

A Framework of Individually-Focused Teleconferencing (IFT) via an Efficient 3D Reprojection Technique

Qi Li^{†*}, Chris Brown[‡], Chandra Kambhampettu[‡], Tao Li[‡] and Shenghuo Zhu[‡]
[†] Computer Sci. Dept, Univ. of Delaware, Newark, DE, 19716,
{qili,chandra}@cis.udel.edu
[‡] Computer Sci. Dept, Univ. of Rochester, Rochester, NY, 14627,
{brown,taoli,zsh}@cs.rochester.edu

ABSTRACT

In this paper, we propose a framework on individually-focused teleconferencing (IFT) which is supported by an efficient 3D reprojection technique. The goal of designing IFT is to provide sufficient 3D pose information to each teleconference participant in order to establish lively communication in a teleconference on low-bandwidth internet. A novel IFT deployment and an efficient 3D reprojection technique are two major contributions of this paper. Our 3D reprojection technique uses a mirror reference view in three-view epipolar geometry. It overcomes the inefficiency in previous techniques and provides visually good recovery of pose information of teleconference participants even though it is theoretically an approximation scheme.

Keywords

Individually-focused teleconferencing (IFT), three-view epipolar geometry, 3D reprojection, mirror reference view/image, fundamental matrix

1. INTRODUCTION

There are two challenges in designing a video teleconferencing system. One is the transmission of huge amounts of video data. A user who blasts a high-bandwidth video signal, say greater than 128 Kbps, can cause severe and widespread network problems ([14]). The other challenge is visual realism. Most commercial video teleconferencing systems ([19, 5]) deal only with two-point communication, where individual pose is not a big issue because there is no confusion as to who they are talking with. Multiple-

*Tel: 302-831-0556. Fax: 302-831-8458. This work was initiated when the author did summer research internship in IBM Almaden Research Center in 2001. It was developed at University of Rochester and finally completed at University of Delaware.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC 2003, Melbourne, Florida, USA

point teleconferencing is trickier. It seems natural to apply augmented reality and computer vision techniques (such as pose tracking, facial tracking, head-mounted display and so on) to this kind of teleconferencing ([16, 9, 8, 6]). In what follows, teleconferencing is always assumed to be multiple-point video teleconferencing.

In paper [18], Raskar *et al* present a design of the future office by using the common office hardware: cameras, LCD projectors and desktops. More precisely, they use "cameras and projectors that can be operated in either a capture or display mode. When in capture mode, the projectors and cameras can be used together to obtain per pixel depth and reflectance information from designated display surfaces. When in display mode, the image-based models may be used to render and then project geometrically and photometrically correct images onto (potentially irregular) display surfaces". This work [18] has many good ideas in introducing virtual reality to a video teleconferencing, especially on the deployment of projectors and cameras. However, some aspects in their system can be simplified to meet virtual video teleconferencing for our current life.

According to a common complaint on current commercial teleconferencing or telepresentation systems, it is hard for a participant to tell whom he/she is talking with. We simplify the system in [18] and concentrate on preserving individual pose rather than other factors such as depth and spatially immersive displaying. Being a teleconference participant, one can accept display on a flat board as well as on a 2D film or TV screen, and also accept some confusion on how far the speaker is away from him/her. But a participant cannot tolerate a scenario where all other participants look like they are all talking with him/her. Teleconferencing which can preserve individual pose is characterized as *individually-focused teleconferencing* (IFT). Another simplification is on the image-based modeling process. The image-based modeling in [18] belongs to 3D reconstruction category and it is a model-based reconstruction scheme, i.e., a concrete 3D object is built before a teleconference using a camera calibration technique. The camera calibration is known to be a time-consuming process. There are other methods that do not involve camera calibration such as: triangulation [10] depth map [3, 15, 20], and parameterized mesh model [17, 13]. Image-based 3D modeling is usually accompanied with pose tracking technique in the context of virtual teleconferencing [18, 16, 13]. Pose tracking is another time-consuming process and may involve severe accumulation error.

The technique we are going to use in IFT is called *3D reprojection* which provides a model-free and tracking-free scheme. Once the parameters for a view adjustment are pre-computed, a teleconferencing system always reprojects a real frame captured by a stationary camera from the camera's view point to a virtual view point by the pre-computed parameters. Three-view epipolar geometry is the underlying theory for 3D reprojection. Many works have been done on three-view epipolar geometry [21, 11, 7]. Most of these works only considered reprojection of feature/corner points in an image rather than dense reprojection. Avidan and Shashua [1] studied dense reprojection via an optical flow between a source view and a reference view where the baseline between the source and reference view is assumed to be tiny. Theoretically, optical flow is the optimal method for dense reprojection. However, computing optical flow computation itself is an ill-posed problem which is very sensitive to image noise and is also a time-consuming process.

The goal of this paper is to propose a framework of teleconferencing which has low network communication cost by reducing the number of cameras for each participant while still preserves individual pose of each participant (so that he/she can easily tell whom he/she is talking with) via an efficient 3D reprojection technique. The rest of this paper is organized as follows: Section 2 proposes a framework of individually-focused teleconferencing (IFT). Section 3 presents an efficient dense 3D reprojection technique by introducing a mirror reference view in three-view epipolar geometry. The experiments in Section 4 show the accuracy and efficiency of our 3D reprojection technique. Finally, Section 5 gives our conclusions and future work.

2. EFFICIENT IFT

A straightforward but inefficient solution for IFT is to arrange cameras in each participant's office spatially corresponding to the other participants. For n participants this requires $n - 1$ cameras per participants.

We aim to achieve IFT with fewer than $n - 1$ cameras in each participant's office by simulating the geometry of the teleconference, using reprojection to simulate camera angles. Fig. 1 (a) shows the network architecture that we assume for teleconferences and (b) shows the associated design of the virtual conference table. In the scenario described in Fig. 1 (a), there are 6 participants. A and B are a group sharing LAN1; C is an individual using LAN2; D, E are a group sharing LAN3; F is an individual using LAN4; We use the principle that **participants in the same LAN are seated close to each other at the virtual table** — yielding a geometry like that of Fig. 1 (b).

Since A and B “sit” close to each other, there should only be slight difference between the views of F being watched by A and B. Therefore, it is reasonable to send only one view to the group of participants A and B. We call this view a “master view” for the group. An IFT system has the responsibility of generating two individual views from the master view after it reaches LAN1 so that A and B each see an individually-focused view. To capture one master view, we need one real camera. So in the scenario in Fig. 1, there are three cameras in F's office to capture all the master views of F.

But the number of cameras in A's office is not exactly the same as F's because A is not the unique participant who is using LAN1. Besides providing three master views to

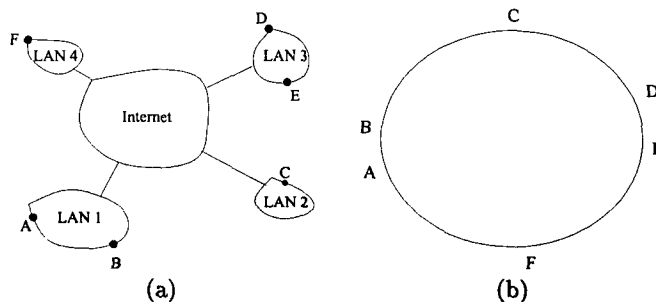


Figure 1: (a) Communication architecture of a teleconference; (b) A virtual conference table for nine participants in four LANs; the same LAN implies physical proximity.

three LANs, the IFT system still needs to provide a master view to the group member B. If the number of its group members increases, the IFT system will assign the same increasing number of master views for all group members. In other words, assuming a low traffic in a given LAN, the IFT system of some participants do not involve any reprojection/synthesis process of views for the participants in the same LAN. This is because the bandwidth of a LAN is high enough to tolerate many images.

The deployment of IFT system in the site (office) of a participant is showed in Fig. 2. The major equipment of an IFT system include a projector, a laptop/desktop and multiple video cameras. The wall of a participant's office is the screen for an IFT system.

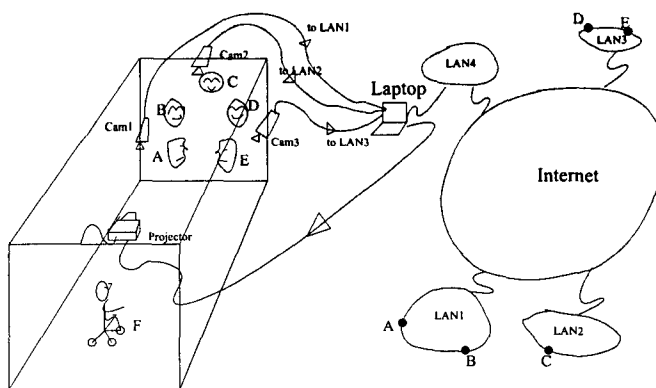


Figure 2: A scenario of deployment of an efficient IFT system including a projector, a laptop/desktop and three video cameras. The image captured by i -th camera is only sent to i -th LAN. The images of other participants are displayed in the wall of participant F's office after an associated pose adjustment by the 3D reprojection technique.

It is worth to point out that the one-to-many camera-participant property in an efficient IFT exactly matches the one-to-many package delivery scheme of multicast addressing. Actually, it is multicast addressing that provides network support for an IFT system. However, to concentrate on the geometry framework of an efficient IFT system (which is the kernel and attractive problem in designing an efficient

- Compute the fundamental matrix F from a source view to a target view
- Compute the fundamental matrix F' from the mirror view of the source view to the target view.

- Online part

- For any image point² $p = (y, x, 1)$ in the source view, and $p' = (y, cols - x, 1)$ in the mirror image, where cols is the number of columns of an image. Compute the intersection of two epilines Fp and $F'p'$ (i.e., their cross product) and render this intersection by the color of the point p .

This is a forward rendering algorithm with which some image points in target view may not get rendering. In practice, we usually use the backward rendering. Since their ideas are similar (even though backward rendering seems to be a little tricky), we present the backward rendering algorithm in Appendix.

In summary, we build fundamental matrices before a teleconference, and apply these parameters to adjust real views of participants during teleconference. All computations only involve the basic algebraic operations and the computation complexity is of the image size, therefore, it is a simple and efficient 3D reprojection technique.

4. EXPERIMENTS ON THE 3D REPROJECTION SCHEME

In this section, we will first test the accuracy of our 3D reprojection technique on model images of Santa's face captured by Panasonic PVDV 400 camcorder by evaluating the r.m.s. error, and its efficiency by the computation time. Then we apply the parameters (i.e., two fundamental matrices) acquired from Santa images to a real face image captured by a web camera and get an adjusted virtual view. The image size is 320×240 . Our computer is IBM laptop, T21 CPU 800, MEM 128M.

Fig. 5 (a) and (b) show a source view and a target view of Santa face respectively. The physical distance between the center of Santa's face and the optical center of our camera is around 30cm, and the depth range of the face is around 5cm. Fig.5 (c) shows the synthesized Santa by our 3D reprojection technique. We can see that visually the pose of the synthesized face looks very close to the target image. The r.m.s reprojection error is around 2.0 pixels. The reprojection speed is around 0.5 second to synthesize one image. A reference result is from [2] where 5 seconds is needed to synthesize an image in the context of SGI Indy computer and 260×480 image size. There are two ways to further improve the reprojection speed to meet the real-time displaying requirement. The first way is to use subsampling, and the second one is to embed this algorithm into a circuit.

With parameters computed from the source and target images of Santa face, we now apply the online algorithm of the 3D reprojection algorithm to a view of a real face which is captured by a web camera. A unique requirement on the deployment of the web camera is that the cross angle between the orientations of the web camera and the real face should be roughly equal to the one between the camcorder

²We use the homogeneous coordinates of image points in the context of epipolar geometry computation.

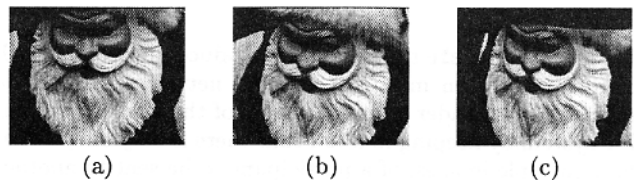


Figure 5: Synthesized view of Santa face is very similar to the target view of Santa face. (a) A source view; (b) A target view; (c) A synthesized view.

and Santa face (We can easily achieve this requirement by manually deploying the web camera). Fig. 6 shows a real view of a student in University of Delaware and its adjusted view. We observe that this view adjustment matches pretty well with the view change from the source view to the target view of Santa face.



Figure 6: View adjustments of a real face by the epipolar geometry computed from the images of Santa's face. (a) A real view captured by a web camera; (b) An adjusted view;

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a theoretical framework of our IFT system supported by an efficient 3D reprojection technique. An IFT system contains many important factors such as the deployment of equipment, the accuracy and efficiency of 3D reprojection algorithm, and network addressing. We focus on the first two factors and design an IFT system deployment scheme, and propose an efficient 3D reprojection algorithm which overcomes the limitation of traditional 3D reconstruction schemes and some other current reprojection techniques by introducing a mirror image (of a source/target image in the forward/backward algorithm) as a reference view in a three-view epipolar geometry.

Since our 3D reprojection technique is based on basic algebraic operations without having to invoke any other computer vision or image processing routines, it is extremely viable to design a circuit that stores geometry information represented by a family of fundamental matrices. The performance of an IFT system thus depends on the accuracy and completeness of the geometry information obtained via the fundamental matrices. Our future work involves computing "arbitrary" view adjustment for a more practical IFT system on a multicast enabled networks.

6. ACKNOWLEDGMENTS

Sincere thanks to the User Group, IBM Almaden Research Center (ARC) that supported the 2001 summer research internship to the first author, and thanks to Dr. Tanveer S. Mahmood and all other researchers in User Group, ARC for the discussions.

7. REFERENCES

- [1] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *Proc. IEEE CVPR*, pages 1034–1040, 1997.
- [2] S. Avidan and A. Shashua. Novel view synthesis by cascading trilinear tensors. *IEEE Transactions on Visualization and Computer Graphics*, 4(4):293–306, October/December 1998.
- [3] S. Chen and L. Williams. View interpolation for image synthesis. In *Computer Graphics (SIGGRAPH'93)*, pages 279–288, 1993.
- [4] S. Deering. Host extensions for ip multicast. In *Request For Comment 1112*. <http://asg.web.cmu.edu/rfc/rfc1112.html>, Aug, 1989.
- [5] J. Duran and C. Sauer. Mainstream videoconferencing: A developer's guide to distance multimedia. Addison-Wesley, 1997.
- [6] Kishino F. Virtual space teleconferencing system - real time detection and reproduction of human images. In *Proc. Imagina 94*, pages 109–118, 1994.
- [7] Andrew W. Fitzgibbon and Andrew Zisserman. Automatic camera recovery for closed or open image sequences. In *ECCV (1)*, pages 311–326, 1998.
- [8] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, H. Farid, and T. Kanade. Virtual space teleconferencing using a sea of cameras. In *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, pages 161–167, 1994.
- [9] C. Greenhalgh and S. Benford. Massive, a collaborative virtual environment for teleconferencing. *ACM Transactions on Computer, Human Interaction*, 2(3):239–261, 1995.
- [10] R. Hartley and P. Sturm. Triangulation. In *American Image Understanding*, pages 957–966, 1994.
- [11] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2001.
- [12] R.I. Hartley. In defense of the eight-point algorithm. *IEEE Trans. on PAMI*, 19(6):580–593, 1997.
- [13] Z. Liu, Z. Zhang, C. Jacobs, and M. Cohen. Rapid modeling of animated faces from video. Technical Report MSR-TR-2000-11, Microsoft Cooperation, Feb 2000.
- [14] M.R. Macedonia and D.P. Brutzman. Mbone provides audio and video across the internet. In *IEEE computer magazine*, pages 30–36, April 1994.
- [15] L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. In *Computer Graphics (SIGGRAPH'95)*, pages 39–46, 1995.
- [16] J. Ohya, Y. Kitamura, H. Takemura, F. Kishino, and N. Terashima. Real-time reproduction of 3d human images in virtual space teleconferencing. In *Proceedings of the IEEE Virtual Reality Annual International Symposium*, pages 408–414, 1993.
- [17] F.I. Parke. A parametric model of human faces. In *PhD thesis, University of Utah*, 1974.
- [18] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The office of future: A unified approach to image-based modelling and spatially immersive displays. In *ACM SIGGRAPH*, pages 179–188, 1999.
- [19] E. Rosen. Personal videoconferencing. Manning, 1996.
- [20] R. Szeliski and H.Y. Shum. Creating full view panoramic image mosaics and environment maps. *Computer Graphics*, 31(Annual conference series):251–258, 1997.
- [21] P. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.

APPENDIX

A. 3D REPROJECTION IN BACKWARD RENDERING SCHEME

Fig. 7 shows a new design of three-view epipolar geometry where the opposite view of a target view (rather than source view) is introduced. And now we consider the intersection of epilines in the source image plane rather than the target image plane.

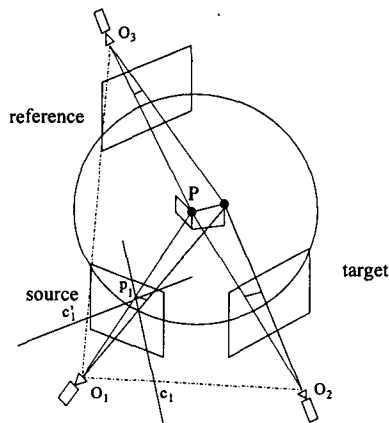


Figure 7: A three-view epipolar geometry for backward rendering.

Backward rendering scheme is a little tricky since it starts from a target view which seems to be unknown for us. But note that each reprojection algorithm is always divided into an offline part and online part, and the offline part is to acquire fundamental matrices from model images whose target views are completely known to us. So a 3D reprojection in backward rendering scheme can be summarized as follows:

- Offline part
 - Compute the fundamental matrix F from a target view to a source view
 - Compute the fundamental matrix F' from the mirror view of the target view to the source view.
- Online part
 - For any image point $p = (y, x, 1)$ in a target view, assume $p' = (y, cols - x, 1)$. Compute the intersection of two epilines Fp and $F'p'$ and render point p by the color of the intersection of two epilines.