H.263 Facial Video Image Coding with the MELP Voice Codec

Kunio Takaya, Li Ding and Naofumi Aoki*

Electrical Engineering, University of Saskatchewan
57 Campus Drive, Saskatoon, Sask. S7N 5A9 , Canada
Fax: (306) 966-5407
Telecommunications Research Laboratories (TRLabs)
108-15 Innovation Blvd., Saskatoon, Sask. S7N 2X8 , Canada
Fax: (306) 668-1944

Abstract—
The International Telecommunication Union has put forward a recommendation H.263 "Video Coding for Low Bitrate Communication". This recommendation defines the rules for compressing the moving picture component of audio-visual services at low bit rates. The H.263 video codec developed from MPEG-1 is primarily designed for low bit rate applications. However, there is still room to improve its encoding speed if its application is limited to videophone and videoconferencing which uses the internet or ISDN phone lines. The basic encoding operations involved in motion compensation, discrete cosine transform (DCT) and bit rate control were modified and tailored specifically for facial video images. The strategies to improve the encoding speed are presented in this paper. Experimentally, we achieved an ecoding speed approximately 10 times faster than that of the generic H.263. While imposing a restriction on the video object bing facial images, a low bit rate sound layer based on the MELP vocoder was added by sparing 2.4 kbps for voice communication as a part of H.263 video codec. Fluctuations both in terms of the waveform and pitch period of voiced speech was incorporated in the MEPL vocoder to enhance the naturalness of sysnthsized speech.

I. Introduction

The purpose of video image compression is to cram more image information into less space and bandwidth. The fundamental idea behind video compression is to remove both of the spatial and temporal redundancy, to reduce the data rate, ideally without introducing visual quality degradation.

MPEG is a very popular compression format in which two basic techniques are used. One is block-based motion compensation for the reduction of the temporal redundancy and another one is DCT transform coding for the reduction of spatial redundancy. In recent years, many research efforts have been made to fit the concept of MPEG, which was intended for the use in digital TV and DVD (Digital Video Disk), into the low and very low bit rate applications such as Internet video broadcasting, video conferencing, and videophone in PCS (Personal Communication Systems). The International Telecommunication Union (ITU) has put forward a new standard H.263 [1] which has been developed from MPEG-1 by reducing the bandwidth requirement to fit in that of the standard telephone lines. The basic configuration of the H.263 video source coding algorithm is based on ITU-T recommendation H.261[2], but with some changes to improve performance and error recovery. The H.263 standard is a widely used international video compression standard for video telephony applications. Due to this bandwidth limitation, the standard picture size for H.263 is CIF(352x288) or QCIF(176x144).

There are two issues to be addressed and resolved to successfully implement in real time. One is the issue further increase the compression gain. A CIF video is equivalent to 36.495 Mbps if frame rate is 15 frame/s. The best compression gain achievable without significantly degrading the picture quality, is about 700:1. Even with this much of compression, the bit rate is still well above 32 kbps (full duplex in ISDN). The other issue is encoding speed. H.263 and other MPEG based video compression techniques perform very sophisticated tasks requiring processing time. This is true more in encoder than in decoder. The implementation of H.263 standard can be hardware-based or software-based. In this paper, we develop strategies to improve the encoding speed important for real-time videophone and videoconferencing, and at the same time to further compress video data into the H.263 bit stream if possible. For this particular application, the scene of the video sequences is usually composed of a moving head and one still background. We focus on H.263 specific to facial images.

H.263 does not specify how to handle sound as other MPEG standards define how its sound layers should be structured. For example, MPEG-1 aiming at the DVD bit rate of 1.5 Mbps uses perceptual subband/transform coding which exploits the characteristics of the ear, i.e. psychoacoustic masking effect. The sound bit rate is one of 192 kbps, 128 kbps and 64 kbps. In the low bit rate case of H.263, the allowed bit rate is 32 kbps in full duplex and 64 kbps in half duplex. Unless the target of sound compression is limited to human voice, it is practically impossible to include sound within the framework of H.263. Since we limit ourselves only to facial images for video, it is natural to limit sound only to human voice. Among various speech coding methods, namely, waveform codecs, LPC (Linear Predictive Coding) vocoder and Hybrid codecs, we chose MEPL (Mixed Excitation Linear Prediction) codec as the most suitable codec to go along with H.263 video codec. MELP can produce speech comparable to CELP (Code Excitation Linear Prediction) operating at 4.8 kbps

* Visiting scholar from Research Institute for Electronic Science, Hokkaido University, N12 W6 Sapporo, Japan

in cellular phones at a lower bit rate of 2.4 kbps.

Since motion vector search, DCT transform and bit rate control are the heart of the H.263 video encoding, we describe the modifications made and some intelligence added to each of the basic encoding operations to improve the compression gain as well as to achieve faster encoding speed. Then, we discuss further improvements possible for MELP speech codec.

II. Face Tracking

Since facial expressions and head movements are of our primary interest in video sequences of a talking person, we keep track of the movements of a face within the video frame. The face of a talking person on a videophone occupies only a subspace of the video frame. In order to show the head movement or changes in facial expressions, it is not necessary to send one whole frame. Only the area where a face is residing, i.e. a subsize area is sufficient and the rest of the frame can be considered as a background. This allows us to drastically reduce the information to encode and transmit while keeping the original image size still the same. If once the encoder sends the whole image including the background to the decoder, the encoder can discard the unnecessary still background until any significant background change will occur. The face tracking must be implemented in the stream of H.263 video encoding process on the ongoing basis. This face tracking requires head boundary analysis by means of edge detection techniques.

One generic approach to edge detection is differential detection. There are two types of differential edge detection: first order derivative and second order derivative. For the first order derivative edge detection method, spatial first differentiation is performed, and the resulting edge gradient is compared to a threshold value. An edge is judged present if the gradient exceeds the threshold. For the second derivative edge detection method, an edge is judged present if the second derivative changes its sign and crosses zero [3]. A crude head detection based on the first derivative is employed to keep the time for face tracking minimal. A simplified orthogonal gradient edge detection method used to detect the head is defined by equation (1).

$$G(j,k) = |G_r(j,k)| + |G_c(j,k)| \quad (1)$$

$$\text{Where,} \quad G_r(i,k) = F(j,k) - F(j,k-1),$$
$$G_c(i,k) = F(j,k) - F(j+1,k).$$

G_c is the difference between adjacent row pixels and G_r is the difference between adjacent column pixels. After the edge gradient is formed according to equation (1), the gradient is then compared to a threshold to determine if an edge exists. For the Claire image, the threshold value equal to four times the average pixel value over the image works well. Figure 1 (a) is a Claire image. Figure 1 (b) gives the edge detection result after thresholding.

Since H.263 standard treats Macroblocks as a basic compression unit, the head detection needs to be accurate only to the extent that one can tell which macroblocks contain the head boundary. The macroblocks where the head boundary is residing can be found by searching the head outline image resulting from the edge detection, from outside inward starting on the left side, right side and top of the head. After the left most, right most and top most positions of the outlined head boundary are found, we take advantage of the fact that the proportion of the width to the height is about 7:9 to find the bottom of the head. All the macroblocks containing the head are thus found. Figure 1 (c) is a Claire image which shows the macroblocks within the image. Figure 1 (d) shows all the macroblocks containing the head. Note that the head detection method keeps extra surrounding macroblocks in order to make room for head movement. This edge detection method assumes that the background is a plain background. If the background contains patterns or objects, the edge detection would result in erroneous head edges. The proposed head detection method were tested to work for the Claire Image and other images having a plain background. The past history of motion vectors can provide useful information as to which macroblocks have moved associated with the head movement. This helps identify the macroblocks which contain the head.

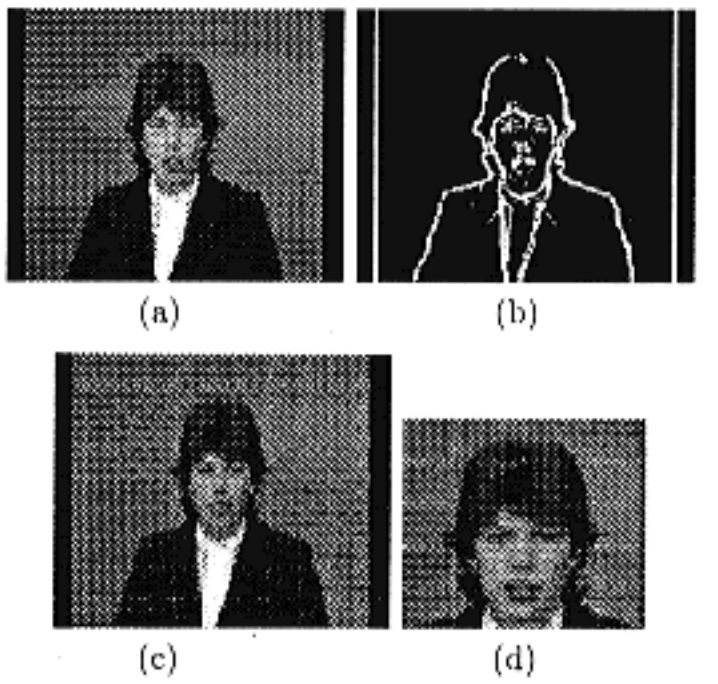

Fig. 1. Head detection for face tracking method, (a) A Claire image, (b) Detected edge image, (c) Macroblocks in the Claire image, and (d) Macroblocks containing the head.

III. Fast Motion Estimation

One important component in H.263 video data compression is block-based motion vector search which plays a key role in reducing the temporal redundancy applicable only to video images. Since video frames represent a series of snap shots successively taken with a time interval of 1/30 seconds, they are very similar, particularly between two contiguous frames. It is possible to predict any one frame from its previous frame. Noticeable differences exist only in the locations moving objects included in the two pictures.

In H.263, frames are logically divided into 16x16 pixel

macroblocks for which motion compensated prediction is performed. Motion in the picture usually implies that a group of pixels in the previous picture will be moved to a different position in the current macroblock. This displacement is called motion vector which is encoded in the bitstream. The motion vector associated with a macroblock is obtained by matching a 16x16 macroblock in the reference frame. The data in a macroblock will be more compressed if the motion vector plus the difference between the predicted macroblock and the current macroblock are encoded. However, the motion estimation which determines the best matched macroblock and is performed by the encoder presents a time consuming computational challenge. It is one of the reasons why the encoder is slow.

In the video sequences of a talking person in front of a video camera, when the person moves his/her head, all the macroblocks in the head area move in the same direction. Facial expressions are shown by the eyes and mouth. We can take this observation into account to perform the motion vector search. Motion vector search can be performed in much more restricted regions of eyes and mouth which contain far more details related to facial expressions than other regions. The motion vectors for other regions of the head such as forehead and hair areas can be assumed to be the same as the motion vectors found in the regions of mouth and eyes. For other macroblocks, since they are mainly the background, we can set the motion vectors as zero.

After the face tracking, we can roughly estimate which macroblocks cover the eyes and mouth within the macroblocks containing the head. Figure 2 (a) shows a subsize image where the head boundary are found. Figure 2 (b) shows the macroblocks which contain the head and Figure 2 (c) shows the motion vector search and prediction for the macroblocks within the head area. In Figure 2 (c), MS means that Motion vector Search is performed for the current macroblock, PS is Predicted Motion vector meant to use the motion vector found from MS macroblocks. For other macroblocks, we set the motion vectors zero.

Another important feature of the videophone sequence is that there are only small head movements from one frame to another frame, thus the motion vector search range can be reduced to a small range in order to save the search time. For example in the Claire image, the search window was reduced to 5 from the default search window size of 15. See Figure 3. This searching procedure specific to the facial images reduces the computation time necessary to find motion vectors.

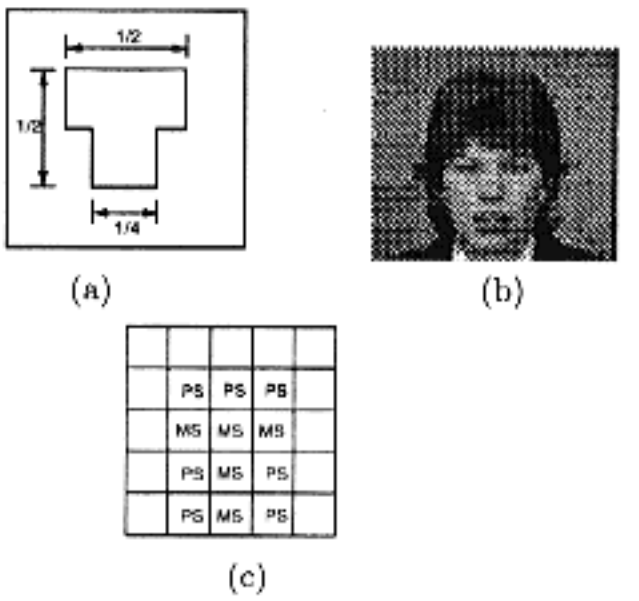

Fig. 2. Fast Motion Estimation, (a) Subsize image containing the head boundary, (b) Macro blocks containing the head, (c) MS: Motion vector search area, PS: Prediction search area

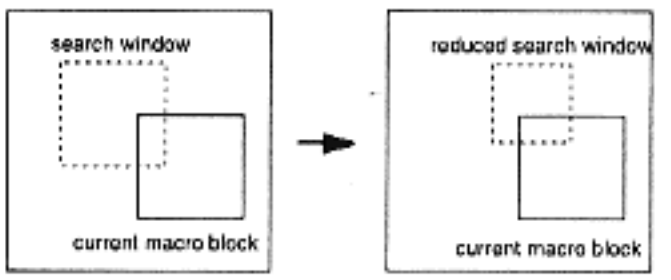

Fig. 3. Reduction of search window range

IV. Fast Ddiscrete Cosine Transform

DCT (Discrete Cosine Transform) is a mathematical transform which translates digitized image data from its spatial domain into the frequency domain. In H.263, a picture is divided into 16x16 pixel macroblocks, each of which is further divided into four 8x8 luminance blocks. No subdivision is applied to chrominance. The DCT is performed on each of the 8x8 luminance blocks and on two 16x16 chrominance blocks. Given a block of 2D data, the DCT calculates 2D frequency components. The rationale behind the DCT is that picture information thinly spread thinly over a large number of pixels becomes more concentrated around the lower frequency components after the transformation. Low frequency components appear in the upper left corner of the DCT block as shown in Figure 4. The lower right values represent higher frequencies.

High frequency components of an image represent the details of the image which may not be required to reproduce its approximation since high frequency components are usually very small - small enough to be neglected with little visual distortion. So, the DCT coefficients are zig-zag scanned from the lowest frequency to higher frequencies until the coefficient value becomes less than a selected quantization level. Thus, the selection of the quantization level play an important role to control the trade-off between

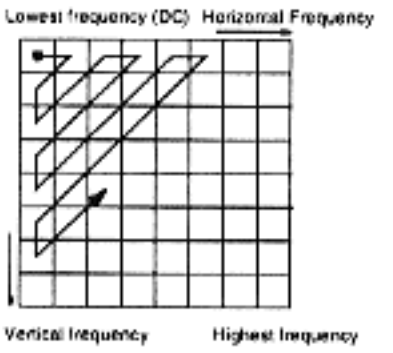

Fig. 4. A 8x8 DCT block and its zig-zag encoding sequence

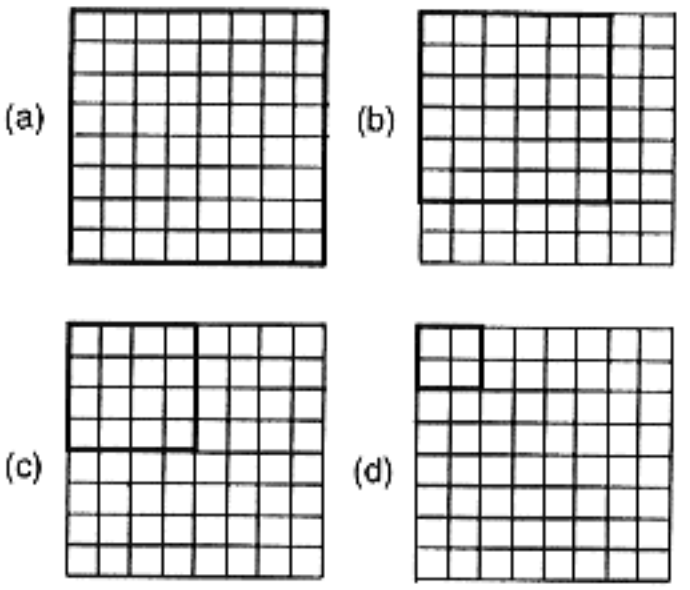

Fig. 5. Fast DCT Transform

image quality and compression gain.

DCT transform is a heavy computational burden in H.263 encoding process, since every block in every frame needs the DCT. If we can reduce the computation time for the DCT, the encoding speed can be improved. We have found that the smaller is the difference between the current macroblock and the corresponding predicted macroblock, the more zero coefficients will result in the DCT. Since those zero coefficients are in high frequencies, we only need to compute a limited number of coefficients but not all of the DCT coefficients. According to the magnitude of the difference between the current macroblock and predicted macroblock in P and B frames, we choose one of the three different sizes of DCT transform: 2x2 point DCT, 4x4 point DCT and 6x6 point DCT as depicted in Figure 5. Smaller point DCT is performed depending on the magnitude of the difference. For facial images in general, the magnitude of the difference is not large since the head does not move quickly from one frame to another. Since many DCT coefficients are zero for high frequencies, performing 6x6 point DCT can maintain almost the same image quality as that produced by 8x8 point DCT.

The difference is smaller in less detailed regions such as background, shoulder and hair. It is greater in the regions of eyes and mouth, which have been already identified for motion vector search. So, the size of the DCT is varied depending on which part of the facial image is transformed by the DCT. The encoding speed is thus improved by using the variable DCT block size. This approach however does not significantly change the compression gain or bit rate.

V. Bit Rate Control

Bit rate control is important in the encoder to achieve the target bit rate requirement while maintaining the good video quality. H.263 does not specify a bit rate control method[1]. The bit rate in H.263 can be fixed or variable. Depending on how a generated bit stream is utilized, constant bit rate (CBR) control or variable bit rate (VBR) control can be chosen [4]. In applications to use a file to store the bit stream, VBC is more adequate since H.263 decoder can reproduce video frames faster than the encoder. The decoder can display the video clip in sync with the real timing. In videophones, the speed of the transmission media is constrained by a certain bit rate. None of the frames are allowed to exceed one frame time strictly speaking. In practice, the average bit rate of H.263 output cannot exceed the bit rate of the channel. CBR is suitable under circumstances.

For the CBR control method, the quantizer scale is controlled to achieve the target bit rate by considering a hypothetical rate control buffer at the encoder's output. The buffer occupancy level is used as feedback to control the quantizer scale. In the software simulator for H.263 standard developed by Telenor Research and Development in Norway[4], both of the two methods, VBR and CBR are implemented. The default bit rate for VBR control suitable for low delay teleconferencing video is set according to the TMN5 document[5]. The CBR method, which is standard in MPEG-4, uses a fixed frame rate to achieve the target bitrate as a mean bit rate for the whole sequence. For our H.263 encoder tailored for videophone applications, the CBR control can be simplified in order to improve the encoding speed.

Clearly, when more bits are assigned to encode a given scene, the image quality is better. The smaller a step size of quantization, the more bits are produced by the encoder and the better is the image quality. For the facial images, the eyes and mouth need details to show facial expressions, thus require more bits to be allocated. A finer quantization scale is applied to the region of eyes and mouth, whereas a coarse quantization scale is applied to the background and shoulders not requiring much details. For the remaining regions such as hair, the quantization scale is updated according to the target bit rate.

For VBR control, the default value of the quantization step, which applies to the DCT coefficients in macroblocks, is 10. It can be adjusted from 1 to 31 so the average bit rate over a sequence can stay close to the target bit rate. The encoding speed of VBR is faster than that of CBR as this method needs not to consider the complexity of the scene in a frame. If the quantization step is properly chosen, VBR can produce a bit rate close to the desired, keeping image quality consistent throughout the entire video sequence.

VI. Experimental Results

This section gives the results of the experiments conducted to test the performance of the added features to H.263 to deal with facial video images. The experiments were mainly to compare the encoding speed and image quality.

Since the achieved encoding speed will vary greatly depending on which platform the program runs on, the experiments were conducted on DEC 3000 AXP and Alpha Station 255. The video image size used in the experiments were all QCIF (176x144) which is the standard H.263 input size. The experiments were done using the software simulator modified to include (1) fast face tracking, (2) fast motion estimation, (3) fast DCT and (4) bit rate con-

TABLE I
Testing on DEC 3000 300AXP for CBR Control

Bit Rate: 28.8 Kbits/sec		
Methods	Encoding Speed (fps)	Image Quality SNR Y Cb Cr (dB)
Original	1.1	37.7 38.6 40.3
Face Tracking	3.8	37.7 38.6 40.3
Fast Motion Est.	5.9	37.4 38.3 39.9
Fast DCT	7.0	34.7 37.4 40.1
Bitrate Control	7.4	34.1 36.6 39.2
Bit Rate: 64 Kbits/sec		
Methods	Encoding Speed (fps)	Image Quality SNR Y Cb Cr (dB)
Original	1.1	41.9 42.2 43.5
Face Tracking	3.8	41.9 42.2 43.5
Fast Motion Est.	5.5	41.6 41.9 43.2
Fast DCT	6.2	35.4 38.9 42.2
Bitrate Control	7.5	33.9 37.4 40.2

TABLE II
Testing on Alpha Station 255 for CBR Control

Bit Rate: 28.8 Kbits/sec		
Methods	Encoding Speed (fps)	Image Quality SNR Y Cb Cr (dB)
Original	1.8	37.7 38.6 40.3
Face Tracking	6.3	37.7 38.6 40.3
Fast Motion Est.	10.3	37.4 38.3 39.9
Fast DCT	13.0	34.7 37.4 40.1
Bitrate Control	13.3	34.1 36.6 39.2
Bit Rate: 64 Kbits/sec		
Methods	Encoding Speed (fps)	Image Quality SNR Y Cb Cr (dB)
Original	1.9	41.9 42.2 43.5
Face Tracking	6.4	41.9 42.2 43.5
Fast Motion Est.	9.8	41.6 41.9 43.2
Fast DCT	12.5	35.4 38.9 42.2
Bitrate Control	13.9	33.9 37.4 40.2

trol in the H.263 standard originally developed by Telenor Research and Development. Although the encoder of the software simulator was not optimized for speed, the encoder ran reasonably fast.

Table 1 and Table 2 show the encoding speeds obtained by DEC 3000, 300AXP and Alpha Station 255, respectively. The CBR control method was specified in this experiment. The newly introduced methods specific to facial video images replaced the equivalent component in the original program one at a time. The row of "fast motion estimation", for example, used the fast face tracking (background elimination) and the fast motion estimation in place of the original functionality.

It is apparent from Table 1 and Table 2 that Alpha station 255 provides twice as large encoding speed as DEC

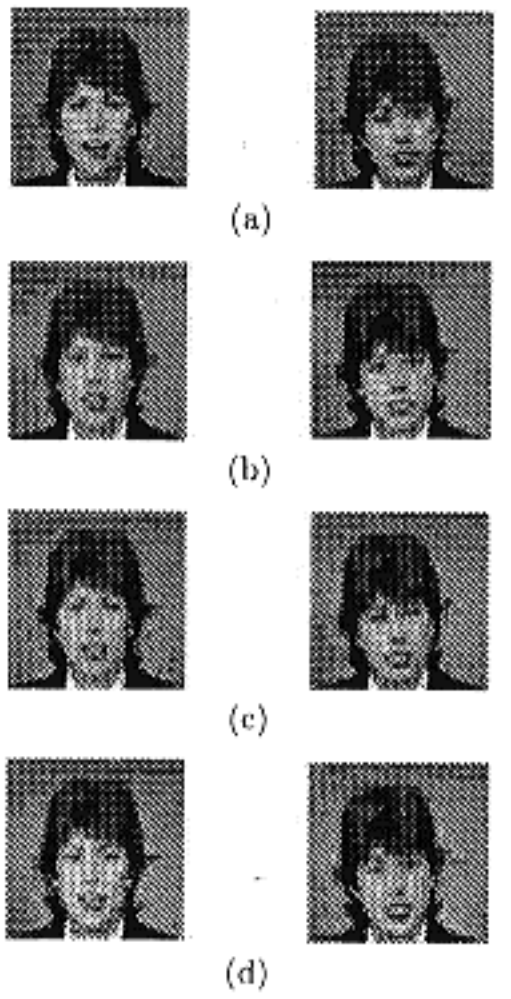

Fig. 6. Comparison of image quality for CBR control for two images, (a) after face tracking, (b) after fast motion estimation, (c) after fast DCT transform, and (d) after bit rate control.

3000 can provide. The facial image specific methods increasingly speed up the encoding speed as more functions are added. Image quality is moderately decreased by adding the facial motion vector estimation and the spatially constrained DCT. Human eyes are, however, not sensitive to such image degradation as reflected to SNR listed in the tables. Figure 6 shows two sets of Claire's images which went through the four facial image specific processings mentioned above.

Table 2 compares the encoding speed between VBR control and CBR control methods. The VBR control tends to produce a longer bit stream because of the fixed quantizer step size. This in turn decreases the bit rate. Since there is no need to check the buffer occupancy level, VCR is faster. The proposed methods applicable only to the facial video images increased the encoding speed approximately by one order of the magnitude. The compression gain on the other hand depends more on the quantization step size and the acceptable image quality.

VII. Sound Layer

Unlike MPEG-1 designed for 1.5 Mbit/sec. of which 350 kbps is reserved for audio, H.263 low bit rate video encoder does not define sound layers specifically. Due to its targeted applications typically at 64 kbps, H.263 cannot accomodate MPEG-I sound codecs which occupy 192, 128 and 64 kbps per channel for the sound layer-1, 2 and 3 respectively. Choice of a sound codec for H.263 is thus limited

TABLE III
Comparison of VBR and CBR Control on Alpha Station 255

Methods	Bit Rate for VBR (kbits/s)	Encoding Speed (fps)	Image Quality SNR (dB) Y Cb Cr
Original	26.9	1.8	32.8 34.4 36.3
Face Tracking	18.2	6.9	32.8 34.4 36.3
F. Motion Est.	18.6	12.3	32.7 34.4 36.2
Fast DCT	17.3	15.1	32.2 34.3 36.4
Methods	Bit Rate for CBR (kbits/s)	Encoding Speed (fps)	Image Quality SNR (dB) Y Cb Cr
Original	28.8	1.7	37.7 38.6 40.3
Face Tracking	28.8	6.3	37.7 36.8 40.3
F. Motion Est.	28.8	10.4	37.4 38.3 39.9
Fast DCT	28.8	13.0	34.7 37.4 40.1

to the speech codecs, such as waveform codec, LPC (Linear Predictive Coding) vocoders, or hybrid codecs. The voice quality of the LPC vocoder is known to be poor compared with that of hybrid vocoders. CELP (Code Excitation Linear Prediction) used in cellular phones works at 4.8 kbps with satisfactory intelligibility. Another recently proposed scheme called MELP (Mixed Excitation Linear Prediction) claims 2.4 kbps providing intelligible and natural human speech, drastically improved from the classical LPC vocoders. The heart of the MELP is the mixed excitation comprised of periodic pulse train and white noise, which is applied to the all-pole vocal tract LPC model. As shown in Fig. ??, additional functional blocks of (1) periodic or aperiodic pulses, (2) adaptive spectral enhancement, (3) pulse dispersion filter are used to mimic the characteristics of natural human speech. The mixed excitation is realized by two sets of a filter bank consisting of five bandpass filters. One filter band is for the periodic pulse train and the other for white noise. These two filter banks are complementary to each other maintaining the sum of the two gains to be constant for each subband.

Pitch period fluctuation represented as jitter in pulse train causes significant influence to the naturalness of synthesized voice. This pitch period fluctuation is always observed even in the steady part of voiced human speech. Although the early MELP vocoder [6] has provided a control block to insert jitter to the periodic pulse, the pitch period fluctuation that follows the $1/f$ spectral characteristic [7] is included in our MELP vocoder to enhancement the naturalness compared with employing a completely random sequence.

The generic MELP vocoder does not take care of the waveform fluctuation in the steady part of voiced speech. Our experimental result shows that waveform fluctuation, which is viewed as cyclic waveform changes, is also significant factor for the naturalness. The $1/f^2$ frequency characteristic that describes Brownian motions represents the waveform fluctuation. Psychoacoustic experiments [7] confirmed that the synthesized voice more naturally sounds in our vocoder which positively inserts $1/f^2$ noise to the periodic pulse train.

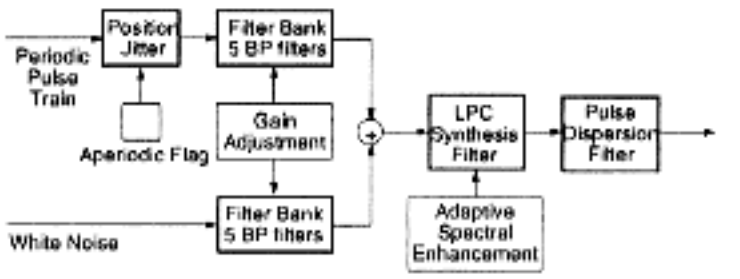

Fig. 7. Block Diagram of MELP Voice Codec (Synthesizer)

VIII. Conclusions

In this paper, we have presented the strategies to improve the encoding speed of H.263 low bit rate video codec dedicated specifically to facial video images keeping the videophone applications in mind. The priori knowledge about the object being a face allowed us to bring in some intelligence pertaining face tracking, fast motion estimation, fast DCT transform and bit rate control to accomplish a faster encoding speed. The improvement in the encoding speed, about 10 times faster, is significant for implementing such H.263 either in hardware or in software. Since these proposed methods are implemented in the general framework of the H.263 video encoding scheme on the ongoing basis, they are fully compatible with the H.263 standard. Taking advantage of the nature of the videophone, an improved version of the MELP vocoder was incorporated into this facial video specific H.263 by sparing 2.4 kbps allowing simultaneous inband voice communication.

References

[1] ITU-T Recommendation H.263, video Coding for Low Bitrate Communication, International Telecommunication Union, May, 1996.
[2] CCITT Recommendation H.261, Video Codec for Audiovisual Services at px64 kbits/s, December 1990.
[3] William.K.Pratt, Digital Image Processing,2nd ed., Toronto: John Wiley and Sons Inc, 1991.
[4] Karl Olav Lillevold, The Software of a Very Low Bitrate Video Encoder Producing H.263 Bitstream, Telenor Research and Development, Norway, June 14, 1996.
[5] ITU Telecommunication Standardization Sector LBC-95, Expert's Group on Very Low Bitrate Visual Telephony, Video Codec Test Model, TMN5, Telenor Research (TR), Jan. 31, 1995.
[6] Alan V. McCree and Thomas P. Barnwell III, A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding, IEEE Trans. on Speech and Audio Processing, vol. 3, no. 4, pp. 242-250, 1995.
[7] N. Aoki and T. Ifukube, Fractal modeling of fluctuations in the steady part of sustained vowels for high quality speech synthesis, IEICE Trans. Fundamentals, vol.E81-A, no.9, pp. 1803-1810, 1998.