International Journal of Innovative Computing, Information and Control
Volume **10**, Number **2**, April **2014**

ICIC International ©2014 ISSN 1349-4198
pp. 495–503

MODEL-BASED APPROACH FOR 3D POSE ESTIMATION WITH APPLICATION TO AUGMENTED REALITY

Huei-Yung Lin[1], Chin-Chen Chang[2,*] and Ting-Wen Chen[1]

[1]Department of Electrical Engineering and
Advanced Institute of Manufacturing with High-Tech Innovations
National Chung Cheng University
No. 168, University Road, Min-Hsiung, Chiayi 62102, Taiwan

[2]Department of Computer Science and Information Engineering
National United University
No. 1, Lien-Da, Kung-Ching Li, Miaoli 36003, Taiwan
*Corresponding author: ccchang@nuu.edu.tw

Received January 2013; revised May 2013

Abstract. *In this paper, we present a model-based approach for 3D pose estimation with an application to augmented reality. A monocular camera is used to acquire the images of the user's motion for 3D pose estimation. In the proposed technique, a graphical 3D human model is first constructed. Its projection on a virtual image plane is then used to match the silhouettes obtained from the image sequence. By iteratively adjusting the 3D pose of the graphical 3D model with the physical and anatomic constraints of the human motion, the human pose and the associated 3D motion parameters can be uniquely identified. The obtained 3D pose information is then transferred to the reality processing subsystem and used to achieve the marker-free interaction in the augmented environment. Experimental results demonstrate the feasibility of the proposed approach.*

Keywords: Augmented reality, Pose estimation, Cost function, Perspective projection

1. **Introduction.** One important issue of augmented reality (AR) is to design an interface for seamless interaction between the virtual objects and the real world. Researchers have proposed various types of techniques to increase the interactivity in the augmented space [19]. In its early development, 3D AR interfaces focus on providing spatially seamless interaction with special-purpose input devices. Recent advances on tangible AR interfaces, on the other hand, emphasize the use of physical objects as tools for projecting the virtual objects onto the surfaces [1]. Nevertheless, both approaches are not capable of "tool-free" interaction only with the bare hands.

The objective of this work is to develop an AR interface with marker-less human body interaction. It consists of 3D motion capture of the human body and the processing of 3D human poses for augmented reality applications. Although there exist some approaches for human computer interaction (HCI) using commercially available motion capture system, the underlying technologies are usually expensive, obtrusive, and require the users to wear special markers for joint or body parts identification [3,5]. The proposed AR system uses only a video camera and a head mounted display (HMD) as the input and output devices, respectively. Since the application domain is less restrictive with only a single camera, especially for low-cost AR systems, the human pose estimation from monocular image capture has become an emerging issue to be properly addressed.

The major difficulties of monocular human pose estimation include the high dimensionality of the pose configuration space, lacking of depth information, self-occlusion, and perspective effect of the camera model [9]. These problems are caused by the inherent ambiguity in 3D to 2D mapping, and have to be resolved with additional constraints [2]. In this paper, we present a model-based method for marker-less 3D pose estimation from a single camera view. The contributions of the proposed approach are as follows: (a) The modified 3D model is created and adjusted such that its projection is aligned with the image silhouette to estimate the pose of a subsequent image; (b) We propose a cost function to facilitate the shape fitting and fast movement of the body part; (c) The high dimensionality of search space for alignment and the ambiguities in 3D pose reconstruction are reduced by anatomic and physical constraints of human motion, as well as the appearance information from the intensity image; (d) The resulting pose parameters and an articulated graphical 3D model are then used for full body interaction in the augmented environment.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 presents the proposed approach. Experimental results are given in Section 4. An application is given in Section 5. Section 6 provides conclusions and future works.

2. Related Works.

Some techniques with gesture or finger tracking were proposed for augmented desk applications. Although there is no need for specific tools, the interaction is still restricted to 2-dimensional or requires markers for vision-based tracking [7,17]. Marker-free interaction for AR interfaces is only adopted for the pose estimation between the objects and the camera [6,11].

Loy et al. [12] presented an approach for the monocular 3D reconstruction of human action in long action sequences. They adopted a keyframe-based approach to estimate the 3D pose of human motion in sports sequences. The 3D reconstruction is derived using a video footage, which is not capable of on-site processing. Chen et al. [4] presented a method to reconstruct 3D human motion parameters using image silhouettes. A weighted-XOR cost metric was used for object alignment, shape fitting and motion tracking. Their approach was used to synthesize motion from archival footage.

Nguyen et al. [13] proposed a 3D human capture system for a mixed reality environment. For a subject, nine cameras are set up to capture images. They constructed the 3D images of the subject using a shape-from-silhouette technique. Their system shows many technologies in human computer interaction. Hahn et al. [8] presented an approach for 3D pose estimation and tracking of human body parts using the shape flow algorithm. They integrated the shape flow algorithm into a tracking system and examined its suitability for tracking human body parts in 3D.

Ye et al. [18] presented a system for estimating body pose configuration from a single depth image. Their approach combines both pose detection and pose refinement which can accommodate both pose difference and body-size difference. Simo-Serra et al. [16] proposed an approach for 3D human pose estimation from noisy observations. They introduced a sampling technique to propagate the noise from the image plane to the shape space. This can produce a set of ambiguous 3D shapes and disambiguation is then performed by setting kinematic constraints.

3. 3D Pose Estimation Approach.

The proposed 3D pose estimation algorithm is based on the comparison of the projected graphical 3D human model and the captured image. An articulated graphical human model is created and adjusted iteratively to align with the input image based on the silhouette and color information of the object region.

3.1. Modeling a human body.

Due to the lack of 3D information from the input images, a graphical 3D model of the human body has to be generated for 3D pose estimation. Most articulated 3D human model is generated with a number of rigid body parts and joints. The number of degrees of freedom is thus a key factor to the construction of the graphical 3D human model.

The 3D human model is created using OpenGL library. It consists of 10 body parts, 9 joints and 22 degrees of freedom. The body parts are represented by spheres, ellipsoids and cylinders. Different colors are assigned to different body parts to facilitate the pose recognition process. Since the graphical 3D model is projected to a virtual image plane for template matching and alignment with the real scene image, the object regions in both images should have a similar size and orientation. Thus, a canonical 3D human model is created first, and an on-site model initialization process is carried out for the user in the scene.

3.2. **Silhouette-based pose estimation.** Given the foreground silhouette image of a human body, the associated pose is estimated by minimizing the difference between the silhouette in the real scene image and the projection of the 3D model on the virtual image. To find the best pose of the graphical model which matches the human pose, suitable metric and cost functions should be provided. Chen et al. [4] presented a Euclidean distance transform approach to calculate the pixel-wise distances between the real and virtual image silhouettes. A cost function defined by the summation of pixel-wise distances was then used to adjust the 3D model. Since both of the entire images were used for comparison, the computational cost was relatively high and the results tended to converge to a local minimum.

Different from their whole silhouette matching approach, we propose a multipart alignment technique. The body parts in the real and 3D model projected silhouette images are compared and adjusted one by one using a core-weighted XOR operation. The pixel difference is processed locally for each body part so that better alignment results with less computation can be achieved. Furthermore, it is suited for articulated 3D models with a number of joints and rigid body parts.

To perform the multi-part pose estimation, the most significant body part, i.e., the trunk, is identified first. It is the central part of the foreground silhouette, connecting the rest of the body parts. Once the trunk is extracted, the regions of the head, upper and lower limbs can be easily acquired. To identify the trunk, an erosion operation is first carried out recursively to remove the limbs in the foreground silhouette. The projected 3D model is then overlaid on the center of the silhouette, followed by a 3 DOF rotation to minimize the difference between the trunk of the foreground silhouette and the 2D projection of the 3D model.

After the 3D pose of the trunk is derived, the upper and lower limbs are processed in the order of arms, wrists, thighs and legs. The identification of the limbs is carried out by comparing the foreground-background ratio of the graphical model. For these body parts, we define 2 DOF for rotation (without the rotation along their main axes). As shown in Figure 1(a), a limb is capable of rotating 360° on the image plane (represented by the angle θ) and 180° off the image plane (represented by the angle φ). When searching the pose of a limb, the angle θ is identified first by rotating the corresponding body part in the 3D model. Several initial orientations separated by 45° are used to avoid the full range search and speed up the alignment process. The angle φ is then calculated by detecting the size change of the projected body part due to the foreshortening as shown in Figure 1(b).

3.3. **Appearance constraints.** The foreground silhouette does not provide the self-occlusion information of the object. To make the pose estimation algorithm more robust, one commonly used approach is to take the color and edge information of the object into

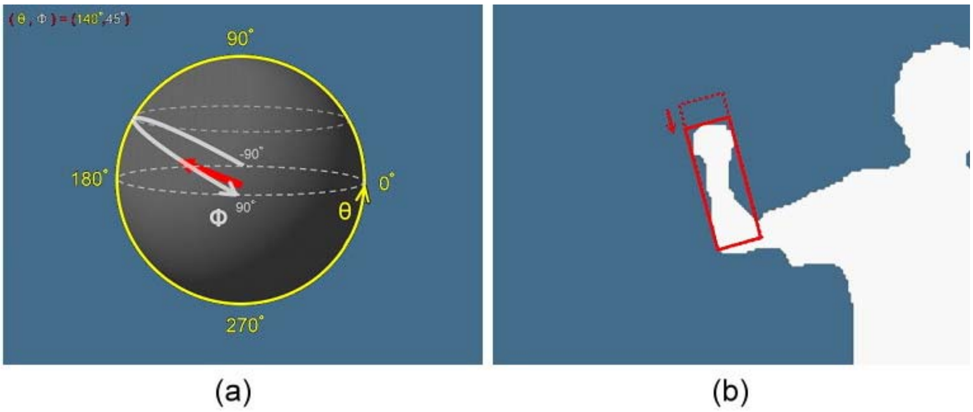

Figure 1. Silhouette matching between images. (a) 2 rotation DOF of a limb; (b) changes due to foreshortening.

account [15]. By extracting the individual parts of the object, the associated poses can then be identified.

The physical and kinematic constraints are enforced on the motion of an initial 3D human model [14]. Thus, self-occlusions of the body parts need not be properly extracted prior to the pose estimation process. One can identify the end of each limb, and combine with the above constraints to estimate the 3D human pose up to a projective ambiguity. In this case, each body part is considered as a link of the human skeleton model, and the positions of the hands and feet will be identified within the foreground silhouette.

4. **Implementation and Results.** Figure 2 shows the experimental environment. The client (motion capture) and server (reality processing) systems are illustrated in the left and right images, respectively. The motion capture subsystem consists of a PC with Intel Core 2 Quad processor, a Logitech QuickCam Sphere AF camera with image resolution of 1600×1200, and a green background for facilitating the foreground human segmentation. The camera is connected to the PC via USB 2.0 interface with the frame rate of 30 fps. The reality processing subsystem consists of a PC with Intel Pentium D processor, a Cyberman HMD (GVD-310A) attached with a Logitech QuickCam Pro 5000 camera, and a marker for creating the program initialization interface. The input and output image resolutions of the camera and HMD are 640×480 and 800×255, respectively. The distance between the user and the camera for motion capture is about 3.8 meters, and the dimension of the marker is 70×70 cm^2.

The first step of model-based pose estimation is to extract the image silhouette of the foreground region. For a given background image sequence, the intensity distributions of each image pixel are calculated for the red, blue, green, and hue channels. Two times of the standard deviations for each pixel are used to model the channel intensity ranges for segmentation. Since the RGB model is more sensitive to the illumination change and the HSV model is better for color discrimination, we use the hybrid approach to derive the robust background model for foreground segmentation. To make the resulting silhouette image more suitable for model-based template matching, morphological operations and median filtering are carried out to remove the holes inside the foreground region. Although the foreground region is not perfectly extracted in most cases, the noise presented in the image is not significant enough to affect the subsequent pose estimation stage. This result also suggests that the sophisticated segmentation algorithm is not always required for the proposed pose estimation technique.

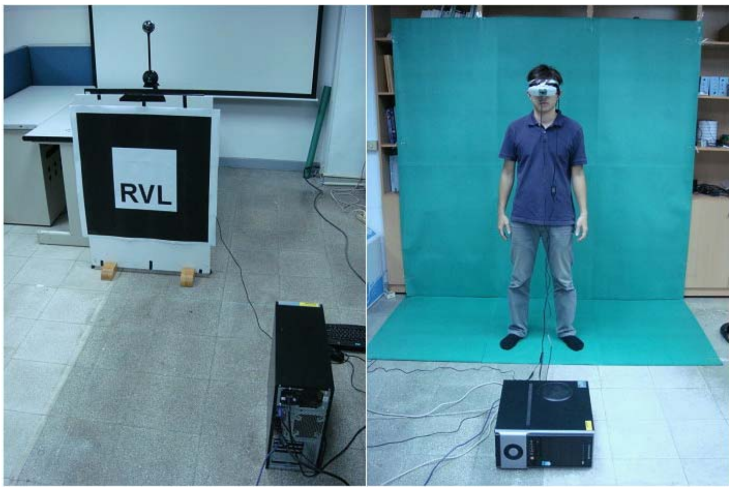

FIGURE 2. The experimental environment

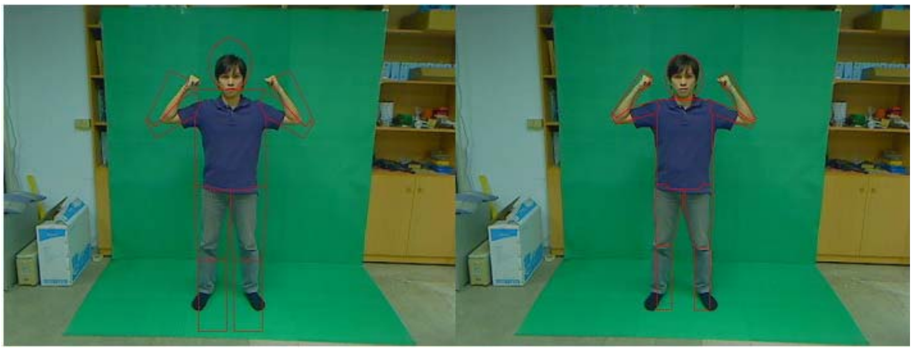

FIGURE 3. On-site model initialization

The location of a body part within the foreground silhouette is identified by the color information. The most significant feature in the foreground region is the skin color of the hands. To extract the associated color model, a simple and robust method is to detect the face color in the initialization stage. The head in the foreground silhouette is first identified using model-based template matching. Histogram analysis on the head region is carried out to separate the skin and hair colors. The threshold for the face color segmentation is then used to extract the hand regions in the foreground.

Figure 3 shows the 3D models prior to and after the on-site model initialization step. As shown in Figure 4(a), the GUI provides the program control by real-time full body interaction for selecting the available options. We initiate several balls at the marker position and make them bounce in the environment with different velocities and directions as shown in Figure 4(b). The balls will be bounced back if they are hit by the user according to the vision-based 3D pose estimation results. Otherwise, the balls will disappear when they pass beyond the user location. Figure 4(c) shows an image capture of marker-less interaction with the virtual objects.

Table 1 shows the processing time of matching body parts in milliseconds. The proposed approach can process three to seven images per second in average for matching body parts. Hence, we can apply the proposed approach for real-time interactive applications.

5. **An Application: Augmented Reality System Architecture.** The proposed marker-less augmented reality interface consists of two subsystems: one is for 3D motion capture of the human body, and the other is for the processing of augmented reality

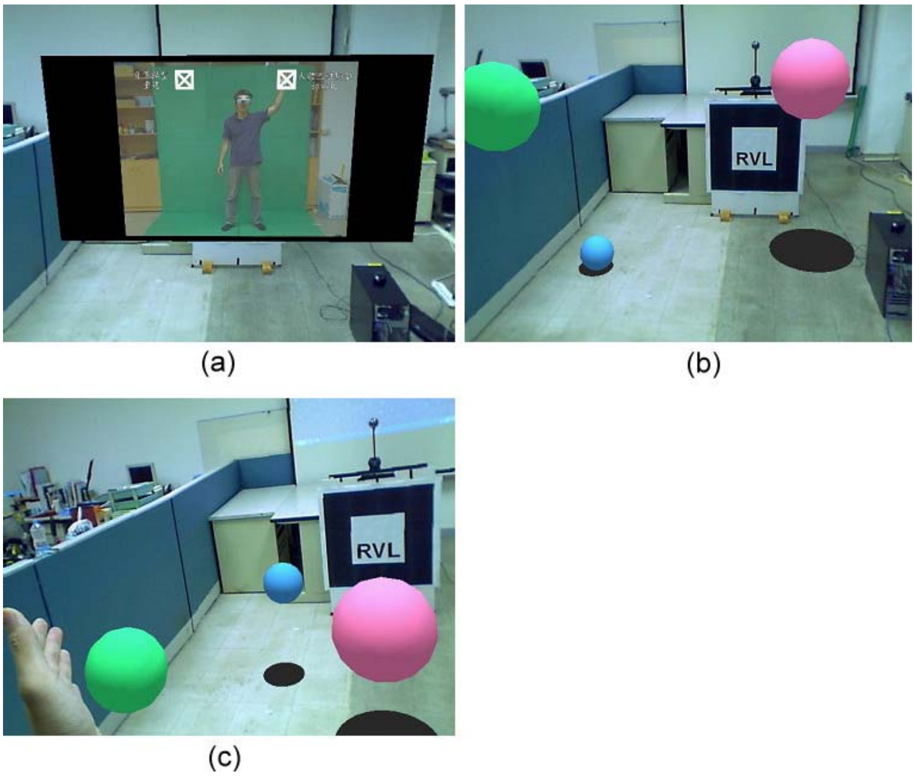

FIGURE 4. The results of augmented reality with marker-less human body interaction. (a) The GUI for system initialization; (b) the augmented environment; (c) interaction with the virtual objects.

TABLE 1. Processing time of matching body parts in milliseconds

Body Parts	Best Time	Worst Time	Average Time
Trunk	4	16	5
Head, Upper Limbs, Thighs	57	123	79
Lower Limbs, Legs	71	166	103
Total Time	132	305	187

applications. The 3D pose estimation and the augmented image synthesis are accomplished by two separate computers communicated via a local area network. The input for the marker-less human body interaction is the image sequence captured by a video camera, and the output for the augmented reality system is through the head mounted display.

The data transmission between the motion capture and the reality processing subsystems is developed on the TCP/IP protocol using WinSock interface. It includes the information requests, the transmission of the 3D pose parameters and the captured image sequence. In general situations, the reality processing subsystem requests the immediate 3D pose information from the motion capture subsystem. Thus, the former and the latter computer systems are defined as the client and the server, respectively.

As described previously, the motion capture and reality processing subsystems are in charge of heavy image processing tasks with the high frame rate constraint. To reduce the data streaming overhead, multi-threading using POSIX thread libraries is adopted on both computers. A single thread is used exclusively for the data transmission task. Similar to most augmented reality systems, marker identification and tracking are used

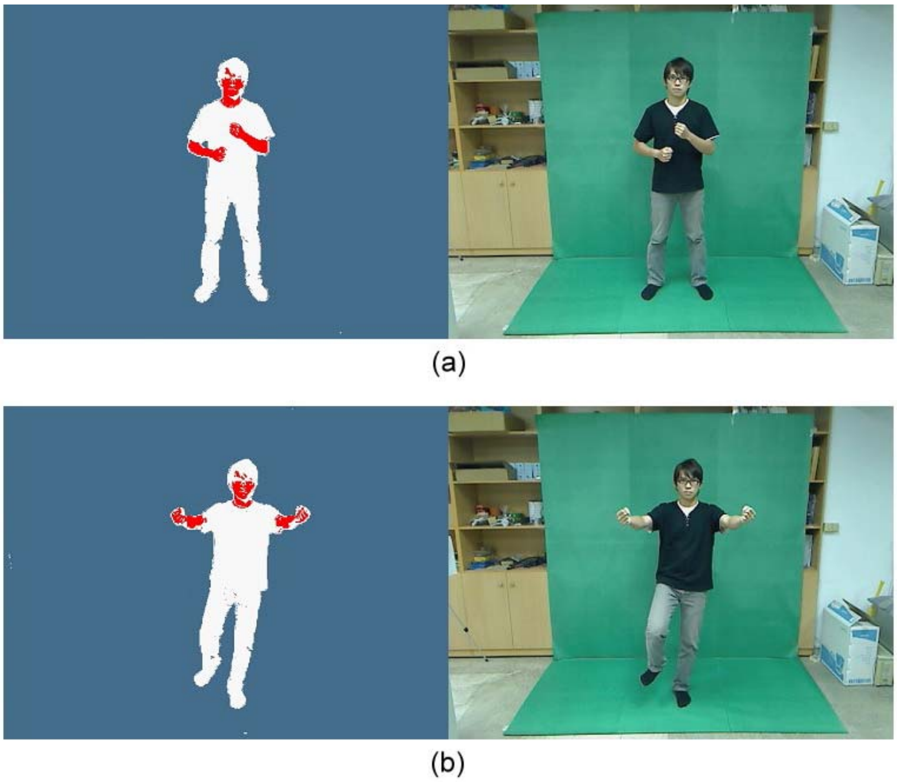

Figure 5. 3D pose estimation results. The left figures show the foreground silhouettes and skin color detection. The original images with estimated 3D graphical model overlay are shown in the right figures. (a) User pose with self-occlusion; (b) user pose with foreshortening.

for manipulating the virtual objects in this work. We adopt the ARToolKit for marker tracking and HMD calibration in our implementation [10]. However, the marker in the application scenario is mainly used for generating a virtual interface at the program initialization stage. The interaction between the human and the virtual objects can be completely marker-free. The motion capture information is used to register the 3D human pose to the virtual environment. Through the HMD, the real scene and virtual objects are simultaneously accessible for manipulation.

In the reality processing subsystem, the real world scene is captured by the camera attached on the HMD. The images are transferred to the client PC for virtual objects overlay, and then transferred back to the HMD for display. At the system initialization stage, a user-friendly interface is displayed at the marker position. For more accurate comparison between the foreground silhouette and the projection of the 3D model, it is clear that there exists a similarity transformation between the graphical 3D model and the real human body. That is, the dimension of each body part should be identical up to a unique scale factor for both the graphical model and real object. Since only one canonical 3D model is created for all situations, it has to be modified for different users according to their shapes. We refer to this step as an “on-site model initialization”.

To perform the on-site model initialization, an image of the user with a pre-defined pose is captured. After extracting the foreground object region, the run-length encoding is used to scan the silhouette image and derive the features of the body parts. Since the initial human pose is pre-defined, the dimension and orientation of each body part in the 3D model can be easily identified by the image features such as head, shoulder and elbow.

In our application scenario, the pose estimation results and the augmented reality with full body interaction can be illustrated separately. Since the 3D motion parameters are essential to the proposed augmented reality system, we have tested several image sequences with various types of human postures. Two results of non-trivial tasks with the arms occluding the body silhouette and severe foreshortening of the arms are shown in Figures 5(a) and 5(b), respectively. In both cases, the 3D poses are correctly identified with the assistance of skin color.

6. **Conclusions.** We have presented a monocular vision-based human pose estimation technique and its application to augmented reality. An articulated graphical human model is created for 3D pose estimation of each body part. The foreground silhouette and color information are used to evaluate the 3D parameters of the graphical 3D model under the anatomic and physical constraints of the human motion. Experimental results of marker-less human body interaction in the augmented environment are presented.

Acknowledgment. This work is partially supported by the National Science Council of Taiwan under Grant NSC-96-2221-E-194-016-MY2. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier and B. MacIntyre, Recent advances in augmented reality, *Computer Graphics and Applications*, vol.21, pp.34-47, 2001.

[2] C. Bregler, J. Malik and K. Pullen, Twist based acquisition and tracking of animal and human kinematics, *Int. J. Comput. Vision*, vol.56, pp.179-194, 2004.

[3] J. Chan, H. Leung, K. T. Tang and T. Komura, Immersive performance training tools using motion capture technology, *Proc. of the 1st International Conference on Immersive Telecommunications*, pp.1-6, 2007.

[4] Y. Chen, J. Lee, R. Parent and R. Machiraju, Markerless monocular motion capture using image features and physical constraints, *Computer Graphics International*, pp.36-43, 2005.

[5] P. T. Chua, R. Crivella, B. Daly, N. Hu, R. Schaaf, D. Ventura, T. Camill, J. Hodgins and R. Pausch, Training for physical tasks in virtual environments: Tai Chi, *Proc. of the IEEE Virtual Reality*, 2003.

[6] A. I. Comport, E. Marchand, M. Pressigout and F. Chaumette, Real-time markerless tracking for augmented reality: The virtual visual servoing framework, *IEEE Transactions on Visualization and Computer Graphics*, vol.12, pp.615-628, 2006.

[7] K. Dorfmuller-Ulhaas and D. Schmalstieg, Finger tracking for interaction in augmented environments, *IEEE and ACM International Symposium on Augmented Reality*, pp.55-64, 2001.

[8] M. Hahn, L. Kruger and C. Wohler, Spatio-temporal 3D pose estimation and tracking of human body parts using the shape flow algorithm, *International Conference on Pattern Recognition*, pp.1-4, 2008.

[9] N. R. Howe, Silhouette lookup for monocular 3d pose tracking, *Image Vision Comput.*, vol.25, pp.331-341, 2007.

[10] H. Kato and M. Billinghurst, Marker tracking and hmd calibration for a video-based augmented reality conferencing system, *Proc. of the 2nd IEEE and ACM International Workshop on Augmented Reality*, pp.85-94, 1999.

[11] T. Lee and T. Hollerer, Multithreaded hybrid feature tracking for markerless augmented reality, *IEEE Transactions on Visualization and Computer Graphics*, vol.15, pp.355-368, 2009.

[12] G. Loy, M. Eriksson, J. Sullivan and S. Carlsson, Monocular 3d reconstruction of human motion in long action sequences, *Proc. of ECCV, LNCS*, vol.3024, pp.442-455, 2004.

[13] T. H. D. Nguyen, T. C. T. Qui, K. Xu, A. D. Cheok, S. L. Teo, Z. Zhou, A. Mallawaarachchi, S. P. Lee, W. Liu, H. S. Teo, L. N. Thang, Y. Li and H. Kato, Real-time 3D human capture system for mixed-reality art and entertainment, *IEEE Transactions on Visualization and Computer Graphics*, vol.11, pp.706-721, 2005.

[14] H. Ning, T. Tan, L. Wang and W. Hu, Kinematics-based tracking of human walking in monocular video sequences, *Image Vision Comput.*, vol.22, pp.429-441, 2004.

[15] R. Poppe, Vision-based human motion analysis: An overview, *Comput. Vis. Image Underst.*, vol.108, pp.4-18, 2007.

[16] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras and F. Moreno-Noguer, Single image 3D human pose estimation from noisy observations, *IEEE Computer Society Conference in Computer Vision and Pattern Recognition*, 2012.

[17] T. Starner, B. Leibe, D. Minnen, T. Westyn, A. Hurst and J. Weeks, The perceptive workbench: Computer-vision-based gesture tracking, object tracking, and 3d reconstruction for augmented desks, *Machine Vision and Applications*, vol.14, pp.59-71, 2003.

[18] M. Ye, X. Wang, R. Yang, L. Ren and M. Pollefeys, Accurate 3D pose estimation from a single depth image, *IEEE International Conference on Computer Vision*, pp.731-738, 2011.

[19] F. Zhou, H. B. L. Duh and M. Billinghurst, Trends in augmented reality tracking, interaction and display: A review of ten years of ismar, *Proc. of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pp.193-202, 2008.