Therapeutic Interaction Detection for Serious Games in Physical Rehabilitation

Lubos Omelina[1,3], Bart Jansen[1,2], Bruno Bonnechere[4], Milos Oravec[3], Serge Van Sint Jan[4]

[1]Department of Electronics and Informatics, Vrije Universiteit Brussel
Pleinlaan 2, 1050, Brussel, Belgium

[2]iMinds,
Dept. of Future Health, Ghent, Belgium
{lomelina,bjansen}@etro.vub.ac.be

[3] Institute of Computer Science and Mathematics
Slovak University of Technology in Bratislava, Slovakia
Ilkovičova 3, 812 19 Bratislava
milos.oravec@stuba.sk

[4]Laboratory of Anatomy, Biomechanics and Organogenesis, Université Libre de Bruxelles - Belgium
Lennik Street, 808
1070 Brussels
{bbonnech, sintjans}@ulb.ac.be

ABSTRACT

Serious games gained popularity in recent years together with the use of modern input devices. Mainly marker-less motion tracking cameras play a special role in the automation of physical rehabilitation. These inexpensive cameras can provide accurate information about the movements and poses of the subject without complicated setup. However, these cameras are still not perfect and experience problems in particular poses, setups or when users are interacting. Interaction between a patient and the therapist is a crucial and inevitable aspect of the therapy and results in frustrations when using new technologies. In this paper we propose a method that can identify whether a therapist is interacting with a patient or not, in order to improve not only the therapy sessions but also the quality of the data collected during the gameplay or assessment, automated with the modern input sensors. We compare our measurement results with a marker based motion tracking system (Vicon) and additional scores to demonstrate the importance of identifying interactions between a therapist and patients.

Categories and Subject Descriptors

I.5.5 [**Pattern Recognition**]: Implementation (C.3) – Interactive systems,

General Terms

Measurement, Reliability, Security

Keywords

User identification, Interactions detection, Rehabilitation, Face recognition

1. INTRODUCTION

A 3D camera combined with serious games (SG) for physical rehabilitation seems to be a perspective tool to in advanced rehabilitation sessions. Several systems use Kinect not only to control the games but also as a measuring tool or to provide feedback to the patient [5–7].

Although the use of 3D cameras for gaming purposes has gained popularity, there are still problems related to these technologies. Current 3D cameras can provide reliable recognition and tracking of human skeletons when a single player is in the scene. However, the occurrence of another person in front of the camera who is interacting with the player leads to problems; e.g. switching between tracked users, tracking of wrong people or significant decrease in the quality of the recognized skeletons. For instance, the Kinect camera can track up to 6 people in terms of positions, but can track only two skeletons simultaneously. The presence of more than 2 people leads to unstable user selection for skeletal tracking. We can overcome this by creating constrains (e.g. only two closest people are tracked) and thus decreasing the ergonomics of the system.

Problems related to skeletal tracking are even more frustrating in serious games for physical therapy due to the frequent occurrence of a therapist in the scene. Therapists need to intervene and help the patient in case of problems or difficulties to play. Detection of interactions between a patient and a therapist has several advantages in modern serious games. Depth cameras (like MS Kinect) experience a significant performance and skeleton stability drop when two people are interacting (touching) together. SG systems in physical rehabilitation should reliably recognize the quality of skeletons for later biomechanical (medical) analysis.

We propose a simple, fast and robust method for detecting human interactions in 3D video, in order to improve serious gaming experience in therapeutic practice. In our method, we employ continuous face recognition to detect, recognize and track the patient with other people in the scene (e.g. the therapist, or clinician). We use a method based on local binary patterns (LBP) that is considered to be state of the art in facial recognition [1]. After identifying users in the scene we identify interactions between the patient and other people in the scene.

PervasiveHealth 2014, May 20-23, Oldenburg, Germany

DOI 10.4108/icst.pervasivehealth.2014.255352

2. RELATED WORK

A significant part of the current research in human interactions is devoted to recognizing actions [9]. Yung et al. [11] proposed a method for recognizing different types of interactions from RGBD (RGB + depth) sources using Support Vector Machines (SVMs) and Multiple Instance Learning (MIL). However, in automated therapy sessions we need to detect in real time whether the therapist is helping the patient, without a need to classify the type of interaction.

3. METHOD

Our method works as a preprocessing step for any motion analysis system that is tracking/analyzing movements of a particular patient and where support by the therapist needs to be detected. The patient is at first recognized based on his face in the color image and afterwards the positions of people closeby are detected from the depth map. The method assumes also labeled regions in the depth-map that specify clusters of points where each cluster represents a human body in the scene. Based on the depth map and labeled regions we identify whether the therapist is supporting or interacting with the patient.

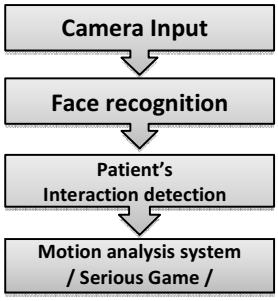

Figure 1. Schematic overview of detection.

3.1 Recognition of a patient

Facial recognition is a well-studied area and there are many different methods with varying accuracy based on a specific use case. We decided to use a method based on LBP features with measuring Chi-square distance which is described in details in [1][3]. The chosen method provides a trade-of between accuracy and computational complexity; thus it can run in real-time while maintaining state-of-the-art recognition accuracy [8].

Before the face is recognized we preprocess the image as follows:

- conversion of the color image to grayscale,
- alignment of the face based on the position of eyes,
- scaling of the face image to unified size,
- equalization of the image histogram.

In order to compute the LBP histogram, the image is segmented into several non-overlapping regions and from each of these regions a histogram of uniform LBP patterns is computed (Fig. 1). Histograms are concatenated from left to right and from top to bottom.

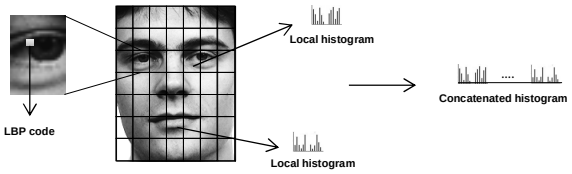

Figure 2. Process of creating concatenated LBP histograms.

The method requires training of the patient's face. In this step the camera captures multiple images of the patient's face and creates a training set. Training needs to be done in advance and requires to perform 5 different poses in front of the camera. For each pose the subject needs to keep the pose for one second. The camera captures 30 frames per second and thus the captured image set contains redundancies. In order to choose the most representative images we use the *k-means* algorithm [2].

3.2 Recognition of the interaction

Physical interactions between the patient and the physiotherapist are the essence of physiotherapy. The physiotherapist can, for example, perform the motion (together) with the patient in order to show him the right way to do it, can stabilize the trunk in order to avoid compensatory movement, can palpate some muscles during exercises to be sure that the patient is recruiting the right muscles, can evaluate passive range of motion of a particular joint by helping the patient to perform this motion… Therefore, it is important that the games are robust to the presence of the therapist in the game and even, that the games can detect when the therapist is interacting with the patient.

Let I be a set depth map acquired from a camera and $I(x, y)$ a point from the depth map. We define a set of points $U \subset I$, such that U contains points representing the detected human bodies. We say that two different bodies U_1 and U_2, $U_1 \cap U_2 = \{\}$ are interacting when there exist such points from $p_1 \in U_1$ and $p_2 \in U_2$ that $\|p_1 - p_2\| < \lambda$ where the threshold λ represents a critical distance (aka comfort zone). The comfort zone represents a parameter of our model.

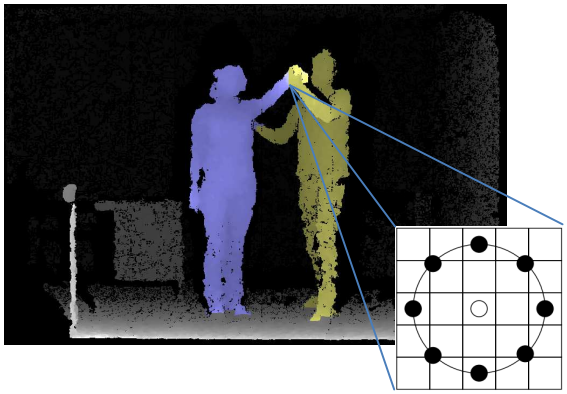

Figure 3. An example of a circular (8,2) neighborhood that is explored for a particular point in the depth map.

Since the depth map I is represented as a grid/matrix of depth points, we assume uniform distances in the X and Y axes. In order to detect collisions, we explore a circular neighborhood $(N, f(\lambda))$ of each point $p \in U$ where N is the number of points being explored and f a function mapping metric space to pixel space (Fig. 3). Exploring only a limited amount of points in the neighborhood provides only an approximation of the intersection between U_1 and U_2, but is computationally less expensive and can run in real-time, leaving resources for other tasks (the game and actual skeleton processing).

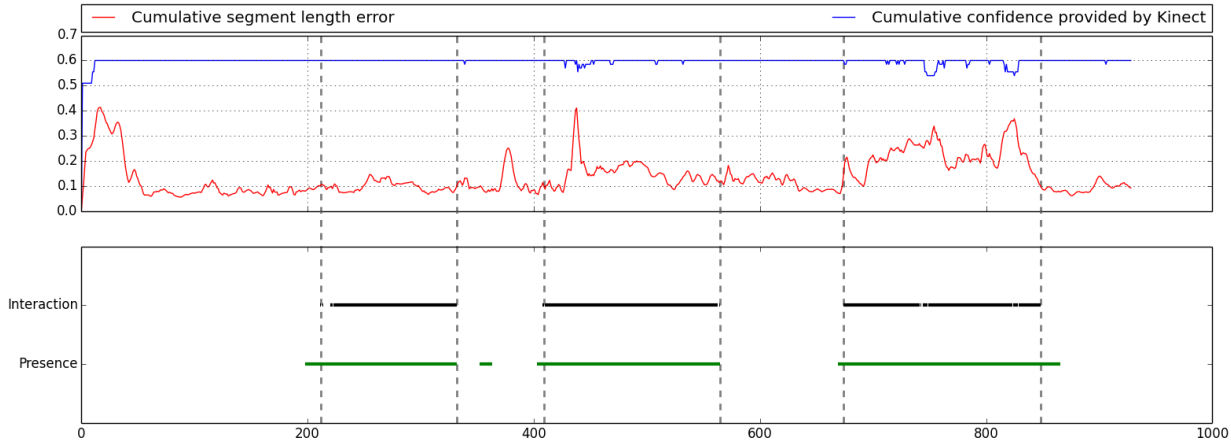

Figure 4. Evolution of skeletal stability and detection of interactions. The top part shows variability of segment lengths measured by Kinect camera (in red) together with cumulative confidence based on the tracking state of each joint (in blue). The bottom part shows 3 detected interaction periods (in black) and periods when the therapist was present in the scene (in green).

$$G_{neighbor}(p_c) = \sum_{n=1}^{N} s(p_c, p_n) \max(0, \|p_c - p_n\| - \lambda) \quad (1)$$

where $s(p_1, p_2)$ is defined as follows

$$s(p_1, p_2) = \begin{cases} 1 & \text{if } id(p_1) != id(p_2) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The resulting image that describes the local neighborhood of every pixel, may still contain misdetections caused by the noise of the depth sensor. In contrast to positive areas caused by noise, interaction areas occur in blobs. We detect these interaction blobs by applying a Laplacian of Gaussian filter of an appropriate size:

$$LoG(x, y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2 + y^2}{2\sigma^2}\right] e^{-\frac{x^2+y^2}{2\sigma^2}}. \quad (3)$$

The result is an image containing positive values in blobs that correspond to the interaction points.

4. RESULTS

To test our method we used the MS Kinect camera as it is the most popular 3D camera available. We also used the Vicon motion tracking system, where we tracked only the patient's skeleton. Both systems recorded the skeleton simultaneously while the patient was interacting with a therapist. In Fig. 4 measurements are shown from one session. During this session the therapist approached the patient 3 times and left the scene afterwards.

In order to demonstrate the performance of our method, we compared events from our method (start/stop of the interaction) with stability of the skeleton over time as the frames become available. We used the following measures to evaluate stability:

Confidence value of the skeleton tracker – The Kinect SDK provides 3-state confidence information about tracking for each joint. A joint can be *(i) tracked*, *(ii) inferred* or *(iii) not tracked*. Our observations show that while two people are interacting in the scene, the quality of the recognized skeletons slightly decreases. It is important to note that the decreased quality does not relate only to interactions but also to the pose, occlusions. In order to compute the cumulative confidence value we simply sum values for each joint while the *tracked* state has weight 0.06, the *inferred* state has weight 0.015 and the *not tracked* state has weight 0. The influence of an interaction between a patient and a therapist on Kinect's confidence values is shown in Fig.4. We can see that the confidence values are rather stable even during the interaction when the skeleton segments are unstable.

Stability of segments lengths – a common problem of skeleton detecting cameras is that segment lengths are varying over time. Since there is no prior knowledge about the skeleton, the camera continuously makes new hypotheses about poses and segment lengths [10]. We know that the lengths should remain the same and their increased variability indicates low skeleton stability. For the comparison purpose, we compute a cumulative error from all segments as a sum of absolute differences of all segments.

Stability of segments lengths measured with Vicon system – we tracked the skeleton of the patient with a marker-less camera (Kinect) and a marker based system (Vicon) simultaneously in order to compare their performance when two people are interacting. It is important to note that although the Vicon marker based system is considered to be the gold standard for human motion tracking, this system might also experience errors due to the lack of visibility of markers while interacting (Fig. 5). We used the Vicon system to track only the patient, not the therapist.

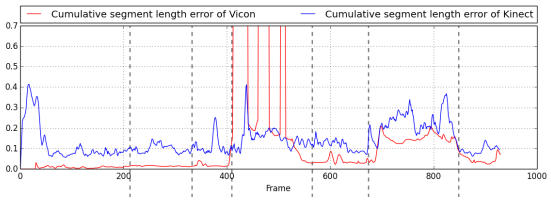

Figure 5. Comparison of cumulative segment length errors for Vicon system and Kinect.

A part of the data processed during the experiments is shown in Fig. 4. This recording contains three separate interactions and in between the therapist leaves the scene. We can see that the segment length variability is increasing in each interaction. This is caused by the decreasing distance between the therapist and the patient. In the third interaction the segments vary the most.

Although there is a significant error, the Kinect reports confident results about the tracked joints.

Fig. 5 shows that the Vicon system provides very accurate results when only a single person is in the scene. Segment lengths are stable even during the first interaction when the therapist approached the patient from a side. In the second and third interaction the therapist occluded several markers – the Vicon system reported significantly worse results. We can see that in case of close interaction stability of both systems decreases. In case of close interactions the Vicon system loses track of markers and thus cannot provide any information about joints connected to those markers. Results also demonstrate that Kinect can even provide more stable results due to presence of all joints inferred from the depth map.

5. CONCLUSION

SGs are still perceived as regular games. However, based on the used technology, SGs could be used for more than performing exercises. Since data of the patients can be recorded during the rehabilitation (games), this data can be used to follow the patients' evolution and to be sure that they are doing the right exercises [4]. In order to have a precise follow up of the patient, the clinician must be sure that the patient was playing alone and was not helped by friends, parents or others (therapist, clinicians). The presented method allows to detect interactions between the patient and other people, and also helps to discriminate between active and passive motion. In addition, during an interaction the precision of the capturing devices decreases and a good discrimination can help to filter out erroneous measurements.

A possible extension of this work might be more detailed segmentation of the scene in order to detect other supporting objects. Games played with balance boards could also benefit from recognizing users standing on the board and thus tracking authenticity of measured data.

6. ACKNOWLEDGMENTS

Research described in the paper was done within the RehabGoesHome and ICT4REHAB projects (www.ict4rehab.org) funded by Innoviris and within the grant No. 1/0529/13 of the Slovak Grant Agency VEGA.

7. REFERENCES

[1] Ahonen, T., Hadid, A. and Pietikäinen, M. 2004. Face recognition with local binary patterns. *Computer Vision-ECCV 2004*. 3021, (2004), 469–481.

[2] Ban, J., Feder, M., Jirka, V., Loderer, M., Omelina, L., Oravec, M. and Pavlovicova, J. 2013. An Automatic Training Process Using Clustering Algorithms for Face Recognition System. *Proceedings ELMAR-2013 : 55th International Symposium. Zadar, Croatia* (2013), 15–18.

[3] Ban, J., Pavlovicova, J., Feder, M., Omelina, L. and Oravec, M. 2012. Face recognition methods for multimodal interface. *Wireless and Mobile Networking Conference (WMNC), 2012 5th Joint IFIP* (2012), 110–113.

[4] Bonnechère, B., Jansen, B., Omelina, L., Da Silva, L., Mouraux, D., Rooze, M. and Van Sint Jan, S. 2013. Patient follow-up using Serious Games. A feasibility study on low back pain patients. *Proceedings of the 3rd european conference on gaming and playful interaction in health care* (2013), 185–195.

[5] Chang, C.-Y., Lange, B., Zhang, M., Koenig, S., Requejo, P., Somboon, N., Sawchuk, A.A. and Rizzo, A.A. 2012. Towards pervasive physical rehabilitation using Microsoft Kinect. *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on* (2012), 159–162.

[6] Chang, Y.-J., Chen, S.-F. and Huang, J.-D. 2011. A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Research in Developmental Disabilities*. 32, 6 (2011), 2566–2570.

[7] Clark, R.A., Pua, Y.-H., Bryant, A.L. and Hunt, M.A. 2013. Validity of the Microsoft Kinect for providing lateral trunk lean feedback during gait retraining. *Gait Posture*. 38, 4 (2013), 1064–1066.

[8] Oravec, M., Pavlovičová, J., Mazanec, J., Omelina, Ľ., Féder, M. and Ban, J. 2011. Efficiency of Recognition Methods for Single Sample per Person Based Face Recognition. *Reviews, Refinements and New Ideas in Face Recognition*. Rijeka : InTech, 2011. 181–206.

[9] Poppe, R. 2010. A survey on vision-based human action recognition. *Image and Vision Computing*. 28, 6 (2010), 976–990.

[10] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. and Blake, A. 2011. Real-time human pose recognition in parts from single depth images. *In CVPR* (2011).

[11] Yun, K., Honorio, J., Chattopadhyay, D. and Berg, T.L. 2012. Two-person interaction detection using body-pose features and multiple instance learning. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on* (2012), 28 – 35.