Automatic 2D to 3D footage conversion for 3D glasses-free display systems

Juliang Shao, Julien Flack, Simon Fox, Phil Harman
Dynamic Digital Depth Research

1.1 Introduction

The emergence of a new generation of commercial 3D glasses-free display systems (e.g. [12], [15]) is driving the demand for stereoscopic content. In the lead-up to the development of formal standards [1] the industry is increasingly using depth maps as an ad-hoc standard for generating stereoscopic images [5][10]. Depth maps can be acquired through active range sensors [2], passive analysis of 2D video signals or some combination of active/passive techniques such as stereo analysis [11] or depth-from-focus/defocus [13]. In this demonstration we will describe a passive technique for recovering depth maps from monoscopic image sequences.

The ability to convert existing 2D footage into 3D is essential for supporting 3D displays with both compelling and current content. Manual techniques for generating depth maps, such as outlining objects, are laborious, time consuming and error-prone. There is significant commercial interest in the ability to generate depth maps automatically. Technically the automatic approach underlying this demonstration is composed of three main steps: 1) camera path reconstruction by structure-from-motion, 2) image correspondence and 3) depth propagation.

It should be emphasized that our objectives is to deploy this automatic conversion system in a commercial postproduction environment. It is therefore essential that the system is both robust (failure tolerant) and fast. To this end, our structure-from-motion algorithms were designed to be distributed across a number of processing nodes. Our structure-from-motion algorithms are based on *plausible* Euclidean reconstruction, which does not require self-calibration over all frames and can accommodate varying and unknown intrinsic camera parameters over long sequences (hundreds or even thousands of frames). These aspects are the focus of our demonstration.

To generate dense depth maps using structure-from-motion a standard rectified image-matching step, using polar rectification [9] and pyramid based SSD (sum of square differences) were used. As these techniques are well known in the industry we do not describe them in detail. Matching was used to generate quasi-dense depth maps and a novel depth propagation algorithm was used to generate fully dense depth maps.

1.2 Distributed design in structure-from-motion

For a long image sequence, there are many shots with each representing a scene. Different shots have to be separated into respective groups. A shot is then cut into many overlapped sections through key frames, which are extracted fundamentally via tracking points.

In individual sections, the trifocal tensors [14] can be computed independently. Due to the overlap of two neighboring sections, the merge of such two sections can be processed by a homograph, e.g. [4]. The whole merging process is a hierarchical strategy. For example, there are 4 basic sections, numbered 01, 02, 03 and 04 (0 means the bottom level). Every two sections in order are merged and become sections 11 (from 01-02), 12 (from 02-03) and 13 (from 03-04), respectively. This process is repeated until we reach the top level, section 31, when one large section remains. The individual merging of two sections is independent of the other sections, though the result will be used during merging at higher levels.

Since trifocal tensor calculation and section merging can be calculated independently, it is possible to use distributed processing. As soon as any two neighboring sections' trifocal tensor calculation are complete, higher level processing can commence. The same parallelism applies to section merging. This design enables structure-from-motion calculations on long sequences to be implemented efficiently on a number of processors.

1.3 Plausible Euclidean reconstruction in structure-from-motion

In free environment shots intrinsic camera parameters, such as focal length, skew, aspect ratio, principal point and lens distortions are generally unknown. For non-metric cameras lens distortion (radial & decentering) can be significant. In practice, without ground truth or accurate calibration, it is difficult to recover all of the intrinsic parameters of the camera. Without loss of generality, some of parameters are supposed to be known, e.g. skew = 0, aspect ratio = 0 or

principal point locates at image center, and regardless of lens distortion [6][9]. When these assumptions are inaccurate it is only possible to recover structure over short sequences – the calibration often fails on longer sequences.

Since self-calibration can be conducted accurately over a short sequence with reasonable assumption, we calibrate over very basic sections (i.e. triplets). As these sections are very short this is generally successful. So every basic (bottom level) section is a purely Euclidean reconstruction, with probably varying camera intrinsic parameters. During section merging, the homograph is calculated by using the previously derived projection matrices. The merged projection matrices could be considered to lie in a plausible Euclidean space, since the two sections were calibrated before merging. When many sections are merged together and bundle adjustment for plausible Euclidean matrices is finally carried out, varying camera intrinsic parameters can be reasonably recovered without their assumptions, which is still necessary for quasi-Euclidean reconstruction [3]. From this view of point, our plausible Euclidean reconstruction has advantages over a quasi-Euclidean reconstruction.

1.4 Depth Propagation

In order to synthesise images from virtual viewpoints for display on multiview autostereoscopic 3D displays it is necessary to generate a dense disparity map. Traditionally, structure-from-motion algorithms rely on sparse feature-based techniques. One approach to the generation of dense disparity information is to use the epipolar constraints derived from structure-from-motion analysis to warp two frames into a stereo pair and use traditional stereo matching algorithms to generate a dense disparity map [8].

Stereo matching is notoriously difficult in images containing low texture and occluded regions. The errors introduced during matching cause distortions in the synthesised views and lead to unpleasant artifacts in the resulting stereoscopic images. The human observer is particularly sensitive to sudden changes in disparity over time. In order to address these problems we have developed a unique approach to dense disparity calculation for image sequences. Essentially we use an initial rectified stereo matching step to provide quasi-dense disparity estimation by retaining disparity calculations only in areas of high matching confidence. This approach is similar in nature to Lhuiller and Quan's initial step of their quasi-dense reconstruction algorithm [7].

We use these initial seed points as training samples in a learning algorithm to determine the relationship between an image and its associated disparity map. This approach is motivated by the high correlation between an image and its disparity map. Once trained the learning algorithm is used to propagate the quasi-dense disparity map. This process is repeated at specific key frames within an image sequence. To generate disparity maps for frames in between two key frames the learning algorithms associated with each key frame are combined in a time-weighted relationship. A detailed description of this approach is beyond the scope of this document, but the interested reader may refer to [5].

Although it is easy to imagine situations in which this propagation technique will generate inaccuracies this approach has a number of characteristics that make it a fit-for-purpose approach in the context of stereoscopic 3D display devices for applications in entertainment and advertising. In particular, we are implicitly introducing a spatial and temporal smoothness constraint on the depth map sequence by essentially using the original video image as a regularization function. In a similar vein to the goal of perceptually lossless image compression, we may claim that the error characteristics of the depth maps are designed to produce stereoscopic images with minimal apparent defects to the human observer.

1.5 Demonstration contents

To illustrate the effectiveness of plausible Euclidean reconstruction we will demonstrate reconstruction on a variety of image sequences including difficult camera motions such as zooming, focusing and forward/reverse motion. The resulting synthesized stereoscopic images will be shown on a glasses-free 3D display demonstrating how this approach provides a practical and automatic means of converting content from 2D to 3D.

1.6 Acknowledgment

We are grateful for Richard Hartley of Australian National University, who provided invaluable consultation on the relevant project. Time-Slice Films Ltd is also appreciated to provide an included footage for the research.

References

[1] 3D Consortium, http://www.3dc.gr.jp/

[2] 3DV Systems, http://www.3dvsystems.com/
[3] P. Beardsley, A. Zisserman, and D. Murray, “Sequential Updating of Projective and Affine Structure from Motion”, In International Journal of Computer Vision (23) 3: 235–259, 1997.
[4] S. Gibson, J. Cook, T. L. J. Howard, R. J. Hubbold, and D. Oram., “Accurate camera calibration for off-line, video-based augmented reality”, In IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2002), Sep 2002. Darmstadt, Germany.
[5] P. Harman, J. Flack, S. Fox and M. Dowley, “Rapid 2D to 3D Conversion”, in Stereoscopic Displays and Virtual Reality Systems IX, Andrew J. Woods, John O. Merritt, Stephen A. Benton, Mark T Bolas, Editors, Proceedings SPIE Vol. 4660, pp 78-86.
[6] R. Hartley and A. Zisserman, “Multiple View Geometry in Computer Vision”, Cambridge University Press, 2000.
[7] M. Lhuillier and L. Quan, “Quasi-Dense Reconstruction from Image Sequence”, 7th European Conference on Computer Vision, Copenhagen, Denmark, May 2002, pp 125-139.
[8] M. Pollefeys “3D Modeling from Images”, ECCV 200 tutorial
[9] M. Pollefeys, R. Koch, and L. Van Gool, “Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters”, In Proc. Int. Conf. on Computer Vision, 1998.
[10] A. Redert, M. Op de Beeck, C. Fehn, W. Ijsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton and P. Surman, “ATTEST: Advanced Three-dimensional Television System Technologies”, Proceedings of the First International Symposium on 3D Data Processing Visualization and Transmission, 2002.
[11] J. Shao, “Generation of Temporally Consistent Multiple Virtual Camera Views from Stereoscopic Image Sequences”, International Journal of Computer Vision. 47(1/2/3): 171-180, 2002.
[12] StereoGraphics SynthaGram, http://www.stereographics.com/products/synthagram/synthagram.htm, Retrieved on 6 August 2003.
[13] M. Subbarao and G. Surya, "Depth from Defocus: A Spatial Domain Approach", Technical Report No. 92.12.03, Computer Vision Laboratory, Electrical Engineering Department, SUNY, Stony Brook, NY.
[14] P. H. S. Torr and A. Zisserman, “Robust parameterization and computation of the trifocal tensor”, Image and Vision Computing, 15: 591--605, 1997.
[15] X3D Technologies GMBH, http://www.4d-vision.com/cgi-bin/change_l.pl?lang=en&count=1, Retrieved on 6 August 2003.