Interaction Detection with Depth Sensing and Body Tracking Cameras in Physical Rehabilitation

L. Omelina1,2,3; B. Jansen1,2; B. Bonnechère4; M. Oravec3; P. Jarmila3; S. Van Sint Jan4

1Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels, Belgium; 2IMinds, Dept. of Medical IT, Ghent, Belgium; 3Institute of Computer Science and Mathematics, Slovak University of Technology, Bratislava, Slovakia; 4Laboratory of Anatomy, Biomechanics and Organogenesis, Université Libre de Bruxelles, Brussels, Belgium

1. Introduction

A motion capture system combined with serious games (SG) for physical rehabilitation seems to be a perspective tool in advanced rehabilitation sessions. Several systems use markerless skeletal tracking with low cost cameras not only to control the games but also as a measuring tool or to provide feedback to the patient and the therapist [1–4].

Although the use of markerless systems (MLS) for gaming purposes has gained popularity, there are limitations implied by underlying technologies. Current low cost MLS, which are based on depth sensing, can provide reliable recognition and tracking of human skeletons when only a single player or clearly separated players are in the scene. However, the occurrence of another person in front of the camera who is interacting with the player(s) 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 first version of the Microsoft Kinect camera can track only two skeletons simultaneously and needs approximately 2.5 m clear space in between players and the sensor. The presence of more than two persons leads to unstable user selection for skeletal tracking, especially when they are interacting. The second version of the Kinect camera can track up to six skeletons but stability of the skeletal tracking remains low in case of interaction between the people in the scene.

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 a patient in case of problems or difficulties to play. This intervention leads to difficulties in skeletal detection and thus the ability to detect this interaction would help to in-

Keywords
User identification, interaction detection, rehabilitation, face recognition

Summary
Introduction: This article is part of the Focus Theme of Methods of Information in Medicine on “Methodologies, Models and Algorithms for Patients Rehabilitation”.

Objectives: This paper presents a camera based method for identifying the patient and detecting interactions between the patient and the therapist during therapy. Detecting interactions helps to discriminate between active and passive motion of the patient as well as to estimate the accuracy of the skeletal data.

Methods: Continuous face recognition is used to detect, recognize and track the patient with other people in the scene (e.g. the therapist, or a clinician). We use a method based on local binary patterns (LBP). After identifying users in the scene we identify interactions between the patient and other people. We use a depth map/point cloud for estimating the distance between two people. Our method uses the association of depth regions to user identities and computes the minimal distance between the regions.

Results: Our results show state-of-the-art performance of real-time face recognition using low-resolution images that is sufficient to use in adaptive systems. Our proposed approach for detecting interactions shows 91.9% overall recognition accuracy which is sufficient for applications in the context of serious games. We also discuss limitations of the proposed method as well as general limitations of using depth cameras for serious games.

Conclusions: We introduced a new method for frame-by-frame automated identification of the patient and labeling reliable sequences of the patient’s data recorded during rehabilitation (games). Our method improves automated rehabilitation systems by detecting the identity of the patient as well as of the therapist and by detecting the distance between both over time.
There are two approaches to tackle the problems:

Creating constraints – creating a set of rules on how to use the SG system in order to avoid situations that are causing problems, e.g. people should not stand closer than a specific distance. This solution significantly decreases usability and ergonomics of the system since it limits ways of using the systems and users need to learn additional rules on how to use it.

Improving system intelligence – additional processing that can identify and track the right person and can identify whether the people in the scene are interacting or not, e.g. the system can notify people about problematic poses. This may increase computational complexity, but does not decrease usability and ergonomics.

Detection of interactions between a patient and a therapist has several advantages in modern SG applied to physical rehabilitation. For instance, discriminating between active and passive motion or the identification of unreliable and inaccurate skeletal measurements would be helpful to the therapist in order to assess the patient’s compliance to planned therapy (i.e., a typical physical therapy scheme must be adapted to the patient’s problems in order to be efficient). In addition, modern SG systems in physical rehabilitation should not only motivate the patient, but also reliably recognize the quality of skeletons for simultaneous biomechanical (medical) analysis.

We propose a simple, fast and robust approach for detecting human interactions in the RGB-D (color + depth) image stream and increased system intelligence, in order to improve the serious gaming experience in therapeutic practice. The captured depth is a matrix with elements representing distance of a particular point from a camera. 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 the clinician). We use a method based on local binary patterns (LBP) that is considered to be state-of-the-art in facial recognition [5]. After identifying users in the scene we identify interactions between the patient and other people in the scene. We use a depth map/point cloud for estimating the distance between two people. Our method uses association of depth regions to user identities provided by the underlying framework and computes the minimal distance between the regions. The same method for user disambiguation and for interaction detection can also be integrated (not only) in home based rehabilitation systems.

The presented method is designed with in the context of serious games used in physical rehabilitation. In this article we focus on detecting a particular type of interaction where a therapist assists the patient to play serious games. In this scenario skeletal tracking faces difficulties due to occluding bodies and interactions. The skeleton recognition algorithm cannot always recover the joint positions correctly and thus the skeletal data can become corrupted. We use depth maps and labeled user regions to detect the intervals when this corruption occurs in order to exclude such data from further processing. On the other hand, skeletal data may be corrected (e.g. by using accelerometers to improve the positions and rotations of body segments) in which case the detected interaction could be a valuable source of information in order to analyze passive range of motion of different joints.

2. Background

Topics covered in face recognition are very broad. We focus only on feature based methods as they are reaching state-of-the-art performance and have the ability to run in real-time. M. Turk and A. Pentland [6] introduced eigenfaces, the first successful approach capable of running in real time. Although a lot of approaches were proposed since then, there are still open problems related mostly to changes in poses and lighting [7]. To address this problem current approaches use light-invariant features like LBP, HOG (Histograms of Oriented Gradients) or Gabor features [8, 9]. Especially the method based on LBP proposed by T. Ahonen [5] gained a lot of attention and belongs to state-of-the-art in face recognition [9]. A strong advantage of LBP features in contrast to other (e.g. Gabor) features is its low computational complexity.

For matching features a hand-crafted distance metric with nearest neighbor classifier or a learning algorithm can be used. Methods based on eigenfaces typically use Euclidean or Mahalanobis distance [6]. T. Ahonen used Chi-square distance with the nearest neighbor classifier to match LBP features. Although machine learning approaches (e.g. neural networks or support vector machines) reach better performance than nearest neighbor matching, they need time consuming training that makes them unusable for some use-cases.

A significant part of the current research in human interactions is devoted to recognizing actions [10]. Yun et al. [11] proposed a method for recognizing different types of interactions from RGBD sources using Support Vector Machines.
and Multiple Instance Learning. However, in automated therapy sessions, we need to detect in real time whether the therapist is supporting the patient, without a need to classify the type of interaction.

3. Methods

Our method can be used 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. An overview of the approach is shown in Figure 2. The patient is at first recognized based on his face in the color image and afterwards the positions of people close by are detected from the depth map. The method labels the depth-map regions with identities of users. Based on the depth map and labeled regions we identify whether the therapist is supporting or interacting with the patient.

3.1 Environment and Use-case Constraints

The general unconstrained face recognition is still an open problem mainly due to varying illumination, different facial expressions, poses, complex background, etc. However, we can avoid some of the open problems by considering constraints of the environment in which the real system is used. Our approach (assuming its use in therapy and serious games) contains several inherited constraints (i.e. constraints given by the use-case and environment in which users interact with the system) that help us improve the face recognition accuracy and thus make the overall system more reliable.

Playing a video game requires users to face the screen in order to interact with the game. In our system we assume that the user’s head is oriented towards the screen what significantly limits the poses performed during the game-play. Occasional cases when the patient’s pose exceeds the tolerance can be easily filtered out by temporal filtering. A longer lasting change of facial pose means that the patient does not look at the camera and is not paying attention to instructions on the screen (and thus not following the exercises). In this case, the system does not collect data because the movement is not the result of the exercise, but rather a free (not relevant) motion that should be excluded from exercise analysis.

Current gaming MLSs are made exclusively for indoor environments and are hence faced with limited variability in lighting. In addition, the games are played in the therapist’s practice or at patient’s home making the position of the camera static (i.e. no large lateral displacements will occur). Most popular MLSs are using depth cameras based on the time-of-flight (ToF) or structured light principle. These cameras illuminate the scene uniformly with near infra-red light. The source of the NIR light is typically placed near the sensor and thus all acquired facial images are lightened under the same conditions.

In our approach we also assume that the user is located at the center of the field of view of the camera and thus the probability of appearing at the sides is rather low.

3.2 Recognition of a Patient

Facial recognition is a well-studied area and there are many different methods with varying accuracy based on the specific use cases. We decided to use a method based on LBP features with Chi-square distance as a similarity metric which is described in details in [5, 12]. The selected method provides a trade-off between accuracy and computational complexity; thus it can run in real-time while maintaining state-of-the-art recognition accuracy (13).

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 features, the image is segmented into several non-overlapping regions and from each of these regions a histogram of uniform LBP patterns is computed (Figure 3).
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. In order to reduce negative influence of the pose variability we collect training images for each face under different poses. Training needs to be done in advance and requires the patient to perform five 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 around 150 samples per view. In order to choose the most representative images per view, we use the k-means clustering algorithm [14]. By proper placement of the camera we can limit poses of the face that occur while the patient is performing the exercises. From our preliminary experiments where patients played the games, we observed that the variance of the horizontal head movement is 42 degrees and 33 degrees for vertical head movement (Figure 4). Although the facial pose variation is limited, it decreases the face recognition accuracy.

3.3 Recognition of the Interaction

Physical interactions between the patient and the physical therapist (PT) are often crucial in physical rehabilitation. The PT 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 support the movement of a particular joint by helping the patient to perform this motion, etc. Therefore, it is important that the games are robust to the presence of the therapist in the scene and that the games can detect when the therapist is interacting with the patient.

In this section we describe a method for detecting interaction from a depth map camera stream. In our method we assume that part of the preprocessing, i.e. the detection of humans from a depth map, is done by the camera itself or the underlying framework. The input to our method is the set of points associating people with depth regions.

Let I be a depth map acquired from a camera and I(x, y) a point from the depth map. We define a set of points I \( \subseteq U \), 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 \( p_1 \in U_1 \) and \( p_2 \in U_2 \) that \( \|p_1 - p_2\| < \lambda \), where the threshold \( \lambda \) represents a critical distance (aka comfort zone). The comfort zone represents a parameter of our model.

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 (Figure 4). 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).

The resulting image \( G_{\text{neighbor}} \) that describes the local neighborhood of each 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:

\[
G(x, y) = -\frac{1}{\pi\sigma^2} \left[1 - \frac{x^2 + y^2}{2\sigma^2}\right] e^{-\frac{x^2 + y^2}{2\sigma^2}}
\]

(3)

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

4. Experimental Setup and Results

4.1 Recognition of a Patient

In our method we test the face recognition and detection interaction as two separate aspects. We integrated the proposed method to a previously developed gaming system used in rehabilitation of children suffering from cerebral palsy [15]. This system uses the Microsoft Kinect sensor as the MLS and thus no additional hardware is necessary.

Each patient needs to enroll to our system. In enrollment we collected facial images from 10 different people using a Kinect camera. The images are scaled to a size of 36 x 48 pixels. The system requires...
users to pass the training procedure in which they demonstrate five different facial poses in order to make face recognition pose invariant. From each facial image we compute the LBP code image that is divided into 12 non-overlapping regions (three divisions vertically and four regions horizontally) – these are used to compute the resulting concatenate LBP histograms. Features are clustered with k-means algorithm to five different canonical centers that are stored in the database as templates for a user. The face recognition performance is shown in Figure 6. Our system can reach more than 90% recognition accuracy with less than 10% false accept rate. Since we perform face recognition on each frame, we consider this result as sufficient for the purpose of identifying the patient.

4.2 Recognition of the Interaction

In this section we first provide an experimental comparison of our method with other approaches where we motivate the need of our approach. Thereafter, we provide detailed evaluation of the proposed interaction detection method.

4.2.1 Comparison with Existing Approaches

To demonstrate the need of interaction detection we analyze the stability of segment lengths during interaction. We use a MLS (MS Kinect) together with a marker-based stereophotogrammetry system (MBS) (Vicon) to capture the patient’s skeleton. The MBS was using eight high resolution

Figure 5 An example of a circular (8,2) neighborhood that is explored for a particular point in the depth map

Figure 6 ROC curves of the face recognition system tested on the dataset collected with MS Kinect v.1 containing 1464 facial images of 10 people in two sessions
calibrated NIR cameras. Both systems recorded the scene simultaneously while the patient was interacting with a therapist. In Figure 7 measurements are shown from one session. During this session the therapist approached the patient three times and left the scene afterwards.

We compared events from our method (start/stop of the interaction) with (i) stability of the skeletal segment lengths (measured by both MBS and MLS) over time as the frames become available, (ii) Kinect’s confidence values. In Figure 7 we can see that the confidence values are rather stable (comparing to the time periods without interaction) even during the interaction when the skeleton segments are unstable. A common problem of body tracking cameras is that segment lengths are varying over time as the camera redetects skeleton in every frame. As the segment lengths should remain the same, their increased variability indicates low skeleton stability. The MBS, considered to be the gold standard for human motion tracking, also experiences errors due to the lack of visibility of markers while interacting (Figure 8).

The recording in Figure 7 contains three separate interactions where in-between the therapist leaves the scene. We can see that the segment length variability is increasing in each interaction due to a decreasing distance between the therapist and the patient. In the third interaction, the segments vary the most, but the Kinect reports confident tracking and thus doesn’t reflect the skeleton’s accuracy. Figure 11 shows results of detecting interactions based on

![Figure 7](image-url) Evolution of skeletal stability and detection of interactions. The top part shows the variability of segment lengths measured by the MLS (in red) together with cumulative confidence based on the tracking state of each joint (in blue). The bottom part shows three detected interaction periods (in black) and periods when the therapist was present in the scene (in green).

![Figure 8](image-url) Comparison of cumulative segment length errors for MBS (Vicon, in red) and MLS (Kinect, in blue)
decreased confidence value provided by the framework – the confidence value is not sufficient for detecting interactions.

►Figure 8 shows that the MBS provides accurate results when only a single person is in the scene and even during the first interaction (the therapist approached the patient from a side). In the second and third interaction the therapist occluded several markers – the MBS reported significantly worse results.

### Table 1

<table>
<thead>
<tr>
<th>Patient’s pose</th>
<th>Therapist’s pose</th>
<th>Exercise</th>
<th>Position of therapist’s hands</th>
<th>Recognition accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standing</td>
<td>Beside</td>
<td>Elbow flexion/extension (from shoulder flexion)</td>
<td>Upper and lower arm</td>
<td>98.1%</td>
</tr>
<tr>
<td>Standing</td>
<td>Beside</td>
<td>Elbow flexion/extension (from shoulder abduction)</td>
<td>Upper and lower arm</td>
<td>97.2%</td>
</tr>
<tr>
<td>Standing</td>
<td>Beside</td>
<td>Shoulder flexion/extension (with elbow in extension)</td>
<td>Shoulder and lower arm</td>
<td>95.9%</td>
</tr>
<tr>
<td>Standing</td>
<td>Beside</td>
<td>Shoulder add/abduction</td>
<td>Shoulder and lower arm</td>
<td>98.1%</td>
</tr>
<tr>
<td>Standing</td>
<td>Beside &amp; behind</td>
<td>Sideways trunk bending</td>
<td>Lower back</td>
<td>91.5%</td>
</tr>
<tr>
<td>Standing</td>
<td>Beside &amp; behind</td>
<td>Sideways trunk bending</td>
<td>Shoulders</td>
<td>98.6%</td>
</tr>
<tr>
<td>Standing</td>
<td>Beside &amp; behind</td>
<td>Sideways trunk bending</td>
<td>Head</td>
<td>99.2%</td>
</tr>
<tr>
<td>Standing</td>
<td>Behind</td>
<td>Sideways trunk bending</td>
<td>Upper back</td>
<td>31.6%</td>
</tr>
<tr>
<td>Sitting</td>
<td>Beside</td>
<td>Elbow flexion/extension</td>
<td>Upper and lower arm</td>
<td>97.6%</td>
</tr>
<tr>
<td>Sitting</td>
<td>Beside &amp; behind</td>
<td>Elbow flexion/extension</td>
<td>Upper and lower arm</td>
<td>98.9%</td>
</tr>
<tr>
<td>Sitting</td>
<td>Beside &amp; behind</td>
<td>Shoulder flexion/extension</td>
<td>Shoulder and lower arm</td>
<td>97.2%</td>
</tr>
<tr>
<td>Sitting</td>
<td>Beside &amp; behind</td>
<td>Shoulder add/abduction</td>
<td>Shoulder and lower arm</td>
<td>97.5%</td>
</tr>
<tr>
<td>Sitting</td>
<td>Beside &amp; behind</td>
<td>Sideways trunk bending</td>
<td>Upper back</td>
<td>98.3%</td>
</tr>
<tr>
<td>Sitting</td>
<td>Beside &amp; behind</td>
<td>Sideways trunk bending</td>
<td>Shoulders</td>
<td>99.1%</td>
</tr>
<tr>
<td>Sitting</td>
<td>Beside &amp; behind</td>
<td>Sideways trunk bending</td>
<td>Head</td>
<td>89.7%</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinect</td>
<td>64.8%</td>
<td>79.6%</td>
<td>98.4%</td>
</tr>
<tr>
<td>Kinect 2</td>
<td>58.0%</td>
<td>91.9%</td>
<td>96.8%</td>
</tr>
</tbody>
</table>

We can see that in case of close interaction, the stability of both systems decreases. Results also demonstrate that MLS can even provide more stable results in case of close interaction due to the presence of all joints inferred from the depth map.

#### 4.2.2 Evaluation of the Proposed Interaction Detection Method

To collect a database of interaction from the gameplay we participated in different

---

**Figure 9** Visualization of depth images captured by the depth camera. Images show four types of interactions that were correctly recognized. a) Interactions with a sitting patient, b) interactions with a standing patient.
rehabilitation sessions that involved six different serious games played with the MLS or the balance board. All games were focused to practice upper limbs, trunk/head balance, or sitting balance. We created a list of 15 most frequent types of interactions (Table 1), we performed them and recorded them in the laboratory with two different depth sensing cameras (we used the MS Kinect, and MS Kinect 2). Then we manually labelled the recordings and marked occurrences of interactions and thus performance of passive motion by the patient. Eight exercises were performed with the patient standing and seven in sitting position.

We recorded the experiments with both, Kinect v. 1 and Kinect v. 2 cameras sequentially (to prevent their mutual influence on the measurements). We summarize the results in Table 2. In our recordings, interaction occurred in 58.9% of all 9266 frames while using Kinect v. 2 camera. We are able to detect correct state of interactions in 91.9% frames. The precision of the interaction detection was 96.8% with 89% of recall. In Figure 9 examples are shown of interactions that were correctly detected by our system. We observed that it is important for both the patient and the therapist to be clearly visible. The best performance of the interaction detection is reached when the therapist’s body is visible and thus not standing behind the patient. However there are cases of interaction that cannot be detected with our approach. Figure 10a shows a patient standing and a therapist reaching to his shoulder – due to discontinuity of the therapist’s body in the depth image our method does not detect interaction (the therapists hand is incorrectly classified as part of the patient’s body). In Figure 10b a therapist is standing behind the patient and hence the body of the therapist cannot be tracked.

5. Conclusion

SGs and depth sensing cameras became a popular tool in physical therapy despite technological problems related to patient-therapist interactions. Physical interaction between patient and therapist negatively influences the accuracy of the sensory de-

Figure 10 Visualization of interactions that were not correctly recognized. a) A therapist supporting the patient’s shoulder, his hand was occluded by patient’s body and thus not part of the depth map. b) A therapist standing behind a patient is not correctly recognized by the underlying framework which detects only one person – thus our method is not able to recognize the interaction.

Figure 11 ROC for classification of interaction based on the confidence values of the joints provided by the Kinect SDK (red) and using the proposed method (green). We can observe that the proposed method offers significantly better results in per-frame classification of interaction.
vices. The accuracy of the sensors is sufficient to be applied in games, however for diagnostic or monitoring purposes the quality of the input data has to be improved. One possibility is to focus only on the sequences and situations where the skeletal tracking is reliable. That approach requires the detection of interacting people in front of the camera. By automatically labeling reliable sequences of the patient’s data recorded during rehabilitation (games), we can follow the evolution and correctness of the performed exercises [16]. Also, in order to have a precise follow up of the patient, the clinician must be able to discriminate active motion; motion with support; passive motion and thus make sure that the patient was playing alone and was not helped by others.

We described a method that improves automated rehabilitation systems by detecting the identity of the patient as well as of the therapist and by detecting the distance between both over time. In the presented study we focused on the detection of interactions during the gameplay within physical therapy. The method also helps to discriminate between active and passive motion. We show that during an interaction, the precision of the capturing devices decreases and thus a good detection for these interactions can help to filter out erroneous measurements.

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

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.

References

3. Clark RA, Pua YH, Bryant AL, Hunt MA. Validity and reliability of the measured data.
14. Ban J, Fedor M, Jirka V, Loderer M, Omelina L, Oravec M, et al. An Automatic Training Process for Patient follow-up; Passive motion and thus make sure that the patient was playing alone and was not helped by others.