Real-Time Human Pose Detection and Tracking for Tele-Rehabilitation in Virtual Reality

Štěpán Obdržálek[a], Gregorij Kurilo[a], Jay Han[b], Ted Abresch[b] and Ruzena Bajcsy[a]
[a] *University of California, Berkeley, USA, {xobdrzal, gregorij, bajcsy@eecs.berkeley.edu}*
[b] *University of California at Davis Medical Center, Sacramento, USA*
{jay.han@ucdmc.ucdavis.edu, tabresch@ucdavis.edu}

Abstract. We present a real-time algorithm for human pose detection and tracking from vision-based 3D data and its application to tele-rehabilitation in virtual environments. We employ stereo camera(s) to capture 3D avatars of geographically dislocated patient and therapist in real-time, while sending the data remotely and displaying it in a virtual scene. A pose detection and tracking algorithm extracts kinematic parameters from each participant and determines pose similarity. The pose similarity score is used to quantify patient's performance and provide real-time feedback for remote rehabilitation.

Keywords. Tele-rehabilitation, human pose estimation and tracking, stereo vision

Introduction

Virtual reality based rehabilitation provide controllable environment in which various training tasks can be planned. Simultaneous use of different sensory devices provides data for real-time feedback and off-line analysis of patient's performance. In this paper we focus on a human guided program where patient and therapist are geographically dislocated. We present a real-time human pose detection, which is used to provide objective means for comparison of patient's performance with respect to the therapist's movement. We have built our framework upon our research work in teleimmersion [11]. The teleimmersion technology allows remote users to interact through 3D virtual environment as if they share the same physical space. In the past we have demonstrated benefits of immersive training in teaching of Tai-Chi movements [1], dance [7] and remote interaction of basketball players in wheelchairs [2]. This project extends our pilot work presented in [5] where a stepping-in-place task was used to evaluate movement patterns of healthy individuals from a stereo camera. We have now included automatic marker-less pose-detection and full-body tracking, which will be applied in different rehabilitation scenarios. In this paper we provide details on the algorithm and describe its application for tele-rehabilitation.

1. Framework

In our teleimmersion framework, geographically displaced users are visualized in a shared virtual environment via their real-time 3D avatars generated from stereo data. We have implemented an adaptive meshing algorithm which provides fast and robust computation of surfaces from stereo images. The stereo reconstruction can achieve the frame rate of more than 25 frames per second (FPS) on images with the resolution of 320x240 pixels or about 10 FPS on images with the resolution of 640x480 pixels.

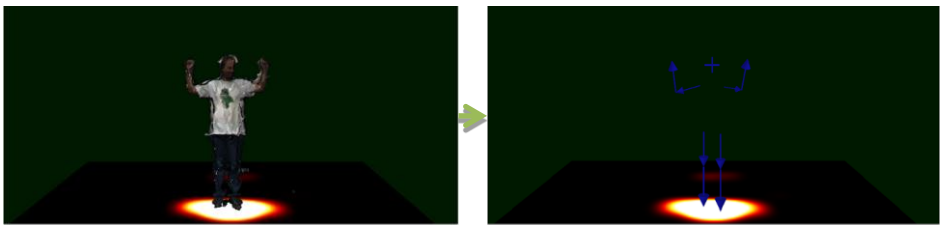

Figure 1: The problem of pose estimation. Given an automatically obtained 3D surface mesh, what human body configuration does it represent?

Further details on the stereo algorithm can be found in [11]. The 3D surface data are used to visualize the users of the system and to provide input to markerless analysis of the pose. The developed framework creates the necessary pipeline for transfer of video, audio and pose data between the sites and visualization of the two displaced persons (therapist and patient) in the same virtual environment.

2. Human pose estimation

The problem of human body pose estimation can be stated as follows. Given a set of input observations, identify the body configuration of human(s) present in the scene. See Figure 1 for an illustration.

Related work: Pose capture systems can be roughly divided into three groups, differing in requirements put on the user and/or environment. Traditionally, pose detection & tracking requires dedicated markers attached to the body. Such systems are the most reliable and precise; however, they are less appropriate in medical setting due to their high price and rather cumbersome setup - attaching the markers or donning a suit. Second group consists of marker-less systems with active sensors, such as range scanners or structured light systems (e.g. the Kinect [10]). Active systems are less obtrusive to the user and no special setup is needed. Their main drawback is a limited applicability due to limited range of the active component and possible interference with the environment (e.g. sunlight) or other sensors. Finally, passive and marker-less systems form the third group [8] which consists of vision-based sensors and are as such relatively inexpensive, modular (e.g. additional cameras can be added to cover larger areas) and non-obtrusive. Unfortunately, the latter group provide the least reliable and precise pose estimation due to appearance and illumination ambiguities.

Perquisites: A pose estimation running within the 3D tele-immersion framework needs to comply with several limitations. To achieve immersive experience, the pose must be extracted in real-time, i.e. at 10 frames per seconds minimum, desirably at about 20 FPS. Due to constrained bandwidth between remote sites, the information shared between individual components of our tele-immersion system is limited to reconstructed 3D meshes only. The original captured images are not transmitted and thus not accessible to the pose estimation algorithm. The algorithm also has to deal with different imperfections and errors of the input data, such as large noise on boundaries of segments, incomplete reconstruction due to self-occlusions or ambiguous appearance. See Figure 3 for examples of challenging 3D input.

In the tele-rehabilitation scenario the appearance of users (patient and therapist) does not change during a session. The algorithm takes advantage of this, learns statistics about the current users on the fly, and uses them in subsequent pose

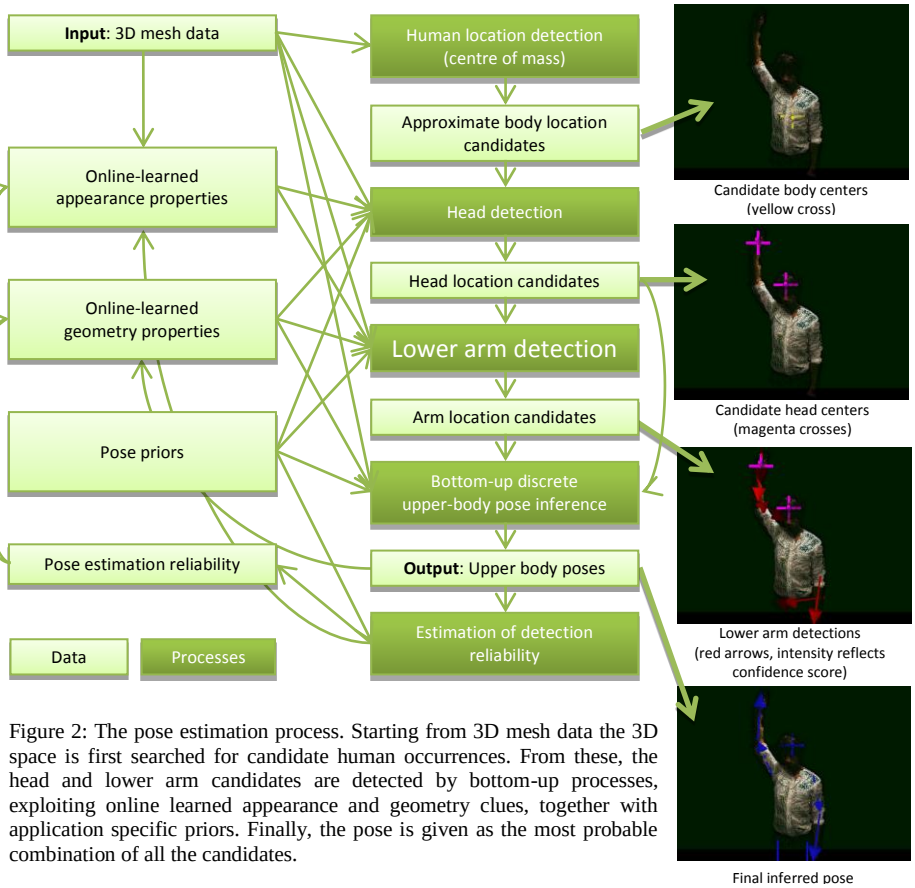

Figure 2: The pose estimation process. Starting from 3D mesh data the 3D space is first searched for candidate human occurrences. From these, the head and lower arm candidates are detected by bottom-up processes, exploiting online learned appearance and geometry clues, together with application specific priors. Finally, the pose is given as the most probable combination of all the candidates.

estimation. In other words, starting with a generic model of a human, the model is progressively adapted to the actual person in the scene.

Following from the intended application, input data detail and noise, and run time allotment, we estimate the poses in only a rather low level of detail. Our pose consists of thirteen 3D points: 2x6 points for shoulders, elbows, wrists, hips, knees and ankles, plus the head centre. These define ten body components: head (a sphere), torso, upper and lower arms, and upper and lower legs (all represented as cylinders).

Method: The pose estimation procedure is schematically illustrated in Figure 2. Both geometric and appearance properties of persons in the scene are maintained and continually adapted, as e.g. in [3][9]. The geometric properties consist of probability distributions of the person's height and limb lengths, which are at the beginning initialised with Gaussian distributions centred on generic lengths. The appearance is represented by probability distributions of colours observed at individual body parts, and are initialised with uniform distributions. Additionally, an application-specific set of prior pose constraints is defined, restricting the variety of poses considered. The most common prior is that of an upright pose, i.e. of a person standing or sitting.

The input to the pose estimation algorithm is a 3D triangle mesh reconstructed from one or more stereo camera pairs. Each vertex of the mesh is attributed with coordinates in the 3D world and colour. In the first step, the 3D mesh is analysed and candidate person (or human-sized object) occurrences are approximately localised.

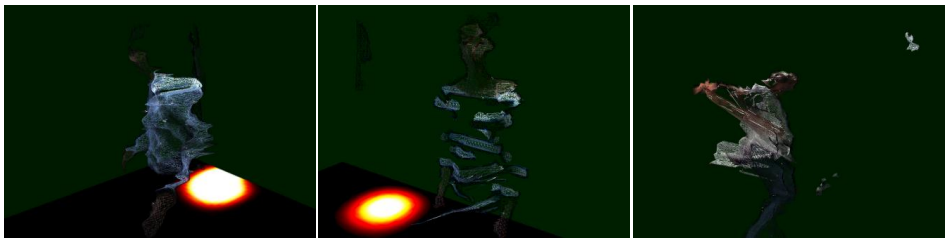

Figure 3: Examples of challenging 3D surfaces: noise, incompleteness, and self-occlusions.

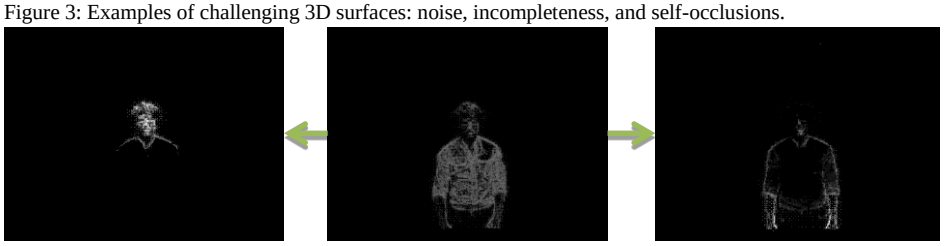

Figure 4: Probability maps. Each mesh vertex is assigned a probability of belonging to a specific body part. The images show probabilities as intensities, with black for zero and white for one. Starting with uniform probability distributions (middle), the probabilities are adjusted based on learned geometric and appearance properties and used in bottom-up detection of body parts. The left image shows probabilities of vertices belonging to head; the right image is for lower arms.

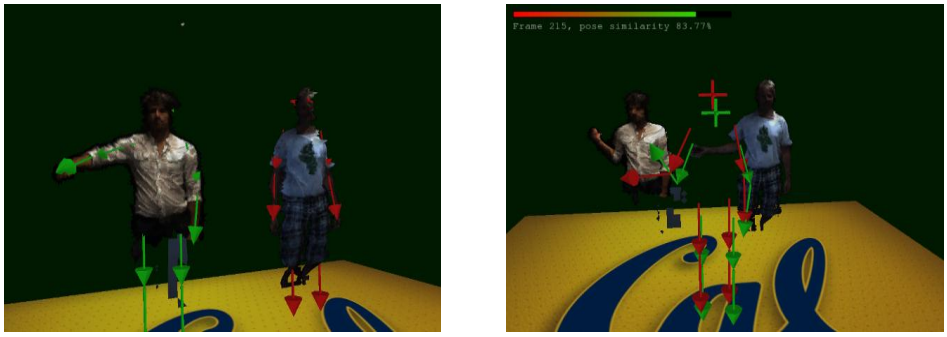

Figure 5: Application in tele-rehabilitation. Poses are extracted for both therapist and patient (left), and aligned and evaluated (right).

Both the appearance properties and the pose priors assign to each mesh vertex a probability of it belonging to a body component. Figure 4 shows the joint probabilities (a product of appearance and priors) for the head and lower arms. Second step of the algorithm uses these probabilities to detect several most probable candidate locations of individual body parts.

In third step a single most probable pose is combinatorially inferred, considering only the discrete set of body part candidates proposed in the second step. Several aspects are taken into account for the inference: detection scores from the second step, the prior pose constraints, the adapted body geometry (limb lengths) and feasibility constraints (body parts cannot intersect). All the constraints are soft, expressed as probability distributions.

Finally, the appearance and geometry models are updated. The update is weighted by our confidence that the pose was extracted correctly, a product of two terms: probability of the pose given the input, *i.e.* reflecting how well the pose fits the 3D mesh, and a term characterising unambiguity of the pose. The second term favours poses with body parts detected in locations out of reach of other body parts. An example is a T-pose, with arms wide open, in which the wrist locations are unreachable by any other part of the body.

3. Conclusions

The proposed algorithm has been integrated with our teleimmersion framework [11] to allow for side-by-side visualization of two remote users (e.g. therapist and patient) with real-time pose evaluation (Figure 5). Although we demonstrate the pose detection and tracking algorithm on the image-based stereo images, the framework could be applied to any depth ranging sensor which can provide visual and spatial information (e.g. Microsoft Kinect).

Examples of readily envisioned tele-medicine and tele-rehabiliation applications of the developed algorithm will be remote patient diagnosis and rehabilitation therapy, where a care provider and the patient are dislocated, in particular in rural areas lacking access to specialist care. This may be pertinent to near future medical landscape where it is expected that aging population demands will far exceed the availability of therapists and physicians.

At this stage we are setting up a system of cameras at UC Davis Medical Center with focus on the evaluation and training of upper extremities. Within this work we will explore how to present and visualize the target pose to the remote user. Furthermore, we will investigate how the collected data can be used to generate computer controlled motion patterns of an avatar which would automatically adapt and promote patient's performance in addition to human therapist training.

Acknowledgements

The research was supported by Center for Information Technology Research in the Interest of Society (CITRIS) at University of California, Berkeley.

References

[1] Bailenson, J., Patel, K., Nielsen, A., Bajscy, R., Jung, S.-H., & Kurillo, G. (2008). The Effect of Interactivity on Learning Physical Actions in Virtual Reality. *Media Psychology* , 354-376.

[2] Bajcsy, P., McHenry, K., Jung, H., Malik, R., Spencer, A., Lee, S. K., et al. (2009). Immersive Environments For Rehabilitation Activities. *International Conference on Multimedia* .

[3] Eichner, M., Marin-Jimenez, M., Zisserman, A., & Ferrari, V. (2010). *Articulated Human Pose Estimation and Search in (Almost) Unconstrained Still Images.* ETH, D-ITET, BIWI, Zurich.

[4] Jack, D., Boian, R., Merians, A., Tremaine, M., Burdea, G., Adamovich, S., et al. (2001). Virtual reality-enhanced stroke rehabilitation. *Neural Systems and Rehabilitation Engineering, IEEE* , 308-318.

[5] Kurillo, G., Koritnik, T., Bajd, T., & Bajcsy, R. (2011). Real-Time 3D Avatars for Tele-rehabilitation in Virtual Reality. *Medicine Meets Virtual Reality Conference*, (pp. 290-296). Newport Beach, CA.

[6] Moline, J. (1997). Virtual reality for health care: a survey. In G. Riva, Virtual Reality for Neuro-Psycho-Physiology: cognitive, clinical and methodological issues in assessment and rehabilitation.

[7] Nahrstedt, K., Bajcsy, R., Wymore, L., Sheppard, R., & Mezur, K. (2008). Computation Model of Human Creativity in Dance Choreography. *Association for the Advancement of Artificial Intelligence* .

[8] Pope, R. (2007). Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding* , 4-18.

[9] Roberts, T., McKenna, S., & Risketts, I. (2002). Online Appearance Learning for 3D Articulated Human Tracking. *International Conference on Pattern Recogntion* , 425--428.

[10] Shotton, J., Fitzgibbon, A., Cook, M., & Blake, A. (2011). Real-time Human Pose Recognition in Parts from Single Depth Images. *Computer Vision and Pattern Recognition*.

[11] Vasudevan, R., Kurillo, G., Lobaton, E., Bernardin, T., Kreylos, O., Bajcsy, R., et al. (2011). High Quality Visualization for Geographically Distributed 3D Teleimmersive Applications. *IEEE Transactions on Multimedia* , 573-584.

[12] Vasudevan, R., Zhou, Z., Kurillo, G., Lobaton, E. J., & Bajcsy, R. (2010). Real-time stereo-vision system for 3D teleimmersive collaboration. *Multimedia and Expo,IEEE* , 1208 - 1213.