INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

3D Video Conferencing Using Monoscopic Camera

Neha Raste
Student, Cummins College of Engineering for Women, Pune, India

Abstract:
Rise of population has increased the necessity to communicate without commutation. Videoconferencing plays a vital role in this regard. The proximity between the communicating parties further increases if it is possible to actually interact with each other virtually. Tele-immersive environments facilitate this feature, although with expensive, bulky and normally inaccessible setups. A 3D tele-immersive video conferencing based solution seems affordable due to easily accessible infrastructure and efficient operation to virtually collaborate with remote parties. The necessity of multiple camera-clusters for 3D input is greatly reduced by using 2D to 3D conversion of images frames as suggested in the proposed system.

Keywords: *Tele-immersive, 3D rendering, networking, virtual*

1. Introduction

Teleconferencing enables 2 or more parties to communicate without commuting. A dynamic view or a video representation of the communicating parties enhances the communication. Nonetheless, the actual presence of the communicating member is far from felt. The system focuses on rendering a 3D virtual presence in a collaborative environment. Such environments can be described as tele-immersive environments. Tele-immersive environments are being developed from the early 21st century. Following is an overview of the technologies developed since.

Video Conferences are fundamentally software based multi-party video meetings that run on PCs, laptops, mobile devices with network connectivity. Such systems are portable, cheap and easily accessible. A typical tele-presence system may also utilize some additional features like large (human size) displays, audio systems equipped to record qualitative sound with large capture range, HD cameras to fit multiple people in the same room, etc.

However, such systems lack a sense of actual presence, as it does not facilitate with some of the very obvious experiences during actual meet, like palpable bodily movements, gaze orientation, etc.

On the other hand, a human-friendly conferencing environment demands every user to get a perspective in the conference space, which intuitively needs synthesis of 3D world.

An actual tele-immersive system however, enables interaction between remotely located users in real-time, hence creating a need of a shared space for collaboration. The TEEVE project of UIUC and UCB together, enables virtual interaction with 3D camera clusters [3]. Their model involved usage of multiple 3D camera clusters to enable perception from different points of observation; output of which are reconstructed to one 3D video stream, which are then compressed and sent to the other collaborating party over internet. Lastly, the received data is rendered in real time into a virtually- shared space, thereby calling it an 'immersive' environment. [3] Overall, it involves an expensive and non-portable setup which is computationally powerful but not easily accessible.

The proposed system is a tele-immersive system that incorporates normal webcam – based image acquisition to convert it real – time into 3D figures in a virtual environment. It provides the basic 3D feature enabling the participants to interact with each other virtually. In a one to one video-call, people see both of them in a shared virtual space as if on a movie screen. The background can be a synthesized environment which can be provided in the application in form of a library, to choose from.

2. System Architecture

The system consists of 3 parts : Image Acquisition, Conversion of 2D image to 3D and placing them in the virtual environment.

The acquired video frames are segmented, i.e. objects are identified from the image as a whole to distinguish a person's silhouette from its surroundings, and to convert the 2 Dimensional shapes into a 3 Dimensional one, using 2-D to 3-D modeling algorithm later described. Then, the conferees are enabled to be represented by arbitrarily shaped video objects, which can be integrated seamlessly into virtual space. Subsequently, these shaped video objects can be encoded, using an MPEG-4 codec, or any other suitable compression scheme which is a perfect combination of low data rates as well as better quality of picture frames in terms of resolution. MPEG-4 can be used as it allows coding of arbitrarily shaped video objects and also provides auxiliary grey-scale alpha planes that may be used to transmit disparity maps along with the video 3D objects.

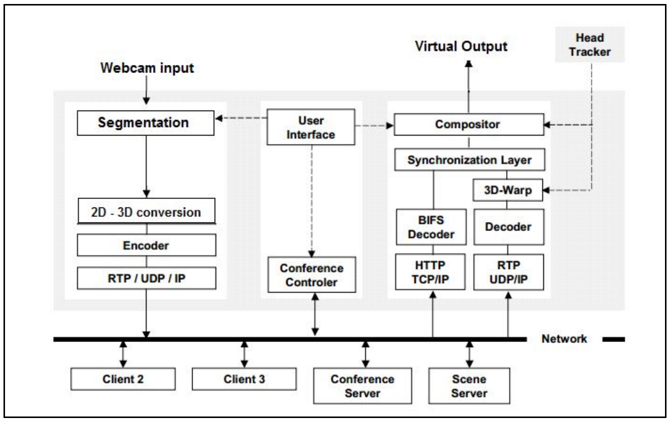

Figure 1: System Block Diagram

Post encoding, the data is sent to other party by RTP through an IP network. The received video streams are decoded using suitable scheme synchronous with encoding system used. So obtained video objects are then synchronized with other scene objects (audio, events, etc.) and integrated into the virtual environment which can be represented using MPEG-4 description language BIFS. The MPEG-4 video objects are processed via a 3D warp using image-based rendering techniques before they are integrated into the artificially synthesized scene. BIFS data can be loaded via hyper text terminal protocol during the initialization phase itself. Also MPEG – 4 facilitates dynamic participation from different terminals. BIFS allows dynamic updating of scene description thus a new object, can also be a person, be easily added to a scene, and so be removed. A conference server dynamically coordinates these changes by receiving requests from particular terminals and by sending related BIFS updates to all parties.

2.1. Video Acquisition and Pre-Processing

Since the image to be included in the final virtual environment is 3D distinct object, it needs to be extracted from a 2D image and preprocessed. Video segmentation is performed to obtain 2 dimensional image frames.

a) Background Subtraction: The conferee's body is extracted from the overall frame, by background subtraction, also called foreground extraction. The rationale behind this algorithm is that the portion of the frame that stays stationary, doesn't change is background. And, a region bounded by particular edge or boundary that moves dynamically is the foreground or object of concern. Thus it is extracted subtracting the rest of the image, leaving only a silhouette of the object, here, the conferee. This computation takes place real time, i.e. the current input pixel data is correlated with the background image. Here the computations are reduced by initializing the conference session when a background image is captured (without the conferee(s).) Due to lesser number of pixel data accompanying the foreground image, computations reduce.

b) Underlying hardware and network requirements: A webcam connected by USB or serial port to a computer should suffice. A front webcam is mostly inbuilt in all laptops and mobile devices. Network connectivity with bandwidth requirement of 2Mbps is required taking in consideration rich media sharing. An enlightened room with minimally cluttered background is suggested.

2.2. 2D-3D Conversion

As mentioned in the former section, an initialization phase is necessary to model a 3D picture from 2D images. During the foreground extraction phase, the silhouette or the boundary of the object in consideration, here, the body is considered. The entire boundary is considered to be a poly-line, i.e. a join of small equal lines forming a curve. Thus, the silhouette is now a polygon which is further divided by constrained Delaunay triangulation algorithm. A chordal axis is determined by joining the mid-points of sides of so formed triangles. The chordal axes are trimmed to define a significant axis. This silhouette is inflated about this significant axis and 3D faces are generated along the spline, i.e. the 3D object is wrapped.

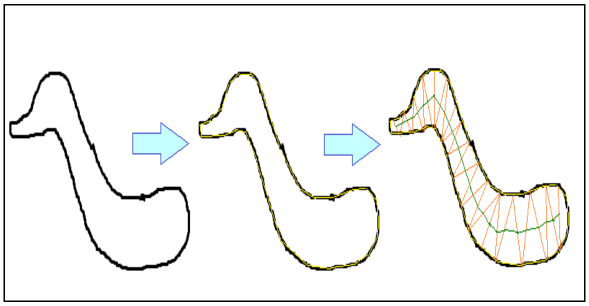

Figure 2: Delaunay triangulation and formation of principal axix

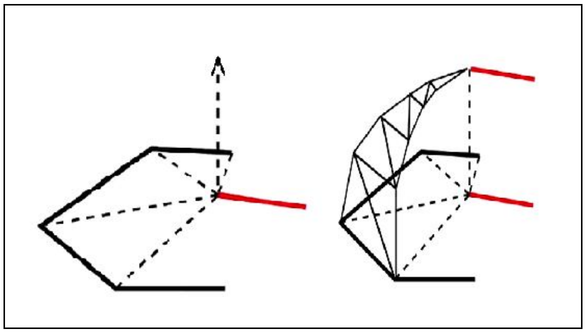

Figure 3: Development of 3D shape along the spline

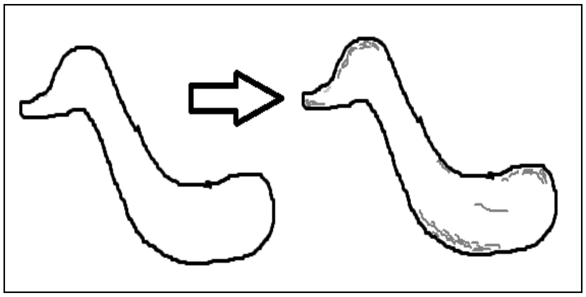

Figure 4: Conversion of 2D to 3D object

2.3. Warping and Composition Of Final Virtual Scene

A 3D graphics card is used to compose the virtual space with artificially synthesized 3D objects. The virtual scene is a mix of 3D objects including the conferees and artificial 'objects' that forms part of the environment. This virtual space is now encoded using some scene description language like BIFS. It involves manipulation of conferees' 3D images and that of synthesized objects.

a) Head tracking: The accurate position of the conferee in the 3D space is accurately estimated by head-tracking. The system is made to identify eyes, which is used as a handle to calibrate positions of people in the virtual space.
b) Graphics: The 3D synthesised virtual rendering of the conferees is considered to be pixel-based textures and are transferred directly to the graphics card through an AGP bus.[1]
c) The virtual scenes may include interactive elements which can be palpated by the virtually immersed participants. For e.g. a user may pick up a ball from the scene and throw it upward. The ball now automatically comes back to the ground of virtual space as it can be programmed to follow the law of gravity.

3. Results

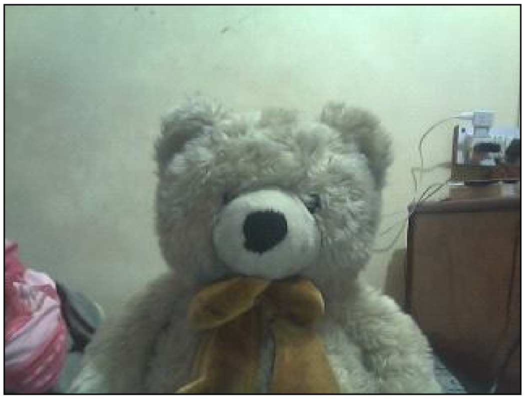

Figure 5: Picture frame captured through webcam

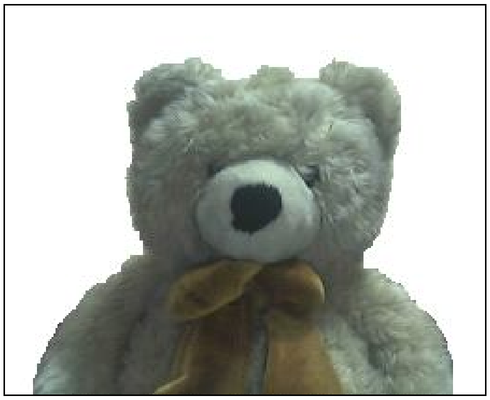

Figure 6: After foreground extraction

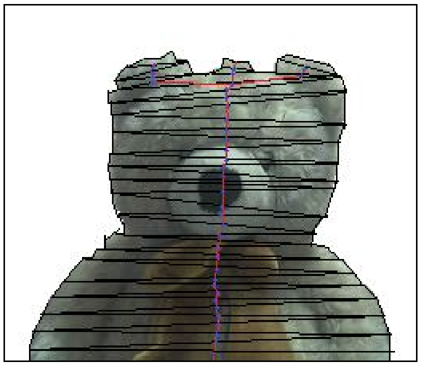

Figure 7: Delaunay triangulation and formation of principal axix

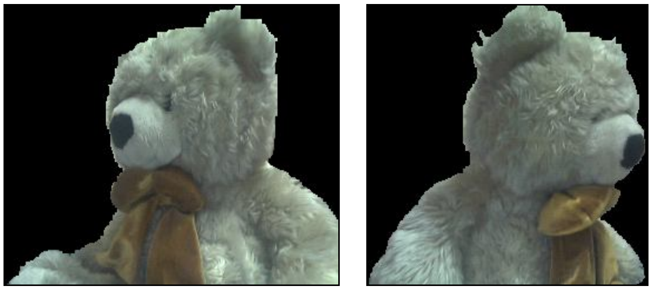

Figure 8 & 9: Left view and Right view of the synthetic 3D object

Such objects can be warped and composed into the desired virtual scene.

4. Conclusions

The processing of 3D objects is done by image based rendering, highly reducing computational load incurred by actual 3D reconstruction, calibration, synchronization of cameras etc. Also, depth map considerations are not required due to image based representation of 3d objects. This makes the system plausible for real time considerations. Virtual synthesis and integrating video objects with the environment by warping facilitates natural and realistic rendering.

5. Future Work

The system currently designed considering only domestic and business use of collaborative environments. Had this system been used in environments such as surgical scientific usage, the system may not work as expected. The 2D to 3D modeling is designed to work for discrete body parts and not for minute features such as those that may be encountered during surgical or other such situations.

6. References

1. Peter Kauff, Oliver Schreer. “An Immersive 3d Video-Conferencing System Using Shared Virtual Team User Environments”.
2. Takeo Igarashi, Satoshi Matsuoka, Hidehiko Tanaka “Teddy: A Sketching Interface for 3D Freeform Design” Siggraph 99.
3. Zhenyu Yang, Nahrstedt, K.,Yi Cui, Bin Yu, Jin Liang, Sang-Hack Jung, Bajscy, R. “TEEVE: The Next Generation Architecture for Tele-immersive Environments”
4. Mun Wai Lee ; Ram Nevatia “Body Part Detection for Human Pose Estimation and Tracking” IEEE Workshop on Motion and Video Computing.
5. Ramanarayan Vasudevan, Gregorij Kurillo, Edgar Lobaton,Tony Bernardin, Oliver Kreylos, Ruzena Bajcsy, Klara Nahrstedt, “High-Quality Visualization for Geographically
   Distributed 3-D Teleimmersive Applications” IEEE transactions on multimedia, vol. 13, no. 3, june 2011.
6. Luis Almeida, Paulo Menezes, Jorge Dias “On-Line 3d Body Modelling For Augmented Reality” Unpublished.
7. www.coursera.edu/icg-001
8. http://mmgt.com/wp-content/
9. http://www.pkeconsulting.com/