm onto the downsampled binary edge image. The algorithm detects the outline of a face location geometrically modeled as an ellipse, using as preprocessed input data binary thresholded gradient magnitude images. Our face location detection algorithm was designed to locate both oval shapes (i.e., 'filled') as well as oval contours partially occluded by data. The algorithm is organized in a hierarchical three-step procedure: coarse scanning, fine scanning, and ellipse fitting. A final step consists of selecting the most likely among multiple candidates. This decomposition of the recognition and detection task in three steps, along with the small input image size, make the algorithm attractive for its low computational complexity. Each step is described and illustrated in.[3] Note that there is no specific algorithm for the shoulders area above. Since we need one more area for shoulders, we choose the area by a artificial rectangle by around 3 times the width of face for convenience and upper starting point of the height by the second lowest lip area in macroblock so that we don't need to find the shoulder area using any kind of specific algorithm. Note that our objective of considering the shoulder area is for the absorbing the movement adaptively. Therefore, this shoulder doesn't affect the output rate trends (target bitrate) until a relatively big motion happens. Figure 7 depicts a thresholded image resulting from the previous algorithms for face outline and Figure 8 shows a bilevel morphological erosed operation result in order to improve the face symmetry for the eyes-nose-mouth area. The area inside the closest boundary rectangle for upper one third W_u is considered as eye area, while other two third rectangle area is thought as lip one. Similarly, the area inside the closest boundary head outline is considered face area. Figure 9 shows macroblock labeling of each automatic area detected image. If there is any ambiguity in labeling macroblocks between eyes and nose regions, we assign the macroblocks into the lip area. Figure 10 shows the shoulder area artificially taken. In our example in Figure 9, the number of the lowest macroblocks of lip area is 3. Therefore, 9 macroblock is the width for shoulders area, while the starting point of the area begins right under the lip area macroblocks. Note that now we have all model areas by previous algorithms in[3] with the consideration of shoulders area.

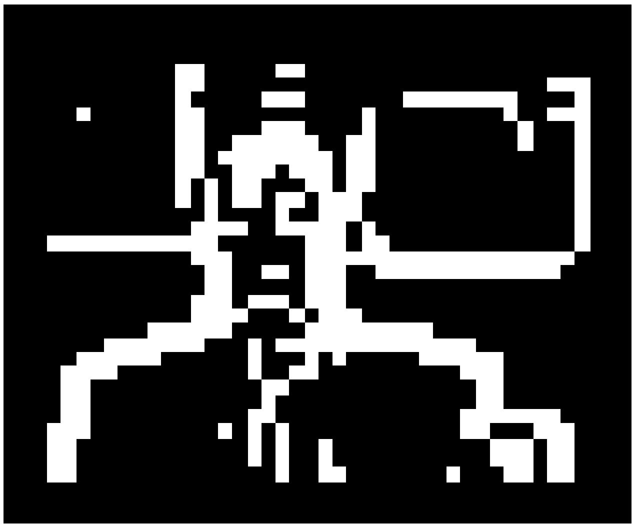

Figure 7: Thresholded edge image for face ellipse detection.

Figure 8: Bilevel morphological erosed image for face features detection.

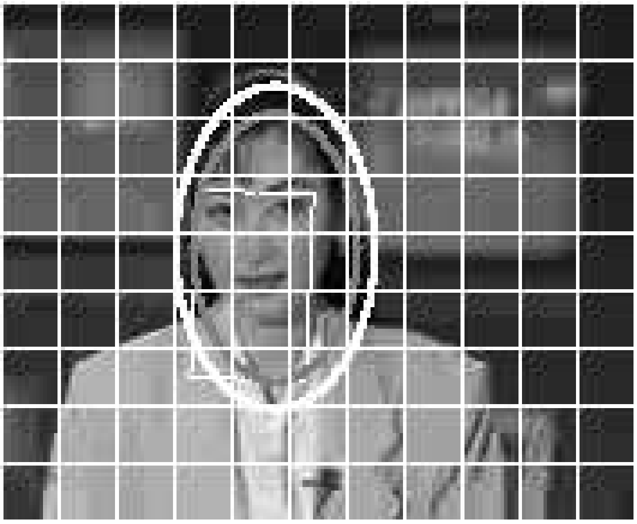

Figure 9: Automatic area detected image.

For the motion adpatation, first of all, we perform the model area detection. Then, if there is any big motion of human face, it will refresh original regions and obsolete areas which are not perceptually pleasing to human eyes. If the motion is relatively small, we apply regular STMAC coding technique. Note that this is similar to conventional "Conditional Replenishment" coding technique.[7]

For the background area in MA-STMAC codig, we also need a basic functionality to suspend a macroblock data transmission of the selected regions as is in regular STMAC. Once again, we use a " not coded" mode for decoders just to copy the macroblock from the previous reference frame at the exactly same position.

We encode the entire model area in order to absorb the motion of human face and refresh the obsolete background. Therefore, we use regular STMAC and use MA-STMAC with selective area refresh when the motion is big.

4 TEMPLATE ADAPTIVE RATE CONTROL

In the MA-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, we have 4 different step sizes for 5 different areas: eye, lip, face, and backround area which includes the shoulder area. In addition, we also have 4 different frame rates for the temporal model: for the same areas such as eye, lip, face and background which includes shoulder. For MA-STMAC coding rate control, we should control these 8 different spatial scalabilities (QS) and temporal scalabilities (fps), given a target bit rate. In our simulations, we use the same rate control scheme, the Template Adaptive Rate (TAR) control [6], 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, lip, face, 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 order to obtain this table, we use the exact same rate control algorithm as TMN4, and then applied this formula to obtain template index TI:

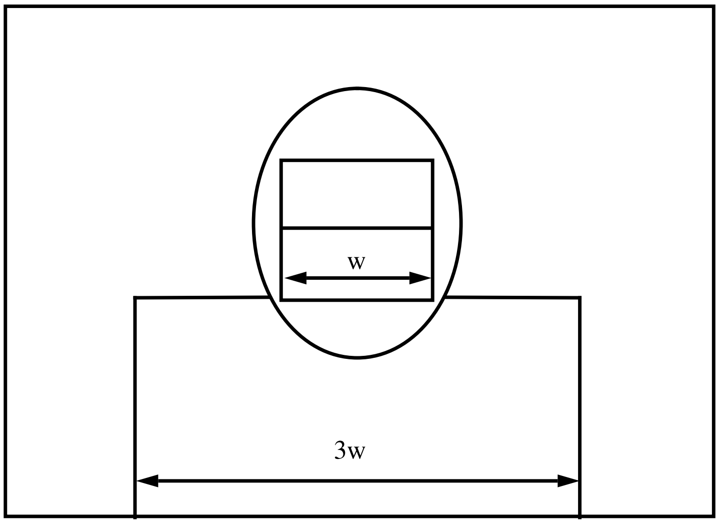

Figure 10: Artificially taken shoulder area.

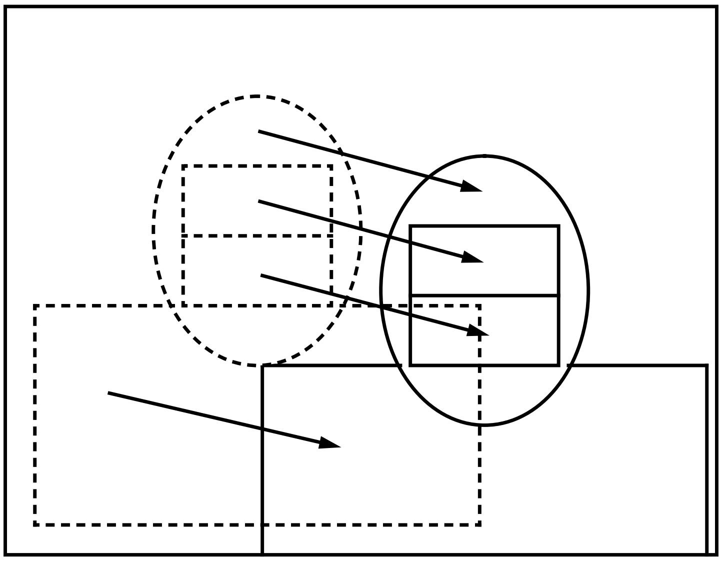

Figure 11: Selective area refresh for the moving human face model.

$$TI_i = \left\lfloor \frac{PQ_i}{4} \right\rfloor \tag{1}$$

Note that quantization step size (QS) is twice the quantization parameter (QP). Since the range of the quantization parameter is limited by 31, a given TI is also bounded by 8.

5 CONCLUDING REMARKS

The compression efficiency of MA-STMAC depends on the base codec efficiency. For example, if we have a very efficient H.263 encoder in which 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.

We provide simulation results comparing the MA-STMAC scheme with conventional H.263 compression. We made a moving Akiyo sequence artificially from foreground and background Video Object Planes (VOPs) which are given as a MPEG-4 test sequence. Our experiments show that we can make better quality video sequences around face area using MA-STMAC coding at 70Kbps in Figure 13 with consideration of motion. Figure 12 depicts a conventional H.263 encoder result operating at 78Kbps with the same test video sequence. Note that MA-STMAC is transmitting 30 frames per second, while H.263 is transmitting 10 frames per second. Conventional H.263 sequences have less quality than MA-STMAC, since it encodes every area including the background as well. In MA-STMAC, background macroblock is not encoded, just copied from previous frame thus saving a lot of bits in order to assign more bit budget to foreground. Note that we use a general model for STMAC which has 4 different bit budget areas. A simpler case includes two areas of face and background. However, the policy of assigning bits is most important work in the STMAC coding. Sometimes, if we assign very low frame rates and give much finer QP for eye area, there may be perceptually noticeable artifacts (discontinuity). From this case, we recognize there is another factor to be considered, which is "area discontinuity in motion". This is especially important for face area including eyes and lips since these areas are the focus of human observers. We believe this is a trade-off problem. For example, if we use two area model(in which face has finer QP and more frequent frame refresh rate and background has rougher QP and less frequent frame refresh rate), there are no discontinuity artifacts. But that is not the best policy for improving compression. In the contrary, if we use these four areas (which correspond to the eyes, lips, face, and background), we could save the maximum number of bits, but sometimes experience "area discontinuity in motion".

In conclusion, for very low bit rate image coding area selectivity is inevitable. Once an appropriate model is selected and matched to the input image, we assign various spatial and temporal scalabilities in each of the area: 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. Since the different temporal scalabilities, we need to consider the model movement. In this paper, we try two basic modification on conventional STMAC coding. First modification is for "automatic model area detection". Second modification is for "motion adaptation of model". The MA-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 considering motion adaptation.

Acknowledgements

The authors wish to thank Sunil Patel for performing some of experiments.

Figure 12: Conventional H.263 at 78Kbps (10 fps).

Figure 13: Motion Adaptive STMAC at 70Kbps (30 fps).

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.

6 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.

[6] J. B. Lee and A. Eleftheriadis. Spatio-temporal model-assisted compatible coding for low and very low bitrate video telephony. *IEEE Proc. ICIP '96*, September 1996.

[7] A. Netravali and B. Haskell. *Digital Pictures: Repesentation, Compression, and Standards.* Plenum Press, New York, 1994.