Express Letters

Face Location in Wavelet-Based Video Compression for High Perceptual Quality Videoconferencing

Jiebo Luo, Chang Wen Chen, and Kevin J. Parker

Abstract— **We present a human face location technique based on contour extraction within the framework of a wavelet-based video compression scheme for videoconferencing applications. In addition to an adaptive quantization in which spatial constraints are enforced to preserve perceptually important information at low bit rates, semantic information of the human face is incorporated to design a hybrid compression scheme for videoconferencing since the human face is often the most important portion within a frame and should be coded with high fidelity. The human face is detected based on contour extraction and feature point analysis. An approximated face mask is then used in the quantization of the decomposed subbands. At the same total bit rate, coarser quantization of the background enables the face region to be quantized finer and coded with higher quality. Simulation results have shown that the perceptual image quality can be greatly improved using the proposed scheme.**

I. Motivations

In videoconferencing, a video sequence often contains head-and-shoulder images. The human face and the associated subtle facial expressions need to be transmitted and reproduced as faithfully as possible at the receiving end. The perceptual quality of the images in videoconferencing can be improved by masking the face region for a 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.

There are two advantages in developing a discriminative quantization in a wavelet-based coding scheme. First, the good spatial and frequency localization properties resulting from wavelet decomposition allow the compression to be adjusted to each individual subband and any desired local region within a subband. Second, since interpolation and filtering are involved in the wavelet synthesis stage at the receiving end, a face mask will not produce visible artificial boundaries between the regions of different quantization accuracies. While the image content within the masked region is reproduced with relatively higher fidelity, the transition around the region boundary appears gradual and natural. On the contrary, a block discrete cosine transormation (DCT)-based coding scheme often generates visually annoying blocking artifacts around the boundaries of a masked region.

II. Contour-Based Face Location Method

There have been several face location approaches [2]–[4], mainly for various computer vision applications. We propose in this paper a face location approach based on edge detection for the purpose of facilitating a discriminative quantization. Through developing an edge-based approach, we try to make the algorithm robust to lighting conditions, skin color, clothing, etc. The proposed approach consists of the following steps.

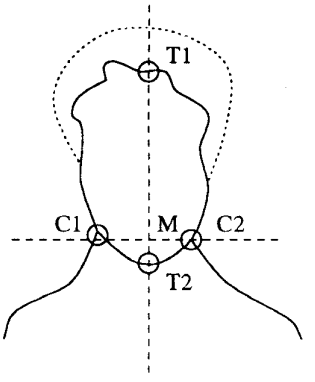

Fig. 1. Face contour feature points.

1) Edge Detection in the Baseband Image: Removal of the excessive details in the original image would make the extraction of face area easier while smoothing of the image would suppress some noise. In this aspect, the wavelet decomposition already provides a very good low resolution representation of the original image—the baseband. Note that this is different from simple subsampling which would generally cause loss of important detail information. To detect edges in the baseband image, a robust edge detection algorithm proposed by Canny [5] is applied.

2) Knowledge-Based Contour Following: Based on the knowledge of the shape of the human face, head, and upper body, we search for Ω-shaped contour segment(s) in the obtained edge image. In some cases where long hair generates another Ω-shaped contour outside the face contour, only the inside contour is retained.

3) Feature Point Extraction: Further identification of the face part is done by finding the feature points along the extracted Ω-shaped contour. A local curvature measure called adaptive K-cosine [6], given by

$$k\cos_i = \frac{|(\vec{X}_i - \vec{X}_{i+k}) \odot (\vec{X}_i - \vec{X}_{i-k})|}{|\vec{X}_i - \vec{X}_{i+k}||\vec{X}_i - \vec{X}_{i-k}|} \tag{1}$$

is calculated where $\vec{X}_i$ denotes a point in a chain-coded curve and k is the window size. $\odot$ represents inner product. The K-cosine is adaptive in the sense that the size of the window used in computing the K-cosine, instead of being fixed, is optimally selected according to the local characteristics of the curve. A fixed window size would make the algorithm very sensitive to scale. The corner points are extracted as the local maxima of the K-cosine, as shown in Fig. 2(b).

4) Knowledge-Based Facial Contour Feature Point Identification: Based on the knowledge of the human face, several feature points can be identified on the face contour. Top point T1 is the top feature point in the Ω-shaped contour. Two chin points, C1 and C2, as shown in Fig. 1, are critical in separating the face from the neck or collar. The reason is that the chin contour is often blurred and hard to extract because usually there is not much discontinuity between the chin and

Manuscript received April 5, 1995; revised October 31, 1995. This paper was recommended by Associate Editor Y.-Q. Zhang. This work is supported by NSF Grant EEC-92-09615 and a New York State Science and Technology Foundation Grant to the Center for Electronic Imaging Systems at the University of Rochester.

The authors are with the Department of Electrical Engineering and Center for Electronic Imaging Systems, University of Rochester, Rochester, NY 14620 USA.

Publisher Item Identifier S 1051-8215(96)05193-2.

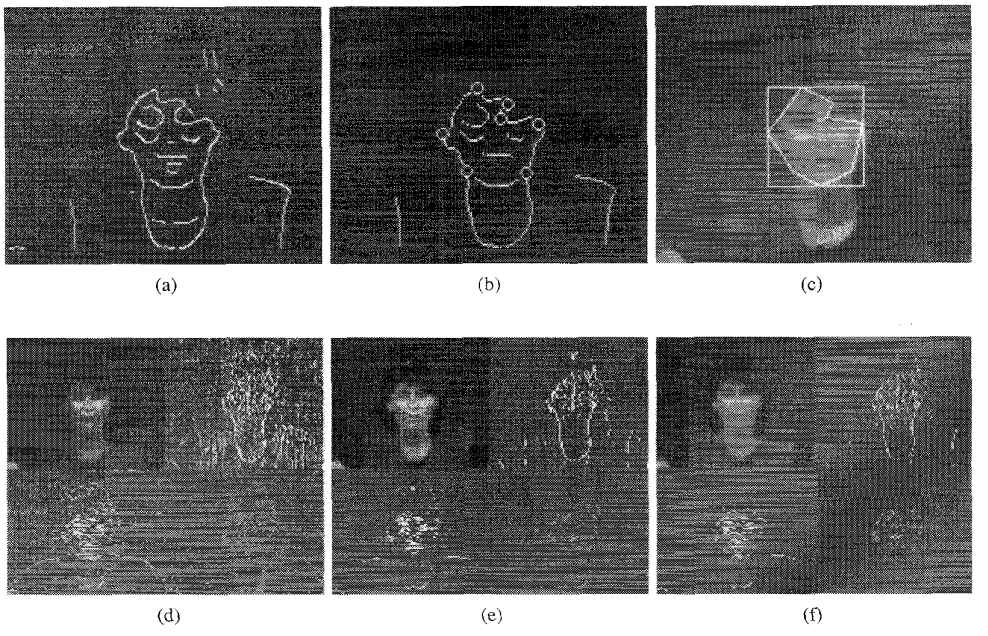

Fig. 2. Construction of a face mask for "Miss America." (a) Edge image, (b) feature points, and (c) face mask. Quantization of the subbands (d) decomposed subbands, (e) quantized subbands, and (f) subbands quantized with face mask.

the neck. We have to rely on the two chin points which can often be robustly located. These two points are almost symmetric about the long (vertical) axis of the Ω-shaped contour. In particular, they have similar negative K-cosine values which correspond to obtuse angles at the transition from face to neck. For the same reason, the tip point of the chin, T2, is also hard to locate. However, its approximate location can be decided according to the normal proportion of the human face. Results of the face location are given in Fig. 2(c).

5) Face Representation: We use the smallest rectangular box which encompasses all the facial feature points to represent the face mask since it requires the least side information for image transmission. A precise contour is very complex and costly for the compression. Furthermore, the rectangular box provides some extra transition region between the face and the background, which is desired in producing natural-looking images in visual communication.

III. Compression with Face Mask

We have recently proposed a novel adaptive quantization approach [1] for the coding of the high frequency subbands based on the concept of adaptive clustering with spatial constraints. In the adaptive quantization, a Gibbs random field (GRF) is employed to enforce spatial constraints so that visually important structures can be identified. This quantization, on one hand, improves coding efficiency through removing those "impulses" whose contributions to the reconstruction and the perception are negligible but may otherwise need considerable bits to code. On the other hand, it preserves visually important components. The entropy of the subbands is significantly reduced after such a quantization. The quantized subbands (with contrast-boosting histogram equalization for viewing purpose) are shown in Fig. 2(e).

In addition to the adaptive quantization, which can be considered as a GRF-based prioritization, semantics-based prioritization can also be achieved by allocating more bits to the region of interest. With the obtained face mask, we develop an optimal prioritization strategy. To achieve high fidelity coding, the quantization strategies, i.e., the assignment of quantization levels and the selection of quantizers for each subband, should all be determined according to the perceptual importance and the bit rate constraints. Within each subband, we perform a discriminative quantization. Inside the masked region, we use a fine uniform quantizer with considerably large number of quantization levels, or small quantization step sizes. The reason to use a uniform quantizer is that the information within the masked region is considered as equally important no matter of what intensity value at what gray level. Meanwhile, only weak spatial constraint is imposed. Outside the masked region, a coarser nonuniform Lloyd-Max quantizer which minimizes mean square error (MSE) is used in combination with a relatively strong GRF to accomplish the adaptive quantization. Table I gives an example of such a discriminative quantization which maintains the same bit rate and the results are shown in Fig. 2(f). In comparison to Fig. 2(e), the quantization is designed to be in favor of the masked region in all the subbands.

IV. Simulation Results

We have obtained some simulation results based on the proposed discriminative quantization strategy. The proposed face location method is sufficient for typical video sequences with slightly cluttered background, such as "Miss America," "Claire," and "Trevor." For a comparison with JPEG, only spatial subband decomposition is applied to an individual frame of the videoconferencing sequence "Miss America." The first observation is that JPEG performs extremely poorly at such a low bit rate of 0.2 bpp since the image has been distorted to the extent that the face can no longer be recognized [Fig. 3(d)]. If the spatiotemporal subband decomposition described in [7] is applied, the bit rate would be about 0.1 bpp since the highpass temporal (HPT) band does not consume much of the bit rate. The second observation is that head-and-shoulder images can be coded with much higher subjective quality at the same overall bit rate by introducing a face mask for a discriminative compression [Fig. 3(b) and (e)]. In Fig. 3(b), the face is of poor quality since the entire image is coded at a low bit rate. In Fig. 3(e), the face is reproduced with much higher fidelity while the quality of the background degrades somewhat further with the discriminative quantization. Furthermore,

TABLE I
EXAMPLE OF THE DISCRIMINATIVE QUANTIZATION FOR THE SUBBANDS

Adaptive quantization without face mask		Adaptive quantization with face mask	
LL band quantization step size = 16	HL band quantization level number = 5 β = 0.1	LL band quantization nonmasked region: step size1 = 20 β = 0.0 masked region: step size2 = 4 β = 0.0	HL band quantization nonmasked region: level number = 3 β = 0.2 masked region: step size = 4 β = 0.0
LH band quantization level number = 5 β = 0.1	HH band quantization level number = 3 β = 0.1	LH band quantization nonmasked region: level number = 3 β = 0.2 masked region: step size = 4 β = 0.0	HH band quantization nonmasked region: level number = 1 β = 0.2 masked region: step size = 4 β = 0.0

Fig. 3. *Comparison of the reconstructed images.* (a) Original, (b) quantization with no mask, and (c) enhanced image from Fig. 3(b). (d) JPEG, (e) quantization with mask, and (f) enhanced image from Fig. 3(e).

TABLE II
IMAGE QUALITY EVALUATION

Compression	PSNR (db)	Subjective quality	PSNR (db) after enhancement	Subjective quality
No face mask	33.78	very poor	35.07	fair
With face mask	31.62	good	33.09	very good
JPEG	32.98	extremely poor		

to remove the quantization artifacts in the reconstructed image, the HMRF-based image enhancement technique is applied to the reconstructed images [7]. With the face mask, the enhancement can also be adjusted accordingly. Stronger spatial constraints can be applied to the background region, while the spatial constraints for the masked region should be very weak in order to preserve the image details. The artifacts caused by coarse quantization, such as the blotchiness in the neck region, are greatly reduced with minimum blurring of the important image details, especially in the face region [Fig. 3(c) and (f)]. In terms of PSNR, the discriminative quantization scheme is inferior at the same bit rate because the majority of the pixels other than those within the face mask are quantized more coarsely so that the overall MSE increases. However, the PSNR of the face area is much higher in the case of the discriminative quantization. Meanwhile, the enhancement produces significant improvement in PSNR and visual quality in both cases.

The image quality evaluation using both subjective and objective criteria is given in Table II.

The rate-distortion performance of the result is not among the best for current schemes (e.g., H.263 standard). In this simulation, we achieved a bit rate of 48 kb/s at a peak signal-to-noise ratio (PSNR) of 33 dB. However, the simulation has not taken full advantage of the potential of this wavelet-based discriminative coding scheme. First, we use only one level of wavelet decomposition. Consequently, efficient zerotree coding technique is not applicable [8]. In general, more decomposition levels enable better energy compaction and more efficient zerotree coding. Second, two-tap Haar filter is inadequate to exploit the temporal redundancies in the video signal. In spite of these facts, the simulation clearly demonstrates that the wavelet-based discriminative coding scheme can outperform nondiscriminative schemes by a significant margin, especially in terms of the perceptual quality of the reconstructed images.

V. Conclusion and Future Work

A new face location technique based on the extraction of the facial contour is proposed. It has been applied to a wavelet-based image and video compression scheme designed primarily for video-conferencing. An approximate face region mask is generated and used for a discriminative quantization of the decomposed subbands. Simulation results have shown that this approach is promising in videoconferencing applications since the perceptual image quality of the face region can be greatly improved. Motion information can certainly be incorporated for robust location and tracking of the face region in a more complex background using complete three-dimensional information. Semantic information, such as the detection of the mouth and the eyes, can also be incorporated to confirm the face detection [3]. We are convinced that computer vision techniques can be incorporated into compression schemes to improve the performance of the overall coding system.

References

[1] J. Luo, C. W. Chen, K. J. Parker, and T. S. Huang, "Adaptive quantization with spatial constraints in subband video compression using wavelets," in *Proc. Int. Conf. Image Processing,* Washington, DC, Oct. 1995, pp. I 594–597.

[2] K. Lam and H. Yan, "Fast algorithm for locating head boundaries," *J. Electron. Imaging*, vol. 3, no. 4, pp. 351–359, Oct. 1994.

[3] G. Yang and T. S. Huang, "Human face detection in a complex background," *Pattern Recognition,* vol. 27, pp. 53–63, Jan. 1994.

[4] S. Shimada, "Extraction of scenes containing a specific person from image sequences of a real-world scene," in *Proc. IEEE TENCON '92,* Melbourne, Australia, Nov. 1992, pp. 568–572.

[5] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. PAMI-8, pp. 679–698, Nov. 1986.

[6] M. Fischler and R. C. Bolles, "Perceptual organization and curve partitioning," in *Readings in Computer Vision,* M. A. Fischler and O. Firschein, Eds. San Mateo, CA: Morgan Kaufmann, 1987, pp. 210–215.

[7] J. Luo, C. W. Chen, K. J. Parker, and T. S. Huang, "A scene adaptive and signal adaptive quantization for subband based image and video compression using wavelets," to appear in *IEEE Trans. Circuits Syst. Video Technol.*

[8] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing,* vol. 41, pp. 3445–3462, Dec. 1993.

Multiviewpoint Video Coding with MPEG-2 Compatibility

Belle L. Tseng and Dimitris Anastassiou

***Abstract*—An efficient video coding scheme is presented as an extension of the MPEG-2 standard to accommodate the transmission of multiple viewpoint sequences on bandwidth-limited channels. With the goal of compression and speed, the proposed approach incorporates a variety of existing computer graphics tools and techniques. Construction of each viewpoint image is predicted using a combination of perspective projection of three–dimensional (3-D) models, texture mapping, and digital image warping. Immediate application of the coding specification is foreseeable in systems with hardware-based real-time rendering capabilities, thus providing fast and accurate constructions of multiple perspectives.**

I. Introduction

Recent interests in three-dimensional (3-D) technologies prompt the addition of depth impression onto the otherwise common two-dimensional (2-D) video signals. Two major processes to perceiving 3-D can be categorized. The first is due to the two slightly different perspectives of the world offered simultaneously to our left and right eyes. The human visual system then converts these two stereo images into one single fused 3-D perception. The other approach to sense depth, even with only one eye-viewpoint, is through motion parallax. Due to motion from our head movements, the relative object displacements of the resulting perspective view are sufficient cues in deriving the 3-D sensation. Accordingly, in our presentation, the depth appearance is contributed by both processes where construction of stereo and virtual viewpoint images is possible.

A multiviewpoint video, multiview for short, is a 3-D extension of the traditional movie sequence, in that there are multiple perspectives of the same scene at any one instance in time. Comparable to a movie made by a sequence of holograms, a multiview video offers a similar look-around capability. An ideal multiview system allows any user to watch a true 3-D stereoscopic sequence from any perspective the viewer chooses. Such a system has practical uses in interactive applications, medical technologies, educational and training demonstrations, remote sensing developments and is a step toward virtual reality.

With the development of digital video technology, a video data compression standard, namely the second Motion Picture Experts Group specification (MPEG-2), has been adopted by the International Standards Organization (ISO) and the International Telecommunications Union (ITU). MPEG-2 specifies the coding process for one video sequence; detailed descriptions can be found in [1]. Recently, MPEG-2 has also been shown to be applicable to two sequences of stereoscopic signals through the use of spatial and temporal scalability extensions [2]–[5]. However, extending the number of video viewpoints beyond two cannot be done practically by using the same methodology. For this motivation, a novel multiview codec is presented to complement the MPEG-2 standard.

II. Geometric Definitions and Notations

An ordinary 2-D video sequence offers only one perspective of the acquired scene. For a 3-D viewing experience, at least two viewpoints are required to obtain the depth impression from the left and right

Manuscript received August 15, 1995; revised February 8, 1996. This paper was recommended by Associate Editor Y.-Q. Zhang.

The authors are with Columbia University, New York, NY 10027 USA.

Publisher Item Identifier S 1051-8215(96)05192-0.