EG UK Theory and Practice of Computer Graphics (2011)
Hamish Carr, Ian Grimstead (Editors)

Advantages of 3D Extraction and Spatial Awareness within a Videoconferencing Environment

M. M. Rodriguez-Frias †[1], T. Morris[1], M. Turner[2] and A. Rowley[2]

[1] School of Computer Science, The University of Manchester, UK
[2] Research Computing Services, The University of Manchester, UK

Abstract

This work-in-progress paper describes some of the issues, and lists proposed use-case scenarios, for the development of a real-time multi-participant 3D enhanced videoconferencing system. We are interested in creating a framework capable of reconstructing an approximate 3D model of the physical environment, from a collection of images taken via the same system from probably unknown camera viewpoints. This reconstruction framework creates a partial 3D world that is embedded back within the videoconferencing environment and transmitted to all participants. We hypothesise that detailed 3D positional information combined with this augmented 3D world information of remote sites, can be useful to participants over and above the usual audio and video streams; it is believed that 3D reconstruction can be a rich tool to enable analysis and spatial awareness, moreover facilitating interactions with participants at remote sites.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—I.3.4 [Computer Graphics]: Graphics Utilities—I.3.6 [Computer Graphics]: Methodology and Techniques—

1. Introduction

In recent years, there has been an ever increasing interest and development in systems that allow multiple participation through audio-video collaboration around the world. These videoconferencing software suites often include collaborative environments and simple ones are becoming pervasive within many researchers' or workers' lives. All these systems try to replace and possibly even improve upon the ability to communicate freely mimicking a level of presence that a face-to-face or group meeting in a physical location would provide. Advantages provided by these systems include: decreasing cost, saving time, minimizing travel while still facilitating a similar communication experience. The disadvantages have also been documented, as these systems often have a reduced level of co-presence and do not have the similar immersive experience as a face-to-face encounter has.

For this work in progress paper we specify an advanced collaborative system as one that can be used by multiple users, usually from remote locations around the world, working together and interacting closely with each other, sharing information and exchanging requests. As a minimum we consider each physical room node to have remotely controllable cameras; at least two, with a minimum of pan, tilt and zoom facilities. This enables 3D computer graphic information to be captured that is the focus for this research.

Spatial awareness is the well thought-out organized awareness of the objects in the space around oneself, and also an awareness of one's body position in this space. Spatial awareness requires a participant to have a model of the three dimensional environment and be able to have a shared experience of this. We hypothesise that increasing spatial awareness within a videoconference environment will help to reproduce extra physical presence cues to recreate the feeling of co-presence among the participants, regardless of their actual location; creating a strong sense of immersion. Immersion in conferencing applications implies that the users can rely on natural sight having the sensation of being in the place depicted by the system and the sensation that what is happening is really happening.

Having more than one controllable camera allows for the

† CONACyT

DOI: 10.2312/LocalChapterEvents/TPCG/TPCG11/097-100

opportunity for stereoscopic features to be extracted from a physical room node and this can provide depth information of the scene creating an increased level of spatial awareness for a remote participant. The objective of this work is to integrate this information from pairs of camera views creating a pseudo three-dimensional model of the scene that can be transmitted to the other physical room locations via embedding the data in the videoconferencing data streams. There are some current technologies implementing and transmitting 3D information which include the "Pseudo-3D Video Conferencing with a Generic Webcam" [HH08], an accessible tool as it only requires a single webcam; "The Coliseum technology" [BTS*02], unfortunately requiring 5 cameras in each user point; "Achieving Eye Contact in a One-to-Many 3D Video Teleconferencing System" [ALF*09] an innovative One-to-Many method of communication, and the "Kinected conference: augmenting video imaging with calibrated depth and audio" [DYIR11], that enhances communication beyond 2D video, exploring spatial calibration depth.

Despite the fact that the tools mentioned above are usable and have advantages, we propose to enhance the spatial awareness with an innovative and robust system, able to acquire information, encode, transmit, and reconstruct 3D scenes in real-time in a multi-participant environment, over a low performance network. This will use commodity type resources without specialist equipment. The components of the project can be summarized in the following four stages; data acquisition, data encoding/decoding, transmission, and processing. Figure 1 shows a system overview diagram, using the collaborative system based on the Access Grid Toolkit (AGTkit) [CDO*00] which acts as a testbed for the work and is described in the next section. The stages including calibration, are then described in sections 3-5, before a set of use case scenarios are presented. These case scenarios each show how a participant can gain an enhanced experience through the use of the partial and full 3D computer graphical representations of the remote physical room nodes.

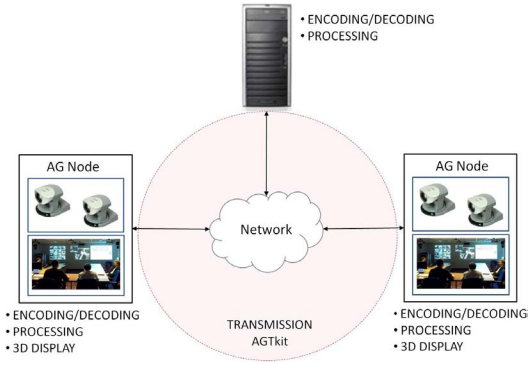

Figure 1: *System overview diagram highlighting the distributed nature of the AGTkit; in that additional processing modules can be at any location, for example at a physical room based node machine, or attached to the network on a remote site.*

2. Advanced Videoconferencing System

The AGTkit is an open source software environment, designed to support wide-area, real-time, computer-mediated communications. As an advanced videoconferencing software application it uses modular audio and video tools to allow people in different locations worldwide to collaborate within virtual meeting rooms. Within these virtual rooms, participants can see and speak to each other in real-time, use on-line chat and share their desktop. Also users can upload and share data sets and remotely run a set of bespoke shared applications. The toolkit is flexible and allows other applications and services to be built on top of the environment, extending the modular and grid based architecture.

Having multiple cameras, multiple projection screens, and echo-cancelling microphones and loudspeakers, at each physical room node, as well as special purpose computer servers and networking, makes every person in a physical room fully able to engage in a session. Using multi-cast communication the Access Grid (www.accessgrid.org) can support very large-scale distributed meetings, collaborative work sessions, seminars, lectures, tutorials, and training requiring only a linear scale in bandwidth cost.

The main reason why this technology was selected as a suitable testbed is because it is scalable to the available infrastructure; from a desktop computer to a large meeting room. The number of participants that can be invited to a session is constrained only by network bandwidth, decoding and encoding compute time and the size of the screen to be used for display of the resulting video feeds. A subsidiary reason is that the system treats communication as simple tagged data streams, so any number of audio, video or desktops screens can be transmitted. Therefore adding extra communication streams, for example one for 3D geometry and another for remote camera orientation and control, is possible.

3. Data Acquisition

Whilst a large amount of information can be gleaned using a single camera, three dimensional reconstructions of a scene may be possible using data captured from multiple cameras. In our current implementation we gather the visual information using two pan-tilt-zoom Canon VC-C4 cameras, but there is, in principle, no limit to the number of cameras that could be used.

The relative position and orientation of the cameras (the extrinsic parameters) and their intrinsic parameters (focal length, aspect ratio, etc) is required for the reconstruction. The intrinsic parameters are obtained through calibration. (A metric reconstruction requires measurements in length units, a reconstruction correct to scale can be generated without absolute measurements, and is sufficient for our purposes.) We calibrated the cameras using the calibration toolbox for Matlab [Bou01], using images of a planar checkerboard.

Once the cameras are calibrated, we acquire video data from each camera, and store images at predefined times.

4. Encoding, Transmission and Decoding

The information is preferably transmitted as UDP (User Datagram Protocol) packets, which reduces the amount of data to be transmitted thereby reducing overall latency. UDP applications must generally be willing to accept some loss, errors or duplication in packets as data can be lost, delayed, or arrive out of order. Currently data loss is managed with buffers, delaying transmission fractionally, but additional applications are planned to further minimise and manage data loss. The frames will be transmitted over UDP to the receiver, where every frame will be decoded and processed to render into the scene as appropriate.

5. Processing

Our system uses computer vision and computer graphics algorithms to reconstruct the scene incrementally. The stages of our image-based reconstruction are:

- Feature detection.
- Feature matching between pair of images.
- Scene geometry and 3D points calculation.
- Mesh and texture generation.

Whilst any feature detector could be used to locate points for the reconstruction, the Scale Invariant Feature Transform (SIFT) [Low04] is used as it is robust to changes of scale and rotation. Figure 2 shows feature detection and feature matching between a pair of images.

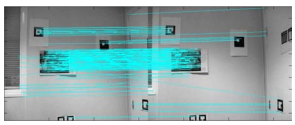

Figure 2: *Feature matching between a pair of images.*

The results shown in this paper are computed using Bundler [SSS06]. Bundler is a structure-from-motion system for unordered image collections, it is used as an initial approach but it is not a real-time system. The algorithms Clustering Views for Multi-view Stereo (CMVS) and Patch-based Multi-view Stereo Software (PMVS -version 2) [FP08] are planned to be used to create denser points.

6. Visualization and 3D Geometric Case Studies

With the aim to provide understanding of the 3D remote environment in order to heighten co-location and increase the level of presence, we propose to embed in the AGTkit a module capable of controlling image capture, of processing the images transmitted and creating geometrical data that can be transmitted itself as an extra data stream. This is embedded at the consumer service end of the AGTkit allowing for novel modes to be displayed and four of these are described in the following use-case scenarios.

Figure 3: *An embedded videoconferencing stream within an initially empty panorama of the physical room node.*

6.1. Embedded Video in Panorama Mode

The first scenario controls the remote camera to take a series of still 2D images that are stitched together to create a panorama image [BL07]. Each remote camera has then a single pre-defined panorama associated with it. To aid a remote participant's comprehension of the room, the real-time video stream is superimposed onto the panorama at the appropriate place. Therefore, the panoramic image is used as a background, enhancing the knowledge of the remote site environment, giving context and enhanced co-location, as shown in figure 3. The remote camera now needs to transmit the pan, tilt and zoom values alongside their video stream, which is done over a separate data channel.

Despite the potential increase of spatial awareness, this basic approach has its limitations. As shown in the figure, one of the main downsides of this scenario is the limited size of the video stream, and items, for example in this figure the hands, may become cropped and disjoint.

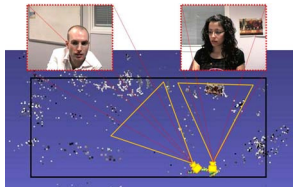

Figure 4: *Reconstruction with a 3D point cloud and viewing frustrums displayed.*

6.2. 3D point-cloud remote site projected views

The second scenario aims to create a 3D geometrical spatial projection of the remote environment. As we have at least two cameras at each site, we can capture all possible views from each camera, initially independently. 3D locations of features can be computed from the sets of different 2D views. This results in a 3D point cloud of features, and simultaneously extracts the relative locations of the cameras. The 3D geometrical point cloud and the fixed camera (x, y, z) position values can then be transmitted as a new data channel to all sites.

Figure 4 shows the initial result of reconstruction, which were obtained using 100 images from a pair of remote cameras. MeshLab [CCR08] was used to visualize the reconstructed 3D scene, and superimposed is a description of the

current camera frustum. Having the 2D video stream viewable alongside the 3D point cloud with the animated camera frustum superimposed, allows for the next level of co-location for remote participants in that they can understand the spatial location of the remote audience.

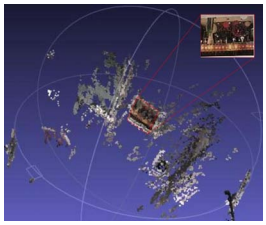

Figure 5: *3D reconstruction.*

6.3. Remote Director mode and Full 3D Experience mode

The third and fourth scenarios are work-in-progress concepts and only have mock-ups available. The third scenario is to add interactive 3D geometrical selection, incorporating functionalities allowing the user to operate the projection. This is termed as the remote director mode. So if in the case shown in figure 4 a remote participant wishes to look at the image on the wall behind the left shoulder of the user, then by clicking on the relevant 3D features in the 3D point cloud which represents this image, the live remote camera can change its pan, tilt and zoom values to highlight this feature. This, with a realigned view of the point cloud is shown in figure 5.

To enable this form of remote interactivity the manipulation techniques to control the 3D model will include: moving around, position, scaling and rotating the scene to view it from different angles. A mapping is then required between projected (x,y,z) values in the 3D point cloud space to camera pan, tilt and zoom values which is the inverse transform matrix of the original feature extraction algorithm.

The final and fourth scenario is the creation of a full-immersive experience where the 2D video streams are projected within a denser 3D feature rich point cloud environment. In this use-case scenario, all the functionalities mentioned above will be brought together, within a more traditional virtual reality environment; addressing 3D reconstruction, spatial awareness, avatars and identity. Interactivity and data communication will still then be managed through the AGTkit environment.

7. Further Work

The future use-case studies are the third and fourth scenarios. We propose to create an interactive tool allowing attendants to manipulate the 3D scene of the remote environment. The user will be able to select the area to be viewed in the display, control the remote cameras and make visible a 3D projection of just the desired area.

The ultimate use-case scenario proposed is the 3D full immersion system, an ensemble of the above scenarios, integrating all their functionalities. The 3D full immersion environment will be capable of recreating a virtual reality venue, where attendants will feel as if they were there, this will be possible using the appropriate 3D hardware. Some activities to be undertaken to evaluate the framework are:

- Real-time performance analysis
- Most important is the measurement of the immersion experience through interviews and simple tasks [SSC10]
- Evaluate the quality of video transmitted over the network

Simulation results will be used to demonstrate the feasibility of the proposed method. Due to the work being still in progress, we have not yet performed real testing experiments.

References

[ALF*09] ANDREW J., LANG M., FYFFE G., YU X., BUSCH J., MCDOWALL I., BOLAS M., DEBEVEC P.: Achieving eye contact in a one-to-many 3d video teleconferencing system. *ACM Transactions on Graphics 28*, 3 (August 2009). 2

[BL07] BROWN M., LOWE D.: Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision 74*, 1 (2007), 59–73. 3

[Bou01] BOUGUET J.-Y.: Camera calibration toolbox for matlab, 2001. `http://www.vision.caltech.edu/bouguetj/calib_doc/index.html`. 2

[BTS*02] BAKER H. H., TANGUAY D., SOBEL I., GELB D., GOSS M. E., CULBERTSON W. B., MALZBENDER T.: The coliseum immersive teleconferencing system. *Proc. Int. Workshop Immersive Telepresence* (2002). 2

[CCR08] CIGNONI P., CORSINI M., RANZUGLIA G.: Meshlab: An open-source 3d mesh processing system. *ERCIM News 73* (Apr 2008), 45–46. `http://meshlab.sourceforge.net/`. 3

[CDO*00] CHILDERS L., DISZ T., OLSON R., PAPKA M. E., STEVENS R., UDESHI T.: Access grid: Immersive group-to-group collaborative visualization. *4th Int. Immersive Projection Technology Workshop* (2000). 2

[DYIR11] DEVINCENZI A., YAO L., ISHII H., RASKAR R.: Kinected conference: augmenting video imaging with calibrated depth and audio. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (New York, NY, USA, 2011), CSCW '11, ACM, pp. 621–624. 2

[FP08] FURUKAWA Y., PONCE J.: Patch-based multi-view stereo software, 2008. `http://grail.cs.washington.edu/software/pmvs`. 3

[HH08] HARRISON C., HUDSON S. E.: Pseudo-3d video conferencing with a generic webcam. *Tenth IEEE International Symposium on Multimedia* (2008), 236–241. 2

[Low04] LOWE D. G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60* (2004), 91–110. 3

[SSC10] SLATER M., SPANLANG B., COROMINAS D.: Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. *ACM Trans. Graph. 29* (July 2010), 92:1–92:9. 4

[SSS06] SNAVELY N., SEITZ S. M., SZELISKI R.: Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph. 25* (July 2006), 835–846. 3