Volumetric Hand Reconstructions for Non-verbal Communication in Shared Virtual Environments

C. John [1,2] and H. Regenbrecht[1] and U. Schwanecke[2]

[1]University of Otago, New Zealand
[2]RheinMain University of Applied Sciences, Germany

Abstract

Enhancing desk-based computer environments with virtual reality technology requires natural interaction support, in particular hand tracking and visualization. In shared virtual environments hand tracking does not only support interaction with the virtual environment, but also provides a channel for non-verbal communication.
This paper presents a tabletop user interface and its integration into a three-dimensional teleconferencing system. Dense volume reconstructions of hands based on a probabilistic Shape from Silhouette algorithm are applied for tracking and visualization. The interface integration employs the derived probabilistic hand reconstructions to interactively generate mesh models characterized as isosurfaces of probabilistic volume data. It will be shown, that meshes resulting from probabilistic volume reconstructions thereby exhibit superior detail in comparison to mesh models generated from non-probabilistic comparative approaches which represent the current state of the art.

Categories and Subject Descriptors (according to ACM CCS): H.5.3 [Information interfaces and presentation]: Group and Organization Interfaces—Computer-supported cooperative work; I.4.8 [Image processing and computer vision]: Scene Analysis—Shape; I.3.6 [Computer Graphics]: Methodologies and Techniques—Interaction techniques

1. Introduction

The implementation of (tele) presence, which is the perceived sense that an users own body and body parts belong to the virtual world presented, is essential for virtual and mixed reality environments. Telepresence thereby involves a consistent and congruent interplay of tracking and visual display of a human operator in an efficient, accurate and robust manner [HD91]. Of paramount importance here is the accurate and efficient registration, tracking, reconstruction and display of hands. Hand data are important for visualization of artificial representations within a virtual environment (presence) and for interaction with virtual objects or subjects (grasping, gestural communication).
In our current work we focus on the use of hands for non-verbal communication. A vision based tabletop user interface [JSR09a] therefore has been integrated into a shared virtual environment called cAR/PE! [RLK*04]. The interface allows the computation and tracking of probabilistic hand volume reconstructions that are employed here for mesh model generation and display. We will present hand models which show, that in comparison to non-probabilistic approaches, superior quality meshes can be generated from volumes computed with a probabilistic *Shape from Silhouette* (SfS) algorithm.
SfS type algorithms can be subdivided into two major categories, surface driven (polyhedral) and voxel driven methods. Voxel driven methods thereby exhibit a higher robustness against segmentation errors in which case polyhedral methods produce incomplete or corrupted surface meshes. Due to this aspect they are commonly preferred in the literature [HLS04] and were also the matter of choice in the tabletop interface presented. The following overview therefore will be limited to voxel based SfS.
In voxel based SfS approaches the user space is subdivided into voxels and volumetric reconstructions are usually computed in a two step procedure. In a first step the silhouettes of an object are segmented out of each camera image. These silhouettes are then back projected into the user space in a second step (see figure 1) to find voxels which lie within the

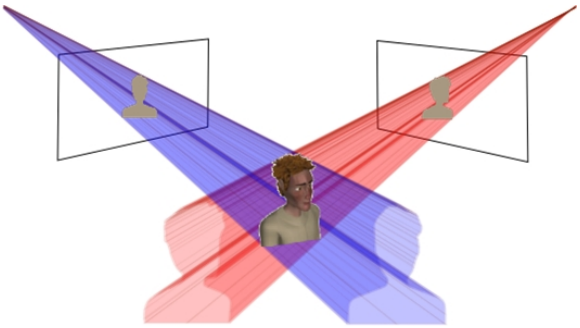

Figure 1: *Shape from Silhouette principle*

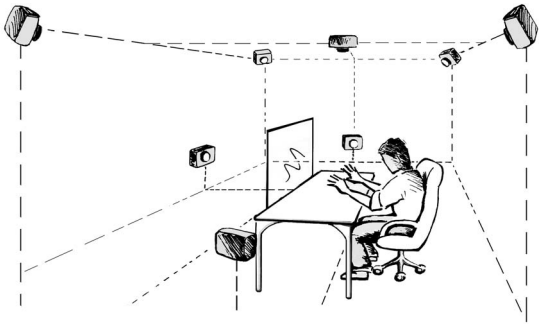

Figure 2: *Proposed tabletop environment*

common intersection volume of all back projections. The resulting intersection volume can now be considered as a dense approximation of the real object shape. It is commonly denoted as the objects visual hull [Lau94].

2. Related Work

Visual hull reconstructions recently gained popularity in different fields of research like body tracking applications [MCA06, dAST*08] or appearance capture for free viewpoint video [GM03, HLS04, dAST*08]. For display purposes of reconstructed objects voxels are usually plotted directly [MCA06], a mesh is generated from volume reconstructions [HLS04, GM03], or in a more recent approach [dAST*08] a given base mesh is aligned and deformed to fit the reconstructed object. What is common to all these system implementations is, that their tracking and/or visualization components are based on visual hull input derived from a non-probabilistic SfS type approach.

In non-probabilistic SfS algorithms the computation of reconstructions is subdivided into two strictly separated steps. First is to segment the silhouettes out of each camera image by drawing a binary decision for each image pixel, and second is to project the silhouettes back into the volume and compute the visual hull by again drawing a binary decision for each voxel in the user space. Each time a binary decision has to be made information is lost. In addition, due to the strict two step separation, visual hull reconstructions become more sensible against image segmentation errors than they actually have to be. A more robust probabilistic approach addressing these issues has been first introduced by Landabaso *et al.* [LP08] who replaced the image based first classification part of SfS with soft assignments of likelihoods. They casted the voxel segmentation task of the second SfS step as maximum a posteriori (MAP) assignments in a discriminative model. This resulted in more robust visual hull reconstructions, because the probabilistic formulation allowed the introduction of a 3D background model description.

An extended real time capable GPU version of this algorithm has been presented in [JSR09b]. The algorithm was reformulated to deal with different classifiers and to introduce an additional 3D foreground model, thus leading to a higher probabilistic distance between the fore- and background process. Besides this also an interactive hand tracking application on top of these reconstructions has been presented which showed their feasibility for scene analysis purposes.

3. Tabletop User Interface

In the following an overview of the tabletop interface to be integrated is given. The hardware setup of the interface is based on eight cameras which are placed around the desktop of an user and are connected to a single host computer, see figure 2. The camera setup is calibrated in a three step approach. First, the intrinsic camera parameters are estimated using the MATLAB camera calibration toolkit, second the extrinsics are computed using the "Self-Calibration" toolkit from Svoboda *et al.* [SMP05] and third an affine alignment to a defined metric world frame centered at the users desktop is computed.

The reconstruction component of the system is based on a probabilistic SfS algorithm which has been implemented on a GPU to allow the computation of visual hulls in real-time. For visual hull reconstruction now different image based classifiers for fore-/background and skin color detection are combined in a Bayesian framework to compute the MAP estimate for each voxel location that it belongs to a skin colored foreground object, using

$$\max_{n=1,\ldots,N} P(\Gamma_n|\mathbf{c}_1,\ldots,\mathbf{c}_S) = \max_{n=1,\ldots,N} P(\Gamma_n)\cdot\prod_{s=1}^{S}\frac{P(\Gamma_n[s]|\mathbf{c}_s)}{P(\Gamma_n[s])}.$$

Here Γ_n is defined as one of N possible skin colored foreground super classes given as the combinations of observed skin colored foreground or background in the S different camera sensors. Multiple foreground classes exist in case of known or tolerated occlusion situations between cameras and an object to be reconstructed. $P(\Gamma_n)$ and $P(\Gamma_n[s])$ are super class priors which can be derived from example visual hull reconstructions. That way scene knowledge is integrated into the assignment rule. Finally $P(\Gamma_n[s]|\mathbf{c}_s)$ is the conditional probability of a certain skin colored foreground

Figure 3: *Three user setup in cAR/PE! environment*

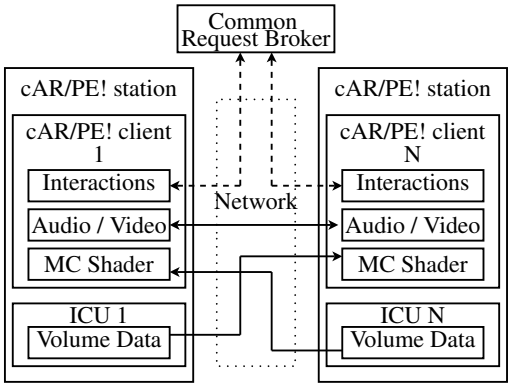

Figure 4: *Information flow between cAR/PE! and ICU*

or background observation $\Gamma_n[s]$ in view s given an observed color vector $\mathbf{c}_s$ in this view. This probability is obtained as a fusion of the different image based classifiers. See [JSR09b] for a more detailed description. Summarized the pSfS algorithm can be described as follows: Based on all image observations and a 3D model of the fore- and background process of the scene, a likelihood is assigned to each voxel location that characterizes it's probability to belong to a skin colored foreground object.
The computed reconstructions are now internally used to detect and track centroids of head and face volumes. The mean centroid tracking error of the system has been measured with 15.8mm and a standard deviation of 7.1mm. The worst case tracking error is usually bound to an interval of ±5cm. The bounding volumes for hand reconstructions were therefore specified with a size of $30^3 cm^3$, assuming a hand size of 20cm±5cm variability.

4. Integration into cAR/PE!

The 3D teleconferencing application cAR/PE! essentially provides a virtual room in which people can meet for collaboration. In the current, integrated scenario a person is represented as a video plane with 3D virtual hands. A possible three user setup is depicted in figure 3. The system consists of several instances of the cAR/PE! client application and a "Common Request Broker" (CRB) which manages link connections and message passing between the different clients. Bandwidth intensive communication channels based on audio, video and volume data are transmitted directly via point to point client links, whereas state and control messages are distributed via the CRB.
The tabletop interface is controlled by its own thread, the Interface Control Unit (ICU). This unit segments the tracked bounding volumes of hands ($2 \times 30^3 cm^3$) out of the probabilistic scene reconstruction and compresses as well as dispatches the data to subscribed clients. The current resolution of each segmented hand volume is 83^3voxel, which leads to an additional bandwidth requirement of approximate 1.5MBit/s per compressed ICU volume data channel.
The actual interface integration into cAR/PE! is finally implemented as a cAR/PE! module which establishes compressed volume data links to the ICU's of all teleconferencing participants. For display purposes the module implements a marching cubes [LC87] shader unit which generates and visualizes mesh models from the received probabilistic hand volumes. Figure 4 represents the general data flow within the system.

5. Results

The presented cAR/PE! hand volume integration results in a completely interactive system implementation. Thereby an ICU with eight cameras (25 fps with 640×480 pixel resolution), a CRB and a cAR/PE! client instance (scene rendering performance approximately 240 fps) are running on a desktop computer with an Intel Q6600 2.4GHz processor and a NVIDIA GeForce GTX280 graphics card.
For the system integration two different mesh generation variants based on the marching cubes shader have been compared. In the first version probabilistic volume data were thresholded and binarized for mesh generation. Thus the resulting dataset is a discrete volumetric scalar field $V(x,y,z) : \mathbb{R}^3 \rightarrow \{0,1\}$ which is equivalent to a volume obtained from a non-probabilistic SfS type approach. Resulting reconstructions are depicted in the first row of figure 5. For the second variant probabilistic volume data $V(x,y,z) : \mathbb{R}^3 \rightarrow [0,1]$ were utilized directly to allow a more accurate isosurface extraction. See bottom row of figure 5 for resulting mesh models.
From a decision theoretic point of view a natural choice of the isosurface threshold is given with a 50% skin-colored foreground likelihood. However in our application an empirical measure of 60% was found to be more feasible, as this resulted in more realistic reconstructions.

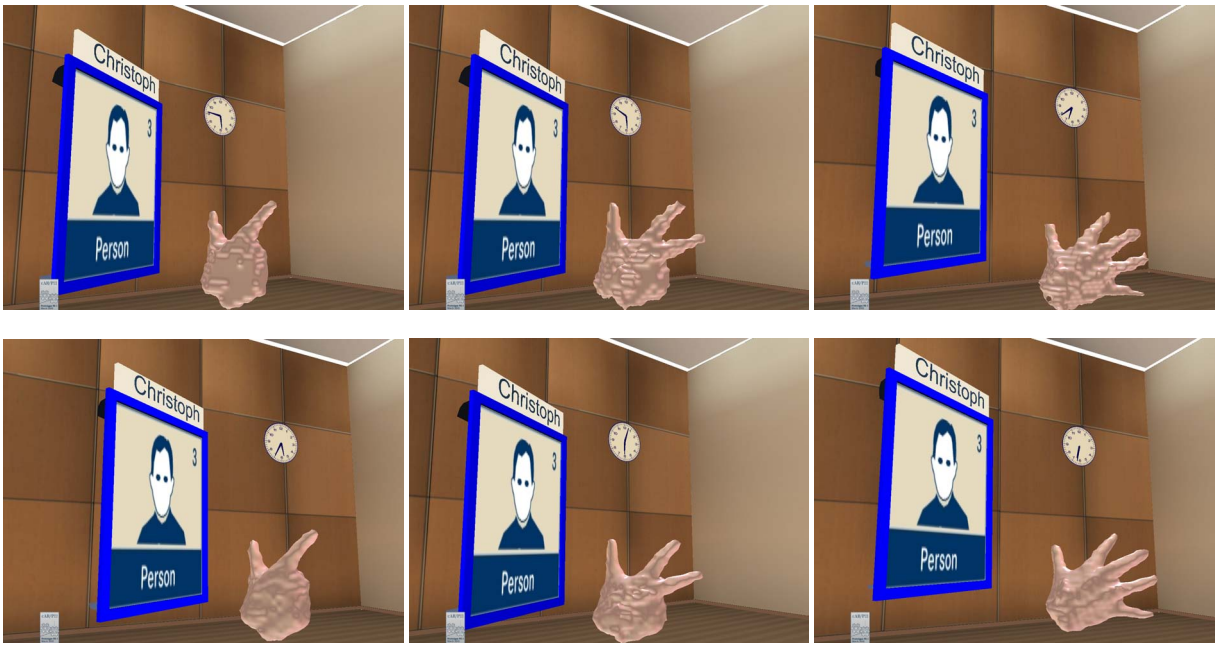

Figure 5: *First row: Non-probabilistic SfS reconstructions. Second row: Probabilistic SfS reconstructions*

6. Conclusion

A new tabletop user interface for volumetric reconstruction and tracking of hands and its integration into a shared virtual environment has been presented to introduce hands as an additional non-verbal communication channel for 3D teleconferencing applications. The use of a GPU for visual hull computation and mesh extraction thereby results in a real-time system implementation with off-the shelf hardware components.

Besides the system integration in its own right, it also has been pointed out that in comparison to non-probabilistic SfS type approaches probabilistic SfS can result in improved isosurface approximations based on probabilistic scalar fields.

References

[dAST*08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. In *ACM SIGGRAPH 2008 papers* (2008), ACM, pp. 1–10. 2

[GM03] GOLDLÜCKE B., MAGNOR M.: Real-time, free-viewpoint video rendering from volumetric geometry. In *in Proc. Visual Computation and Image Processing (VCIP* (2003), pp. 1152–1158. 2

[HD91] HELD R., DURLACH N.: Telepresence, time delay and adaptation. In *Pictorial communication in virtual and real environments*. Taylor & Francis, Inc., Bristol, PA, USA, 1991, pp. 232–246. 1

[HLS04] HASENFRATZ J.-M., LAPIERRE M., SILLION F.: A real-time system for full body interaction with virtual worlds. *Eurographics Symposium on Virtual Environments* (2004), 147–156. 1, 2

[JSR09a] JOHN C., SCHWANECKE U., REGENBRECHT H.: Real-time volumetric reconstruction and tracking of hands and face as a user interface for virtual environments. In *Virtual Reality Conference, 2009. VR 2009. IEEE* (March 2009), pp. 241–242. 1

[JSR09b] JOHN C., SCHWANECKE U., REGENBRECHT H.: Real-time volumetric reconstruction and tracking of hands in a desktop environment. In *Computer Analysis of Images and Patterns* (2009), pp. 1053–1060. 2, 3

[Lau94] LAURENTINI A.: The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell. 16*, 2 (1994), 150–162. 2

[LC87] LORENSEN W. E., CLINE H. E.: Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph. 21*, 4 (1987), 163–169. 3

[LP08] LANDABASO J., PARDAS M.: A unified framework for consistent 2-d/3-d foreground object detection. *Circuits and Systems for Video Technology, IEEE Transactions on 18*, 8 (2008), 1040–1051. 2

[MCA06] MÜNDERMANN L., CORAZZA S., ANDRIACCHI T.: Markerless human motion capture through visual hull and articulated ICP. In *Proc. NIPS Workshop* (2006). 2

[RLK*04] REGENBRECHT H., LUM T., KOHLER P., OTT C., WAGNER M., WILKE W., MUELLER E.: Using augmented virtuality for remote collaboration. *Presence: Teleoper. Virtual Environ. 13*, 3 (2004), 338–354. 1

[SMP05] SVOBODA T., MARTINEC D., PAJDLA T.: A convenient multi-camera self-calibration for virtual environments. *PRESENCE: Teleoperators and Virtual Environments 14*, 4 (2005), 407–422. 2