A Framework for Realistic 3D Tele-Immersion

P. Fechteler, A. Hilsmann, P. Eisert
Fraunhofer HHI
Berlin, Germany

S.V. Broeck, C. Stevens
Alcatel Lucent
Antwerp, Belgium

J. Wall, M. Sanna
Queen Mary University
London, Great Britain

D. A. Mauro
Telecom ParisTech
Paris, France

F. Kuijk, R. Mekuria, P. Cesar
Centrum Wiskunde & Informatica
Amsterdam, Netherlands

D. Monaghan, N.E. O'Connor
Dublin City University
Dublin, Ireland

P. Daras, D. Alexiadis
Informatics and Telematics Institute
Thessaloniki, Greece

T. Zahariadis
Synelixis Solutions
Chalkida, Greece

ABSTRACT

Meeting, socializing and conversing online with a group of people using teleconferencing systems is still quite different from the experience of meeting face to face. We are abruptly aware that we are online and that the people we are engaging with are not in close proximity. Analogous to how talking on the telephone does not replicate the experience of talking in person. Several causes for these differences have been identified and we propose inspiring and innovative solutions to these hurdles in attempt to provide a more realistic, believable and engaging online conversational experience. We present the distributed and scalable framework REVERIE that provides a balanced mix of these solutions. Applications build on top of the REVERIE framework will be able to provide interactive, immersive, photo-realistic experiences to a multitude of users that for them will feel much more similar to having face to face meetings than the experience offered by conventional teleconferencing systems.

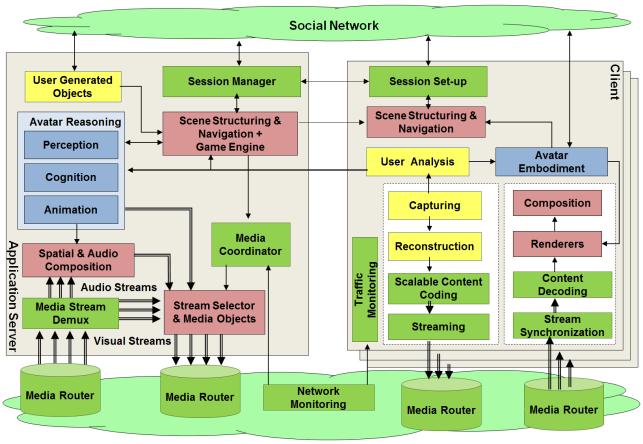

Figure 1: Overall REVERIE system architecture.

1. INTRODUCTION

Teleconferencing systems enable participants in different locations to share a common experience, where specially designed tables and real-time high-definition video increase the feeling of proximity. Current-generation teleconferencing systems allow participants to talk to each other as if they were in the same location. But, their shortcoming is that participants cannot collaboratively perform a task together: they remain captives in a 2D screen projection. The next challenge in tele-presence is tele-immersion (TI), which will enable individuals that are geographically apart to interact naturally with each other in a shared 3D virtual environment. Where teleconferencing allows participants to share a common space, TI allows them to share an activity. REVERIE [24] is a project that aims at such an immersive environment: an ambient, content-centric Internet-based environment where people can work, meet, participate in live events, socialize and share experiences as they do in real life, but without time, space and affordability limitations.

In one of the definitions of Virtual Reality [25], simulation does not involve only a virtual environment but also an immersive experience; according to another author [22], instead of perception based on reality, Virtual Reality is an alternate reality based on perception. An immersive experience takes advantage of environments that realistically reproduce the worlds to be simulated.

An example use case of the REVERIE project: A teacher may take his class on a trip in the virtual world of REVERIE; he and the students can stay in the classroom, sitting behind their desks. Each of them has a PC with a capturing device that captures the upper part of their bodies. The REVERIE

Figure 2: Example 3D reconstructions in various poses, using the implicit volumetric approach (the three on the left) and the explicit mesh-zippering approach (the three on the right).

system automatically blends in instructions for control of uncaptured features (such as their legs). Being represented in the virtual world by a personalized avatar, the students and the teacher see on their display what their representation in the virtual world is supposed to see, the gaze of the avatar corresponds to whom or what the student is looking at on the screen. A virtual guide, a fully autonomous agent, may take them on a tour through a building. In a "follow-me" mode, this agent takes care of navigation of the whole group. By raising a hand, students may get the opportunity to pose a question. The facial expression, the gaze and emotional behavior of the participants are analyzed and reflected in the appearance of their avatar. Assisted by path planning and collision detection, the students can wander around in the environment after the tour. They can inspect objects or elements of the building and they can have private discussions with the teacher or fellow students they meet, their avatar showing their frustration, enthusiasm or indifference, just as in the real world.

2. SYSTEM ARCHITECTURE

Basis of the REVERIE architecture, depicted in Fig. 1, is a model where various REVERIE clients communicate, using an Application Server. The server, that shows up in the figure on the left side, hosts application related modules such as a game engine utilized for scene navigation and modules for avatar reasoning, and networking related modules like session manager, network monitoring and media stream synchronization. At the top of the figure, the social network is highlighted, which will be used for extracting static information of the participants (e.g. public information of their profiles) and dynamic information (e.g. participants availability). User generated content (e.g. virtual environments) could be stored and downloaded or shared on demand from the social network. At the bottom of the figure, the Future (public) Internet is highlighted, which will provide the means for efficient delivery of content.

The REVERIE project exploits and enhances technologies and tools integrated into one common framework to enable end-to-end processing and efficient distribution of 3D, immersive and interactive media over the Internet. This framework takes care of 3D data acquisition and processing (capturing of the human user's gestures and emotional expressions), 3D sound processing, networking, and real-time rendering. An avatar reasoning module deals with autonomous behavior, physical interaction and emotional engagement regarding the virtual humanoid characters.

3. CAPTURING RAW 3D DATA

Capturing and 3D reconstruction of full-body humans in real-time, an important task for multi-party TI applications, is addressed in this section. The main idea is that multiple users at distant sites are reconstructed on-the-fly; their 3D representations are compressed, transmitted and finally rendered inside a common virtual space in real-time.

Many accurate methods for 3D reconstruction can be found, using either active direct-ranging sensors [7] or multiple RGB cameras [9]. However, they have not been applied in real-time TI applications or they are not applicable in a TI framework due to processing time limitations. On the other hand, relevant TI-oriented approaches [26, 20] fuse partial 3D data only at the rendering stage, in order to synthesize intermediate 2D or stereo views. In contrast, the here described methods produce full-geometry 3D textured meshes that are rendered using standard computer graphics techniques. The methods discussed in this section take input from multiple consumer depth cameras, specifically Kinect sensors. Two methods have been developed: A volumetric method that implicitly fuses depth-map measurements and an explicit fusion method based on mesh-zippering ideas (see below).

The capturing system is composed of five Kinect sensors, connected to a computationally strong host PC. The sensors are placed vertically, at a height of approximately 1.80m to capture the whole human body, on a circle of diameter 3.60m and all pointing to the center of the working volume. In order to fully calibrate each single Kinect the method of [12] is used, while for the external calibration a custom-made calibration object with three intersecting large planar surfaces is used by exploiting the depth information from the sensors.

Reconstruction is performed for each captured frame, resulting in a time-varying mesh sequence. The methods were implemented based on the CUDA parallel computing architecture and are running in near real-time (approx. 10fps).

Data preprocessing is similar for both reconstruction meth-

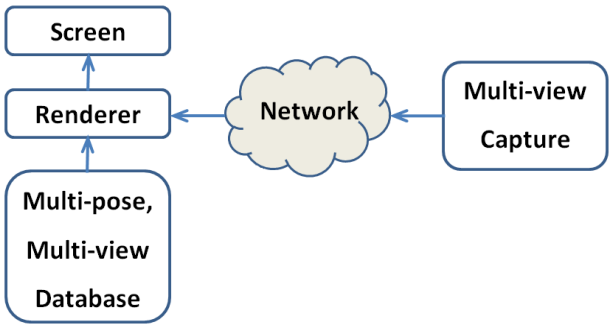

Figure 3: Concept of model based capturing and rendering.

ods and can be summarized as follows: (a) A weak 2D bilateral filter is applied to the depth maps to reduce Kinect measurements noise; (b) A binary "human silhouette" map is generated for each depth map, by background subtraction; (c) For each depth map, a surface-normal map is calculated; (d) Additionally, a "distance-to-background" confidence map is computed, which accounts for the fact that Kinect measurements near object boundaries are noisy. (e) Finally, using the normal-map and "distance-to-background" map, a "confidence" map is calculated.

The data from multiple depth sensors can be implicitly fused in a weighted manner, following a volumetric approach. The objective is to compute an appropriate volumetric function $V(\mathbf{X})$, which is negative (empty) outside the captured solid, positive (occupied) inside the solid and (almost) zero near its surface. Then, the final surface can be reconstructed by the extraction of the zero-level isosurface of $V(\mathbf{X})$, using a marching cubes algorithm. The employed approach is similar to the volumetric Signed Distance Function [7], but takes into account the fact that opposite sides of the solid objects are captured by opposite-facing Kinects and produces almost watertight models. All of the depth camera observations contribute to $V(\mathbf{X})$ in a weighted manner, using the "confidence" maps. The reconstruction results shown in Fig. 2 are with a 3D space discretization of $2^7 \times 2^8 \times 2^7$ voxels.

The explicit fusion via mesh-zippering ideas is based on the notion of Step Discontinuity Constrained (SDC) triangulation. It can be shortly summarized as follows: (a) Initially, five separate meshes are generated, each one corresponding to one Kinect, using terrain triangulation on the 2D depth image plane. More specifically, the idea of SDC triangulation is realized, which assumes that pixels that are adjacent in the 2D depth image generate connected vertices, unless their Euclidean 3D distance is higher than a threshold. (b) The generated separate meshes contain redundancy in the sense that they have a significant overlap. Additionally, the overlapping mesh regions are practically quite noisy, because they mainly contain regions near the captured object boundaries and due to interference between corresponding Kinects. Therefore, the method continues with the decimation of the overlapping mesh regions, working in pairs of adjacent meshes. Finally, (c) a "clipping" step, which is based on the detection of adjacent mesh regions and local constrained Delaunay triangulation, "stitches" together the multiple meshes. Reconstruction results are given in Fig. 2.

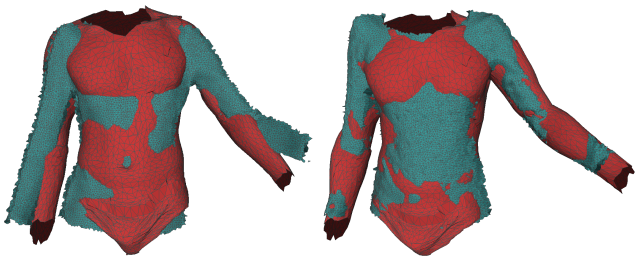

Figure 4: Input (left) and output (right) of the Kinematic Template Fitting algorithm for tracking, with the template model colored in red and scan data in green.

The Kinect RGB camera allows only automatic white-balance and exposure control. This often results in color values that vary significantly between adjacent RGB views. An on-the-fly technique is used that builds color matching/correction functions during reconstruction, which aims at minimizing the color difference between pairs of pixels in two cameras that capture (approximately) the same 3D points.

Then, the color of each vertex is obtained as the weighted average of the pixels in all visible cameras that the vertex projects to. The "confidence" maps are used to weight the color measurements. Practical experiments have shown that the use of this weighted color mapping scheme significantly improves the visual quality.

4. MODEL BASED CAPTURING AND RENDERING

A major limitation of directly streaming the captured data to allow for free viewpoint rendering, e.g. images-plus-depth or pure multi-view, is the very high bandwidth consumption, which drastically limits the achieved level of quality.

To overcome this limitation, we exploit an underlying articulated template model and a pose-dependent database of appearance to render arbitrary viewpoints solely based on pose parameters, see Fig. 3. Such, only parameters describing a person's pose (typically, a very limited number of skeleton joint angles) need to be sent over the network. A pose-dependent, multi-view database of high-resolution appearance examples of a participant is created in an offline fashion, see Fig. 5. This database contains intensively preprocessed data, like alpha masks, depth maps, skeleton information as well as image warp functions between the database images to transform one pose into another one. Following the approach presented in [13], we synthesize highly realistic animations from this database solely based on the joint angles of the underlying template model for the desired pose by warping and merging the pre-captured database images. The main idea is that small pose-dependent details are modeled by appearance/images, enabling highly realistic visualization while coarse 3D information allows animation and free-viewpoint rendering.

In order to use this animation scheme for TI purposes, the required parameters of all joint angles and global similarity have to be extracted on-the-fly from a live capture. For this purpose we capture the user with multiple calibrated

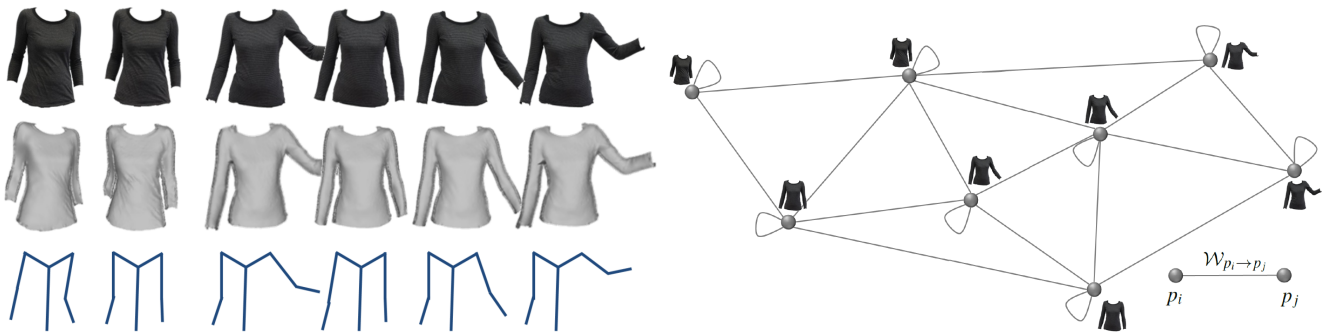

Figure 5: High-resolution, multi-view, multi-pose database: Appearance, depth map and pose per frame (left) and pose-to-pose warp function (right).

HD cameras in a green-screen environment and subtract the background using chroma-keying techniques. Then, using the ICP-based Kinematic Template Fitting algorithm [8], the rigged human template model mentioned above is fitted in a globally consistent manner against the 3D reconstructed depth maps of the segmentation results in a shape aware manner. Fitting results are depicted in Fig. 4. The extracted joint configuration currently consists of 51 float values and completely describes the pose of the user. With the usual 4 bytes per float value and a desired framerate of 25fps, a bandwidth of 5.1 kbyte/sec is sufficient for this approach. These parameters are streamed to all other participants, in order to synthesize their virtual view onto this user.

5. 3D AUDIO RENDERING

3D recordings and audio, namely techniques that aim to create the perception of sound sources placed anywhere in 3D space, are becoming an interesting resource for composers, live performances and augmented reality systems. It is important to note that a properly rendered audio environment can enhance the sense of immersion and presence in the scene, but the problem of recreating an immersive experience is not trivial. With a standard headphones system, sound seems to have its origin inside the listener's head. This problem is solved by binaural spatialization, which gives a realistic 3D perception of a sound source located somewhere around a listener.

Binaural spatialization is a technique that aims at reproducing a real sound environment using only two channels (like a stereo recording). It is based on the assumption that our auditory system has only two receivers, namely the ears, and to obtain a representation of the acoustic environment it exploits some physical parameters of the signal called "cues" [29, 5]. If it is possible to deliver a signal equal (or nearly equal) to the one which a subject would receive in a real environment, this will lead to the same perception.

Currently, most projects using binaural spatialization aim at animating the source while keeping the position of the user fixed. However, for an immersive experience this is not sufficient: it is necessary to know at any time the position and the orientation of the listener within the virtual space in order to provide a consistent signal [23], so that sound sources can remain fixed in virtual space independently of movements, as they are in natural hearing [4].

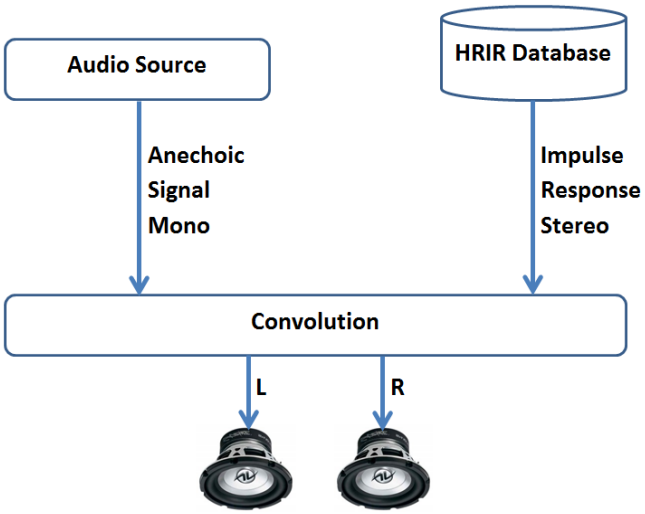

Figure 6: The workflow diagram of the audio subsystem.

The REVERIE framework proposes the integration of an audio subsystem that receives the information regarding the position of the sound source with respect to the listener, chooses an appropriate head related impulse response (HRIR) from a database and then performs the necessary operations: FIR (Finite Impulse Response) filtering or frequency-domain convolution. This choice depends mostly on the "size" of the filter kernels. For short filters FIR filtering performs very well while, given the complexity, moving to longer impulse responses (such as the ones of reverberant environments) could lead to performance degradation. For this reason convoluting in frequency domain via FFT is a convenient strategy. A schematic overview is depicted in Fig. 6.

The REVERIE audio subsystem is also extending the system by adding room acoustic simulation (reverb). This helps to enhance even more the sense of immersion and the "coherence" with the virtual environment.

6. REAL-TIME STREAMING

3D TI provides a common virtual space, where distributed participants can naturally interact. Advances on 3D reconstruction and rendering - and the success of Microsoft's

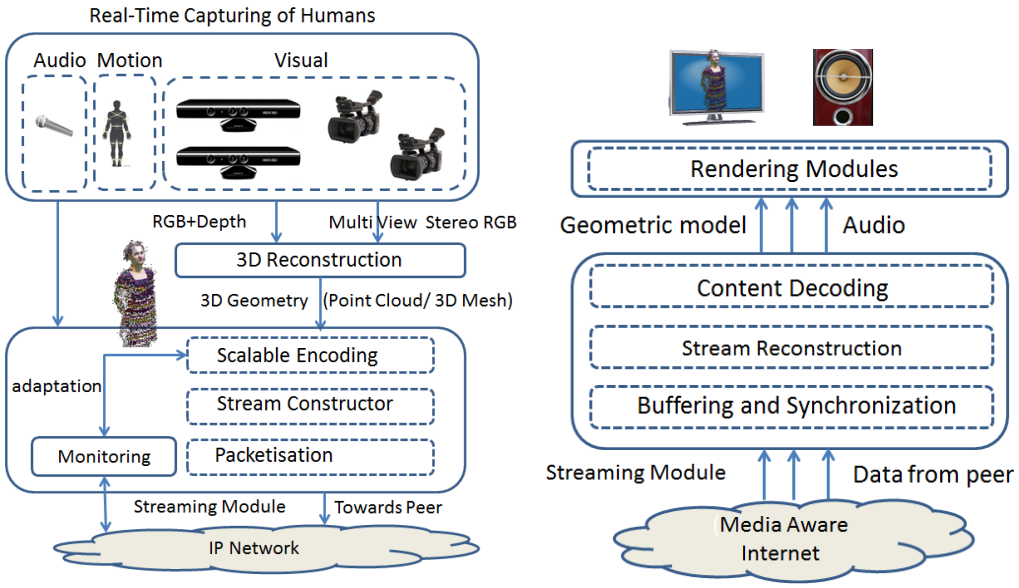

Figure 7: Source location (left) and destination location (right) of the media pipeline.

Kinect - enable, in real-time, the creation of highly realistic representations of the participants (see sections 3 and 4). Compared to previous approaches, which typically use image-based content representations (e.g. 3D videos) [14], this new generation of 3D TI systems ensures the immersion of users, enables the integration with virtual worlds, allows for adaptation mechanisms that make an efficient use of the network resources, and takes advantage of advanced rendering methods provided by modern graphic cards. Unfortunately, due to its novelty, existing video codecs and packetisation schemes do not support such representations (neither do geometry streaming mechanisms intended for downloading and interacting with remotely stored geometry-based objects). While research in the past has provided some solutions for streaming geometric objects (e.g. [3, 18, 6]), none of them handled the critical real-time constraints imposed by 3D TI. In these types of infrastructures, synchronization between media streams is particularly challenging, as the pipeline for visual data introduces large delays compared to the pipeline for immersive audio. Moreover, current solutions intended for video (e.g. RTP/RTSP or MPEG-TS) are not suitable for captured geometric representations, unless they are extended with support for specific 3D graphics codecs [16].

This section presents a streaming engine that meets real-time encoding/decoding constraints. In particular, this streaming engine includes both a novel fast local compression algorithm and a rateless packet protection scheme. These components have been successfully integrated into a larger TI environment that includes beyond the state of the art 3D reconstruction and rendering modules. This resulted in a prototype that can capture, compress, transmit, and render triangle mesh geometry in real-time over the Internet. Fig. 7 shows the media pipeline. At the sender side, different types of media are captured: audio, motion (for model-based avatars), and visual data. In this section we are interested in the visual pipeline for transporting geometry data when capturing users. As users interact, this part of the pipeline is the most challenging in terms of real-time streaming and synchronization. After capture, the representation has to be encoded using an efficient compression method. Subsequently, the streams are packetized and adapted, so they can be sent over the network. Channel coding is applied to cope with lossy transmission over UDP. TCP could be used to avoid losses, but the end-to-end delay would be compromised. The data is then transmitted to the remote sites. At the receiver side, the received packets are first buffered and synchronized. Then, re-construction of the stream takes place. The stream is decoded and rendered into one common virtual environment by a set of rendering modules, depending on the type of streamed data.

In terms of encoding, the system takes advantage of the 3D reconstruction phase [2], where a sequence of triangular meshes with different numbers of vertices and changing topology are produced. This implements an optimized encoding algorithm that exploits the regularities introduced in the connectivity and geometric data. It achieves a compact representation for each reconstructed mesh. Tests on a large dataset show an encoding and decoding speed-up of over 10 to 20 times at similar compression and quality rates, when compared to the high-end MPEG-4 SC3DMC mesh encoder [21]. Generally, mesh geometry transmission over lossy networks has been challenging, as triangle mesh compression have not been designed with resilience to losses in mind. In the presented system, a rateless code to achieve minimal end-to-end delay and protection against packet losses has been implemented. This implemented rateless code, similar to fountain codes [19], ensures complete packet loss protection of the triangle mesh object and avoids delay introduced by retransmissions. This approach is compared to a streaming mechanism over TCP and outperforms it at packet loss rates over 2% and/or latencies over 9ms in terms of end-to-end transmission delay [21].

3D TI pushes the limits of current infrastructures because

of the high volume of synchronized data that needs to be transmitted in real-time between different locations. The presented streaming engine supports various types of 3D representation, including 3D audio. First, live-captured full geometry (triangle mesh / point clouds) can be delivered. In particular, we use octree compression for point clouds [15] (available in PCL) and MPEG-4 SC3DMC coding [18] for triangle meshes. Second, MPEG BBA/AFX is used for transmitting motion commands, if a model-based representation (e.g. the avatar) has been downloaded in advance. Finally, the engine can stream stored 3D videos (video plus depth) for extra material to be shown in the virtual world (e.g. a movie in the common virtual space). The next challenge is to support media synchronization, independently of the media representations being transmitted. For this, a time client is required in the form of a virtual clock for ensuring time coherence. Moreover, both sender and receiver need to provide support for synchronization. At the sender side, timestamps (based on a common clock) are generated, and later used by the receiver side. Delay estimation can be used for easing or enforcing encoding and scheduling parameters (rate estimation, priority encoding), so a targeted end-to-end delay is achieved. At the receiver side, streams should be aligned for ensuring synchronization between the different data streams (audio, geometry, and movement, so a coherent scene comprising the actions at the right moment is constructed.

7. HUMAN-LIKE BEHAVIOR IN AN IMMERSIVE VIRTUAL WORLD

So far we presented capturing, modeling, rendering and real-time streaming as enabling techniques for TI. To make the users feel immersed we need more than that. The virtual environment of REVERIE is populated with (semi-autonomous) avatars that represent users and autonomous agents that serve for some kind of support. Just as in the real world, these virtual humans should react on their environment, especially on humanoids they meet.

Human perception is very sensitive to the behavior and the expression of emotion in others. As a result, users will only feel immersed in a virtual environment if the behavior of humanoid characters in that environment adheres to their expectations, based on their life-long observations in the real world. Any discrepancies can be disturbing [11, 27]. Within the context of the REVERIE project, an avatar reasoning framework has been developed that serves to model the mind of a virtual human as well as analyzing the captured data from a real human and modeling the interaction between virtual and real. This framework allows people to be represented in an Internet based virtual world. This component brings accuracy of the behavior of avatars and agents as being the most essential element of feeling immersed.

The architecture of the avatar reasoning framework that supports the use case presented in the Introduction is shown in Fig. 8. The different instantiations of each class of component that show up reflect the multi-modality required for REVERIE, each instantiation dealing with a specific aspect of human behavior. Analyzer modules receive input from external data capturing components and perform basic processing on the incoming data, ranging from transformation of data to gesture recognition; it does not include interpretation. The Interpreter modules are responsible for high-level interpretation of input produced by the analyzers in the context of appropriate states, and subsequently update relevant states. In other words, the Interpreter modules add intelligence; they are responsible for how input is perceived subjectively, rather than objectively. Proposer modules produce candidate actions based on the current state of the avatar and its context. Realizer modules check for consistency and resolve conflicts in proposed actions. Preferred actions are selected and turned into a behavior plan. The Navigation, Animation Blending and Speech components produce the low level graphics data needed for rendering.

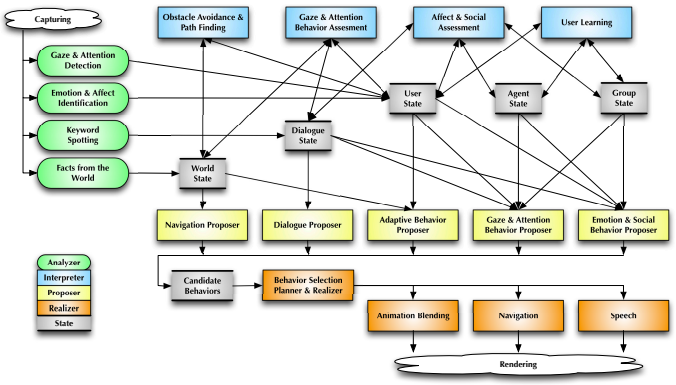

Figure 8: Functional model of the framework, consisting of Analyzers, Interpreters, Proposers, Realizers and State modules.

Each user's semi-autonomous avatar will learn from their user's input to allow the avatar to act independently in certain scenarios [10]. This semi-autonomous behavior needs to be visually indistinguishable by emulating the behavior of the user. This is achieved by applying intelligent techniques to learn from the user's behavior in response to their actions with different objects and characters in the virtual world. Fuzzy logic theory is an ideal medium for embedding this autonomous behavior in the users' avatar. There are numerous rationales for using fuzzy logic: it can be used to obtain a reasonable model when real world data is too complex to be utilized in a dynamic way in real-time; it will provide a degree of uncertainty which will result in more believable and realistic actions for the semi-autonomous avatar; and fuzzy logic can control behaviors at run-time speeds and meet the requirements of real-time 3D simulations [1]. In the current avatar reasoning framework, fuzzy logic is being employed by developing fuzzy inference systems which describe the user's interaction with distinct objects in the virtual world. To fully describe the user's behavior as they interact with the world as a whole, the individual fuzzy systems are grouped into an ensemble. As a rule, combining individual intelligent systems into an ensemble improves both robustness and accuracy [17].

8. GAZE TRANSLUCENCY

During the hunt for food, people got specialized over time in acquiring high visual acuity of things they are focussing at and in sensing motion in the peripheral of their view [28]. These unique features of our human vision system were es-

Figure 9: Video based reference prototype realizing gaze translucency for a multi-media conference.

sential to us then and are still important to us now for instance in traffic situations. We use this ability also in our daily conversations where we are able to somehow feel when someone is looking at us and to observe who someone else is gazing at. This ability to see what other people are currently seeing in real-time is coined gaze translucency.

Current video communication tools like Skype or Google Hangout do not provide gaze translucency which makes a video session with several people very different from our face to face conversations. Often, a mosaic-like composition of the individual users is shown in which everyone in the mosaic is staring directly at the user. As a result, the user feels as if he is getting more attention then he deserves.

Based on their collective gaze information, the REVERIE framework implements gaze translucency by dynamically adapting the rendering of the users participating to a remote multimedia session. By reusing the modality of the multimedia session to also visualize who is looking at whom, the remote multimedia session should feel like a natural immersive experience.

The view adaptation provided by the REVERIE framework also reuses the unique features of our natural human vision system in terms of visual acuity and motion sensing. Though more appropriate alternatives may eventually be provided, the video based reference prototype depicted in Fig. 9 makes use of the following view adaptations.

- The person someone else is looking at is immediately shown with normal brightness while all other users will slowly fade out to a lower brightness.
- People looking at the current user are immediately shown in colour while people looking away from the current user slowly fade out to monochrome.
- When people are looking directly at each other, an additional green border is shown.

Alternative view adaptations can make use of different configurable and personalized view adaptations like glooming or blurring. In the geometry based REVERIE system, specific properties of an avatar may be adapted to realize gaze translucency. By making reuse of the multimedia modality being used and the natural properties of the human vision system, the REVERIE framework realistically reproduces the world to be simulated.

9. CONCLUSION

With the ever increasing availability of high powered computers, high-speed broadband and low-cost 3D capture technologies, truly immersive online interactions are becoming more of a reality. In this paper we have outlined some of the most important aspects of providing a realistic and believable online immersive experience. We have presented unique and novel methods to address these aspects utilizing affordable and readily available technology, all integrated in one common framework called REVERIE. This framework consists of high-quality real-time 3D capturing based on a five-Kinect setup as well as an intelligent combination of an articulated template model and image-based rendering techniques, which requires only a very limited set of parameters to be transmitted in order to still achieve photo-realistic free-viewpoint visualizations of participants. The feeling of 3D immersion is complemented with spatialized 3D audio sound of all participants with consideration of the acoustics of the geometry of the virtual room. Networking strategies have been implemented in order to fulfill the real-time constraints for the occurring diversely natured and hugely sized data, which need to be streamed. On top of the pure visual reproduction of the participants, a user's view onto the other participants is adapted according to their gaze directions, in order to visualize who is looking at whom. Besides that, we have seen in the use case how the REVERIE system allows a teacher to take his class on a virtual trip into the REVERIE virtual environment. The avatar reasoning framework serves to bring in accuracy of the behavior of avatars and agents, as studies have shown that this is the most essential element of feeling immersed. In that sense, having a discussion with virtual fellow-students that look you in the eye outperforms the experience of using teleconferencing systems that display

participants with their eyes fixed on a screen.

10. ACKNOWLEDGMENTS

The research that lead to this paper was supported in part by the European Commission under the Contract FP7-ICT-287723 REVERIE.

11. REFERENCES

[1] R. Aguilar, V. Muoz, and Y. Callero. *Control application using fuzzy logic: Design of a fuzzy temperature controller*, pages 379–396. Fuzzy Inference System - Theory and Applications. May 2012.

[2] D. Alexiadis, D. Zarpalas, and P. Daras. Real-Time, Full 3-D Reconstruction of Moving Foreground Objects From Multiple Consumer Depth Cameras. *IEEE Transactions on Multimedia*, 15(2):339–358, 2013.

[3] G. AlRegib and Y. Altunbasak. 3TP: An application-layer protocol for streaming 3-D graphics. *IEEE Transactions on Multimedia*, 7(6):1149–1156, 2005.

[4] D. R. Begault. *3-D Sound: For Virtual Reality and Multimedia*. Academic Press Professional, Cambridge, MA, 1994.

[5] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, Cambridge, MA, 1996.

[6] W. Cheng, W. Ooi, S. Mondet, R. Grigoras, and G. Morin. Modeling progressive mesh streaming: Does data dependency matter? *ACM Transactions on Multimedia Computing, Communications, and Applications*, 7(2), 2011.

[7] B. Curless and M. Levoy. A Volumetric Method for Building Complex Models from Range Images. In *SIGGRAPH '96*, 1996.

[8] P. Fechteler, A. Hilsmann, and P. Eisert. Kinematic ICP for Articulated Template Fitting. In *Proc. International Workshop on Vision, Modeling and Visualization (VMV'12)*, Magdeburg, Germany, 12-14 November 2012.

[9] Y. Furukawa and J. Ponce. Carved Visual Hulls for Image-Based Modeling. *International Journal of Computer Vision*, 81:53–67, 2009.

[10] M. Gillies and D. Ballin. Integrating Autonomous Behavior and User Control for Believable Agents. In *Proc. International Joint Conference on Autonomous Agents and Multi-agent Systems*, volume I, pages 336–343. IEEE, 2004.

[11] J. Gratch and S. Marsella. A Domain-independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.

[12] D. Herrera, J. Kannala, and J. Heikkila. Joint Depth and Color Camera Calibration with Distortion Correction. *IEEE Trans Pattern Anal Mach Intell.*, 34, 2012.

[13] A. Hilsmann, P. Fechteler, and P. Eisert. Pose Space Image Based Rendering. *Computer Graphics Forum (Proc. Eurographics)*, 32(2), 6-10 May 2013.

[14] Z. Huang, A. Arefin, P. Agarwal, K. Nahrstedt, and W. Wu. Towards the Understanding of Human Perceptual Quality in Tele-Immersive Shared Activity. In *Proc. ACM Conference on Multimedia Systems*, pages 29–34. ACM, 2012.

[15] B. Jovanova, M. Preda, and F. Preteux. MPEG-4 Part 25: A graphics compression framework for XML-based scene graph formats. *Signal Processing: Image Communication*, 24(1-2), 2009.

[16] K. Mamou and T. Zaharia and F. Preteux. TFAN: A low complexity 3D mesh compression algorithm. *Computer Animation and Virtual Worlds*, 20(2-3), 2009.

[17] M. Korytkowski, R. Nowicki, L., Rutkowski, and R. Scherer. Merging Ensemble of Neuro-Fuzzy Systems. In *Proc. International Conference on Fuzzy Systems*, pages 1957–1957. IEEE, 2006.

[18] H. Li, M. Li, and B. Prabhakaran. Middleware for Streaming 3D Progressive Meshes over Lossy Networks. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(4), 2006.

[19] M. Luby, A. Shokrollahi, M. Watson, and T. Stockhammer. Raptor Forward Error Correction Scheme for Object Delivery. RFC 5053 (Proposed Standard), Oct. 2007.

[20] A. Maimone and H. Fuchs. Encumbrance-free Telepresence System with Real-time 3D Capture and Display using Commodity Depth Cameras. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Basel, Switzerland*, 2011.

[21] R. Mekuria, M. Sanna, S. Asioli, E. Izquierdo, D. Bulterman, and P. Cesar. A 3D Tele-Immersion System based on Live Captured Mesh Geometry. In *Proc. ACM Multimedia Systems Conference*. ACM, 2013.

[22] K. Osberg. *But What's Behind Door Number 4??? Ethics and Virtual Reality: A Discussion* . Human Interface Technology Lab Technical Report R-97-16, 1997.

[23] S. P. Parker, G. Eberle, R. L. Martin, and K. I. McAnally. Construction of 3-D Audio Systems: Background, Research and General Requirements. Technical report, Victoria: Defence Science and Technology Organisation, 2000.

[24] REal and Virtual Engagement in Realistic Immersive Environments (REVERIE). Project web site: www.reveriefp7.eu.

[25] J. Steuer. Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication*, 42(4):73–93, 1992.

[26] R. Vasudevan, G. Kurillo, E. Lobaton, T. Bernardin, O. Kreylos, R. Bajcsy, and K. Nahrstedt. High-Quality Visualization for Geographically Distributed 3-D Teleimmersive Applications. *IEEE Transactions on Multimedia*, 13(3), June 2011.

[27] V. Vinayagamoorthy, M. Gillies, A. Steed, E. Tanguy, X. Pan, C. Loscos, and M. Slater. Building Expression into Virtual Characters. In *Eurographics Conference State of the Art Reports*, Vienna, Austria, 4-8 September 2006.

[28] E. Werner. *Manual of Visual Fields*. New York, Churchill Livingstone, 1991.

[29] W. A. Yost. *Fundamentals of Hearing: An Introduction*. Academic press London, 1994.