

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 videoconferencing. 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. 1 594-597.
- [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.

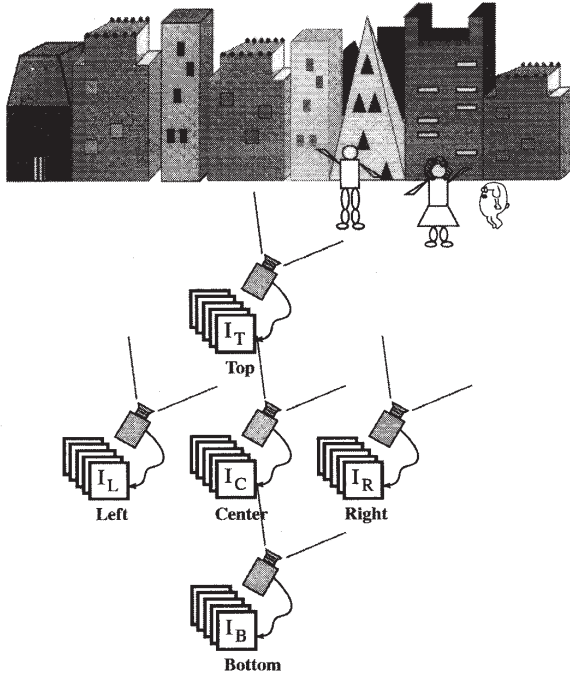


Fig. 1. Camera configuration of multiple viewpoint images.

perspectives of a stereoscopic signal. In a multiview system, multiple viewpoints are involved. Let N be the number of viewpoint sequences where the minimum number of viewpoints is equal to three for the presented specification, i.e., $N \geq 3$. The number of viewpoints must be extendable in the future, thus accommodation of additional viewpoints is an essential feature.

One central viewpoint image is designated to be the principal view from which the other multiview images are predicted. Let the central image be denoted by I_C from viewpoint V_C . The central viewpoint, usually the middle viewpoint, is chosen so that its image has the highest collection of overlapping objects with each of the other viewpoint images. In this manner, the central viewpoint image can be used to interpolate and predict most of the other views.

The other multiview images are designated by I_X from corresponding viewpoints V_X . An example of four additional views is illustrated in Fig. 1, where cameras positioned at viewpoints $V_X = \{V_L, V_R, V_T, V_B\}$ capture respective images $I_X = \{I_L, I_R, I_T, I_B\}$ corresponding to the left, right, top, and bottom. These camera-captured images available at the encoder are referred to as *real* views, whereas images not directly taken by a camera, but derived by prediction or interpolation methods, are called *virtual* views. Virtual pictures include those viewpoint images seen between two real cameras, thus having such virtual constructions allows the viewer to see a smoother video transition between two real views.

The image coordinate system corresponding to each viewpoint is independently defined as (X_i, Y_i) , where viewpoint index $i = \{C, L, R, T, B\}$ for the five multiview configurations of Fig. 1. Let the global rectangular coordinate system (X, Y, Z) be defined as corresponding to the image coordinate system (X_C, Y_C) of the central viewpoint, where Z defines the orthogonal axis from the central image plane.

Each multiview camera position is represented by the global coordinates (vx_i, vy_i, vz_i) , with the camera zooming parameter

described by F_i . In addition, the camera rotations are given by the vertical tilting angle α_i , the horizontal panning angle β_i , and the axial revolving angle γ_i . The total viewpoint vector for some viewpoint index i is denoted as $V_i = [vx_i, vy_i, vz_i, F_i, \alpha_i, \beta_i, \gamma_i]$. Viewpoint vectors are obtained relative to the central viewpoint, where $V_C = [0, 0, 0, 1, 0, 0, 0]$. Thus for a given acquisition camera configuration, viewpoint directions are relative 3-D translations and rotations with respect to the central view. If camera configurations are unknown, then locating several fixed points between multiviews allows determination of the relative orientation transformation [6], [7].

III. DEPTH AND DISPARITY ESTIMATION

Possessing the depth position of an object in one image allows for geometric prediction of the object location in all other viewpoint images. Consequently, our desire to determine the depth of every object in a scene permits construction of the scene from any viewpoint. The depth of an object can be geometrically calculated if two or more perspectives of the object are given, as in a collection of multiviews. First, the positions of the object in each of the available viewpoint images must be located; this problem is widely known as the *correspondence problem*. After locating the object positions from two views, the difference in image coordinates is termed *disparity*. Following, it can be shown mathematically that the depth is inversely proportional to the derived disparity [7].

Under the block-based motion characterization constraint of MPEG-2, one disparity vector corresponding to each block of an image can be incorporated for prediction of a second image. Block-based disparity compensation is the dominant approach for the coding of stereoscopic video sequences. These approaches are sufficient for coding and transmission of two stereo signals without entailing true depth details to the decoding system. For acceptable predictions of multiviewpoint images however, accurate depth information for every pixel is required for multiple spatially-continuous interpolations of intermediate viewpoint images. Thus a depth map is required covering every pixel whose quantized values can be transmitted as the gray-level intensity of a second "image." Since the majority of the depth image is quite flat and relatively object oriented, the depth map can be considerably compressed as illustrated in [8]. To obtain a dense depth map, many efficient techniques have been developed for stereo image pairs as well as for multiple views, including [7], [9], and [10].

IV. MPEG-2 COMPATIBLE MULTIVIEWPOINT VIDEO CODING

The delivery of two viewpoint video sequences is accomplished by the scalability extensions of the MPEG-2 video coding standard, where spatial predictions are obtained by disparity-compensation on a macroblock basis. As mentioned in Section III, however, for multiview video coding, a dense depth map is required for accurate predictions of multiple views. Seeking compatibility with the compression standard, an ad hoc group of MPEG-2 is established to investigate a new profile for multiviewpoint systems [11]. For conformity, the multiview profile is to supplement the main profile of MPEG-2 so that all standardized systems are capable of deciphering at least one video sequence from one viewpoint. Consequently, the central viewpoint is selected to be processed and transmitted in accordance with the main profile. The remainder of this section is devoted to describing the proposed encoding and decoding systems for the processing of the other viewpoints.

A. Multiviewpoint Encoding System

A block diagram of the proposed multiview encoding system is illustrated in Fig. 2 for five multiview sequences: $I_C^t, I_L^t, I_R^t, I_T^t, I_B^t$,

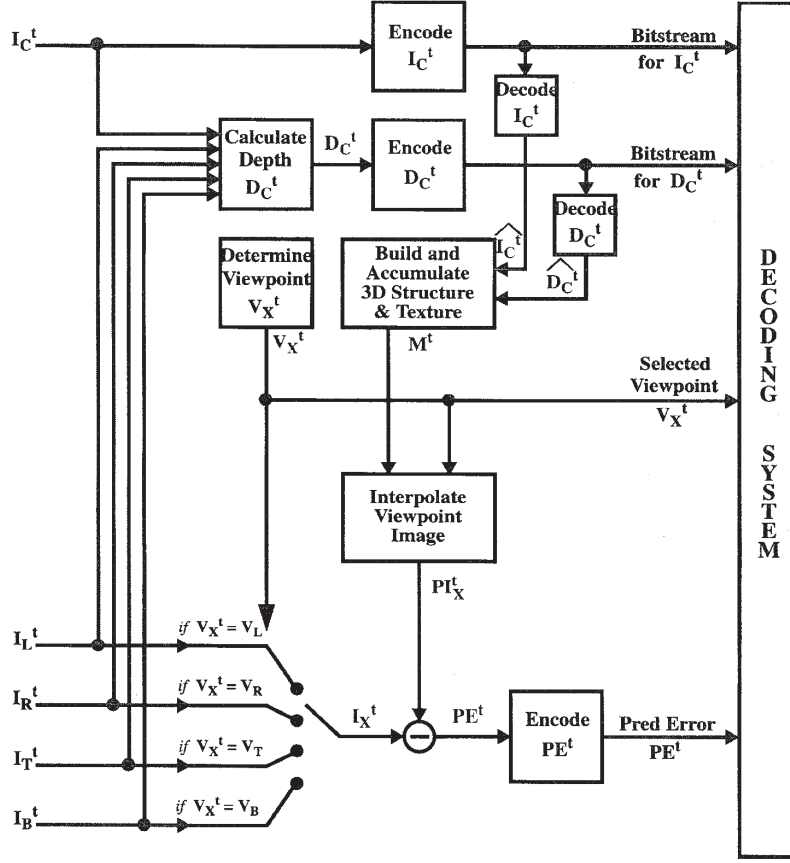


Fig. 2. Block diagram of multiview encoding system.

and I_B^t . First, determine all viewpoint vectors, $V_C^t, V_L^t, V_R^t, V_T^t$, and V_B^t from the acquisition camera configurations positioned in Fig. 1. Find the central viewpoint image, in our case I_C , achieving the best predictions for the other views. The central image sequence I_C^t is then processed in the main profile of MPEG-2. Next, the encoded bitstream for I_C^t is ready for transmission and in parallel is decoded to determine the receiver-reconstructed central images, denoted $\hat{I}_C^t$.

Starting with the central viewpoint image I_C^t at time t , a depth value $z_C = D_C^t(x_C, y_C)$ is calculated for every pixel $I_C^t(x_C, y_C)$, thus forming a depth map image, named D_C^t , corresponding to the central viewpoint image. Afterwards, encode the depth map D_C^t as a secondary highly compressible image, and transmit the encoded bitstream for D_C^t . Simultaneously, the encoded bitstream for D_C^t is decoded to determine the receiver-reconstructed depth map, assigned $\hat{D}_C^t$.

Obtain a 3-D mesh model representation [12], called M^t , for the central image $\hat{I}_C^t$ of time t by associating each coordinate pair (x_C, y_C) with its corresponding depth value $\hat{D}_C^t(x_C, y_C)$. Consequently, the 3-D surface texture is derived from the 2-D image intensity $I_C^t(x_C, y_C)$, and a graphical model of the scene is obtained. This geometric representation offers ease in interpolating different viewpoints. Similar to rendering approaches in computer graphics [13], [14], given a 3-D geometrical model and its associated

texture intensity, every viewpoint can be effectively and efficiently constructed by simple geometric transformations followed by 3-D texture mapping. Furthermore, interpolating virtual viewpoint images is facilitated and feasible.

Select a noncentral viewpoint, designated as V_X^t , chosen from the set of original real viewpoints in which the best construction of its image I_X^t is desired for time t . The selection of viewpoint V_X^t is based on a round-robin schedule where every viewpoint is successively selected in a cycle. In Fig. 3, a round-robin rotational scheme for four noncentral views is shown where the selection of each image is related to time t in the following manner: $I_L^t, I_R^{t+1}, I_T^{t+2}, I_B^{t+3}, I_L^{t+4}, I_R^{t+5}, I_T^{t+6}$, etc., distinguished by a double border. Next, transmit the selected viewpoint vector V_X^t for time t .

The next step is to interpolate the selected noncentral viewpoint image, referred to as predicted image PI_X^t , by rendering the 3-D mesh model M^t in the specified viewpoint V_X^t . This interpolation step requires rendering a wireframe image of the desired viewpoint by simple geometric transformations with perspective projections of the mesh model, followed by texture mapping the corresponding areas of the central image onto the appropriate wireframe substructures.

Subsequently, calculate the prediction errors PE^t required for the final reconstruction of the selected view by examining the difference between the original image I_X^t and the predicted image PI_X^t . Finally, encode and transmit the residual prediction errors PE^t . Before

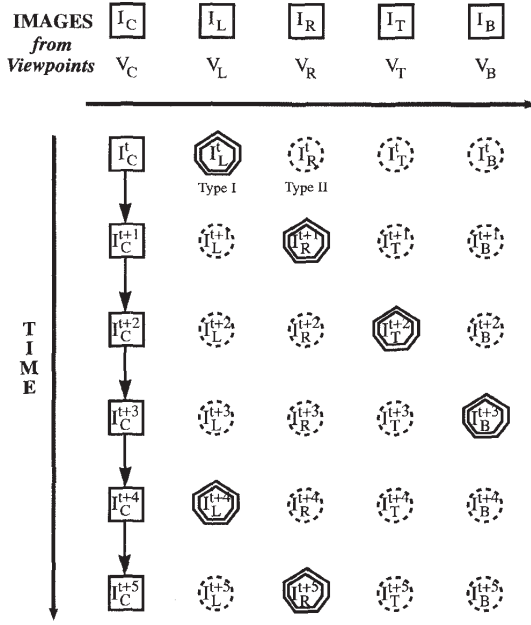


Fig. 3. A round-robin rotational schedule for encoder selection of one viewpoint.

moving on to the next frame, determine if the allocated bit rate allows transmission of an additional noncentral image. If bandwidth allows, the prediction errors for another view can be determined and sent. Alternatively, a new round-robin schedule can be adopted where two viewpoints are selected for every time, thus perfect reconstruction is always achievable with unlimited bandwidth. Subsequently, the entire process starts all over for the next time frame.

B. Multiviewpoint Decoding System

A block diagram of the proposed multiview decoding system is illustrated in Fig. 4 for construction of any viewpoint images. First, the received bitstream of the central view is decoded on the main profile of MPEG-2, whose decoded images, denoted $\widehat{I}_C^t$, are stored in memory. Second, the bitstream of the depth map is decoded and hereafter referred to as $\widehat{D}_C^t$. Next, obtain a 3-D mesh model, called M^t , for the central image $\widehat{I}_C^t$ as performed in the encoding system. Consequently, store the model M^t for time t in memory.

Upon receiving the encoder-selected viewpoint vector V_X^t for time t , store the vector in memory. Next, the prediction errors associated with the selected viewpoint V_X^t are decoded, denoted as $\widehat{PE}^t$. Now, determine the viewpoint requested from the user at the decoding system for time t , assigned V_U^t . Subsequently, the user-requested viewpoint image (whether real or virtual) is interpolated, referred to as predicted image PI_U^t , where the index U designates the appropriate viewpoint V_U^t . The prediction PI_U^t is constructed by rendering the 3-D mesh model M^t in the specified viewpoint V_U^t by the same viewpoint image interpolation procedure as in the encoder.

Determine if the encoder-selected noncentral viewpoint V_X^t is similar to the user-requested viewpoint V_U^t . If the same, $V_U^t = V_X^t$, then the user can immediately construct the desired image. The final reconstructed image $\widehat{I}_X^t$ is obtained by combining the prediction errors $\widehat{PE}^t$ with the predicted image PI_X^t . This kind of reconstruction is thereafter defined as *Type I Prediction*. Next,

the reconstructed noncentral image $\widehat{I}_X^t$ is stored in memory for later reference. Satisfying the user's request for this viewpoint image, the whole process is repeated for the next time frame.

On the other hand, if the user did not request the same viewpoint as the one selected by the encoder, $V_U^t \neq V_X^t$, the following steps are carried out to generate such a view. Given the user-requested viewpoint V_U^t , retrieve from memory the image $\widehat{I}_U^{t-f}$ corresponding to the nearest past image I_U reconstructed by a Type I Prediction. Similarly, retrieve the image $\widehat{I}_U^{t+b}$ corresponding to the nearest future image I_U reconstructed by a Type I Prediction. For example, referring to Fig. 3, if the user requests viewpoint V_R at time $t+2$, then the front image is I_R^{t+1} ($f=1$) and the back image is I_R^{t+5} ($b=3$). Due to the round-robin schedule for one noncentral viewpoint selection during each time frame, the maximum value of f and b is limited by $N-2$, where N is the number of viewpoints. Note that a delay may be required in order to access the future image.

Next, load the 3-D mesh model M^{t-f} created at time $t-f$ from $\widehat{I}_C^{t-f}$ and $\widehat{D}_C^{t-f}$. In parallel, load the 3-D mesh model M^{t+b} created at time $t+b$ from $\widehat{I}_C^{t+b}$ and $\widehat{D}_C^{t+b}$. Next, generate the front mesh image MI_U^{t-f} by locating a grid of fixed points on the image $\widehat{I}_U^{t-f}$. Also, generate the back mesh image MI_U^{t+b} by locating a corresponding grid of fixed points on the image $\widehat{I}_U^{t+b}$. Afterwards, generate the intermediate mesh-predicted image MPI_U^t for time t by digitally image warping between the front mesh image MI_U^{t-f} and the back mesh image MI_U^{t+b} . The image warping technique and generating mesh images by locating appropriate fixed points are well explained in [15].

Finally, a construction for $\widehat{I}_U^t$ is obtained by combining the intermediate mesh-predicted image MPI_U^t with the predicted image PI_U^t . The method for this final combination is left to the decoding system using any image fusion techniques [16], [17], e.g., XOR operator, average, etc. Construction of this final image $\widehat{I}_U^t$ is termed as *Type II Prediction*. At this point, if other viewpoint images are requested by the users, the prediction steps may be repeated to generate other multiviews. Otherwise, the decoding process starts at the beginning to decode the next time frame.

V. CONCLUSION

The proposed MPEG-2 compatible video coding specification for multiviewpoint systems offers many advantages. In addition to providing a structural framework for the new multiview profile of the MPEG-2 coding standard, this codec retains the same encoding and decoding processing for one basic viewpoint sequence based on the main profile. Furthermore, because the standard only defines the bit-stream syntax and the decoding process, there is freedom in designing very high quality encoders and very low cost decoders. Similarly, the presented specification suggests the same flexibility, providing for creativity and enhancements in accordance with advanced technologies.

To accommodate the delivery of an adaptive number of viewpoint sequences in an MPEG-2 compatible coding structure, a resilient and flexible multiview codec is required where reducing the image data by exploiting redundancy among the multiple views becomes crucial. Our novel contribution to the multiview coding research is mainly due to the concept of combining 2-D image processing with 3-D computer graphics, where 3-D model representations of the captured scene are incorporated with computer graphics tools and image warping techniques to construct the desired views. Many existing graphical tools and animation facilities, currently available in hardware, offer speedy rendering capabilities. Thus, interpolation of any viewpoint image sequence can be generated quickly and

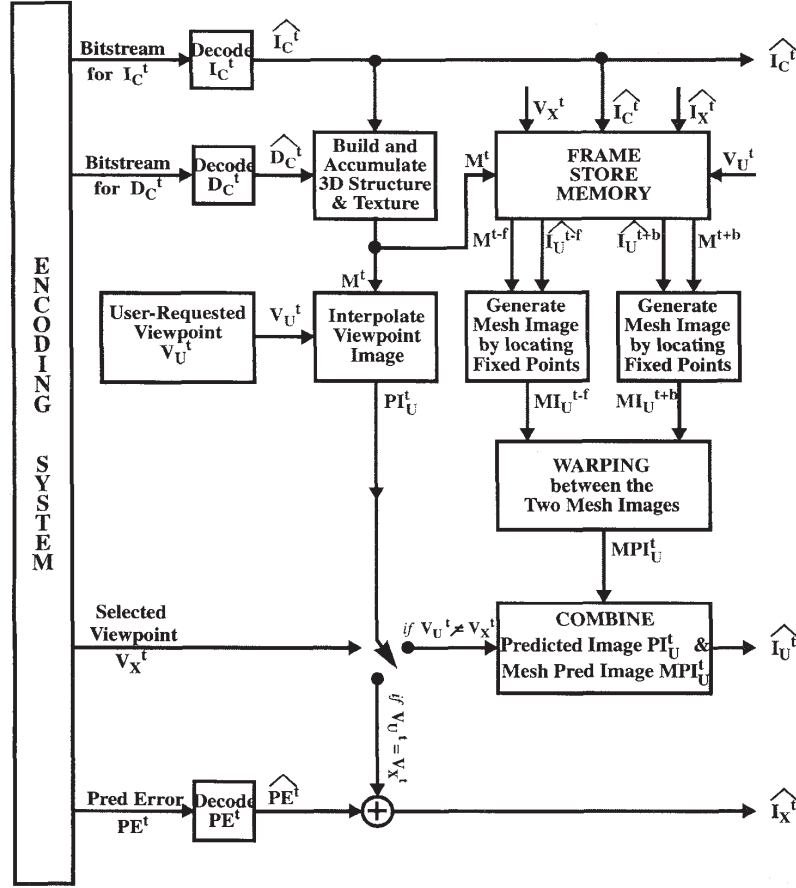


Fig. 4. Block diagram of multiview decoding system.

efficiently. Also, construction of nontransmitted virtual viewpoints is possible due to availability of 3-D models.

In rendering an image of a 3-D surface structure, the textured image data are mapped onto corresponding 3-D geometric meshed surfaces. Consequently, texture mapping automatically incorporates the fore-shortening effect of 3-D surface curvatures. In addition, since the texture is a 3-D function of its position in the object, tracking of image points is facilitated by knowing the real world 3-D coordinates of each point. Furthermore, additional enhancements on the constructed images may be obtained by incorporating other graphical tools, e.g., illumination and shading. Thus, innovative ideas can be freely augmented onto our system as technologies progress.

REFERENCES

- [1] "MPEG Draft International Standard. ISO/IEC 13818-2: Generic Coding of Moving Pictures and Associated Audio: Video," 1995.
- [2] B. L. Tseng and D. Anastassiou, "Compatible video coding of stereoscopic sequences using MPEG-2's scalability and interlaced structure," presented at *Int. Workshop HDTV '94*, Torino, Italy, Oct. 1994.
- [3] A. Puri, R. Kollarits, and B. Haskell, "Stereoscopic video compression using temporal scalability," in *Proc. SPIE Visual Commun. Image Processing '95*, Taipei, Taiwan, May 1995.
- [4] T.-H. Chiang and Y.-Q. Zhang, "Stereoscopic video coding," presented at *Symp. Multimedia Commun. Video Coding*, New York City, Oct. 1995.
- [5] B. L. Tseng and D. Anastassiou, "Perceptual adaptive quantization of stereoscopic video coding using MPEG-2's temporal scalability structure," presented at *Int. Workshop Stereoscopic and Three Dimensional Imaging IWS3DI '95*, Santorini, Greece, Sept. 1995.
- [6] W. Burger and B. Bhanu, "Estimating 3-D egomotion from perspective image sequences," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 1040-1058, Nov. 1990.
- [7] B. K. P. Horn, *Robot Vision*. Cambridge, MA: MIT Press, 1986.
- [8] D. Tzovaras, N. Grammalidis, and M. G. Strintzis, "Disparity field and depth map coding for multiview 3D image generation," *IEEE Trans. Image Processing*, to appear in 1996.
- [9] S. Barnard and M. Fischler, "Computational stereo," *ACM Computing Surveys*, vol. 14, pp. 553-572, Dec. 1982.
- [10] B. L. Tseng and D. Anastassiou, "A theoretical study on an accurate reconstruction of multiview images based on the Viterbi algorithm," presented at *Int. Conf. Image Processing ICIP '95*, Washington, DC, Oct. 1995.
- [11] T. Homma, "MPEG Contribution 95/N0861: Report of the Ad Hoc Group on MPEG-2 Applications for Multi-Viewpoint Pictures," Mar. 1995.
- [12] G. Bozdagi, A. M. Tekalp, and L. Onural, "3-D motion estimation and wireframe adaptation including photometric effects for model-based coding of facial image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 4, pp. 246-256, June 1994.
- [13] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practice*. 2nd ed. Reading, MA: Addison-Wesley, 1990.
- [14] J. Neider, T. Davis, and M. Woo, *OpenGL Programming Guide*. Reading, MA: Addison-Wesley, 1993.

- [15] G. Wolberg, *Digital Image Warping*. Los Alamitos, CA: IEEE Computer Society Press, 1990.
- [16] C.-P. Yeh, "Depth perception based on fusion of stereo images," *SPIE Vol. 1778 Imaging Technologies and Applications*, pp. 221–226, 1992.
- [17] Y. T. Zhou, "Multi-sensor image fusion," presented at *Int. Conf. Image Processing*, Austin, TX, Nov. 1994.

A Block-Based Gradient Descent Search Algorithm for Block Motion Estimation in Video Coding

Lurng-Kuo Liu and Ephraim Feig

Abstract—A block-based gradient descent search (BBGDS) algorithm is proposed in this paper to perform block motion estimation in video coding. The BBGDS evaluates the values of a given objective function starting from a small centralized checking block. The minimum within the checking block is found, and the gradient descent direction where the minimum is expected to lie is used to determine the search direction and the position of new checking block. The BBGDS is compared with full search (FS), three-step search (TSS), one-at-a-time search (OTS), and new three-step search (NTSS). Experimental results show that the proposed technique provides competitive performance with reduced computational complexity.

I. INTRODUCTION

The high correlation between successive frames of a video sequence makes it possible to achieve high coding efficiency in a video coding system by reducing the temporal redundancy. Motion compensated video coding technique, which predicts current frame from previous frame (or reference frame), has been used to exploit the temporal redundancy between successive frames. Motion estimation plays an important role in such an interframe predictive coding system. Among many types of motion estimation algorithms, block-matching technique has been adopted in many video compression standards, such as MPEG [1], [2], H.261 [3], and H.263 [4], due to its simplicity. In block-matching technique, frames are divided into blocks and one motion vector is associated with each block. For each block in the current frame, the motion estimation searches for a motion vector which points to the best match block in the reference frame. The best matching block is then used as the predictor for the current block.

The full search (FS) block-matching algorithm is the simplest, but computationally very intensive. It provides an optimal solution by exhaustively evaluating all the possible candidates within the search range in the reference frame. Several fast algorithms, such as the three-step search (TSS) [5], the 2-D logarithmic search (LOGS) [6], one-at-a-time search (OTS) [7], and the new three-step search (NTSS) [8] have been developed to reduce the computational complexity by reducing the number of checking points. Of these, the first three can be easily trapped into a local minimum, thereby degrading performance; see, for example [9]. The NTSS takes into account the fact that the distribution of global minimum in real world video sequences is centered at zero. This is especially true for head-and-

Manuscript received September 21, 1995; revised March 10, 1996. This paper was recommended by Associate Editor Y. Wang.

The authors are with IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA.

Publisher Item Identifier S 1051-8215(96)05194-4.

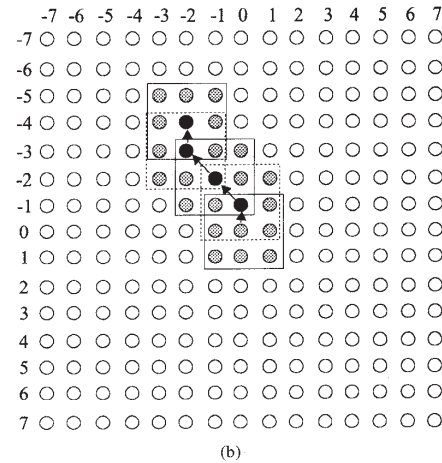
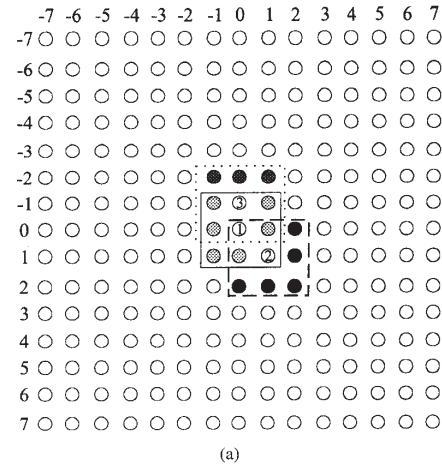


Fig. 1. (a) The first pass of BBGDS is centered around pixel (1); the second pass is centered around either pixel (2) or pixel (3). The shading highlights those new pixels that have to be checked in the second pass. (b) Example of the BBGDS search procedure, where motion vector $(-2, -4)$ is found.

shoulder sequences typical in video conferencing. The NTSS adds checking points which are centered about zero to the first step of the TSS. But usually the search stops after the first step, and then the second most likely scenario has the search ending after a second step which searches close to the first. So, on average, NTSS is somewhat faster than TSS. More significantly, for head-and-shoulder sequences, NTSS yields significant SNR gains over TSS.

In the present work we push the idea of utilizing the statistical nature of the motion even further. We further assume that the global minimum has a monotonic distortion in its neighborhood. In our first step we only search around the center point. If the optimum is found at the center, the procedure stops. This will be more than 80% of the time. Otherwise, we proceed to search around the point where the minimum was found. The procedure continues until the winning