AN INTELLIGENT FRAMEWORK TO ASSESS EMBODIED COGNITION FROM PHYSICAL ACTIVITIES IN CHILDREN

by

ASHWIN RAMESH BABU

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington,
in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
May 2021

To my family and my beloved friends for being always by my side.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincerest gratitude to my supervisor, Prof. Fillia Makedon, for believing in me and for her continuous and unconditional support in my research. Her guidance has been extremely valuable towards completing my Ph.D.

Next, I would like to thank my collaborators and committee members, Dr. Vassilis Athitsos (Professor at UT Arlington), Dr. Morris Bell (Professor Emeritus, Yale School of Medicine), Dr. Ramez Elmasri, and Dr. Farhad Kamangar, for providing valuable feedback during the course of my Ph.D.

Furthermore, I would like to thank my colleagues from Heracleia - Human Centered Computing Lab and Vision Learning and Mining lab for the numerous discussions and collaborations we have had over the years. They all played a unique role in the successful completion of this thesis.

Finally, I would like to thank my parents, family, and friends for being the greatest source of emotional support and encouragement I could ever ask for.

March 25, 2021

ABSTRACT

AN INTELLIGENT FRAMEWORK TO ASSESS EMBODIED COGNITION FROM PHYSICAL ACTIVITIES IN CHILDREN

Ashwin Ramesh Babu, Ph.D.

The University of Texas at Arlington, 2021

Supervising Professor: Prof. Fillia Makedon

Cognition refers to *"The mental actions or process of acquiring knowledge and understanding through thought, experience, and the senses"*. It encompasses many aspects of intellectual functions and processes such as attention, working memory, response inhibition, motor functions and more. Humans start to develop these cognitive skills right from their childhood and become fully developed through their adulthood.

Impairments in these cognitive functions, specifically in Executive Functions (Higher-order cognitive functions), disrupt their everyday life leading to a troubled childhood and lifelong difficulties in family, employment, and community functioning leading to socio-economic repercussions. Identifying such impairments at the right age (early childhood) provides the best opportunities for remedial intervention, as brain plasticity is highest in children and diminishes with age. Attention Deficiency Hyperactivity Disorder (ADHD) is one of the common psychiatric neuro-developmental disorders that often could cause cognitive impairments, specifically with executive abilities/functions. They are commonly found in children and young adolescents, starting at the age of 6, and occur three times more frequently in boys than in girls.

There is a need for assessments to estimate the level of cognitive development so that proper intervention can be offered when problems with executive functions arise.

The main aim of this research is to develop an automated and non-intrusive system to measure the level of cognitive development in children (e.g., early, middle, full development) with various cognitive tasks assessing different cognitive skills. The Activate Test of Embodied Cognition (ATEC) is an assessment test designed to measure executive functions in children through physically and cognitively demanding tasks and provides measurements for attention, working memory, response inhibition, self-regulation, rhythm, and coordination as well as motor speed and balance. The proposed tool takes advantage of state-of-the-art knowledge from both the fields of Artificial Intelligence and Cognitive Sciences to provide accurate measures of cognitive development. The tool aims to assist therapists in decision-making by providing performance metrics regarding the subject's performance. This work also advances computational methods for human action recognition to provide automatic measurements of various metrics of performance. These metrics are related to generic motion features as well as metrics explicitly defined by cognitive experts.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

TECHNOLOGY AS A TOOL FOR COGNITIVE ASSESSMENT

1.1 Introduction

The term ***Cognition*** can be defined as mental actions or process of acquiring knowledge and understanding through thought, experience, and senses. It comprises several intellectual functionalities such as attention, engagement, long and short-term memory, cognitive flexibility and task-switching ability, problem-solving and decision-making skills, and many others. These cognitive processes have their impact on every aspect of life, from school to work to relationships. When individuals have difficulty in successfully performing one or more of the above cognitive functionalities, they are considered to have deficits/impairments in their cognitive functionalities. Deficits or impairments in these cognitive functions are not associated with any particular condition but could be one of the manifestations of multiple underlying conditions. Some of the causes of cognitive deficits in childhood include Autism Spectrum Disorder (ASD), Attention-deficit/hyperactivity disorder (ADHD), etc. At the same time, for adults, a decline in their cognitive performance can be seen as their age progresses. Additionally, certain conditions such as Alzheimer's disease, Multiple Sclerosis, Parkinson's disease, etc., could cause cognitive and motor impairments in adults as well [14]. It is necessary for healthcare professionals to identify these cognitive impairments in their early stages and help patients with the proper intervention.

As part of identifying these cognitive deficits, cognitive assessments and rehabilitation become a significant part of the process. These assessments can diagnose major cognitive impairments and can provide more detailed information on the deficits in the functionalities. In addition, work has shown that these cognitive assessments help in improving their cognitive functionalities [15]. Assessing and observing these cognitive functionalities demand long-term assessment and monitoring of individuals with different sets of tasks to assess various cognitive behaviors. Most of the time, these assessments take place in a controlled environment. NIH toolbox[16] and the psy-toolkit[17] library are some of the very popular cognitive assessment tool-kits that comprise of multiple tests aiming to assess various cognitive functionalities.

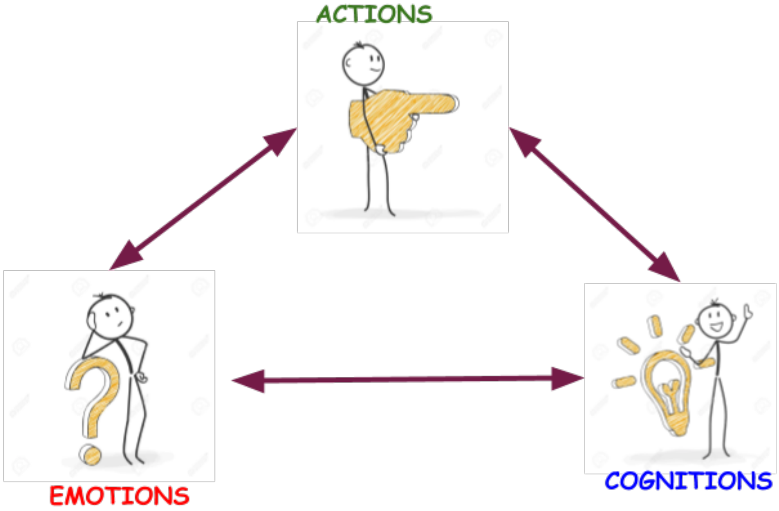


Figure 1.1. Relation between Actions, Cognition, and Emotions in Human Behavior. Actions can trigger emotions and thoughts, while at the same time, they can be the result of our feelings and cognition. Their highly dependent relation and interaction enable us to perceive the world around us and respond to the different stimuli of our surroundings. [18].

With the improvement in technology, researchers are focusing on interpreting such functionalities in the wild, which is still a challenging problem in the field of behavioral and cognitive science [19]. Artificial Intelligence (AI) has been used widely to understand and analyze human behaviors in the environments in which they interact, as it is one of the important indicators of their cognition. **As humans, we can consider ourselves as active agents that are continuously interacting with their environment, producing and perceiving countless information at any given moment. A non-stop process that eventually affects drastically our bodily needs, our reactions and our mental desires** (Figure 1.1) [18]. Hence, human behavior is one of the important indicators of cognition in individuals.

1.2 Technology to Understand Human Behavior

Given a controlled environment with a defined set of possibilities, current technologies can provide good results in understanding the behaviors of the individuals such that this information could be used to extract useful cognitive information. This could further assist the therapists in estimating their cognitive state and thus provide useful insights on possible cognitive impairments [20]. For example, extracting the behavioral information could help workers in the assembly line improve productivity, monitoring patients during rehabilitation, and assessments towards detecting various cognitive impairments.

1.2.1 Understanding Physical Actions

Human Action Recognition (HAR) is one of the popular approaches to understand individuals' behavior. Some of the commonly used approaches include using cameras (RGB and Depth), wearable sensors such as accelerometers, gyroscope, etc., towards recognizing actions performed by individuals. In recent times, methods that are images and video based are preferred due to their non-invasive nature and their ability to track multiple people in the scene. This is achievable with improvement in the machine and deep learning techniques. In the past, researchers have worked on extracting features that are oriented to actions to understand the behaviors [21, 22, 23]. But in recent times, deep learning has revolutionized the area of human action recognition which are capable of modeling the behaviors by automating the complete recognition process from the input source (RGB video frames, depth frames) to the action classes [24, 25, 26, 27]. Additionally, depth information can be considered as an enhanced vision-based device since it can additionally provide depth data that can facilitate the detection of human movements. This helps to implement crucial processes such as the extraction of a human silhouette, reducing the dependencies of shadows, light reflection, and color similarity [28, 29].

On the other hand, action recognition based on wearable sensors provides more accurate representation of the user's motions and activity as they directly contact the user's body. This approach also facilitates avoiding other noise in the scene, such as objects, the place or recording, etc. Work in this area has provided a variety of features that can be extracted from these sensors that can describe what action is being performed. Although the extracted features are generally dependent on the type of sensors, standard statistical features from both time and spectral domain are sufficient to distinguish different activities that are performed. [30, 31, 32, 33, 34].

1.2.2 Understanding User Emotions

Emotions are one of the important indicators of cognition, and it has been researched lately in the field of behavioral analytics. As mentioned in Figure 1.1, actions, emotion and cognition are highly interconnected. Cognition can trigger emotions, and emotions can trigger actions. Hence, recognizing emotions could provide information about the cognitive state of the user. Still, recognizing emotions is a challenging problem due to the great variability that is observed across different subjects when expressing the same emotions. In recent times, many approaches have been proposed for emotion detection, with the most popular one from the audio sentiment analysis [35] extracting and video data. Specifically from their facial expressions and body postures [36, 37, 38, 39]. Some of the other commonly used methods include analyzing fEMG [40], monitoring arousal using ECG [41], galvanic skin response (GSR) [42], respiration sensors [43] or EEG based approaches [44].

1.3 Motivation and Thesis Outline

1.3.1 Motivation

The main motivation of this research is to develop an automated and non-intrusive system that is capable of measuring the level of cognitive development in children with various novel cognitive tasks that are both physically and cognitively demanding. Such systems are required to be low-cost and easy to use by medical experts. To the best of our knowledge there is no such tool that can measure various cognitive skills from physical activities that are intended to assess various neurological conditions.

1.3.2 Thesis Outline

This section discussed the overview of cognition and how cognitive assessment evolved through the years. We discussed some of the common challenges in identifying deficits in cognitive functionalities and how users' behavioral information can be used to solve this problem. Further, an extensive review of some of the existing methods and approaches to learn user behavior through various technologies was presented in this chapter.

In the following chapters, we will look into how the mentioned technology has been used in multiple scenarios, such as in the workplace and at schools to measure cognitive functioning and deficits in these cognitive functionalities. Chapter 2 we explore the need for cognitive assessment and training in the workplace. As part of this work, we devised an elaborate study on cognitive training in the workplace using the Towers of Hanoi (TOH) task and the different training approaches that could provide the most benefit in the training process. Conclusions were based on how well the participants performed the task and user study as well. Further, the need to understand the users' emotional state while performing cognitive training was identified. Other than just the metrics from the task, emotional information is extracted from the user from their body postures, facial expressions, and from their EEG signals which were recorded through a non-invasive headband to predict the emotional state of the user from which the task outcome was predicted. This information was used to build an adaptive system.

Subsequently, Chapter 3 presents the importance of cognitive assessment for children and the need for such assessment and evaluation right from the young age. An extensive study is presented on cognitive functions and their association with various neurological conditions, such as ADHD and Autism. Further, this chapter discusses about some of the popular and commonly used approaches to evaluate deficits in various cognitive functionalities and discusses their current drawbacks. One central research question that arises is the association between cognitive functionalities and physical movements. Hence, in Chapter 4, the ATEC system is introduced to evaluate various cognitive functionalities through tasks that are both physically and cognitively demanding. As part of this work, a set of tasks and an automated system were developed that can automatically capture the movements performed by the children and automatically score the correctness of the task performed. Chapter 5 uses state-of-the-art computer vision techniques to evaluate the finger opposition task that has been proved to identify deficits in motor functions. Further, Chapter 6 introduces some of the popular computer vision-based motion analysis techniques that have been used to evaluate cognitive functionalities such as attention and response inhibition through a novel "Ball drop to the beat" and "Tandem Gait" tasks. Finally, this work discusses some of the recent works in self-supervised techniques that take advantage of the unlabelled data to build effective representation, which could further be used towards human action recognition. We extend some of the new approaches on images to videos and present their results on our dataset. Finally, we summarize our findings, highlight the takeaways of this research and provide suggestions for future directions in the area of understanding cognition from physical movements.

CHAPTER 2

COGNITIVE ASSESSMENT AND TRAINING IN WORKPLACE

2.1 Introduction

In Chapter 1 we discussed the individual components that compose human behavior, namely: actions, thoughts, and emotions. We highlighted the dependency that is observed between the actions and emotions and reviewed recent work in human behavioral analysis with technologies such as motion capture, sensors, etc. This chapter discusses how cognitive training and assessment have the potential to help employees in the workplace [45]. Additionally, multiple training methodologies and the impact of those methods on employees were investigated. Further, the behavioral and physiological information of the participants, such as their facial expressions, body postures, and EEG signals, were extracted to understand the emotional state of the participant during the cognitive training and assessment. The correlation between emotion and cognition was studied, which was later used towards personalizing the assessment task based on the individual's cognitive state.

2.2 Need for Cognitive Assessment in Workplace

Cognitive assessment and cognitive rehabilitation in the workplace are becoming very popular/necessary in recent times, especially for workers who are involved with jobs requiring high precision and concentration. Generally, cognitive training is predictive of job performance across all domains. Employees with higher cognitive ability tend to adjust better to new tasks with their ability to learn and apply new information. Work has proved that this cognitive training and frequent assessment have sustainably improved executive functions in middle-aged industry workers [46]; hence, companies that are involved in manufacturing are enforcing their employees to take such tests in frequent intervals as the industries are making their environment smarter where humans and machines collaborate in achieving their goals. With the increase in implementation of technical systems to work with humans comes the complexity for humans' to maintain a proper overview. Usually, new employees undergo extensive training with existing experts tutoring them. These trainings are common to all employees to provide the required knowledge on how to handle the system but do not consider the mental/physical health, history, etc. These cognitive rehabilitation techniques not only help workers towards their roles but also provide assistance for workers with any kind of cognitive impairments [47]. Also, cognitive/mental health decline over aging [48]. With the improvement in Augmented Reality and Virtual Reality, people have started to use them in various training scenarios as it helps to simulate the environment [49, 50] and computer game-based training is one of the commonly used methods for cognitive training [51, 52, 53, 54, 55].

2.3 Towers of Hanoi as a Tool for Cognitive Training and Assessment

Towers of Hanoi(TOH) is a well-known executive function task that is used to assess cognitive skills, such as working memory, procedural learning, problem-solving, and inhibition process [56, 57, 58]. This work utilizes TOH tasks to simulate cognitive training for employees at the workplace. In addition, this work compares three different approaches of training to find the most effective training method that involves procedural learning and problem-solving. The three training approaches used are personal trainer, computer-based training, and game-based training. To evaluate the effectiveness of the task and the training method, different evaluation metrics such as time taken to solve the complete task, the average time taken for every move, the total number of moves made, errors made during task, etc.

2.3.1 Experimental Setup

The experimental setup consists of a webcam, a computer, a physical TOH as represented in Figure 2.1. A computer game replicating the physical experimental setup was developed using Unity game engine. The number of disks in the TOH game was set to 5 as we found that it was neither too difficult nor easy. The minimum number of steps taken to solve a TOH task is represented as $2^n - 1$, where n represents the number of disks. In our case, the minimum number of steps is 31. The TOH were in a stationary position facing the webcam with the pegs labeled as column 1, column 2, and column 3.

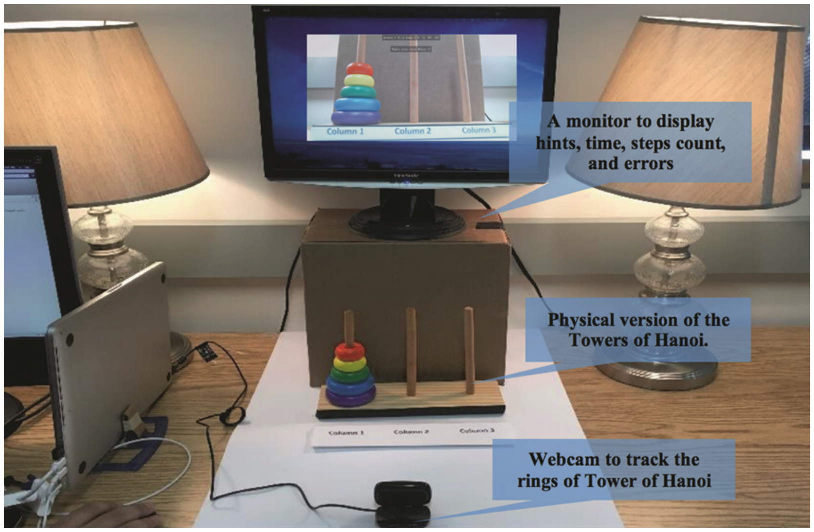

Column 1
Column 2
Column 3
A monitor to display hints, time, steps count, and errors
Physical version of the Towers of Hanoi.
Column 1
Column 2
Column 3
Webcam to track the rings of Tower of Hanoi

Figure 2.1. Experimental setup for cognitive training and assessment with Towers of Hanoi(TOH) task.

2.3.2 Towers of Hanoi Rules

By default, there are some standard rules to solve the Tower of Hanoi problem. No additional rules were added to make the game easier for participants. These include,

- Move only one disk at a time.
- All disks, except the one being moved, must be on a tower.
- User will use only one hand to deal with the disk.

2.3.3 Experiment

The participants were undergraduate students in the Department of Computer Science and Engineering at the University of Texas at Arlington. There were no restrictions based on their age and gender. The age of the participants ranged between 18 to 30. A total of 30 participants were divided randomly for each of the three training methods, and they did not have prior knowledge about the TOH task. The experiment was divided into training phase and testing phase.

2.3.3.1 Training Phase

During the training phase, the participants were provided with the rules of the TOH. Each participant was presented with one of the training methods, which were randomly assigned prior to the study.

Human Trainer

With this method, the participants were trained with a personal human trainer as shown in Figure 2.2. The trainer went through the steps verbally to solve the TOH with the participants. This training was timed, and the number of errors while solving was recorded.

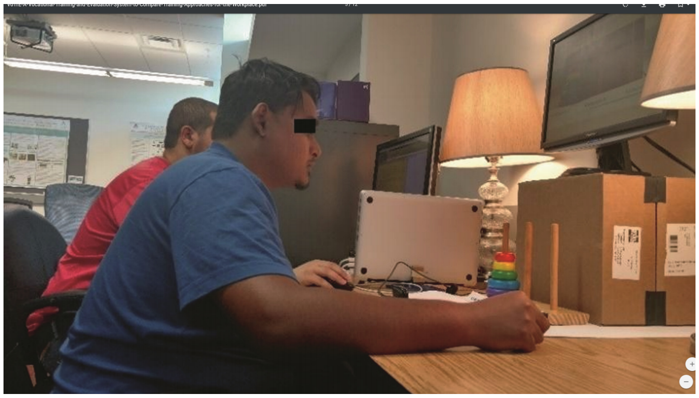


Figure 2.2. Human Trainer. Human trainer(left) gives verbal instructions to the participant(right) to solve the TOH task.

Game Based Training As mentioned before, a game simulation of the TOH task was made with Unity game engine. The participants solved the game using mouse clicks on the disks. Instructions were given in the form of text on the screen. An example instruction will look like, ”Move the red disk to column 3”. To notify wrong moves, error messages were displayed on the screen along with sound notifications. The setup is represented in Figure 2.3.

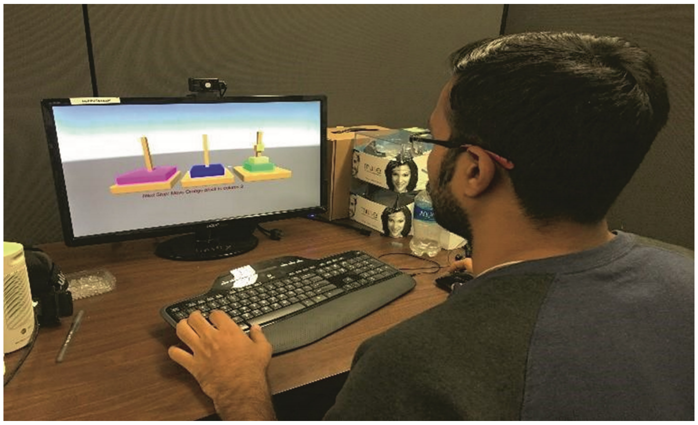


Figure 2.3. Game-based training. Participant plays the unity game.

Computer Based Training

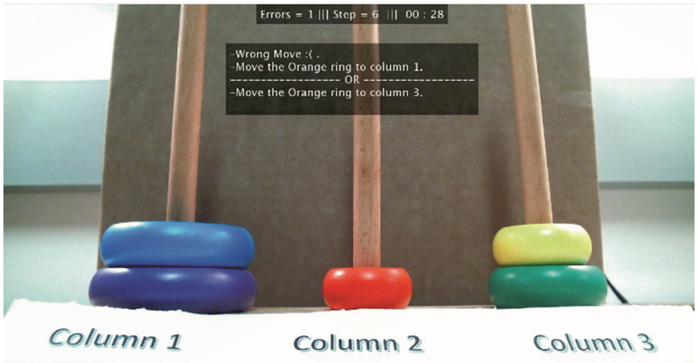
In this method, the participants were trained with computer-aided instructions. That is, instead of an individual trainer, participants were asked to solve TOH with instructions flashing on the screen. This is similar to the Game-based training, but this will have a physical TOH setup rather than a game.

The system was implemented using MATLAB. A webcam placed in front of the TOH capturing ten frames per second, recognizing the disks and their positions. To identify individual colors, HSV (Hue, Saturation, and Value) color space was used. The size of the disks was considered to avoid shadow based errors, noise, and other objects that impeded the frame. Next, to identify the position of each disk, the centroid of each disk was calculated, and the tower to which it belonged was identified. The system considered the disks for evaluation only when they were in a stationary position. Each step was considered a separate state. Every time a change is made, the system compared the current position of the disks with the position of the expected state. If they matched, it was considered a successful move, and the current and the past states were updated. If they did not match, it meant that the disks were not in the expected position and were considered an error. In such cases, the system asked the participants to go to the previous move or to the actual move, as shown in Figure 2.4, and then the system proceeded.

2.3.3.2 Test Phase

Once the training was completed, all participants were asked to complete the task(without any assistance) using the physical TOH as shown in Figure 2.5. The results of this phase helped us evaluate the functional capacity of the participants and the effectiveness of the training.

The system kept track of the past and of current state. Initially, when the test began, there was no past state, and the starting position was updated as the current

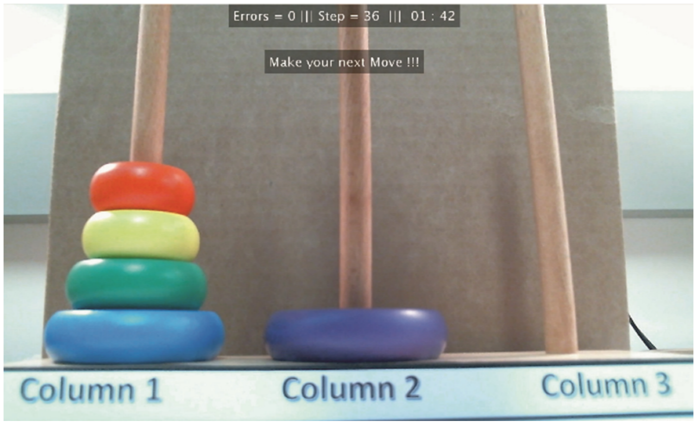


```
Errors = 1 ||| Step = 6  |||  00 : 28

-Wrong Move :( .
-Move the Orange ring to column 1.
-------------------- OR --------------------
-Move the Orange ring to column 3.
```

Figure 2.4. Computer based training. The system asks the participant to go to the previous move or to the actual move to be made when they fail to follow the instructions.

state. For every move, the system checked the position of the disks and compared it with the rules. If the position of the rings satisfies the rules, the system updates the current position as the new present state and updates the past state.



```
Errors = 0 ||| Step = 36  |||  01 : 42

Make your next Move !!!
```

Figure 2.5. Test phase. The system does not provide the participant any instructions.

2.3.4 Data Analysis

The criteria on which the analysis was made include the total number of moves, total time taken to solve TOH, the time taken for each step, and the number of errors made during the training and testing phase. Figure 2.6 shows the average number of steps each group of participants needed to complete the TOH. All participants were trained to finish the task in 31 moves, and any extra moves were considered as an error during testing. The average number of moves in the testing phase was nearly the same in all groups, and the participants performed extra moves in the testing phase compared to the training phase.

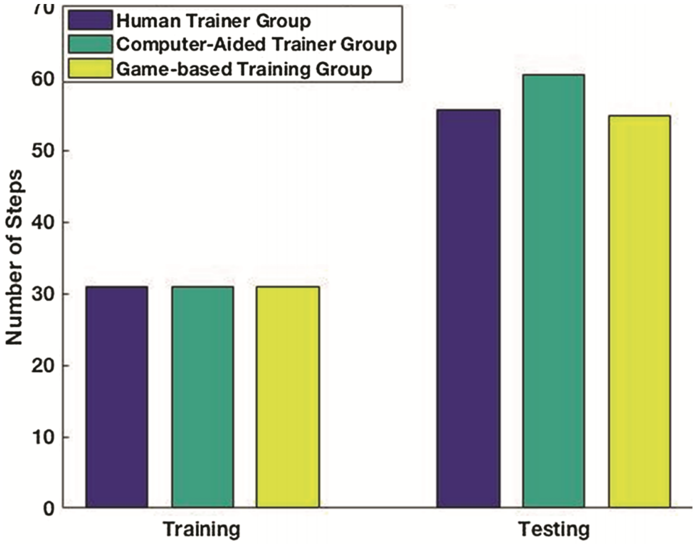


Figure 2.6. Average number of moves(steps) each group of participants performed in both the training and testing phase.

Figure 2.7 shows that the participants took greater time to complete the testing phase compared to the training phase. It also shows that the participants took less time to complete the game-based task in the training phase. The reason may be explained by the fact that the participants did not interact with the physical TOH in the Game-based Training, which resulted in less physical effort and time. However, the completion time in the testing phase was very similar in all three groups. The results of the number of moves and the completion times might indicate that the different training approaches did not have a major effect on how the trainees performed. It also indicates that performing the training twice had a very low practice effect since the participants took longer time and more moves to complete the testing phase compared to the training phase.

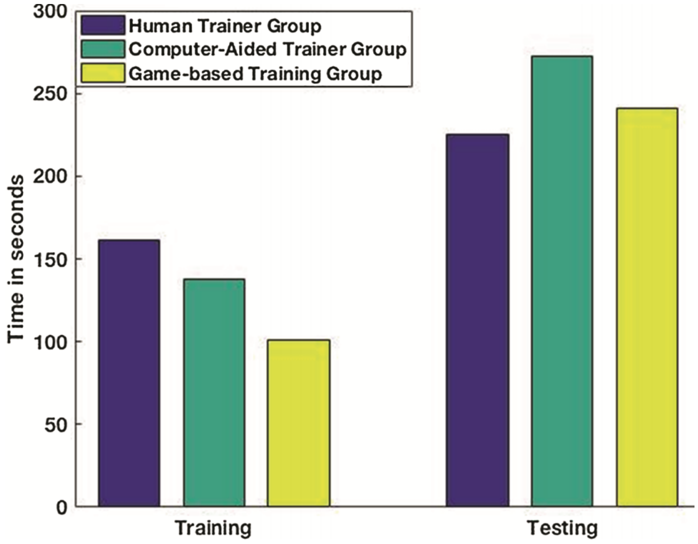


Figure 2.7. Average time to finish the TOH task by each group of participants in both the training and testing phase.

Figure 2.8 shows the average number of errors performed by each group. In the training phase, both extra and illegal moves (i.e., large disk placed on smaller disk) are considered errors, and in the testing phase, only illegal moves are considered errors. The participants in the human training group performed with few errors when compared with the older groups. This may indicate, although the performance is very similar in all the three groups, the participants who were given the rules for the task by the human trainer could follow the rules better. In the Game-based training group, the game restricted the participants from making illegal moves. When the participants tried to make illegal moves, the disks stayed in the original towers/columns and did not move to the wrong columns. The lack of hands-on experience during game-based training may have contributed to the increased number of errors in the testing phase.

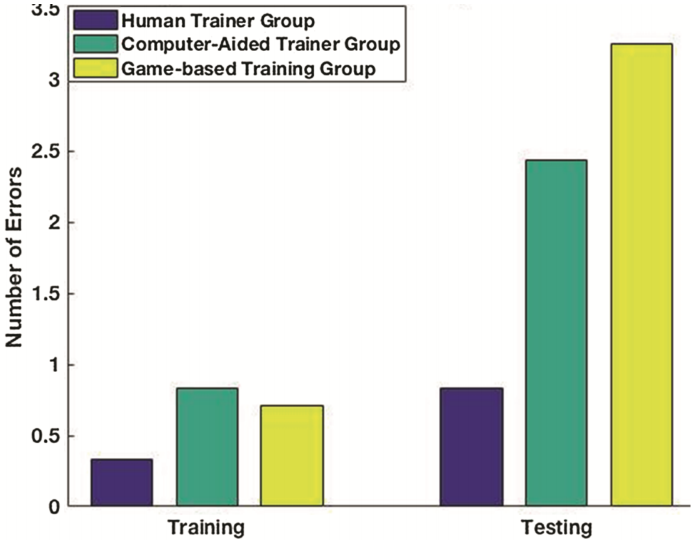


Figure 2.8. Average number of wrong/illegal moves(steps) each group of participants performed in both the training and testing phase.

2.3.5 User Survey Results

At the end of every experiment, participants were asked to fill out a survey form about their experience with the system. From the survey, more than 90 percent of the participants liked the experiment as they were new to the TOH. They also stated that they required complete focus and concentration while solving the TOH with minimum steps and errors. When the participants were asked how they liked their training method, computer-aided training group had the highest rating, followed by the human trainer training group, and then the Game-based training (GBT) group. Contrary to the above statement, when the participants were asked to rate how much

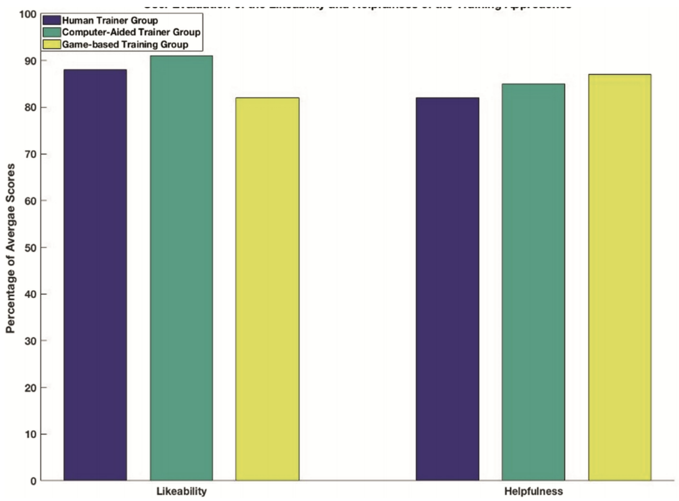


Figure 2.9. User survey result for the likeability and helpfulness of the training approaches.

their training method helped to complete the task, the highest ratings were received from the GBT group, whereas the human trainer group had the lowest rating as shown in Figure 2.9. Specifically, one of the trainees in the human trainer group commented that ‘listening to a human trainer is helpful, but it does not help me think on my own.’

2.3.6 Inference from Participants’ Performance

One of the advantages of the Computer-based training and testing is that the system provides feedback based on user performance while the user gets the actual experience of the TOH task. The data that is collected during the testing phase can be used to measure cognition for various disorders and can be tracked over time with slight modifications to the came, such as changing the number of disks to track improvements. However, understanding the user’s cognition cannot be achieved with just the metrics from the task. It requires more information, such as the user’s emotional state, which impacts their behavior. The following section describes how user's behavior can be understood using data from multiple external sensors.

2.4 Multi-Modal Data for Cognitive Assessment

Designing a Multi-modal system is not a trivial task given the possibility of multiple outcomes and their dependence on multiple input combinations. Depending on the application and the data that is being considered, the weights for the modalities differ. Work proposed by Huang et al. [1] explains data fusion as the process of joining data from multiple modalities with the aim of extracting complementary and more complete information for better prediction. The authors explain three main data fusion strategies, namely, early, joint and late fusion.

Early fusion is commonly known as feature level fusion points to combining multiple input modalities into a single feature representation before feeding into a machine learning model. These modalities can be fused in a variety of ways, including concatenation, pooling, and using a gated unit. Fusing the original features represents early fusion type 1 represented in Figure 2.10 (Left) while fusing the extracted features from another neural network represents early fusion type 2.

Joint Fusion is the process of joining learned feature representations from intermediate layers of neural networks with features from other modalities as input to the final model. The difference in this approach compared to the early fusion is that the loss is propagated back to the features extracting Neural Network, which can also be termed as end-to-end learning. This approach is represented in Figure 2.10 (Middle).

Late Fusion approach represented in Figure 2.10 (Right), refers to the process of leveraging predictions from multiple models to make a final decision, which is why it is often known as decision-level fusion. Typically, different modalities are used to

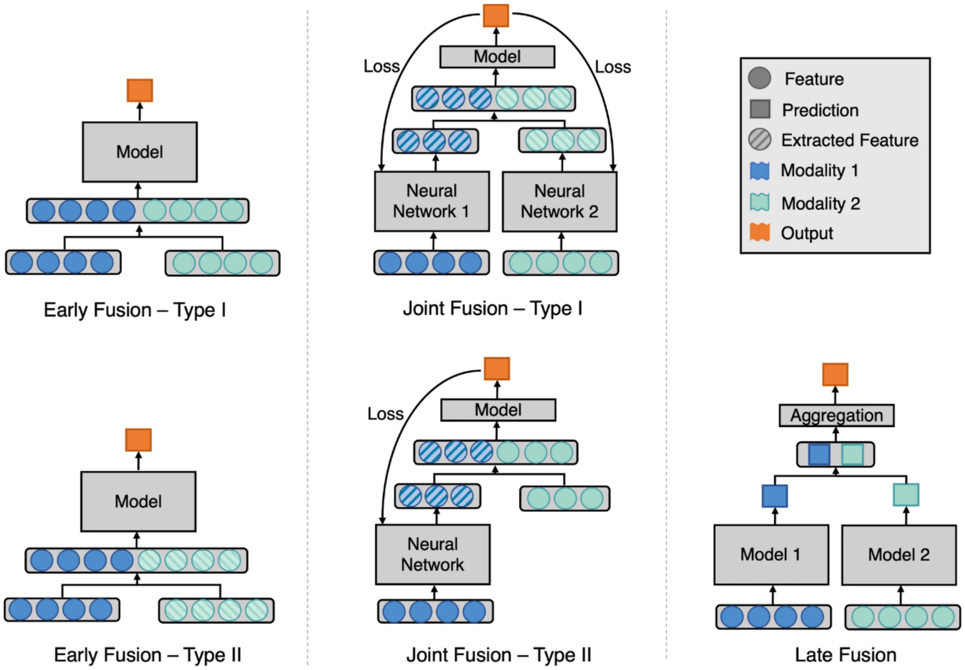


Figure 2.10. Model architecture for different fusion strategies. Early fusion (left figure) concatenates original or extracted features at the input level. Joint fusion (middle figure) also joins features at the input level, but the loss is propagated back to the feature extracting model. late fusion (right figure) aggregates predictions at the decision level [1].

train separate models, and the final decision is made using an aggregation function to combine the predictions of multiple models. Some examples of aggregation functions include: averaging, majority voting, weighted voting, or meta-classifier based on the predictions from each model. The choice of the aggregation function is usually empirical, and it varies depending on the application and input modalities.

2.4.1 Multi-Modal User Monitoring for Cognitive Assessment and Rehabilitation

Based on the mentioned Multi-modal data fusion theories, an intelligent system for cognitive assessment and rehabilitation is proposed, which not only assesses the participant based on the scores from the task but also monitoring his behavior such as emotions for better understanding. In this proposed approach, a late fusion strategy has been used where the decision from multiple modalities are combined to make the final decision. The experimental approach was based on the theory that Actions can impact cognition which in turn will be reflected through emotion [59]. With the cognitive task, we increase the difficulty level of the task so as to induce stress which is reflected in their emotion which in turn triggers both physiological and behavioral signals [60, 61]. Positive or negative emotions triggered while performing these tasks affect the outcome of the task as they consume their cognitive resources [62, 63, 64].

The proposed system uses a combination of non-invasive sensors such as electroencephalography(EEG) headband and image sensors to capture participants' brain wave patterns, body postures, facial expressions to analyze their emotions and stress. The proposed system uses Machine Learning and Computer Vision techniques to automatically analyze the data recorded and predicting the user's stress. We also built a GUI to visualize the signals that were recorded.

2.4.2 Sequence Learning Task for Cognitive Assessment

This is part of the work published by [38] where we used the sequence learning(SL) task as an assessment tool. The SL task is recognized as an important tool for assessing cognitive load and its relation to training by therapists and performance experts [65, 66, 67]. The SL task involves listening or seeing a set of character sequences and hearing able to repeat them correctly in a certain amount of time. The sequence could be delivered via speech or image on a computer screen. Performance outcomes from SL task can help therapists and other experts to determine what particular treatment or rehabilitation an individual might need to enhance his/her performance in a given domain or application.

For the task setup, the user had three buttons with labels (”A”,”B”,”C”). The NAO robot dictated the alphabetical sequence of either 5, 7, 9 characters in length. Each comprised of only the above-mentioned characters. The user was expected to reproduce the sequence dictated by the NAO robot by pressing the button within a time duration. The NAO robot verified the sequence reproduced by the user and recorded the outcome. The outcome of this task is a binary data where ’0’ represents failure, and ’1’ represents success. To complete the task, the participant had to reproduce 12 such sequences. The task was built in such a way that the complexity of the task increases gradually to induce stress. Figure 2.11 represents the task setup for the experiment.

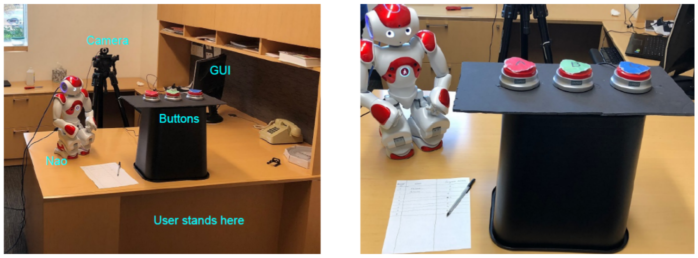


Figure 2.11. Sequence Learning Task Setup.

2.4.3 Experimental Setup

For the experiment, we used a socially assistive humanoid robot, NAO, to instruct and monitor the user. An image sensor and an EEG headband, MUSE was used for the experiment. Figure 2.12 represents the experimental setup of the assessment. Data is collected from two sources, the MUSE headband that collects EEG signals of the user and the image sensor, which records the person performing the task. The image sensors monitor both the facial expressions of the participant and the body postures of the user to detect stress.

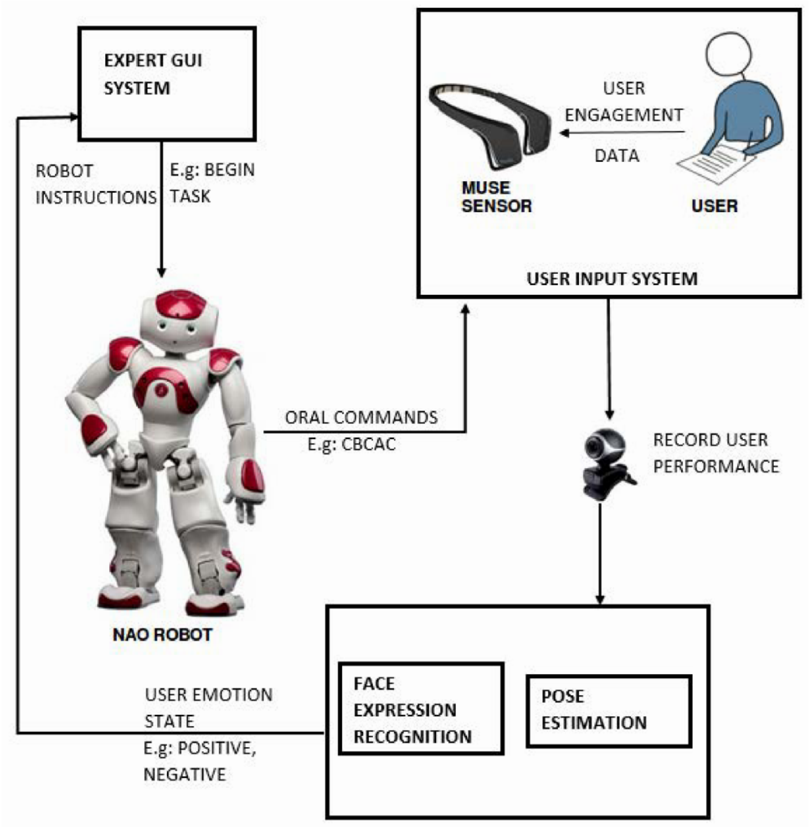


Figure 2.12. The proposed system for cognitive assessment system with multi-modal data.

2.4.4 Data Collection

For the data collection, 15 graduate students were recruited. They were in the age group of 22 to 35. Each user completed one full session (12 sequences) of the SL task. The MUSE headband was used to collect the EEG signals from each participant. The MUSE headband consists of a total of 7 sensors, two forehead sensors, two sensors near the ears, and three reference sensors. The signals are generated at a sampling rate of 220Hz and provide access to the raw EEG signals. The frequency bands provided by the device are $\alpha(9to13Hz)$, $\beta(12to30Hz)$, $\gamma(30to50Hz)$, $\theta(5to8Hz)$ and $\delta(1to4Hz)$. The data were separately stored during the listening phase and performance phase. Research shows that information listening is the key element in the proposed task.

Hence, to predict the user performance, only the data collected during the listening phase was considered. For the first and second modalities, which are the emotion recognition system from the body pose and the facial expression, a camera was setup in front of the user. The camera collects RGB images at a rate of 30 frames per second, and they were also collected separately for the listening and performance phases throughout the session. Data from the image sensors were used to predict the emotions of a user using two separate modalities: facial expressions and body postures.

2.4.5 Emotion Recognition with Image Sensor

The data from the image sensors were used to predict the emotions of a user using two separate modalities: facial expression and body posture. From the detected emotions, task performance was predicted. The input images collected were RGB images in the format ($Format : Width$ x $Height$ x $Channels$).

2.4.5.1 Emotion recognition from Facial Expression

To predict emotions from facial expressions, a Convolutional Neural Network-based architecture was used. The architecture was inspired by Arriaga et al. [2]. The model was trained with multiple publically available datasets to recognize stress and emotion from facial expression. The model consists of 4 residual depth-wise separable convolutional operations. Each convolution is followed by a batch normalization operation and a ReLU activation function. The last layer of the neural network consists of a global average pooling and a softmax operation. Adam optimizer was used for optimization. The complete sequence consists of a face detection module and an emotion classification module. The emotion classification network extracts facial features such as the eyes, eyebrows, mouth, etc., to classify the frames into one of the three classes, positive, negative, and neutral. All levels of stress were combined to negative emotion class. Figure 2.13 represents the model architecture for emotion prediction from facial expression.

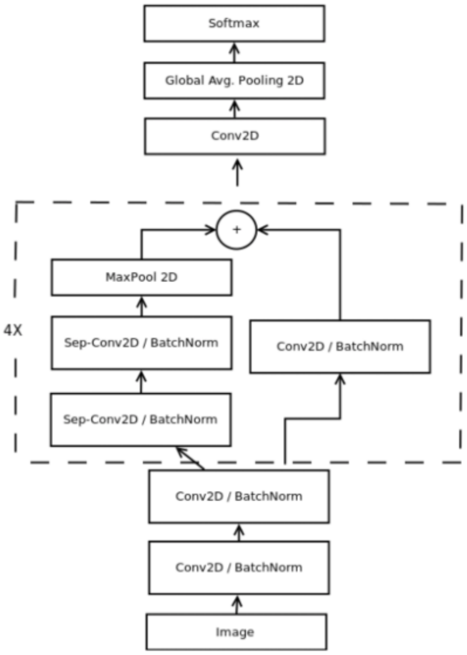


Figure 2.13. Architecture for emotion prediction from facial expression (inspired from [2]).

2.4.5.2 Emotion Recognition from Body Poses

As represented in Figure 2.14, the pipeline for the emotion recognition from the body poses consists of two stages, body key-points detection and emotion recognition part. A deep convolution-based architecture was used for keypoints detection [68, 69]. The system takes in an RGB image ($Format : Width$ x $Height$ x $Channels$) and produces as output the 2-dimensional locations of the anatomical key-points for each person in the image. The feed forward network of the neural network predicts a set of 2-Dimensional confidence maps of body part locations and a set of 2-Dimensional vector fields of part affinity fields which holds the degree of association between the parts. The confidence maps and the affinity fields are parsed by greedy inference for the final output of the 2-Dimensional key-points. In the architecture, the network is split into two branches; the top branch predicts the confidence maps, and the bottom branch predicts the affinity fields. The input image is first sent to a VGG-19 model to generate a set of feature maps which is sent as input to the first stage of each branch. In the subsequent stage, the predictions from the two branches are concatenated with the original features for better results.

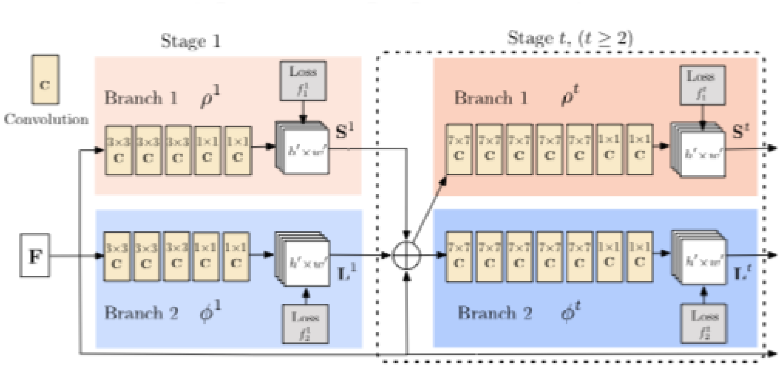


Figure 2.14. Architecture for body Key-points detection [3].

The second part is the emotion recognition from the key points detected. The human body is a complex structure with several degrees of freedom. Research has shown that people can exhibit emotion through their body postures. Extensive research has been performed on how humans behave and exhibit emotions under various conditions [70, 71, 72]. Wallbott and Harald, in their paper [71], mention the various possible postures for different emotions such as anger, sadness, happiness, disgust, etc. For example, the authors mention that humans tend to lean their body forward when angry, and they tend to bend their head down when sad, etc. Based on the information from these papers and the results extracted from body key-points detection, intermediate data were extracted, such as the position of the hands, head tilt angle, body angles, etc. The system then detects the postures and classifies them to one of the categories, i.e., ”positive,” ”neutral,” or ”negative,” and produces an output of 1, 0, or -1, respectively. The system was tested with 200 annotated images of different human body postures and produced an accuracy of 71 percent.

2.4.5.3 Calculating Task Outcome Prediction from Image Sensors

The above sub-systems described in Section 2.4.5.1 and Section 2.4.5.2 classify the input image frames into one of the three classes. For each sequence, both the modules predict the emotional state of the user in each frame, and an array of emotion state is generated. Figure 2.15 represents the output of emotion recognition from facial expression and body postures in one frame. It shows the body key points represented with colored dots joined by lines and a bounding box around the face for facial expression. The image also shows the final output of the two sub-systems, which displays a ’0’, meaning that the emotion recognized was neutral.

The above sub-systems (Sections 2.4.5.1, 2.4.5.2) were individually trained to find the optimum threshold cutoff of negative and positive emotions to maximize successful predictions. The data was split into 90% training and 10% test data. The system was trained and tested multiple times with different test data to check the consistency of the method. The first modality (EFE - Emotions from facial expression, Section 2.4.5.1) achieved a maximum accuracy of 75%. The second modality (EBP-Emotions from the body postures, Section 2.4.5.2) was able to achieve maximum accuracy of 62.5 %. The accuracy is also mentioned in Table 2.2.

In Algorithm 1, for every sequence, the total neutral frames, positive frames, and the negative frames are computed. With the threshold and the estimation of emotions for every frame in a sequence as parameters, this data is sent to Algorithm 1 which predicts the user performance outcome. For each sequence, the algorithm finds the prediction class and confidence of prediction from which the final prediction is made in Algorithm 2.

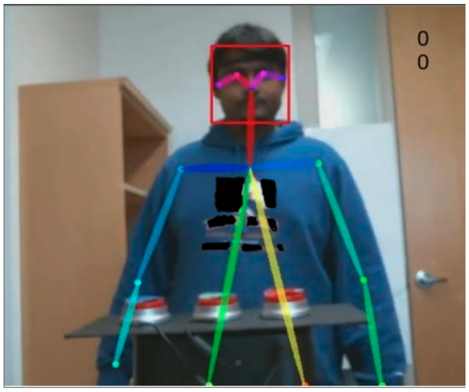


Figure 2.15. Output of emotion from facial expression and body postures. Zeros represent the prediction of the individual modalities.

2.4.6 Predicting Task Performance Outcome from Physiological Data

The third modality is to predict task outcome from EEG signals collected with the MUSE headband. The MUSE headband collects five different bands of EEG signals (alpha, beta, gamma, delta, and theta) based on which the signal's features were extracted with a CNN. With the ground truth data from the NAO robot and the features extracted, the network was trained to predict the task performance. Usually, EEG signals are noisy. The noise was removed by applying an Exponentially Weighted Moving Average (EWMA) filter and was normalized. Figure 2.16 represents the architecture for training and predicting the EEG signals. The architecture

Algorithm 1 Task Performance Prediction from Behavioral Data

```
Input: Frames of a sequence with emotions predicted, Threshold from training
Output: Individual modality prediction, classes(Success, Failure)
$neutral\_frames$ = total frames with neutral emotions in a sequence
$positive\_frames$ = total frames with positive emotions in a sequence
$negative\_frames$ = total frames with negative emotions in a sequence
$Total\_frames$ = total number of frames in a sequence
if (number of predicted negative emotions >= Threshold) then
    $Prediction = Failure;$
    $$Confidence = \frac{negative\_frames + neutral\_frames}{Total\_frames} \tag{2.1}$$
end
if (number of predicted negative emotions < Threshold) then
    $Prediction = Success;$
    $$Confidence = \frac{positive\_frames + neutral\_frames}{Total\_frames} \tag{2.2}$$
end
```

consists of two convolutional layers, and each of them followed by a batch normalization operation with a ReLU activation function. They are followed by two fully connected layers and a softmax layer which produces the probabilities of the participant succeeding and failing the task. The class which has the highest probability will be considered. The network weights were initialized with Xavier initialization [73], and Adam optimizer was used as an optimizer. The system was trained with 90% of the data and tested with 10% of the data. The cross-validation process was performed ten times with different validation sets to check for consistency of the model across all samples. The proposed model produced 83% accuracy as mentioned in Table 2.1 and Table 2.2.

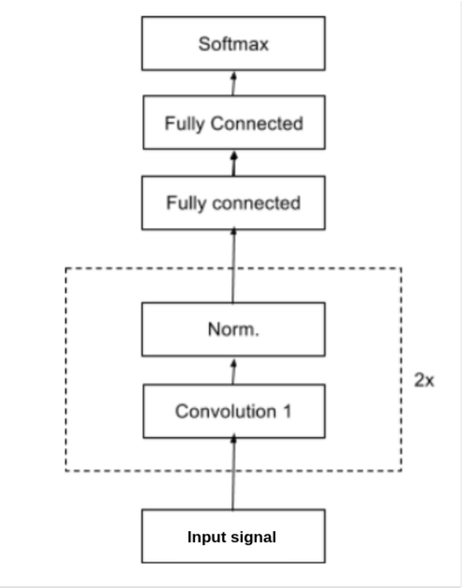


Figure 2.16. Network to predict task outcome from EEG signal.

2.4.7 Final Task Performance Prediction

From the individual predictions of each of the three modalities, a combined decision is made for the final prediction as mentioned in Algorithm 2. The individual modalities predict user performance outcome which is a success or a failure. With the prediction and the confidence values of the individual modalities, the total score is calculated which can be found in Equation 2.4, Algorithm 2. A negative confidence value is assigned if prediction is ”failure” and a positive confidence value is assigned if prediction is ”success”. If the total score is still positive, the final prediction will belong to the ”success” class and if the total score is negative, the final prediction will belong to the class ”failure”.

Algorithm 2 Final Performance Prediction Combined from Three Modalities

```
Input: Prediction output and confidence from 3 modalities, pred(EFE),
       pred(EBP), pred(ENN)
Output: Final Prediction, classes(Success, Failure)
if (any of the predictions is Failure) then
    $Confidence\_of\_that\_Modality \ * = -1;$    (2.3)
end
$Score = Confidence\_EFE + Confidence\_EBP + Confidence\_ENN$    (2.4)
if (Score is positive) then
    $FinalPrediction = \text{Success};$
else
    $FinalPrediction = \text{Failure}$
end
```

2.4.8 Results and Discussion

The collected data was split to 90 percent training and 10 percent validation. The training and testing procedure was performed multiple times with different training and testing samples to check the consistency of the system. The mentioned results in Table 2.2 are the average of the results. Accuracy and F1 score were calculated for individual modalities, EFE(Emotion from Facial Expression module), EBP(Emotions

Table 2.1. Task outcome prediction from EEG signal. Abbreviations: SVM-Support Vector Machines, GB-Gradient Boosting, RF-Random Forests, ET-Extra Trees

	SVM	GB	RF	ET	ENN
F1 Score	0.62	0.69	0.56	0.54	**0.82**
Accuracy	0.65	0.74	0.67	0.75	**0.83**

from body postures), and ENN(EEG signal with Neural Network) the combined results were calculated as mentioned in Algorithm 2. In addition, the results from [74] where task outcome prediction was performed on EEG signals with SVM(Support Vector Machines), GB(Gradient Boosting), RF(Random Forests), ET(Extra Trees) were also compared in Table 2.1. It is clear from the results that the accuracy of the combined system has outperformed the accuracies of the separate modalities. Also, the results provide higher accuracy than the existing traditional algorithms.

From the above experiment and results, it was observed that when we induced stress by increasing the length of the sequence, both the facial expression and the body posture modalities predicted more negative emotion frames, thus explaining that stress is reflected in the emotions. Moreover, there was a change in the extracted features of the EEG signals when there was a sudden increase in complexity of the sequence. A combined prediction having the highest accuracy in the task performance prediction provides some evidence that cognitive stress are observed in human behaviors, and combining physiological and behavioral information helps us understand more about the individuals.

	EFE	EBP	ENN	EFE+EBP+ENN
F1 Score	0.738	0.540	0.820	**0.870**
Accuracy	0.75	0.625	0.83	**0.875**

Table 2.2. Prediction from individual modalities and combined. Abbreviations: EFE-Emotion from facial Expression module, EBP-Emotions from body postures and ENN-EEG signal with Neural Network

2.4.9 Future Work

There is plenty of scope to extend this research down this line. The Emotion recognition from the body pose system produced the lowest accuracy of 62 percent. It could be because of various factors. The system considered only the current position and the state of the user during every prediction. Moreover, bodily expressions change based on cultural background. There is a need to detect motions and gestures, which might increase the accuracy. There was a shortage of annotated data for this specific setup. Increasing the number of training samples will have a positive impact on accuracy. Additionally, based on the information from the modalities, personalization will be implemented. This means the task aims at improving/training individual's cognitive capabilities. Since its subjective, the information will be used to personalize the task for the individual.

CHAPTER 3

ASSESSING COGNITIVE SKILLS IN CHILDREN

3.1 Introduction

In the previous chapter, we investigated how technology and Machine Learning, in particular, can be used to understand various aspects of human behavior, primarily related to emotion recognition in assessing cognitive skills in adults. Most of our applications were to evaluate them in their workplace. Therefore, monitoring the game/task metrics was comparatively easy, from which multiple meaningful cognitive measures were extracted. Yet, it was not complete with just the metrics from the task being performed, and hence there was a need to understand the behavior of the adult while performing the given task. Therefore, several methods were proposed, such as monitoring the physiological signals and the external behaviors while the adult was performing the task to extract information about the participant's emotion with which the adult's mental state was predicted.

In recent years, some researchers have proved that assessing cognition and diagnosing various cognitive impairments in the early stage (childhood) could positively impact the person's growth [75]. However, children with these cognitive problems are also more likely to grow up to have substance use disorders, impulse disorders, and other types of mental illness, which need to be detected early in life. In addition, identifying cognitive problems early childhood provides the best opportunities for remedial intervention, as brain plasticity is highest in children and diminishes with age. Therefore, it is essential to improve our understanding of how to assess cognitive functionalities in children better.

Executive Functions include cognitive processes that coordinate, integrate and control cognition, particularly in novel situations, and are necessary for high-order problem solving and goal-directed behavior [76, 77, 78]. They are predominantly divided into three major categories, Inhibitory Control, Cognitive/mental flexibility, working memory, as represented in Figure 3.1. Cognitive impairments, particularly in executive functions, can lead to poor academic performance and lifelong difficulties in family, employment, and community functioning. Ackerman et al. [79] analyzed the role executive function plays in preschoolers' academic performance and development. Attention Deficiency Hyperactivity Disorder (ADHD) is a common psychiatric neurodevelopmental disorder that often could cause cognitive impairments, specifically with executive abilities/functions. They are commonly found in children and young adolescents, starting at the age of 6, and occur three times more frequently in boys than in girls [80]. ADHD is generally associated with greater risks for low academic achievement, poor school performance, retention in grade, school suspensions and expulsions, poor peer and family relations, anxiety and depression, aggression, conduct problems and delinquency, early substance experimentation and abuse, driving accidents and speeding violations, as well as difficulties in adult social relationships, marriage, and employment [81]. Children with ADHD often exhibit a slower growth in certain cognitive skills known as Executive Functions such as working memory, cognitive flexibility, response inhibition, planning, and sequencing, etc. [82]. In addition, in 2009, researchers found that the brains of students with ADHD mature more slowly than their peers, and the part of the brain that enables students to work on "boring tasks," such as school work, has a reduced number of dopamine receptors and transporters, which explains why students can play video games for hours but struggle to complete their homework on time [83]. Thus, it is vital to improving our understanding of how to assess cognitive capabilities in children better.

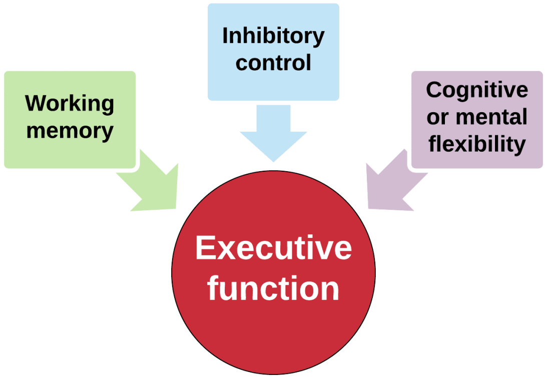


Figure 3.1. Three category model of Executive Function.

3.2 Traditional Tests to assess ADHD and Cognitive Impairments in Children

Research on cognitive impairments has been conducted for more than five decades to understand the relationship between cognitive impairments due to various psychiatric neuro-developmental disorders. Work has been done with various approaches for a proper diagnosis and intervention in the past. Traditionally, the diagnosis starts with gathering comprehensive background information through interviews with the children, parents, and school teachers, followed by trained psychologists administering standardized tests and a feedback session on the performance to explain the findings, provide recommendations for possible treatments or intervention. Figure 3.2 represents the steps involved in the assessment of suspected ADHD. The procedure usually starts with a psycho-social intake questionnaire to understand the medical history, developmental history, immediate environment, social history, family psychiatric history, etc., followed by tests to understand the child's behavior to various circumstances. The following sections provide a brief introduction to some of the commonly used approaches to assessing ADHD and cognitive impairments.

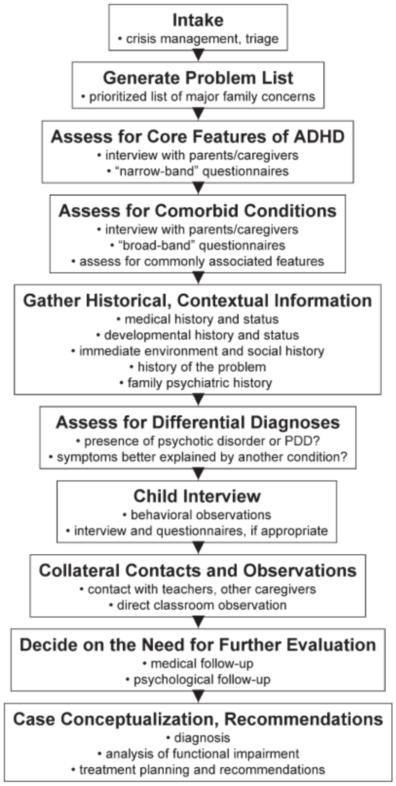

Intake
• crisis management, triage
Generate Problem List
• prioritized list of major family concerns
Assess for Core Features of ADHD
• interview with parents/caregivers
• "narrow-band" questionnaires
Assess for Comorbid Conditions
• interview with parents/caregivers
• "broad-band" questionnaires
• assess for commonly associated features
Gather Historical, Contextual Information
• medical history and status
• developmental history and status
• immediate environment and social history
• history of the problem
• family psychiatric history
Assess for Differential Diagnoses
• presence of psychotic disorder or PDD?
• symptoms better explained by another condition?
Child Interview
• behavioral observations
• interview and questionnaires, if appropriate
Collateral Contacts and Observations
• contact with teachers, other caregivers
• direct classroom observation
Decide on the Need for Further Evaluation
• medical follow-up
• psychological follow-up
Case Conceptualization, Recommendations
• diagnosis
• analysis of functional impairment
• treatment planning and recommendations

Figure 3.2. Overview of procedural heuristic for the assessment of suspected ADHD [4].

3.2.1 Swanson Nolan and Pelham (SNAP) Questionnaire

The Swanson, Nolan, and Pelham Teacher and Parent Rating Scale (SNAP), developed by James Swanson, Edith Nolan, and William Pelham, is a 90-question self-report inventory designed to measure Attention Deficit Hyperactivity Disorder (ADHD) and Oppositional Defiant Disorder (ODD) symptoms in children and young adults [84]. Each question measures the frequency of a variety of symptoms or behaviors, in which the respondent indicates whether the behavior occurs ”not at all,” ”just a little,” ”quite a bit,” or ”very much.” The questionnaire is designed for use with children and young adults ages 6-18. The results provide insights on measures such as inattention, hyperactivity, impulsivity, etc.

3.2.2 Computerized Assessment Tests

Computerized tests to assess various cognitive measures provide a more significant benefit of speed, accuracy, and low cost. Furthermore, automated tests have several advantages over paper-based tests. These include standardized administration of the test across a wide range of subjects, automatic scoring, and reporting, self-paced instructions. In addition, computerized tests are consistent in providing quantitative analysis of performance, allow for frequent assessment of cognitive function, and in some instances can also be self-administered inexpensively at home. One such toolkit which is widely used is the NIH toolbox cognitive Battery (CB) [5]. In this section, we review existing computerized neurocognitive tests and their effectiveness for an executive behavior assessment. In recent years, many assessment tests have been published to assess various aspects of Executive Functions (EFs) such as Inhibitory Control, Working memory, Cognitive or mental flexibility.

Eriksen Flanker Test is a set of response inhibition tests that are used to assess the ability to supress responses that are inappropriate in a particular context. In this test, the participants are required to indicate the left-right orientation of a centrally presented stimulus while inhibiting attention to the potentially in-congruent stimuli surrounding it (i.e., the flankers, two on either side), in this case, the stimuli are arrows pointing left or right. The modified flanker task [5] is a version of the Eriksen Flanker [85] task was adapted from the Attention Network Test [86]. In the modified version, which is used with children, the stimuli are fish (designed to be more engaging and larger, making the task easier). The version created for the CB includes both an easier fish block and a more difficult arrow block. On some trials, the orientation of the flanking stimuli is congruent with the orientation of the central stimulus, and on others, it is in-congruent. Performance on the in-congruent trials provides a measure of inhibitory control in the context of selective visual attention (which can also be considered a measure of executive attention). Figure 3.3 represents a trial sequence for the flanker inhibitory control and attention test.

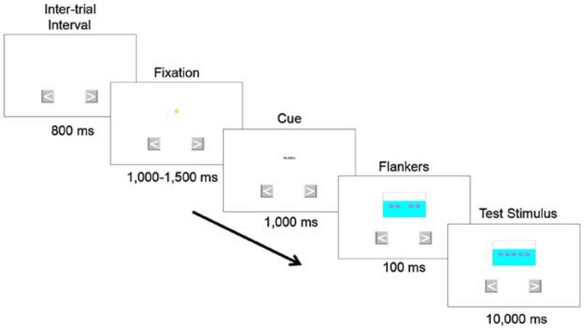


Figure 3.3. Trial sequence for the Flanker inhibitory Control and Attention Test [5].

The Dimensional Change Card Sort (DCCS) Test was proposed to measure cognitive flexibility, also known as task switching or set-shifting. This task, designed by Zelazo and colleagues based on Luria's seminal work on rule use, has been used extensively to study the development of cognition in childhood. In the standard version of the DCCS, children are shown two target cards (e.g., a blue rabbit and a red boat) and asked to sort a series of bivalent test cards (e.g., red rabbits and blue boats) first according to one dimension (e.g., color), and then according to the other (e.g., shape). Most 3-year-olds perseverate during the post switch phase, continuing to sort test cards by the first dimension, whereas most 5-year-olds switch flexibly. Both the standard version of this task and a more challenging version show excellent test-retest reliability in childhood. Figure 3.4 represents a trial sequence for the DCCS test.

The List Sorting Working Memory Test is a sequencing task requiring children and adults to sort information and sequence it. Items are presented both visually and auditorily. First, participants are presented with a series of illus-

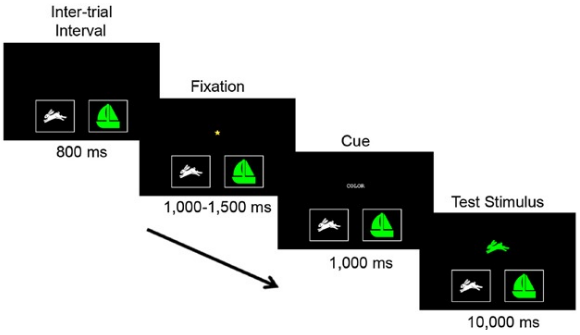


Figure 3.4. Trial Sequence for the Dimensional Change Card Sort Test [5].

trated pictures, each depicting an item on the computer, along with their auditory names—each item is displayed for 2 seconds. Next, participants are instructed to remember the stimuli and repeat them verbally to the examiner in order of size, from smallest to largest. The number of objects in a series increase on successive items, thereby taxing the working memory system when longer sequences need to be remembered. Furthermore, the task starts with a "1-list" version where the children have to sequence one type of stimuli according to size order and then switch to a "2-list" version where two types of stimuli have to be sequenced, each in size order.

The stop-signal task has been built to assess response inhibition and response time where the participants are expected to respond based on visual cues [6]. Participants were asked to press a button as quickly as possible when the target appeared. If participants did not respond within 650 ms, they received negative feedback. During 'Stop trials,' a centrally displayed red stop signal was presented after the green target appeared at various delays, and participants were asked not to respond. Figure 3.5 represents a sequence of the stop-signal task.

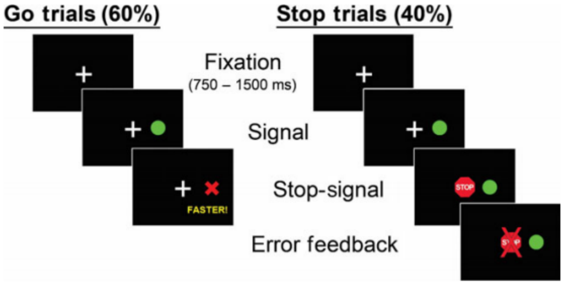


Figure 3.5. Schematic representation of stop-signal task. During GO trial, participants responded to a peripheral target appearing on the right or left side by pressing the corresponding button. During STOP trial, a centrally presentaed red stop-signal appeared at a variable delay follwoing the GO cue. Participants were instructed to inhibit their response when the STOP cue appeared [6].

3.3 Drawbacks of the Current Methods in Measuring Cognitive Skills

The discussed traditional and Computer-Based Assessment tests in this chapter have proved to be very effective in assessing cognitive skills. The NIH toolbox has been a standardized tool for such assessment tests in recent years. However, these tasks, being computer based do not require much body movements of the participants as part of the assessment process. Studies have shown significant improvement in cognitive skills of children who are more active physically, which establishes a stronger connection between physical movements and cognitive skills [87, 88]. Thus, there is a need for assessment tests that are both physically and cognitively demanding and are closer to the daily functions of children. The next chapter introduces a novel assessment test, Activate Test for Embodied Cognition which intends to assess cognitive skills through physical tasks.

CHAPTER 4

THE ATEC SYSTEM

4.1 Introduction

In the previous chapter, we presented some of the popular techniques to assess cognitive impairments in children and how cognitive impairments are among the common effects of ADHD. We discussed how these assessments were made in earlier times, which was more of a subjective measure that demanded objective measures to assess cognitive aspects in children. Advances in technology and Artificial Intelligence (AI) facilitated building computerized assessments to produce more accurate measurements to determine various cognitive elements, specifically executive functions. This chapter presents the impacts and effects of physical exercises and activities on cognition and cognitive training.

4.2 Physical Tasks to Assess Cognition in Children

Physical activities are an essential manifestation of cognitive functions [89]. Consequently, physical activities can be used to assess and train cognitive skills [90] and such assessments are easy to implement in school settings. At the same time, understanding how physical manifestations of cognitive skills correlate with other types of manifestations (such as response to problem-solving computer-based tasks) remains far from complete [91]. A key hurdle in improving this understanding is the difficulty and time expense of measuring performance in physical activities.

The inclusion of physical exercises in cognitive training is motivated by research illustrating that physical fitness and activity in children lead to measurable improvements in cognitive skills and academic performance [87]. The physical tasks should be designed so that their cognitive demands should be similar to the cognitive demands made by the computer-based training tasks. Thus, these physical exercises can be used to train sustained attention, self-regulation, working memory, cognitive flexibility, and multiple simultaneous attention [92].

4.3 The Head-Toes-Knees-Shoulders (HTKS) Test

Head-Toes-Knees-Shoulders (HTKS) is one of the prevalent and established tasks to assess various cognitive measures. HTKS measures behavioral self-regulation integrates aspects of Executive Function into a short game appropriate for children aged 4-8 years. Using no materials but rather relying on interactions between the examiner and the child, the HTKS has three sections with up to four paired behavioral rules: ”Touch your head,” ”touch your shoulders,” ”touch your toes,” and ”touch your knees.” Children first respond naturally and then are instructed to switch rules by responding in the ”opposite” way (e.g., touch their head when told to touch their toes) [93]. If children respond correctly after all four paired behavioral rules are introduced, the pairing is switched in the third session (i.e., head goes with knees and shoulders go with toes). HTKS measures behavioral self-regulation by requiring children to integrate into their behavior the following EF skills: (a) Paying attention to the instructions, (b): using working memory to remember and execute new rules while processing the commands, (c) using inhibitory control through inhibiting their natural response to the test command while initiating the correct, unnatural response, and (d) using cognitive flexibility and working memory when rules accumulate and then change in the second and third sessions. Performance in the HTKS task was shown to correlate with academic achievement outcomes in young children [94]. HTKS is not a task that children participants explicitly receive training on; hence, HTKS is a valuable measure of improvement in physical manifestations of cognitive skills, independent of the specific training tasks.

4.4 Activate Test of Embodied Cognition

The Activate Test of Embodied Cognition(ATEC) system [95] has been specifically built to assess various cognitive measures such as working memory, response inhibition, coordination from physical exercises which children perform as part of the test. Table 4.1 represents the list of ATEC tasks that has been formulated for various measures. It consists of 17 physical exercises with different variations and difficulty levels, designed to provide measurements of executive and motor functions, including sustained attention, self-regulation, working memory, response inhibition, rhythm and coordination, and motor speed and balance. The measurements are converted to ATEC scores which describe the level of development (early, middle, full development).

Table 4.1. ATEC tasks to assess various Cognitive Measures

Category	**Test**
Gross Motor Gait and Balance	Natural walk, gait on toes, Tandem Gait, Stand eyes closed hands outstretched, stand on One Foot
Synchronous Movements	March Slow, March Fast
Bilateral Coordination and Response Inhibition	Bi-Manual Ball Pass with Red Light, Green Light, and Yellow Light
Visual Response Inhibition	Sailor Step Slow, Sailor Step Fast
Cross Body Game	Cross your body (Ears, Shoulders, Hips, Knees)
Finger-Nose Coordination	Hand Eye Coordination
Rapid Sequential Movements	Foot Tap, Foot-Heel, Toe Tap, Hand Pat, Hand Pronate/Supnate, Finger Tap, Appose Finger Succession

4.4.1 Gross Motor Gait and Balance

Generally, people with cognitive impairments can experience motor dysfunctions, including deficits in gait and balance. The tasks include walking forward, where the participants were made to walk in a straight line for x number of steps. The main goal was to analyze if there is any abnormality in walking patterns and balancing.

4.4.1.1 Gait on Toes

In this task, the participants are asked to walk for eight steps on their toes(sneaky toes). The task assesses how many correct steps the child can make out of 8 expected number of steps. In this scenario, the correct number of steps means a correct step on toes.

4.4.1.2 Tandem Gait Forward

In this task, the participants are asked to walk in a straight line where for every step, the heel of the foot moving forward is expected to touch the toes of the leg behind. The correct steps here is calculated through the total number of correct steps performed out of the total number of expected steps(i.e. eight steps)

4.4.1.3 Stand-Arms Outstretched

This task mainly aims to assess balance in children. The participants are expected to stand still with their both hand out-stretched for 10 seconds. Participants are scored based on their ability to stand for the given time duration. In addition, scores were provided based on the number of seconds they can stand without giving up.

4.4.1.4 Stand on One Foot

In this task, the participants are expected to stand on one foot for 10 seconds. Participants are scored based on their ability to stand for the given time duration. Scores were provided based on the number of seconds they can stand without giving up. In the first trial, this was performed with the left foot, and in the second trial, the same was performed with the right foot.

4.4.2 Bilateral Coordination

Bilateral coordination refers to the ability to coordinate both sides of the body at the same time in a controlled and organized manner. For example, a child who is delayed in developing bilateral coordination skills may prefer to use one hand alone rather than both hands together. Good bilateral integration/coordination indicates that both sides of the brain are communicating effectively and sharing information. Children who have difficulty coordinating both sides of their body can have difficulty completing daily living takes(dressing, tying shoes), fine motor activities(stringing beads, buttoning), visual-motor tasks (drawing, writing) and gross motor activities such as walking, climbing stairs, etc.

Bi-manual Ball Pass

During this task, the participants are expected to juggle a ball based on auditory cues. During the task, the participant is expected to pass the ball from one hand to another hand for every beat accurately and in a timely manner. This is done for two trials, with the first trial where the commands are provided for every 1.5 seconds (slow) and the commands are provided for every 1 second (fast) for a total of 8 repetitions.

4.4.3 Attention, Response Inhibition

Attention is one of the important components of cognitive functions. Attention is defined as the ability to focus and process information in the environment. In this process, it is important to ignore and filter out unrelated information and to be able to perform a task despite the presence of a distraction. Similarly, Inhibitory control, also known as response inhibition, is a cognitive process and, more specifically, an executive function that permits an individual to inhibit their impulses and natural, habitual, or dominant behavioral responses to stimuli in order to select a more appropriate behavior that is consistent with completing their goals. Self-control is an important aspect of inhibitory control. For example, attempting to cross a road and to stop on seeing a vehicle on the way.

Ball Pass to the Beat

Similar to the bi-manual Ball Pass, in this task, the participants are expected to juggle the ball based on both auditory and visual commands. In addition to just passing the ball, the task had additional constraints such as ”No ball pass” and ”hand raise.” The commands were provided in the form of lights,

”Green light” - Make the pass

”Yellow light” - Hand Raise

”Red Light” - No Pass

The commands are provided in the form of both ”auditory” and ”visual” ques. There are two trials for each kind of cue: one where the commands are fired for every 1.5 seconds (slow) and the other where the commands are provided for every 1 second(fast) as represented in Figure 4.3.

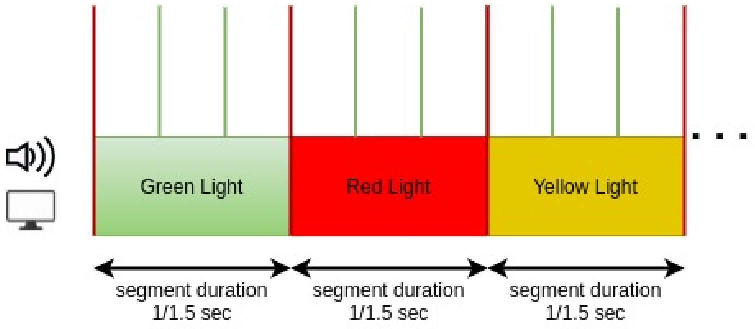


Figure 4.1. Audio visual stimuli during the task. The speaker represents the audio cues and the display represents the colors displayed on a screen.

4.4.4 Motor speed

Foot Tap - Tap the toes of the feet as fast as possible for 10 seconds.

Heel-to-Toe Tap - Alternate between heel and toe-tapping

Hand Pat - Tap the hand to the leg as fast as possible for 10 seconds.

Hand Pronate/Supinate - Alternate between the back of the hand and palm

Finger Tap - Tap the index and thumb together as fast as possible for 10 seconds.

Oppose Finger Succession - Tap the thumb and each finger together in succession.

4.5 ATEC Data Acquisition System

In this section, the data collection setup for the ATEC system is defined. Children between the age of 6-10 years were invited to participate in the ATEC assessment after the required parent consenting and screening procedure required by the study protocol. The ATEC administration includes a recording and administrative interface created to streamline assessments with as little distraction and interruption as possible. The ease of use is paramount as the assessment suite will be by both experts and non-experts. Video data is preferred as the sensor-based data collection can be more expensive and distracting, especially with child participants. Two Microsoft Kinect V2 cameras record a front and side view of the participants. RGB, depth, audio, and skeleton data are stored. The recording modules are connected to the Android-based administrative interface, which controls the flow of the assessment. It allows the administrator to select between all the tasks in the ATEC suite. Figure 4.2 represents the data collection setup. Each task has an instructional video and one or more assessment videos, while there are also practice videos to ensure that the child has understood the rules. An instructional video gives a brief demonstration of the current exercise and how it is performed. Selecting an assessment video triggers the recording modules to activate while Aliza, the on-screen instructor, guides the children through each task. An annotation software was developed to enable both computer scientists and cognitive experts to visualize and annotate the collected data. The software performs automated segmentation given the timestamps of the presented stimuli for each task. For each assessment recording, an expert evaluates the performance against a set of task-specific criteria. The annotation and scoring guidelines were designed considering both computer vision and related cognitive aspects of the task. This expert annotation is then used as the benchmark for automated approaches. Figure 4.3 represents a screenshot of the annotation software.

4.6 ATEC Dataset

Data were collected from 55 children between the age of 6 - 10 years. The procedure starts with the parents completing pre-screening paperwork which collects information about the child's history and family history. This is followed by paper-based assessment tests such as the Psycho-social intake questionnaire, Social Responsiveness Scale, Swanson, Nolan, and Pelham questionnaire, etc., that were discussed in the previous chapter.

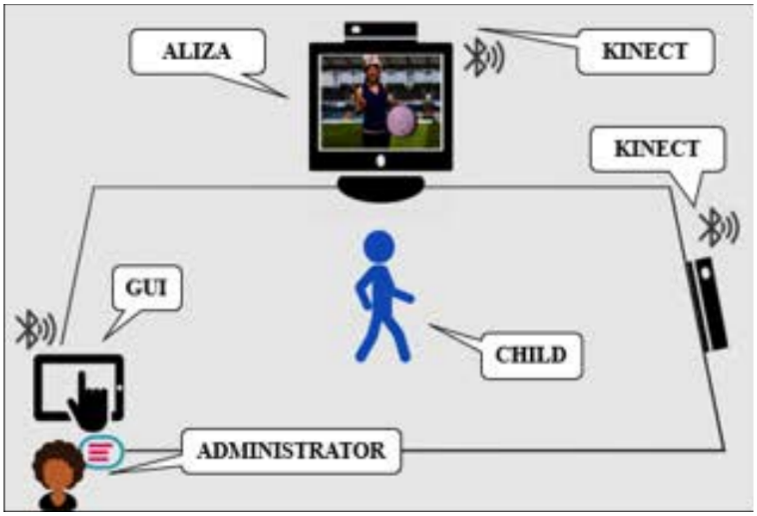


Figure 4.2. The ATEC setup includes two kinect cameras, a large screen and a tablet interface for the administrator. Administration takes place in classroom environments.

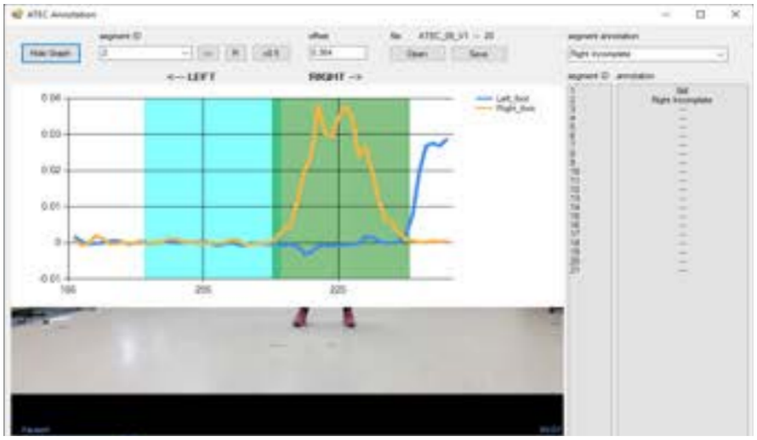


Figure 4.3. Annotation software was developed to enhance manual scoring and annotate the collected data, given the task rules and the cognitive measures to be assessed.

This is followed by standard computer tests from the NIH toolbox to measure various cognitive measures such as attention, response inhibition with tests such as go/no go tests, flanker tests, etc. Finally, the children perform all the tasks from the ATEC program. Data were collected for two trials, with each trial a week apart.

4.7 Automated System to Assess Physical Exercises

Unlike computer-based exercises, there are currently no automatic methods for assessing individual performance during training from physical activities. Therefore, it isn't easy for these embodied cognitive exercises to collect a large amount of data to study correlations between performance in those exercises and overall level and improvement in specific cognitive skills. Furthermore, there is also no way to do precise error analysis for feedback and correction. Thus, a computational goal of the proposed research is to overcome these limitations by developing human motion analysis algorithms that measure performance for the exercises mentioned above through various computational techniques.

CHAPTER 5

ASSESSING MOTOR SKILLS WITH RAPID SEQUENTIAL MOVEMENTS

5.1 Introduction

In this work, published by Babu et al. [7], a novel intelligent system to monitor and assess motor speed in children while they perform rapid sequential hand movements through which the cognitive development of the child can be estimated. The system uses computer vision techniques to detect hands automatically and predict the gesture as they are performed. At the end of the task, the system provides statistics on the performance of the subject. We use a task from the proposed ATEC system, which is the "Finger Opposition" task. The task is a well-established task to assess sensorimotor function for various neurological disorders. Students are instructed to tap the index, middle, ring sequentially, and little finger against their thumb during this task, as shown in Figure 5.1. The subjects are expected to perform the sequential movement for every count/beat provided by the system. The proposed approach has a user interface that can automatically record performed actions using a simple camera, predict actions, and visualize the performance statistics. The proposed system was built with state-of-the-art deep learning techniques for hand detection and action recognition.

5.2 Finger Opposition Dataset

As part of this work, a dataset was created for the finger opposition task with five subjects (2 males and 3 Females) and combined with an existing dataset [96]. The combined dataset consists of data from 10 subjects (approximately 4500 images)

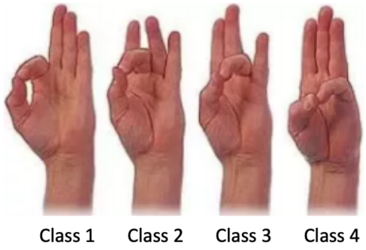


Figure 5.1. Finger Opposition task: The four hand represents the four classes [7].

with various hand angles to increase the robustness of the system. The image frames were manually divided into sequences(set of image frames determining a class) and were annotated as depicted in Figure 5.2. A sequence is complete when one of the four fingers touches the thumb and returns to the original position. We annotated a total of 200 training sequences with sequence lengths ranging from 10 frames to 28 frames. Similarly, there were 50 validation sequences, and the system was tested with 30 sequences that were collected in real-time.

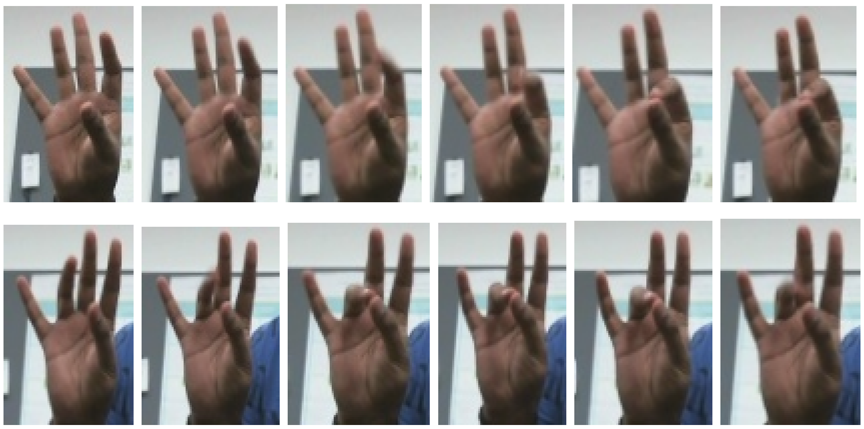


Figure 5.2. Sample image sequence for class 1 (top) and class 3 (bottom) [7].

5.3 Proposed System

The primary goal of the proposed method is to build an intelligent system that can automatically capture the actions performed by the subjects, classify them and generate scores as shown in Figure 5.3. This will serve as a tool to evaluate and assess physical activities which can disclose executive function disabilities. The proposed system consists of multiple parts functioning together to achieve this goal.

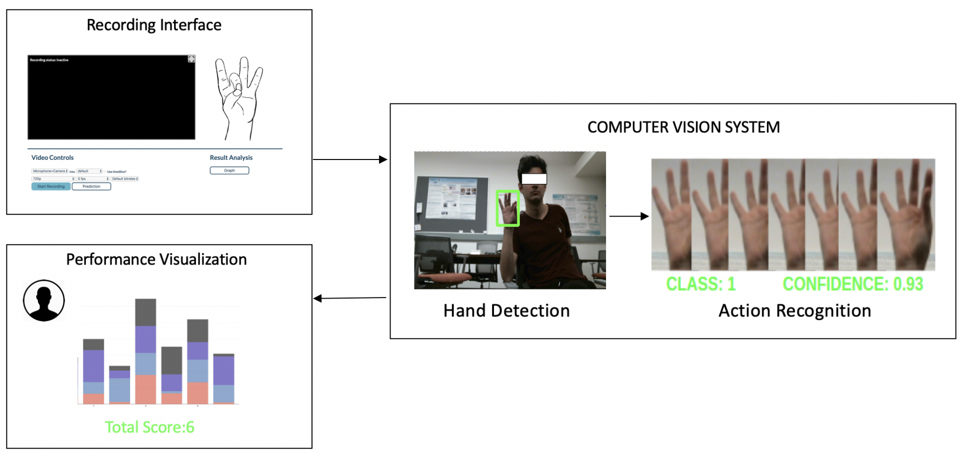


Figure 5.3. Proposed Architecture.

5.3.1 Intelligent GUI

The GUI that has been developed allows the therapist to have control over the complete system. The GUI has two main components, Recording and Analysis. The interface was developed with Flask, a web development framework for Python. The modular design provides a comfortable and easy-to-use interface for both recording and analysis.

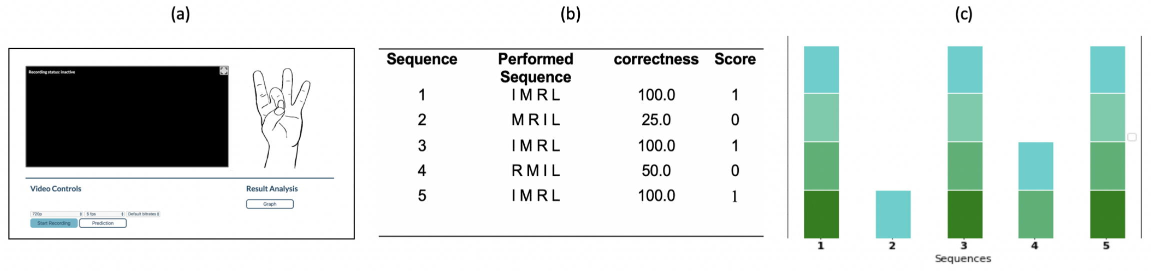
5.3.1.1 Recording Interface

This module of the interface is primarily used by the therapist to initiate or to stop the recording. The interface can produce beeps while recording to guide the subject. It also facilitates managing the frame rate and resolution of the image frame is recorded. The interface also provides a view of what is being recorded during the task. This way, the therapist can keep track of how and what is being recorded. For the image frames recorded through the interface, we define a temporal window with its size based on the average time taken to complete a sequence (i.e., Switch from one finger to another finger). Predictions are made on the image frames at every step of the sliding window. The predictions are reflected on the interface for real-time visualization. Figure 5.4(a) shows a sample screenshot of the fully functional interface.

5.3.1.2 Analysis Interface

Once the recording is stopped, the therapist has the option to visualize the performance statistics of the subject. The interface provides a consolidated score for the subject and step-wise scores allowing the therapist to see where the subject missed the sequence. From Figure 5.4(b), which shows a screenshot of the score table, The therapist will be able to see a detailed view for every sequence (consists of all four subsequences). Column 2 (Performed Sequence) explains the order in which the sequence was performed. The abbreviations represent I - Index Finger, M - Middle Finger, R - Ring Finger, L - Little finger. For example, in sequence 1, the order is "IMRL," which means the participant has performed it in the correct order. Hence, the correctness for that sequence is 100 percent, with a full score of 1 point. In sequence 2, the participant has correctly performed only one subsequence;

hence the correctness percentage is 25 percent with a score of zero. Similarly, Figure 5.4(c) allows the therapist to visualize the correctness as a graph where each column represents a sequence, and each colored square represents a finger. The column is generated based on the fingers that were correctly performed in a sequence. For example, in Figure 5.4(b), sequence 2, the subject performed only the little finger in the correct order, and hence only its respective color is shown in Figure 5.4(c)



Sequence	Performed Sequence	correctness	Score
1	IMRL	100.0	1
2	MRIL	25.0	0
3	IMRL	100.0	1
4	RMIL	50.0	0
5	IMRL	100.0	1

Figure 5.4. (a) Screenshot of the Recording Interface, (b) Screenshot of the Score table, (c) Screenshot of the performance table.

5.3.2 Hand Detector

The first part of the computer vision pipeline is the Hand Detector. This part aims to detect the active hand in the scene and pass it on to the action recognition system. Even though there are numerous traditional approaches such as HOG classifiers, Ada-boost classifier, etc., it is possible to build a more robust, accurate, and time-efficient system for real-time processing by exploiting recent developments in Deep Neural Networks.

For hand detection, we use an approach called Single Shot Multi-Box Detector (SSD)[97] which was built using EgoHands Dataset [98, 99]. The feature that sets SSD apart is that it uses only a single deep neural network for the entire process of detecting the hands. In contrast, other methods such as Fast-RCNN [100] employ multiple elements in their pipeline, which makes SSD more time-efficient [97]. The algorithm splits the given input frame into a grid of size $N \times N$, where for every cell in the grid, a set of default boxes with different ratios and scales are generated. During prediction, the network generates scores for the presence of objects(hand) in each of the default bounding boxes. If the score is greater than a certain threshold, the system assumes a hand in that particular generated default box. The system finally performs non-maximum suppression to remove duplicate predictions. In addition, this procedure is performed at various scales of the feature map to capture hands of different sizes.

5.3.2.1 Training, Validation, and Testing

As mentioned above, the system was trained with EgoHands dataset [98] which contains more than 4800 image frames with approximately 15000 ground-truth labeled hands. This dataset was chosen because it includes images of people performing different activities that involved hand movements (playing chess, playing cards, solving puzzles, etc.).

The dataset was split into train(80%), validation(10%) and Test(10%). In addition, the system was tested on numerous image frames collected throughout the Finger Opposition task. The detector was evaluated with Mean Average Precision (mAP). The system's mAP was the highest of 96 percent when tested at a 0.5 threshold. The system works at a rate of 15 frames per second on a single GPU. Only the detected hand is passed on to the next stage as other parts of the captured frames are not relevant for the classification. Figure 5.5 represents a sample output of the Hand Detector system.

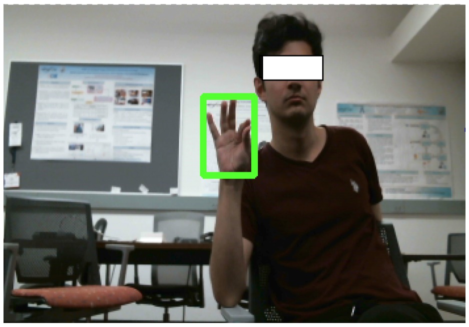


Figure 5.5. Prediction from Hand Detector.

5.3.3 Action Recognition System

Our system classifies a sequence of images of a specific window size into one of the four classes mentioned above, as shown in Figure 5.1. Our action recognition system is based on 3D Convolutional Neural Networks (CNNs) [101]which is a natural successor of standard 2D CNNs [102]. CNNs have achieved state-of-the-art results in many computer vision applications. They are a variation of artificial neural networks, which are translation invariant, and weights of spatial filters in each layer are shared across the entire image. In 3D CNNs, instead of 2D spatial filters, 3D Spatio-temporal filters are employed, which means they extract features from both the spatial and the temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in the multiple adjacent frames.

In this work, we used a particular variant of CNNs called Residual Deep Neural Networks (ResNet)[103] which has been one of the most successful architectures for image classification and feature extraction. The residual neural network utilizes skip connections or short-cuts to jump over some layers. The motivation for skipping over layers is to avoid vanishing gradient problems, facilitating building more deeper networks that are easy to train and optimize.

One of the main challenges of 3D CNNs is that they have a huge number of learnable variables, and as a result, they require a huge amount of data for training. Training such deep networks with fewer data leads to overfitting of the model. Thus in our experiment, we used a comparatively shallow network 3D-ResNet10 [104] which is basically a ResNet with ten layers. We also attempted to pretrain the network on general activity recognition datasets like kinetics [105], but it did not contribute much towards better performance. Each block of 3D-ResNet10 comprises convolutional layers with a 3D kernel of size $3 \times 3 \times 3$, Batch Normalization (BN), and Rectifier Linear activation units (ReLu).

5.3.3.1 Training, Validation and Testing

During training, we split our dataset into training and validation with a ratio of 4 to 1. We performed K-fold cross-validation during our training process. For testing, we used 30 sequences that were collected in real-time. During the training, the length of the sequence was 8, i.e., we trained our network with 8 images for every sample. The image frames were RGB images with a resolution of 64×64

Our network was optimized with Adam optimizer with a learning rate of 0.1, which was divided by 10 when validation loss saturated. Figure 5.6 represents the Confusion Matrix generated for the test dataset with the best performing model based on Table 5.1. From the confusion matrix, we observed that the system classified class 1 and 2 perfectly while there was some miss classification for class 3 and class 4.

5.3.4 Scoring System

The Scoring system calculates the scores for the task performed by the subject based on the task rules and guidelines. Before calculating the scores, the predictions are pre-processed in order to smooth the prediction made by the action recognition

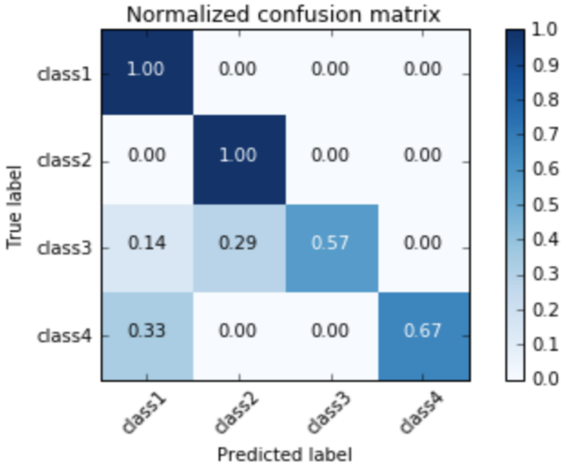


Figure 5.6. Confusion Matrix for the Action Recognition tested real time.

Table 5.1. Comparison with 3D ResNet and its variants

Model	Val.	Pre.	Recall	Test acc.	(Sec.)
ResNet10	**0.89**	**0.88**	**0.88**	**0.80**	0.12
ResNet10 Pre.	0.92	0.91	0.91	0.625	0.11
ResNet18	0.39	0.17	0.39	0.37	0.18
ResNext 50	0.26	0.28	0.24	0.20	0.23

system. The smoothing stage is expected to handle any of the misses in the right prediction from the computer vision system.

5.3.4.1 Smoothing Process

During testing, we use the Sliding Window Approach to make predictions for the frames that are being collected. Depending on the subject's speed performing the task and the frame rate at which the system works, there is a possibility for multiple predictions when the subject switches from one finger to another. An example output of prediction will look similar to (1,1,1,1,2,2,2,3,3,3,3,4,4,4,4), where the number represents the class predictions of a window, and a confidence score is associated with each prediction is also generated. Then, we perform a smoothing process similar to the Moving Average approach, where it helps fix any miss classification with the help of the confidence score. For every prediction, if the confidence score is less than a certain threshold (i.e., the system is not confident about its prediction) and if the current prediction is different from its neighbors, then the current prediction will be updated. After the smoothing process, the duplicate predictions are combined into one prediction by averaging their confidences.

5.3.5 Calculating Scores

The Scoring System generates a score for the exercise performed. A complete sequence is when the subject performs all four subsequences, i.e., Thumb to the Index finger, Thumb to Middle finger, Thumb to the Ring finger, and Thumb to Little finger. The subject receives a full point when a complete sequence is achieved. The score ranges from a minimum of zero to a maximum of the number of times the complete sequence is performed (The maximum is considered as five in this case). The subject is expected to switch from a subsequence to another only after hearing the beat from the system.

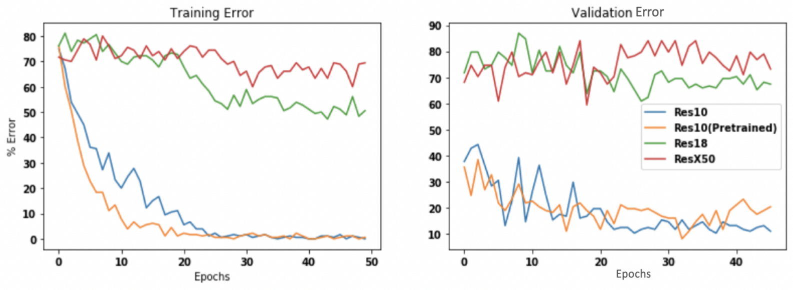


Figure 5.7. (Left): Training curve comparisons for different architectures considered. (Right): Validation curve comparisons for the architectures considered.

5.4 System evaluation, analysis, and Discussion

As mentioned, action recognition is still an unsolved problem, and it is highly dependent on large datasets. Therefore, multiple attempts were made to choose the suitable method for the Finger Opposition task. Based on the recent surveys on action recognition [106, 26, 107], we built methods that have been proven to work best on public datasets, trained, fine-tuned them. Table 5.2 shows the attempted ways, and the optimal method was chosen based on the validation results.

3D Convolutional Networks was the right approach for this problem, and it has been observed that residual networks work better for action recognition [13], still, we attempted to train on other popular networks such as DenseNet, which did not yield good results. ResNet 18 was built and trained; the model performed poorly with validation and test accuracy less than 40 percent. One of the reasons for this to happen could be the model's inability to generalize. Hence, an attempt was made to vary the depth of the network by using ResNext-50 and ResNet-10. While ResNext-50 performed poorly as predicted, we observed that ResNet-10 achieved both validation and testing accuracy and precision and recall more than 80 percent, which was the best among all the attempted approaches. We also tried to pretrain ResNet-10 with other public datasets such as Kinetics, UCF-101, etc. but failed to beat the one trained from scratch. We also evaluated the system on its time taken to process a sub-sequence. Table 5.1 shows the time taken for each of the models to predict one sequence.

Also, we plotted the error percentage of the different ResNet models during training and validation. ResNet-18 and ResNext-50 shows high error rate. Based on all the above observations, we concluded that ResNet-10, which was trained from scratch, performed better than all other attempted models and methods, including CNN+RNN(LSTM, GRU), Multi-Stream Networks.

Table 5.2. Table showing the comparison with other state-of-the-art methods

Method	Val. Accuracy
2D-CNN+LSTM	0.65
2D-CNN + GRU	0.60
Multi-Stream Network	0.76
3D Convolutional Networks	**0.89**

5.5 Evaluation of the End-to-End System

The accuracy of the end-to-end system is measured in terms of the performance attained by the complete pipeline (Hand Detector, Action Recognition, and Scoring) in successfully scoring the exercise. The input to the system will be a stream of RGB image frames, and the output of the system will be scored. We collected data from 5 new subjects who were not involved in the initial data collection to assess the complete system. This was done to ensure that the system is user-independent. With the manual annotations, we calculated sequence-wise and total scores for all the participants. Figure 5.8 represents the performance of all 5 participants. In the figure, the columns (pair of bars) represent individual participants, the left bar in each column represents what the participant actually performed, the right bar represents system predictions, and the colors in each bar represent different sequences. As mentioned above, every participant performed the sequence five times; hence, the maximum step-wise score can be attained 20 ($5 \times (Index, Middle, Ring, Little)$). The scoring in the graph is similar to the correctness column in Figure 5.4(B). For example, in Figure 5.8 first column, the ground truth score is a perfect 20 as the subject performed all the five sequences perfectly while the prediction system correctly predicted the actions. Similarly, we can observe that in column 2, even though the participant performed all the sequences perfectly with a ground truth score of 20, the system missed classifying sequences 3 and 4 correctly. From Figure 5.8 we can observe that the system generated score accurately 96 percent of the time.

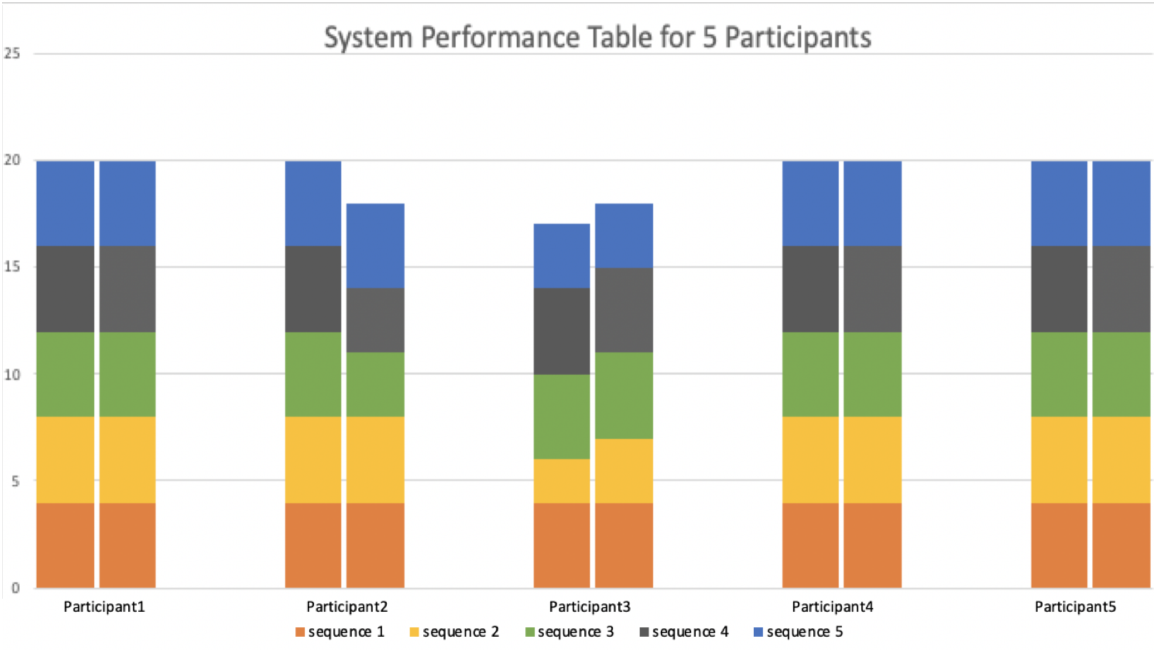


Figure 5.8. End-to-End System Performance. Left bar represents the actual performance of the subjects and right bar represents the system's prediction. Each column(2 bars) represents a participant .

5.5.1 Observations

From Figure 5.8, we can observe that the system scored the performance accurately for most of the sequences. While measuring the system accuracy, we observed a few different scenarios where the system did not perform as expected. Such scenarios include a drop in frames which sometimes missed temporal information leading to wrong predictions. Also, during the sliding window process, there were times when the window had frames from more than one class leading to ambiguity in classification for the system. These could be avoided by improving the model performance with additional data during the training, specifically for classes three and four.

CHAPTER 6

AN AUTOMATED SYSTEM TO ASSESS GAIT, ATTENTION AND RESPONSE INHIBITION

6.1 Introduction

This chapter introduces some of the important cognitive functionalities such as gait, attention, and response inhibition, followed by an automated system that can measure these functionalities through a physical task. The preliminary approach that uses a simple threshold-based algorithm is introduced, followed by the drawbacks of such approaches. Next, we introduce some of the work that has been proposed in recent years to solve the action recognition problem and evaluate their performance on our dataset, followed by a multi-modal approach that uses different features extracted from the RGB images. Finally, we compare the performance of the proposed multi-modal approach with some of the state-of-the-art approaches published in the recent past.

6.2 Computer Vision for Motion Analysis

Human activities are a sequence of body configurations and postures. Therefore, physical activity recognition can be characterized as a simultaneous alignment and recognition problem. This consists of recognizing body gestures from videos and measuring the correctness of the movement performed based on the alignment of these body configurations with expected movement. Specifically, in computer vision, this has been a long-existing problem. The motion boundaries around the person's contour seem to contain information for the action recognition that is as important as the optical flow within the region of the body. This hints at the idea that flow may help extracting the motion and form that is the shape of the object performing the action [9]. Similarly, human skeleton information also has been proved to capture the motion dynamics of people in the scene as the human body can be viewed as an articulated system of rigid bones connected by hinged joints. As the area evolved, multiple approaches have been attempted and applied to solve this problem.

In addition to working directly on the RGB video/sequence of images, work has proved that extracting valuable features from the video/sequence of images plays a vital role in solving this problem. In the following section, we will look into some of the commonly extracted features for Human action recognition.

6.2.1 Body Pose Estimation

One popular approach in recent years for human action recognition is to extract the body pose information from the scene, which is used as features to solve the Human action recognition problem. The extracted features from the sequence of frames are used to solve the temporal alignment problem. Many techniques have been proposed in recent years to model human body models. Some of the popular ways of representing/modeling humans are skeleton-based models, contour-based models, and volume-based models, as mentioned in Figure 6.1. Skeleton-based models, also known as stick figures or kinematic models, represent joint locations and the corresponding limb orientations following the human body skeleton structure. The skeleton-based model can also be described as a graph where vertices indicate joints and the edges encoding constraints or prior connections of the joints within the skeleton structure. Generally, for skeleton-based pose estimation, two different approaches are followed;

The top-down approach employs person detectors to obtain a set of the bounding box of people in the input image. It then directly leverages existing single-person

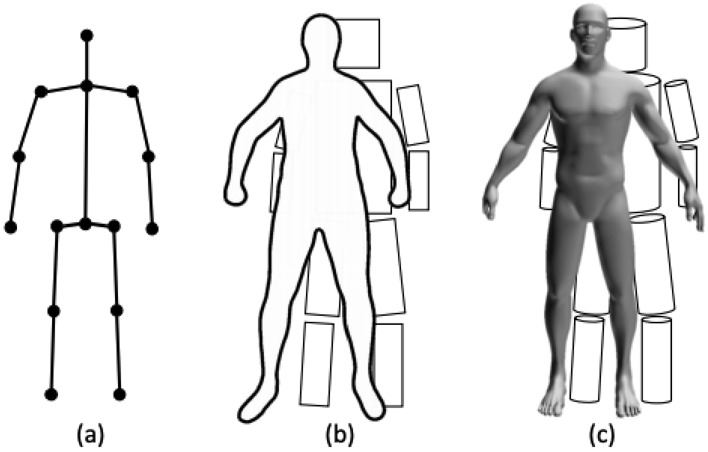


Figure 6.1. Commnly used human body models. (a) Skeleton-based models; (b) contour-based models; (c) volume-based models [8] .

pose estimation to predict human poses. The expected poses heavily depend on the precision of the person's detection. While bottom-up methods directly indicate all the 2D joints of all persons and then assemble them into independent skeletons. Some of the popular bottom-up approaches existing today include Deep Cut employing Fast R-CNN-based body part detector first to detect all the body part candidates, labeling each part to its corresponding part category, and assembling these parts integer linear programming to complete the skeleton. Similarly, OpenPose uses Convolutional Pose Machines (CPM) to predict candidates of all body joints with part affinity fields (PAFs). The proposed PAFs can encode locations and orientations of the limbs to assemble the estimated joints into different poses of persons. Similarly, Pose Partition Network (PPN) aims to conduct both joint detection and dense regression for joint partition. In addition, the PPN performs local inference for joint configurations with joint partitions.

6.2.2 Optical Flow

Optical flow is one of the popular motion representations for action recognition [9]. Optical flow is often formulated as estimating the 2D projection of the true 3D motion of the world. The field of optical flow has made significant progress by focusing on improving the numerical accuracy on the standard benchmarks. The first end-to-end trainable deep convolutional network is FlowNet [108] which was trained on synthetic data. To fill the shortcomings of the above method, SpyNet [109] was proposed a combination of the traditional pyramid approach and convolutional networks.

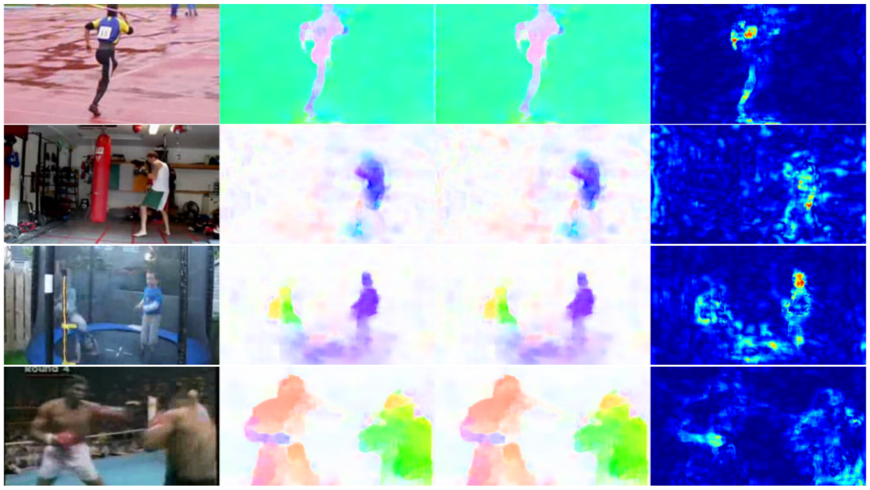


Figure 6.2. Sample results from SpyNet. The first image represents the input image, the two flow fields and the euclidean distance between the two flow vectors at each pixel. Flow vectors noticeably change more around motion boundaries and where humans are located [9].

6.3 Assessing Response Inhibition and Attention with Ball Drop to the Beat Task

Inhibitory control, also known as Response Inhibition, is a cognitive process and, more specifically, an executive function that permits an individual to inhibit their impulses and natural responses to stimuli to select more appropriate behavior that satisfies the required goal. Some of the existing popular neuropsychological tests include Flanker task, go/no-go, primarily computer-based, which were discussed in Chapter 3. Ball Drop to the Beat is a physical task that has been built to assess response inhibition. As part of this task, the child is expected to juggle the ball based on auditory and visual commands. In addition to just passing the ball, the task had additional constraints such as "No ball pass" and "hand raise." The commands were provided in the form of lights,

- **Green light" - Make the pass**
- **Yellow light" - Hand raise**
- **Red Light" - No pass**

Instructions were provided through visual and auditory cues and at a different pace. The automated system aims to detect the actions performed and score them according to the rules. The scores were generated based on the following two criteria,

Accuracy: In the given time, for the given command, what action was performed by the child.

Rhythm: How accurately in terms of time the child performed the task.

The ball drop dataset consists of a total of 3300 video segments that were extracted from the video recordings of 25 children. These video segments were manually segmented and annotated.

6.3.1 Preliminary Approach

The complete pipeline to score the ball drop task consists of multiple parts as explained in [110] and is represented in Figure 6.3. First, videos are recorded at the rate of 30 frames per second. Second, the input video is broken down into image frames which are decoded. As the first step, for feature extraction, we extract the body key points of the participants. The body key points considered for this experiment are wrist points, elbow points, and shoulder points. A Convolutional Neural Network (CNN)-based approach is used to extract the body keypoints [111]. The system uses the decoded image as input of size $w \times h$. The Deep Neural Network-based model predicts 2D confidence maps of the body joint locations and a set of 2D vector fields of part affinity fields which is the degree of association between the parts.

With the key points extracted for every frame in the segment, more detailed features such as the x-distance, y-distance between the wrist points, elbow points, wrist, and shoulder points were extracted. These features were used to detect various events during the exercise. For every segment, the features were pre-processed to remove noise. Noise includes any wrong detection of body key-points, key-points not being detected, etc. First, the moving average is computed on the features for every segment, followed by applying a low pass filter to remove the high-frequency components caused by hand jitters and minor movements.

A ball pass event occurs when the participant moves the hand holding the ball towards the other hand, makes the transfer, and moves back to the original position. In such a scenario, the distance between the wrist points decreases until the transfer happens and increases again. Similarly, a hand raise event occurs when the participant moves the hand holding the ball towards the hand's shoulder holding the ball and retreats to the original position where the distance between the wrist and

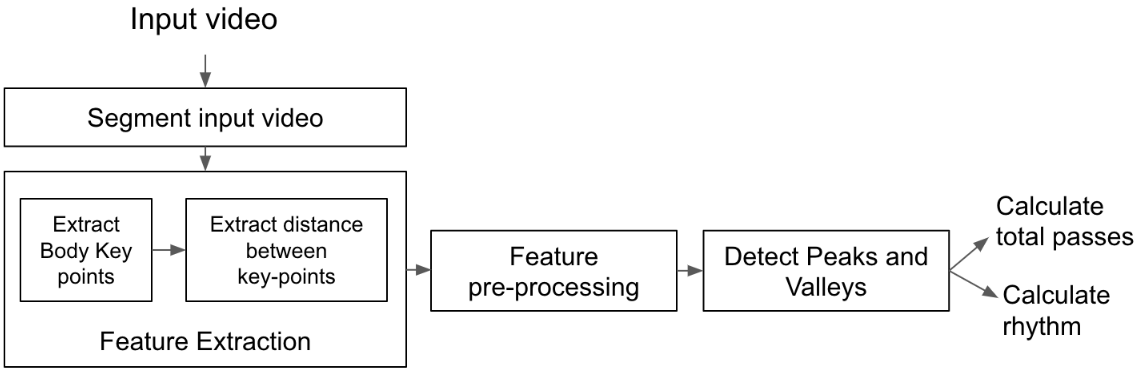


Figure 6.3. The overall architecture for threshold-based approach to score ball-drop task.

the shoulder joint initially increases and starts to decrease while retreating. A peak is formed every time such an event occurs.

After processing the segmented features, peak and valley detection is performed for the segment. Mathematically, peaks and valleys represent local maxima and minima. A video segment T which consists of n image frames, with x being the features for every image frame, is defined by

$$T = \{(f_1, x_1), (f_2, x_2), ..., (f_n, x_n)\} \tag{6.1}$$

Where f represents the frame number. Peaks (P) and valleys (V) for a segment are defined by,

$$\begin{aligned}
&P = \{(f_i, x_i) | (x_{i-1} < x_i > x_{i+1}) \vee (x_1 > x_2) \vee (x_n > x_{n-1}) \\
&and \\
&V = \{(f_i, x_i) | (x_{i-1} > x_i < x_{i+1}) \vee (x_1 < x_2) \vee (x_n < x_{n-1}), \\
&\forall i = 2, 3, ...n - 1\}
\end{aligned} \tag{6.2}$$

With the above equations, peaks and valleys are detected, which correspond to the respective events. Figure 6.4 (left) represents a ball pass event in a video

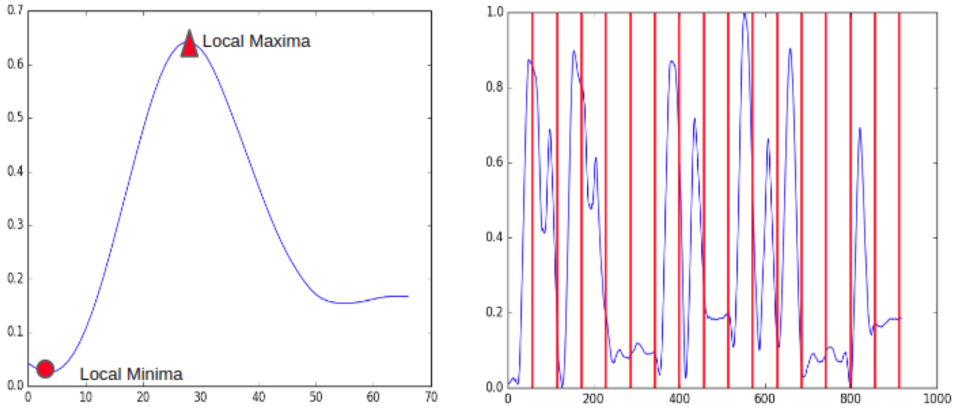


Figure 6.4. Ball pass event in a video segment (left). Segmentation of features, given the stimuli timestamps (right). The x-axis represent the frames in the video segment and y-axis represent the vertical distance between the joint locations.

segment. Figure 6.4 (right) represents the task being divided into different segments. In this approach, noise in the signal was detected as a ball pass when there was no pass. Hence, a threshold on the height of the peak was considered. The height of the peak is the distance between a peak and a valley in the segment. We used 998 segments from our dataset for this experiment. Fifteen percent (144 segments) of the data was used to identify the right threshold for different events, and 85 percent (854 segments) of the data was used for evaluation. For these evaluations, we use data from 7 subjects performing ten ball drop-related tasks. Each subject performed the exercises twice, one week apart, to determine test-retest reliability.

6.3.2 Rhythm Detection

The goal of detecting rhythm was to identify how timely the child performs the action in the given time duration. In order to identify this, the complete segment was considered, along with the key points being detected for every frame in the segment. Since the goal here was to identify at what point an action took place in a segment, only the upper body key points were considered. This includes the shoulder, elbow, and wrist of both hands. In Figure 6.5(b), the colored keypoints(hands) represent the features considered for the detection. For a given time t, $f1$ represents the features of the left hand and $f2$ representing the features of the right hand and they are computed as follows:

$$\begin{gathered} f1 = [z_{11}, y_{11}, z_{12}, y_{12}, ...z_{1n}, y_{1n}], \\ f2 = [z_{21}, z_{21}, z_{22}, y_{22}, ...z_{2n}, y_{2n}] \end{gathered} \tag{6.3}$$

where n represents the number of keypoints being considered for the detection. The distance d_t between the feature $f1$ and $f2$ is computed as follows:

$$d_t = f(f1, f2) \tag{6.4}$$

Hence, for a given video segment containing sequence of images, the distance between $f1$ and $f2$ in every image can be represented as $D = [d_1, d_2, ..., d_t]$. After a series of processing, such as normalizing, inverting the distance, smoothing, and plotting the distances for every frame, it was observed that a peak formed when an action took place. It was also observed that whenever there was no action happening in a segment, there would be a flat line representing no action. Figure 6.5(a) represents a sample segment with the curve representing D. The red lines in the graph represent the upper bound and the lower bound, between which if a peak occurs, it is assumed that the action has started and completed correctly in the given time.

6.3.3 Priliminary Results

The proposed method performs highest with 78 percent accuracy. Figure 6.6 represents the confusion matrix of the validation set using the proposed method. The rhythm detection system was evaluated on the test set. The data in the training set was used to empirically identify the optimal upper bound and lower bound in order

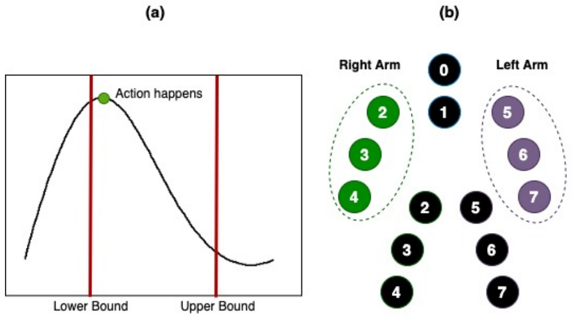


Figure 6.5. Rhythm Detection. (a) represents the distances D plotted in a segment. (b) represents the keypoints considered for rhythm detection .

to maximize the prediction accuracy. On the test set, the system was able to achieve an average accuracy of 88.5 percent in detecting the rhythm score.

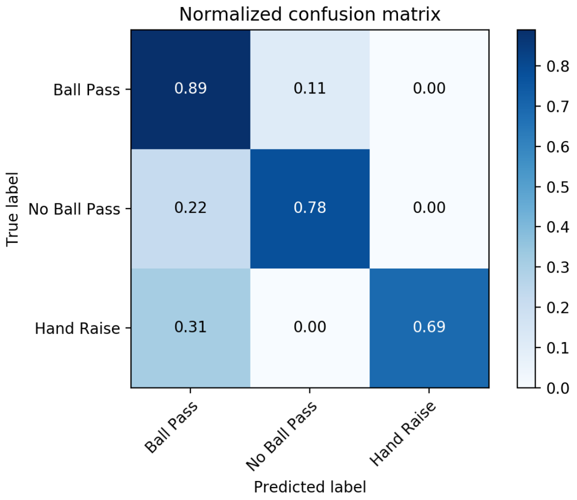


Figure 6.6. Confusion Matrix of Proposed method for Ball Drop to the Beat task.

The score for Response Inhibition (RI) is calculated with,

$$RI = \frac{number\, of\, correct\, nopass/redlight\, actions}{Total\, number\, of\, nopass/redlight\, commands} \tag{6.5}$$

Similarly, the score for attention (Attn.) is calculated with,

$$Attn. = \frac{number\, of\, correct\, raise\, \&\, pass\, actions}{Total\, number\, of\, raise\, \&\, pass\, commands} \tag{6.6}$$

6.4 Tandem Gait to Assess Motor and Gait Functions

The Tandem gait is a task that is part of the ATEC system to assess gait and balance in children. Children generally affected neurological conditions such as ADHD exhibit motor and gait abnormalities [112]. The target is for the kids to walk in a straight line for every beat provided with the heel of the moving foot touching the toe of the stationary leg as represented in Figure 6.7. The dataset of the Tandem gait task consists of 432 video segments that were manually extracted and annotated. For all the experiments, the dataset was split such that 90 percent were used for training and 10 percent for testing.

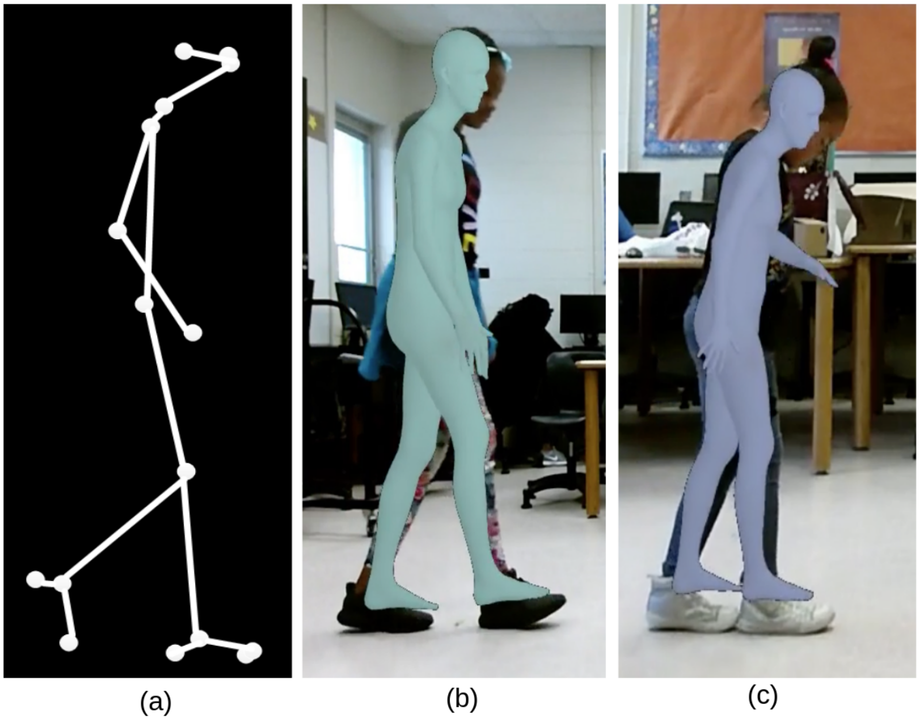


Figure 6.7. (a): Skeleton keypoints, (b) Example of a invalid step, (c): Example of an valid step.

6.4.1 Priliminary approach and Results

To automate the evaluation for the tandem gait task, the Video Inference for Body Pose and Shape Estimation (VIBE) system [113] was used to extract the 3D joint locations of the child from the recorded videos. VIBE is a video pose and shape estimation method that predicts the parameters of the SMPL body model. The VIBE system estimates a total of 25 body joints shown in Figure 6.7 (a). Since a valid step is only when the participant's heel touches their opposite leg's toe, the horizontal position of both foot heels and big toes was extracted. Further in our pipeline, the distance between left foot toe and right foot heel and the distance between right foot toe and left foot heel were calculated, represented in Figure 6.8. These two signals were first filtered (by a moving average filter) to remove the high-frequency components caused by foot jitters and minor movements and then combined to form a single signal. Finally, a peak detection algorithm was used on both the signals to detect valid valleys (which infer valid steps) in each signal, similar to the approach mentioned in Section 6.3.1. The sum of the valleys that are extracted from the mentioned algorithm provides the total score. However, for detecting the valid steps, only valleys with certain characteristics were chosen. We set up a threshold in our algorithm that differentiates all valid steps from the invalid ones. First, the local minima (valley) value which corresponds to the distance between heel and toe, should be below the threshold to be considered a valid valley. Secondly, if multiple small valleys are close to each other in each signal, they were considered one valid valley. For example, in Figure 6.8 there are four significant red valleys and four significant blue valleys. The dataset was divided into 10 percent for training and 90 percent for testing. The training data was used in identifying the right threshold for detecting valid valleys was chosen as 0.2. Comparing the scores predicted by the computer vision system with the manual scores shows that the proposed method can achieve an accuracy of 81.25%.

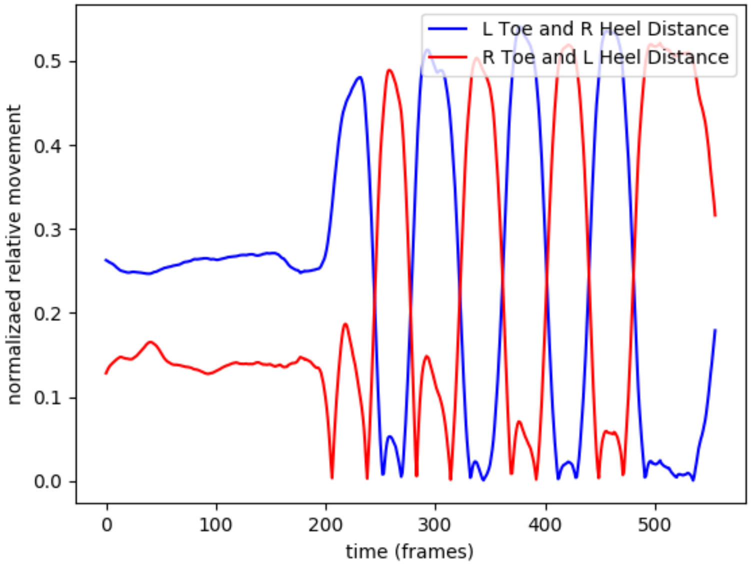


Figure 6.8. Distance between children heels and toes. The blue curve represents the distance between Left toe and right heel. The red curve represents the distance between right toe and left heel.

6.5 Supervised Learning for Human Action Recognition

As observed from the previous section results, a simple threshold-based approach is not robust in classifying the actions performed for the above-described tasks. Therefore, this section discusses the various state-of-the-art deep learning methods in human action recognition and its performance on the ATEC dataset and a multi-modal supervised approach that extracts multiple features from the RGB video to perform the classification. As part of this, some of the most popular video architectures for action recognition were implemented to see its performance on our dataset.

6.5.1 Convolutional Neural Network with Recurrent Networks

This approach [114] takes advantage of convolutional networks that have been proved in the past decade for their high performance to extract features from the individual image frames in a video clip independent of each other. A recurrent layer such as Long Short Term Memory units or Gated Recurrent Units LSTM/GRUs with batch normalization is used to process the output of the convolutional layer based architecture such as the Inception model. The combined model is trained using cross entropy loss on the outputs at all time steps. The high level architecture has been represented in Figure 6.9 (a).

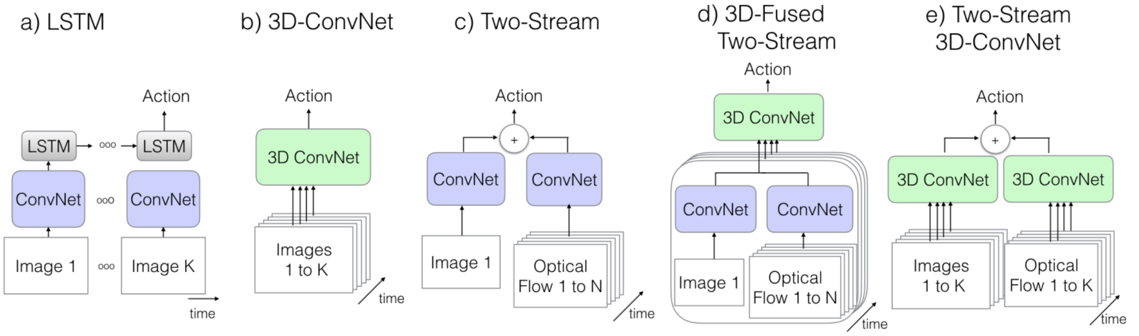


Figure 6.9. Existing State-of-the-art video processing architectures. K represents the total number of frames in a video, N stands for a subset of neighboring frames in the video [10].

6.5.2 3D-Convolutional Networks

3D convolutional networks have been commonly used to extract features from videos in recent times. This is very similar, but in place of 2-D kernels, 3D Spatiotemporal kernels were used [13]. Their unique characteristics to directly create hierarchical representations of spatiotemporal data have made them very successful. One of the possible drawbacks of this approach is its number of parameters compared to traditional 2D convolutional architectures. The high level architecture has been represented in Figure 6.9 (b).

6.5.3 Two-Stream Networks

This approach uses a two-stream convolutional network based architecture that incorporates spatial and temporal networks [115]. The spatial stream is used to take advantage of the spatial features, whereas for the temporal features, an off-the-shelf optical flow extractor was used for the givens video clip. The input to the spatial stream is an image frame that is randomly sampled from the given video clip. The output from the optical flow was stacked as input to the temporal stream. The output of the temporal stream and the spatial stream were averaged towards the final prediction. The architectural representation of this approach is in Figure 6.9 (c). The extension of this approach uses a 3D-Fused two-stream method that stacks multiple two-stream modules [116]. 3D kernels are used to combine information from the individual modules. The architecture of this method is represented in Figure 6.9 (d).

6.5.4 2-Stream 3D Convolutional Networks

This approach is an improvisation of the above method as explained in [10]. The approach uses two separate 3D convolutional Networks two extract information from the RGB-based video clips and optical flow-based video clips. The individual models were trained separately, and during inference time, the results from both models were combined to make the final prediction. The approach is represented in Figure 6.9 (e).

6.5.5 Multi-modal Data Fusion for Human Action Recognition

This section presents a multi-modal approach for human action recognition which was published by Babu et al. [117]. Inspired by the previous works [118, 119, 120], using multiple modalities has resulted in improving the results specifically for human action recognition. Literature has shown that extracting various information from the raw RGB input such as Body Keypoints, Optical flow information, tracking objects in the scene provides a varied set of information that one modality cannot provide, thus improving the final prediction results. Inspired by the previous work, multiple modalities (optical flow, object trajectories, body pose) were used to classify the actions performed by the kids. We use an attention mechanism to combine features from individual modalities towards the final prediction.

6.5.5.1 Individual Modalities

Modality 1: Optical Flow

With an extensive review of the advantages of using optical flow to understand motion as described in section 6.2.2, an off-the-shelf implementation [121] of optical flow was used to capture the motion information between the consecutive frames in a video.

With optical flow being computed, a deep neural network based architecture inspired from [13] was used to extract useful information from the optical flow data. The architecture is based on 3D Convolutional Neural Networks (CNNs), which is a natural successor of standard 2D CNNs. In 3D CNNs, instead of 2D spatial filters, 3D Spatio-temporal filters are employed to extract features from both the spatial and the temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. In addition, for this modality, a special variant of the CNNs called Residual Deep Neural Networks (ResNet) was built with 3D filters ($3x3x3$). Figure 6.10 (a) represents the architecture used, with the dotted blocks representing the residual blocks. Every convolutional operation was followed by a batch normalization operation to reduce the internal covariate shift and a Rectified Linear Activation Unit (ReLU). Down-sampling of the inputs is performed at conv 3_x, conv 4_x, and conv 5_x while increasing the feature size. A comparatively shallow network with 18 layers was empirically selected as represented in Figure 6.10(a). After training, the features were extracted from the pre-logit layer, which was used during fusion.

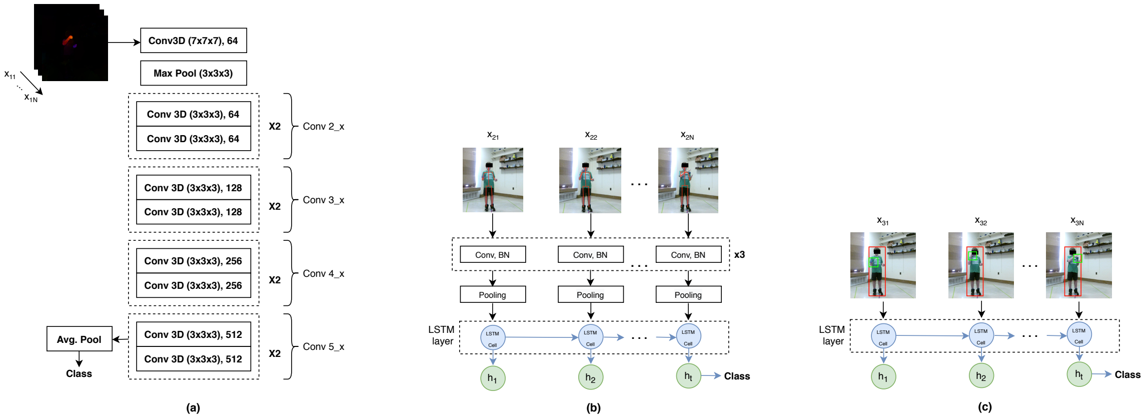


Figure 6.10. (a) Optical flow based method to predict actions performed. The dotted blocks represents residual block. (b) Action prediction using sequence of body keypoints. Conv. represents a convolutional operation and BN represents a batch normalization operation. (c) Objects coordinates in the scene based action prediction.

Modality 2: Human Poses

An open-source pose estimation framework was used to detect 3D joint locations from the RGB video [122]. As a pre-processing step, any missing key points in a given image frame are fixed with information from the previous frames. This top-down method first detects humans in the scene and subsequently performs pose estimation on each detected region.

For a given video segment containing n frames in it, 18 key points are extracted from each frame that represents various body joint positions, including facial key points such as eyes, ears, and nose. In this work, only nine key points (only upper body excluding facial key points) out of the 18 key points are considered as the remaining key points do not contribute significantly towards predicting actions in this scenario. Each keypoint is represented as a 3D coordinate (z, y, v) on the image plane. Hence, a given frame P at time t is represented by the coordinates of the nine key points as shown in the following equation:

$$P_t = [(z_{1,t}, y_{1,t}, v_{1,t}), (z_{2,t}, y_{2,t}, v_{2,t}), ..., (z_{9,t}, y_{9,t}, v_{9,t})] \tag{6.7}$$

Where z denotes the coordinate extending from left to right and y extending top to bottom, and v representing the depth for each keypoint, hence, for a given frame, the input dimension is of size (9,3).

The proposed subnet to extract spatial and temporal features from skeletal points is comprised of a series of 1D convolutional layers and batch normalization followed by a pooling layer. A single-layered Long-Short Term Memory (LSTM) unit with a hidden state (h) dimension of 32 is used to capture the temporal relation among the frames. The architecture is initially trained with a softmax layer at the end. During the fusion process, features h_t, which is the hidden state of the last LSTM block, are extracted. The subnet is represented in figure 6.10(b).

Modality 3: Object detection

This modality aims at detecting objects in the scene. It is essential to identify the objects being interacted with along with their positional information at a given point of time to predict the actions. Identification of the positional information of objects in the scene provides a sequence of coordinates. This sequence of coordinates is fed into a subnet to identify the trajectories of the objects being interacted with, leading to the identification of the actions [120]. Objects recognized in the scene $o_i = \{ l_i, s_i \}$ consists of a bounding box l_i and its category $s_i \in$ S, where S is the set of all possible object categories (e.g., ball, person) being encoded in the form of Binary Presence Vector (BPV) and i ranging from 0 to k with k representing the total number of objects detected in the scene. A popular object detection algorithm YOLO V3 [123] is used to identify the objects of interest in the scene at any time t. During detection, any missing objects in a given image frame were fixed with information from the previous frames.

For every image frame, the object's coordinates are normalized and concatenated along with the class vector. A single layered LSTM layer with a hidden state(h) size being 32 is built to capture the temporal relation between the frames. The architecture is initially trained with a softmax layer at the end. During the fusion process, features h_t, which is the hidden state of the last LSTM block, are extracted. The subnet is represented in Figure 6.10(c) represents the architecture to predict actions through objects in the scene.

6.5.5.2 Multi-Modal Fusion

In a multi-modal scenario, not all modalities equally contribute towards the final prediction. Identifying the modalities and features within them that have the most contribution and prioritizing them have proved to be very effective in every domain. In order to solve this problem, a self-attention-based fusion approach is proposed inspired by [124]. In this approach, every feature within each modality is provided with a corresponding weight which learns during the training process based on their contribution towards predicting the target.

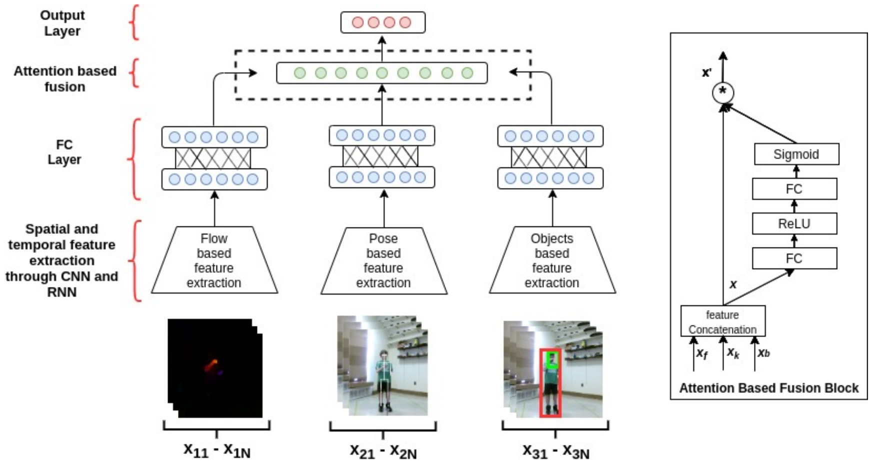

Output Layer
Attention based fusion
FC Layer
Spatial and temporal feature extraction through CNN and RNN
Flow based feature extraction
Pose based feature extraction
Objects based feature extraction
x11 - x1N
x21 - x2N
x31 - x3N
x'
*
Sigmoid
FC
ReLU
FC
x
feature Concatenation
xf
xk
xb
Attention Based Fusion Block

Figure 6.11. Multi-modal fusion.

The overall architecture, including the attention-based fusion module is represented in Figure 6.11. In order to calculate the weights of features of each modality, first all features are concatenated into one vector as follows:

$$x = [x_f, x_k, x_b] \tag{6.8}$$

where $x_f \in \mathbb{R}^{C_f}$ is the feature vector obtained from optical flow subnet, Figure 6.10(a), $x_k \in \mathbb{R}^{C_k}$ is the feature vector from the pose subnet, Figure 6.10(b), $x_b \in \mathbb{R}^{C_b}$ is the feature vector from objects position based subnet, Figure 6.10(c) and finally $x \in \mathbb{R}^{C}$ $(C = C_f + C_k + C_b)$ comprising of features from all modalities. Further, to calculate attention weights for features of x, function F_w is introduced as represented in equation 6.9. For F_w to fully capture feature-wise dependencies, it should meet two criteria. First, it must be capable of learning nonlinear interaction between features. Second, it must learn a non-mutually-exclusive relationship that ensures multiple features are allowed to be emphasised. To meet these criteria, a gating mechanism with a sigmoid activation is employed.

$$\alpha = F_w(x, W) = \sigma(g(x, W)) = \sigma(W_2\delta(W_1x)) \tag{6.9}$$

where δ refers to the ReLU [125] function, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. In order to generalize, the gating mechanism is parameterized by forming a bottleneck with two Fully-Connected (FC) layers (W_1 & W_2) around the non-linearity, i.e., a dimensionality-reduction layer with reduction ratio r, a ReLU and then a dimensionality-increasing layer returning to the original feature dimension of X. The final output is obtained by element-wise product of combined feature vector X and calculated attention weights vector α:

$$x' = F_a(x, \alpha) = \alpha x \tag{6.10}$$

Where x' represents the output of the attention block with the features from the modalities combined and weighted, which in turn is succeeded by a softmax layer for final prediction.

6.5.5.3 Experimental Results

Table 6.1 conveys that many state-of-the-art methods did not perform as expected on the ball drop dataset, with the proposed method outperforming all of them, which could be because of the nature of the data. For example, the ball drop task contains actions that are very similar to each other such as raising the hand and passing the ball, unlike actions in other popular datasets, requiring multiple modalities to solve the problem. It can be observed in Table 6.1 that two-stream I3D has produced second to the best results showing that optical flow could play a vital role in solving the problem.

In Table 6.2 it can be observed that the body keypoint-based model has achieved the highest accuracy as a single modality. Although usage of three modalities has produced satisfactory results when compared to the previous works for action recognition, extensive tests were necessary with a different combination of modalities and

Table 6.1. Existing State-of-the-Art methods (VS) proposed approach on Ball drop dataset. The results are averaged over 5-folds. KP - Key points, flow - Dense optical flow, RGB - RGB image frames, Object Pose - Objects in the scene

Method	**Test. Acc.**	**Features**
1D CNN	0.59	KP
3D CNN [13]	0.73	RGB
Two-Stream Network [115]	0.76	RGB+flow
Two-Stream 3D ConvNet [10]	0.82	RGB+flow
CNN + RNN(LSTM) [126]	0.69	RGB
DeepGRU [127]	0.61	KP
Dillhoff et. al. [110]	0.78	KP
Attnsense [128]	0.81	flow+KP
Proposed approach	**0.89**	KP+Object Pose+flow

fusion strategies to find an optimal solution with a much less complex method. It was observed that no other combination of modalities and fusion methods outperformed the proposed method. Adding object detection as an additional modality has improved the accuracy by 5.2 percent for attention-based fusion. Moreover, it can be observed that the combination of optical flow and object position and the combination of optical flow and body keypoints provide similar accuracy, and it is higher than the combination of keypoints with object position. This result verifies the important contribution of optical flow as an additional modality. When looking at the fusion strategies, literature has proven that usage of attention to weigh features based on their importance has worked, similarly, Table 6.2 proves the same. Irrespective of what the modalities are being combined, the attention-based fusion produces slightly better results.

Table 6.2. Experimental results for multi-modal approach. nat. Concat: Natural Concatenation, bal. Concat: Balanced Concatenation, Self-Attn.: Self Attention. All results averaged over multiple folds.

Method	**Test.**	**Time(Sec.)**
Optical Flow (opt_flow)	0.720	0.229
Body Keypoints(KP)	0.760	0.106
Objects Trajectories (Obj_Pos)	0.680	0.103
opt_flow+KP(nat. Concat.)	0.820	0.236
opt_flow+KP(bal. Concat.)	0.839	0.239
opt_flow+KP(Self-Attn.)	0.846	0.240
opt_flow+Obj_Pos (nat. Concat.)	0.841	0.232
opt_flow+Obj_Pos (Bal. Concat.)	0.839	0.236
opt_flow+Obj_Pos (Self-Attn.)	0.840	0.241
KP+Obj_Pos (nat. Concat.)	0.790	0.118
KP+Obj_Pos (bal. Concat.)	0.763	0.123
KP+Obj_Pos (Self-Attn.)	0.795	0.139
KP+Obj_Pos+flow (nat. Concat.)	0.890	0.254
KP+Obj_Pos+flow (Bal. Concat.)	0.875	0.259
KP+Obj_Pos+flow (Self-Attn.)	**0.898**	**0.260**

6.6 Conclusion

This section concludes with the results of some of the existing state-of-the-art action recognition approaches in the literature, followed by a multi-modal fusion approach to recognize the actions performed by the children. To identify the right combination of modalities and various fusion strategies, multiple combinations were attempted from which the architecture with three modalities produces the best results.

CHAPTER 7

SELF-SUPERVISED LEARNING FOR HUMAN ACTION RECOGNTION

7.1 Introduction

In chapter 6, we observed that Deep Neural Networks, with their powerful ability to learn different levels of general visual features, forming the backbone for many approaches we discussed. However, one of the significant drawbacks of the previous supervised techniques is that they rely heavily on expensive manual labeling and suffer from generalization error, spurious correlations, etc. [129]. Moreover, the performance of deep convolutional neural networks dramatically depends on the amount of training data.

As a promising alternative, self-supervised learning has drawn massive attention for its data efficiency and generalization ability [130, 131]. Many self-supervised methods were proposed to learn visual features from large-scale unlabelled images or videos without using any human annotations. To learn such visual features from unlabeled data, a typical solution is to formulate ***pretext tasks*** using Deep Neural Network architectures where these networks can learn objective functions. These pretext tasks utilize the structure of the data as a supervisory signal such that the method is unsupervised in the sense that it does not require human annotation. The pretext tasks are generally built not to solve the intended problem but an additional step to learn good data representation. Some of the popular pretext tasks to learn image/video representation include colorizing grey-scale images [132], image in-painting [133], image jigsaw puzzle [134], Generative models [135]. Compared to supervised learning methods, which require a data pair (X_i, Y_i) where X_i represents the data sample, and Y_i represents the label for the respective data sample, pseudo label P_i is automatically generated for a pre-defined pretext task without involving any human annotation. The pseudo label P_i can be generated using attributes of the images or by traditional hand-designed methods. Given a set of N training data $D = \{P_i\}_{i=0}^{N}$, loss function can be defined as:

$$loss(D) = \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} loss(X_i, P_i)$$

One major advantage of such self-supervised learning approaches is that they are generic in learning the features so that they could be used for any downstream tasks (E.g., Classification, object detection, etc.).

7.2 Context based Video Representation Learning

As videos consist of frames that are stacked together, they contain both spatial and temporal information. The temporal information between frames is used as supervision criteria for self-supervision feature learning.

7.2.1 Temporal Order Verification

The goal is to learn a feature representation using only the Spatio-temporal features that are available naturally in the videos inspired from the work of Misra et al. [136]. Figure 7.1 represents the overall architecture used for this method. In this approach, a tuple of frames are extracted from the video sample, and the model investigates whether the frames are in the correct temporal order or not. This task aims to encourage the model to learn the dynamics and motion of objects in the scene, thus learning the temporal structure of the tuple. Some of the critical challenges we observed were, a) choosing the total number of frames in the tuple, b) how a subset of frames can be sampled from the data to form a tuple. During our experiments, we observed that sampling the tuple uniformly from the input data did not yield good results as there were parts of the sample that had very minimum motion. Hence, the tuple was sampled from the middle region of the data sample that contained maximum motion in the dataset. For our dataset, we empirically identified that a tuple length of 16 yielded the best results.

Further, we used a ResNet-based architecture with 3D kernels to extract features from the input tuple for the encoder. Although, this could be replaced with other types of encoders such as Convolution + Recurrent network-based models, transformers, etc.

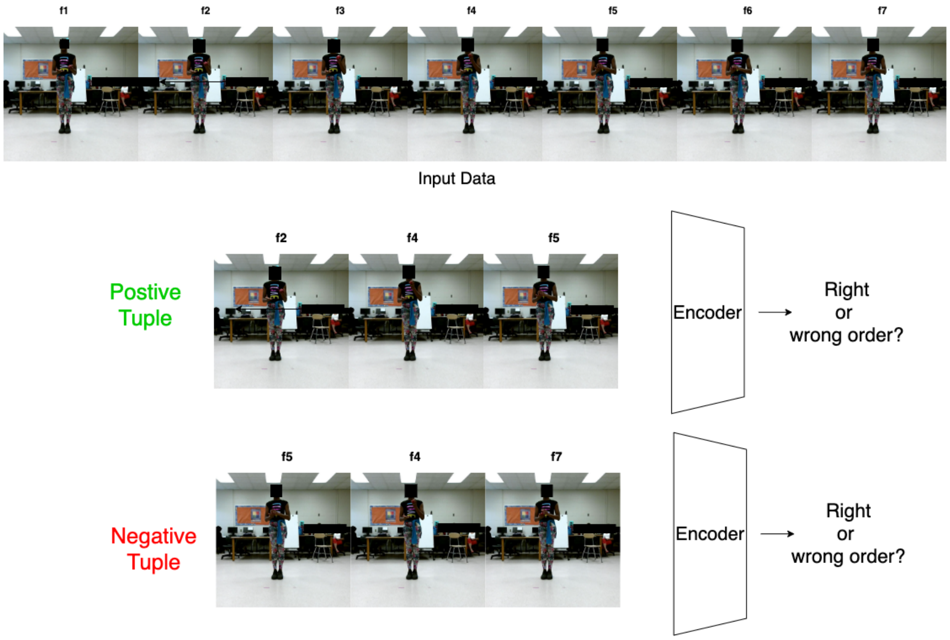


Figure 7.1. Overall architecture for temporal order verification .

7.2.2 Order Prediction Network (OPN)

This approach, inspired by the work presented by Lee et al. [137] proposes that successfully solving the sequence sorting task allows the model to learn useful representation by observing the dynamics of objects in the scene. As part of this work, a tuple of length four were extracted from the data sample and was shuffled. Given the length of the tuple, the maximum possible combination will be $4! = 24$ combinations. Similar to the previous approach, given the frames with maximum motion lies in the middle of the data sample, the tuples were a sample from the middle region of the input data sample. Furthermore, temporally consistent spatial augmentation for the extracted tuple that includes random cropping and color transformation was applied. Figure 7.2 represents the overall architecture. The red dotted box represents temporally consistent random crop-based spatial augmentation. The numbers represent the frames that were chosen from the input data.

As an extension of this work, multiple non-overlapping clips were extracted from the data sample and were shuffled. The overall architecture of this method is represented in Figure 7.3. Then, a 3D encoder was used to extract features from the individual clips, which were further pairwise concatenated, and a fully connected layer was placed to predict the order of the clips. The length of the individual clips was empirically selected as 16, and the number of clips being 3 yielded the best results for the ATEC dataset. The significant advantage of this approach over the traditional OPN network is its ability to include more frames while training which helps the model learn long-term dependencies.

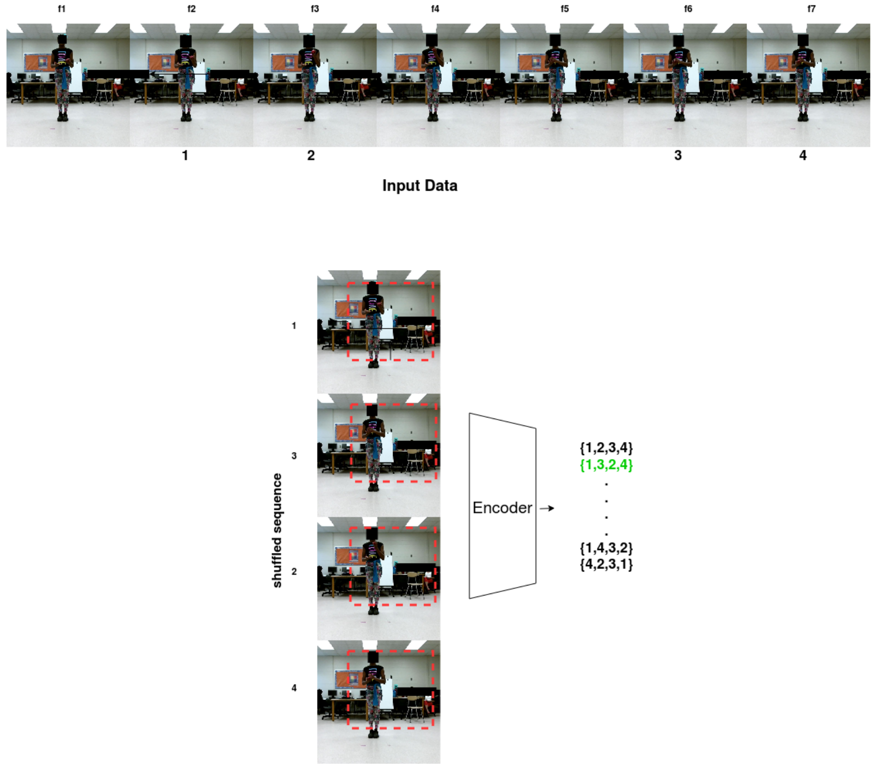


Figure 7.2. Overall architecture for Order Prediction Network (OPN).

7.3 Self-Supervised Contrastive Learning

Contrastive learning has recently become a dominant component in self-supervised learning for computer vision and other domains. It aims at generating embeddings of samples such that embeddings of similar classes should be similar to each other and so encouraged to be close to each other while trying to push away embeddings from different samples. To achieve this, a similarity metric is used to measure how close two embeddings are. For instance, one sample from the training dataset is taken from which two transformed/augmented versions are created, which are called "positive samples." During the training phase, the model aims to generate embeddings for the samples in the batch/dataset. The model encourages the embeddings of the positive samples to be close to each other while the rest of the samples (Negative samples)

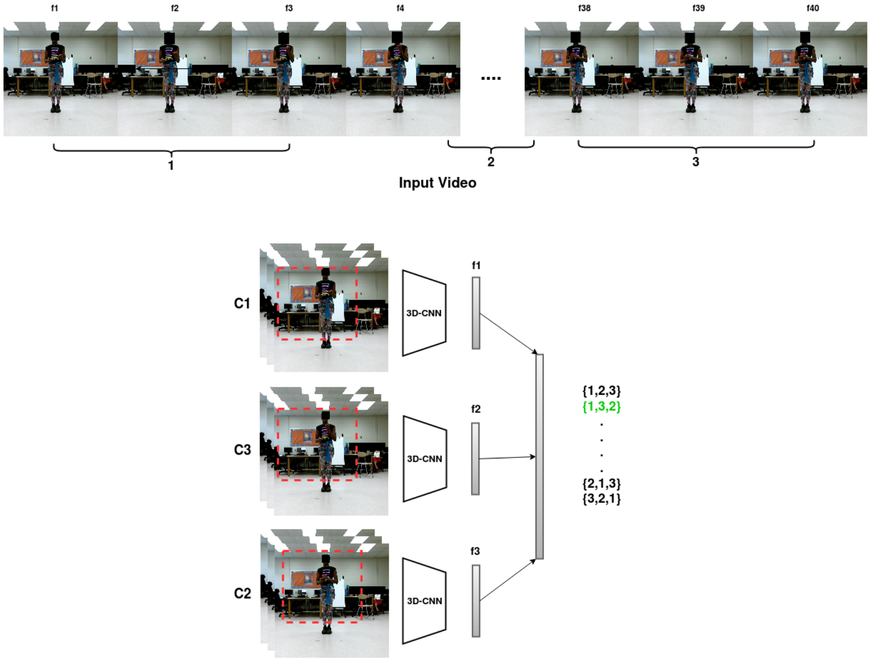


Figure 7.3. Overall architecture for Clip Prediction Network (CPN).

to be far. The model learns effective representations of the samples and is used later for transferring knowledge to downstream tasks. In Image classification, some of the recent contrastive techniques have produced results comparable to state-of-the-art results on the ImageNet dataset.

7.3.1 Data Augmentation

Given the scenario, the type of data augmentation plays a vital role during the training process. The choice of spatial and temporal augmentation is important as sometimes, changes induced by stronger augmentation could change the structure of the data such that it cannot be viewed as the same anymore. For example, learning image representation by identifying the degree of rotation of the image has been proved to be effective [138]. Still, In our experiments, we found that rotation does not work well in predicting the actions. This could also be due to the fact that the dataset has action classes such as "passing the ball from right hand to left hand" and "Passing the ball from left to right," and applying rotation-based transformation alters the meaning of the videos. Similarly, Work from Yamaguchi et al. [139] points out some drawbacks of rotation-based transformations affecting the model's ability to learn. Further, in our experiments, we found that applying spatial augmentation independent of the video frames breaks the natural motion in the scene, resulting in much poorer performance.

Hence, as part of this work, data augmentations were carefully designed to consider spatial and temporal cues. We applied temporally consistent spatial augmentation during the training process to not break the natural motion in the scene. The spatial augmentation during training contains a random factor in which transformations such as cropping, color jittering, blurring are applied. Similarly, for temporal augmentation, a straight forward approach was to take two clips from the input video being the positive pairs but, given the nature of the dataset, this approach did not work, as there is a possibility of these two video clips containing different motions which would not help in the learning process. Hence the entire video was divided into n bins where $'n'$ represents the clip length. So, to build a clip from the video, one image frame was randomly sampled from each bin for which the spatial augmentation was applied.

7.3.2 Methodology

A self-supervised contrastive architecture was built to learn effective representation as represented in Figure 7.4. The approach utilizes an infoNCE contrastive loss [140] for optimization during training.

From the equation 7.1, N represents the total number of videos considered in the batch, from where $2N$ clips are generated with augmentation. z_i and $z_i^{'}$ as the encoded

$$L_i = -log\frac{exp(sim(z_i, z_i^{'})/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq i]} exp(sim(z_i, z_i^{'}))} \qquad (7.1)$$

representations of the positive sample, and z_k represents the negative sample for the i^{th} input video. sim in the equation represents the similarity calculated between two embeddings. $sim(u, v) = u^T v/||u||_2|v||_2$ is the inner product between the two embeddings.

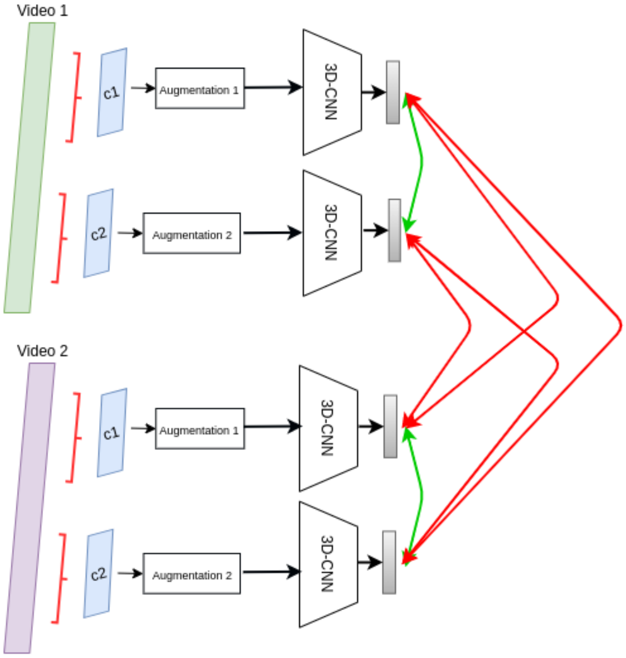
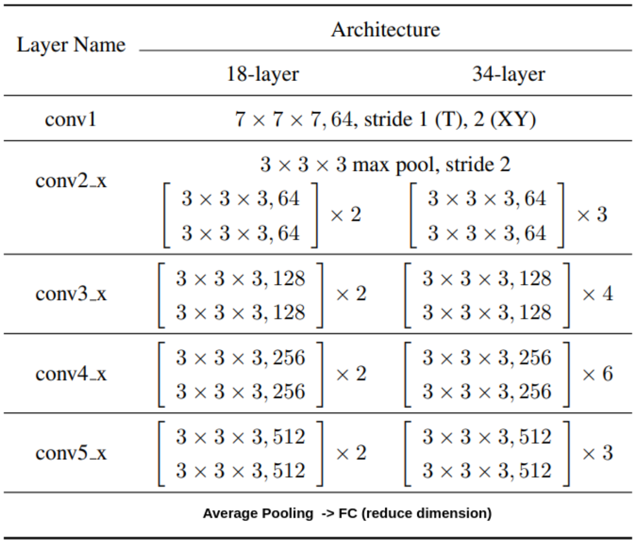


Figure 7.4. Overview of Self-supervised Contrastive approach for video representation learning.

7.3.3 Encoder

3D Convolution based architecture is used for the encoder due to its ability to capture information from the adjacent frames in contrast to the traditional 2D convolutional kernels. Specifically, a ResNet-based architecture [13] with 34 layers was used as we empirically found them to yield better results compared to the other architectures. The details of the architecture is shown in Figure 7.5. During the evaluation, only the encoder is considered where a layer replaces the final layer with the actual number of classes in the dataset followed by a softmax function.



Layer Name	Architecture	
	18-layer	34-layer
conv1	$7 \times 7 \times 7, 64$, stride 1 (T), 2 (XY)	
conv2_x	$3 \times 3 \times 3$ max pool, stride 2	
	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 3$
conv3_x	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 4$
conv4_x	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 6$
conv5_x	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 3$
Average Pooling -> FC (reduce dimension)		

Figure 7.5. 3D ResNet Architecture. Residual block are shown in brackets. Each convolutional Layer is followed by batch normalization[11] and ReLU activation function[12]. Downsampling is performed by conv3_1, conv4_1 and conv5_1 with stride of 2. The dimension of the last fully connected layer is set to the number of class in the dataset proposed by [13].

7.3.4 Downstream Task

Downstream tasks are generally used to evaluate the quality of features learned through the pretext tasks. Given that this scenario being a classification problem, the pretext task was trained using the pseudo labels to learn the video representation from which, encoder part is extracted from the architecture. The final layer of the encoder is replaced with the total number of classes in the dataset to make the model suitable for classification problem. We froze the weights of the pretrained encoder, and only the final linear layer was trained with 10 percent of the labeled data.

Table 7.1. Self-supervised approaches on the Ball Drop dataset.

Method	**Test. Acc.**
Temporal Order Verification[136]	0.60
OPN [137]	0.56
Clip Order Prediction [141]	0.69
Odd one out network [142]	0.51
ST Puzzle [143]	0.64
Self-Supervised Contrastive Approach	0.72

7.3.5 Implementation Details

For the implementation, Stochastic Gradient Descent (SGD) with a momentum was used with momentum set to 0.9. The starting learning rate was set to 0.01 and divide by 10 every time the validation loss saturates. The mentioned spatial and temporal augmentation techniques was used during the training of the pretext task. The size of the clips were empirically identified as 16 frames from the experiments. InfoNCE loss was used during the training with the temperature coefficient set to 0.1.

7.3.6 Experimental Results and Conclusion

From Table 7.1 we observed that the contrastive learning approach inspired by [144] produced the highest accuracy of 72 percent in our Ball Drop dataset, which is followed by the Clip Order prediction method produced by [141]. Similarly, for the Tandem Gait task, the Contrastive approach has produced an accuracy of 0.70, which is the highest accuracy achieved so far when compared to other self-supervised approaches.

This section concludes by discussing the results of some of the recent work in self-supervised approaches to learning video representation. Although the best results from the self-supervised approaches are comparatively lower than the supervised learning approaches discussed in the previous chapter, they have a huge advantage when it comes to learning representation from unlabelled data. Unlike the supervised approaches, which we were able to integrate with the automated tool, work is required to improve the performance of the self-supervised approaches.

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 Conclusion

The first part of this work introduces the benefits of cognitive assessment and training in the industrial workplace. First, it introduces a computer-based assessment tool that involves the physical setup yet monitored and scored by an intelligent system. Further, three different training approaches were compared to understand the effectiveness and the impact of the training approaches on the individual's learning. As a next step, extracting physiological and behavioral information from the participants while performing a cognitive task proved to help create a more personalized assessment system that could adapt based on the participant's cognitive state.

In the next part of the work, I have developed an automated and non-intrusive system to estimate cognitive development in children with multiple novel physical tasks. These systems assist cognitive therapists in diagnosing various neurodevelopmental conditions, such as Executive Function Disorder, ADHD, etc. The developed system consists of two main components, a) A simple user interface capable of recording and giving instructions simultaneously, b) An analysis module that uses state-of-the-art motion analysis and Deep learning techniques to extract information from the collected data. Preliminary data were collected from 55 children as part of the initial work, which was used to build an automated tool using state-of-the-art supervised deep learning techniques. The drawbacks of the threshold-based approach were solved with supervised deep learning approaches. Specifically, we built a multi-modal fusion method with a self-attention mechanism which provided the best results so far in our ball drop dataset. Further, we explored various self-supervised approaches to take advantage of the available un-annotated data to learn data representation. Although the results of the self-supervised approaches were comparatively lower than the supervised approaches on the ”Ball Drop” and ”Tandem Gait” dataset, it has a greater advantage in learning data representation without labels. Hence exploring self-supervised approaches in the future is definitely a promising direction in solving real world problems. To our knowledge, this is one of the few end-to-end systems that are capable of assessing various cognitive functionalities from physical tasks. Furthermore, given that annotating the data for machine learning models is time-consuming, this work explores some recent and state-of-the-art self-supervised learning approaches to build models to learn visual features to learn effective data representation. Given that these systems are extremely simple to use and low in cost, they are ideal for deploying across schools in the country towards early detection of such conditions to provide proper intervention.

8.2 Future Work

The future goals of this work can be focused on in multiple directions. First, we have implemented the system to evaluate Ball-Drop-to-the-Beat, Finger Opposition, and tandem gait as part of this work. Further, building an automated system to include more tasks mentioned in Table 4.1. Adding additional tasks to the system makes the system more accurate in estimating cognitive development. Secondly, given the data being collected frequently, to take advantage of the unlabelled data, more self-supervised approaches should be built to learn effective data representation, which will help improve the overall accuracy. Currently, work has been done to identify psychosis in adults with the proposed assessment systems.

Another functionality that can be integrated with the system is the ability to provide remote monitoring and assistance. The current setup requires a controlled environment where cameras are setup in a specific position to collect data which reduces the ability to be used in home environments. The ability to collect the data in home environments and analyze such data might be more helpful to provide remote assistance.

REFERENCES

[1] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–9, 2020.

[2] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *arXiv preprint arXiv:1710.07557*, 2017.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.

[4] D. W. Beebe, "The psychological evaluation of attention deficit hyperactivity disorder in school-aged children," in *Attention Deficit Hyperactivity Disorder*. Springer, 2005, pp. 143–163.

[5] P. D. Zelazo, J. E. Anderson, J. Richler, K. Wallner-Allen, J. L. Beaumont, and S. Weintraub, "Ii. nih toolbox cognition battery (cb): Measuring executive function and attention," *Monographs of the Society for Research in Child Development*, vol. 78, no. 4, pp. 16–33, 2013.

[6] L. M. Schmitt, S. P. White, E. H. Cook, J. A. Sweeney, and M. W. Mosconi, "Cognitive mechanisms of inhibitory control deficits in autism spectrum disorder," *Journal of Child Psychology and Psychiatry*, vol. 59, no. 5, pp. 586–595, 2018.

[7] A. R. Babu, M. Zakizadeh, J. R. Brady, D. Calderon, and F. Makedon, "An intelligent action recognition system to assess cognitive behavior for executive

function disorder," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2019, pp. 164–169.

[8] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, 2020.

[9] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," in *German Conference on Pattern Recognition*. Springer, 2018, pp. 281–297.

[10] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[12] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[13] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3154–3160.

[14] A. Rajavenkatanarayanan, V. Kanal, K. Tsiakas, D. Calderon, M. Papakostas, M. Abujelala, M. Galib, J. C. Ford, G. Wylie, and F. Makedon, "A survey of assistive technologies for assessment and rehabilitation of motor impairments in multiple sclerosis," *Multimodal Technologies and Interaction*, vol. 3, no. 1, p. 6, 2019.

[15] Y.-Y. Tang, Y. Ma, J. Wang, Y. Fan, S. Feng, Q. Lu, Q. Yu, D. Sui, M. K. Rothbart, M. Fan, *et al.*, "Short-term meditation training improves attention

and self-regulation," *Proceedings of the National Academy of Sciences*, vol. 104, no. 43, pp. 17 152–17 156, 2007.

[16] R. C. Gershon, M. V. Wagster, H. C. Hendrie, N. A. Fox, K. F. Cook, and C. J. Nowinski, "Nih toolbox for assessment of neurological and behavioral function," *Neurology*, vol. 80, no. 11 Supplement 3, pp. S2–S6, 2013.

[17] G. Stoet, "Psytoolkit: A software package for programming psychological experiments using linux," *Behavior research methods*, vol. 42, no. 4, pp. 1096–1104, 2010.

[18] M. Papakostas *et al.*, "From body to brain: Using artificial intelligence to identify user skills & intentions in interactive scenarios," Ph.D. dissertation, 2019.

[19] P. Thagard, *Mind: Introduction to cognitive science.* MIT press Cambridge, MA, 1996, vol. 4.

[20] A. Lioulemes, M. Papakostas, S. N. Gieser, T. Toutountzi, M. Abujelala, S. Gupta, C. Collander, C. D. Mcmurrough, and F. Makedon, "A survey of sensing modalities for human activity, behavior, and physiological monitoring," in *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments.* ACM, 2016, p. 16.

[21] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 716–723.

[22] L. Xia and J. Aggarwal, "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, pp. 2834–2841.

[23] J. Waleed, T. M. Hasan, and Q. K. Abed, "Eye-gaze estimation systems for multi-applications: An implementation of approach based on laptop webcam," *Diyala Journal For Pure Science*, vol. 14, no. 02, pp. 153–173, 2018.

[24] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[25] X. Li and M. C. Chuah, "Rehar: Robust and efficient human activity recognition," *arXiv preprint arXiv:1802.09745*, 2018.

[26] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv preprint arXiv:1806.11230*, 2018.

[27] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[28] E. S. Ho, J. C. Chan, D. C. Chan, H. P. Shum, Y.-m. Cheung, and P. C. Yuen, "Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments," *Computer Vision and Image Understanding*, vol. 148, pp. 97–110, 2016.

[29] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018.

[30] R. J. Kate, A. M. Swartz, W. A. Welch, and S. J. Strath, "Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data," *Physiological measurement*, vol. 37, no. 3, p. 360, 2016.

[31] A. Phinyomark, F. Quaine, S. Charbonnier, C. Serviere, F. Tarpin-Bernard, and Y. Laurillau, "Emg feature evaluation for improving myoelectric pattern recognition robustness," *Expert Systems with applications*, vol. 40, no. 12, pp. 4832–4840, 2013.

[32] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, “Deep learning for sensor-based activity recognition: A survey,” *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[33] A. Jalal, M. A. K. Quaid, and A. S. Hasan, “Wearable sensor-based human behavior understanding and recognition in daily life for smart environments,” in *2018 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2018, pp. 105–110.

[34] V. Kanal, M. Abujelala, S. Gattupalli, V. Athitsos, and F. Makedon, *APSEN: Pre-screening tool for sleep apnea in a home environment*, 2017, vol. 10287 LNCS.

[35] E. Spyrou, T. Giannakopoulos, D. Sgouropoulos, and M. Papakostas, “Extracting emotions from speech using a bag-of-visual-words approach,” in *Semantic and Social Media Adaptation and Personalization (SMAP), 2017 12th International Workshop on*. IEEE, 2017, pp. 80–83.

[36] H. Gunes, C. Shan, S. Chen, and Y. Tian, “Bodily expression for automatic affect recognition,” *Emotion recognition: A pattern analysis approach*, pp. 343–377, 2015.

[37] A. Majumder, L. Behera, and V. K. Subramanian, “Automatic facial expression recognition system using deep network-based data fusion,” *IEEE transactions on cybernetics*, vol. 48, no. 1, pp. 103–114, 2018.

[38] A. R. Babu, A. Rajavenkatanarayanan, J. R. Brady, and F. Makedon, “Multimodal approach for cognitive task performance prediction from body postures, facial expressions and eeg signal,” in *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, 2018, pp. 1–7.

[39] A. Rajavenkatanarayanan, A. R. Babu, K. Tsiakas, and F. Makedon, “Monitoring task engagement using facial expressions and body postures,” in *Proceedings*

of the 3rd International Workshop on Interactive and Spatial Computing, 2018, pp. 103–108.

[40] M. R. Wrobel, “Applicability of emotion recognition and induction methods to study the behavior of programmers,” *Applied Sciences*, vol. 8, no. 3, p. 323, 2018.

[41] S. Brás, J. H. Ferreira, S. C. Soares, and A. J. Pinho, “Biometric and emotion identification: An ecg compression based method,” *Frontiers in psychology*, vol. 9, p. 467, 2018.

[42] A. Verma, A. Dogra, K. Malik, and M. Talwar, “Emotion recognition system for patients with behavioral disorders,” in *Intelligent Communication, Control and Devices*. Springer, 2018, pp. 139–145.

[43] Q. Zhang, X. Chen, Q. Zhan, T. Yang, and S. Xia, “Respiration-based emotion recognition with deep learning,” *Computers in Industry*, vol. 92, pp. 84–90, 2017.

[44] W.-L. Zheng and B.-L. Lu, “Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks,” *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[45] A. R. Babu, A. Rajavenkatanarayanan, M. Abujelala, and F. Makedon, “Votre: A vocational training and evaluation system to compare training approaches for the workplace,” in *International Conference on Virtual, Augmented and Mixed Reality*. Springer, 2017, pp. 203–214.

[46] P. D. Gajewski, G. Freude, and M. Falkenstein, “Cognitive training sustainably improves executive functioning in middle-aged industry workers assessed by task switching: a randomized controlled erp study,” *Frontiers in human neuroscience*, vol. 11, p. 81, 2017.

[47] D. Gorecky, S. F. Worgan, and G. Meixner, "Cognito: a cognitive assistance and training system for manual tasks in industry," in *Proceedings of the 29th Annual European Conference on Cognitive Ergonomics.* ACM, 2011, pp. 53–56.

[48] B. Klimova, "Computer-based cognitive training in aging," *Frontiers in aging neuroscience*, vol. 8, p. 313, 2016.

[49] S. Webel, U. Bockholt, T. Engelke, N. Gavish, M. Olbrich, and C. Preusche, "An augmented reality training platform for assembly and maintenance skills," *Robotics and Autonomous Systems*, vol. 61, no. 4, pp. 398–403, 2013.

[50] N. Gavish, T. Gutiérrez, S. Webel, J. Rodríguez, M. Peveri, U. Bockholt, and F. Tecchia, "Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks," *Interactive Learning Environments*, vol. 23, no. 6, pp. 778–798, 2015.

[51] D. E. Barnes, K. Yaffe, N. Belfor, W. J. Jagust, C. DeCarli, B. R. Reed, and J. H. Kramer, "Computer-based cognitive training for mild cognitive impairment: results from a pilot randomized, controlled trial," *Alzheimer disease and associated disorders*, vol. 23, no. 3, p. 205, 2009.

[52] P. Pivec, "Game-based learning or game-based teaching?" 2009.

[53] S. J. Johnstone, S. J. Roodenrys, K. Johnson, R. Bonfield, and S. J. Bennett, "Game-based combined cognitive and neurofeedback training using focus pocus reduces symptom severity in children with diagnosed ad/hd and subclinical ad/hd," *International Journal of Psychophysiology*, vol. 116, pp. 32–44, 2017.

[54] J. W. Rice, "The gamification of learning and instruction: Game-based methods and strategies for training and education," *International Journal of Gaming and Computer-Mediated Simulations*, vol. 4, no. 4, 2012.

[55] A. Rajavenkatanarayanan, V. Kanal, M. Kyrarini, and F. Makedon, "Cognitive Performance Assessment based on Everyday Activities for Human-Robot

Interaction," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 398–400.

[56] M. C. Welsh, T. Satterlee-Cartmell, and M. Stine, "Towers of hanoi and london: Contribution of working memory and inhibition to performance," *Brain and cognition*, vol. 41, no. 2, pp. 231–242, 1999.

[57] T. E. Goldberg, J. A. Saint-Cyr, and D. R. Weinberger, "Assessment of procedural learning and problem solving in schizophrenic patients by tower of hanoi type tasks." *The Journal of Neuropsychiatry and Clinical Neurosciences*, 1990.

[58] M. Bustini, P. Stratta, E. Daneluzzo, R. Pollice, P. Prosperini, and A. Rossi, "Tower of hanoi and wcst performance in schizophrenia: problem-solving capacity and clinical correlates," *Journal of Psychiatric Research*, vol. 33, no. 3, pp. 285–290, 1999.

[59] G. Kielhofner, *A model of human occupation: Theory and application.* Lippincott Williams & Wilkins, 2002.

[60] L. Pessoa, *The cognitive-emotional brain: From interactions to integration.* MIT press, 2013.

[61] C. D. Salzman and S. Fusi, "Emotion, cognition, and mental state representation in amygdala and prefrontal cortex," *Annual review of neuroscience*, vol. 33, pp. 173–202, 2010.

[62] R. Pekrun, "The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators," *Applied Psychology*, vol. 41, no. 4, pp. 359–376, 1992.

[63] ——, "The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice," *Educational psychology review*, vol. 18, no. 4, pp. 315–341, 2006.

[64] R. Pekrun and L. Linnenbrink-Garcia, "Academic emotions and student engagement," in *Handbook of research on student engagement.* Springer, 2012, pp. 259–282.

[65] J. J. Van Merrienboer and J. Sweller, "Cognitive load theory and complex learning: Recent developments and future directions," *Educational psychology review*, vol. 17, no. 2, pp. 147–177, 2005.

[66] B. Weiner, "Theories of motivation: From mechanism to cognition." 1972.

[67] M. Bannert, "Managing cognitive load—recent trends in cognitive load theory," *Learning and instruction*, vol. 12, no. 1, pp. 139–146, 2002.

[68] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[69] N. Fourati and C. Pelachaud, "Emilya: Emotional body expression in daily actions database." in *LREC*, 2014, pp. 3486–3493.

[70] N. Bianchi-Berthouze, P. Cairns, A. Cox, C. Jennett, and W. W. Kim, "On posture as a modality for expressing and recognizing emotions," in *Emotion and HCI workshop at BCS HCI London*, 2006.

[71] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[72] M. M. Gross, E. A. Crane, and B. L. Fredrickson, "Methodology for assessing bodily expression of emotion," *Journal of Nonverbal Behavior*, vol. 34, no. 4, pp. 223–248, 2010.

[73] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[74] M. Papakostas, K. Tsiakas, T. Ginnakopoulos, and F. Makedon, "Towards predicting task performance from eeg signals."

[75] C.-Y. Chen, J. P. Lawlor, A. K. Duggan, J. B. Hardy, and W. W. Eaton, "Mild cognitive impairment in early life and mental health problems in adulthood," *American Journal of Public Health*, vol. 96, no. 10, pp. 1772–1778, 2006.

[76] J. H. Bernstein and D. P. Waber, "Executive capacities from a developmental perspective," *Executive function in education: From theory to practice*, pp. 39–54, 2007.

[77] C. Hughes and A. Graham, "Measuring executive functions in childhood: Problems and solutions?" *Child and adolescent mental health*, vol. 7, no. 3, pp. 131–142, 2002.

[78] T. Shallice and P. W. Burgess, "Deficits in strategy application following frontal lobe damage in man," *Brain*, vol. 114, no. 2, pp. 727–741, 1991.

[79] D. J. Ackerman and A. H. Friedman-Krauss, "Preschoolers' executive function: Importance, contributors, research needs and assessment options," *ETS Research Report Series*, vol. 2017, no. 1, pp. 1–24, 2017.

[80] M. G. Sim, E. Khong, G. Hulse, *et al.*, "When the child with adhd grows up," *Australian family physician*, vol. 33, no. 8, p. 615, 2004.

[81] M. Adamou, M. Arif, P. Asherson, T.-C. Aw, B. Bolea, D. Coghill, G. Gujónsson, A. Halmøy, P. Hodgkins, U. Müller, *et al.*, "Occupational issues of adults with adhd," *BMC psychiatry*, vol. 13, no. 1, pp. 1–7, 2013.

[82] R. A. Barkley, "Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of adhd." *Psychological bulletin*, vol. 121, no. 1, p. 65, 1997.

[83] C. Dendy, "Executive function... "what is this anyway?"," 2008.

[84] M. S. Atkins, W. E. Pelham, and M. H. Licht, “A comparison of objective classroom measures and teacher ratings of attention deficit disorder,” *Journal of abnormal child psychology*, vol. 13, no. 1, pp. 155–167, 1985.

[85] B. A. Eriksen and C. W. Eriksen, “Effects of noise letters upon the identification of a target letter in a nonsearch task,” *Perception & psychophysics*, vol. 16, no. 1, pp. 143–149, 1974.

[86] M. R. Rueda, J. Fan, B. D. McCandliss, J. D. Halparin, D. B. Gruber, L. P. Lercari, and M. I. Posner, “Development of attentional networks in childhood,” *Neuropsychologia*, vol. 42, no. 8, pp. 1029–1040, 2004.

[87] C. L. Davis and S. Cooper, “Fitness, fatness, cognition, behavior, and academic achievement among overweight children: do cross-sectional associations correspond to exercise trial outcomes?” *Preventive medicine*, vol. 52, pp. S65–S69, 2011.

[88] E. E. Davis, N. J. Pitchford, and E. Limback, “The interrelation between cognitive and motor development in typically developing children aged 4–11 years is underpinned by visual processing and fine manual control,” *British Journal of Psychology*, vol. 102, no. 3, pp. 569–584, 2011.

[89] J. E. Donnelly and K. Lambourne, “Classroom-based physical activity, cognition, and academic achievement,” *Preventive medicine*, vol. 52, pp. S36–S42, 2011.

[90] D. P. Van Dusen, S. H. Kelder, H. W. Kohl III, N. Ranjit, and C. L. Perry, “Associations of physical fitness and academic performance among schoolchildren,” *Journal of School Health*, vol. 81, no. 12, pp. 733–740, 2011.

[91] M. E. Hopkins, F. C. Davis, M. R. VanTieghem, P. J. Whalen, and D. J. Bucci, “Differential effects of acute and regular physical exercise on cognition and affect,” *Neuroscience*, vol. 215, pp. 59–68, 2012.

[92] B. E. Wexler, "Integrated brain and body exercises for adhd and related problems with attention and executive function," *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, vol. 5, no. 3, pp. 10–26, 2013.

[93] M. M. McClelland, C. E. Cameron, R. Duncan, R. P. Bowles, A. C. Acock, A. Miao, and M. E. Pratt, "Predictors of early growth in academic achievement: The head-toes-knees-shoulders task," *Frontiers in psychology*, vol. 5, p. 599, 2014.

[94] D. R. Becker, M. M. McClelland, P. Loprinzi, and S. G. Trost, "Physical activity, self-regulation, and early academic achievement in preschool children," *Early Education & Development*, vol. 25, no. 1, pp. 56–70, 2014.

[95] B. Muppala, K. Tsiakas, C. Fleury, A. Weinstein, and M. D. Bell, "1.39 activate test of embodied cognition (atec): A new automated assessment system using cognitively demanding physical tasks to assess development of executive function," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 58, no. 10, p. S159, 2019.

[96] S. Gattupalli, A. R. Babu, J. R. Brady, F. Makedon, and V. Athitsos, "Towards deep learning based hand keypoints detection for rapid sequential movements from rgb images," in *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, 2018, pp. 31–37.

[97] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[98] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[99] D. Victor, "Real-time hand tracking using ssd on tensorflow," https://github.com/victordibia/handtracking, 2017.

[100] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[101] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," 2017.

[102] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[103] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[104] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" 2017.

[105] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017.

[106] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[107] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[108] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[109] A. Ranjan and M. J. Black, “Optical flow estimation using a spatial pyramid network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4161–4170.

[110] A. Dillhoff, K. Tsiakas, A. R. Babu, M. Zakizadehghariehali, B. Buchanan, M. Bell, V. Athitsos, and F. Makedon, “An automated assessment system for embodied cognition in children: from motion data to executive functioning,” in *Proceedings of the 6th international Workshop on Sensor-based Activity Recognition and Interaction*, 2019, pp. 1–6.

[111] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “Openpose: realtime multi-person 2d pose estimation using part affinity fields,” in *arXiv preprint arXiv:1812.08008*, 2018.

[112] H. Naruse, T. X. Fujisawa, C. Yatsuga, M. Kubota, H. Matsuo, S. Takiguchi, S. Shimada, Y. Imai, M. Hiratani, H. Kosaka, *et al.*, “Increased anterior pelvic angle characterizes the gait of children with attention deficit/hyperactivity disorder (adhd),” *PLoS one*, vol. 12, no. 1, p. e0170096, 2017.

[113] M. Kocabas, N. Athanasiou, and M. J. Black, “Vibe: Video inference for human body pose and shape estimation,” 2019.

[114] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[115] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *arXiv preprint arXiv:1406.2199*, 2014.

[116] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[117] A. Ramesh Babu, M. Z. Zadeh, A. Jaiswal, A. Lueckenhoff, M. Kyrarini, and F. Makedon, "A multi-modal system to assess cognition in children from their physical movements," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 6–14.

[118] W. Wang, J. Zhang, C. Si, and L. Wang, "Pose-based two-stream relational networks for action recognition in videos," *arXiv preprint arXiv:1805.08484*, 2018.

[119] Z. Cai, H. Neher, K. Vats, D. A. Clausi, and J. Zelek, "Temporal hockey action recognition via pose and optical flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[120] G. Kapidis, R. Poppe, E. Van Dam, L. Noldus, and R. Veltkamp, "Egocentric hand track and object-based human action recognition," in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI).* IEEE, 2019, pp. 922–929.

[121] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European conference on computer vision.* Springer, 2004, pp. 25–36.

[122] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5253–5263.

[123] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[124] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[125] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML’10. Madison, WI, USA: Omnipress, 2010, p. 807–814.

[126] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, “Two stream lstm: A deep fusion framework for human action recognition,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV).* IEEE, 2017, pp. 177–186.

[127] M. Maghoumi and J. J. LaViola Jr, “Deepgru: Deep gesture recognition utility,” in *International Symposium on Visual Computing.* Springer, 2019, pp. 16–31.

[128] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, “Attnsense: multi-level attention mechanism for multimodal human activity recognition,” in *Proceedings of the 28th International Joint Conference on Artificial Intelligence.* AAAI Press, 2019, pp. 3109–3115.

[129] X. Liu, F. Zhang, Z. Hou, Z. Wang, L. Mian, J. Zhang, and J. Tang, “Self-supervised learning: Generative or contrastive,” *arXiv preprint arXiv:2006.08218*, vol. 1, no. 2, 2020.

[130] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A survey on contrastive self-supervised learning,” *Technologies*, vol. 9, no. 1, p. 2, 2021.

[131] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[132] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European conference on computer vision.* Springer, 2016, pp. 649–666.

[133] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[134] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *European conference on computer vision.* Springer, 2016, pp. 69–84.

[135] M. Z. Zadeh, A. R. Babu, A. Jaiswal, and F. Makedon, “Self-supervised human activity recognition by augmenting generative adversarial networks,” *arXiv preprint arXiv:2008.11755*, 2020.

[136] I. Misra, C. L. Zitnick, and M. Hebert, “Shuffle and learn: unsupervised learning using temporal order verification,” in *European Conference on Computer Vision.* Springer, 2016, pp. 527–544.

[137] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang, “Unsupervised representation learning by sorting sequences,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 667–676.

[138] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” *arXiv preprint arXiv:1803.07728*, 2018.

[139] S. Yamaguchi, S. Kanai, T. Shioda, and S. Takeda, “Multiple pretext-task for self-supervised learning via mixing multiple image transformations,” *arXiv preprint arXiv:1912.11603*, 2019.

[140] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

[141] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang, “Self-supervised spatiotemporal learning via video clip order prediction,” in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 334–10 343.

[142] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-supervised video representation learning with odd-one-out networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3636–3645.

[143] D. Kim, D. Cho, and I. S. Kweon, "Self-supervised video representation learning with space-time cubic puzzles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8545–8552.

[144] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," *arXiv preprint arXiv:2008.03800*, 2020.

BIOGRAPHICAL STATEMENT

Ashwin Ramesh Babu was born in Salem, India, in 1992. In 2014, he received his B.Tech in Information Technology from Anna University, India and his M.S degree from the University of Texas at Arlington, USA in 2016.

In September 2016, he joined the HERACLEIA Human-Centered Computing Laboratory at the University of Texas at Arlington as a Ph.D. student. He participated as a Graduate Research Assistant in multiple NSF-funded projects under the supervision of Prof. Fillia Makedon. Since 2018, he led the NSF-funded project from the Cyber Human Systems (CHS) program with the title *"Computational Science for Improving Assessment of Executive Function in Children"*. Further, he served as a "Graduate Teaching Assistant" for courses such as "Introduction to Programming" and "Advanced Topics in Human-computer Interaction" classes. During summer 2019, Ashwin worked as a Deep Learning/Machine Learning Intern at Hewlett Packard Enterprise, where he was awarded for "Best Technical Presentation" for his contribution as an intern.

In addition, Ashwin received the "Outstanding Doctoral Dissertation Award 2021" by the Computer Science and Engineering Department at University of Texas at Arlington. During his time at HERACLEIA lab, he co-authored several peer-reviewed papers published in technical conferences and has served as a reviewer in several international conferences and journals. Ashwin's research interests revolve around applying Deep Learning and Computer vision in healthcare and cognitive science, specifically towards data-driven personalization.