Adaptive reference frame selection for generalized video signal coding

J. S. McVeigh[1], M. W. Siegel[2] and A. G. Jordan[1]

[1]Department of Electrical and Computer Engineering
[2]Robotics Institute, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213

ABSTRACT

In this paper, we present a new algorithm that adaptively selects the best possible reference frame for the predictive coding of generalized, or multi-view, video signals, based on estimated prediction similarity with the desired frame. We define similarity between two frames as the absence of occlusion, and we estimate this quantity from the variance of composite displacement vector maps. The composite maps are obtained without requiring the computationally intensive process of motion estimation for each candidate reference frame. We provide prediction and compression performance results for generalized video signals using both this scheme and schemes where the reference frames were heuristically pre-selected. When the predicted frames were used in a modified MPEG encoder simulation, the signal compressed using the adaptively selected reference frames required, on average, more than 10% fewer bits to encode than the non-adaptive techniques; for individual frames, the reduction in bits was sometimes more than 80%. These gains were obtained with an acceptable computational increase and an inconsequential bit-count overhead.

Keywords: pre-processing for video compression, motion estimation, multi-view image sequence compression.

1. INTRODUCTION

We define a *standard video signal* as a sequence of images obtained by sampling a scene in the time domain. A *generalized video signal* is an extension of this concept to the sampling of a scene in multiple domains. An example of a generalized video signal is a stereoscopic image sequence, where a scene is sampled in both the temporal and perspective domains. Possible applications of generalized video signals are remote inspection, tele-operation, and entertainment – or any application that may benefit from additional information of a scene, provided by multiple views.

Efficient compression of generalized video signals is required to avoid the linear increase in the bandwidth otherwise needed to transmit the signal with each additional view. We assume that each view is sampled in the time domain; thus, a generalized video signal can be viewed as a parallelization of individual standard video signals. A naive approach to coding a generalized video signal would be to treat each view, or stream, independently and to employ any available standard video signal compression technique (e.g., MPEG[1,2] or H.26P[3]). A more effective approach would attempt to exploit the potentially large correlations between the sampling domains. This approach has been demonstrated previously for both stereoscopic sequences and multi-spectral imagery, where compression is performed using both inter- and intra-view predictive coding.[4,6,7,9,15,16,17] However, in all prior work known to us, the reference frames used to predict a desired frame were both fixed and heuristically chosen. These reference frames do not necessarily yield the best prediction; compression performance suffers accordingly.

Prediction performance is related to the notion of *prediction similarity* between two image frames. Two frames have maximum similarity if the scene is static during the sampling interval. For the case of a standard video signal, the previous and next (or future) frames in time generally can be assumed to be the most similar to the current frame. The similarity relationship between frames of a generalized video signal is not as straightforward; it depends on the structure and motion of both scene objects and cameras, and it varies as the signal progresses. To improve compression performance, we desire a method to estimate the prediction similarity among frames of a generalized video signal that avoids the computationally costly process of motion estimation for each possible reference frame.

After a brief discussion on the generation of generalized video signals, we quantify prediction similarity as the absence of occlusion, and provide an estimate of this quantity from the variance of composite displacement vector maps. These novel

steps lead to the key contribution of this paper: the application of the similarity estimate to the adaptive selection of the best possible reference frame for the predictive coding of generalized video signals. We also provide comparisons with non-adaptive reference frame schemes, and we conclude with potential directions for future research.

2. GENERALIZED VIDEO SIGNAL PREDICTION

A standard video signal is obtained from a camera that samples the visual information of a scene in both a two-dimensional spatial grid (i.e., the image raster) and also temporally. The camera capturing this signal can be parameterized by its position and orientation in three-dimensional space, its zoom or scale factor, and its spectral band selectivity. While this video signal provides an enormous amount of information on the scene, a single camera can provide visual information from only one orientation/scale/wavelength-band at any given time instant. The benefits of added realism, improved scene analysis, and selective viewing can be achieved when multiple views of the scene are available.

The concept of a generalized video signal unifies the individual, standard video signals of applications that require more than one view of a scene under a common framework that illustrates the relationship between and within the various views. This multi-dimensional signal is indexed not only by the 2-D spatial grid and time axis of standard video signals, but also by the parameters that uniquely describe the individual cameras. Possible sampling domains (and applications) of generalized video signals include: perspective (binocular imagery), scale (multi-resolutional imagery), and wavelength (multi-spectral imagery). For simplicity, in this paper, we only will consider generalized video signals obtained from spatially displaced cameras.

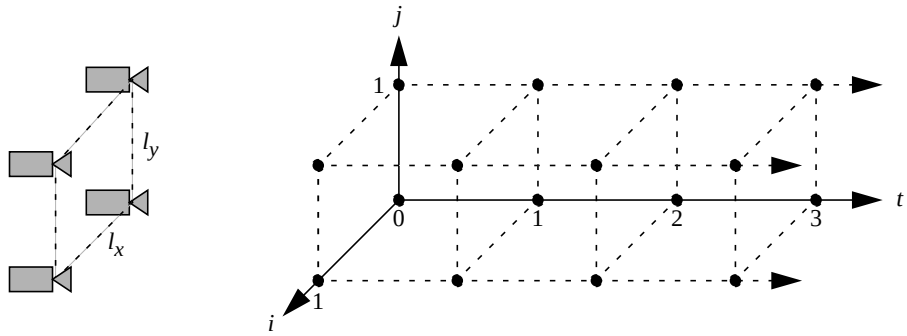
A possible application requiring multiple views from spatially displaced cameras, with identical scale and wavelength parameters, is that of a system designed to provide the viewer with simulated horizontal and vertical motion parallax[a]. Two-dimensional motion parallax can be synthesized by presenting the appropriate view, selected from a continuum of intermediate views within a bounding planar surface, according the observer's position. To avoid increases in camera complexity and transmission bandwidth, only the extremum views on the bounding surface need to be captured and the intermediate views can be generated via common image interpolation techniques.[5,8,13] The camera configuration and sampling structure of such a generalized video signal are depicted in Fig. 1. The bounding surface is an imaginary rectangle and the extremum views are obtained from four cameras positioned on the corners of the rectangle. We denote each frame of the resulting generalized video signal by its discrete sampling domain indices as: $F(i, j, t)$, where $x(i, j)$ and $y(i, j)$ respectively denote the horizontal and vertical coordinates of the camera that captured the frame.

While content- or object-based compression techniques may provide extremely compact representations of video[11], we feel that real-time and robust systems employing these techniques are still quite a few years away from reality. Therefore, we take the approach of a hybrid coder framework for the coding of generalized video signals, and consider the special characteristics of these multi-view signals to achieve superior compression performance. A hybrid coder consists of a prediction operation followed by a residual image encoding step. This framework is the basis for the recently proposed multi-view extension to the MPEG standard, which uses the temporal-scalability option of the standard to accommodate for multiple views of the scene.[14,15] In keeping with the MPEG class of video compression standards, the multi-view extension merely provides the allowable bit-stream syntax and, hence, the decoder structure. The encoder can make signal dependent decisions during the encoding process as long as the resulting bit-stream is compliant with the syntax.[10] For generalized video signals, one such decision that could provide substantial performance gains would be the adaptive selection of the best reference frame for the prediction process.

Under certain conditions, two frames of a generalized video signal, offset in some sampling domain(s), will be very similar or even identical. Consider, for example, the tetrocular camera configuration of Fig. 1 and the two streams denoted by

---

a. Motion parallax is the phenomena whereby changes in the viewer's position result in objects appearing to move, with the amount of displacement related inversely to the object's distance from the viewer.



a. Camera configuration b. Sampling structure

Figure 1: Four camera configuration (a) and sampling structure (b) of generalized video signal obtained from tetrocular set-up. The scene is sampled in the horizontal perspective (x), vertical perspective (y), and temporal (t) domains. The dots at each corner of the sampling structure represent the sampled image frames, and the horizontal lines represent the four image streams.

$F(0, 0, t)$ and $F(1, 0, t)$. The two cameras that captured these streams were separated by a horizontal spacing of l_x. We assume that the two cameras have equal temporal sampling periods of T. Two frames, offset in both time and perspective, will be identical if the scene objects are static and if the cameras move with constant horizontal velocity v, where l_x equals an integer multiple of vT. The same relationship would hold for the streams $F(0, 1, t)$ and $F(1, 1, t)$. For this situation one stream could be reconstructed exactly from another stream and knowledge of the temporal offset; thus, this generalized video signal could be compressed extremely compactly. While the scenario is an extremum where alignment is perfect, it illustrates the gain that may be achieved, for more realistic situations, through the exploitation of inter-stream correlations.

We now consider the situation where inter-stream correlation is exploited by predicting an image frame from a reference frame offset in one, or more, sampling domains. The prediction process for this signal is depicted by,

$$\tilde{F}(i, j, t) = \Psi\{\hat{F}(i - i_d, j - j_d, t - t_d)\} \tag{1}$$

where the prediction operator (Ψ) generates the predicted frame ($\tilde{F}$) from the reconstructed reference frame ($\hat{F}$), offset in the three sampling domains by (i_d, j_d, t_d). The offsets that produce the best prediction, according to some criteria, depend on the structure of the scene, the camera configuration, and the motion of both scene objects and cameras. Since at least some of these quantities will most likely vary considerably as the scene evolves and changes, the reference frames used in the prediction process should not be pre-selected. We find the optimum offsets through the maximization of estimated prediction similarity for all candidate reference frames. Since we are performing the prediction in the context of compression, the criteria for ranking the offsets that yield the best prediction should be based on the number of bits required to encode the residual image.

3. PREDICTION SIMILARITY

A brute force solution to the problem of adaptively selecting the optimum offsets would be to predict the desired frame from all possible reference frames and then just use the frame that yielded the best prediction. While this method is guaranteed to yield the best prediction performance, its implementation is impractical. In many video coders the prediction operation consumes the largest share of the processing cycles available. Each new view added to a generalized video signal would increase the number of prediction operations in proportion, i.e., the process is exponential.

Prediction similarity between two frames can be quantified by examining the process of image frame prediction. Points, or regions, within the desired frame can be accurately predicted only if corresponding points are also present in the reference frame. Conversely, regions of the desired frame occluded in the reference frame cannot be accurately predicted. Therefore, we wish to find the reference frame that has minimum occlusion with the desired frame to be predicted. Our definition of occlu-

sion incorporates all regions that prove difficult for the particular prediction process used, regardless of the source of this difficulty. For example, if the displacement of a region between two frames is described by an affine transformation and the prediction process only can handle translational motion (as is the case for common block-based techniques), we characterize the region as occluded.

The prediction process is completely described by a displacement vector map, which specifies the set of vectors that map pixels in the reference frame to each pixel in the desired frame. If a displacement vector map contains all zero vectors, the reference frame can be assumed to be identical to the desired frame since no displacement occurred between the frames. A high level of similarity also would occur if all of the displacement vectors were a constant, non-zero value; all points are displaced by a constant amount and occluded regions are present only at the borders of the images. Extending these observations leads to a fundamental property of displacement vector maps: an occlusion can occur only when there exists a discontinuity between the displacement vectors of neighboring pixels. The degree and size of discontinuity can be approximated from the variance of the displacement vector map.

Generating displacement vector maps and calculating the variance for all possible reference frames, however, is equivalent to the described brute force solution. We propose to estimate displacement vector maps for each possible reference frame through the synthesis of composite displacement vector maps. The composite maps are obtained through the vector addition of *single-step* displacement vector maps, which are analogous to partial derivatives of the pixel displacement with respect to the various sampling domains. Prediction similarity then is taken as the inverse of the variance of the composite vector map.

We require that one of the streams of the generalized video signal be predicted and coded independently of the other streams. This stipulation does not result in an overall loss of performance since prediction cannot be performed circularly. The independent stream frames are predicted using any conceivable combination of forward and backward predictions in the time domain. Each frame in the remaining, dependent streams is predicted from only one reference frame in another stream. The reference frames used in the prediction of the dependent stream frames are organized in a grid-like pattern to ensure prediction relationships between all frames. The displacement vector maps describing all predictions are retained. Since displacement estimation is performed only once for each dependent stream frame, the number of prediction operations increases only linearly with each addition view.

Since we are concerned with the variance of the actual displacement of image points between two frames, we wish to eliminate erroneous displacement estimates. These false estimates may be due to errors in the prediction process or the meaningless calculation of displacement for occluded regions. For simplicity, we attempt to minimize false estimates by reversing the estimated displacement vector maps. The first step in the reversal process is to count the number of times each pixel in the reference frame is used to predict the desired frame. If a pixel is referenced only one time, we assign the reversed-map pixel a displacement vector equal to the negative of the vector that referenced it, and we mark the vector as valid. If a pixel is referenced more than once, we select the vector that yielded minimum prediction distortion and mark the other vectors as invalid. We assume that the prediction process provides the distortion (e.g., the mean-squared error) for each vector. Finally, if a pixel is not referenced by any vectors, we also mark it as invalid. Only valid vectors are used in subsequent processing. Reversing the vector maps achieves the goals of reducing erroneous estimates and it allows for similarity calculations in both directions, without requiring bidirectional displacement estimation.

The next step in our prediction similarity estimation procedure is to calculate composite displacement vector maps for each reference frame to be examined. The composite maps are generated through the addition of valid vectors from the processed single-step displacement maps that relate the reference frame to the desired frame[b]. The relative presence of occlusion (r), in one image frame dimension, is estimated from (2) and (3),

---

b. These composite displacement vector maps cannot be used to provide an accurate estimate of prediction similarity through direct prediction. The composite maps are generated for regions that are assumed to be unoccluded, and do not cover the entire frame. The resulting prediction performance from using these maps will likely be unrelated to the actual prediction performance for the complete frame.

$$r = \beta\sigma^2_{\text{composite}} \tag{2}$$

$$\beta = \frac{\sigma^2_{\text{composite}}}{\sigma^2_{\text{previous}} + \sigma^2_{\text{current}}} \tag{3}$$

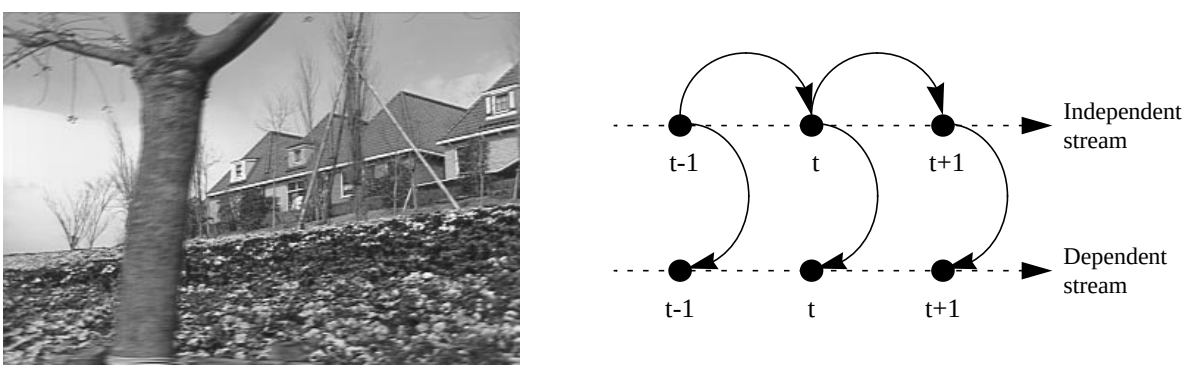
where $\sigma^2_{\text{composite}}$, $\sigma^2_{\text{previous}}$, and $\sigma^2_{\text{current}}$ are the variances of the composite, previous composite, and current displacement vector maps, respectively. The quantity β weights the composite map variance by the relative increase in variance due to the addition of the two vector maps. This weighting attempts to compensate for the accumulation of errors when generating composite maps from multiple single-step displacement vector maps. The prediction similarity measure is taken as the inverse of the L_2-norm of the relative presence of occlusion in both horizontal and vertical image dimensions (r_u and r_v). The reference frame that yields the maximum estimated similarity is selected and used for the final prediction of the desired frame.

4. EXPERIMENTAL RESULTS

The algorithm was tested on three generalized video signals. All predictions were performed using a block-based technique with $16x16$ blocks.[12] For simplicity, the independent streams were predicted using forward prediction only, similar to the H.261 standard.[3]

As a proof of concept, the first signal examined was generated by modifying the standard video signal, *Flower Garden*. A sample frame of this sequence is shown in Fig. 2a. A two-view generalized video signal was simulated by using the same sequence twice with a variable temporal offset. The independent stream consisted of the continuously numbered 149 frames of the original sequence. The dependent stream had a relative offset that randomly varied from $[-2 \ldots 2]$ frames. The single-step prediction structure used to generate the composite displacement vector maps and to estimate the prediction similarity is shown in Fig. 2b. The known, optimum reference frame was selected correctly approximately 75% of the time. When an error did occur, the similarity measure for the correct reference frame was only slightly less than that of the selected reference frame.

For a more meaningful evaluation, we examined the algorithm's performance on two stereoscopic sequences (*Buggy* and *Finish Line*) captured by the authors. The stereoscopic camera consisted of two cameras offset by a horizontal distance of 50 mm. The odd-line fields of a standard NTSC frame were captured by the left-eye camera, and the even-line fields by the right-eye camera. The resulting temporal sampling of the two streams was offset by $\frac{T}{2} = \frac{1}{60}$ of a second. For both sequences the



a. b.

Figure 2: *Flower Garden* sequence. a) Original stream 0, frame 25, b) Prediction structure used to compute single-step predictions. Curved arrows represent prediction of desired frame (at arrow head) from reference frame.

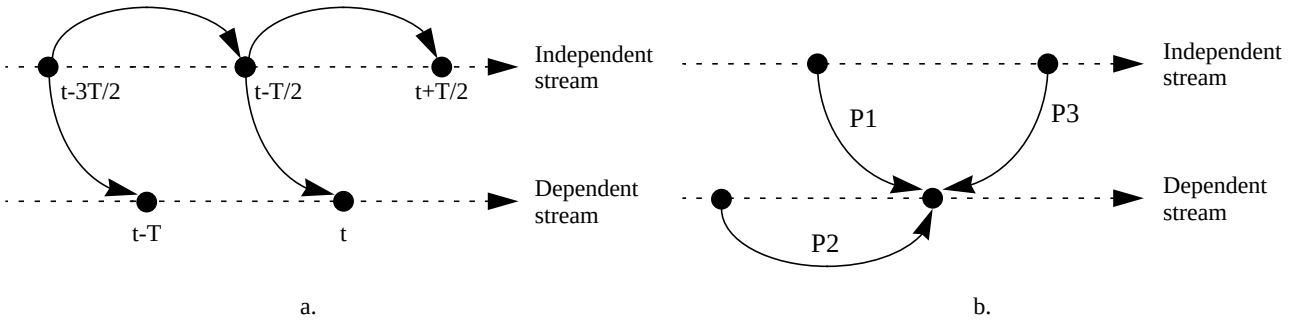


Figure 3: Stereoscopic sequence prediction structure. a) Single-step predictions b) Relative positions of possible reference frames for dependent stream, for both fixed and adaptively selected reference frame schemes.

left-eye image stream was predicted independently of the right-eye stream. The single-step predictions used to generate the composite displacement vector maps are shown in Fig. 3a. The relative positions of the three possible reference frames for the prediction of the dependent stream are shown in Fig. 3b. The reference frame used by the prediction denoted by P1 was used to generate the single-step displacement vector map relating the dependent and independent streams. The prediction similarity for the reference frames of all three possible predictions (P1, P2 and P3) were estimated using the described procedure. If the maximum estimated similarity was obtained for the reference frames of P2 or P3, the frame was predicted using the specified reference frame; otherwise, the initial prediction was used. Sample frames for both sequences are shown in Figs. 4a and 4b.

The *Buggy* sequence is characterized as containing a large degree of both camera and object motion. Prediction and bit count performance results for the adaptive reference frame selection algorithm are shown in Figs. 5a and 5b, respectively. The performance results for prediction schemes that used fixed reference frames, denoted by P1, P2, and P3, are also included. Prediction performance was quantified with the peak signal-to-noise ratio (PSNR) between the luminance component of the predicted and original images. The frame bit counts were obtained from a modified MPEG-1 simulation that performed conventional DCT-based compression on the residual between the predicted and original frames. The bits-per-frame values include the bits required to describe the selected reference frame, displacement vectors, and header information. The quantization step-size of the coder was fixed, resulting in a fixed reconstructed image quality of approximately 35 dB (PSNR). We believe the eight sharp peaks in both plots are due to a malfunction in our image digitization mechanism that resulted in some

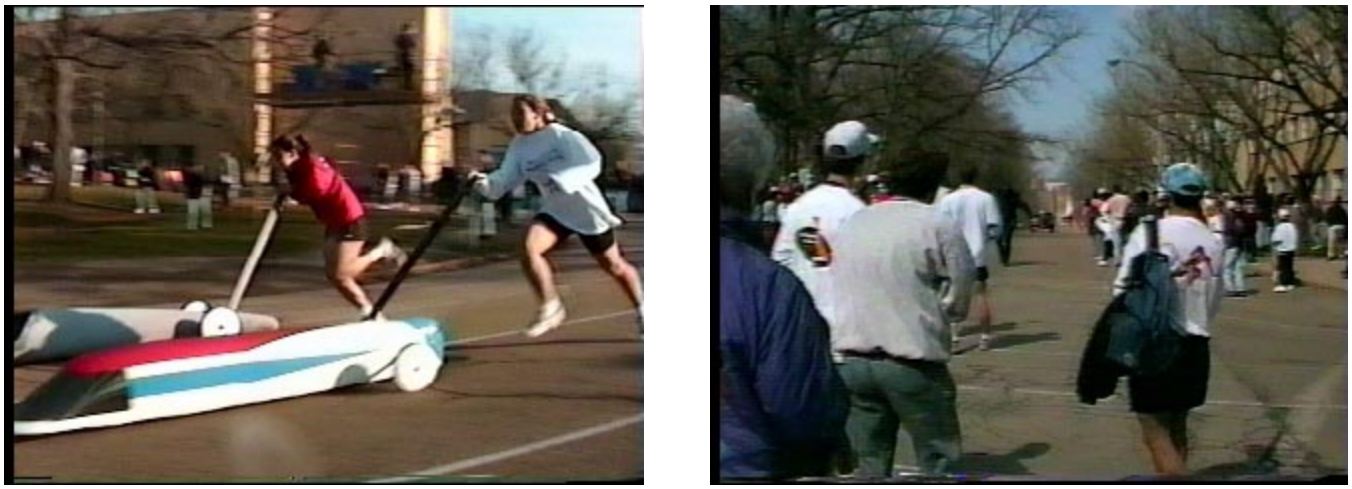
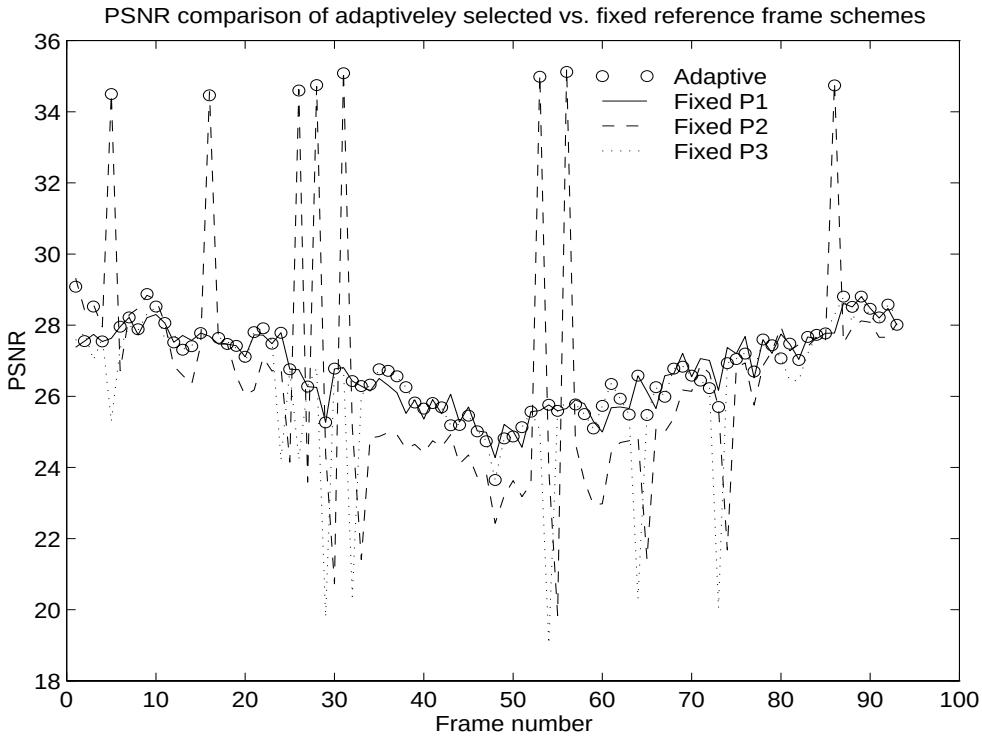
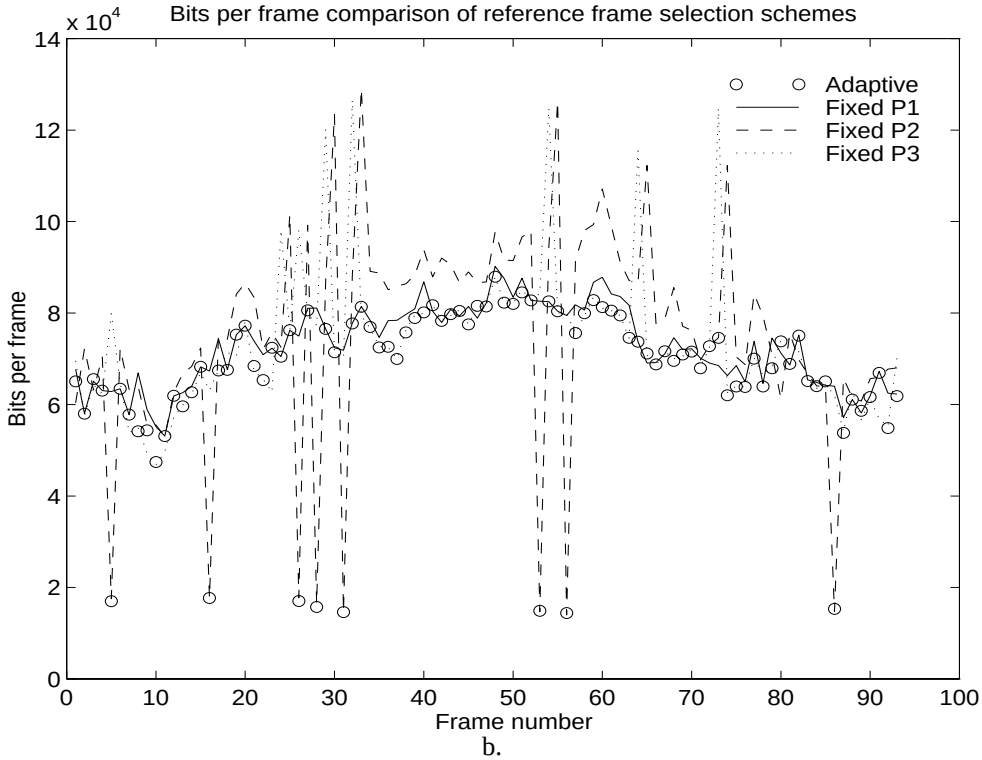


Figure 4: Sample frames. a) *Buggy* stream 0, frame 40, b). *Finish Line* stream 0, frame, 100



a.



b.

Figure 5: *Buggy* sequence performance comparison of adaptively selected versus fixed reference frame schemes. The reference frames used in the three fixed schemes are illustrated in Fig. 3b. The best reference frame is almost always selected by the adaptive scheme. When the optimum reference frame is not correctly selected, the chosen reference frame is most often only slightly inferior.a) Prediction PSNR comparison, b) Bits per frame comparison.

frames being recorded twice. Regardless of the source of these anomalies, the algorithm correctly selected the best frame for the prediction process.

The *Finish Line* sequence also was predicted and encoded using both adaptively selected and fixed reference frames schemes denoted by P1, P2, and P3. The average prediction PSNR and frame bit counts for these fixed schemes are shown in Table 1. This sequence contains almost no camera or object motion over time. However, the two views have a large degree of occlusion across the perspective sampling domain due to the proximity of objects to the stereo-camera. The algorithm, consequently, always selected the inter-view reference frame (i.e., prediction P2) as the frame with maximum similarity.

Prediction	**PSNR**	**Bits-per-frame**
P1	25.73	92284
P2	29.94	64130
P3	25.57	94291

Table 1: Average per-frame prediction and compression performance for the non-adaptive reference frame schemes. The adaptive algorithm always selected the reference frame used by the P2 prediction structure.

5. CONCLUSION

We have presented a simple scheme for adaptively selecting the best possible reference frame for the predictive coding of generalized video signals. One stream of the generalized video signal is predicted and coded independently of the other dependent streams. Single-step displacement vector maps are estimated for each frame in the dependent streams. Vector estimates for occluded regions are discarded through processing of the single-step maps, and composite maps are generated for each candidate reference frame. The composite maps estimate the prediction operation for unoccluded regions from the given reference frame. The reference frame with the estimated maximum similarity with the desired frame is chosen for the final prediction, where similarity is defined as the absence of occlusion. This scheme requires a maximum of two frame predictions for each image, as opposed to an exponential increase, with each additional view, for a brute force solution.

The results for the signal obtained by modifying the *Flower Garden* sequence indicate that this scheme most often selects the known, optimum reference frame. The correct reference frame was not selected only on a few occasions when the relative temporal offset between the two streams was ± 2 frames.

Prediction and compression of the stereoscopic sequence, *Buggy*, was performed using both adaptive and fixed reference frame schemes. The relative location of the optimum reference frame varied greatly throughout the sequence. This variation is most likely the result of the large degree of motion contained within this sequence and it validates the hypothesis that the reference frame should be adaptively selected. The average PSNR prediction gain of the adaptive technique over the fixed schemes P1, P2, and P3 was 0.7 dB, 1.0 dB, and 1.2 dB, respectively. The average reduction in bits per frame was approximately 9%, 13%, and 10% over P1, P2, and P3, respectively. While the average reduction in bits is modest, for certain frames, the number of bits required to encode the residual was reduced by over 80%. Also, we anticipate improved performance gains with increased signal dimension, and, hence, an increased number of candidate reference frames.

While the relative location of the best reference frame did not vary for the *Finish Line* sequence, this location likely would not be known a priori. In fact, if either fixed prediction schemes P1 or P3 were used in the predictive coding of this sequence, the prediction PSNR would have decreased by over 4 dB and the bit rate would have increased by more than 44% when compared to scheme P2, which was adaptively selected by our algorithm.

These experimental results indicate that the algorithm works and that it should be performed to improve the compression performance for generalized video signals. Future work includes the application of this technique to more elaborate generalized video signals, and the possible refinement of the estimated similarity measure to further improve the accuracy of the selection. Throughout this discussion we have conveniently neglected the issues of storage and delay required to allow for adaptively selected reference frames. The cost/benefit analysis of this flexibility will be addressed in a future paper.

6. ACKNOWLEDGMENT

This work was supported by the Advanced Research Projects Agency under ARPA Grant. No. MDA 972-92-J-1010.

7. REFERENCES

1. ISO/IEC JTC1/SG29/WG11, ISO/IEC 11172-2, “Information Technology - Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s - Part 2: Video”, May 1993.

2. ISO/IEC JTC1/SC29/WG11 Test Model Editing Committee, “MPEG-2 Video Test Model 5", ISO/IEC JTC1/SC29/WG11 Doc. N0400, April 1993.

3. ISO/IEC JTC1/SC29/WG11, “Information Technology - Generic Coding of Moving Pictures and Associated Audio, Recommendation H.262, ISO/IEC 13818-2, Draft International Standard,” March 1994.

4. G. P. Abousleman, M. W. Marcellin and B. R. Hunt, “Compression of hyperspectral imagery using the 3-D DCT and hybrid DPCM/DCT,” *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 33, No. 1, pp. 26-34, January 1995.

5. H. Aydinoglu and M. H. Hayes, “Compression of multi-view images,” *Proc. IEEE Internat. Conf. on Image Processing*, Vol. 2, pp. 385-389, Austin, TX, 13-16 November 1994.

6. R. Chassaing, B. Choquet and D. Pele, “A stereoscopic television system (3D-TV) and compatible transmission on a MAC channel (3D-MAC)”, *Signal Processing: Image Communication*, Vol. 4, No. 1, pp. 33-43, November 1991.

7. S. Gupta and A. Gersho, “Feature predictive vector quantization of multispectral images,” *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 30, No. 3, pp. 491-501, May 1992.

8. R. Hsu, K. Kodama and H. Harashima, “View interpolation using epipolar plane images,” *Proc. IEEE Internat. Conf. on Image Processing*, pp. 745-749, Vol. 2, Austin, TX, 13-16 November 1994.

9. M. E. Lukacs, “Predictive coding of multi-viewpoint image sets,” *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 521-524, Tokyo, Japan, 1986.

10. J. N. Mailhot and H. Derovanessian, “Grand Alliance HDTV Video Encoder,” *IEEE Trans. on Consumer Electronics*, Vol. 41, No. 4, pp. 1014-1019, November 1995.

11. F. C. M. Martins and J. M. F. Moura, “3-D video compositing: Towards a compact representation for video sequences,” *Proc. IEEE Internat. Conf. on Image Processing*, pp. 550-553, Washington, DC, 23-26 October 1995.

12. J. S. McVeigh and S.-W. Wu, “Partial closed loop versus open loop motion estimation for HDTV compression,” International Journal of Imaging Science and Technology, Vol. 5, No. 4, pp. 268-275, 1994.

13. J. S. McVeigh, M. W. Siegel and A. G. Jordan, “Intermediate view synthesis considering occluded and ambiguously referenced image regions,” *Signal Processing: Image Communications*, accepted.

14. A. Puri and B. Haskell, “Straw man proposal for multi-view profile,” ISO/IEC JTC1/SC29/WG11 MPEG95/485, November 1995.

15. A. Puri, R. V. Kollarits and B. G. Haskell, “Stereoscopic video compression using temporal scalability,” *Proc. SPIE Internat. Conf. on Visual Communications and Image Processing*, Vol. 2501, pp. 745-756, Taipei, Taiwan, 23-26 May 1995.

16. A. Schertz, “Source coding of stereoscopic television pictures”, *Proc. IEE Internat. Conf. on Image Processing and its Applications*, pp. 462-464, Maastricht, Netherlands, 7-9 April 1992.

17. S. Sethuraman, M. W. Siegel, and A. G. Jordan, “A multiresolution framework for stereoscopic image sequence compression,” *Proc. IEEE Internat. Conf. on Image Processing*, Vol. 2, pp. 361-365, Austin, TX, 13-16 November 1994.