

Coding Face at Very Low Bit Rate via Visual Face Tracking

Jilin Tu¹, Zhen Wen¹, Hai Tao², Thomas Huang¹

Beckman Institute¹
University of Illinois at Urbana Champaign
Urbana, IL61801, USA
{jilintu, zhenwen, huang@ifp.uiuc.edu}

Department of Computer Engineering²
Baskin School of Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
{tao@soe.ucsc.edu}

ABSTRACT

Facial expressions convey very important information which is sometimes indispensable for people to fully understand the speech during communication. When the volume of video data is overwhelming and channel resources are limited, very low bit rate video coding becomes a solution for teleconferencing. In this paper, we propose an efficient and robust very low bit rate face coding method via visual face tracking. After the face is located in the video frame, the generic facial geometric model is adapted to the face, and facial texture for the model is extracted from the first frame of the video. The facial motion is then tracked and synthesized. The facial motion parameters, synthesized residual error and video background are transmitted at very low bit rate. Experiments show that our method can achieve better PSNR around facial area than H.26L at about the same low bit rate and have better subjective visual effects.

1. INTRODUCTION

Human face is a very unique and crucial communication tool people rely on heavily in every day life. The information contained in the facial expression incorporating with the speech is sometimes indispensable for people to fully understand each other during communication. Because of this reason, there is a substantial growing of the demand for videophone, teleconferencing applications in which the key problem is how to transmit video sequences at low bit rate given the foreground is known as human face and shoulder, and possibly simple background.

To accommodate the needs of transmission of large volume of video data over limited channel, several video coding standards, such as MPEG-1, MPEG-2, H.261, H.263 and H.26L[1][2][3][4] have been proposed. These standards are characterized by coding schemes that consist of block-based motion-compensated prediction, and quantization of prediction error with discrete cosine transform. Because these approaches only utilize the

spatial-temporal redundancy statistics of the video signal without a prior knowledge of the semantic content of the video, they are well applicable for general purpose video data compression where the scene in video frame is arbitrary. In the mean time, due to the difficulty to extract redundancy from video, a high coding rate usually also accompanies a high coding latency for certain video quality[4]. This hinders these approaches from applications where real-time video transmission is needed.

For the applications where human face is known as the major foreground object, model-based coding has been proposed to improve coding efficiency[5][6][7]. In these approaches, the human face geometry is characterized with a 3D mesh model. The facial motion is parameterized as rotation and translation for rigid motion, and action unit or facial muscle weights for non-rigid facial expression. These parameters together with the video background can be transmitted over channel at very low bit rate, and the video can be reconstructed via synthesis of the facial area based on the transmitted parameters. While this idea seems simple and intuitive, currently there is still not completely model-based coder presented because it is difficult to extract these facial geometry and motion parameters from video automatically and robustly.

People have proposed many approaches to extract face parameters from video. In [8], a face model *Candide* is fitted to the face by localizing the facial features in the video frame by frame, such as eye, mouth, nose, chin, etc. The facial features are estimated by image segmentation and deformable template matching, and the 3D face model parameters (size, orientation, and local facial feature geometry) can be adapted based on perspective projection and human face symmetry assumption. In [9], the head pose and facial expression can be estimated iteratively based on successive scaled orthographic approximation after key facial features are identified. In stead of only employing spatial information and adapting face model to face in video frame by frame, another popular approach is to employ the temporal motion flow information in video and track the face dynamically after face model is adapted to the face in the first frame. In these approaches, the rigid and non-rigid motion are

usually estimated simultaneously [10][11][12][13][14]. The facial deformation and head motion is characterized as a rigid face plus some linear combination of non-rigid facial deformation bases subjected to spatial translation and rotation. Depending on how rotation is parameterized, the parameters are retrieved in different ways. One approach [11][14] parameterizes rotation as angular changes. A linearized model between optical flow of the mesh node on the face and motion parameter displacements can be established, and the motion parameters can be calculated using a least square estimator given the optical flow as input. The second approach [12][13] parameterizes rotation as three orthonormal vectors, which forms a rotation matrix, using factorization, the rotation matrix and the vector of facial deformation coefficients can be obtained given rank deficiency condition of the optical flow and orthonormality constraints of the rotation matrix.

In this paper, we propose an efficient and robust very low bit rate face coding method via visual face tracking that can achieve real-time data compression. After face is located in the video frame, the generic facial geometric model is adapted to the face, and the facial texture for the model is extracted from the first frame of the video. The facial motion is then tracked and synthesized thereafter. The facial motion parameters, synthesized residual error and video background are transmitted at very low bit rates. Experiments show that our method can achieve better PSNR around facial area than H.26L at about the same low bit rate and have better subjective visual effects. In Section 2, we will describe our system in details, and the experimental results will be shown and analyzed in Section 3.

2. CODING FACE AT VERY LOW BIT RATES VIA FACE TRACKING

2.1 SYSTEM OVERVIEW

The architecture of the system we developed is shown in Figure 1. The video is first sent to a face tracker that extracts face motion parameters, and the facial motion is synthesized via a face synthesizer based on the motion parameters. The synthesized face and the original video frame are then sent to a Foreground/Background splitter, which split the foreground (facial area) and background into two regions. The foreground is replaced with residual errors between the synthesized face and the original face in video. The video background and synthesized foreground residuals are then sent to H.26L coder. The background will be coded as ordinary video frames by H.26L codec, and the foreground residuals are coded by Intra_16X16 mode. At receiver, the decoder synthesizes the facial motion according to the received face motion

parameters, reconstructs the foreground and background regions, and recovers the foreground facial area by summing up the synthesized face and foreground synthesized residuals. As most of the facial motion details are carried in the facial motion parameters, the foreground residual tends to have small amplitude, we can choose to code the video with very lot bit rates without losing much information of the foreground.

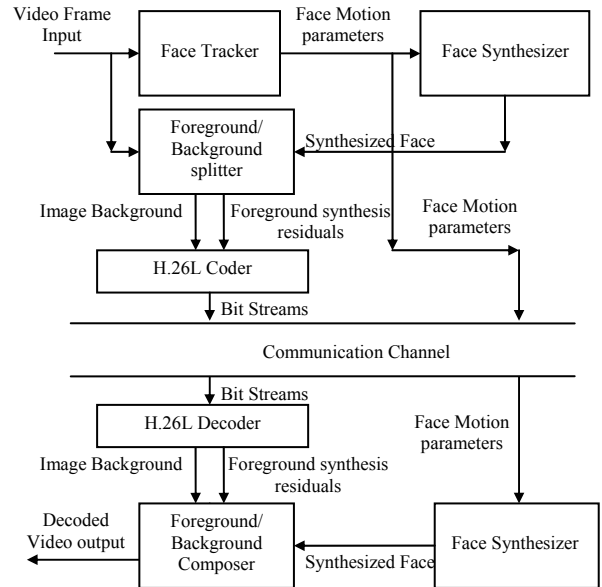


Figure 1. The System Architecture Overview

2.2 FACE TRACKER AND SYNTHESIZER

We have developed a robust 3D facial motion tracking system based on a piecewise Bezier Volume deformation model[11][15]. This model characterizes non-rigid facial motion as linear combination of some action unit bases which are composed of key facial expressions and visemes, namely, $V=Lp$ where V is the non-rigid facial motion, and L is a matrix with each column as a action unit basis, and p is the coefficient vector for the action units. And the key facial expressions and visemes are characterized with mapping from 3D facial motion control points to the 3D deformation of the facial surface piece-wisely using Bezier volume techniques, namely $L=BD$, where B describes the mapping function composed of Bernstein polynomials, and each column of D stands for the 3D displacement of control points of one specific action unit. The composition of the piece-wise Bezier volume is shown in Figure 2.

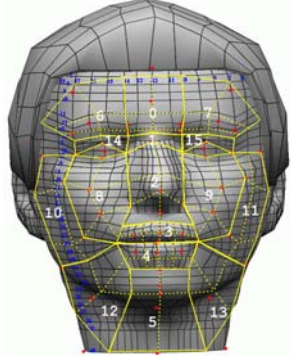


Figure 2. The composition of Piece-wise Bezier volume

This model allows the user to generate action units from facial fiducial points, which is independent of the topology of the actual mesh model. This property is crucial when deforming face models of different geometry. Based on this model, the facial motion can be characterized as $\mathbf{R}(\mathbf{V}_0 + \mathbf{L}p) + \mathbf{T}$, where \mathbf{V}_0 is the neutral face, $\mathbf{L}p$ defines the non-rigid deformation, and \mathbf{R} , \mathbf{T} stand for 3D rotation and translation respectively. After a linearized model between the derivatives of head pose parameters \mathbf{R} , \mathbf{T} , motion unit weights P and low-level optical flow of facial area is derived, the motion parameters can be retrieved by solving a set of linear equations. The optical flow calculation is done in a multi-resolution manner so that the robustness and accuracy are largely improved [11].

The face synthesizer takes facial texture in the first frame of the video, and maps it to the generic face model. The texture mapped face model is then deformed and rendered thereafter based on the motion parameters \mathbf{R} , \mathbf{T} , p .

2.3 EMBEDDING SYNTHESIZED FACE IN H.26L CODEC

After the synthesized face is obtained, the residual error can be calculated by subtracting it from the original frame, and the video frame is divided into foreground residual and background region. For the background, we can still employ the advantage of H.26L and do the coding based on spatial-temporal redundancy analysis. For the foreground residual, it contains mostly high frequency signal components, therefore we specify Intra_16X16 mode as the coding mode for macroblocks in the foreground residual region. The advantage of this is that the facial motion details mostly captured by the motion parameters will not be lost when the video is coded at the lowest bit rate.

3. EXPERIMENT RESULTS

The face tracker runs at 25 fps in rigid tracking mode and 14 fps in non-rigid tracking mode. The H.26L codec software was obtained from International Multimedia Telecommunications Consortium (IMTC) FTP website [16]. After the tracking system is integrated with the H.26L codec software, some reconstructed video frames are shown in Figure 3. We took a sequence of about 147 frames of size 352 by 240 as input. The video is coded at about the same bit rate of 18-19kb/s using our approach and H.26L codec respectively. The first row shows two key frames of the synthesized facial motion after face tracking. The second row shows the reconstructed frames using synthesized facial motion, and the third row shows the reconstructed frames using H.26L codec. It is easy to see that the facial motion details are preserved better in the reconstructed frames using synthesized facial motion than the reconstructed result of H.26L codec, and thus have better subjective visual effect. For the reconstructed video with facial motion synthesis, the Peak Signal to Noise Ratio (PSNR) around facial area is 29.28, while for the reconstructed video using H.26L codec, the PSNR is about 27.35. Besides, we specified Intra coding and forward predictive (P) coding modes to code the background, while H.26L utilizes bidirectional coding approach as default option and explores a bigger space to search for an appropriate mode for the macroblocks. Therefore it took on average only 1.4 seconds to code a frame in our approach, while 5 seconds to code a frame for H.26L codec.

4. CONCLUSIONS AND FUTURE WORK

This paper presents an efficient and robust very low bit rate face coding method via visual face tracking. With facial motion parameters extracted from the video via face tracking technique and transmitted over the channel, the facial motion synthesis residual errors and video background can be coded at very low bit rates. The video can be reconstructed according to reconstructed background, residual errors, and synthesized facial motion based on received facial motion parameters. Experiments show that our method can achieve better PSNR around facial area than H.26L at about the same low bit rate and have better subjective visual effects. As the performance of face tracking is the bottle-neck for model-based coding approach, our next step is to further improve on face tracking. In the mean time, we will also look into better approach to code the synthesis residual errors.

5. REFERENCES

- [1] Video coding for low bit rate communication, ITU-T Recommendation H.263 Version 2(H.263+), Jan. 1998.
- [2] Generic Coding of Audio-Visual Objects: (MPEG-4 video), Final Draft International Standard, Document N2502, 1999.

- [3] Joint Final Committee Draft (JFCD) of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC), July, 2002.
- [4] John Watkinson, "The MPEG Handbook: MPEG1, MPEG2, MPEG4", Focal Press, 2001.
- [5] K. Aizawa and T. Huang, "Model based image coding: Advanced video coding techniques for very low bit-rate applications", Proc. IEEE, vol. 83, pp. 259-271, Feb. 1995.
- [6] H. Musmann, "A layered coding system for very low bit rate video coding", Signal Processing: Image Communication, Vol. 7, pp. 267-278, Nov. 1995.
- [7] W. J. Welsh, S. Searby, and J. B. Waite, "Model-based image coding", British Telecom Technology Journal, Vol. 8, no. 3, pp.94-106, July 1990.
- [8] M. Kampmann, "Automatic 3-D Face Model Adaptation for Model-Based Coding of Videophone Sequences", IEEE Trans. On Circuits and Systems for Video Technology, Vol. 12, No. 3, March 2002.
- [9] Chin-Chun Chang, et al, "Determination of Head Pose and Facial Expression from a Single Perspective View by Successive Scaled Orthographic Approximations", Intl. Journal of Computer Vision 46(3), 179-199, 2002.
- [10] Li, H. and Forchheimer, R. "Two-view facial movement estimation". IEEE Trans, on Circuits and Systems for Video Technology 4(3):276-287.
- [11] Tao, H. and Huang, T.S. "Bezier volume deformation model for facial animation and video tracking", In Proc. of Intl. Workshop, CAPTECH'1998: Modelling and Motion Capture Techniques for Virtual Environments, Berlin, Germany, Springer Verlag:Berlin, pp. 242-253.
- [12] M. Brand and R. Bhotika, "Flexible flow for 3D nonrigid tracking and shape recovery", In Proc. CVPR, 2001
- [13] Matthew Brand, "Morphable 3D models from video", Technical report: TR-2001-27, MERL.
- [14] Douglas DeCarlo and Dimitris Metaxas. "Optical Flow Constraints on Deformable Models with Applications to Face Tracking",. In IJCV, July 2000, 38(2), pp. 99-127
- [15] Thomas S. Huang, and Hai Tao, "Visual Face Tracking and its Application to 3D Model-based Video Coding", Picture Coding Symposium 2001, pg. 57-60.
- [16] H.26L codec reference software: <http://ftp.imtc-files.org/>



(a) The synthesized face motion



(b) The reconstructed video frame with synthesized face motion



(c) The reconstructed video frame using H.26L codec

Figure 3. Comparison of the reconstructed video frames