Compression of stereo image pairs and streams

M. W. Siegel[1]
Priyan Gunatilake[2]
Sriram Sethuraman[2]
A. G. Jordan[1,2]

[1]Robotics Institute, School of Computer Science
[2]Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA, 15213

ABSTRACT

We exploit the correlations between 3D-stereoscopic left-right image pairs to achieve high compression factors for image frame storage and image stream transmission. In particular, in image stream transmission, we can find extremely high correlations between left-right frames offset in time such that perspective-induced disparity between viewpoints and motion-induced parallax from a single viewpoint are nearly identical; we coin the term "WorldLine correlation" for this condition. We test these ideas in two implementations, (1) straightforward computing of blockwise cross- correlations, and (2) multiresolution hierarchical matching using a wavelet- based compression method. We find that good 3D-stereoscopic imagery can be had for only a few percent more storage space or transmission bandwidth than is required for the corresponding flat imagery.

1. INTRODUCTION

The successful development of compression schemes for motion video that exploit the high correlation between temporally adjacent frames, e.g., MPEG, suggests that we might analogously exploit the high correlation between spatially or angularly adjacent still frames, i.e., left-right 3D-stereoscopic image pairs. If left-right pairs are selected from 3D-stereoscopic motion streams at different times, such that perspective-induced disparity left-right and motion-induced disparity earlier-later produce about the same visual effect, then extremely high correlation will exist between the members of these pairs. This effect, for which we coin the term "WorldLine correlation", can be exploited to achieve extremely high compression factors for stereo video streams.

Our experiments demonstrate that a reasonable synthesis of one image of a left-right stereo image pair can be estimated from the other uncompressed or conventionally compressed image augmented by a small set of numbers that describe the local cross-correlations in terms of a disparity map. When the set is as small (in bits) as 1 to 2% of the conventionally compressed image the stereoscopically viewed pair consisting of one original and one synthesized image produces convincing stereo imagery. Occlusions, for which this approach of course fails, can be handled efficiently by encoding and transmitting error maps (residuals) of regions where a local statistical operator indicates that an occlusion is probable.

Two cross-correlation mapping schemes independently developed by two of us (P.G. and S.S.) have been coded and tested, extensively on still image pairs and more recently on some motion video streams. Both methods yield comparable compression factors and visual fidelity; which can be coded more efficiently, and whether either can be coded efficiently enough to make it practical for real time use, is under study.

The method developed by P.G. is based on straightforward computing of blockwise cross-correlations; heuristics that direct the search substantially improve efficiency at the price of occasionally finding a local maximum rather than the global maximum.

The method developed by S.S. is based on multiresolution hierarchical matching using wavelets; efficiency is achieved by doing the search for the best match down a tree of progressively higher resolution images, starting from a low resolution highly subsampled image.

In the following sections we discuss the need and opportunity for compression of 3D-stereoscopic imagery, discuss the correlations that can be exploited to achieve compression, describe and refine the approach, summarize the content and performance of the two implementations we have prototyped to date, and outline several topics we have targeted for ongoing research.

This paper is intended as a high level introduction to our thoughts about and our progress toward compression for 3D-stereoscopy. The specific references that we cite in the text and the general references that we also include in the bibliography point to background literature, as well as to three recent papers [1, 2, 3] in which we document the low level details of our recent work.

2. NEED AND OPPORTUNITY

The scenario we imagine is that binocular 3D-stereoscopy is grafted onto "flat" (monoscopic) display infrastructures; we regard the alternative scenario, that 3D-stereoscopy is built into the foundations of the infrastructure, as being somewhat farfetched in light of the cost and effectiveness of the current generation of 3D display devices and systems.

Displays become rapidly more expensive as their spatial resolution and temporal frame rate increases. Thus in any application the display is usually chosen to meet but not to exceed substantially the application's requirements. In flat applications each eye sees, at no cost to the other eye, the full spatial and temporal bandwidth that the display delivers. When a 3D-stereoscopic application is grafted onto a flat infrastructure the display's capabilities must be divided between the two eyes. The price may be extracted in either essentially the spatial domain, e.g., by assigning the odd lines to the left eye and the even lines to the right eye, or in essentially the temporal domain, e.g., by assigning alternate frames to the left and right eye. The distinction is in part semantic, since the "spatial" method of this example is often implemented in practice via sequential fields in an interlaced display system. The fundamental issue is that when 3D-stereoscopy is implemented on a single display each eye gets in some sense only half the display. A user contemplating using 3D-stereoscopy must thus acquire a display (and the underlying system to support it) with twice the pixel-per-second capability of the minimal display needed for the flat application; the alternatives require choosing between a flickering image or a reduced spatial resolution image.

As indicated, lower level capacities of the system's components must also be doubled. In particular, all the information captured by two cameras (each equivalent to the original camera) must be stored or transmitted or both. Doubling these capacities may be more difficult than doubling the capability of the display, inasmuch as (except at the very high end) the capability of the display can be increased by simply paying more. The most difficult system component to increase is probably the bandwidth of the transmission system, which is often subject to powerful regulatory as well as technical

constraints. Nevertheless, the bandwidth must apparently be doubled to transmit 3D-stereoscopic image streams at the same spatial resolution and temporal update frequency as either flat image stream.

In fact, because the two views comprising a 3D-stereoscopic image pair are nearly identical, i.e., the information content of both together is only a little more than the information content of one alone, it is possible to find representations of image pairs and streams that take up little more storage space and transmission bandwidth than the space or bandwidth that is required by either alone. The rest of this paper is devoted to an overview of how this can be done, some details of our early implementations, and a discussion of possibilities for the future.

2.1. Background

We remind the reader that image compression methods fall into two broad categories, "lossless" and "lossy". Lossless compression exploits the existence of redundant or repeated information, storing the image in less space by symbolically rather than explicitly repeating information, and by related methods such as assigning the shortest codes to the most probable occurrences. Lossy compression exploits characteristics of the human visual system by discarding image content that is known to have little or no impact on human perception of the image.

Our approach to compression of 3D-stereoscopic imagery has two components, related to there being two perspective views in a 3D-stereoscopic pair. One component may be either lossless or slightly lossy, as in conventional compression of flat imagery; the other component is by itself a very lossy (or "deep") method of compression. The intimate connection between the two views makes it possible to synthesize a perceptually acceptable image from a compression so deep that, by itself, it would be incomprehensible.

The left and right views that comprise a 3D-stereoscopic image pair or motion stream pair are obviously very similar. There are various ways of saying this: they are often described as "highly redundant", in that most of the information contained in either is repeated in the other, or as "highly correlated" in that either is for the most part easily predicted from the other by application of some external information about the relationship (the relative perspective) between them. We can thus synthesize a reasonable approximation to either view given the other view and a little additional information that describes the relationship between the two views. A useful form for the additional information is a disparity map: a two dimensional vector field that encodes how to displace blocks of pixels in one view to approximate the other view.

Fortunately a "reasonable approximation" is enough: perfection is not required. This is the case because of two psychophysical effects, one well known, the other less so.

It is well known that one good eye and one bad eye together are better than the good eye alone, i.e., the information they provide in a sense adds rather than averages. The resulting perception is sharper than the perception provided by the better eye alone. Thus presenting one eye with the original view intended for it, and presenting the other eye with a synthetic view (which might be imperfect in sharpness and perhaps even missing some small features), the perception of both together is better than the perception of the original view alone.

A related perceptual effect that we have observed informally has been documented in several controlled experiments: a binocular 3D-stereoscopic image pair with one sharp member and one blurred member successfully stimulate appropriate depth perception.

Thus we expect that if one member of a 3D-stereoscopic image pair is losslessly or nearly losslessly compressed and the other is (by some appropriate method) deeply compressed, the pair of decompressed (higher resolution) and synthesized (lower resolution) views will together be perceived comfortably and accurately.

In the following section we describe several approaches to compression, ultimately focusing on the method we are now developing along two complementary implementation paths.

2.2. Correlations

We identify four kinds of correlations or redundancies that can be exploited to compress 3D-stereoscopic imagery. The first two make no specific reference to 3D-stereoscopy; they are conventional image compression methods that might (inefficiently!) be applied to two 3D-stereoscopic views independently. The third kind applies to still image pairs, or to temporally corresponding members of a motion stream pair. The fourth kind, which is really a combination of the second and third kinds, applies to motion stream pairs.

- ***Spatial correlation***: Within a single frame, large areas with little variation in intensity and color permit efficient encoding based on internal predictability, i.e., the fact that any given pixel is most likely to be identical or nearly identical to its neighbors. This is the basis for most conventional still image compression methods.
- ***Temporal correlation***: Between frames in a motion sequence, large areas in rigid-body motion permit efficient coding based on frame-to-frame predictability. The approach is fundamentally to transmit an occasional frame, and interpolation coefficients that permit the receiver to synthesize reasonable approximations to the intermediate frames. MPEG is an example.
- ***Perspective correlation***: Between frames in a binocular 3D-stereoscopic image pair, large areas differing only by small horizontal offsets permit efficient coding based on disparity predictability. If one imagines the two perspective views as being gathered not simultaneously but rather sequentially by moving the camera from one viewpoint to the second, then perspective correlation and temporal correlation are to first order equivalent.
- ***WorldLine correlation***: We borrow the term "worldline" from the Theory of Special Relativity, where the worldline is a central concept that refers to the path of an object in 4-dimensional space-time. Observers moving relative to each other, i.e., observers having different perspectives on space-time, perceive a worldline segment as having different spatial and temporal components, but they all agree on the length of the segment. Analogously in 3D-stereoscopic image streams, when vertical and axial velocities are small and horizontal motion suitably compensates perspective, time-offset frames in the left and right image streams can be nearly identical. WorldLine correlation is the combination of temporal correlation and perspective correlation; the most interesting manifestation of WorldLine correlation is the potential near-identity of appropriately time-offset frames in the left and right image streams respectively.* The concept is useful for situations in which the camera is fixed and parts of the scene are in motion, the scene is fixed and the camera is in motion, and both the camera and parts of the scene are in motion.

WorldLine correlation is depicted pictorially in Figure 1.

*Thinking in a suitable generalized fourier domain, simultaneous pairs from different perspectives and pairs from one perspective at different times are characterized by nearly identical amplitude spectra but substantially (although systematically) different phase spectra.

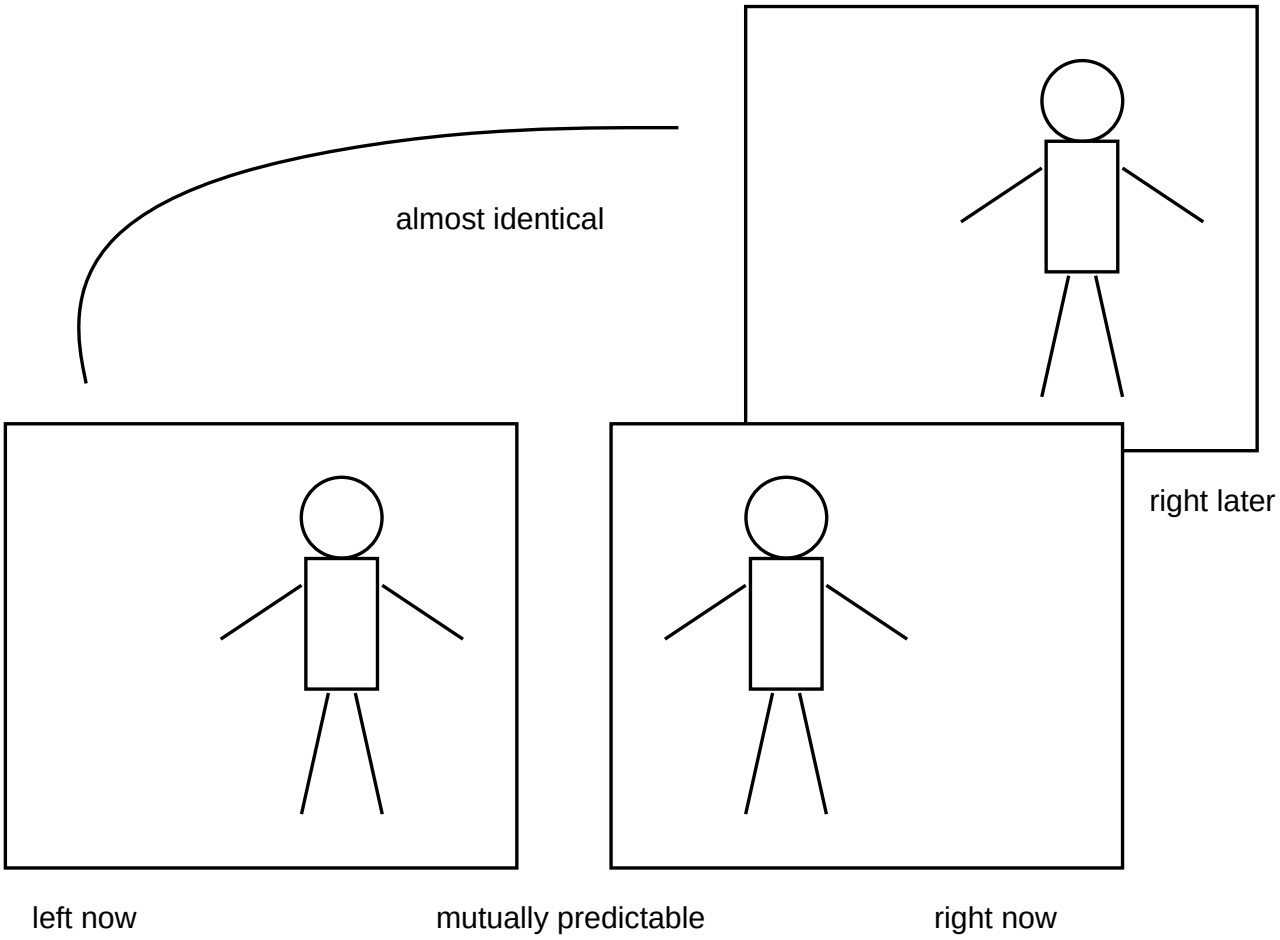

Figure 1: Pictorial depiction of WorldLine correlation.

3. APPROACH

3.1. Basic Approach

Our basic approach to compression of 3D-stereoscopic imagery is based on the observation that disparity, the relative offset between corresponding points in an image pair, varies only slowly over most of the image field. Given the validity of this assumption, either member of an image pair can be synthesized (or "predicted") given the other member and a low-resolution map of the relative disparity between the two members of the pair. It is the possibility that the disparity map can be low resolution, combined with the fact that the disparities vary slowly and can be represented by small numbers (few bits) that permits deep compression.

As a numerical example, suppose that over most of the image field the disparity does not change significantly over eight pixels. Then a disparity map can be represented by a field with 1/64 the number of entries as the image itself. Each disparity is a vector with two components, horizontal and vertical, so the net compression has an upper bound of 1/32, about 3%. In fact further significant advantages can be obtained by recognizing that the disparity components can be encoded with fewer bits than the original intensities, e.g., perhaps three bits for the vertical disparities (four pixels up or down) and perhaps five

bits for the horizontal disparities (sixteen pixels left or right). Removal of redundancy in this map, e.g., run length encoding, leads to even further gains.

Our basic approach to coding 3D-stereoscopic image pairs, or corresponding pairs of a 3D-stereoscopic image stream, is easily seen from the following outline:

- Generate:
  - Code either image conventionally
  - Code the disparity map
- Store/Move:
  - Transmit the coded components
- Use:
  - Decode the conventionally coded image
  - Decode the disparity map
  - Synthesize the other image
  - Display 3D-stereoscopically

3.2. Problem with the Basic Approach

The basic approach has a basic fault: it cannot cope with occlusions, i.e., features that can be seen from only one of the two perspectives. This follows simply from the fact that the synthesized view is just a "rubber sheet" map of the conventionally compressed view. Thus features that are occluded in the conventionally compressed view (visible only in the view that is subsequently deeply compressed) cannot be synthesized. Similarly, features that are visible in the conventionally compressed view but are occluded in the subsequently deeply compressed view do not fit comfortably into this scheme.

The human visual perception system has an effective way to deal with occlusions: we have a detailed understanding of the image semantics, from which we effortlessly and unconsciously draw inferences that fill in the missing information. If this capability could be duplicated in a computer algorithm it would be essentially the solution to the general image understanding problem; its pursuit, let alone its solution, is beyond the scope of the present work.

Fortunately a pragmatic alternative exists: we can code and transmit the residuals (a map of the pixel-by-pixel differences between the original and its prediction from the disparity map). The differences are usually small, permitting it to be coded efficiently by conventional methods. In fact we can achieve a particularly efficient implementation in either of two equivalent ways. Both approaches work by coding and transmitting the residuals only in limited regions. In one approach the residuals are preserved only where they exceed a predetermined threshold. In the other approach a local statistical operator is used to identify regions in the image where occlusions are probable, and the residuals are computed, coded, and transmitted only for these regions.

3.3. Resulting Hybrid Method

The result is a hybrid algorithm whose flow should be clear from the preceding discussion, but which we will outline explicitly for completeness:

- Generate:
  - Code one image conventionally
  - Code the disparity map

    - Code the residuals of the prediction
- Store/Move:
    - Transmit the coded components
- Use:
    - Decode the conventionally coded image
    - Decode the disparity map
    - Synthesize the other image
    - Decode the residuals
    - Add the residuals to the prediction
    - Display 3D-stereographically

We are also conducting several subsidiary experiments aimed at understanding how the detailed coding scheme can be optimized for the human perceptual system. For example, it seems plausible that rapidly alternating which eye sees the conventionally compressed view and which eye sees the deeply compressed view may be more comfortable than fixing this choice. We are testing this and comparable hypotheses.

4. IMPLEMENTATION AND RESULTS

We have implemented two methods and are experimenting with them in parallel.

The first method, implemented by P.G., uses straightforward blockwise cross-correlation. This is the obvious candidate for initial experiments because it is easy to code and because we have a strong intuitive understanding of its parameters. It is thus straightforward to experiment with and understand the results of varying the parameters. In this implementation simple heuristics efficiently direct the matching search, decreasing the run time of the algorithm; however, as expected, avoiding exhaustive search makes the method somewhat prone to finding erroneous local matches.

The second method, implemented by S.S., uses a wavelet-based multiresolution hierarchical matching approach. The high spatial frequency content of the image is preserved at half the initial resolution; despite its high resolution, it can be coded efficiently because pixel values differ significantly from zero only in the immediate vicinity of the edges in the original image. The low spatial frequency content of the image is preserved in reduced resolution images. High and low frequency sub-images are computed down several hierarchical levels. The disparity map is built from the bottom up in a coarse-to-fine updated search; it is thus robust against finding incorrect local matches. It is computationally efficient, essentially because compression and disparity map building make use of the same intermediate results. Its hierarchical structure permits graceful degradation with lower-capability displays or noisy channels.

To date we have demonstrated in both implementations:

- Acceptable binocular perception with 1 to 2% of the total bandwidth allocated to disparity coding, and
- Excellent binocular perception with 10 to 20% of the total bandwidth allocated to disparity and residual coding.

For example, Figure 2 shows an original right and left 3D-stereoscopic image pair, and Figure 3 shows the right image after conventional compression and decompression and the left image synthesized from the left member of Figure 2 and the disparity map computed (by the simple block matching method) from the left and right members of Figure 2.

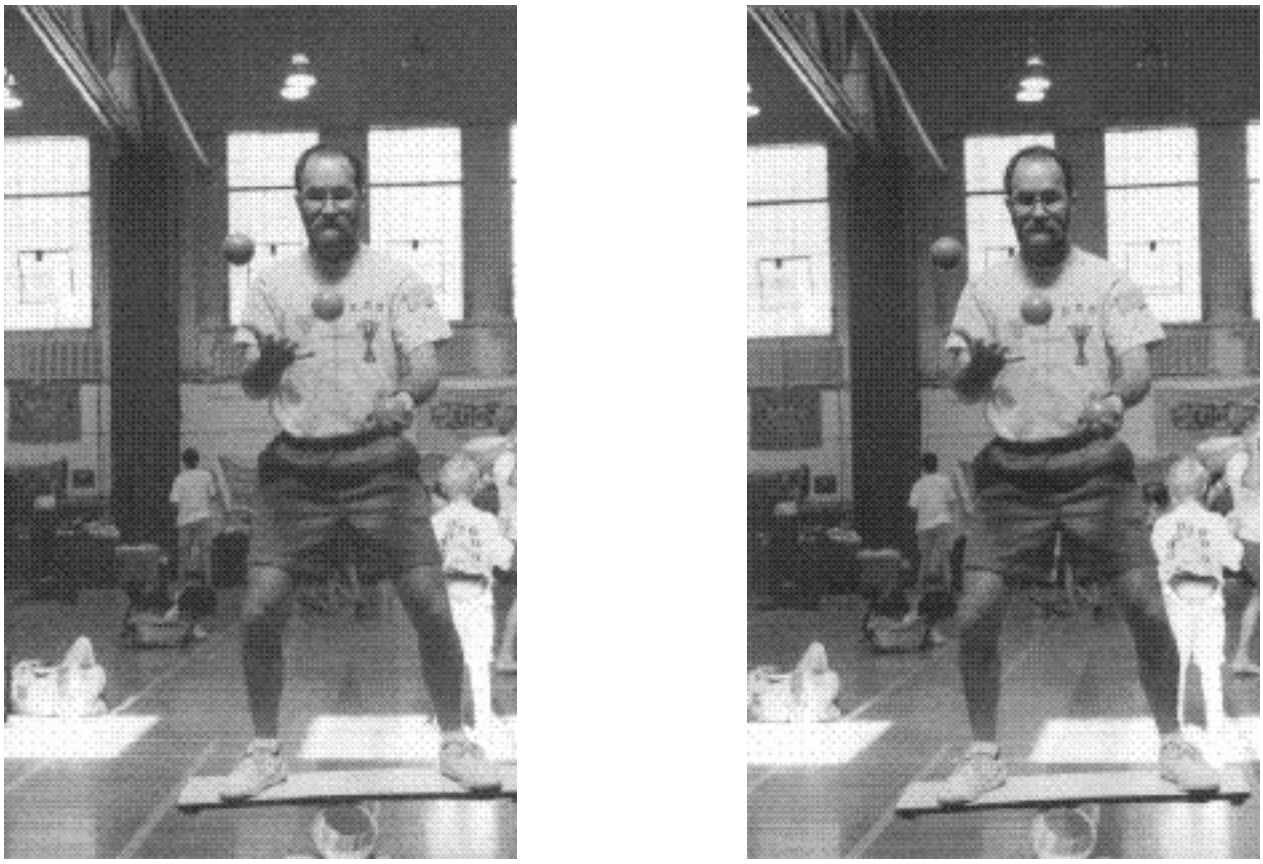

Figure 2: Original Left and Right Views

We expect that in our ongoing work compression depth and synthesis fidelity will both increase substantially.

Topics that we need to address in the context of compression of 3D-stereoscopic imagery include:

- Optimizing implementation of the WorldLine approach.
- Optimizing the left-right alternation sequence of conventionally coded and synthesized views.
- Addressing asymmetric resource issues (consequences of the fact that we can afford more hardware at the coding side than at the decoding side)
- Addressing delay penalties (which are relatively unimportant for unidirectional broadcast, but which are a serious problem for real-time two-way communication and teleoperation)
- Implementing formal performance evaluation using appropriate statistical measures of compression effectiveness.
- Implementing psychophysical performance evaluation using appropriate human factors experimental methods and measures.

Topics we intend to pursue later with a view toward long-term payoffs include:

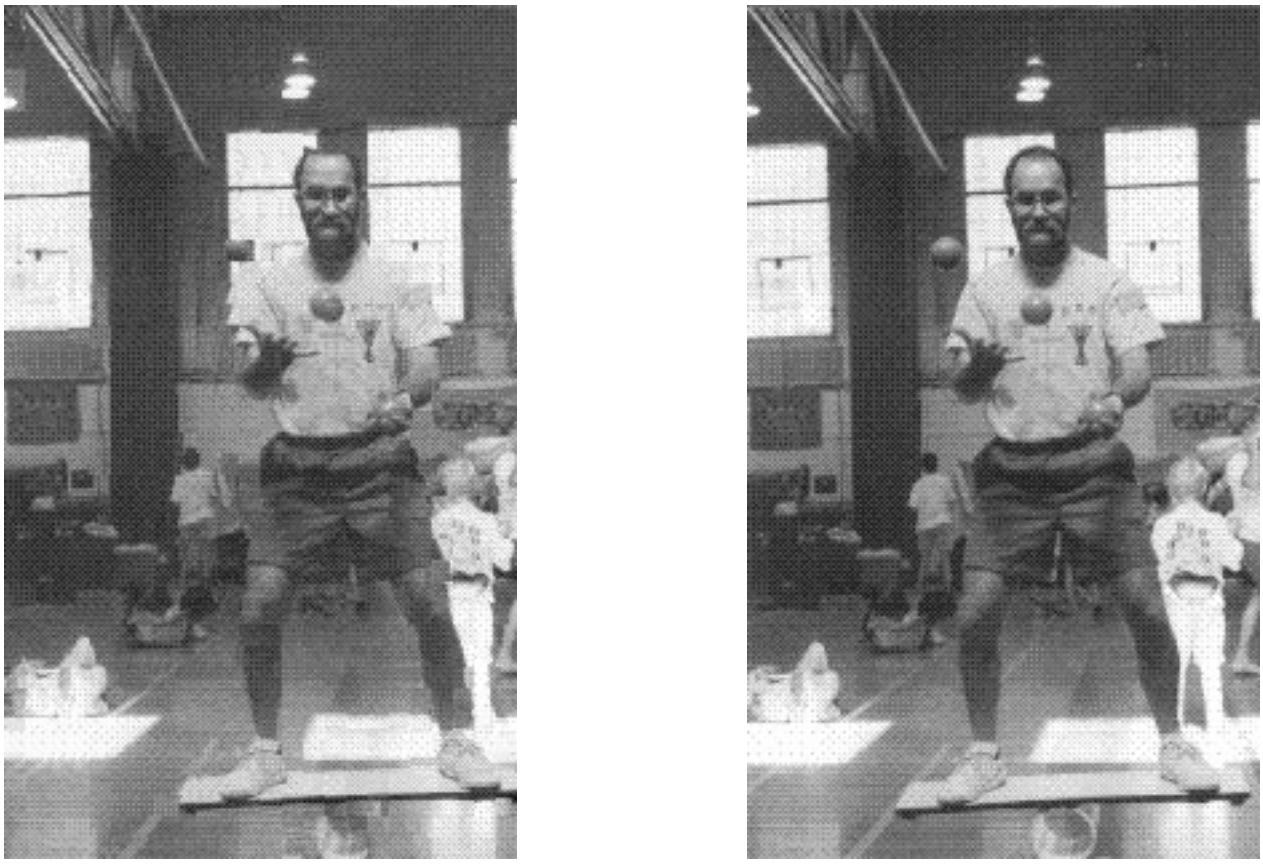

Figure 3: Synthesized Left and Decompressed Right Views

- Using three cameras: compute predictors for left and right views given the middle view, transmit the middle view and the predictors, synthesize 3D-stereoscopic views at the receiver. This approach has several practical advantages including compatibility with flat display systems and ease of adapting the 3D-stereoscopic rendering to the preferences and visual abilities of the viewer.
- Object based methods: apply the methods of machine vision and automated image understanding to augment deeply compressed imagery with semantic information that is used at the receiver to synthesize apparently losslessly transmitted imagery; it should be obvious that this is an extremely ambitious goal.

5. CONCLUSIONS AND PLANS

Because they are highly redundant, binocular 3D-stereoscopic image streams can be encoded in little or no more space (transmitted in little or no more bandwidth) than either component stream.

Single step and hierarchical encoding methods produce psychophysically pleasing imagery.

Future research will address in the short term fine-tuning the architectures and algorithms and understanding their fundamental mathematical and psychophysical efficiencies, and in the long term issues such as multiple camera schemes and object based compression methods.

6. ACKNOWLEDGEMENTS

The ideas discussed in this paper were refined in the course of many discussions with (alphabetically) Tom Ault, Victor Grinberg, Alan Guisewite, Joe Mattis, Jeff McVeigh, Steve Roth, and Scott Safier. This work was funded by ARPA High Definition Systems Grant MDA972-92-J-1010.

7. REFERENCES AND BIBLIOGRAPHY

1. Priyan D. Gunatilake, M. W. Siegel, A. G. Jordan, ‘‘Compression of Stereo Video Streams’’, *International Workshop on HDTV ’93, Ottawa, Ontario, CANADA*, Elsevier, Oct 1993.

2. Sriram Sethuraman, Angel G. Jordan, and M. W. Siegel, ‘‘Multiresolution based hierarchical disparity estimation for stereo image pair compression’’, *Applications of Subbands and Wavelets (Newark, NJ)*, NJIT, Mar 1994.

3. Sriram Sethuraman, M. W. Siegel, and A. G. Jordan, ‘‘A multiresolution framework for stereoscopic image sequence compression’’, *Proceedings of the First IEEE Int’l Conference on Image Processing (Austin, TX), Vol II*, IEEE, Nov 1994, pp. 361-365,

4. I. Dinstein, J. Tselgov, et al., ‘‘On Stereo Image Coding’’, *Ninth International Conference on Pattern Recognition*, IEEE Computer Society, Beer Sheva, Israel, 1988.

5. I. Dinstein, J. Tselgov, et al., ‘‘Compression of Stereo Images and the Evaluation of Its Effects on 3-D Perception’’, *SPIE Applications of Digital Image Processing*, Polytechnic University, Electrical Engineering Dept. and Ben-Gurion University, Behavioral Sciences Dept., Brooklyn, NY and Beer Sheva, Israel, 1989, pp. 522-530.

6. Michael G. Perkins, ‘‘Data Compression of Stereopairs’’, *IEEE Transactions on Communications, Vol. 40, No. 4*, Apr 1992, pp. 684-696.

7. Oliver Rioul and Martin Vetterli, ‘‘Wavelets and Signal Processing’’, *IEEE SP Magazine*, Oct 1991, pp. 16-38.

8. R. Skerjanc and J. Liu, ‘‘A three camera approach for calculating disparity and synthesizing intermediate pictures’’, *Signal Processing: Image Communication 4*, Elsevier, Heinrich-Hertz Institute, Berlin, GERMANY, 1991, pp. 55-64.

9. A. Tamtaoui and C. Labit, ‘‘Schemas de compression de sequence d’images stereoscopiques par compensation de mouvement et disparite’’, *Journees de la Television en Relief*, Elsevier, CCETT, Rennes, FRANCE, 1990, pp. .

10. A. Tamtaoui and C. Labit, ‘‘Constrained disparity and motion estimators for 3DTV image sequence coding’’, *Signal Processing: Image Communication 4*, Elsevier, IRISA/INRIA, Rennes Cedex, FRANCE, 1991, pp. 45-54.

11. A. Tamtaoui and C. Labit, ‘‘Coherent disparity and motion compensation in 3DTV image sequence coding schemes’’, *ICASSP ’91*, Elsevier, IRISA/INRIA, Rennes Cedex, FRANCE, 1991, pp. 2845-2848.

12. K. Metin Uz, Martin Vetterli, and Didier J. LeGall, ‘‘Interpolative Multiresolution Coding of Advanced Television with Compatible Subchannels’’, *IEEE Transactions on Circuits and Systems for Video Technology, Vol. 1, No. 1*, Mar 1991, pp. 86-99.

13. Hiroyuki Yamaguchi, et al., ‘‘Stereoscopic Images Disparity for Predictive Coding’’, *Proceedings ICASSP 1989*, Osaka, JAPAN, 1989, pp. 1976-1979.