Proc. Int. Conf. on Image Process. and its Applications, July 12-15, 1999, Manchester, UK

VIEW RECONSTRUCTION FOR 3-D VIDEO ENTERTAINMENT: ISSUES, ALGORITHMS AND APPLICATIONS

Janusz Konrad

INRS-Télécommunications, Institut National de la Recherche Scientifique, Montréal, Canada

ABSTRACT

Significant advances in stereoscopic imaging in the last decade have lead to viable applications in medicine, teleoperation and, more recently, in entertainment. Although the stereoscopic technology is still mostly analog, the migration to the digital domain is inevitable. Such a migration creates new challenges for stereoscopic video entertainment, but at the same time offers new opportunities. One particular challenge is the reconstruction of intermediate views (between the left and right cameras), that finds various applications. Below, several algorithms aiming at high-quality view reconstruction, recently developed at INRS, are described, and their relative merits are discussed. Since a practical implementation requires low complexity, results of a study of various models and parameters aiming at computational simplicity are reported.

1. INTRODUCTION

Analog stereoscopic imaging has been used in medicine, teleoperation and entertainment for a number of years. The analog format, however, does not lend itself naturally to versatile (different display technologies), flexible (manipulation) and efficient (bandwidth) applications. The ongoing transition to digital storage and transmission of visual information creates new challenges but also opens new opportunities for stereoscopic and 3-D imaging systems. Among the challenges that have clearly emerged are: compression of stereoscopic and multi-view data to meet capacity constraints, and signal processing to enhance viewer comfort and system functionality. The opportunities are multiple, but suffices it to mention stereoscopic image/video delivery over standard communications channels thanks to JPEGS (JPEG-stereo) and MPEG-2's temporal-scalability mode. Another example is Digital Dynamic Depth Inc.'s proprietary digital transmission of stereoscopic TV [1].

Among the major issues in stereoscopic visualization is viewer discomfort. It can be caused by the non-robustness of human perception or by excessive 3-D image cues. Viewers usually experience no fusion problems with properly-acquired stereoscopic images, i.e., with moderate parallax. However, when the parallax is large, although some viewers have no problem fusing the left and right images into a meaningful 3-D cue, others can feel a significant discomfort. This is due to a substantial variation among viewers in terms of their sensitivity to stereo cues [2]. In order to minimize this discomfort, the amount of parallax (or "3D-ness") within each stereo pair may need to be reduced, e.g., by reconstructing intermediate images between the left and right cameras (Fig. 1.a). One can also imagine a scenario where the amount of parallax is too small; to enhance the viewing experience it may need to be increased by stretching the camera baseline[1] (Fig. 1.b). In general, a future 3-D TV or computer screen may need to be equipped with a "3D-ness" knob similar to "brightness" used today.

Another issue, closely related to viewer discomfort, is naturalness of the perception. A stereoscopic camera registers a 3-D scene from two perspectives. If the corresponding images are presented on a stereoscopic display, viewer will perceive a viable 3-D rendition of the scene, but with any lateral motion he/she will see an incorrect perspective due to the independence of camera and viewer motion. This is known as motion-parallax conflict. For example, a still object registered by a stationary stereoscopic camera and presented on a stereoscopic screen seems to be turning in presence of viewer head motion. This problem can be solved by computing proper views in response to viewer movements and presenting them on the screen.

In addition to assuring viewer comfort, view reconstruction finds application in 3-D photography. The idea is to produce a faithful 3-D rendition of a scene by printing, under a lenticular-like sheet, about two dozen closely-spaced views. Observed under a slight rotation, the continuously-changing views give a 3-D impression of a camera pan/translation. Clearly, in order to print the two dozen views from a stereoscopic or multi-view image, view reconstruction is needed.

2. RECONSTRUCTION OF VIEWS

Let's recall first some definitions. *Homologous points* in the left and right images are the projections of one 3-D point (Fig. 2). *Disparity* $\boldsymbol{d}$ is a 2-D vector equal to the difference in coordinates of the homologous points. Disparity is a 1-D vector if cameras have parallel optical axes.

As shown in Fig. 1, view reconstruction is concerned

This work was supported by the the Natural Sciences and Engineering Research Council of Canada under strategic grant STR192788.

[1] Stereoscopic camera baseline is the distance between optical centers of the cameras.

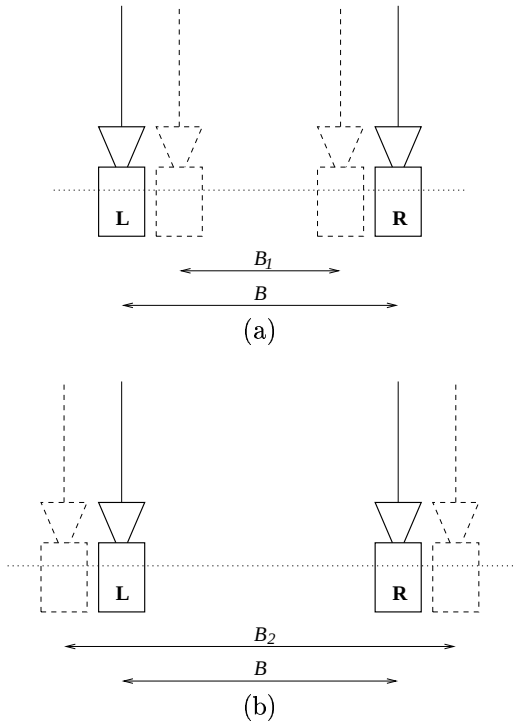


Figure 1: Position of *virtual* cameras (dashed lines) with respect to *true* cameras (solid lines) for parallel camera geometry: (a) baseline reduction; (b) baseline increase.

with computation views that would have been obtained by *virtual* cameras had they been used. This computation must be based on the images acquired by the *true* cameras. Methods developed to date to perform such computation can be classified as either based on 3-D modeling [3] or on 2-D signal processing [4].

Methods based on 3-D modeling attempt to recover 3-D representation of the scene from the left and right images, and then perform a suitable projection onto an arbitrarily-positioned virtual camera. This approach is flexible as it allows arbitrary location of the true and virtual cameras, however it is highly constrained by the complexity of the viewed scene and usually requires calibrated cameras. In practice, 3-D modeling works for simple environments only.

2-D signal processing methods do not attempt to recover a 3-D representation but stay in the realm of 2-D signals. Usually, such methods first establish a correspondence between homologous points *via* disparity estimation. This correspondence could be used for the recovery of a 3-D representation, but since the cameras are usually uncalibrated, disparity-compensated interpolation is used instead to directly reconstruct the virtual camera images. This approach usually works well only for small-baseline stereo cameras (limited disparity). By not relying on a 3-D model of the world this approach is capable of handling quite complex scenes.

For entertainment applications, such as 3-D TV or 3-D photography, images are usually acquired by

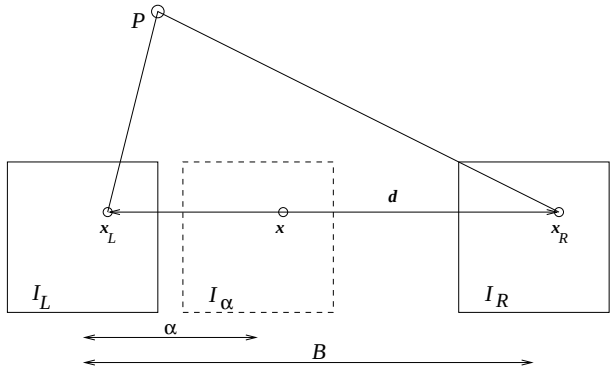


Figure 2: Homologous points $\boldsymbol{x}_L$ and $\boldsymbol{x}_R$ (projections of 3-D point P). $\boldsymbol{x}$ would be another homologous point should a virtual camera be positioned at distance α from the left camera.

closely-spaced cameras (baseline $\approx$ 6.5-7.0cm) with parallel or almost parallel optical axes. Moreover, for baseline adjustment (Fig. 1) and for limited-range tracking of viewer head motion (e.g., involuntary head movements within one sitting position in front of a screen) the virtual cameras are located on the line joining the true cameras. Clearly, under these constraints and the preference of arbitrary scene complexity, signal processing-based methods are more practical.

3. ALGORITHMS TESTED

In the last few years, several algorithms for intermediate view reconstruction have been developed at INRS, primarily for entertainment applications. These algorithms address different issues pertinent to high-quality view reconstruction.

Below, four methods are described. The first three methods share the two-stage approach whereby disparity estimation is followed by disparity-compensated linear interpolation:

$$I_\alpha(\boldsymbol{x}) = \mu_L I_L(\boldsymbol{x} - \alpha\boldsymbol{d}) + \mu_R I_R(\boldsymbol{x} + (1-\alpha)\boldsymbol{d}) \qquad (1)$$

where I_L, I_R and I_α are the left, right and intermediate (reconstructed) image intensities, α is the distance between the left and virtual cameras (with baseline B normalized to 1) and $\boldsymbol{x}$ is a spatial coordinate. The filter coefficients μ_L, μ_R are α-dependent so that for small α's μ_L is close to 1 (I_L contributes more to I_α than I_R), and for α's close to 1 μ_R is close to 1 (I_R contributes more) [5]. The fourth method below combines the disparity estimation and interpolation into one step and is more involved computationally.

Fixed-block reconstruction

This method uses an MPEG-style disparity representation (one vector per 16×16 block of pixels) and exhaustive search block matching for disparity estimation [6]. The linear filtering (1) is used for computation of intermediate pixel values within each reconstructed block.

Quadtree-block reconstruction

This approach uses variable-size blocks to model disparities. The disparity field is first estimated over 16×16 blocks. Then, blocks that cannot be modeled by a single disparity vector are split into four 8×8 blocks; a robust splitting method based on outlier detection was developed for this purpose [6]. The disparities are re-estimated for the 8×8 blocks and the process is repeated. The process stops at 4×4 blocks since further splitting is unreliable due to too few pixels upon which to base the decision. Disparity estimation at each level incorporates an anisotropic disparity smoothness constraint [6]. Based on the estimated disparities $\boldsymbol{d}$, the linear interpolation (1) is performed to reconstruct I_α.

Pixel-based reconstruction

This method is based on the concept of optical flow estimation. Compared to the original method [7], our approach incorporates three important improvements: more flexible intensity matching error (vector pivoting at α instead of at 0), 2-D disparities instead of 1-D, and additional constraint on the amplitude of the vertical disparity component. The method is implemented hierarchically over multiresolution pyramid of images and results in floating-point disparities. The linear interpolation (1) is applied to the computed disparities.

Winner-take-all reconstruction

Unlike the above methods, this approach solves the image reconstruction and the disparity estimation in one step [8]. In fact, the disparity estimates can be considered to be byproducts of the reconstruction process. The image I_α is reconstructed as a tiling of fixed-size blocks coming from various positions (disparity compensation) of *either* the left image I_L *or* the right image I_R; a combination of I_L and I_R is not allowed. This winner-take-all approach is motivated by the fact that linear filtering (combination of I_L and I_R) causes edge busyness and texture blur if disparity estimates are imprecise; only exact disparities allow a precise reconstruction of object boundaries.

4. RESULTS AND DISCUSSION

We have evaluated the four methods on 4 stereoscopic video sequences from CCETT and NHK[2]. The tests were performed in the context of parallax adjustment (Figs. 1 and 3) and continuous look-around (20 images estimated for $0 < \alpha < 1$ and played in time, resulting in a simulated camera pan). An analysis of the reconstructed images allowed us to evaluate the impact of various models and parameters used. The conclusions reached are described below.

[2]The author would like to thank Dr. B. Choquet of the CCETT, Rennes, France, the RACE DISTIMA project of the European Community, and the NHK of Japan for providing INRS with the stereoscopic sequences used in this work.

Disparity scale Out of the three disparity models used, the pixel-based model performed best. The fixed-block method suffered from reconstruction artifacts at object boundaries due to the constant-disparity (depth) model for all pixels in a block; a block at object boundary overlaps two different depths. This can be seen in Fig. 4.a where part of the pole on right and many parts of the tulips are missing. The quadtree-block method, designed to avoid this problem, showed a remarkable improvement at object boundaries; the lower part of the pole and most of the tulips are restored (Fig. 4.b). However, some distortions remain (double tulip stem and small blocks at object boundaries that do not show up in print very well) due to suboptimal splitting and the final 4×4 block structure. The pixel-based model performed best (Fig. 4.c) although some distortions could still be perceived around object boundaries. These distortions are visible for scenes with large object/background depth differences that induce large disparity discontinuities. Since the optical-flow-type disparity estimation is based on the assumption of local disparity smoothness, i.e., neighboring disparity vectors are assumed similar, any discontinuities in the true disparities are smoothed out in the disparity estimates. Thus, incorrect disparity vectors are present in the vicinity of a depth discontinuity (usually object boundary), especially on that side of the discontinuity that is less textured. In consequence, the reconstructed pixels in the band around an object boundary are affected. Subjectively, in the continuous look-around application this is seen as "rubber stretching" of texture on either side of object boundary, e.g., object "pulls" background close to its boundary. To correct this problem, anisotropic disparity smoothing should be used.

Disparity precision Although the disparity estimation algorithms from Section 3 can deliver various precisions of $\boldsymbol{d}$ (typically 1/4-, 1/2- or full-pixel precision for block-based methods and floating-point precision for the pixel-based method), it is not clear whether a higher precision is necessarily needed for the reconstruction step (1). In a series of tests, images reconstructed from the original disparities with sub-pixel precision were compared with those reconstructed from the same disparities but quantized to full pixels. There were virtually no differences between the two cases; only a very close inspection revealed a few pixels slightly changing when images were switched "in-place". This is an important conclusion since disparity precision greatly affects the computational complexity of matching algorithms (first two methods).

Intensity interpolation model To reconstruct image I_α (Fig. 1), the disparity $\boldsymbol{d}$ is divided into two parts: $\boldsymbol{d}_L = \alpha\boldsymbol{d}$ and $\boldsymbol{d}_R = (1-\alpha)\boldsymbol{d}$. The disparity-compensated coordinates $\boldsymbol{x} - \boldsymbol{d}_L$ and $\boldsymbol{x} + \boldsymbol{d}_R$ in I_L and I_R, respectively, are likely to be off the image sampling grid ($0 \le \alpha \le 1$), and thus spatial interpolation is needed. Usually, a separable 2-D operator based on

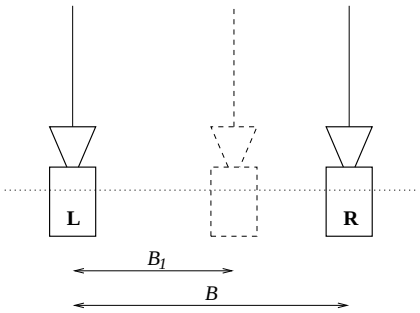


Figure 3: Camera baseline reduction from B to B_1 with only one virtual camera.

small-kernel 1-D interpolators, such as 2-coefficient linear or 4-coefficient cubic, is used. In the tests carried out, there was no visual difference between the interpolation based on the optimal bi-cubic interpolator [9] and a simple bi-linear interpolator. Clearly, the reconstruction step can benefit from the low-complexity of the bi-linear interpolation. Note, however, that a continuous-kernel/continuous-derivative bi-cubic interpolator [9] is still needed for any optical-flow-type disparity estimation.

Luminance and color balancing Due to the difficulty of perfectly matching two cameras (optics, gain adjustment, color calibration, white balancing, etc.), the images of a stereo pair usually differ in luminance and color, and at times even in the amount of blur present. Since luminance and color differences between the left and right images are interpreted as spatial displacements[3], the disparity estimate may be biased if cameras are mismatched. To counteract this, a preprocessing of either I_L or I_R based on scaling and shifting of luminance [10], or luminance and color [6], has been shown to be very effective. Without such preprocessing the disparity estimates are severely biased if mismatch is significant and introduce very objectionable distortions in the reconstructed images.

Two- or one-image parallax adjustment As shown in Fig. 1, to reduce parallax, images from virtual cameras with a reduced baseline B_1 need to be computed. Suppose that the virtual cameras had baseline B_1, but one of them would coincide with one of the true cameras (Fig. 3). Then, only one image would need to be reconstructed. The viewpoint in the scenario from Fig. 3 would be slightly different than that for Fig. 1.a, but this could be acceptable in some applications, such as "3D-ness" adjustment in a 3-D TV set. Note, that the computational complexity would be somewhat reduced, although this reduction would be marginal as the bulk of computations rests with the disparity estimation. Moreover, the perceived quality of the 3-D image may increase since only one image is processed[4].

[3]The underlying assumption in most disparity estimation algorithms is that luminance of homologous points is identical.

[4]It is known that the human visual system tends to suppress information from either left or right image in such a way that the final 3-D percept inherits the better quality of the two images.

We have confirmed this expectation in practice by running a number of experiments with camera baseline reduced from B=1.0 to B_1=0.8, 0.6, 0.4 and 0.2. In each case we observed a clearly improved sharpness and slightly less distortion for a single-image reconstruction as compared with a two-image reconstruction. The improved sharpness is due to the fact that spatial interpolation (resulting in detail loss) is applied to one image only, while the reduction in distortions is caused by their presence in one image instead of two. It is interesting that a better quality can be attained with lower computational complexity, but this scenario may be appropriate for certain applications only. Fig. 5 shows windows from the original images (center row) and from reconstructed images at B_1=0 (top row, no 3-D is perceived) and B_1=2B (bottom row, 3-D is exaggerated). Note, how the tulips are displaced outwards for baseline reduction and inwards for baseline increase.

Blur from disparity errors As expected, the nonlinear filtering of the winner-take-all method reduced edge busyness and texture blur compared to the linear filtering (1). The improvements were clear but mostly subtle; they do not show up in print very well. Due to the tessalation of the reconstructed image into blocks coming from either left or right image, the method results in block flicker when applied to continuous look-around. This may be alleviated to a large extent by additional constraints on the decision labels, but at a significant increase in complexity. The approach seems to be interesting only for the reconstruction of single-α images.

5. CONCLUSIONS

Our experience shows that pixel-based disparity estimation is the most appropriate approach for small-baseline linear view interpolation. However, the the approach is not flawless due to the disparity oversmoothing at object boundaries. To counteract this, one option is to use discontinuity-preserving anisotropic smoothness [11], while another is joint disparity segmentation and estimation. The interpolation step is quite forgiving; bi-linear interpolator and full-pixel disparities are sufficient. It is interesting that a single-image parallax adjustment results in higher-quality 3-D percepts than a two-image adjustment. This is encouraging for some applications, e.g., "3D-ness" or "depthness" knob on a 3-D TV set.

6. REFERENCES

[1] P. Harman, "An architecture for digital 3-D broadcasting," in *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems*, vol. 3639, Jan. 1999.

[2] W. Tam and L. Stelmach, "Stereo depth perception in a sample of young television viewers," in *Int. Workshop on Stereoscopic and 3D Imaging*, (Santorini, Greece), pp. 149–156, Sept. 1995.

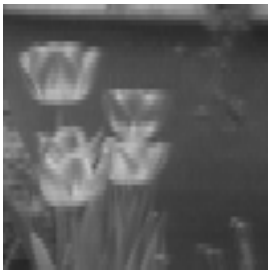
(a) Block-based

(b) Block-quadtree

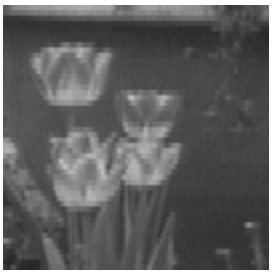
(c) Pixel-based

Figure 4: Image *Flower* reconstructed at α=0.5.

(a) α=0.5



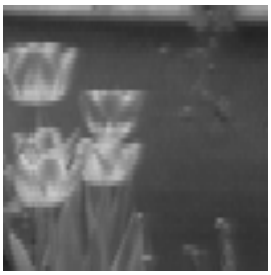
(b) α=0.5



(c) left original

(d) right original

(e) α=-0.5



(f) α=1.5

Figure 5: Parallax adjustment for *Flower* (100×100 window): (a-b) reconstructed images for B_1=0, (c-d) original images, (e-f) reconstructed images for B_1=2B. Top-row images are identical; no depth is perceived.

SPIE Stereoscopic Displays and Virtual Reality Systems, vol. 3295, pp. 53–64, Jan. 1998.

[3] N. Chang and A. Zakhor, "View generation for three-dimensional scenes from video sequences," *IEEE Trans. Image Process.*, vol. 6, pp. 584–598, Apr. 1997.

[4] E. Izquierdo, "Stereo matching for enhanced telepresence in three-dimensional videocommunications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 4, pp. 629–643, 1997.

[5] J. Konrad, "Enhancement of viewer comfort in stereoscopic viewing: parallax adjustment," in *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems*, vol. 3639, Jan. 1999.

[6] A. Mancini and J. Konrad, "Robust quadtree-based disparity estimation for the reconstruction of intermediate stereoscopic images," in *Proc.*

[7] R. March, "Computation of stereo disparity using regularization," *Pattern Recognit. Lett.*, vol. 8, pp. 181–187, Oct. 1988.

[8] A.-R. Mansouri and J. Konrad, "Block-based winner-takes-all reconstruction of intermediate stereoscopic images," in *Proc. SPIE Visual Communications and Image Process.*, vol. 3309, pp. 922–933, Jan. 1998.

[9] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, pp. 1153–1160, Dec. 1981.

[10] R. Franich and R. ter Horst, "Balance compensation for stereoscopic image sequences." ISO/IEC JTC1/SC29/WG11 – MPEG96, Mar. 1996.

[11] L. Robert and R. Deriche, "Dense depth map reconstruction using a multiscale regularization approach which preserves discontinuities," in *Int. Workshop on Stereoscopic and 3D Imaging*, (Santorini, Greece), pp. 32–39, Sept. 1995.