CogniLearn: A Deep Learning-based Interface for Cognitive Behavior Assessment

Srujana Gattupalli
CSE, University of Texas at Arlington
Arlington, Texas, USA
srujana.gattupalli@mavs.uta.edu

Dylan Ebert
CSE, University of Texas at Arlington
Arlington, Texas, USA
dylan.ebert@mavs.uta.edu

Michalis Papakostas
CSE, University of Texas at Arlington
Arlington, Texas, USA
michalis.papakostas@mavs.uta.edu

Fillia Makedon
CSE, University of Texas at Arlington
Arlington, Texas, USA
makedon@uta.edu

Vassilis Athitsos
CSE, University of Texas at Arlington
Arlington, Texas, USA
athitsos@uta.edu

ABSTRACT

This paper proposes a novel system for assessing physical exercises specifically designed for cognitive behavior monitoring. The proposed system provides decision support to experts for helping with early childhood development. Our work is based on the well-established framework of Head-Toes-Knees-Shoulders (HTKS) that is known for its sufficient psychometric properties and its ability to assess cognitive dysfunctions. HTKS serves as a useful measure for behavioral self-regulation[22]. Our system, CogniLearn, automates capturing and motion analysis of users performing the HTKS game and provides detailed evaluations using state-of-the-art computer vision and deep learning based techniques for activity recognition and evaluation. The proposed system is supported by an intuitive and specifically designed user interface that can help human experts to cross-validate and/or refine their diagnosis. To evaluate our system, we created a novel dataset, that we made open to the public to encourage further experimentation. The dataset consists of 15 subjects performing 4 different variations of the HTKS task and contains in total more than 60,000 RGB frames, of which 4,443 are fully annotated.

Author Keywords

Computer Vision; Deep Learning; Human Computer Interaction (HCI); Head-Toes-Knees-Shoulders (HTKS); Cognitive Assessment

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces; I.2.1. Artificial Intelligence: Applications and Expert Systems; I.2.6. Artificial Intelligence: Learning; I.2.10. Artificial Intelligence: Vision and Scene Understanding; I.4.8 Image Processing And Computer Vision: Scene Analysis; K.3.1 Computers And Education: Computer Uses in Education

IUI 2017, March 13–16, 2017, Limassol, Cyprus.

http://dx.doi.org/10.1145/3025171.3025213

INTRODUCTION

Cognitive impairments in early childhood can lead to poor academic performance and require proper remedial intervention at the appropriate time [22]. ADHD affects about 6-7% of children [15, 25] and occurs about three times more frequently in boys than in girls [31]. According to [5, 11, 19] ADHD is a psychiatric neurodevelopmental disorder that is very hard to diagnose or tell apart from other disorders. There are specific symptoms that can be observed in individuals suffering from the disease including inattention, inability to follow instructions, distractibility, hyperactivity or acting impulsively [21, 28]. Such cognitive insufficiencies hinder the development of working memory and can affect school success and even have long term effects that can result in low self-esteem and self-acceptance[8]. As shown in [23], the traditional game called Head-Toes-Knees-Shoulders (HTKS) can provide sufficient psychometric observations and can be used as a measure of behavioral self-regulation. According to the authors in [23] and their extended research in the task, HTKS is significantly related to cognitive flexibility, working memory, and inhibitory control. The game has three sections with up to four paired behavioral rules: "touch your head" and "touch your toes;" "touch your shoulders" and "touch your knees." Subjects first respond naturally, and then are instructed to switch rules by responding in the "opposite" way (e.g., touch their knees when told to touch their shoulders). HTKS, along with other similar cognitive assessment methods are being already practiced in around 300 schools around the US [35]. Our work aims to provide a computerized infrastructure for performing and evaluating the aforementioned assessments. We propose a system that will be responsible for training, monitoring and building cognitive abilities, which helps with identifying and overcoming such cognitive impairments. Our framework aims to provide useful performance measures like accuracy and measure of correctness. We also

aim to provide qualitative and quantitative performance summary. The ultimate goal is to deliver meaningful information to cognitive experts and help develop skills in children that can result in overall improvement of child's academic performance.

To that end, we exploit the demonstrated efficiency of the HTKS on predicting cognitive behavior and we provide a prototype user interface for recording and analyzing motion, when children perform the HTKS in front of a *Microsoft Kinect V2* camera. Our system operates based on state-of-the-art machine learning techniques on pose estimation from RGB video streams. In particular, we build upon the deep-learning architecture proposed in [14] and we use this as a base module for evaluating our task-specific application. The selected deep-architecture exploits a Convolutional-Neural-Network (CNN) and performs pose estimation providing highly accurate results regarding the relative position of human body-joints. We monitor and analyze hand-position with respect to the body joints of our interest (i.e. Head,Toes,Knees and Shoulders) and we evaluate if user's action complies with the expected motion (i.e. What body part did he/she touch, if he/she were asked to touch his/her head?). To evaluate our system, we capture a novel dataset with 15 subjects performing four different modes of the HTKS game. As these are our initial experimentation on the task, and access to real subjects was not yet feasible, our dataset consists of 15 adult individuals, mainly undergraduate students. However all subjects had no previous experience on the game or how their motion would be evaluated, thus avoiding unwanted biases. To illustrate results and motion analysis, we developed a novel user interface that can facilitate therapists and human experts who monitor the performing subjects and provide them with the desired measurements.

The rest of the paper is structured as follows: First we present related work on similar technologies and efforts and we justify our choices on the technical part. In following sections we discuss in detail the methodology of our approach, the experimental architecture, the proposed user interface, and finally our experiments by describing our data collection process and the system accuracy obtained on our dataset. The paper concludes with the presentation of our overall conclusions and our next experimentation steps on the task.

RELATED WORK

Emerging technologies have significantly influenced several medical related processes, such as diagnosis, rehabilitation and treatment. The effect of computer science in the medical domain is observable not only in the level of human computer interaction but also on the quality and the quantity of useful data that a modern system can automatically capture and provide to the experts as assistive material to the diagnosis.
The implementation of such systems must meet two major criteria: a) keep the user motivated and b) provide meaningful and understandable data to the domain experts[24]. Towards that direction various works have been proposed that try to access different but similar medical conditions. In [13], the authors proposed after extended research, that active video game play (i.e. consoles like Microsoft Kinect or Nitendo Wii) can promote physical activity and rehabilitation of children with Cerebral Palsy. In [18], an interactive game was developed to assist stroke-patients improve their balance. In [2], the authors focused on extending the attention span of children with ADHD by designing a computer software application that constantly monitors the users attention state using an eye-tracker and adapts its user interface for incorporating multiple stimuli. In [3], the researchers deployed a humanoid robot to teach children who suffer from complex developmental disabilities, simple coordinated behaviors. The authors in [32] designed a virtual-reality game for upper-limb motor-rehabilitation, while in [27] a virtual reality environment was deployed for the assessment and rehabilitation of attention deficits in children, and especially ADHD. In [6], a set of new game designs is presented, that is based on psychological tests or tasks and aims to monitor or improve ADHD related symptoms.
The system we propose in this paper fits very well as a component to the framework suggested in [33], where the authors proposed a system that combines two types of feedback to the therapists (one directly from the user and one automatically generated by a computer vision-based mechanism). Our work is based on the well-established framework of HTKS as a tool for assessing cognitive dysfunctions [22], while at the same time employs state-of-the-art vision based techniques for activity recognition and evaluation. In it is packaged with a carefully designed UI that is intuitive and motivates both users (therapists and patients) to interact with. Computation-wise we exploit the remarkable performance reported on such tasks by deep-learning approaches. Deep-learning and especially CNNs have shown very good results in activity recognition tasks [9, 12] compared to traditional approaches based on shallow classifiers and hand-crafted features [20, 36].
Finally as other similar systems [7, 26, 29, 34] our application outputs metrics that are valuable to the experts and can add significant information to the level of treatment and diagnosis. Those metrics are defined in the HTKS protocol and are mainly related to the number of errors (if the subject performed the expected motion) and the delay (how much time did the subject require to perform a specific command).

METHODOLOGY

Our primary goal has been to build a novel framework, to automate capture and assessment of performance of the HTKS self-regulatory task in order to provide a platform to our users to monitor and advance participants' cognitive skills. The primary purpose of this work is to provide a novel framework that can serve as a tool for evaluating and assessing physical activities, which can reveal potential cognitive dysfunctions. In particular, our systems deploys the HTKS framework, which has proven value as a cognitive assessment tool. More specifically, we provide a sophisticated system that is easy to use by general users and provides valuable data and meaningful measures and information that can benefit advanced users such as therapist and cognitive experts. For the purpose of this paper, we use the terms 'users' and 'instructor' synonymously and the person being observed while performing the HTKS task as the 'participant'. The proposed methodology is based on a well-defined, modular and well-

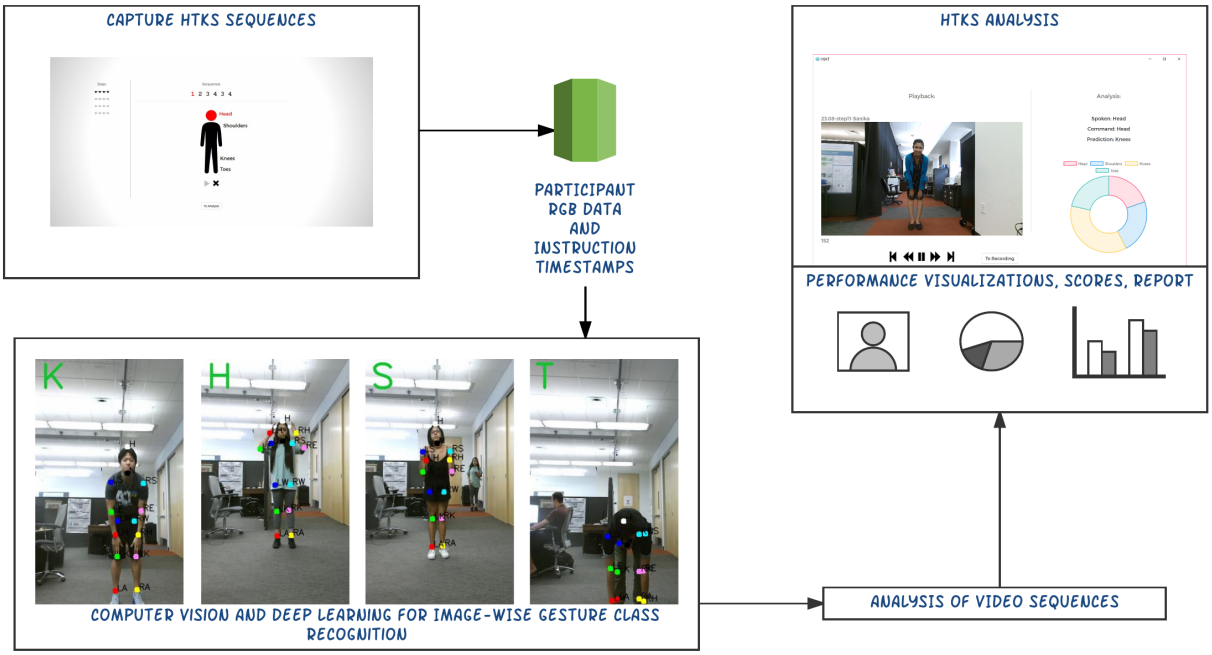

Figure 1: System Architecture

structured framework that consists of a user friendly front-end and a robust and dynamic back-end and provides outcomes based on latest computer vision and machine learning advances. We provide an automized module in our interface to capture RGB data from subjects while performing self-regulatory tasks according to the HTKS cognitive study protocol[23]. The interface to capture this data aims to enhance and support collaboration between the instructor and the participant during the study. Visual data is made available to the instructor to demonstrate the actual and intended motion and audio instructions are delivered to the participant to perform the preferred tasks and motion. Evaluations on this captured data are performed solely on the captured RGB dataset. The evaluation module of our interface gathers these information and provides systematic feedback regarding the assessment of these physical exercises to the human experts.

EXPERIMENTAL ARCHITECTURE

Our experimental setup is based on our framework as described above and is shown in Figure 1. We perform evaluation of HTKS motion accuracy over long sequences and provide performance measures such as whether the performed movement was consistent with the instructions, the response time of the participant and the steps where the participant made the most incorrect movements. The analysis of this motion is performed in a frame-by-frame manner by recognition of the HTKS gestures in each RGB frame and then performing analysis over longer video sequences. For this, first we perform human body pose estimation and then we apply our algorithm that classifies each RGB image frame into one of the four HTKS gestures classes(Head, Shoulders, Knees and Toes). Detailed description of this recognition and our algorithm is provided in the Physical Activity recognition section.

Physical Activity Recognition

In our approach, we classify each frame separately as belonging to one of the four H,T,K or S classes. We do this in two steps: 1.) we localize full-body joints co-ordinates and 2.) we use these co-ordinates to classify frames and assign image-wise gesture labels.

Full-body Joints Localization

The quality of the second approach depends highly on the skeleton detection and tracking algorithm. In many cases, skeleton tracking offered by Kinect provides poor results due to self-occlusions[37], This is observable in our dataset when tracking the limbs while the person bends to touch knees or toes. Some example visualizations of these self-occulusions during Kinect's skeleton tracking are shown in Figure 2 and because of this we decided on following our own way of pose recognition instead of using Kinect's skeleton tracking.

RGB data can be more consistent and less noisy than Kinect's skeleton tracking data and can work better in different lighting conditions and interference. Depth data could also be a valuable source of information to experiment with which we have not yet explored but we would like to in our further research. We decided on creating a computer vision based system using the RGB data and taking advantage of the latest technological advances in deep learning to obtain a pose estimator that works well for our problem.

For the joint localization problem on RGB data, we explored various existing state-of-the-art methods and considered different pose estimator models and decided on using a pose estimator called DeeperCut[14]. While some of the other pre-trained models works well with upper-body pose[12] and useful for certain other applications, our problem requires an efficient and accurate pose estimator for full-body human joint

localization as well as that could work with multi-person pose estimation. Using this we track 14 body joint locations *LS, LE, LH, LW, LK, LA,H, N, RS, RE, RH, RW, RK, RA* that correspond to body parts 'left shoulder', 'left elbow', 'left hand', 'left waist', 'left knee', 'left ankle', 'head', 'neck', 'right shoulder', 'right elbow', 'right hand', 'right waist', 'right knee', 'right ankle' respectively. We choose to use the pose estimator model stated above as it is the current state-of-the-art pose estimator for multi-person pose estimation and also provides competitive performance on single-person pose estimation as demonstrated by their results on popular datasets for these problems. Here, the pose estimation is performed by joints extraction using CNN-based body part detectors and then by further using deep networks to perform conditioning for these detected body parts pairwise for each image.
It is further beneficial for us to use this model as it is fast and accurate to aid with robustness of our user interface and its performance for multi-person pose estimation would be helpful in our next steps of enhancing our interface for the different settings that we mentioned before. This pre-trained model works on recognizing poses of adult participants as well as children as proven by the results on the popular *MPII human pose*[1] dataset for single and multi-person pose estimation which consists of annotated classes of adults as well as children, for example images in the MPII dataset with activity labels "Playing with children", "Child care".

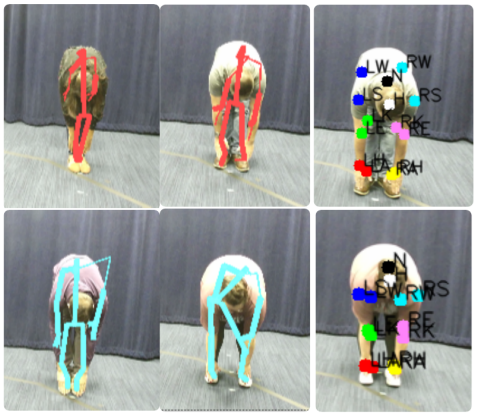

Figure 2: Example visualizations of self-occlusions in Kinect2 skeleton tracking (Left) Vs Vision-based pose estimation (Right)

Note that we use this pre-trained deep learning model and perform experiments using Caffe framework[16] for which we do not need to supply any training data from our own dataset. Now that we obtain a pose estimation method to localize full body joints (top of head, neck, shoulders, elbows, hands, waist, knees and toes) we formulate our algorithm to obtain image-wise HTKS gesture labels.

CogniLearn HTKS Recognition Algorithm

The captured RGB data is directly input to the pose tracker which outputs pixel locations for the 14 body joints. We use these body joint joint locations obtained to measure the distance between the hands and each of the the four body parts of interest (Head, shoulders, knees and toes). To obtain confidence scores for each of the four gestures (1–Head, 2–Shoulder, 3–Knees and 4–Toes) we follow these steps:

- Calculate r_1 and l_1 as euclidean distance from right hand(*RH*) to head(*H*) and distance from left hand(*LH*) to head(*H*) respectively.
- Calculate euclidean distances r_2, r_3, r_4 as distances from right hand(*RH*) co-ordinate to the three body parts shoulder(*RS*), knee(*RK*) and toe(*RA*) respectively. Similarly, calculate euclidean distances l_2, l_3, l_4 as distances from left hand(*LH*) co-ordinate to the three body parts shoulder(*LS*), knee(*LK*) and toe(*LA*) respectively.
- Calculate distances d_1, d_2, d_3 and d_4 where each d_i is average of r_i and l_i for the corresponding body parts:
- Compute scores z_1, z_2, z_3 and z_4 from the distances (where, $x \in 1$ to 4):

$$Scores(z_x) = 1/d_x$$

- Perform softmax over the scores to obtain probabilities:

$$\sigma(l)_x = \frac{e^{z_x}}{\sum_{x=1}^{4} e^{z_x}}$$

In this way we obtain probability for the given frame of the image to belong to each of the gesture classes H,T,K and S. The gesture class with the highest P(gesture/frame) $\sigma(l)_x$ is chosen as the gesture class classified for that frame.

Figure 1 describes this approach and its interaction flow with the interface. Further, the modularity of our interface would easily allow replacing the current to a pose estimator with a better one technological advances in the state-of-arts for this problem.
The gesture classes obtained from this algorithm are incorporated in the calculations for overall scores of instructions in longer sequences of HTKS steps performed. The details about calculations of these scores for the steps and sequences that we record is specified in next sections. Visualizations of this approach and gesture class obtained through this algorithm are shown in the link: `https://www.youtube.com/watch?v=w63cqZtIeIk&t=14s`. Accuracy of our system that we obtain by using this approach is shown in the system accuracy section of our paper.

COGNILEARN HTKS INTERFACE

Our interface consists of two main components: Recording and Analysis. The recording module is for use by the instructor to conduct the HTKS task and collect data. The analysis module is used to observe the data analysis provided by the system. Both these modules are built following the self-regulatory task protocol similar to that in [23] established by cognitive experts and are part of a single program using Electron: a Node JS framework used for using Node's powerful webapp tools to develop a standalone app with access tools normally reserved for the server. The modular approach of our GUI facilitates ease of using the interface for either

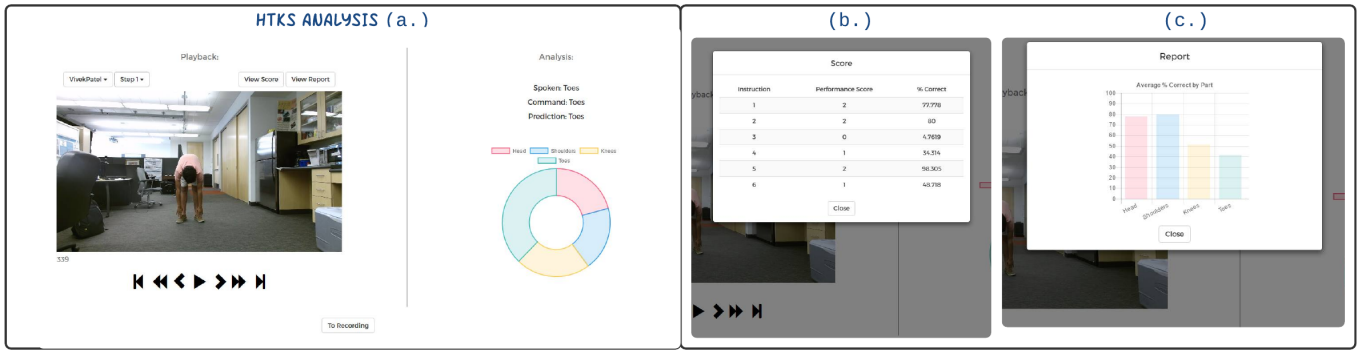

Figure 3: CogniLearn GUI Screenshot - Analysis module. a.) Screenshot that shows visualization and playback of participant's performance, it also contains prediction, instruction, command and confidence chart, b.) View Score section displays scores for instructions for the step, c.) View Report generates report of Participant's performance corresponding to body parts(H,T,K and S).

or both of these tasks of capturing and analyzing participant data. The interface is built keeping in mind good HCI practices to provide better visual features and enable ease of use to capture and deliver information that would be accessible and understandable for the user.

Recording Module

This part of the interface is built to be used by the instructor to capture participant data while they perform self-regulatory physical exercises from the HTKS task. In the recording module, the avatar in the middle shows the names currently corresponding to the body part, this way the instructor can keep track of current tracking of words in the audio to the body parts intended in the movement for the step that is currently recorded. Instructors also have the options to move between the steps that they desire to record and to cancel or restart the recording. Audio instructions are provided to the participant by the system and not by the instructor. These instructions convey to the participant which physical exercises to perform. The sequence displayed above this avatar in Figure 4 corresponds to the sequence of moves to be made. For example, 1 corresponds to head, 2 to shoulders, 3 to head, and 4 to toes, as in the first step shown. When play is pressed, instructions are given, and the correct sections are highlighted for the instructor to see.

Timestamped RGB data is collected simultaneously, which is the primary advantage of using Electron. In our interface, Electron communicates via a socket with a C# application enables collection of Kinect data with proper metadata and timing as provided by the front-end application. Once the initial step is completed, the system moves to the next step, which shuffles the body parts: the task switching. Each step selected on the left corresponds to a different set in which different body parts are swapped, and a different sequence to be followed. Instructor can move between these steps and allow recording for all instructions within the step or a subpart of it. The steps contained in our recording module and the audio instructions and actual intended motion instruction provided within those steps is provided in Table-1.

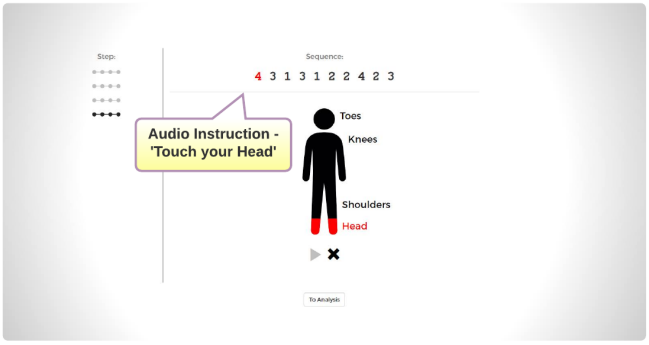

Figure 4: Recording module of HTKS Interface - This screenshot shows Step-4 which has task switching for 'Toes' and 'Head'. The comment on top of the screenshot is just to show the audio instruction given by the system.

Analysis Module

The analysis interface provides front-end visualizations of predictions obtained from our algorithm that runs at the back-end. It allows for selection of participant and provides step-wise as well as consolidated summary of participant performance in terms of speed and accuracy of movement. In the Figure 3(a.) we can see that the frame numbers are shown to the bottom left of the image, which allows for precise manual annotation, if needed. Shown to the right are three categories of text: Spoken, Command, and Prediction. Spoken refers to the word that is spoken by the system. Command refers to the part that the participant is intended to move to, when swaps are accounted for. Prediction refers to what the system thinks the participant is touching. The graph below this is a prediction confidence visualization as a doughnut chart that shows the relative certainty for each body part. In the image, knees has the highest relative certainty in the chart, so "Knees" is shown as the prediction above. The scores are given similar to protocol on paper[23] where 0- incorrect, 1- almost correct, 2- correct. The evaluation that we perform are based on timestamps of the image frames captured. The subject is supposed to perform the instructed gesture in those 3 seconds

Step #	Sequence Displayed	Audio Instruction (Spoken) "Touch your - "	Actual Intended Movement (Command)
1.	123434	H,S,K,T,K,T	H,S,K,T,K,T
2.	2432123413	S,T,K,S,H,S,K,T,H,K	K,T,S,K,H,K,S,T,H,S
3.	4321323423	T,K,S,,H,K,S,K,T,S,H	H,K,S,T,K,S,K,H,S,K
4.	1242433132	H,S,T,S,T,K,K,H,K,S	H,S,T,S,T,K,K,H,K,S

Table 1: CogniLearn Interface Steps

so based on how long the subject takes to get to the desired gesture determines the correctness. A demo-video showing an initial prototype of the proposed UI can be found here: `https://www.youtube.com/watch?v=lkSk1pCzXHM`

EXPERIMENTS

For our experiments, we collect data from 15 participants and perform analysis of participants' performances, using our CogniLearn interface. Here we present our algorithm for score calculations and demonstrate participants' scores using this algorithm. We also provide our evaluations measures that we use to calculate CogniLearn system accuracy and the our results from these calculations.

Data Collection

Using the CogniLearn interface, we capture data following steps similar to the well established self-regulatory protocol[23] from cognitive experts and provide analysis of this data to deliver valuable performance measures to our users. We use our interface for data collection from our participants while performing the HTKS tasks as instructed by the audio instructions. We collected RGB data from 15 participants (9 Male and 6 Females) of age group 18 to 30 that we recruited to follow the instructions provided by the interface and perform the task sequences. All participants were required to perform all the four steps (see Table 1) that we follow in the recording module of the HTKS interface to the best of their cognitive abilities. RGB image frames are captured while the participants perform these steps. They are stored systematically in our database along with the timestamps for each frame. Additional timestamps for start and end times of every audio instruction delivered are stored along with the image data collected. These start and end timestamps are useful for analyzing participants' performance and for score calculations. During this study and data collection, the analysis interface provides to the us valuable measures of performances of these participants such as scores for individual instructions, summarized report for overall performance per gesture class and visualizations of the steps performed. Description of the participants' performances from this dataset for our study can be found in the participant accuracy section. Our dataset and necessary annotations can be found at the provided link: `http://vlm1.uta.edu/~srujana/HTKS/CogniLearn_HTKS_Dataset.html`. This dataset comprises of over 60,000 frames of RGB data captured for the participants which creates a substantial dataset for cognitive analysis as well a baseline to perform gesture recognition for the physical exercises performed during the HTKS task. We provide annotated gesture class labels for 4443 of these frames and our annotation process and experimental details are provided in the next section that describes our system accuracy.

Score Calculations

We analyze participants performance for each instruction within steps and provide scores 0 - incorrect, 1 - almost correct and 2 - correct. For every step, the instructions are given at an interval of 3 seconds, the recording stops 2 seconds after the last instruction is provided. For each instruction, t_{begin} specifies the start of the instruction. This is the timestamp that the audio instruction for this instruction is delivered and the participant is ready to perform the movement. t_{end} for each instruction is the timestamp at which the next audio instruction is delivered. This states that the current instruction is complete and next instruction starts. That means $t_{end_i} = t_{begin_{i+1}}$ where i is the current instruction and $i+1$ is the next instruction. To make up for the time that the participant takes to reach to the body part we add 30 milliseconds to each t_{begin}. The scores for these instructions are calculated according to the metric as follows:

- All frames are capture with a timestamp associated with them. We denote the timestamp for each frame as the t_f.
- Start time (t_{begin}) and End time (t_{end}) of each instruction captured are saved along with participants' recorded videos and data and are available for score calculations.
- For number of frames T, with t_f in the range from t_{begin} to t_{end}, compare the predicted gesture class label with the instruction provided. Keep a count C as the number of frames within the instruction range that match with the instruction. Instruction score is assigned using the equation below:

$$Performance(P) = \frac{C \times 100}{T}\%$$

$$Instruction\_Score = \begin{cases} 1, & \text{if } P > 30\% \ \& \ P \leq 60\% \\ 2, & \text{if } P > 60\% \ \& \ P \leq 100\% \\ 0, & \text{Otherwise} \end{cases}$$

In our system, the scores range from 0 to 72, as a total of scores from all four steps shown in Table 1. Max attainable score for the first step is 12 and max attainable scores for the steps 2,3 and 4 are 20 for each of these steps. Figure 5 depicts scores obtained by the 15 participants from our dataset measured by the CogniLearn system. The graph here shows step-wise and cumulative scored obtained by each participant. This graph also contains confidence intervals (error bars) according to our system accuracy. Such visualizations can be

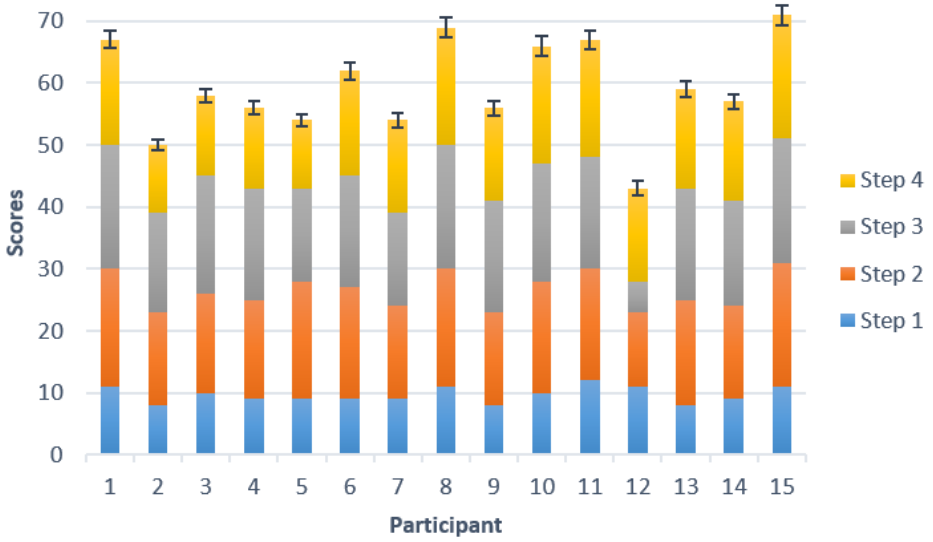

Figure 5: Participants' Performance.

useful to understand participant performance, for. e.g. in this case we can see that participants 15 and 8 performed the steps really well and attained almost the max score whereas participant 12 did well in the task with no task switching but lost most points for the third step which has task switching for 'Head' and 'Toes'.

Evaluation Measures

Accuracy for our system is measured in terms of the performance attained by our algorithm in successful classification the four gestures (Head, Shoulder, Knees and Toes) as their appropriate class. The input to our algorithm are RGB frames from the participants. The unsupervised learning problem of classification of gestures for each of these frames is solved using our computer vision and deep leaning based algorithm that performs image-wise analysis of these frames and deliver confidence scores for the four gesture classes for each frame. These image-wise predictions need to be compared with ground truth image gesture classes to obtain the accuracy measure of the performance of our system. To get these ground truth values we manually annotate of gesture class labels for the images. We annotated all frames from Step-1 recordings of all the 15 participants from our collected dataset. Step 1 consists of instructions H,S,K,T,K,T performed by these participants. Here, first 5 instructions are captured for 3 seconds each and last instruction is of 2 seconds. All recordings are at a frame rate of approximately 30 fps. These labeled gestures are compared to the predictions made by the system for each frame within Step-1 of participants' videos.

Performance is measured based on data from Step-1 recordings for all the 15 participants (9 Male and 6 Female) which gives us a total of 4,443 image frames. Our algorithm does not rely on the captured participant data for training and so all conducted experiments are completely user-independent.

System Observations

While measuring system accuracy we observed a few different cases where the system did not perform well and for which the performance could be easily improved by minor changes to the algorithm. One such failure case is demonstrated in Figure 6(a.). Here, for a human observer it is very clear that the person is touching their head but the distances d_1 which the distance from hand to head (demoted by 'a' in Figure 8a) and d_2 which the distance from hand to shoulders (demoted by 'b' in Figure 8a) are almost the same and thus in some such cases the system predicts the output as "Shoulders". This is due to the fact that our joint localization method outputs joint co-ordinate predictions for the wrist position for the participant and not for the fingers or the tip of the hand. Another reason is that the person can be touching some lower part of the head for this gesture class (Head) during which the wrist can get closer to the shoulder than to the top of the head (H). Similarly, we observe there are such offsets when a person touches their toes but the system predicts "Knees" when the wrist is closer to knees than to toes, but the particiant's finger tips are actually closer to the toes. Using these observations we improve our algorithm by making up for these offset distances due to wrist localization.

Other failure cases are when people have worn hats and this occludes their faces and/or other body joints visible in the RGB images. For these cases the system performs well for cases of instructions "Head", "Shoulder" and "Knees" as the face is visible but occasionally fails to detect correct pose for the "Toes" position. This can be improved as a future work by tracking position of the joints across consecutive frames so when a person bend and the face is no longer visible, the system should still be able to provide a measure of there the joints can be. Similarly, this kind of method would help if a

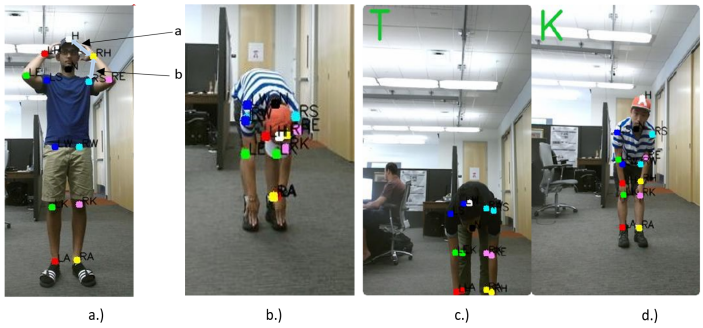

Figure 6: Visualizations of our method - a.) Demonstrates need for offset as distances of wrist from head(*a*) and shoulders(*b*) are almost similar, b.) Pose estimation failure for occlusion due to accessory (hat), c.) Correct estimation with person in background and partially cropped image (Toes not visible), d.) Correct classification of pose for with accessory when face is visible.

person is occluded or a certain body part in occluded in certain frames of the captured data.

Algorithmic Improvements

Experiments with our dataset and observations from visualization of our system performance demonstrated an opportunity for improvement of our algorithm by accounting for certain offset distances from tip of the hand to wrist. Output from our pose estimator provides us joint localizations for wrist positions and not for fingers. These offsets need to be considered while measuring distance of hand to different body parts. It is necessary to consider this while performing calculations for most cases of distance measured from hands for "Head" and "Toes" gesture classes.

We follow an approach similar to [12] where we consider the offset value according to the face height of a person. This provides a better estimate of the offset than setting a fixed value as this offset would be scale invariant. The face height f is calculated with the help of coordinates that we have available for 'top of head - h_{top}' and 'neck - n'. We use half of this face height as the offset value. This obtained offset δ is then deducted from the distances of hand from head and toes. The offset needs to be deducted from distances of both the hands and the HTKS algorithm described previously needs to be modified in the following way:

$$\delta = \frac{f}{2} \text{ Where, } f = d(h_{top}, n)$$

- Calculate d_1 based on updated r_1 and l_1 where r_1 = r_1-δ, l_1 = l_1-δ. Similarly, calculate d_2 based on updated r_2 and l_2 where r_2 = r_2-δ, l_2 = l_2-δ.

- Perform calculations for the remaining steps from the *CogniLearn HTKS recognition algorithm* using these updated values of d_1 and d_2.

These improvements help us avoid the common failure cases stated above which are cause due to localization of wrist position instead of tip of the hand.

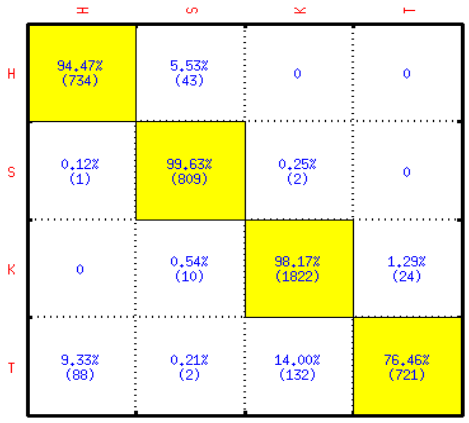

Figure 7: Confusion Matrix for system performance with the algorithmic improvements

System Accuracy

We measure the accuracy of our system by comparing predicted gesture classes for the images with the ground truth gesture class labels for our annotated 4443 image frames. These include frames from all the 15 participants and the experiments are completely user-independent. Figures - 7,8 show the confusion matrix for analyzing system accuracy with and without introducing our algorithmic improvements. Inclusion of the the improvements in the algorithm gives overall system accuracy of 92.54% over the 4443 frames which is a substantial improvement over the system accuracy obtained from algorithm without improvements which was 85.05%. We observe a substantial improvement for correct classification of "Head" gesture class which shows that half of face height works as a good offset value of this gesture class. This offset value also gave us some improvement over "Toes" gesture class where the number of correctly classified frames increased from 667 to 721. As a future work, we can work on further improving these obtained accuracy values.

Comparison to Alternatives

In building an end-to-end system that we want to build, many implementation choices need to be made. In order to help make these choices, we collected a preliminary dataset that covers all the possible subtasks within the HTKS task. There are 16 variations that can be recognized within the HTKS task according to hand movements. These 16 variations are subsets of the HTKS task where each movement begins at one of the four body parts (H,T,K or S) and ends at one of the four body parts (H,T,K or S).

We gathered a preliminary dataset (see Figure 9) by recruiting 5 subjects (3 Males and 2 Females) and recording snippets for the 16 distinct sub-tasks using *Microsoft Kinect for Windows V2*. For this initial dataset, we captured multi-modal data of three different modalities namely RGB, Depth and Skeleton tracking 2D and 3D co-ordinates. For each sub-task, snippets were recorded for approximately 5 to 7 seconds at 30 fps for

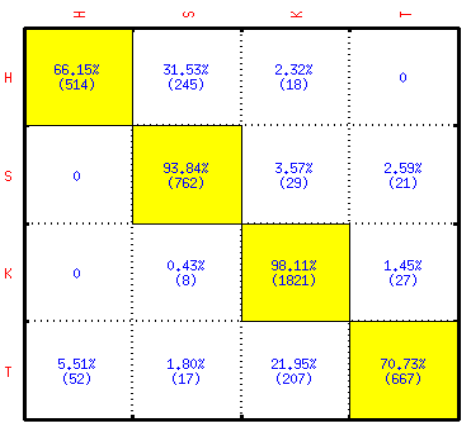

Figure 8: Confusion Matrix for system performance without the algorithmic improvements

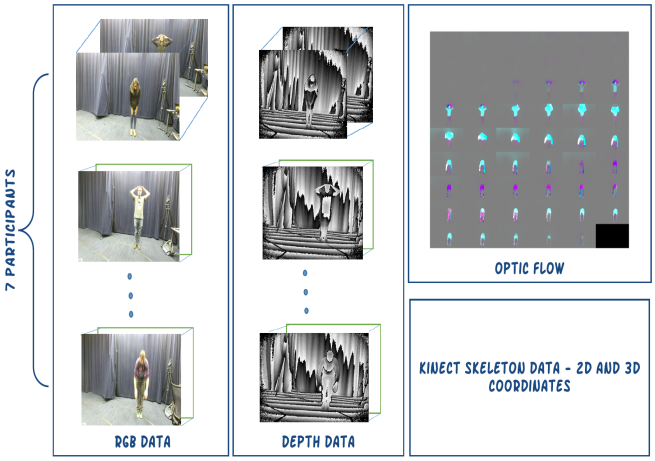

Figure 9: Multi-Modal data for preliminary experiments

all the three modalities and at three different depth positions of the participant from the Kinect V2 camera.

Activity recognition for these sub-tasks can be obtained by different approaches. The alternative approaches that we considered before deciding on our current approach for our problem was to perform holistic recognition using a Convolution Neural Network applied on raw video input to obtain frame-wise gesture labels. For both these approaches, the current and the alternative, input data is raw RGB video data and frame-by-frame analysis of the RGB images of the video is performed for classification into the following four gesture classes: 1–Head, 2–Shoulder, 3–Knees and 4–Toes. Recognition over long sequences can then be performed on the output from any of these gesture class classifiers by the algorithm shown for score calculations.
We considered static frames or also using the motion energy (See Figure 9 optic flow feature) images[4] that could serve as an addition feature for the alternative approach of using CNN model. Further, this approach can also be be improved using Long-Short Term memory(LSTM) in the same manner as proposed by [10].

Alternative Approach

Note that in our current approach there is no training set used. We used a pre-trained deep learning based pose estimator to get body part coordinates and then perform classification based on distances between the body parts(*CogniLearn HTKS algorithm*). The alternative approach that we considered to solve the gesture classification problem was to treat it as a supervised learning problem. To perform experiments according to this we divided our dataset into training, validation and test sections and annotated them image-wise with gesture class labels. The architecture we used was the CaffeNet[17] network and used the model to perform training using the Caffe[16] framework. To get better accuracy we started the network with weights from a model pre-trained on 1.2M image ILSVRC-2012 dataset[30] which is an Imagenet subset in order to obtain a powerful model and avoid overfitting. We performed transfer learning and fine-tuned this model using RGB data from our training set (80% of our initial data and consists of 2 Males and 1 Female participants) and obtained accuracy on our initial test data(10% of our initial dataset and consists of 1 Male and 1 Female participants). We perform user independent experiments on this dataset and achieved 79.98% accuracy after training this network for 500 Iterations. This network performance can be improved in several different ways such as adding more training data, using additional modalities like flow images, depth information and using LSTM as described before. Despite this accuracy obtained, there are certain drawbacks to using this approach to build our CogniLearn system.
The main drawback of using this approach is that this would require fine-tuning of our system for all the different settings in which our system could be used. For example, our system could be used with single-person or with multiple people in frame and for settings for conducting study for different user groups such as adult, elderly and/or child participants. The ultimate goal of our work is to obtain a system that is capable of being deployed in all these different settings and to be used in various environments such as in schools for assessment of cognitive behavior in children, or to be used in clinics with adults/children or elderly that need assessment and enhancement of cognitive abilities, to be used for a single participant or for group study that involves multi-person cognitive assessment. Another drawback of this approach is that it is a supervised learning problem and requires to label all the different training data if we decide to train it to learn all these different settings. This would require us to collect labeled calibration data for every setting.

CONCLUSIONS

We proposed a system that is responsible for evaluating and monitoring cognitive abilities of human subjects, based on the well known framework of Head-Toes-Knees-Shoulders. HTKS task has demonstrably been used as a cognitive evaluation tool for children and young adults in the past. We developed a system that deploys a deep learning architecture, analyzes human activity, and provides informative measures

to the experts regarding the performance of the subject on the task. The proposed framework is supported by a specifically designed user interface that can help human experts, cross-validate and/or refine their diagnosis. To evaluate our system, we created a novel dataset with 15 subjects performing 4 different variations of the HTKS task (in total more than 60,000 frames of which, 4,443 are fully annotated). We illustrate the accuracy of our method and we show detail results for each specific task.

Our immediate future plans include three very important goals. Firstly, we plan to update our dataset with more representative data, captured from real children participants in a more natural environment. Since, deep-learning is well-known for it's high invariant levels and also considering the merits that can be achieved via transfer learning, our expectations are that environmental and other external factors will not significantly affect the system accuracy. Secondly, we plan to focus more on modeling temporal dependencies for recognizing activity. Temporal modeling may provide knowledge that can have important impact on the activity evaluation step, as it may reveal patterns that we are not able to recognize at the moment. Finally, our far-reaching goal is to enrich our target exercising tasks, and incorporate more similar cognitive tests that rely on physical exercising.

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation grants IIS-1055062, IIS-1565328 and CNS-1338118. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Andriluka, M., Pishchulin, L., Gehler, P. V., and Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014* (2014), 3686–3693.
2. Asiry, O., Shen, H., and Calder, P. Extending attention span of adhd children through an eye tracker directed adaptive user interface. In *Proceedings of the ASWEC 2015 24th Australasian Software Engineering Conference*, ACM (2015), 149–152.
3. Billard, A., Robins, B., Nadel, J., and Dautenhahn, K. Building robota, a mini-humanoid robot for the rehabilitation of children with autism. *Assistive Technology 19*, 1 (2007), 37–49.
4. Brox, T., Bruhn, A., Papenberg, N., and Weickert, J. High accuracy optical flow estimation based on a theory for warping. In *Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV* (2004), 25–36.
5. Cormier, E. Attention deficit/hyperactivity disorder: a review and update. *Journal of pediatric nursing 23*, 5 (2008), 345–357.
6. Craven, M. P., and Groom, M. J. Computer games for user engagement in attention deficit hyperactivity disorder (adhd) monitoring and therapy. In *2015 International Conference on Interactive Technologies and Games*, IEEE (2015), 34–40.
7. Da Gama, A., Chaves, T., Figueiredo, L., and Teichrieb, V. Poster: improving motor rehabilitation process through a natural interaction based system using kinect sensor. In *3D User Interfaces (3DUI), 2012 IEEE Symposium on*, IEEE (2012), 145–146.
8. Dendy, C. A. Executive functionwhat is this anyway?. *Retrieved September 18* (2008), 2008.
9. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), 2625–2634.
10. Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., and Saenko, K. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015* (2015), 2625–2634.
11. Dunn, D. W., and Kronenberger, W. G. Attention-deficit/hyperactivity disorder in children and adolescents. *Neurologic clinics 21*, 4 (2003), 933–940.
12. Gattupalli, S., Ghaderi, A., and Athitsos, V. Evaluation of deep learning based pose estimation for sign language. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, ACM (2016).
13. Howcroft, J., Klejman, S., Fehlings, D., Wright, V., Zabjek, K., Andrysek, J., and Biddiss, E. Active video game play in children with cerebral palsy: potential for physical activity promotion and rehabilitation therapies. *Archives of physical medicine and rehabilitation 93*, 8 (2012), 1448–1456.
14. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., and Schiele, B. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. *CoRR abs/1605.03170* (2016).
15. Jadad, A. R., Booker, L., Gauld, M., Kakuma, R., Boyle, M., Cunningham, C. E., Kim, M., and Schachar, R. The treatment of attention-deficit hyperactivity disorder: an annotated bibliography and critical appraisal of published systematic reviews and metaanalyses. *The Canadian Journal of Psychiatry 44*, 10 (1999), 1025–1035.
16. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* (2014).

17. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R. B., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014* (2014), 675–678.
18. Lange, B., Flynn, S., Proffitt, R., Chang, C.-Y., et al. Development of an interactive game-based rehabilitation tool for dynamic balance training. *Topics in stroke rehabilitation* (2015).
19. Lange, K. W., Reichl, S., Lange, K. M., Tucha, L., and Tucha, O. The history of attention deficit hyperactivity disorder. *ADHD Attention Deficit and Hyperactivity Disorders 2*, 4 (2010), 241–255.
20. Leightley, D., Darby, J., Li, B., McPhee, J. S., and Yap, M. H. Human activity recognition for physical rehabilitation. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, IEEE (2013), 261–266.
21. Mayes, R., Bagwell, C., and Erkulwater, J. Adhd and the rise in stimulant use among children. *Harvard review of psychiatry 16*, 3 (2008), 151–166.
22. McClelland, M. M., and Cameron, C. E. Self-regulation in early childhood: Improving conceptual clarity and developing ecologically valid measures. *Child Development Perspectives 6*, 2 (2012), 136–142.
23. McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., and Pratt, M. E. Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in psychology 5* (2014).
24. Najjar, L. J. Principles of educational multimedia user interface design. *Human Factors: The journal of the human factors and ergonomics society 40*, 2 (1998), 311–323.
25. Parrillo, V. N. *Encyclopedia of social problems*. Sage Publications, 2008.
26. Pirsiavash, H., Vondrick, C., and Torralba, A. Assessing the quality of actions. In *European Conference on Computer Vision*, Springer (2014), 556–571.
27. Rizzo, A. A., Buckwalter, J. G., Bowerly, T., Van Der Zaag, C., Humphrey, L., Neumann, U., Chua, C., Kyriakakis, C., Van Rooyen, A., and Sisemore, D. The virtual classroom: a virtual reality environment for the assessment and rehabilitation of attention deficits. *CyberPsychology & Behavior 3*, 3 (2000), 483–499.
28. Ross, R. G. Psychotic and manic-like symptoms during stimulant treatment of attention deficit hyperactivity disorder. *American Journal of Psychiatry 163*, 7 (2006), 1149–1152.
29. Roy, A. K., Soni, Y., and Dubey, S. Enhancing effectiveness of motor rehabilitation using kinect motion sensing technology. In *Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS), 2013 IEEE*, IEEE (2013), 298–304.
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision 115*, 3 (2015), 211–252.
31. Sim, M. G., Khong, E., Hulse, G., et al. When the child with adhd grows up. *Australian family physician 33*, 8 (2004), 615.
32. Sucar, L. E., Orihuela-Espina, F., Velazquez, R. L., Reinkensmeyer, D. J., Leder, R., and Hernández-Franco, J. Gesture therapy: An upper limb virtual reality-based motor rehabilitation platform. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 22*, 3 (2014), 634–643.
33. Tsiakas, K., Papakostas, M., Chebaa, B., Ebert, D., Karkaletsis, V., and Makedon, F. An interactive learning and adaptation framework for adaptive robot assisted therapy.
34. Velloso, E., Bulling, A., Gellersen, H., Ugulino, W., and Fuks, H. Qualitative activity recognition of weight lifting exercises. In *Proceedings of the 4th Augmented Human International Conference*, ACM (2013), 116–123.
35. Wexler, B., and Bell, M. ACTIVATE Program. `http://www.c8sciences.com/`.
36. Xia, L., and Aggarwal, J. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), 2834–2841.
37. Zhou, L., Liu, Z., Leung, H., and Shum, H. P. H. Posture reconstruction using kinect with a probabilistic model. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, VRST '14 (2014), 117–125.