This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2014.2326254, IEEE Transactions on Neural Systems and Rehabilitation Engineering

A Computerized Recognition System for the Home-Based Physiotherapy Exercises Using an RGBD Camera

Ilktan Ar and Yusuf Sinan Akgul

***Abstract*—Computerized recognition of the home based physiotherapy exercises has many benefits and it has attracted considerable interest among the computer vision community. However, most methods in the literature view this task as a special case of motion recognition. In contrast, we propose to employ the three main components of a physiotherapy exercise (the motion patterns, the stance knowledge, and the exercise object) as different recognition tasks and embed them separately into the recognition system. The low level information about each component is gathered using machine learning methods. Then, we use a generative Bayesian network to recognize the exercise types by combining the information from these sources at an abstract level, which takes the advantage of domain knowledge for a more robust system. Finally, a novel post-processing step is employed to estimate the exercise repetitions counts. The performance evaluation of the system is conducted with a new dataset which contains RGB (Red, Green, and Blue) and depth videos of home-based exercise sessions for commonly applied shoulder and knee exercises. The proposed system works without any body-part segmentation, bodypart tracking, joint detection, and temporal segmentation methods. In the end, favorable exercise recognition rates and encouraging results on the estimation of repetition counts are obtained.**

***Index Terms*—Home-based Physiotherapy, Exercise Recognition, Estimation of Repetition Count, Bayesian Network.**

I. 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 [1]. These treatments usually include performing physiotherapy exercises regularly in a controlled manner [2].

Physical therapists provide care to patients by adjusting therapy parameters and supervising the therapy sessions. They perform this task by a combination of verbal instruction, demonstration, and physical guidance during and/or before the execution of the physiotherapy exercise in a session [3]. With the physical guidance, patients can repeat the desired exercise by improving their ability to detect and correct errors [4]. Manual feedbacks obtained at the error detection and correction phase, which are important steps in motor learning [5], improve the success of the rehabilitation process.

Inherent costs, travel required to receive in-clinic treatment, therapy accessibility and availability are the main obstacles that limit one-to-one sessions between the therapists and the patients [6]. Due to these problems, patients are asked to perform some parts of their therapy sessions at home. To gain information about these sessions and improve the treatment process, home-based physical therapy sessions must be analyzed.

Machine learning based motion recognition research has recently made important progress towards building low cost, usable, stable and accurate action recognition systems from video data for controlled environments [7], [8]. The idea of employing such methods for the recognition of the home based physiotherapy exercises has attracted interest among the computer vision community [9], [10], [11]. However, considering the recognition of the physiotherapy exercise problem as an instance of only motion recognition misses some crucial domain knowledge. A physiotherapy exercise has three main components [12]: the motion patterns of the exercise, the stance position of the exercise, and the exercise object. The motion patterns focus on the exercise's speed, acceleration in the motion and the displacement of patients in the sessions. The stance knowledge is needed to analyze the posture information of the patients in the sessions (start position, end position). The object existence in the exercise increases the confidence of patients and accelerates the treatment process. We argue that any automated exercise recognition system should explicitly use this fundamental domain knowledge for a robust real world system.

This paper extents our previous work [1] for the home monitoring of physiotherapy exercises using a Red-Green-Blue-Depth (RGBD) camera with further analysis and experiments. Our system combines the information from the three physiotherapy components to capture domain related information for the exercise type recognition. For the validation of our work, we create an RGBD dataset of Home-based Physical Therapy Exercises (HPTE) to demonstrate shoulder and knee exercises by consulting physiotherapists. We also propose a novel approach to estimate the repetition count of an exercise in a given session.

At the center of our system, we employ a Bayesian network that consists of hidden nodes for the exercise type, motion patterns, stance position, and exercise object (Fig. 2). The observable nodes of the Bayesian network are for the observed image

Manuscript received September 11, 2013; revised April 2, 2014; accepted May 14, 2014.

Ilktan Ar and Yusuf Sinan Akgul are with the Vision Lab, Department of Computer Engineering, Gebze Institute of Technology, Gebze, Kocaeli, 41400, Turkey. (email: ilktana@khas.edu.tr; akgul@bilmuh.gyte.edu.tr)

Ilktan Ar is also with the Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Kadir Has University, Cibali, Istanbul, 34083, Turkey.

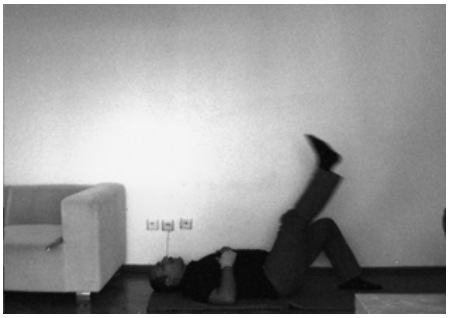
For vision lab see http://vision.gyte.edu.tr

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2014.2326254, IEEE Transactions on Neural Systems and Rehabilitation Engineering

TABLE I
DETAILS OF THE EXERCISE TYPES IN HPTE DATASET

a) Stick Exercise
Object: stick, Stance: standing
The patient stands and holds an object with his hands. While he keeps his elbows in upright position, he raises the object slowly above his head and lowers the object.

b) Diagonal-Stick Exercise (dia-stick)
Object: stick, Stance: standing
The patient stands and holds an object with his hands. Then he holds the object for several seconds as in the above figure and returns to the beginning position.



c) Lie Back Exercise
Object: N/A, Stance: lying down
While the patient lies on her back, she raises her leg without twisting her knee. She holds her leg in stretched position for a while. Then she pulls her leg to the beginning position. Exercise continues with the other leg.

d) Towel Exercise
Object: towel, Stance: standing
The patient holds the object above his shoulder with one hand and holds the object on his back with the other hand. Then he stretches his arm by pulling the object with lower hand and leaves stretching.

e) Straight Pendulum Exercise (str-pen)
Object: chair, Stance: bending
The patient stands and holds the object with one hand. Next she bends forward slightly. Then she dangles her arm forward. Finally the patient starts swinging her arm forward and backward for 30 seconds.

f) Circular Pendulum Exercise (cir-pen)
Object: chair, Stance: bending
The same exercise as straight pendulum only the swinging is in a circular manner.

g) Chair Exercise
Object: chair, Stance: sitting
While the patient sits on the object, he stretches out one of his leg to forward. He holds his leg in stretched position for a while. Then he pulls his leg to the beginning position. Exercise continues with the other leg.

h) Heel Exercise
Object: chair, Stance: sitting
The patient sits on the object. He moves his foot to the back of the object with raising his heel to the up. He holds his foot in position for a while. Exercise continues with the repetition of the same action with the other foot.

i) Depth image of g).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2014.2326254, IEEE Transactions on Neural Systems and Rehabilitation Engineering

features that are calculated for each component separately by running different feature extractors. The Bayesian network enables us to make inferences about the motion, stance, and object components separately while the interactions between these components are handled automatically by the conditional dependency resolution mechanism of the network.

The rest of this paper is organized as follows. Firstly, a survey of relevant literature is presented in Section II. Section III describes the Home-Based Physical Therapy Exercises (HPTE) dataset and a brief definition of the system with the dataflow diagram is sketched in Section IV. The recognition layer of the system that contains the generative Bayesian network is discussed in Section V. Feature extraction based on motion, stance, and object information is represented in Section VI. The estimation of repetition count is defined in Section VII. Finally, Section VIII demonstrates the experimental results of the system and conclusions are drawn in Section IX.

II. Related Work

The available systems which are designed to recognize physical therapy exercises can be categorized as sensor-based and camera based. Although sensor-based systems require specific adjustments and use expensive specialized hardware, they are successful at detecting and tracking joint points. The camera based systems employ common electronic devices such as Microsoft Kinect. They are not as accurate as the sensor-based systems, but they are very practical and inexpensive.

Zhou and Hu [9] surveyed the detection and tracking systems in human motion tracking for the rehabilitation. They reported that this is an active research topic since the 1980s and the existing systems have various problems such as need of engineers and physiotherapists to perform calibration and sampling, hardware cost, and functional problems. The calibration process must be made individually for each patient and adjusted during the recovery time, which is a resource (time, money, etc.) consuming process. In addition such a calibration is not feasible at home environments. Soutscheck *et al.* [13] presented a system to support and supervise fitness and rehabilitation exercises by observing angular measurements of the knee joints. Their system uses specialized sensors to track 2D and 3D knee positions. Fitzgerald *et al.* [14] developed a system which includes a computer game to instruct and analyze an athlete's rehabilitation exercise series. Their system uses ten inertial motion tracking sensors in a wearable body suit and a portable computer which communicates with this suit by a Bluetooth connection. Kesner *et al.* [15] designed a wearable upper body orthotics system for home-based rehabilitation which also contains a limb position sensing system. They dealt with shoulder joint rehabilitation and focused on adjustability, wearability and adaptability of this orthotics system. Roy *et al.* [16] developed a remote monitoring system for physical activities by using body-worn sensors with a neural network and fuzzy logic processing technique. Courtney and de Paor [17] presented a markerless system for analyzing human gait with a single camera. Jung *et al.* [18] introduced a tracking system for upper body movement in 3D space by using wearable inertial sensors. Guralic *et al.* [19] classified some limb movements (activities) in physical therapy processes by using the data collected with wearable wireless transceivers. Most of the sensor-based systems provide only constrained solutions to the problems of exercise recognition due to some inherent constraints. These constraints include the non-portability of the sensors [13], non-ease of use by the patients [13], high system costs [14], being limited by only a few exercise types [19], missing exercise objects [18], requirement of a clinical environment [13], requirement of specialists [13], [15], and difficulties in the system customization for each patient and recovery stage [15].

Timmetmans *et al.* [20] presented a sensor-based system to monitor the task oriented arm training of patients. This system is composed of tracking sensors and an exercise board. The context-specific sensor motor input is measured and evaluated with kinematic information. Li *et al.* [21] designed a multimodal physical activity recognition system for a wearable wireless sensor network. Their system uses both ambulatory electrocardiogram (ECG) and accelerometer signals. Duff *et al.* [22] developed a mixed reality training system for stroke rehabilitation to improve the reaching movements of patients. This system records the movements with a 10-camera 3-D infrared passive motion capture system and outputs audios and visual feedbacks for the therapists. Increasing the number of sensors and sensor types provides increments in the range of exercise types and generates better outputs. However, these increments also limit the sensor-based systems to run on more controlled environments [20], make them more complex to customize [21], and render them more expensive for home use [20], [22].

Microsoft Kinect made a considerable impact as an RGBD camera with a provided software library to detect human joints for the physiotherapy exercises. Chang *et al.* [23] used a Kinect-based system in a public school setting to motivate physical rehabilitation. They worked with two young adults with motor impairments. The system increased patients' motivation to continue the rehabilitation tasks. Lange *et al.* [24] investigated the use of low cost depth sensing technology to project full-body interaction within virtual reality and game-based environments. They developed a system that records body-part positions for the rehabilitation tasks by using these products. Nie *et al.* [25] developed a Kinect based system for rehabilitation. This system tracks full body motion, records the related information and delivers this data to the physical therapists. Note that most Kinect based monitoring systems rely on human joint detection library that comes with the device. This library is very robust under different conditions but it is not designed to be used while the subject is in interaction with exercise objects such as chairs, sticks, and balls. In addition, this library provides only suboptimal results for some patient poses such as lying on the ground. As a result, for the practical physiotherapy exercises, depending on joint positions from this library has robustness problems. Although we also use a Kinect sensor in our system, we use it for only obtaining RGBD data from from the patient and we do not rely on any joint detection information.

The exercise recognition problem can be considered in the context of human action recognition topic which is one of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2014.2326254, IEEE Transactions on Neural Systems and Rehabilitation Engineering

the most challenging topics in computer vision [7], [8], [26], [27]. Most of the above systems are based on action recognition techniques for this task. However, there are differences between physiotherapy exercises and simple human actions. Human action recognition systems generally classify primitive actions such as: walking, bending, hand waving, jumping, etc [9]. Physiotherapy exercises are different from primitive actions in terms of object usage, action duration, and movement complexity. Therefore, recognition of physiotherapy exercises, requires more specific systems than the systems designed to recognize basic human actions which do not tackle object interactions.

III. Dataset

We provide a dataset with 8 classes of exercises in 6 sessions with 5 volunteers, namely Home-based Physical Therapy Exercises (HPTE) dataset. This dataset contains a total of 240 color and corresponding depth videos to demonstrate shoulder and knee exercise sessions, which are the most common exercise types [12]. These videos are captured by a Microsoft Kinect sensor. The duration of each session varies between 15 seconds up to 30 seconds. Table I shows sample frames for each exercise type with the motion, object, and stance information.

For each HPTE session, we assume that only one patient performs one type of exercise with more than two repeats, in front of the Microsoft Kinect sensor. Each actor performed the given exercises completely. Subpar exercises are not allowed. We also assume that only the related exercise object is present on the scene.

Microsoft Kinect sensor provides color and depth videos with 640x480 pixels at 30 fps. We store color and depth videos as 256 gray level images in 320x240 pixels resolution by using The Kinect for Windows SDK version 1.5 [28]. The depth sensor sometimes could not measure 12 bits per-pixel depth information due to surface reflection, shadow, etc. To solve this problem we follow the same procedure as outlined in [10]. This procedure starts with the nearest neighbor interpolation to fill non-measured points and continues with Median filtering (4x4 sized) to smooth the depth frame. The HPTE dataset is available upon request.

IV. The System Overview

The main modules and the flow of the data between the modules of our system are shown in Fig. 1. Briefly, our system gets color and depth videos of physiotherapy sessions and finds the exercise type and then outputs the repetition count of these exercises.

The feature extraction module extracts motion patterns, stance knowledge, and object usage information as low-level features. The exercise recognition module employs a generative Bayesian network to recognize the exercise types in the sessions. This module uses machine learning based classifiers to collect evidences about the global motion, stance, and object availability information. The last module, estimation of the repetition count, gets exercise label, motion and stance representation as input and outputs the repetition count for the given exercise session. The details about each system module

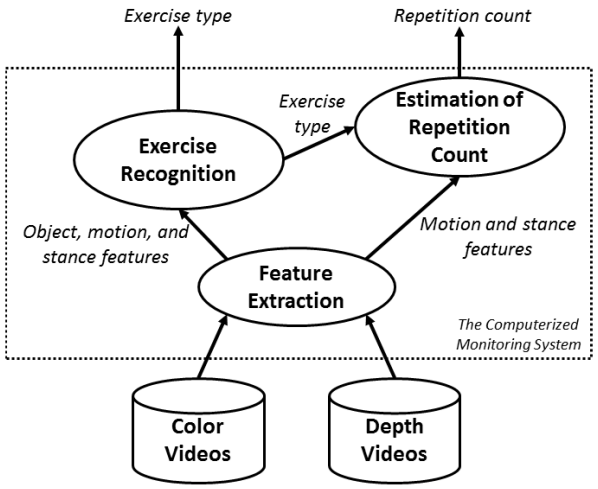


Fig. 1. The dataflow diagram of the proposed system.

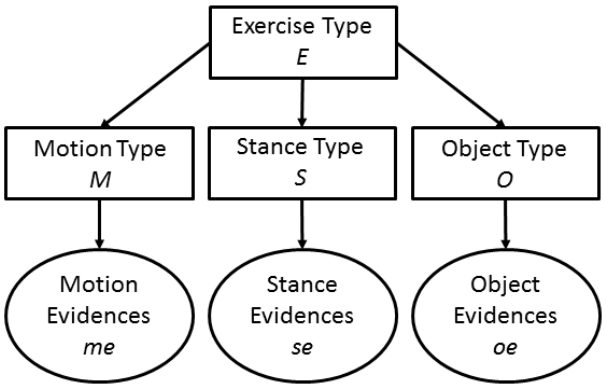


Fig. 2. The graphical model of the generative Bayesian network. In this model, rectangles indicate hidden nodes and circles indicate observable nodes.

are given in the following sections starting with exercise recognition module.

V. Exercise Recognition

We define an 8-element exercise type set E={*stick-exercise, towel-exercise, chair-exercise, diagonal-stick-exercise, lie-back-exercise, straight-pendulum-exercise, circular-pendulum-exercise, heel-exercise*} that contains all exercise types (Table I). The main function of the exercise recognition module is to estimate the exercise type $e \in E$ for the given video session.

We design separate machine learning classifiers for each component of the physiotherapy exercises. Each classifier captures specialized domain related information without getting affected from the complexity of the main task. In order to handle the dependency relations between these components and to estimate the final exercise type, we employ a generative Bayesian network as the graphical model in Fig. 2. In this graphical model; exercise type E, object information O, motion information M, and stance information S are the hidden nodes; the object evidences oe, motion evidences me, and stance evidences se are the observable nodes.

Given the evidences from the video data, the exercise label

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2014.2326254, IEEE Transactions on Neural Systems and Rehabilitation Engineering

assignment process $L(v)$ is defined as

$$L(v) = \underset{e \in E}{argmax} \sum_{S,M,O} P(e, S, M, O, me, se, oe), \quad (1)$$

where v is the given video and $P(e, S, M, O, me, se, oe)$ is the joint probability distribution table. $P(e, S, M, O, me, se, oe)$ is defined by using the conditional dependencies within the graphical model as

$$P(e) \prod P(S|e)P(M|e)P(O|e) \\ P(se|S)P(me|M)P(oe|O). \quad (2)$$

$P(e)$ is 0.125 because there are eight different equally likely exercises in HPTE dataset. $P(S|e)$, $P(M|e)$, and $P(O|e)$ terms are calculated by using the relationships listed in Table I. For example, if the exercise e is *chair-exercise* then $P(S = standing|e = chair)$ is equal to 0 while $P(S = sitting|e = chair)$ is equal 1 and if the exercise e is *heel-exercise* then $P(O = towel|e = heel)$ is equal to 0 while $P(O = chair|e = heel)$ is equal to 1.

$P(se|S)$,$P(me|M)$, and $P(oe|O)$ can be written as

$$P(se|S) = \frac{P(S|se)P(se)}{P(S)}, \\ P(me|M) = \frac{P(M|me)P(me)}{P(M)}, and \quad (3) \\ P(oe|O) = \frac{P(O|oe)P(oe)}{P(O)}.$$

The probabilities of $P(se)$, $P(me)$, $P(oe)$, $P(S)$, $P(M)$, and $P(O)$ in the above equation are the same for a given exercise video v and hence they do not have to be estimated. However, to find the exercise label $L(v)$, we need to calculate $P(S|se)$, $P(M|me)$, and $P(O|oe)$.

Most of the classical exercise recognition systems [9], calculate only $P(M|me)$ to find the exercise type e. Our system, on the other hand, combines the motion information ($P(M|me)$) with the information from exercise object ($P(O|oe)$) and patient stance ($P(S|se)$), which should produce more robust exercise recognition rates. In addition, bringing these types of information under a well known Bayesian Network framework makes our system theoretically sound.

In order to calculate the values of $P(S|se)$, $P(M|me)$, and $P(O|oe)$, we employ supervised machine learning based classifiers. These classifiers obtain features from the video or image data and train them with the given labels. During the testing time, the same features are calculated and an appropriate label is assigned for the given test data. These classifiers are very successful if the number of features extracted from the video or image data is too high for a manual feature selection. For our application, we extract thousands of features from each frame of the video data, so a machine learning based label assignment is appropriate.

We train Support Vector Machine (SVM) classifiers for the label assignment problems of $P(S|se)$ and $P(M|me)$. These classifiers take the features me and se, and they assign labels M and S for the given video. However, classical SVM classifiers do not produce a probability value for the assigned label which is needed to estimate the values of $P(S|se)$ and $P(M|me)$. In order to calculate these probabilities, we employ a Gibbs distribution based approach. The probability of M for the given motion evidence me is calculated as

$$P(M|me) = \frac{1}{Z} exp(-Q(me)) \quad (4)$$

and the probability of S for a given stance evidence se is calculated as

$$P(S|se) = \frac{1}{Z} exp(-Q(se)) \quad (5)$$

where Z is the normalizing constant, and the potential functions $Q(me)$ and $Q(se)$ are the evidence predictions assigned by the SVM classifiers.

Our exercise-object classifier also uses machine learning techniques with a slightly different video sampling approach. We employ an object detection method from the literature that decides on the object availability $OA(o, f)$ by searching an object o in a given frame f and outputs as 1 or 0 (found or not found). We follow the uniform sampling approach to specify the object usage information for the whole video v by taking one frame out of 20 frames. Finally the probability of O for a given object evidence oe is calculated as

$$P(O|oe) = \frac{\sum_{i=1}^{n} OA(oe, f_i)}{n}, \quad (6)$$

where f are the selected frames of v and n is the total number of selected frames.

VI. Feature Extraction

The feature extraction module provides low-level information to the other modules. Although we provide a set of novel features for the exercise recognition task, this task can be completed with other types of feature extractors without affecting the main operations of other system modules. The following subsections describe the extracted features for motion, stance, and exercise object information.

A. Motion-Based Features

Motion information in an exercise session is represented by motion patterns. Each exercise session consists of repetitions of a type of exercise. Each exercise type has different motion patterns as shown in Table I. We propose a novel approach for obtaining the motion patterns in a session by enhancing Haar-like features.

Haar-like features are popularly used in pattern recognition. These features were first proposed by Papageorgiou *et al.* [11] as the difference between the intensity sums of all pixels in rectangular boxes. Ciu *et al.* [29] used the 3D Haar-like features to detect pedestrians. These features are extracted in a space-time volume as cubic filters. They also extended the idea of integral images to integral volumes for computational efficiency.

In our approach, we employ 3D Haar-like features to represent the motion patterns contained in the whole video sequence. 16 different cubic filters are designed to capture variations within patient's 3D spatiotemporal space. We also include filters tuned for exercises performed at different

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2014.2326254, IEEE Transactions on Neural Systems and Rehabilitation Engineering

speeds. Some of these filters are shown in Fig. 3. Filter sizes are determined as 8x8 pixel and 4x4 pixel sizes in spatial domain, 4 frame and 8 frame in temporal domain. Further evaluations on the parameter selection are discussed in Section VIII.

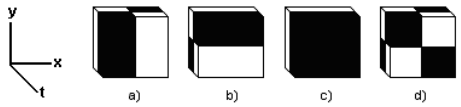


Fig. 3. Examples of cubic filters which are used in extraction of 3D Haar-Like features.

3D Haar-like Features ($3DHFs$) are extracted by applying cubic filters to the video (v) with the convolution process as

$$3DHF_v(x, y, t, f) = v(x, y, t) * CubicFilter_f, \quad (7)$$

where t is the time step or the video frame number, x and y are 2D coordinates of a video frame, and *f=1...16* is the type of the filter. The outputs of $3DHFs$, which are spatial, temporal, and filter-type dependent motion patterns, are normalized to the 0-255 interval for the calculation and storage efficiency. Short-Term Motion Patterns ($STMPs$), which describe spatially independent local motion information between consecutive frames, are defined as histograms of $3DHFs$ as

$$STMP_v(t, f) = Histogram[3DHF_v(x, y, t, f)]. \quad (8)$$

For a given video frame t and filter type f, (8) forms a histogram with 256 bins, which is adequate to differentiate between short term actions (ex. swinging of arm). However, when there is not any locally significant motion, different exercises can have similar $STMPs$. In addition, $STMPs$ are filter-type dependent patterns. To address these shortcomings of $STMPs$, we offer a new descriptor called Concatenated Short-Term Motion Patterns $CSTMPs$

$$\begin{aligned} CSTMP_v(t) = STMP_v(t, f_1)||STMP_v(t, f_2)||... \\ STMP_v(t, f_{n-1})||STMP_v(t, f_n) \end{aligned} \quad (9)$$

where f_i is the i^{th} filter, n=16 is the number of filters, and $||$ is the vector concatenation operation.

For a given video frame t and filter type f, (9) forms a vector with 256x16 elements. Since there is only one $CSTMP$ for each video frame, there are too many of them (for all video sessions) to be processed efficiently. In order to find the most effective and efficient subset of $CSTMPs$ for a given video, we use a recent popular approach [30] to transform $CSTMPs$ into codewords. This process can be accomplished by K-means clustering, mean-shift clustering, or any similar techniques. However, recent studies on tree based codes offer more efficient and quicker solutions [30], [31], [32].

Random Forests (RF) were introduced by Breiman [33] as a collection of decision trees. During the training stages, nodes in the trees are split by using a random subset of input data. Then each random tree in the forest grows and predicts the input test data's class label. Finally, these votes are aggregated to find the final label.

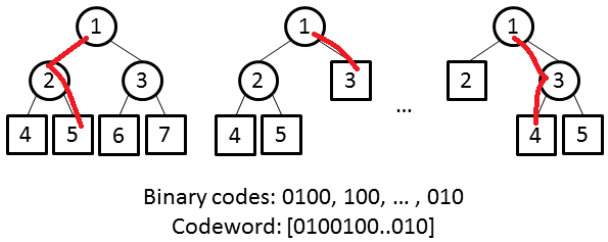


Fig. 4. Demonstration of extraction of codewords for motion information. Random Forest with a number of trees, which is trained with $CSTMPs$ used to obtain codewords for a given input vector. The selected paths for each tree are shown as a red curve.

We employ Random Forests and adopt the methodology in [30] to generate the codewords for $CSTMPs$. In this methodology, each decision tree in the forest generates a binary code for a given feature vector. The length of the binary code is equal to the leaf count of that tree. The value of the binary code is obtained by a simple process. First, each node in the tree is indexed by moving from top-left to down-right starting with the root node. The leaves which are ordered by indices, form the code. The leaf in which the feature vector falls into takes a value of one and the other leaves take value of zero. As the RF contains more than one decision tree, binary codes obtained from each tree concatenated into the final codeword (see Fig. 4 for example). In the end, we calculate a codeword for each frame of the video, which is shown as

$$CW_v(t) = Codeword(CSTMP_v(t)). \quad (10)$$

The size of each vector CW depends on the depth of each tree and the number of trees of the Random Forest. After obtaining codewords (CWs) for each $CSTMP$ in a video sequence, we need a descriptor to obtain the general motion information about the whole video sequence. We call this descriptor as Global Motion Vector GMV and define it as

$$GMV(v) = \mu(CW_v(t), t = 1..T)||var(CW_v(t), t = 1..T) \quad (11)$$

where T is the number of frames in the video v, μ describes the statistical mean of a set of vectors, and var describes the variance. The size of the vector GMV is two times the size of one CW vector. Note that GMV can be built for color or depth video separately or combination of them by concatenating the related GMV. Note also that GMV of a video is our final motion feature vector which is used by (4).

B. Stance-Based Features

The physical therapists view the posture of a person to investigate the patient's disorder. Moreover, stance/pose information about an exercise type is a discriminative property because each exercise type has a well defined posture. For example, consider two scenarios where the first scenario includes a person who is holding a thin stick with *two* hands and raising-lowering the stick. The second scenario includes another person who is holding a thick stick with *one* hand and raising-lowering the stick. Motion information for both scenarios reveals that there is a raising-lowering motion pattern

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2014.2326254, IEEE Transactions on Neural Systems and Rehabilitation Engineering

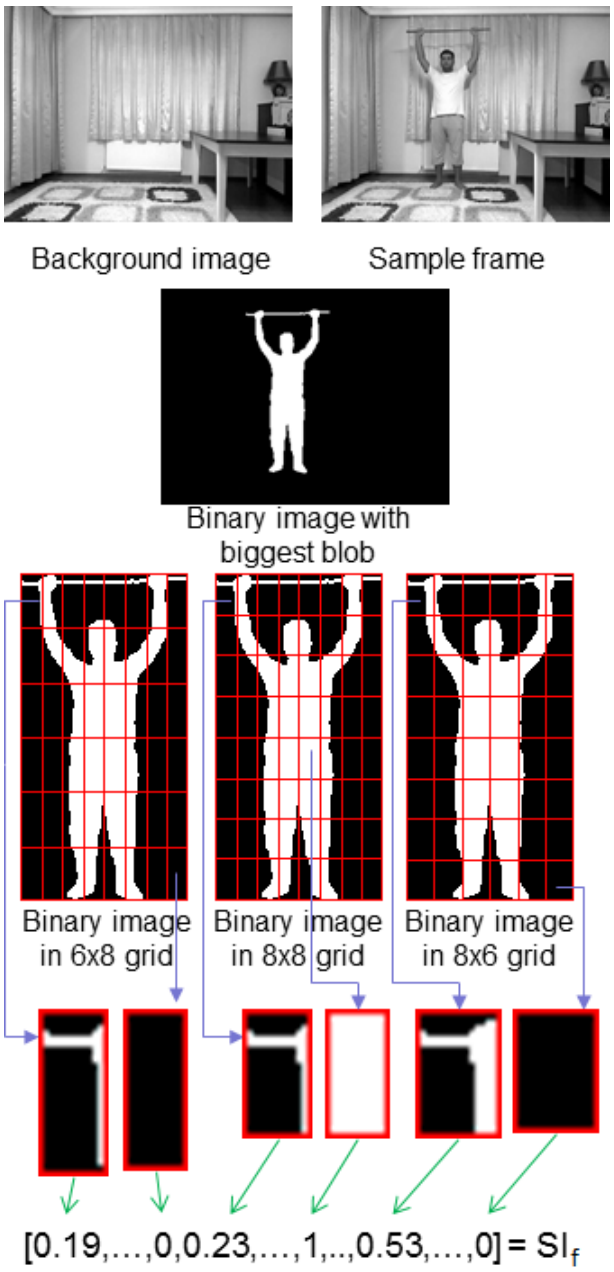


Fig. 5. The extraction of stance-based features. Stance Information SI vector is defined for a given frame f. Note that the binary image can be formed by color, depth, or both frames.

available. The object information reveals that there is a stick in both scenarios. However, the stance information reveals the difference between poses of the first and second person and distinguish the exercises in these scenarios. Also, when the problems such as occlusion, noise, high differences in temporal variances, etc. arise, motion and object information may suffer. Furthermore, studies such as [34] inform that the recognition process benefits from pose-based features even with noise.

We propose a method to extract the stance information of the patient. As mentioned earlier, it is possible to use any existing and suitable method to extract motion, stance, and object based information in our system. This option also increases the system's level of customizability.

In the proposed method, we form a simple but effective approach to represent the Stance Information SI. First, for a given video, we create a static background image using the first few frames of the video where the actor is not on the scene yet. Next, we perform a foreground extraction with this selected frame and the background image. This process yields a binary image by thresholding. If both depth and color videos are used, binary image is merged by the morphological intersection operation. Then, we find the largest blob in the binary image and capture it in a window by assuming that the largest blob corresponds with the actor. We parse this window into 3 different grids with sizes 6x8, 8x8, and 8x6. Finally, the ratio between foreground and background pixels in each cell of the each grid is calculated to form Stance Information $SI_{f,v}$ vector for the given frame f and the given video v. A brief demonstration of the extraction of SI is shown in Fig. 5.

In order to describe stance information about whole v, we can either calculate SI vector for each frame of v or we can select frames with uniform sampling and then calculate SI vectors for these selected frames. We prefer the latter approach which requires less CPU time. We select the frames in v at predefined uniform time-intervals (20 frames out of a video). Then we form the related SI vectors. To obtain a compact representation of stance information for v, we calculate $SI(v)$ by

$$SI(v) = \mu(SI_{v,f}), \tag{12}$$

where f indicates the selected frames in v and μ describes the mean.

The size of the vector $SI(v)$ for the whole video depends on the selected grid sizes (for our case; $48 + 64 + 48 = 160$). Note also that, SI of a video is our final stance feature vector which is used by (5).

As mentioned previously, our stance features use simple background subtraction method which may not be practical for some applications with dynamic backgrounds. In addition, although we did not explicitly used it, the depth information from the RGBD camera can be reliably used for foreground estimation.

C. Object-Based Features

Most of the physiotherapy exercises include object interaction to accelerate recovery period of the patients [35]. Moreover, the patients feel safer and motivated with the exercise objects in physiotherapy sessions. In addition, the type of the exercise object is an important factor in the design of the exercises. Therefore, information about the exercise object in the video reveals important cues about the type of the exercise and supports the two other information sources in our system.

We simply identify the exercise objects in the home environments as sticks, chairs, and towels within Home-Based Physical Therapy Exercises (HPTE) dataset. These objects do not have a standard size, type, color, etc. For example, we use three different objects for sticks: wooden ruler, wooden stick of a mop, and metal stick of a mop. Moreover, patients can also use different exercise objects such as canes, yardsticks,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2014.2326254, IEEE Transactions on Neural Systems and Rehabilitation Engineering

elastic ropes, etc. Note that we assume that only the related exercise object is present in an exercise session.

We represent the exercise objects in the physiotherapy sessions with the object-based features in our system. These features can be extracted by using appearance and/or location information in still images by using object recognition methods in the literature. We adopt three different state-of-the-art object detection and recognition methods to represent object usage information in a given exercise session. Note that these methods may fail for some situations such as occlusions due to people and scene. However, we employ the results from these methods in combination with other information (motion and stance), which makes our system robust against these types of problems.

The first method that we adopted from Fei-Fei [36], extracts SIFT features and then uses these features in a bag of words model to detect objects in a given image by utilizing a training set. We create a sample database of images of exercise object used in the HPTE dataset by storing 30 images for each object. Then, the object detection method is run to check availability of the object o in the given frame f. Finally, we calculate the object-based features by an object availability function $OA(o, f)$ where this function is equal to 1 if the o found in f, else is equal to 0.

We adopt the object detection approach of Viola and Jones [37] as the second method. Briefly, their approach uses a variant of cascaded Adaboost classification technique to find the region of images which contain instances of a certain kind of object. Each image is represented with 2D Haar-like features (rectangular features) by an integral image approach.

We modify the Viola and Jones's [37] approach (which detects the regions of faces in images) to find the region of images which contains exercise objects. For this purpose, we create a training dataset which contains 500 positive images (exercise objects with different scales, rotations, and occlusions), 1000 negative images (part of actors with different poses). Then, we run their method and check the availability of the object o in the given frame f. Similar to the output of the first method, we design an object availability function $OA(o, f)$ which returns 1 if this approach finds a region of o and returns 0 in the other case. Note that we work with foreground RGB images (which contain only the patient and exercise object) which can be extracted with background subtraction using color, depth, or both frames.

The last object detection method is adopted from Gall and Lempitsky [38]. Their method detects instances of an object class, such as horses, pedestrians, or cars in the given images. They developed a class-specific Hough forest (a variant of Random forest) which detects the parts of objects and casts probabilistic votes for the possible locations of the selected object's centroids. Each node of the forest compares the similarity between the current region (window) and the training patches. Finally, the maxima of the Hough image that contains the votes of all parts, is found as object's centroid. We adopt their technique to find the window which captures the region of searched object and modify the last step of their method with a fixed threshold to lower the false positives. We form a training dataset which contains 1500 positive image patches and 3000 negative image patches. At the end of this method, we obtain a bounding box that contains the centroid of the searched object and the related parts or no bounding box (due to thresholding). We follow the same way to generate the object based output vector as in the previous two object detection methods.

VII. Estimation of The Repetition Count

A home-based physiotherapy session consists of several repetitions of the same exercise [25]. The repetition count in a session is assigned by physiotherapists. The repetitive nature of the exercise sessions improves recovery period and provides feedback for further evaluation. As mentioned earlier, the main task of our system is to recognize the exercise type in the given exercise session and the secondary task is to find and record the repetition counts in the given exercise session.

In this study, we form our database with exercise sessions which are fully performed and completed exercises. This means that the HPTE dataset does not contain any subpar (attempted but not completed) exercises. Note that as a secondary task of our system, estimation of the repetition count of exercises in a session (which is not supervised by a physiotherapist, home environment) is required for analysis of patient.

We propose a novel approach to estimate the repetition count. This approach uses the exercise label which is provided by the exercise recognition module and the stance and motion features which are provided by the feature extraction module (Fig. 1). First, a new sub-global representation vector $SGR(\tau)$ for exercise (session) video v is defined as

$$SGR(\tau) = GMV(v_\tau)||SI(v_\tau), \tag{13}$$

where v_τ is the sub-sequence of v from frame 0 to τ. The exercise label, for $SGR(\tau)$ is the same as $L(v)$ of (1). The main idea of the repetition count module of our system is to continuously run the binary SVM classifier on different values of τ. The binary SVM classifier tests the given subsequence wheather the sequence contains the exercise $L(v)$ or not. We then calculate the Confidence Value (CV) of the SVM classifier by the formulation as

$$CV(\tau) = \sum_i a_i k(su_i, SGR(\tau)) + b, \tag{14}$$

where su_i describes the support vectors, a_i describes weights, b describes bias, and k describes the kernel function. Finally, we find the number of peaks of CV graph with increasing τ as the repetition count. We calculate the number of peaks of the CV graph as the count of zero crossings in the derivative of CV(τ) with respect to τ.

Fig. 6 shows a sample CV(τ) curve for a selected exercise session. As the duration of a sub-sequence increases, the system recognizes the exercise more accurately (with a higher confidence value). The local peak points of the confidence curve correspond to exercise repetition startings and ends. Therefore, counting the local peaks on confidence curves would produce the number of exercise repetitions. Note that the confidence curves might become constant for videos with thousands of frames, which can be addressed using a windowed SGR formulation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2014.2326254, IEEE Transactions on Neural Systems and Rehabilitation Engineering

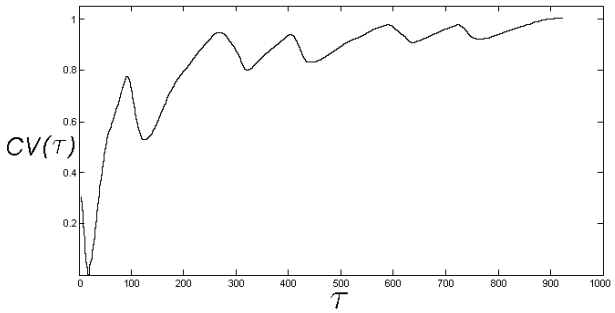


Fig. 6. A sample Confidence Value CV(τ) curve for an exercise session.

VIII. Experimental Results

We design five different sets of experiment to test the proposed system's performance. We use the same parameters for all the experiments. These parameters are: the maximum depth level of Random Forests is 5, the maximum tree count for the forests is 24, and the number of selected features in each node is the square root of the total feature count. The RGB (color) images are converted to 8 bit grayscale images. We complete all the experiments with the leave-one-actor-out manner, so we test the each sample at least once.

In the first set of experiments, we evaluate our system on the Weizmann Human Action dataset [39] using only proposed motion and proposed stance features to compare our action recognition performance with the literature. The Weizmann dataset contains 10 primitive actions performed by 9 people (total 90 sequences) and it is very popular as a baseline benchmark test among the action recognition community. Our system recognizes 84 of the 90 actions with 93.3% recognition rate on this dataset. With these results, our system can compete with the state-of-art methods surveyed in [8], where the recognition rates of these methods change between 84% and 97%.

In the second set of experiments, we test the system's recognition accuracy of the exercise types in the HPTE dataset. We focus on the individual contributions of three exercise components (motion patterns, stance knowledge, and the exercise object information) by using the depth (RGBD) data. Details about features and the related feature extraction methods are given in Section VI. We run the system with various combinations of components and the recognition results are listed in Table II in sorted order with respect to accuracy. This experiment provides valuable data to evaluate different sources of information. We adopt the method of [40] to extract motion based features. In this method, the statistical information about the presence of motion, the location of motion, and the recency of motion is used to define motion based features. We also adopt the method of Cheema *et al.* [41] to extract stance based features. Their method is based on scale invariant contour features which are obtained from silhouette images. We modify their result (representation of stance information) for a given v by using the statistical mean and the variance of each pose descriptor as in (11). The proposed system obtains 89.6% accuracy rate with only motion patterns, 69.58% accuracy rate with only stance knowledge (proposed method), and 72.92% accuracy rate with only adopted stance features [41]. These results demonstrate that the motion patterns are more distinctive than stance knowledge. Moreover, the stance features obtained by [41] is more effective than our baseline stance features. We also observe that motion patterns with stance knowledge provides better recognition rates than motion patterns with object usage information. This shows that stance knowledge provides more valuable information than object based features. Finally, we obtain the best recognition rates with the combination of three main source of information in our generative Bayesian network. Note that the best accuracy of 98.33% is reached two times with different stance methods. We argue that stance information has only limited contribution towards the final results and hence our proposed stance based features can perform as good as other more successful methods such as [41]. The same argument can be made for the object features. Note also that we use both color and depth videos in this set of experiments.

TABLE II
Recognition accuracy of exercises on the HPTE dataset with respect to component types.

Component Type	Rec. Accuracy
Motion + Stance [41]+ Object [38]	98.33%
Motion + Stance + Object [38]	98.33%
Motion + Stance [41] + Object [36]	97.9%
Motion + Stance [41] + Object [37]	97.9%
Motion + Stance + Object [37]	97.9%
Motion + Stance + Object [36]	97.5%
Motion + Stance [41]	95.4%
Motion + Stance	95.0%
Motion + Object [38]	93.33%
Motion + Object [37]	93.33%
Motion + Object [36]	92.9%
Motion	89.6%
Stance [41]+ Object [38]	78.75%
Stance [41]+ Object [37]	77.92%
Stance + Object [38]	76.67%
Stance + Object [37]	75.0%
Stance [41] + Object [36]	74.17%
Stance [41]	72.92%
Stance + Object [36]	72.08%
Stance	69.58%

In the third set of experiments, we evaluate the impact of the depth information. As mentioned in Section III, HPTE dataset contains 240 different RGB and depth videos to demonstrate exercise sessions. We select the best component types from the second set of experiments as proposed motion patterns, stance knowledge obtained by [41], and object usage information obtained by [38]. When the motion, stance, and object information are represented by using only RGB videos, our the system recognizes 220 of the 240 videos successfully with a 91.7% accuracy (without using any depth data). Details about system performance without any depth information is given in Table III as a confusion matrix. We then repeat the same experiment by including the depth information. In this setting our system successfully recognizes 98.33% of the exercise type in 240 sessions. Details of this experiment are given in TableIV as a confusion matrix. While the other two sources of information remaining the same, we also run the baseline method of [40] to extract motion information and

replace our proposed motion patterns with these features. The system obtains recognition rates of 82.50% and 83.33% (in terms of accuracy) on RGB and RGBD data, respectively. Table V summarizes the recognition accuracy results for the third experiment of the third set of experiments. The first experiment of this set shows that our system can be employed in practice without using depth cameras with performance better than 90%. The second experiment of this set shows that, although the proposed system can work just with the RGB videos but using color and depth videos produces much better recognition rates. The general misclassification error is between the straight and circular pendulum exercise types. These errors are reduced by using depth information, which is expected because in the RGB data, the straight pendulum exercise do not show significant motion information due to motion towards the camera. The third experiment of this set demonstrates the superiority of proposed motion patterns to the baseline method of [40].

TABLE III
CONFUSION MATRIX FOR EXERCISE RECOGNITION RESULTS ON HPTE DATASET WITHOUT USING DEPTH VIDEOS.

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

TABLE IV
CONFUSION MATRIX FOR EXERCISE RECOGNITION RESULTS ON THE HPTE DATASET USING BOTH RGB AND DEPTH VIDEOS.

	Stick	Dia-stick	Lie back	Towel	Str-pen	Cir-pen	Chair	Heel
Stick	30	0	0	0	0	0	0	0
Dia-stick	0	30	0	0	0	0	0	0
Lie back	0	0	30	0	0	0	0	0
Towel	1	0	0	29	0	0	0	0
Str-pen	0	0	0	0	29	1	0	0
Cir-pen	0	0	0	0	1	29	0	0
Chair	0	0	0	0	0	0	30	0
Heel	0	0	0	0	0	0	1	29

TABLE V
RECOGNITION ACCURACY OF EXERCISES ON THE HPTE DATASET WITH RESPECT TO DIFFERENT MOTION INFORMATION.

Data Type	Rec. Acc.
Motion + Stance [41] + Object [38] on RGB Data	91.7%
Motion + Stance [41] + Object [38] on RGBD Data	98.33%
Motion [40] + Stance [41] + Object [38] on RGB Data	82.5%
Motion [40] + Stance [41] + Object [38] on RGBD Data	83.33%

The fourth set of experiments investigates the impact of parameter selection in the extraction of motion features. These parameters define the size of filters in spatial and temporal domains. Table VI lists the recognition accuracy of exercise types in the HPTE dataset using only motion features with both of the RGB and depth videos. As we discussed previously in Section VI-A, while the spatial parameter focuses on the area of motion, the temporal parameter covers the recency of motion. We obtain the best results with the selection as 8x8 pixel and 4x4 pixel sizes in the spatial domain, 4 frames and 8 frames in the temporal domain. We use this selection as the motion parameters on all the HPTE experiments.

TABLE VI
RECOGNITION ACCURACY USING ONLY MOTION INFORMATION WITH DIFFERENT SIZED FILTERS.

Spatial domain	Temporal domain	Acc.
2x2 and 4x4 pixels	2 and 4 frames	67.1%
2x2 and 4x4 pixels	4 and 8 frames	72.5%
2x2 and 4x4 pixels	8 and 16 frames	77.1%
4x4 and 8x8 pixels	2 and 4 frames	77.9%
4x4 and 8x8 pixels	4 and 8 frames	89.6%
4x4 and 8x8 pixels	8 and 16 frames	85.8%
8x8 and 16x16 pixels	2 and 4 frames	86.7%
8x8 and 16x16 pixels	4 and 8 frames	88.3%
8x8 and 16x16 pixels	8 and 16 frames	83.8%

TABLE VII
MEAN REPETITION COUNT AND THE AVERAGE ESTIMATION ERROR FOR THE REPETITION COUNT ESTIMATION EXPERIMENT.

Exercise Type	Mean Repetition Count	Average Estimation Error
Stick	5.3	0.14
Dia-stick	4.4	0.24
Lie back	5.2	0.12
Towel	4.6	1.28
Str-pen	5.4	0.43
Cir-pen	4.6	1.46
Chair	4.8	0.72
Heel	6.5	0.60
Average	**5.1**	**0.62**

The final set of experiments analyze the performance of the estimation of the repetition count. In the estimation of repetition count, we use the proposed motion patterns as motion information and adopted method of [41] to form stance knowledge. Since the estimation of the repetitions depends on the recognition accuracy of the exercise label, we examine the estimation performance in two different scenarios. In the first scenario, we manually label the exercise type in each exercise session and then the system estimated the repetition count of the 220 exercise sessions correctly with 91.67% accuracy rate. In the second scenario, we use the output of recognition module as exercise labels and then the system estimated the repetition count of the 210 exercise sessions correctly with 87.5% accuracy rate. In each scenario, most of the estimation errors are observed at the sessions of towel and circular pendulum exercise. Detailed results about the estimation of the repetition count in each exercise type for the second scenario are given in Table VII. The overall average estimation error is 0.62, which is very promising because estimating the repetition counts with less than 1 repetition count error is within the tolerable limits.

IX. CONCLUSIONS

We presented a novel physiotherapy exercise recognition and repetition count estimation system that extents our previous work [1] by new experiments and further analysis. The

main contribution of our system is to model the exercise recognition problem using the base knowledge from the physiotherapy literature, which asserts that a physiotherapy exercise consists of three main components: the motion patterns, the patient stance, and the exercise object. In order to bring these three components together, we employed a generative Bayesian network in our system, which allowed us to develop separate subsystems for each exercise component and handle dependencies between these components transparently. The proposed system also provides a novel module for the estimation of exercise repetition counts which is an important information for the home monitoring systems. We formed a new physiotherapy exercise dataset (HTPE dataset) and made it publicly available to be used by other researchers for comparison.

The experiments performed on HPTE dataset revealed that each component has positive contributions towards very good recognition rates. Our motion pattern analysis module includes many novel contributions and experiments on the Weizmann dataset showed that it is in line with the state-of-art action recognition methods. In the estimation of exercise repetition counts, the proposed system performed favorably in the experiments on physiotherapy data. Our system can be easily extended to use other types of graphical models such as conditional random fields or other physiotherapy exercise components from the physiotherapy literature, such as pressure sensor data or accelerometer data. Moreover, to the best of our knowledge the proposed system is the first system that recognize the physiotherapy exercises which includes patient and object interaction by using a consumer level RGBD camera.

Since our system is designed towards building a physiotherapy monitoring system which recognizes and counts the physical exercises in sessions, it does not produce any instructions to direct the patient to correct any mistakes. However, our system can be extended to perform such tasks by adding a system module for the human computer interactions tasks and forming a database which contains negative samples (subpar, sketchy, wrong, etc. exercises). Finally, our system is currently for offline use, which is not a problem for monitoring home-based exercise sessions of patients. The feature extraction step of the main system takes the most execution time. We emphasize that parallel algorithms can be used to optimize the execution times of the feature extraction processes.

REFERENCES

[1] I. Ar and Y. S. Akgul, "A monitoring system for home-based physiotherapy exercises," in *Computer and Information Sciences III*, 2013, pp. 487–494.

[2] A. Koller-Hodac, D. Leonardo, S. Walpen, and D. Felder, "A novel robotic device for knee rehabilitation improved physical therapy through automated process," in *IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics*, Tokyo, Japan, 2010, pp. 820–824.

[3] P. M. van Vliet and G. Wulf, "Extrinsic feedback for motor learning after stroke: What is the evidence?" *Disability and Rehabilitation*, vol. 28, no. 13-14, pp. 831–840, 2006.

[4] A. Domingo and D. P. Ferris, "Effects of physical guidance on shortterm learning of walking on a narrow beam," *Gait Posture*, vol. 30, no. 4, pp. 464–468, Nov. 2009.

[5] D. E. Rumelhart, G. E. Hintont, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[6] D. M. Brennan, P. S. Lum, G. Uswatte, E. Taub, B. M. Gilmore, and J. Barman, "A telerehabilitation platform for home-based automated therapy of arm function," in *33rd Annual International Conference of the IEEE EMBS*, Boston, Massachusetts, USA, 2011, pp. 1819–1822.

[7] B. T. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, Nov. 2006.

[8] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[9] H. Zhou and H. Hu, "Human motion tracking for rehabilitation - a survey," *Biomedical Signal Processing and Control*, vol. 3, no. 1, pp. 1–18, 2008.

[10] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human detection using depth information by kinect," in *Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR (HAU3D)*, Colorado Springs, USA, 2011.

[11] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. of the Sixth International Conference on Computer Vision*, Washington, DC, USA, 1998, pp. 555–562.

[12] K. Praveen, R. Parvathi, and P. Venkata, *Fundamentals of Physiotherapy*. Jaypee, 2005.

[13] S. Soutschek, J. Kornhuber, A. Maier, S. Bauer, P. Kugler, J. Hornegger, M. Bebenek, S. Steckmann, S. von Stengel, and W. Kemmler, "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*, 2010, pp. 1–4.

[14] D. Fitzgerald, J. Foody, D. Kelly, T. Ward, C. Markham, J. McDonald, and B. Caulfield, "Development of a wearable motion capture suit and virtual reality biofeedback system for the instruction and analysis of sports rehabilitation exercises," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, Aug. 2007, pp. 4870–4874.

[15] S. B. Kesner, L. Jentoft, F. L. Hammond, R. D. Howe, and M. Popovic, "Design considerations for an active soft orthotic system for shoulder rehabilitation," in *33rd Annual International Conference of the IEEE EMBS*, Boston, USA, 2011.

[16] S. H. Roy, M. S. Cheng, S. S. Chang, J. Moore, G. D. Luca, S. H. Nawab, and C. J. D. Luca, "A combined semg and accelerometer system for monitoring functional activity in stroke," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, pp. 585–594, 2009.

[17] J. Courtney and A. M. de Paor, "A monocular marker-free gait measurement system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, pp. 453–460, 2010.

[18] Y. Jung, D. Kang, and J. Kim, "Upper body motion tracking with inertial sensors," in *Proc. of the 2010 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Tianjin, China, 2010.

[19] A. R. Guraliuc, P. Barsocchi, F. Potorti, and P. Nepa, "Limb movements classification using wearable wireless transceivers," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 3, pp. 474–480, 2011.

[20] A. A. A. Timmermans, H. A. M. Seelen, R. P. J. Geers, P. K. Saini, S. Winter, J. te Vrugt, and H. Kingma, "Sensor-based arm skill training in chronic stroke patients: Results on treatment outcome, patient motivation, and system usability," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 3, pp. 284–292, 2010.

[21] M. Li, V. Rozgic, G. Thatte, L. Sangwon, B. A. Emken, M. Annavaram, U. Mitra, D. Spruijt-Metz, and S. Narayanan, "Multimodal physical activity recognition by fusing temporal and cepstral information," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, pp. 369–380, 2010.

[22] M. Duff, Y. Chen, S. Attygalle, J. Herman, H. Sundaram, G. Qian, J. He, and T. Rikakis, "An adaptive mixed reality training system for stroke rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 5, pp. 531–541, 2010.

[23] Y. J. Chang, S. F. Chen, and J. D. Huang, "A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," *Research in Developmental Disabilities*, vol. 32, no. 6, pp. 2566–2570, Nov. 2011.

[24] B. Lange, A. Rizzo, C. Y. Chang, E. Suma, and M. Bolas, "Markerless full body tracking: Depth-sensing technology within virtual environments," in *Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, Orlando, Florida, USA, 2011.

[25] B. Ni, Y. Pei, S. Winkler, and P. Moulin, "Kinect for rehabilitation," in *Proc. of International Convention on Rehabilitation Engineering and Assistive Technology (i-CREATe)*, Singapore, 2012.
[26] L. Wang, W. M. Hu, and T. N. Tan, "Recent developments in human motion analysis," *Pattern Recognition*, vol. 36, no. 3, pp. 585–601, 2003.
[27] P. K. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.
[28] Microsoft. (2013) Kinect for windows sdk version 1.5. [Online]. Available: http://www.microsoft.com/en-us/kinectforwindows/develop/developer-downloads.aspx
[29] X. Cui, Y. Liu, S. Shan, X. Chen, and W. Gao, "3d haar-like features for pedestrian detection," in *The 2007 IEEE International Conference on Multimedia and Expo (ICME)*, 2007, pp. 1263–1266.
[30] X. Zhu, S. Gong, and C. Loy, "Comparing visual feature coding for learning disjoint camera dependencies," in *Proc. of the 23nd British Machine Vision Conference*, 2012, pp. 1–12.
[31] C. Vens and F. Costa, "Random forest based feature induction," in *IEEE International Conference on Data Mining*, 2011, pp. 744–753.
[32] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010, pp. 1324–1332.
[33] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[34] A. Yao, J. Gall, G. Fanelli, and L. V. Gool, "Does human action recognition benefit from pose estimation ?" in *Proc. of the 23nd British Machine Vision Conference*, 2011, pp. 1–11.
[35] R. H. Parry, "The interactional management of patients' physical incompetence: A conversation analytic study of physiotherapy interactions," *Sociology of Health and Illness*, vol. 26, no. 7, pp. 976–1007, 2004.
[36] L. Fei-Fei, "Bag of words models: Recognizing and learning object categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
[37] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.
[38] J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1022–1029.
[39] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. of the 10th IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 1395–1402.
[40] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 257–267, 2001.
[41] S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage, "Action recognition by learning discriminative key poses," in *Int. Conf. on Computer Vision Workshops*, 2011, pp. 1302–1309.

Ilktan Ar received the BS degree in computer engineering from Kadir Has University, Istanbul, Turkey in 2004. He received the MS degree in computer engineering from Yildiz Techinal University, Istanbul, Turkey in 2007. He is currently working as a research asistant in the Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Kadir Has University, Turkey. He is also continuing to work towards his Ph.D degree in the Department of Computer Engineering at Gebze Institute of Technology, Turkey. His research interests include application of computer vision in human motion analysis, video analysis, pattern recognition, and industrial inspection.

Yusuf Sinan Akgul received the BS degree in computer engineering from Middle East Technical University, Ankara, Turkey in 1992. He received the MS and PhD degrees in computer science from University of Delaware in 1995 and 2000, respectively. He worked for Cognex Corporation, Natick, MA between 2000 and 2005 as a senior vision engineer. He is currently with the Department of Computer Engineering, Gebze Institute of Technology, Turkey. His research interests include application of computer vision in medical image analysis, video analysis, object recognition, 3D analysis, and industrial inspection.