Journal of Visual Communication and Image Representation (2020)

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Recognition of occupational therapy exercises and detection of compensation mistakes for Cerebral Palsy

Mehmet Faruk Ongun[a], Uğur Güdükbay[a,*], Selim Aksoy[a]

[a]*Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey*

ARTICLE INFO

Article history:
Received
Received in final form
Accepted
Available online

Communicated by

2000 MSC: 65D18, 65D19, 68T45

Keywords: gesture recognition, cerebral palsy, occupational therapy, compensation mistake, hidden Markov model, depth camera, virtual rehabilitation.

ABSTRACT

Depth camera-based virtual rehabilitation systems are gaining attention in occupational therapy for cerebral palsy patients. When developing such a system, domain-specific exercise recognition is vital. To design such a gesture recognition method, some obstacles need to be overcome: detection of gestures not related to the defined exercise set and recognition of incorrect exercises performed by the patients to compensate for their lack of ability. We propose a framework based on hidden Markov models for the recognition of upper extremity functional exercises. We determine critical compensation mistakes together with restrictions for classifying these mistakes with the help of occupational therapists. We first eliminate undefined gestures by evaluating two models that produce adaptive threshold values. Then we utilize specific negative models based on feature thresholding and train them for each exercise to detect compensation mistakes. We perform various tests using our method in a laboratory environment under the supervision of occupational therapists.

1. Introduction

Cerebral Palsy (CP) is a neurological disorder caused by a non-progressive brain injury or malformation that occurs while the child's brain is under development. CP affects body movement, muscle control, muscle coordination, muscle tone, reflex, posture, balance and cognitive skills. In most cases, it impacts fine motor skills, gross motor skills, and sensory skills. The effects of Cerebral Palsy are long-term, not temporary. The injury and damage to the brain are permanent. The brain does not heal as the way other parts of the body might. On the other hand, associative conditions may improve over time. Rehabilitation, which includes physical and/or occupational therapy, is among the main intervention methods to promote, maintain and restore the physical well-being of CP patients [1].

There are various approaches to CP rehabilitation. Virtual Reality (VR) based rehabilitation generally aims children both because it is effective to perform these exercises at an early age and because they are often put into practice as serious games to make them more attractive and less boring for children. The emergence of depth cameras to be used in schools and homes made it possible to capture the movements of the patients and promoted the use of these types of cameras in virtual rehabilitation [2]. However, for CP patients, exercising games targeting the general population proved problematic in some aspects. First of all, these patients cannot sometimes perform some moves properly and complete the game. They may also have cognitive disabilities that cause them to perform unrelated/undefined moves and sometimes require therapists to step in, which causes the game engine to try and recognize these movements that are out of context. Thus, playing/performing regular exercising games is not practical in this case. Another problem is the compensation mistakes made by the patients during exercises. When the patients have insufficient muscle

*Corresponding author: Tel.: +90-312-290-1386; fax: +90-312-266-4047
e-mail: mehmet.ongun@bilkent.edu.tr (Mehmet Faruk Ongun),
gudukbay@cs.bilkent.edu.tr (Uğur Güdükbay),
saksoy@cs.bilkent.edu.tr (Selim Aksoy)

strength or muscle control, they try to complete the movement by using some other muscles and/or joints, e.g., twisting, bending their elbows during a shoulder exercise. These incorrect exercises are not desired by therapists.

We provide solutions for the described problems to develop a gesture recognition system designed specifically for cerebral palsy patients. First, it should be able to distinguish the movements that are not related to the content of the application from the defined exercises. Second, it should be able to detect and capture the compensation mistakes that are done by the patients. Hence, the scope of this work is to provide a gesture recognition solution to be used in virtual rehabilitation applications for children with cerebral palsy. The contributions are as follows.

- We propose two alternative methods, called *universal negative model* and *universal positive model*, which enable the detection of non-gesture patterns by producing an adaptive threshold value.
- We examine the problem of detecting small mistakes made by patients to compensate for their lack of ability, control, and strength, and devise a new approach to enhance the gesture recognition accuracy in such cases.

The rest of the paper is organized as follows. Section 2 presents background and related work on gesture recognition and occupational therapy. Section 3 describes the proposed framework. Section 4 focuses on detecting non-gesture patterns. Section 5 is about detection of compensation mistakes. Section 6 presents the experimental setup and results. Finally, Section 7 concludes and provides possible future research directions.

2. Background and Related Work

Visual recognition tasks such as object and scene recognition [3], text recognition [4], and video recognition [5] have received significant attention in the computer vision and machine learning literature. In this section, we eloborate on the related work on gesture recognition and occupational therapy exercises, specifically for CP.

2.1. Gesture Recognition

Gesture recognition is concerned with capturing and identifying the motions of human body parts. Gestures may include hand, arm, head and/or body motions. The applications of gesture recognition include medical rehabilitation (e.g., physiotherapy, occupational therapy) [6, 7, 8, 9], human activity recognition [10, 11, 12, 13], sign language recognition [14], virtual reality, forensic identification, and lie detection [15, 16]. Support Vector Machines (SVMs) [17], Dynamic Time Warping (DTW) [18, 19, 20], Adaboost, and Hidden Markov Models (HMMs) [21] are the most common approaches for gesture recognition applications [22, 23].

Regarding SVMs, it should be pointed out that, regular SVMs are capable of classifying a total of two classes, causing the researchers to use multiclass SVMs when more than two gestures are present in the vocabulary. The approaches that use SVMs perform gesture recognition by classifying single frame gestures or poses, not temporal data [24]. Biswas et al. [25] use multiclass SVMs together with Kinect where the gestures are classified using only single frames and histograms of depth values in that frame. A similar research [26] also makes use of single frames when recognizing gestures by SVM.

DTW is another method that gives successful results in gesture recognition. The approach of Sempena et al. [27] is one example of DTW used with depth data for gesture recognition. They achieve a high success rate. However, they mainly used the method for recognizing repetitive and simple human activities like running and waving. Hu et al. [28] use the data from the Kinect sensor for real-time human movement retrieval and assessment. They use nonlinear time warping to retrieve video segments similar to the query performed by the user so that the user learns according to the selected video samples for acting correctly.

Some notable studies focus on hand gesture recognition using depth cameras. Yao and Fuo [29] propose a contour-based approach using the Kinect sensor mainly for human-computer interaction. Zhang et al. [30] utilize RGB and depth data from the Kinect sensor for one-shot learning gesture recognition using a Bag-of-Manifold-Words approach. Wang et al. [31] propose a superpixel-based hand gesture recognition system with the Kinect sensor that uses a superpixel earth mover's distance metric to measure the dissimilarity between hand gestures. Ren et al. [32] propose a part-based hand gesture recognition system using the Kinect sensor and a distance metric, Finger-Earth Mover's Distance (FEMD), to measure the dissimilarity between hand shapes.

We choose HMMs as the gesture recognition model in our study. Before moving onto examining the method in detail, the reasoning behind this choice needs to be explained. Comparing studies that use different approaches does not give accurate information because of the differences in the datasets used. Because of this, research that compares different algorithms using the same set of gestures is inspected thoroughly. There are effective HMM-based gesture recognition solutions for Kinect time series joint data [24, 33, 34, 35]. HMMs are also extensively studied in sign language recognition, which is similar to exercise recognition in principle. Suarez and Murphy [36] emphasize the high classification rate and prevalence of HMM-based solutions in gesture recognition. Comparing the HMM solutions with alternatives in sign language recognition also pictures HMM as a successful approach.

Recurrent Neural Networks (RNN), in particular Long Short-Term Memory Networks (LSTM), have also been recently popular in gesture recognition tasks. However, they require a large amount of training data, which is not available for the gesture recognition tasks and the detection of compensation mistakes for the CP study described in this paper. HMMs have been shown to perform better than LSTMs in settings with small sample sizes [37].

There are various options when it comes to designing the model for the HMM approach [38]. The most popular types of HMMs are as follows:

- *Ergodic model*: A model in which it is possible to reach

any state from any other state.

- *Left-to-right model*: A model in which a state can only be reached from the preceding states. These types of models inherently impose a temporal order and thus widely used in speech and gesture recognition.

- *Parallel left-to-right model*: Similar to the Left-to-Right model, except that it has several paths through the states.

Even though there are many pieces of research on human activity and exercise recognition with HMM, to the best of our knowledge, there is no research that concentrates on the recognition of erroneous exercises practiced by the patients. Lu et al. [39] propose an HMM-based method using Kinect RGB-D camera. They extract the joint information using the depth data provided by the camera and then generate histogram data of joint locations (with the spherical coordinate system). They tested their approach with their dataset and compared their performance with those of other approaches using the MSR Action Three-Dimensional (3D) dataset provided by Microsoft.

Yang et al. [33] focus on hand gesture recognition primarily, and thus involves the segmentation of hand from RGB data before HMM. They do not use depth cameras. The features selected for recognition are hand position, velocity, size, and shape. Another problem they have dealt with is the data aligning problem. It is mainly the time-variance problem and the method they utilized is a simple aligning algorithm. It is asserted that with the use of various features together in recognition, they managed to increase the recognition performance.

Uddin et al. [34] propose an HMM-based approach that uses histogram data. They use both silhouette and joint data as features for different setups and compare them. It is pointed out that their approach gives better results using joint-based features. The type of their HMM is the Left-to-Right Model because of its temporal nature. The requirements in these studies are similar to ours and their methodology provides a good baseline approach for dealing with gesture recognition.

Granger et al. [40] compare the performances of temporal models for gesture recognition, namely Hybrid Neural Network-Hidden Markov Models (NN-HMM) and Recurrent Neural Networks. They conclude that Hybrid NN-HMM models produce better results than RNNs but training hybrid models is difficult than training end-to-end neural networks.

2.2. Occupational Therapy

We focus on the recognition of occupational therapy exercises. In simple terms, occupational therapy is a sub-branch of physiotherapy, that focuses on the daily activities of the patients. Even though occupational therapy practitioners use similar exercises for the rehabilitation of the patients, in terms of context and the evaluation of these exercises, it has different characteristics.

According to the practice framework of The American Occupational Therapy Association, occupational therapy aims to enhance the daily lives of individuals and groups in homes, schools, workplaces, and so on by utilizing the everyday life activities in the therapy. Occupational therapists provide development in body functions, body structures, motor skills, processing, and social interaction skills by putting their knowledge of the transactional relationship among the person and occupations the person is engaged and by creating an occupation-based intervention plan with the aim of successful social participation. Because occupational therapists aim the result of participation, when needed, they manage and modify the environment and objects within to increase the engagement. Habilitation, rehabilitation, the promotion of health and wellness for people with needs related or unrelated to their disability are the objectives of occupational therapy [41].

Rehabilitation exercises are one of the tools that are being used by occupational therapy practitioners in various cases. The target audience for occupational therapy includes all age groups from children to seniors and many different types of disorders, namely cerebral palsy, stroke, and Parkinson's among many others [42].

Taking advantage of the latest technologies is not uncommon in occupational therapy. Especially in recent years, using virtual therapy and/or augmented reality technologies in therapy sessions gained attraction. Wentao et al. [43] use robotic therapy practices for cerebral palsy patients. To train the virtual therapist (robot), they use HMM as a pattern recognition method. With the emergence of commercially available RGB-D cameras, gesture recognition based on data obtained from RGB-D cameras became one of the main focus areas [2].

Chang et al. [44] propose a Kinect-based upper limb rehabilitation system for CP patients. Based on their experimental work, they argue that their system managed to increase the motivation of the test subjects and provided an improvement in the success rate of the exercises. In another publication, Chang et al. [45] test a similar system on young adults and get similar successful results with regards to the rehabilitation of patients. Pedraza et al. [46] describe a Kinect-based virtual reality system and claims to improve patient mobility, aerobic capacity, strength, coordination, and flexibility.

3. The Proposed Framework

We describe our baseline method for gesture/exercise recognition. The focus of the design will be the unique characteristics of the problem at hand and the resulting solution is intended to have the ability to recognize and differentiate different upper-body exercises. Various problems regarding feature selection, model structure, scaling problem, and continuous observation symbols are addressed in detail and the solutions are elaborated in-depth. It is specifically emphasized in [47] that when the recognition problem has untraditional aspects and the devised system does not provide tailored solutions to these, experiments can present diminished results in terms of the recognition accuracy. Fig. 1 depicts the framework of the proposed solution.

Our framework first takes the frames from the depth camera as input and processes the skeleton data provided by the Software Development Kit (SDK) of the depth camera. When the starting position is detected, we record the data at each frame as the gesture data until we detect the end position. Afterward,

we process the gesture data using the universal negative model to see if it is a non-gesture (see Section 4.1). If the gesture is one of the gestures defined in the exercise set, we apply a two-stage process, involving *feature thresholding* (see Section 5.1) and applying *negative models* (see Section 5.2), to improve the accuracy of determining correct and incorrect gestures; i.e, detect compensation mistakes. Finally, if no mistake is detected, we classify the gesture as a correct gesture.

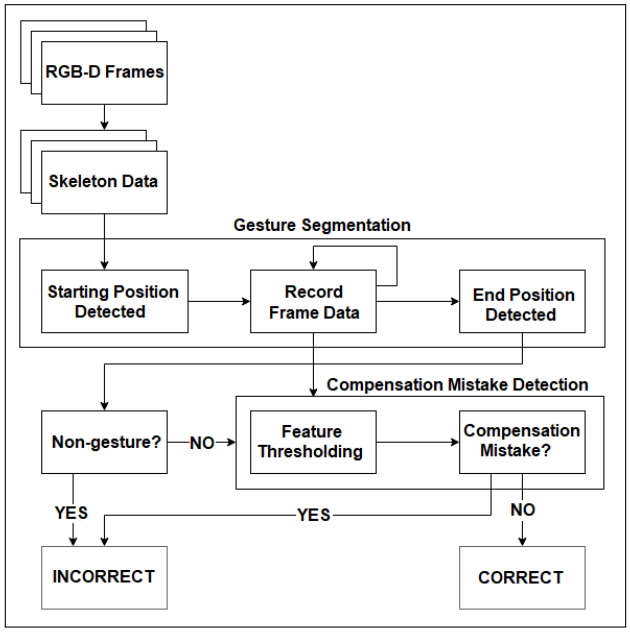

Fig. 1. The framework of the proposed solution.

3.1. Experimental Setup for Cerebral Palsy

CP could be classified according to the way motor skills are affected, which is an indication of the damaged part of the brain: *spastic*, *dyskinetic* or *ataxic*. Another type of classification is dictated by the location of impairment. *Quadriplegia* is when both arms and legs are affected, *diplegia* defines the patients with impairment at both legs, and *hemiplegia* is when one arm and one leg on the same body side is affected, as a result of brain damage that affects one hemisphere [48, 49].

Gross Motor Function Classification System (GMFCS) [50] and Manual Ability Classification System (MACS) [51] are two classification standards that are used to differentiate patients according to the severity of impairment. We focused on children that need occupational therapy exercises to improve their ability to complete their daily activities and be able to perform the exercises correctly. Another requirement was the ability to stand because otherwise, the tracking accuracy of the depth camera reduces dramatically. As a result, children that are classified in levels 1 or 2 of GMFCS and levels 2 or 3 of MACS are targeted during our study.

The depth camera performs well when capturing the upper extremities. The patients often need to sit, lay down or hold onto something when performing the lower extremity exercises, which reduces the camera's accuracy. Because of these reasons, the chosen exercises for this study are upper extremity functional exercises including the movement of arms, specifically shoulder and elbow joints. These exercise groups are considered important by the occupational therapist because the better use of upper extremities affects the daily lives of the patient greatly. We selected five main gestures from the group of upper extremity functional exercises: shoulder flexion (180°), shoulder abduction (90°), shoulder external rotation, elbow flexion and extension, and combined Proprioceptive Neuromuscular Facilitation (PNF) pattern of all other four movements.

3.2. Feature Selection

Feature selection for the recognition task is one of the most significant steps in gesture recognition. Some studies used various sets of features within the same system and the results are drastically different from each other [34]. The first thing to determine for feature selection is the data source for our approach. Because Kinect RGB-D camera is used as hardware, we have two different types of data: *silhouette* and *skeleton (joint)* data.

Silhouette data is widely used in activity recognition systems. However, compared to joint data, the use of HMM with silhouette data is relatively uncommon. In [52], activity data is classified using the nearest neighbor matching. The recognition features are body shape and gait position during walking activity. Zhang et al. [53] utilize a Bag-of-3D points approach for recognition. This is another example of non-HMM gesture recognition. Bobick et al. [54] propose a successful recognition approach that uses modified silhouette data with a non-HMM method. As stated before, the research in [34] uses silhouette with HMM but the skeleton data shows up to 84% performance gain. Hence, it can be concluded that the silhouette data is not suitable for use with HMMs whereas other approaches give better results. A further inspection makes it evident that silhouette data are generally used for daily activity recognition instead of gestures like rehabilitation exercises. These findings suggest that the use of joint data is more suitable for our problem.

MS Kinect provides joint data in 3D space. Xia et al. [39] examine various approaches on gesture recognition utilizing joint data. It is possible to utilize joint locations, joint motions and/or joint angles as features. Campbell et al. [55] examine the advantage and disadvantages of each type of joint data. They focused on the features' shift-invariance and rotation-invariance properties. It is argued that when joint locations are used, the approach becomes vulnerable to expected coordination shifts in 3D space and the rotations of the subjects. When joint angles are used, the system becomes shift/coordinate-invariant, however, it is still affected by rotations in space. Thus, they propose to use joint motion (i.e., derivative of location or angle) as a shift-invariant and rotation-invariant feature set. However, one disadvantage of using derivatives is that it depicts the same gestures performed at different speeds as different gestures.

Coordinate shifts are important in our problem because the position of the subject relative to the camera is not always the same. The body scales of patients are different. This makes it appropriate to choose a shift-invariant feature set. However, rotation-invariance is not needed in this case because the subject directly faces the camera or stands perpendicular to it depending on the gesture during training and recognition phases. Because of these reasons, using joint angles has no disadvantages with regards to shift and rotational variance in our case

and it provides the Viterbi algorithm with time-warping behavior [55].

Another requirement for a joint angle feature is that it is important for the evaluation of occupational therapy exercises because the correctness of each gesture is generally decided by the joint angles [42]. As a result, we chose different joint angles or their 2D projections for each different gesture in our dataset. For each gesture, the requirements and standards for the gesture are taken into account and the joints that should be tracked for the gesture to be classified accordingly are determined by the occupational therapy researchers.

We choose a common set of features for all upper extremity gestures based on the expertise of an occupational therapist and the success of the depth camera on determining various joint angles. However, it is necessary to determine or limit the number of features (i.e., joint angles). To determine the number of features, a series of tests are conducted. In this stage, gestures that are not defined in our exercise set or incorrect exercises are not taken into consideration. We conduct the tests as multi-class classification problems where we try to distinguish each of the *five* exercises from each other. Three different subjects perform each exercise a total of 20 times.

According to the results in Fig. 2, among all joint angles, using four to six features gives F1 scores above 0.91, with four features reaching the highest accuracy of a 0.96 F1 score. By using this range as a guideline for the number of features (i.e., joint angles), and by using the recommendations of the occupational therapists regarding potential compensation mistakes, we finalize the sets of features for each exercise as follows.

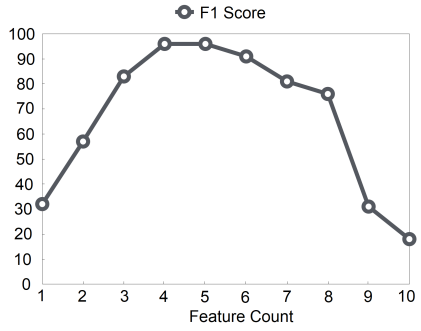

Fig. 2. The graph depicting the F1 score for different feature counts.

- *Shoulder flexion*: the shoulder angle on the xz-plane, the shoulder angle on the yz-plane, the elbow angle on the xz-plane, the elbow angle on the yz-plane, and the body angle on yz-plane (cf. Fig. 3).

- *Shoulder abduction*: the shoulder angle on the xy-plane, the shoulder angle on the xz-plane, the elbow angle on the xy-plane, the elbow angle on the xz-plane, the head angle on the xy-plane, and the body angle on the xy-plane (cf. Fig. 4).

- *External rotation*: the shoulder angle on the xy-plane, the shoulder angle on the xz-plane, the elbow angle on the xy-plane, the elbow angle on the xz-plane (cf. Fig. 5).

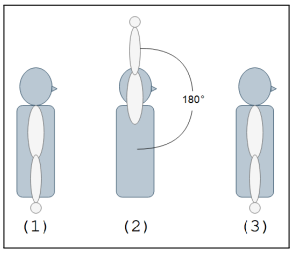

Fig. 3. Shoulder flexion: starting position (1), the arm is raised 180° from the front while keeping the elbow angle as 180° (2), and the end position(3).

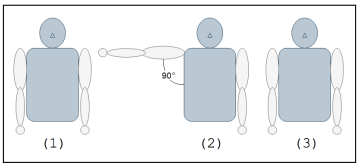

Fig. 4. Shoulder abduction: starting position (1), the arm is raised 90° from the side while keeping the elbow angle as 180° (2), and the end position (3).

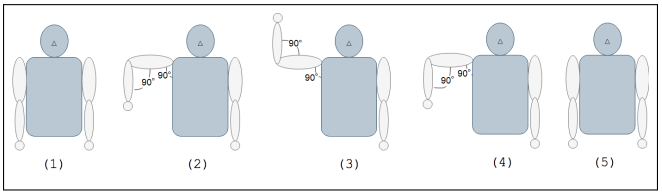

Fig. 5. External rotation: starting position (1), the arm is raised from the side while keeping the elbow and shoulder angles as 90° (2), the arm is rotated using only the shoulder joint (3), the arm is rotated back (4), and end position (5).

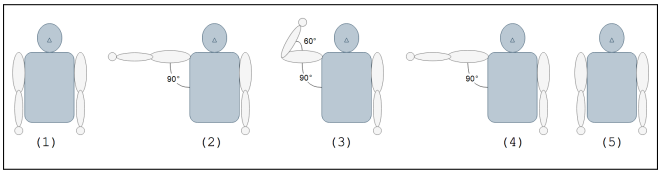

Fig. 6. Elbow flexion and extension: starting position (1), the arm is raised to the side while keeping the elbow straight and the shoulder angle as 90° (2), the elbow is flexed to 60° (3), the elbow is extended back (4), and the end position (5).

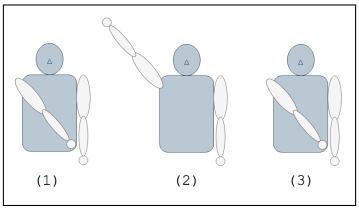

Fig. 7. Combined PNF pattern: starting position (1), the arm is raised diagonally while keeping the diagonal movement line straight (2), and back to the end position (3).

- *Elbow flexion-extension*: the shoulder angle on the xy-plane, the shoulder angle on the xz-plane, the elbow angle on the xy-plane, the elbow angle on the xz-plane, the head angle on the xy-plane, and the body angle on the xy-plane (cf. Fig. 6).

- *Combined PNF pattern*: the shoulder angle on the xy-

plane, the shoulder angle on the xz-plane, the elbow angle on the xy-plane, the elbow angle on the xz-plane, the head angle on the xy-plane, and the body angle on the yz-plane (cf. Fig. 7).

3.3. Dataset

We propose an exercise recognition system that specifically targets the requirements of occupational therapists. In our setup, when a cerebral palsy patient is required to perform a specific exercise, the system should be able to

- eliminate unrelated gestures performed by the subject during the exercise session and do not count them as correct or incorrect, and
- recognize the compensation mistakes of that specific gesture and count them as incorrect exercises.

The goal is not to distinguish exercises from each other, but to distinguish the compensation mistakes for the specified exercise. Hence, each time the patients are directed to perform one predetermined exercise and during the session all the gestures performed are either: non-gestures (gestures that are defined in our exercise set but not the predetermined exercise are also classified as non-gestures), the correct version of the predetermined exercise, or incorrect version of the predetermined exercise.

The training and test data were collected from six patients. The exercises include Shoulder Flexion, Shoulder Abduction, External Rotation, Elbow Flexion and Extension, and combined PNF pattern. The numbers of correct and incorrect gestures performed for each exercise by each patient in the training set are 30 and 150, respectively. Thus, the total number of correct and incorrect gestures for each exercise is 180 and 900, respectively. The number of incorrect exercises is much higher because there are *five* different compensation mistakes defined for each exercise. The total number of nongestures in the training set is 1120, which is the number of motions captured during a data collection session with five other subjects. The number of correct and incorrect gestures performed for each exercise by each patient in the test set are both 5, resulting in a total of 30 correct and 30 incorrect gestures for each exercise. The number of nongestures for each patient in the test set is 10, resulting in a total of 60 nongestures. The correctness of each gesture is determined by the supervising occupational therapists.

The patients who performed the exercises are chosen so that they can perform the exercises without direct physical assistance. In this way, Kinect provides stable data and abrupt movement changes like stopping, accelerating or decelerating during the exercise are prevented. When recording a gesture, we fix the number of frames by adjusting the sampling rate. In other words, we use the same number of frames for each gesture regardless of the duration of the gesture. We reduced the sampling rate to the lowest degree that the accuracy is not disturbed. This process positively affects the computational performance of the recognition algorithm.

3.4. Hidden Markov Model Structure

HMMs have different types and the structure of each type dictates the recognition property of the model devised. Our choice of model type is a Left-to-Right model. It is already stated that the inherent temporal structure of the Left-to-Right model makes it useful for recognition problems that have temporal data like gesture and speech recognition.

The choice of model type is not sufficient to define our model structure. Another point that needs addressing is the number of states. One should consider the properties of the recognition problem concerned and determine the state count accordingly. However, one issue to pay attention to is that when the training data size is constant, increasing the state count result in declined performance [56]. Hence, one needs to find the minimum count of states to represent the gesture. Although the states in HMM are not a direct representation of frames or time intervals of the gesture, they tend to produce observation symbols showing similar feature properties. When we analyze our gesture set, each gesture starts with a "resting pose" [55], then the related upper-limb reaches a starting point and reaches the final pose before taking the same route back to resting pose. This process implies four stages of gesture (excluding resting poses at the beginning and end): rest-to-start, start-to-final, final-to-start, and start-to-rest. Thus, we designed our model with four states between one start and one end states, a total of six states. The difference between start and end states is that they have no self-transitions (see Fig. 8).

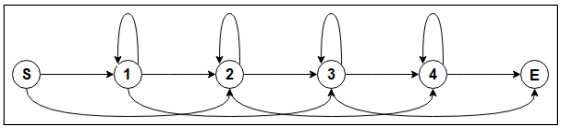

Fig. 8. The proposed HMM structure.

As it is seen from Fig. 8, each state can "skip" the state that supersedes it. These transitions are added to account for possible missing parts/frames of the gesture. The research shows that HMM can learn to skip several states at once, resulting in gestures with 4-5-6 state modeled using a 6-state structure [55]. The initial state transition probabilities are uniformly distributed. Another parameter in the design of the HMM is how the observations are modeled. We use a Gaussian autoregressive model to characterize the joint angles as continuous observations [56].

4. Detection of Non-gesture Patterns

HMM generates the recognition result by comparing the likelihoods of all trained models and selecting the one with the highest value. This approach works in contexts that all the input data is known to be within the predetermined set of gestures. However, when it is possible to have inputs that are not related to any gesture trained, like studies in [57], [54] and [58], HMM does not function as intended. Even though the input is not in any way related to the dataset, HMM picks the model with the highest probability and naturally, this phenomenon causes problems.

The problem we are dealing with also has a similar context. Despite our predetermined dataset with both correct and incorrect gesture performances, it is always expected from a subject to perform an entirely unrelated gesture. Subject walking in/out of Kinect's field of view, resting between exercises, therapists interventions are examples of such situations. Thus, we need to propose solutions to overcome this limitation of a conventional classification model.

Specifying a constant threshold value does not work because the likelihoods of the models fluctuate altogether depending on the input properties, the length of the observation sequence being one of them [54, 35]. Therefore, a mechanism should be devised that would produce an adaptive threshold value. This objective could be achieved by other models that generate a threshold value based on the input gesture. The ideal threshold value for a correct gesture would be less than that of the corresponding model and would be greater than that of all other models when a non-gesture is given as input. We have two solutions that could provide us with adaptive threshold values that are close to the ideal: *universal negative model* and *universal positive model*. We also compared the performance of our two methods with the performance of the threshold model proposed by Lee and Kim [59].

4.1. Universal Negative Model

The universal negative model is the concept of having a trained weak hidden Markov model that encapsulates all gestures that are not included in our dataset. Hereby, it is expected that when the observation sequence is a non-gesture, the model with the maximum likelihood would be the universal negative model. It can also be considered as another model competing with our gesture models with the distinction of representing multiple gestures.

During the training session, we directed a total of five subjects to perform various random gestures in front of Kinect for eight hours. The gestures included possible gestures that can be performed in the subjects' daily lives and possible recognition scenarios. The length, speed and number of joints used in gestures were not restricted during the training session. The only restriction was that none of the gestures should be similar to the ones we have in our dataset. A total of 1120 nongesture motions were recorded for the training of the universal negative model. Only one model is used for all of the correct gestures in our dataset.

The model designed for the universal negative model can be seen in Fig. 9. We used a parallel left-to-right model [56]. It is essentially a left-to-right model that obeys all state transition probabilities of linear left-to-right models. Its difference is that it is a cross-coupled connection of two parallel left-to-right models. We decided to use such a model because the gestures in the training set do not have well-defined properties, so it was not possible to generate a linear left-to-right model that fits the general properties. More importantly; because we used many different gestures for the training of this model, even though none of them was part of our original exercise set, the model was generating high likelihoods even for defined gestures. Thus, we needed a more fitting model and the parallel left-to-right structure we implemented with more hidden states delivered the desired results. The intuition behind the choice of a parallel model is to be able to model the movements of the left and right arms separately.

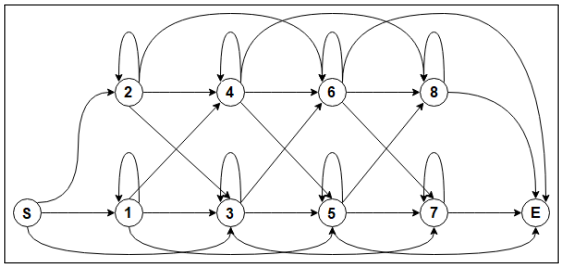

Fig. 9. The structure of the universal negative model.

4.2. Universal Positive Model

Each universal positive model is a model for the superset of all gestures for each exercise in the dataset. When the input belongs to the set of correct or faulty gestures for a specific exercise, this model is expected to generate a likelihood value that is less than that of the corresponding model for that exercise. For this approach to work successfully, the model should be loosely fitted. In this way, the model is more likely to fit better to the non-gesture than the models for the correct or faulty gestures when the given input is a non-gesture. However, if it behaves exactly like an average model, this approach can generate values that are always less than the likelihood produced for some of the gestures (correct or faulty) in our dataset. Because the universal positive model uses gesture examples for training, we used the same type of HMM structure that we used for our defined gestures (see Fig. 8).

4.3. Threshold Model

The threshold model proposed by Lee and Kim [59] is another HMM-based technique for the detection of non-gestures (see Fig. 10). Its purpose is similar to that of the universal positive model. It is a weak model for all trained gestures in the dataset. The difference of the threshold model with our universal models is that it is not a trained model, but rather a "generated" model. We used the same training samples for gestures in the dataset at once to train a universal positive model. However, Lee and Kim used part of the training data to train their threshold model. They first trained their gesture models separately and then combined the hidden states of each gesture model, with their self-transition and observation-transition probabilities fixed, in the threshold model. To provide complete transitivity, it is designed as an ergodic model. It would be disadvantageous to use an ergodic model because it does not have the temporality that the left-to-right models have. This disadvantage is critical because it is what makes this approach works on theory. Because the threshold model will have all the states of the corresponding model, it will also be able to match the positive input gesture. However, the specifically-trained model will have a better fit because it represents the temporal relations between the states better whereas the threshold model is ergodic.

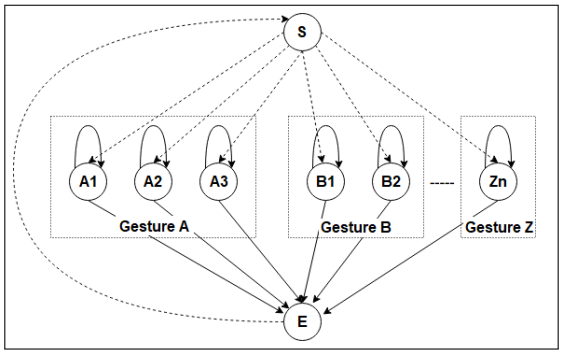

Fig. 10. The structure of the threshold model, based on [59].

One potential weakness of the threshold model is that naturally it has a large number of states and this causes significant performance loss in terms of processing speed. To overcome this weakness, they reduced the number of states based on relative entropy, which has been used as a measure of the distance between two probability distributions [59, 60]. Because we do not need to differentiate gestures from each other and we aim to capture the mistakes in each gesture, we do not have a large number of models in each run. Hence, we did not need to carry out such a reduction in the implementation of the threshold model.

We expect that the threshold model having the same states as the exercise model will generate a reasonable likelihood when the input is a gesture. For the threshold model to detect a nongesture, it must generate a likelihood greater than those of the specific gesture models. This is achieved in the original model by using the combination of a large number of states. However, the threshold model that we use includes only a small number of states obtained from the models of the five exercises and may not be as effective. Hence, the likelihood it generates for a nongesture may be less than that of a particular exercise model and it will be unable to classify nongestures correctly.

4.4. Comparison

In Table 1, we observe that the universal negative and threshold models have similar performances overall; the universal negative model having a higher F1 score in some exercises while the threshold model having higher scores in others. The precision and recall values are also similar, so it can be concluded that these methods perform similarly when generating an adaptive threshold value. However, it should be noted that the performance of the universal negative model depends on the training set.

The universal positive model performs poorly compared to the other two approaches. While the precision score of the universal positive model is lower than those of the universal negative and threshold models, the real difference is in the recall scores. One can observe that the number of false-negatives is much higher in the universal positive model for each gesture, hence, it leads to a significantly low recall value. This could be explained by the universal positive model having a temporal structure like the actual gesture models. One thing that makes the threshold model advantageous is its ergodic structure, which makes the gesture models have high likelihoods for correct gestures. As the universal positive model has a temporal structure (left-to-right model), it sometimes generates higher probabilities for defined gestures than the corresponding gesture models, and as a result, produces false negatives.

Table 1. The comparison of the universal negative, the universal positive, and the threshold models.

	Model	TP	FN	TN	FP	Precision	Recall	F1 Score
Shoulder flexion	Univ. neg.	51	9	58	2	0.9623	0.8500	0.9027
	Univ. pos.	41	19	52	8	0.8367	0.6833	0.7523
	Threshold	53	7	56	4	0.9298	0.8833	0.9060
Shoulder abduction	Univ. neg.	54	6	56	4	0.9310	0.9000	0.9153
	Univ. pos.	45	15	55	5	0.9000	0.7500	0.8182
	Threshold	53	7	58	2	0.9636	0.8833	0.9217
External rotation	Univ. neg.	59	1	60	0	1.0000	0.9833	0.9916
	Univ. pos.	39	21	49	11	0.7800	0.6500	0.7091
	Threshold	53	7	57	3	0.9464	0.8833	0.9138
Elbow flexion and extension	Univ. neg.	60	0	60	0	1.0000	1.0000	1.0000
	Univ. pos.	47	13	54	6	0.8868	0.7833	0.8319
	Threshold	56	4	56	4	0.9333	0.9333	0.9333
Combined PNF pattern	Univ. neg.	49	11	52	8	0.8596	0.8167	0.8376
	Univ. pos.	43	17	51	9	0.8269	0.7167	0.7679
	Threshold	55	5	52	8	0.8730	0.9167	0.8943
Cumulative	Univ. neg.	273	27	286	14	0.9512	0.9100	0.9302
	Univ. pos.	215	85	261	39	0.8465	0.7167	0.7762
	Threshold	270	30	279	21	0.9278	0.9000	0.9137

5. Detection of Compensation Mistakes

Unlike most approaches that focus on gesture recognition, our purpose is not to differentiate multiple gestures from each other, but rather differentiate between correct and incorrect versions of each specific gesture. Thus, we have several types of "incorrect" gestures for each gesture, which are versions of the same gesture performed in an undesired way. These mistakes are determined and added to our dataset according to the guidance of the therapists. Hence, this study aims to differentiate each gesture from their "incorrect" versions.

Because of the nature of this problem, traditional gesture recognition approaches may not perform well. This problem has some peculiarities and the traditional approaches perform relatively poor in similar scenarios [47]. A more tailored method needs to be engineered to achieve improved accuracy in differentiating these correct and incorrect gestures. To this end, we propose a two-stage method, which include *feature thresholding* and *negative models*.

5.1. Feature Thresholding

The feature set for each gesture is determined with the help of extensive testing under the supervision of occupational therapists. The features that would define the gesture in the best way possible and provide the best differentiation are chosen. However, the goal is to differentiate each gesture from its incorrect versions, not from other gestures. As a result, another issue we focus on when determining the features is the characteristics of these incorrect gestures.

For each gesture, the most common and most undesired mistakes were categorized by the occupational therapists and the joint angles that best define these mistakes were identified. In this way, the joint angles and the critical values that should or should not be exceeded by the patient are determined for every incorrect gesture version.

The number of features that we use is restricted to be between four and six. The four features (two shoulder angles and two elbow angles) are often necessary and sufficient to define each gesture. Nevertheless, when the type of compensation moves and the resulting incorrect gestures are examined, extra features need to be added.

Because CP patients often suffer from muscle stiffness, they try to compensate for this stiffness by activating (flexing and extending) other muscles in their bodies. Focusing on upper extremity exercises, the compensation tendencies and common mistakes are defined by the therapists as a result of performing each exercise with CP patients. When performing upper extremity exercises, three main compensation techniques come up: bending the body forward or backward, bending the neck to activate the upper shoulder muscles, and bending the elbow for shoulder exercises. Hence, additional to the elbow and shoulder angles, the body and head angles are also added to the feature set of each exercise so that we can catch the incorrect gestures.

There are two types of restrictions that define a gesture as an incorrect one apart from being classified as a non-gesture:

- *Type 1*: Moving a joint that should stay fixed more than a specified value. For instance, when performing the shoulder flexion exercise; the elbow angle on the xz-plane should stay between 15° and -15° and the shoulder angle on the xz-plane between 75° and 105°. These two movements, i.e., moving the shoulder forward or bending the elbow, take the tension from the shoulder muscles that should complete the exercise and make the exercise ineffective. Similar restrictions also exist for the body and head angles on different exercises.

- *Type 2*: Not reaching or exceeding a target angle. For the shoulder flexion exercise, the shoulder angle on the yz-plane should reach a value between 75° and 105° at its peak point, stay there for a while and then decrease. The real target value is 90° where the 15° tolerance value is determined considering the inaccuracy of the depth camera.

5.2. Negative Models

Negative models are new models that represent incorrect gestures. They are conceptually similar to the Universal Negative Model, but the training sets for these models include specific gestures. These models are trained using determined compensation mistakes as the training set and using the same features like the correct gesture model. We experiment with two types of negative models: *fault-specific negative model* and *gesture-specific negative model*.

5.2.1. Fault-specific Negative Model

The fault-specific gesture model is the basic application of the negative model concept. In this approach, a separate model is trained for each different compensation mistake. The downside of this is that the computational cost is increased as the number of types of mistakes increases.

We compared the fault-specific negative model to the baseline solution that does not use any specific negative model other than the universal negative model. The only possible scenario for a baseline solution to classify a compensation mistake is by classifying it as a non-gesture. Hence, the sum of false negatives and true negatives for each gesture equals to non-gesture count.

The results presented in Table 2 show the superiority of the fault-specific negative model in terms of accuracy. Since the baseline solution, which is similar to gesture recognition solutions that are used in generic exercise recognition problems, is not designed specifically to solve the problem of small mistakes, such a difference in accuracy is expected. One can see that the baseline solution performs better in terms of recall value. This is because the baseline solution classifies most of the incorrect gestures as correct gestures and false positives are not taken into account when calculating recall. However, the objective of the fault specific approach is to reduce false positives, and in that case, the precision value is very important for comparison.

5.2.2. Gesture-specific Negative Model

The gesture-specific negative model encapsulates all types of mistakes related to one exercise in a single model for each different exercise. The reason we applied this approach is to increase the processing speed. Even though we did not have a large number of mistakes to observe a significant performance gain in our tests, such a solution could be needed for different exercises or patient types.

The gesture-specific negative model is compared to the baseline solution in the same way as the fault-specific negative model (see Table 2). The results show us that the gesture-specific negative model performs better than the baseline solution in terms of precision and F1 Score. The baseline solution has a better recall value overall, but as discussed in previous sections, the recall value is not critical in this case.

5.2.3. Comparison

When we compare the results of these two approaches, we see that the fault-specific negative model generates better results than the gesture-specific negative model. This is because the fault-specific negative model learns a separate model for each specific type of mistake for each gesture. However, the gesture-specific model has different gestures in its training set, hence, it is not as successful as the fault-specific model for learning

Table 2. The comparison of the gesture-specific negative model, the fault-specific negative model, and the baseline solution.

	Model	TP	FN	TN	FP	Precision	Recall	F1 Score
Shoulder flexion	Baseline	24	6	9	21	0.5333	0.8000	0.6400
	Gesture-specific	22	8	19	11	0.6667	0.7333	0.6984
	Fault-specific	23	7	22	8	0.7419	0.7667	0.7541
Shoulder abduction	Baseline	25	5	4	26	0.4902	0.8333	0.6173
	Gesture-spec.	23	7	18	12	0.6571	0.7667	0.7077
	Fault-specific	25	5	21	9	0.7353	0.8333	0.7813
External rotation	Baseline	29	1	4	26	0.5273	0.9667	0.6824
	Gesture-specific	25	5	20	10	0.7143	0.8333	0.7692
	Fault-specific	27	3	26	4	0.8710	0.9000	0.8852
Elbow flexion and extension	Baseline	26	4	2	28	0.4815	0.8667	0.6190
	Gesture-specific	21	9	19	11	0.6563	0.7000	0.6774
	Fault-specific	25	5	24	6	0.8065	0.8333	0.8197
Combined PNF pattern	Baseline	27	3	10	20	0.5745	0.9000	0.7013
	Gesture-spec.	22	8	23	7	0.7586	0.7333	0.7458
	Fault-specific	24	6	22	8	0.7500	0.8000	0.7742
Cumulative	Baseline	131	19	29	121	0.5198	0.8733	0.6517
	Gesture-specific	113	37	99	51	0.6890	0.7533	0.7197
	Fault-specific	124	26	115	35	0.7799	0.8267	0.8026

the individual mistakes for each gesture. The rationale for the usage of the gesture-specific approach was to reduce the computational burden. While only one extra model is calculated for the gesture-specific model, the fault-specific model requires as many models as the number of defined mistakes.

6. Evaluation and Results

We proposed gesture recognition methods to provide a better solution to the CP gesture recognition problem. In Sections 4 and 5, we compared our proposed solutions to state-of-the-art approaches as baseline solutions. In this section, we provide the overall results of our proposed solution and present the results of our solution for occupational therapy exercises.

We describe the method we used when conducting the tests as follows. During our study, constant testing with occupational therapists and CP patients took place. Our target users were hemiplegic CP patients that are classified in levels 1 or 2 of GMFCS and levels 2 or 3 of MACS standards. Six children with CP between the ages of 7 and 12 were chosen for performing the exercises. A total of six detailed testing sessions were completed in 24 weeks. Each session lasted 45-60 minutes. The results presented in Sections 4 and 5 belong to the last session. Previous sessions are performed for different reasons: restricting the number of features, selecting features, defining compensation mistakes, tracking the children's progress, and so on.

By collecting the data for all six sessions, we can observe the patients' performance and evaluate the overall benefits of our solution. Nevertheless, it should be noted that the results obtained here do not prove that the improvements to their performance are solely the result of using our solution. During this phase, these children were continuing their conventional rehabilitation programs and were having exercise sections in related facilities. Restricting their rehabilitation program to our solution is not possible and creating control groups having similar levels of complications is demanding medically and requires special permissions.

We use the success rate to evaluate children's progress. It is simply the ratio of correctly performed exercises to all gestures performed by the children. A parallel study conducted by occupational therapists used a different method to measure the progress of children. Dynamic Occupational Therapy Cognitive Assessment for Children (DOTCA-Ch) is used to assess the children. All data were collected strictly anonymously by an experienced therapist who was blind to the treatment protocol. DOTCA-Ch also evaluates the child's cognitive state.

The pre-intervention scores of DOTCA-Ch were 3.81 ± 2.26 in orientation, 5.27 ± 2.09 in motor control and 15.72 ± 8.51 in visuomotor construction. After the last session was completed, the orientation score was improved to 5.09 ± 2.15, the motor control was at 6.09 ± 1.77 and the visuomotor construction was 18.54 ± 7.77. These measures show a significant statistical difference in performance. It should be noted that cognitive abilities are also taken into account.

Fig. 11 shows the improvement in children's success rates when performing the determined five exercises. The presented data are the cumulative result of all children. A gradual increase is observed for all five exercises.

7. Conclusions and Future Work

We propose a new approach that makes it possible to use gesture recognition for occupational therapy exercises with children with cerebral palsy. To differentiate gestures that are not defined exercises, which is an important problem in our case considering the cognitive impairments of the children, we propose an alternative method called the universal negative model and universal positive model. The purpose of these methods is to generate an adaptive threshold model. We were able to get results comparable to a successful method in the literature. We also propose various solutions for capturing the exercise mistakes done by the patients to compensate for their lack of muscle control and muscle strength. These incorrect exercises generally resemble the original exercise and thus classified as a correct exercise by traditional gesture recognition algorithms. With the help of our new approach, it is possible to get reasonable results compared to the conventional approach.

Because this is not a problem dealt with by previous studies, it is not possible to make a direct comparison. Other approaches focus on other aspects of gesture recognition whereas we focus on capturing compensation mistakes. However, the effects of our approach on children's motor control and orientation progress are examined and a significant improvement is observed.

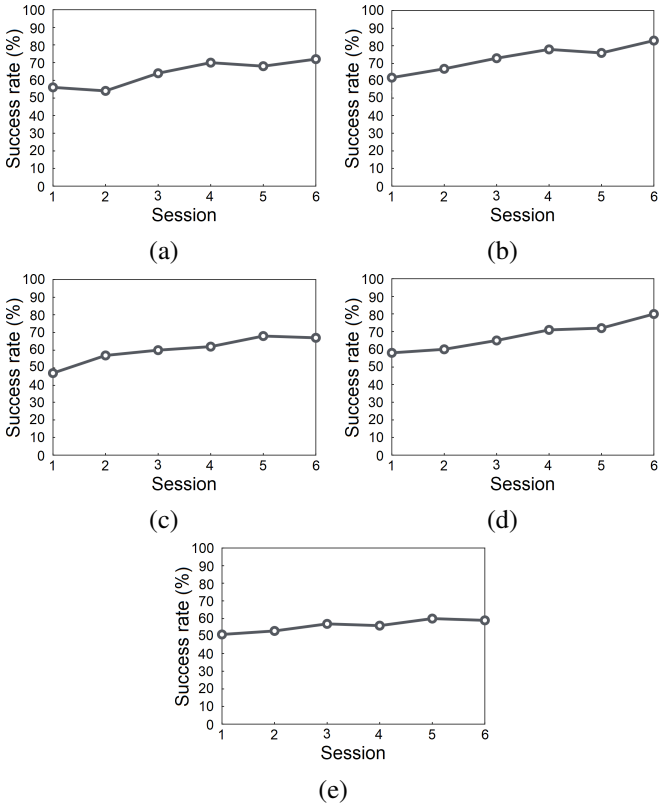

Fig. 11. The success rate graph for (a) shoulder flexion, (b) shoulder abduction, (c) external rotation, (d) elbow flexion and extension, and (e) combined PNF pattern.

A very basic method to separate the time frames of each gesture from one another is implemented. We used the starting and ending poses of each gesture for this purpose. As future work, a sliding-window based method could be used. Recent developments in deep learning methods (e.g., [61],[62]) could also be utilized. Adapting the solutions we proposed to deep neural network approaches could be a good research direction in the future.

Acknowledgments

We are grateful to the personnel and the CP patients of the Department of Occupational Therapy at Hacettepe University, Ankara, Turkey, specifically Prof. Dr. Hülya Kayıhan and her colleagues, who allowed us to perform the experiments in their facilities as well as helped with the experiments.

References

[1] M. Bax, M. Goldstein, P. Rosenbaum, A. Leviton, N. Paneth, B. Dan, B. Jacobsson, D. Damiano, Proposed definition and classification of cerebral palsy, Developmental Medicine & Child Neurology 47 (2005) 571 – 576.

[2] D. Webster, O. Celik, Systematic review of Kinect applications in elderly care and stroke rehabilitation, Journal of NeuroEngineering and Rehabilitation 11 (2014) 108.

[3] C. Yan, L. Li, C. Zhang, B. Liu, Y. Zhang, Q. Dai, Cross-modality bridging and knowledge transferring for image understanding, IEEE Transactions on Multimedia 21 (2019) 2675–2685.

[4] C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, Q. Dai, A fast Uyghur text detector for complex background images, IEEE Transactions on Multimedia 20 (2018) 3389–3398.

[5] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, STAT: Spatial-temporal attention mechanism for video captioning, IEEE Transactions on Multimedia 22 (2020) 229–241.

[6] A. Pérez-Munoz, P. Ingavélez-Guerra, Y. Robles-Bykbaev, New approach of serious games in ludic complements created for rehabilitation therapies in children with disabilities using Kinect, in: Proceedings of the IEEE XXV International Conference on Electronics, Electrical Engineering and Computing, INTERCON '18, 2018, pp. 1–4.

[7] I. Ar, Y. S. Akgul, A computerized recognition system for the home-based physiotherapy exercises using an RGBD camera, IEEE Transactions on Neural Systems and Rehabilitation Engineering 22 (2014) 1160–1171.

[8] J. F. Lin, D. Kulić, Online segmentation of human motion for automated rehabilitation exercise analysis, IEEE Transactions on Neural Systems and Rehabilitation Engineering 22 (2014) 168–180.

[9] L. E. Sucar, R. Luis, R. Leder, J. Hernández, I. Sánchez, Gesture therapy: A vision-based system for upper extremity stroke rehabilitation, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology, EMBS '10, 2010, pp. 3690–3693.

[10] C. Liang, L. Qi, Y. He, L. Guan, 3D human action recognition using a single depth feature and locality-constrained affine subspace coding, IEEE Transactions on Circuits and Systems for Video Technology 28 (2018) 2920–2932.

[11] X. Zhen, L. Shao, D. Tao, X. Li, Embedding motion and structure features for action recognition, IEEE Transactions on Circuits and Systems for Video Technology 23 (2013) 1182–1190.

[12] D. Wu, L. Shao, Silhouette analysis-based action recognition via exploiting human poses, IEEE Transactions on Circuits and Systems for Video Technology 23 (2013) 236–243.

[13] H. Wang, J. Fu, Y. Lu, X. Chen, S. Li, Depth sensor assisted real-time gesture recognition for interactive presentation, Journal of Visual Communication and Image Representation 24 (2013) 1458–1468.

[14] S. Ong, S. Ranganath, Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 873–891.

[15] A. Jaimes, N. Sebe, Multimodal human-computer interaction: A survey, Computer Vision and Image Understanding 108 (2007) 116 – 134.

[16] S. Mitra, T. Acharya, Gesture recognition: A survey, IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews 37 (2007) 311–324.

[17] A. I. Maqueda, C. R. del Blanco, F. Jaureguizar, N. Garca, Human-computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns, Computer Vision and Image Understanding 141 (2015) 126 – 137.

[18] P. Barros, N. T. Maciel-Junior, B. J. Fernandes, B. L. Bezerra, S. M. Fernandes, A dynamic gesture recognition and prediction system using the convexity approach, Computer Vision and Image Understanding 155 (2017) 139 – 149.

[19] S. Ghodsi, H. Mohammadzade, E. Korki, Simultaneous joint and object trajectory templates for human activity recognition from 3-D data, Journal of Visual Communication and Image Representation 55 (2018) 729–741.

[20] H. Cheng, L. Yang, Z. Liu, Survey on 3D Hand Gesture Recognition, IEEE Transactions on Circuits and Systems for Video Technology 26 (2016) 1659–1673.

[21] M. Chen, G. AlRegib, B. Juang, Feature processing and modeling for 6D motion gesture recognition, IEEE Transactions on Multimedia 15 (2013) 561–571.

[22] P. K. Pisharady, M. Saerbeck, Recent methods and databases in vision-based hand gesture recognition: A review, Computer Vision and Image Understanding 141 (2015) 152 – 165.

[23] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, in: Time-of-Flight and Depth Imaging: Sensors, Algorithms, and Applications, Lecture Notes in Computer Science, volume 8200, Springer, 2013, pp. 149–187. doi:10.1007/978-3-642-44964-2_8.

[24] A. D. Călin, Gesture recognition on Kinect time series data using Dynamic Time Warping and Hidden Markov Models, in: Proceedings of the 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC '16, 2016, pp. 264–271.

[25] K. K. Biswas, S. K. Basu, Gesture recognition using Microsoft Kinect, in: Proceedings of the 5th International Conference on Automation, Robotics

and Applications, volume 2, IEEE, 2011, pp. 100–103. doi:10.1109/ICARA.2011.6144864.

[26] N. H. Dardas, N. D. Georganas, Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques, IEEE Transactions on Instrumentation and Measurement 60 (2011) 3592–3607.

[27] S. Sempena, N. U. Maulidevi, P. R. Aryan, Human action recognition using Dynamic Time Warping, in: Proceedings of the Electrical Engineering and Informatics, ICEEI '11, 2011, pp. 1–5. doi:10.1109/ICEEI.2011.6021605.

[28] M. Hu, C. Chen, W. Cheng, C. Chang, J. Lai, J. Wu, Real-time human movement retrieval and assessment with Kinect sensor, IEEE Transactions on Cybernetics 45 (2015) 742–753.

[29] Y. Yao, Y. Fu, Contour model-based hand-gesture recognition using the Kinect sensor, IEEE Transactions on Circuits and Systems for Video Technology 24 (2014) 1935–1944.

[30] L. Zhang, S. Zhang, F. Jiang, Y. Qi, J. Zhang, Y. Guo, H. Zhou, BoMW: Bag of manifold words for one-shot learning gesture recognition from Kinect, IEEE Transactions on Circuits and Systems for Video Technology 28 (2018) 2562–2573.

[31] C. Wang, Z. Liu, S. Chan, Superpixel-based hand gesture recognition with Kinect depth camera, IEEE Transactions on Multimedia 17 (2015) 29–39.

[32] Z. Ren, J. Yuan, J. Meng, Z. Zhang, Robust part-based hand gesture recognition using Kinect sensor, IEEE Transactions on Multimedia 15 (2013) 1110–1120.

[33] Z. Yang, Y. Li, W. Chen, Y. Zheng, Dynamic hand gesture recognition using hidden Markov models, in: Proceedings of the 7th International Conference on Computer Science & Education, ICCSE '12, 2012, pp. 360–365. doi:10.1109/ICCSE.2012.6295092.

[34] M. Z. Uddin, N. Thang, T.-S. Kim, Human activity recognition via 3-D joint angle features and hidden Markov models, in: Proceedings of the Int. Conference on Image Processing, ICIP '10, 2010, pp. 713–716.

[35] Y. Dennemont, G. Bouyer, S. Otmane, M. Mallem, A discrete Hidden Markov models recognition module for temporal series: Application to real-time 3D hand gestures, in: Proceedings of the 3rd International Conference on Image Processing Theory, Tools and Applications, IPTA '12, 2012, pp. 299–304.

[36] J. Suarez, R. R. Murphy, Hand gesture recognition with depth images: A review, in: Proceedings of the IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, IEEE, 2012, pp. 411–417. doi:10.1109/ROMAN.2012.6343787.

[37] G. Lefebvre, S. Berlemont, F. Mamalet, C. Garcia, Inertial gesture recognition with BLSTM-RNN, in: P. Koprinkova-Hristova, V. Mladenov, N. K. Kasabov (Eds.), Artificial Neural Networks, Springer International Publishing, Cham, 2015, pp. 393–410.

[38] H. M. Ertunc, K. A. Loparo, H. Ocak, Tool wear condition monitoring in drilling operations using hidden Markov models (HMMs), International Journal of Machine Tools and Manufacture 41 (2001) 1363 – 1384.

[39] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 20–27.

[40] N. Granger, M. A. el Yacoubi, Comparing hybrid NN-HMM and RNN for temporal modeling in gesture recognition, in: D. Liu, S. Xie, Y. Li, D. Zhao, E.-S. M. El-Alfy (Eds.), Neural Information Processing, Springer Int. Publishing, Cham, 2017, pp. 147–156.

[41] American Occupational Therapy Association, Occupational therapy practice framework: Domain and process, American Journal of Occupational Therapy 56 (2002) 609–639.

[42] H. M. H. Pendleton, W. Schultz-Krohn, Pedretti's Occupational Therapy - E-Book: Practice Skills for Physical Dysfunction, Factsbook, Elsevier Health Sciences, 2013.

[43] W. Yu, R. Dubey, N. Pernalete, Robotic therapy for persons with disabilities using Hidden Markov Model based skill learning, in: Proceedings of the IEEE International Conference on Robotics and Automation, volume 2 of *ICRA '04*, 2004, pp. 2074–2079. doi:10.1109/ROBOT.2004.1308129.

[44] Y.-J. Chang, W.-Y. Han, Y.-C. Tsai, A Kinect-based upper limb rehabilitation system to assist people with cerebral palsy, Research in Developmental Disabilities 34 (2013) 3654–3659.

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

[46] M. Pedraza-Hueso, S. Martín-Calzón, F. J. Díaz-Pernas, M. Martínez-Zarzuela, Rehabilitation using Kinect-based games and virtual reality, Procedia Computer Science 75 (2015) 161–168.

[47] V. Bloom, D. Makris, V. Argyriou, G3D: A gaming action dataset and real time action recognition evaluation framework, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 7–12.

[48] L. Reid, S. E Rose, R. Boyd, Rehabilitation and neuroplasticity in children with unilateral cerebral palsy, Nature Reviews, Neurology 11 (2015) 390–400.

[49] M. Mutsaarts, B. Steenbergen, H. Bekkering, Anticipatory planning of movement sequences in hemiparetic cerebral palsy, Motor Control 9 (2005) 439–58.

[50] C. Morris, D. Bartlett, Gross Motor Function Classification System: Impact and Utility, Developmental Medicine and Child Neurology 46 (2004) 60–65.

[51] A. Eliasson, L. Krumlinde-Sundholm, B. Rösblad, E. Beckung, M. Arner, A. Öhrvall, P. Rosenbaum, The Manual Ability Classification System (MACS) for children with cerebral palsy: Scale development and evidence of validity and reliability, Developmental Medicine and Child Neurology 48 (2006) 549–554.

[52] R. T. Collins, R. Gross, J. Shi, Silhouette-based human identification from body shape and gait, in: Proceedings of the 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR '02, 2002, pp. 366–371. doi:10.1109/AFGR.2002.1004181.

[53] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3D points, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2010) 9–14.

[54] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 257–267.

[55] L. W. Campbell, D. A. Becker, A. Azarbayejani, A. F. Bobick, A. Pentland, Invariant features for 3-D gesture recognition, in: Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996, pp. 157–162. doi:10.1109/AFGR.1996.557258.

[56] L. R. Rabiner, A Tutorial on Hidden Markov Models and selected applications in speech recognition, Proceedings of the IEEE 77 (1989) 257–286.

[57] N. D. Binh, E. Shuichi, T. Ejima, Hand gesture recognition using a real time tracking methods and pseudo hidden Markov model, International Conference on Graphics, Vision and Image Processing (2005) 362–368.

[58] X. Zabulis, H. Baltzakis, A. Argyros, Vision-based hand gesture recognition for human-computer interaction, in: The Universal Access Handbook, CRC Press, 2009, pp. 341–343.

[59] H.-K. Lee, J. H. Kim, An HMM-based threshold model approach for gesture recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (1999) 961–973.

[60] T. M. Cover, J. A. Thomas, Entropy, relative entropy, and mutual information, in: Elements of Information Theory, John Wiley & Sons, 2001, pp. 12–49.

[61] G. Devineau, F. Moutarde, W. Xi, J. Yang, Deep learning for hand gesture recognition on skeletal data, in: Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG '18, 2018, pp. 106–113.

[62] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. Escalante, V. Ponce-Lopez, X. Baro, I. Guyon, S. Kasaei, S. Escalera, A survey on deep learning based approaches for action and gesture recognition in image sequences, in: Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG '17, 2017, pp. 476–483.