Content Based Very Low Bit Rate Video Coding Using Wavelet Transform

Jayashree Karlekar

Hughes Software Systems Ltd.,
Electronic City, Plot 31, Sector 14,
Gurgaon - 122 015, INDIA.
jkarlekar@hss.hns.com

U. B. Desai

Department of Electrical Engineering,
Indian Institute of Technology,
Powai, Mumbai - 400 076, INDIA.
ubdesai@ee.iitb.ernet.in

Abstract

In this paper we present a new hybrid compression scheme for videoconferencing and videotelephony applications at very low bit rates (i.e., 32 kbits/s). In these applications, human face is the most important region within a frame and should be coded with high fidelity. To preserve perceptually important information at low bit rates, such as face regions, skin-tone is used to detect and adaptively quantize these regions. Novel features of this coder are the use of overlapping block motion compensation in combination with discrete wavelet transform, followed by zerotree entropy coding with new scanning procedure of wavelet blocks such that the rest of the H.263 framework can be used. At the same total bit-rate, coarser quantization of the background enables the face region to be quantized finely and coded with higher quality. The simulation results demonstrates comparable objective and superior subjective performance when compared with H.263 video coding standard, while providing the advanced feature like scalability functionalities.

1 Introduction

Video compression at low bit rates has attracted considerable attention recently in the image processing community. This is due to the expanding list of very low bit-rate applications, such as videoconferencing, multimedia, and wireless communications. The bulk of the very low bit-rate coding research has centered around the so called second-generation techniques [2], because existing standards such as H.263 perform rather poorly at lower rates. Second-generation techniques are certainly of interest, they are still characterized by computationally intensive algorithms which are required to solve challenging computer vision problems, and are therefore unlikely candidates for real-time communications over low-capacity communication channels.

The work described in this paper can be seen as a compromise between block-based methods and model-based techniques. The approach taken here is a hybrid approach which blends a waveform-based coding approach with content based techniques, as this facilitates adaptive quantization of the content of input video sequences. The proposed method is based on the discrete wavelet transform (DWT) with number of additional object-oriented features which makes it well suited for very low bit-rate (i.e., < 32 kbits/s) codig of "head-and-shoulder" video sequences, typical of real-time multimedia, videoconferencing, and videotelephony applications. In "head-and-shoulder" sequences, the human face and the associated subtle facial expressions need to be transmitted and reproduced as faithfully as possible at the receiver. The perceptual quality of such "head-and-shoulder" sequences can be improved by masking the face region for discriminative quantization in the process of compression. Higher compression ratio in the background, where the distortion is perceptually less significant allows the human face to be coded with higher fidelity at the same overall bit rate. Replacing DCT with DWT to reduce spatial redundancy makes the underlying video compression scheme highly scalable because of inherent multiresolution structure present in DWT. Also, because of the global nature of the wavelet transform, the region based adaptive quantization does not produce visible artificial boundaries between the regions of different quantization parameters. While the image content within the masked region is reproduced with high fidelity, the transition around the region boundary appears gradual and natural.
The paper is organized as follows: Section 2 describes the face detection method used in the proposed video coder using skin-tone color information, while Section 3 gives the brief description of the video coding algorithm based on wavelet transform. Section 4 gives simulation results. The conclusion is given in Section 5.

2 Face Detection using Skin-tone Color

Face region identification can be looked upon as a preprocessing step for complex applications like, face recognition, identity and credit card verification, face tracking, gaze tracking, indexing huge databases, and for model based coding of images and videos. The algorithms for locating faces based on template matching criterion are described in [9]-[11].

In video transmission and storage, colors are usually separated into luminance and chrominance components to exploit the fact that human eyes are less sensitive to chrominance variations. Human skin tones form a special category of colors, distinctive from the colors of most other natural objects. Although skin colors differ from person to person, they are distributed over a very small area in the chrominance plane [7].

To generate skin-tone colors on chrominance plane, training data of face patches of different images are cut. Fig. 1(a) shows the distribution of skin-tone colors in the (Cb, Cr) chrominance plane. A neural network (NN) is trained for skin-tone and non skin-tone classification. Given an input of (Cb, Cr) pair, the neural network classifies it as skin-tone or non skin-tone. The above trained classifier is used in the proposed video coder to classify each macroblock as a candidate face macroblock or a non-face one. To classify this, averages corresponding to chrominance blocks in a macroblock are fed as an input to the NN. This classification information is then passed on to an adaptive quantizer, so that finer quantization step can be used for face region while coarser quantization step is used for the rest of the image. Examples of face region identification using the proposed NN based approach are shown in Fig. 1 for Miss.America and Suzie video sequences, where white rectangle denotes the detected face regions.

3 Proposed Video Encoder

3.1 Quantization and Zerotree Coding

Zerotree entropy coding (ZTE) [5] [6] is a new tech-
nique for coding wavelet transform coefficients. The technique is based on, but differs significantly from, the EZW [4] and SPIHT [3] algorithms. Like EZW and SPIHT, ZTE algorithm exploits the self-similarity inherent in the wavelet transform of images. ZTE coding organizes wavelet coefficients into wavelet blocks using zerotree. Each wavelet block comprises of coefficients at all scales and orientation that correspond to the frame at the spatial location of that block. The concept of the wavelet block provides an association between wavelet coefficients and what they represent spatially in the frame. Wavelet blocks are scanned from low-to-high frequency into a vector, as proposed in [6], in order to be processed by the remaining parts of the H.263 coder.

3.2 Algorithm of Video Encoder

Each frame of the video sequence is encoded either as intra ("I") or inter ("P", forward prediction) frame, where intra frame is used for the first frame and inter frame is used for the remaining frames of the video sequence. Before that, each frame of the original video sequence, is passed to the NN based face region identification block, which gives information about the location of face region in a frame based on skin-tone color. The output of this block is passed to the control unit, which adaptively selects the quantization parameter. The intra frame is coded using the ZTE algorithm and variable length code tables of H.263.

For inter frames, a block based motion estimation scheme is used to detect local motion on blocks of size $16 \times 16$ with half-pel accuracy. Each block is predicted using overlapping block motion compensation scheme. After all the blocks have been predicted, the residuals are put together to form a complete residual frame for subsequent processing by the DWT. For those blocks where motion estimation fails, the intra mode is selected. To turn this block into a residual similar to the other predicted blocks, the mean of the block is subtracted from it and sent as an overhead [5].

4 Experimental Results

Objective and subjective comparison tests have been carried out between the proposed coding scheme and the emerging ITU-T standard H.263 on various test sequences. Each frame of the video sequence is decomposed into three levels of wavelet transform using orthogonal filter of Daubechies having regularity 4. Frames are coded as I or P using block motion estimation of size $16 \times 16$, overlapping block motion compensation, the discrete wavelet transform, and the ZTE algorithm with region based quantization and modified scan of wavelet blocks [6] so that it can be further encoded using run-length and variable length coding tables of H.263.

Several experiments were carried out using test sequences of "Miss_America", and "Suzie" (QCIF with resolution $176 \times 144$ and having an original frame rate of 30 frames per second (fps)). H.263 results were produced with the publicly available TMN encoder software from Telenor (http://www.nta.no/brukere/DVC). For a target bit rate of 10 kbits/s (Kbps), the encoded frame rate is 7.5 fps for both methods. This target frame rate is achieved by discarding every three out of four frames of the original sequence during the encoding process. The precise bit rate control was achieved by changing the quantization parameter in both methods.

Fig. 2 shows the peak signal to noise ratio (PSNR) versus frame number at 10 Kbps for Miss_America and Suzie video sequences. Note that, the objective measure PSNR is a good indicator of the subjective image quality only at high and medium bit rates. At very low bit rates, this is not always the case. As can be seen from Fig. 2, H.263 and the proposed scheme gives comparable PSNR values. Subjective comparisons between H.263 and the proposed method can be made by examining Fig. 3 and 4 (for luminance component only). Fig. 3 depicts frame 40 of Miss_America and Fig. 4 frame 144 of Suzie at 10 kbits/s. The output of H.263 (Fig. 3(b)) clearly shows the blockiness near eyebrows of Miss_America. Also, lips of Miss_America looks broken and incomplete because of block transform. For Suzie, the blockiness is observed near lower lip and tip of nose. Whereas, the result of the proposed method (Fig. 4(c)) looks more uniform and smooth. The blockiness is absent.

These examples essentially demonstrate that the bit saving arising from the coarse quantization of background have been utilized toward improving the quality of the region where the face is located.

5 Conclusion

For the proposed video coding algorithm, simulation results demonstrate comparable objective and superior subjective performance when compared with H.263. The algorithm provides a better way to address the scalability functionalities because of inherent multi-resolution structure present in the wavelet transform. The algorithm is very efficient for very low bit-rate coding applications. Due to DWT, there are
no visible blocking artifacts present in spite of using region-based adaptive quantization. The region-based enhancements proposed here provide high perceptually quality videoconferencing of head-and-shoulder video material where coding quality of facial areas is desirable.

References


Figure 3: Zoomed images of Miss.America (frame 40) video sequences encoded at 10 kbits/second, 7.5 frames/second: (a) Original; (b) H.263; and (c) Proposed method.

Figure 4: Zoomed images of Suzie (frame 144) video sequences encoded at 10 kbits/second, 7.5 frames/second: (a) Original; (b) H.263; and (c) Proposed method.