A Monitoring System for Home-Based Physiotherapy Exercises

Ilktan Ar and Yusuf Sinan Akgul

Abstract This paper describes a robust, low-cost, vision based monitoring system for home-based physical therapy exercises. Our system contains two different modules. The first module achieves exercise recognition by building representations of motion patterns, stance knowledge, and object usage information in gray-level and depth video sequences and then combines these representations in a generative Bayesian network. The second module estimates the repetition count in an exercise session by a novel approach. We created a dataset that contains 240 exercise sessions and tested our system on this dataset. At the end, we achieved very favourable recognition rates and encouraging results on the estimation of repetition counts.

1 Introduction

Physical therapy (or physiotherapy) is a medical science that concerns with the diagnosis and treatment of patients who have injuries or other problems that limit their capabilities to perform functional activities. Physical therapists provide care to patients by offering a treatment to reduce pain, prevent disability, and restore function. These treatments usually include physiotherapy exercises. However, human power, money, and time resources are not generally sufficient to do one-to-one sessions with all patients. These problems lead to home-based physical therapy exercises and there is a need to monitor this type of treatment.

Major achievements for human motion tracking systems for rehabilitation are surveyed by Zhou and Hu [9]. Soutscheck et al. [7] presented an automatic system

Ilktan Ar

Kadir Has University, Cibali, Istanbul 34083, Turkey, e-mail: ilktana@khas.edu.tr
GIT Vision Lab: http://vision.gyte.edu.tr

Yusuf Sinan Akgul

Gebze Institute of Technology, Gebze, Kocaeli 41400, Turkey, e-mail: akgul@bilmuh.gyte.edu.tr
GIT Vision Lab: http://vision.gyte.edu.tr

to supervise and support rehabilitation and fitness exercises. Their solution outputs angular measurements of the knee joint by 2D and 3D tracking of knee positions using specialized sensors. Fitzgerald et al. [4] developed a system which utilizes ten inertial motion tracking sensors in a wearable body suit and a laptop/computer that communicates with this suit by Bluetooth connection. Jung et al. [5] developed a sensor driven motion tracking system to analyze upper body functions of a person.

In this paper, we propose a robust, low-cost, vision based monitoring system for home-based physical therapy exercises. Instead of expensive systems which require specialized hardware as in the above works, the proposed system use a low-cost Microsoft Kinect sensor which contains a depth and an RGB camera. The novelty of this paper is two-fold. First, we define a generative Bayesian network which combines motion patterns, stance knowledge, and object usage information to recognize the exercise type in the given video sequence. Second, we develop an approach to estimate the repetition count of an exercise in the given session.

2 Dataset

We created a dataset of home-based physical therapy exercises (HPTE) to demonstrate shoulder and knee exercises by consulting physiotherapists. A total of 240 exercise sessions (30 for each exercise type) are stored as gray-level and depth videos. In these exercise sessions, five volunteers performed eight exercises in six series. The exercise sessions are restricted to contain one actor performing one exercise repeatedly. Details of the exercises with sample frames are shown in Fig. 2.

The gray-level and depth videos are captured by Microsoft Kinect sensor with NI framework [6]. The resolution of videos are set to 320x240 pixels. The fps value is selected as 25. Frames of depth and gray-level videos are stored as 256 gray level images. Time duration of exercise sessions varies between 30 seconds up to a minute. The depth sensor sometimes could not measure 11 bits per-pixel depth information due to reflection of surface etc (shown as black pixels in Fig. 2a,f,g). To solve this problem, we follow the same procedure as in [8].
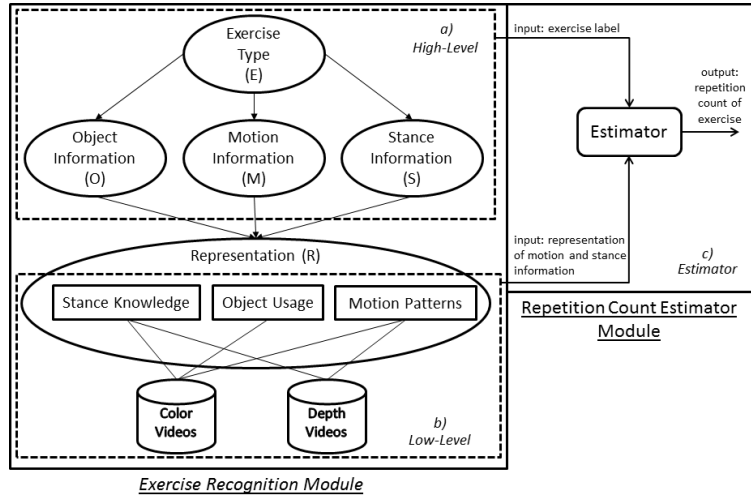
3 The Monitoring System

The design of the monitoring system contains an exercise recognition and a repetition count estimator module; the former is responsible for the exercise recognition process and the latter is responsible for the estimation of the repetition count, as shown in Fig. 1. Exercise recognition module is divided into two parts: low-level and high-level.

We believe that the key patterns in an exercise video sequence are motion, stance, and object information. The low-level part builds representations of these key patterns by using gray-level and/or depth videos. Representations are clustered as R

node in Fig. 1 to provide a better view of the graphical model. The high-level part contains of a generative Bayesian network which uses the graphical model in Fig. 1a to represent conditional independence relation between key patterns. The high-level part also benefits from the relations between object, stance, and motion information as described in Fig. 2.

The repetition count estimator module is dependent on the outputs of the exercise recognition module. This module gets exercise label, motion and stance representation as inputs and outputs the repetition count for the given exercise session.

Fig. 1 The design of the monitoring system.

4 The Low-Level Part of Exercise Recognition Module

The low-level part of exercise recognition module utilizes gray-level and depth videos to form representations of motion, stance, object information.

4.1 Representation of Motion Information

Motion information in videos is the main element of exercise recognition. Exercises have different motion patterns as in Fig. 2. To represent motion information in a video by motion patterns, we employed our previous method in [1]. First local motion information is obtained from depth, gray-level, or both videos by histogram-

ming 3D Haar-like features. Then statistical methods are used to describe motion information in the whole sequence as a global representation. We called the motion information for a given video sequence, which is obtained by our previous method, as *MI*.

4.2 Representation of Stance Information

Stance/pose information about an exercise session supports the other information sources when there are problems like occlusion, noise, high differences in temporal variances or etc.

First static background images are formed for a given depth and/or gray-level video by using the first few frames. Next a foreground extraction is performed for the selected frames (20 frame out of a video) of the given video and silhouette images are produced by thresholding. If both depth and gray-level videos are used, silhouette images are merged by the morphological union operation. Then, the largest blob in each silhouette image is windowed, these windows are parsed into 3 different grids with sizes 6x8, 8x8, and 8x6. Finally, the mean ratio of the foreground pixels in each cell of the each grid is calculated to form Stance Information *SI* vector.

4.3 Representation of Object Information

While most of the physiotherapy exercises include object interaction, object information in the video sequences reveals important clues about the type of these exercises.

First frames are selected at predefined uniform time-intervals (one frame out of 20 frames) in order to represent object information for a given video sequence. Then the object detection algorithm in [3], which uses bag of words models to detect objects, is adopted to check the availability of the corresponding object in the selected frames. The count of frames which includes the corresponding object are calculated and divided by the total number of selected frames. These ratios $OI(v,o)$ (where v is the video id, o is the object id) represent the object information in the video sequence.

5 The High-Level Part of Exercise Recognition Module

The high-level part of exercise recognition module aims to classify the exercise in the given video by using the representations obtained at the low-level part. We prefer to define a generative Bayesian network structure for the assigment of exercise label $e \in E$ to the video of the each exercise session because of the robustness of

Bayesian networks for representing of joint distributions and encoding conditional independence assumptions.

The generative Bayesian network uses the graphical model in Fig. 1a to represent conditional independence relationships between random variables: exercise (E), object information (O), motion information (M), stance information (S), and Representation (R). Label assignment process $L(r)$ is defined as

$$L(r) = \underset{e \in E}{argmax} \sum_{S,M,O} P(E,S,M,O,R), \qquad (1)$$

where r is the representation ($r \in R$) of the given video and $P(E,S,M,O,R)$ is the joint probability distribution table. $P(E,S,M,O,R)$ is defined by using the conditional dependencies in the graphical model (Fig. 1a) as

$$P(E,S,M,O,R) \propto P(E)P(S|E)P(M|E)P(O|E)P(R|S,M,O), \qquad (2)$$

where $P(E) = 0.125$ (because of eight different exercises), $P(S|E)$, $P(M|E)$, and $P(O|E)$ terms can be calculated easily by Fig. 2. $P(R|S,M,O)$ term needs to be converted by using axioms as

$$P(R|S,M,O) = \frac{P(S,M,O|R)P(R)}{P(S,M,O)}. \qquad (3)$$

$P(R)$ and $P(S,M,O)$ values in above equation are neglected because these are the same for any given exercise sessions. $P(S,M,O|R)$ is efficiently represented as

$$P(S,M,O|R) \propto P(S|R)P(M|R)P(O|R). \qquad (4)$$

Finally, the values of $P(S|R)$, $P(M|R)$, and $P(O|R)$ are needed to calculate exercise label $L(r)$. $P(O|R)$ is equal to the $OI(v_r,o)$ obtained at the end of representation of object information process. $P(S|R)$ and $P(M|R)$ are related to SI and MI, respectively. For this relation, linear kernel Support Vector Machines (SVMs) are trained and then the Gibbs distribution is used to translate SVM scores into predictions.

6 Repetition Count Estimator Module

A home-based physiotherapy exercise session consists of a number of repetitions of the same exercise. It is important to record the repetition count for treatment analysis.

A new sub-global representation $SGR(\tau)$ for exercise session s is defined as

$$SGR(\tau) = MI_b(s_\tau)||SI_b(s_\tau), \qquad (5)$$

where s_τ is the sub-sequence of s from frame 0 to τ, $MI_b(s)$ is the motion information about s by using both the gray-level and the depth video of s, and $SI_b(s)$ is the

stance information about s by using both the gray-level and the depth video of s. The exercise label for $GRS(\tau)$ is the same as $L(r)$, where r describes the representation of s, because s contains the same exercise $e \in E$ with different repetition counts. Confidence value (CV) for $SGR(\tau)$ is produced by using the remaining sessions in the dataset as training set and defining a new SVM formulation as

$$CV(\tau) = \sum_i a_i k(su_i, SGR(\tau)) + b, \qquad (6)$$

where su_i describes the support vectors, a_i describes weights, b describes bias, and k describes the kernel function. It is important to mention that the training set are divided into two groups as $L(r)$ labeled videos and the others. Finally, the examination of CV with increasing τ indicates the repetition count. The count of zero crossings in the derivative of $CV(\tau)$ with respect to τ would produce the exercise repetition counts.

7 Experimental Results

We evaluated our system with leave-one-actor-out procedure in each experiment and listed the results in the form of confusion matrix in Table 1.

Table 1 Exercise recognition module's recognition results on HPTE dataset. In the table, x/y means that x is obtained without using depth videos, y is obtained using both gray-level and depth videos.

TABLE 1	Stick	Dia-stick	Lie back	Towel	Str-pen	Cir-pen	Chair	Heel
Stick	28/29	2/1	0/0	0/0	0/0	0/0	0/0	0/0
Dia-stick	3/1	27/29	0/0	0/0	0/0	0/0	0/0	0/0
Lie back	1/1	0/0	28/29	0/0	0/0	0/0	0/0	1/0
Towel	1/1	0/0	0/0	29/29	0/0	0/0	0/0	0/0
Str-pen	0/0	0/0	0/0	0/0	25/28	5/2	0/0	0/0
Cir-pen	0/0	0/0	0/0	0/0	3/1	27/29	0/0	0/0
Chair	0/0	0/0	0/0	0/0	2/1	1/0	27/29	0/0
Heel	0/0	0/0	1/0	0/0	1/0	0/0	1/1	27/29

The exercise recognition module successfully recognized 90.8% of the 240 exercise sessions by using only gray-level videos. The most misclassified exercises were circular and straight pendulum exercises. There is a circular motion in circular pendulum exercise but this motion appears as a straight motion without depth information and caused misclassification.

The exercise recognition module successfully recognized 96.25% of the 240 exercise sessions by using both gray-level and depth videos. The general misclassification error between straight and circular pendulum exercises was greatly reduced

by using depth videos. As a baseline method [2] achieved 80.8% recognition rate on HPTE dataset as the mean of the gray-level and depth sequences.
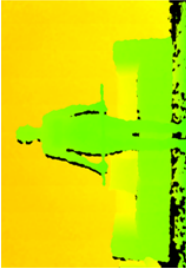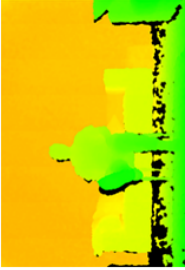
Repetition count estimator module estimated the repetition count of the 211 exercise sessions correctly with 88.0% accuracy rate by using the ground truth labels (manually labeled). Using the obtained labels from exercise recognition module (with depth and gray-level videos), our module estimated the repetition count of 204 exercise sessions correctly with 85.0% accuracy rate. The majority of incorrect estimation of repetition counts were observed in towel and circular pendulum exercise sessions.

8 Conclusions

In this paper, we propose a monitoring system for home-based physiotherapy exercises by using gray-level and depth videos obtained from the Microsoft Kinect sensor. The experimental results showed that the proposed system can effectively recognize the exercise in the given exercise session. We also observed that the monitoring system estimates the repetition count of the exercises in the given exercise sessions with encouraging results. To the best of our knowledge, the proposed system is the first system to monitor home-based exercises that includes objects by using a low-cost Microsoft Kinect sensor.

References

1. Ar I, Akgul YS (2012) A framework for combined recognition of actions and objects. In: International conference on computer vision and graphics, Warsaw
2. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE TPAMI, doi:10.1109/34.910878
3. Fei-Fei L (2007) Bag of words models: recognizing and learning object categories. In: CVPR short courses, Minnesota
4. Fitzgerald D, Foody J, Kelly D, Ward T, Markham C, McDonald J, Caulfield B (2007) Development of a wearable motion capture suit and virtual reality biofeedback system for the instruction and analysis of sports rehabilitation. In: Proceedings of the 29th annual international conference of the IEEE EMBS, Lyon
5. Jung Y, Kang D, Kim J (2010) Upper body motion tracking with inertial sensors. In: Proceedings of the 2010 IEEE international conference on robotics and biomimetics, Tianjin
6. OpenNI, www.openni.org
7. Soutschek S, Kornhuber J, Maier A, Bauer S, Kugler P, Hornegger J, Bebenek M, Steckmann S, Stengel SV, Kemmler W (2010) Measurement of angles in time-of-flight data for the automatic supervision of training exercises. In: 4th international conference on pervasive computing technologies for healthcare, Munich
8. Xia L, Chen CC, Aggarwal JK (2011) Human detection using depth information by kinect. In: Workshop on human activity understanding from 3D data in conjunction with CVPR, Colorado Springs
9. Zhou H, Hu H (2008) Human motion tracking for rehabilitation-A survey. Biomed. Signal Process. Control, doi:10.1016/j.bspc.2007.09.001

Ilktan Ar and Yusuf Sinan Akgul

Fig. 2 Details of the exercises in HPTE dataset. In the each cell of the figure, the exercise types in the dataset are displayed with a sample figure followed by the related object, stance, and motion descriptions. The sample frames are taken from gray-level videos (b,c,d,e,h) and depth videos (a,f,g).