Home-to-home communication using 3D shadows

Tommi Määttä
Eindhoven University of Technology
Eindhoven, The Netherlands
t.maatta@tue.nl

Aki Härmä
Philips Research Europe
Eindhoven, The Netherlands
aki.harma@philips.com

Hamid Aghajan
Stanford University
Wireless Sensor Networks Lab
aghajan@stanford.edu

ABSTRACT

In some visual communication applications it is not possible or even desired to aim at a photorealistic representation of the remote person. One possibility is to aim at stylized visual representations of remote persons, e.g., as avatars shown on a display device or as shadows in lighting. In this paper we introduce a system for persistent and ambient visual communication based on capture, transmission, and rendering of 3D shadow representations of users. The shape of a person is captured using a distributed camera array, compressed, and transmitted over the network. In the receiving end the shape is projected as a shadow on a surface using a lighting device. We demonstrate that the 3D representation of the shape makes it possible to control the 2D visualization at the receiving end in many interesting ways. For example, when controlled by tracking of the observing user the shadow may create a visual illusion of a 3D shape on the wall.

Keywords

social presence, persistent, ambient, multi-view, 3D shape, real-time, video-based capture

1. INTRODUCTION

In video conference applications it is usually desired to transmit a high-quality video image to imitate face-to-face communication. Some of the most important factors for the face-to-face experience are the preservation of facial expressions and eye contact. In a home video telephony system installed in a PC or a TV these can be typically only preserved when the user is seated close to and facing the terminal device. The *terminal-centricity* of video communication limits the user from doing other things while having the call and causes fatigue in long communication sessions. Consequently, video calls are often short in duration. Due to the flat-rate pricing model of home broadband communication it has become economically feasible to have visual home-to-home connections always open, enabling applications to support *persistent* presence.

The requirement for the face-to-face quality can be relaxed, if the goal is to build a visual communication device supporting primarily the social connectedness and awareness of the other side. Understanding the activities and status of others is achieved through social connections and communications. The idea of the system introduced in this paper is to convey the awareness not by direct interaction or sharing the same physical space, but by mediating it through an open-ended, or *persistent* visual communication channel. Such communication systems have been earlier studied in the collaborative working environments and it has been found that they increase the awareness and support other forms of communication [1]. Users of Instant Messengers are already familiar with indicating their presence and keeping the link on for longer periods of time. Persistence and constantly low bit-rate of data when compared to videoconferencing, are the main characteristics of the system discussed in this paper.

If the goal is not to reproduce the physically accurate visual representation of a remote person there are plenty of alternatives. The remote person can be represented in stylized form as an avatar which may take any shape. It is also possible to map the visual representation to some abstract or symbolic representations such as geometric patterns, movements, colors, or even sounds. Which kind of visual information is transmitted from their home to the other side could be controlled by the users themselves. Based on users preferences to different people at different times the user could have several options for communicating, options on devices (projector, screen) and presentations (shadow, avatar) to be used.

In this paper we mainly focus on using a silhouette representation which is projected on a surface such that it looks like a shadow of a person placed between the light source and the surface. Shadow or silhouette representations of remote people have been proposed earlier in [8][11]. The shadow representation preserves many of the subtle non-verbal signals about our mood, attitudes, and feelings. Therefore, people living far apart from each other might still like to maintain the feeling of living close to each other, like neighbours who see each other daily through the windows of their homes. See [3] for an approach to bring distant homes together by merging the two spaces through a spatial system. The benefits of shadow representation are low bit-rate, low rendering complexity and preservation of privacy while still conveying presence.

Immerscom 2009, May 27-29, 2009, Berkeley, USA.

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6226
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6226

2. PROPOSED SYSTEM

The proposed system combines multi-video processing and free-viewpoint visualization to achieve peripheral awareness and persistent connectedness to a close one who will be portrayed as a shadow-like character. As the users 3D shape is captured, walking around the 3D shape is made possible thus enhancing the effect of the other one being here, at ones home. The silhouettes of the person are extracted in one location with multiple cameras observing the location from different viewpoints, the silhouettes are combined into 3D silhouette, the 3D silhouette or derivation of it is coded to a more compact form, transmitted over the network and rendered in another location interactively based on the observing users position. The flow of data in one home is presented in Figure 1.

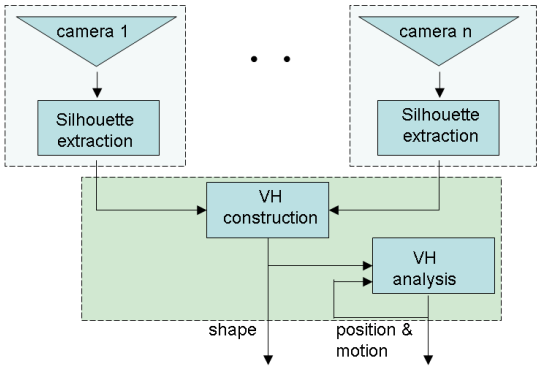

Figure 1: Data processing and flow in one end of the system.

2.1 Multi-view user capture

Human motion capture is to understand human behaviour through detecting, tracking and recognizing users actions. Capture can be performed with only a single view [7], with a pair of cameras [9] or by multiple cameras to achieve even higher robustness. Capture systems are either based on special markers or special suits to mark important spots on the user or they are based solely on the video content. Multi-view, video-based user capture was chosen for reasons inherent to the system designed for the home environment. User capture has to be done unobtrusively to offer ease on long-term use of the daily communication application. In addition to improved robustness a larger area can be covered with multi-view thus giving the user more freedom to move. Silhoutte will also be cleaner as noise outside the actual silhouette is eliminated by forcing correspondences between the multiple views.

The proposed system is designed to detect where the user moves and how fast. The multi-camera system gives video feeds from different viewpoints of the same scene. These feeds are processed locally to a much simpler form by segmenting the user from the background. This process of binary segmentation is hereafter called silhouette extraction. The approach used in our system is based on the following assumptions. The scene is situated indoor and is decently illuminated by ordinary lighting. The cameras are fixed, they do not support zooming, panning or tilting therefore keeping the scene stationary. Only the user will generate movement in the scene. The silhouette extraction is based on using adaptive statistical pixel model to describe the recent history of color at each observed pixel by normal distributions of luminance and green and blue chrominance components [4]. The binary silhouette image is the result of subtracting the model from the video image and thresholding this difference image w.r.t. a predefined threshold. Silhouettes from multiple views are combined into one 3D shape approximation. 3D user capture is performed by extracting certain properties from the shape and these are used in the proposed system for user-movement driven visualization. For real-time performance simple yet intelligent algorithms were needed.

2.2 3D shape reconstruction

There are many approaches for reconstructing 3D shape differing in what image features they use, how the shape is represented and how much detail is preserved of the original object. Shape-from-Silhouette (SFS) [2] technique was used in this paper. SFS is a method for estimating the 3D non-textured shape of an object by using the silhouette images taken from this object from multiple viewpoints. SFS is based on the fact that the separated foreground in form of a silhouette together with the camera viewing parameters can be projected back to the 3D space as a cone that contains the actual foreground object. As every viewpoint forms its own silhouette cone, the intersection of these cones forms the bounding geometry of the actual 3D object. The result is called Visual Hull (VH) [6]. VH is an approximation of the true 3D shape and it depends on the number of views, positions of these views and complexity of the object, see Figure 2.

A straight forward way of implementing SFS is to define a volume-of-interest (VOI), observed by the cameras, within the environment. The VOI is divided into a grid of 3D voxels. Each voxels occupancy is independently tested against the silhouette images from available viewpoints. Occupancy testing is performed by projecting the voxel onto image plane and checking if the voxel projections lie inside the silhouette, see Figure 3. This testing is done for all views. For speeding up the on-line processing the voxel prejections per view were stored in a lookup-table structure for fast access enabling real-time volumetric reconstruction.

2.3 Interactive visualization

In two-way communication the reconstruction of the shape of the local person can also be used to control the visualization of the remote person. In the proposed system the capture of user position and level of activity is performed by analysis of the Visual Hull.

By having the 3D shape, the view can freely rotate around the remote user and go closer or further away. The idea is to give the observing user the power to define the angle and proximity by his natural movements. For example, if local user moves to the right, he will be shown more of the right profile of the remote user. If the observing user moves further away from the screen used for visualization, his distance to the shadow is significantly increased, see Figure 4.

This way by tracking the user's position and showing the corresponding view, the 2D silhouette looks as a 3D object with strong perspective effect. This 3D effect is expected to make the shadow more life-like and further enhance the feeling of the remote user being in the same space with the

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6226
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6226

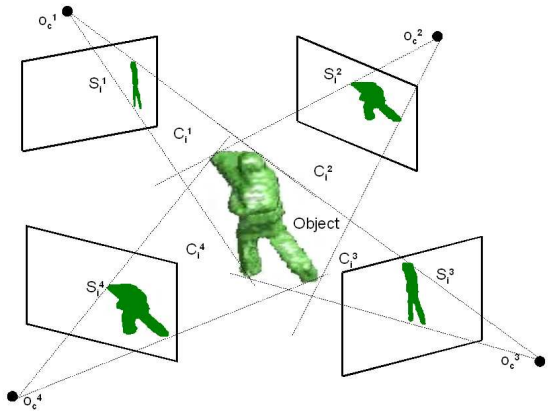

Figure 2: Formation of Visual Hull with four viewpoints: Object forms silhouette image S_i^k on camera k at time instant i.

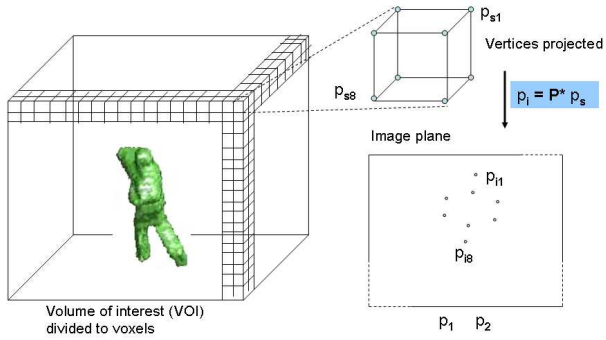

Figure 3: Volume of interest enclosed by a cube formed by regular grid of voxels, whose eight vertices are projected to the image plane. The projections are represented by two uniform points p_1 and p_2 used to test voxels occupancy.

local user. In addition, based on the amount of activity of the remote user the users shadow is shown in different colors, to serve as a status indicator for peripheral awareness.

3. EXPERIMENTS

3.1 Hardware architecture

The proposed system consists of two identical ends, of which one is here used to describe both ends. In the experimental setup built the four cameras situate not exactly at opposite corners to avoid redundant views. As the proposed system is for the home, the cameras used were ordinary mid-price webcameras. All cameras are at same height and their views have no occlusions w.r.t. VOI, see Figure 5. There are four cameras. Based on studies performed on the effect of available silhouette views, by adding a fourth camera a 25 to 140 percent improvement in number of outlier voxels can be expected, and SFS method increases robustness against uncorrelated silhouette noise when increasing the number of views.

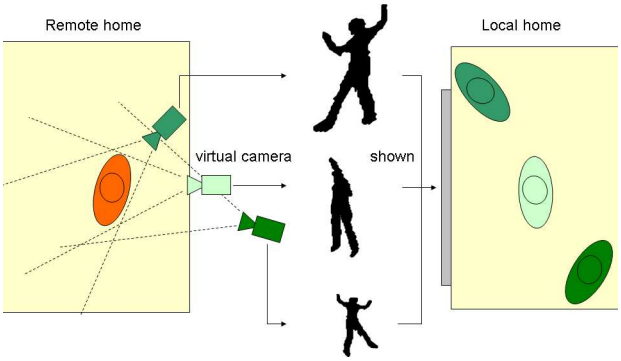

Figure 4: The remote user is shown to the observing user on the right from different angles and proximities based on observing users position. The three user positions and shown viewpoints are illustrated with the same color. The roles of observed and observing are interchangeable.

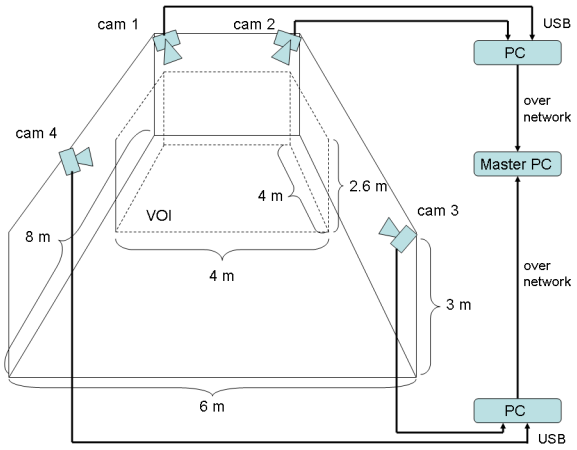

Figure 5: The experimental setup: system architecture and data communications.

The basic blocks of the system are Logitech Fusion USB web-cameras working at 20 fps with YUV 444 format in 320×240 resolution with 73 degrees of horizontal viewing angle, ordinary computers (PCs) with 3 GHz Intel IV processor and Windows XP handling data prosessing/transceiving and the network for inter-PC communication. The two PCs run in real-time two simultaneous applications using the connected cameras in generating silhouette image then sent over the network to the master PC equipped with dedicated graphics card for OpenGL rendering of volumetric reconstruction based on the analysis results on the shape of the observing user. The visualization was performed with a projector in this system, in other scenarios also televisions or screens could be used.

3.2 Data communications and processing

Each silhouette image is sent over the network in one data-packet. Each of the silhouette packets is time-stamped be-

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6226
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6226

fore sending. The average delay between capture of the scene and the volumetric reconstruction visualized locally was between 0.5 to 1.0 second. This is due to the 0.5 second delay of frame buffering and the rest dealing with camera driver and data transmission and coding. Average bandwidth required for a silhouette stream was 11.3 kbytes/s, see Table 1 for bitrates in the 4-camera setup. Such a small local network load is achieved by compressing the binary silhouette data by algorithm called Bzip2. The required data-rate is within the limits for, e.g., Zigbee wireless communication which might have a considerable stature in future home networking. Transmission bitrate of compressed shape, sent between homes, is expected to vary between 7 to 20 kB/s depending on the human shape and its noisiness.

Table 1: Used bytes and corresponding bitrate for a 30 seconds long video sequence with varying amounts of silhouette content.

Camera ID	Kilobytes per 30 seconds	Bitrate kB/s
CAM_1	412	13.7
CAM_2	316	10.5
CAM_3	372	12.4
CAM_4	256	8.5
$CAM_{avg-all}$	339	11.3

3.3 Prerequisite: Calibration

Figure 6: Highlighted calibration object observed from four different viewpoints with the calibration image undistorted for radial distortion.

The aim of our system in using multiple cameras for capturing a scene is to build the three dimensional volumetric reconstruction of the user. To be able to do the reconstruction with the desired metric accuracy the cameras need to be calibrated. Cameras capture the light, emitted by the 3D environment they are focused on, with their image sensor forming the image plane. Through camera calibration one is able to determine which ray in 3D space is associated with which pixel on the 2D image. There are several methods for multi-camera calibration which use different calibration objects, treat the cameras individually or together [10] and are fully-automatic or require manual work.

The classical calibration approach with intrinsic calibration by GML toolbox for radial distortion was used together with the pinhole camera model. The pinhole model is the simplest approximation of camera geometry used in computer vision. The calibration was performed by capturing the projections of known 3D scene points, points whose x,y and z world-coordinates were fixed with a calibration object, on the image plane, see Figure 6, and by using these 3D-to-2D point correspondencies through DTL-method [5] to calculate the projection matrices defining the mapping from 3D scene to 2D image for each camera independently.

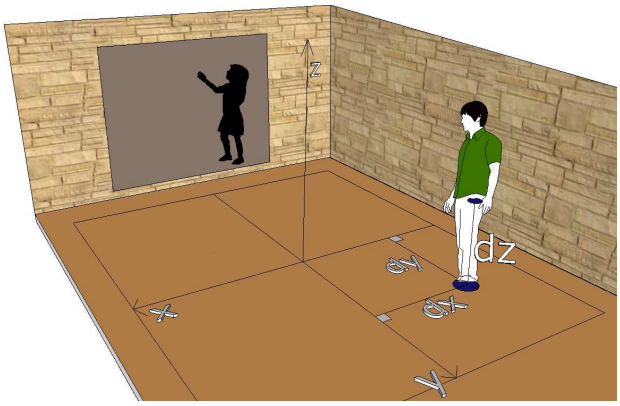

Figure 7: Definition of VOI within environment and based on the fixed world coordinate system (x,y,z) the tracking of user position within the VOI.

3.4 Interactive visualization

In the proposed system the user has the option to move around the shadow of the remote person maximum of 120 degrees to each side, see Figure 9, have a frogs view from below, by crouching, or birds view from above, by raising arms, and go right in front or significantly away from the shadow, see Figure 10. The users position is tracked through shape analysis by calculating the centroid of the voxels forming the VH and using this centroid as the center of mass of the user. Based on the centroid the 3D shape can be kept stationary, thus keeping the shadow on the same spot even as the observed user walks around in the VOI. Based on users natural movement the corresponding view is shown. See Figure 7 for an example how the viewer position is determined and Equations 1, 2 and 3 for how these values determine the viewpoints horizontal and vertical rotation γ_{hor} and γ_{vert} and proximity $zoom$.

$$\gamma_{hor}(dx) = \frac{dx}{Dim_x} \times HRA \quad (1)$$

$$zoom(dy) = \frac{dy}{Dim_y} \times DoC \quad (2)$$

$$\gamma_{vert}(dz) = (dz - CoM) \times s \quad (3)$$

where HRA is the maximal horizontal rotation angle (120 degrees), DoC is the maximal displacement of camera closer to or further away from the center of VOI (200 cm), CoM is the pre-defined hight of center of mass in ordinary standing posture, (60cm), Dim is the VOI dimension in centimeters from the center of the VOI to the border of the VOI either in x (200 cm) or y (200 c,) direction and s is the scaling factor (3) for z-wise movement. The values given for the variables here are the ones used in this proposed system. These variables are used in determining the displacement of the virtual camera from its reference position and orientation, which is initially defined at the end of negative y-axis, just outside the VOI with orientation towards the center of VOI. For creating the most life-like effect the shadows should be projected to a surface. This approach is expected to create a smoother and more life-size visualization than would be achievable without using projection techniques.

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6226
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6226

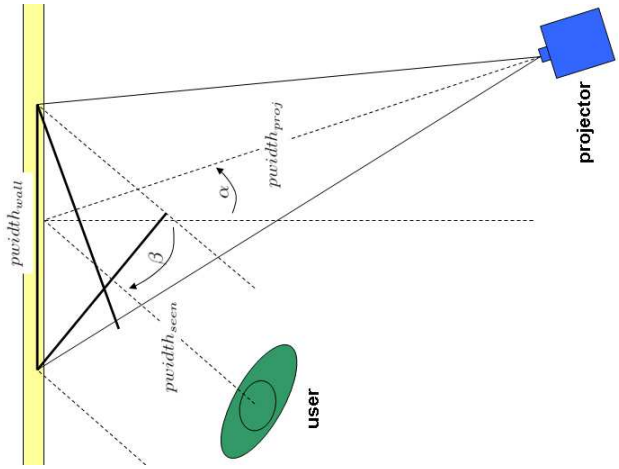

Figure 8: Geometric compensation for projector and observer positions.

3.5 Geometric corrections

For projection to retain its real world feeling by not being deformed by geometric distortions, the projection should also take into account at least the most basic geometric challenges, the horizontal and vertical tilting of the projection surface w.r.t. the projecting device as well to the user, see Figure 8 for illustration of the problem. If the projection surface has a tilting angle α away from the perpendicular orientation against the projector, the relation between original projection width $pwidth_{orig}$ and actual projection width $pwidth_{wall}$ is:

$$\frac{pwidth_{orig}}{pwidth_{wall}} = \cos\alpha \tag{4}$$

Thus for the projection on the wall to have the width of the original, the projection width $pwidth_{proj}$ has to be corrected for angular tilt α before projecting on wall as follows:

$$pwidth_{proj} = pwidth_{orig} \times \cos\alpha \tag{5}$$

The same correction for redefining the width dimension of the projected shadow can also be performed for the vertical hight w.r.t. vertical tilt of the surface.

For compensating also for the effect of users position the similar correction in horizontal direction is required. Otherwise as the local user moves more to one side to see more of the side profile the projection he sees has shrunk in horizontal direction. To compensate for this the projection width on the wall has to be increased accordingly.

$$pwidth_{seen} = \cos\beta \times pwidth_{wall} \tag{6}$$

In order to eliminate the cosine term and thus retain the same width for the observing user the $pwidth_{wall}$ has to be streched as follows:

$$pwidth_{wall_{new}} = \frac{pwidth_{wall}}{\cos\beta} \tag{7}$$

When combined both the shrinking for the angular projection α and expanding for the angular observing β the projection process has to be corrected in horizontal direction as follows:

$$pwidth_{proj} = pwidth_{orig} \times \frac{\cos\alpha}{\cos\beta} \tag{8}$$

4. CONCLUSIONS

The aim was to create a virtual shadow presentation of the users to elicit social connectedness and awareness between the two homes. By using shadow presentation users privacy is more secured and system requirements for data processing and communications can be kept in managable level. The use of multiple cameras makes it possible to have 3D shadows thus increasing the effect of sharing the same space as visualization can be made to match the users natural movements around the room. This is expected to offer more immersive experience of the other being-here, as the user can freely move around the remote persons shadow like the remote person would actually inhabit the same space. To not break the illusion of being-here the geometric distortions introduced by projectors and observers non-perpendicularity against the projection surface are taken into account in the projection process to maintain the right perceptual width of the 3D shadow.

For the real-time system a compromise between more advanced algorithms and performance speed had to be made. As the representation of the user was selected to be a shadow, basic background subtraction based silhouette extraction and as extension of the silhouette concept the Shape-from-Silhouette method were chosen for the task. One benefit of the SFS method is a cleaner shadow as it is generated from the 3D shape. Interaction with the visualization was based on giving the observers movements the ability to control what is shown with what size and from which angle. Movements were tracked by simple and fast calculation of the center of mass of the users shape.

On future work a more sparse camera setup around the home is used, still providing continuous, smooth representation of the user. The use of avatars, body motion based gaming and more advanced interfaces based on more accurate motion capture are possible future extensions of the system proposed.

5. REFERENCES

[1] M. Apperley, L. McLeod, M. Masoodian, L. Paine, M. Phillips, B. Rogers, and K. Thomson. Use of video shadow for small group interaction awareness on a large interactive display surface. In *Proc. Fourth Australasian User Interface Conference (AUIC2003)*, 2003.

[2] K. M. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, volume 2, pages 714 – 720, June 2000.

[3] K. Grivas. Digital selves: Devices for intimate communications between homes. *Pers. Ubiquit. Comput.*, 10:66–76, 2006.

[4] H. Han, Z. Wang, J. Liu, Z. Li, B. Li, and Z. Han. Adaptive background modeling with shadow suppression. In *Proc. of Intelligent Transportation Systems*, pages 720–724, 2003.

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6226
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6226

(a) Frame 1 (b) Frame 31 (c) Frame 68 (d) Frame 98

Figure 9: Local users shape is shown on the front screen and two silhouette masks used to reconstruct the shape on the screen behind on the right all rendered with real-time software. When user clearly moving, activity level increased, the VH is rendered in color. Frames a-c: user takes a step to the right, frame d: user hops back to center.

(a) Frame 77 (b) Frame 200 (c) Frame 294 (d) Frame 491

(e) Frame 694 (f) Frame 797 (g) Frame 894 (h) Frame 942

Figure 10: Simulated action clip of the local observing user walking around the stationary 3D shadow projected on a 2D surface. The first row: observing user goes left and then down to frogs view by crouching, the second row: user goes right and then up to birds view by reaching to the ceiling.

[5] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

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

[7] D. Lee, , and Y. Nakamura. Motion capturing from monocular vision by statistical inference based on motion database: Vector field approach. *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 617–623, Oct. 29 2007-Nov. 2 2007.

[8] Y. Miwa and C. Ishibiki. Shadow communication: system for embodied interaction with remote partners. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 467–476, New York, NY, USA, 2004. ACM.

[9] R. Plänkers and P. Fua. Tracking and modeling people in video sequences. *Comput. Vis. Image Underst.*, 81(3):285–302, 2001.

[10] T. Svoboda, D. Martinec, and T. Pajdla. A convenient multi-camera self-calibration for virtual environments.

[11] S. Yasuda, S. Hashimoto, M. Koizumi, and N. Okude. Teleshadow: feeling presence in private spaces. In *SIGGRAPH '07: ACM SIGGRAPH 2007 sketches*, page 15, New York, NY, USA, 2007. ACM.

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6226
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6226