TOWARD DIGITAL PHENOTYPING: HUMAN ACTIVITY REPRESENTATION FOR EMBODIED COGNITION ASSESSMENT

A Dissertation

Presented to the Faculty of the Graduate School

of University of Texas at Arlington

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Mohammad Zaki Zadeh Ghariehali

May 2023

TOWARD DIGITAL PHENOTYPING: HUMAN ACTIVITY REPRESENTATION FOR EMBODIED COGNITION ASSESSMENT

Mohammad Zaki Zadeh Ghariehali, Ph.D.

University of Texas at Arlington 2023

Cognition is the mental process of acquiring knowledge and understanding through thought, experience and senses. Based on Embodied Cognition theory, physical activities are an important manifestation of cognitive functions. As a result, they can be employed to both assess and train cognitive skills. In order to assess various cognitive measures, the ATEC system has been proposed. It consists of physical exercises with different variations and difficulty levels, designed to provide assessment of executive and motor functions.

This thesis focuses on obtaining human activity representation from recorded videos of ATEC tasks in order to automatically assess embodied cognition performance. Representation learning is a collection of methods that allows a model to be fed with raw data and to automatically encode the representations needed for downstream task like activity recognition. Both supervised and self-supervised approaches are employed in this work, But the emphasis is on the latter which can exploit a small set of annotated data to obtain an effective representation. The performance of different self-supervised approaches are investigated for automated cognitive assessment of children performing ATEC tasks.

This effort is the first step toward building a comprehensive digital phenotyping framework that can collect multi-modal data from variety of sensors such as cameras, wearables, etc., for monitoring human behaviour. Digital phenotyping is the moment by moment, quantification of the individual-level hu-

man phenotype using data from personal digital devices. Digital phenotyping will close the loop between detecting clinical phenomena and taking action by using data to trigger and deliver personalized digital treatment or prevention interventions.

COMMITTEE MEMBERS

The members of the Committee approve the doctoral dissertation of Mohammad Zaki Zadeh Ghariehali.

Supervisor Professor
Fillia Makedon

Vassilis Athitsos

William Beksi

Shirin Nilizadeh

Dean of the Graduate School

ACKNOWLEDGEMENTS

I'd like to thank my supervisor, Dr. Fillia Makedon, for giving me this opportunity and for believing in me. I'd also like to thank my Heracleia Lab colleagues and labmates.

Dr. Morris Bell, Dr. Glenn Wylie, and Dr. Maher Abujelala deserve special recognition for their invaluable assistance in gathering the data used in my thesis. Furthermore, I would like to thank the members of my Ph.D. supervising committee, Dr. Vassilis Athitsos, Dr. William Beksi, and Dr. Shirin Nilizadeh, for their encouragement and valuable insights throughout my Ph.D.

Last but not least, I'd like to thank my parents for their unwavering support and encouragement throughout my Ph.D. journey. It would not have been possible without their assistance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This thesis focuses on extracting human activity representations from recorded video in order to automatically assess embodied cognition performance. This is the first step toward developing a comprehensive digital phenotyping framework capable of collecting multi-modal data from a variety of sensors in order to monitor human behavior and recommend appropriate treatments. In this work, both supervised and self-supervised approaches are used. Supervised methods require all collected data to be carefully annotated by experts, whereas self-supervised methods can use a small set of annotated data to obtain an effective representation. Some of the concepts mentioned above are further explained in the sections that follow.

1.1 Digital Phenotyping

In genetics, phenotype refers to an organism's composite observable traits. Digital phenotyping is quantification of the individual-level human phenotype on a moment-by-moment basis utilizing data from personal digital devices. Its goal is to monitor and measure human behavior using data generated and collected automatically by smartphones, cameras, wearables, and other connected devices [51, 48].

Digital phenotyping will close the loop between detecting clinical phenomena and taking action by using data to trigger, tailor and deliver personalized digital treatment or prevention interventions as shown in Figure 1.1. The multimodal data is recorded from participants based on studies defined by specialists, as shown in this diagram. This data is then examined using an automated algorithm to help experts modify the experiment accordingly [51, 48].

This concept of the extended phenotype is supported by the growth and evolution of digital devices and sensors. There is a growing collection of health-related data that can impact assessments of human illness thanks to social media, wearable devices, and camera sensors. In the discipline of psychiatry, it's difficult to overestimate the value of data-driven, objective measurements of individual behavior. Previously, psychiatrists relied almost entirely on self-reports of mental health symptoms, which have few biological markers and unclear diagnostic categories [48].

In order to identify individuals at risk for depression, anxiety, or even suicide, digital phenotyping can include passive monitoring of activity changes using an accelerometer, phone usage statistics, and natural language analysis of social media posts. The online data gathering can also be used to detect changes in the condition and relapse early on. Finally, medical professionals can use this information to design interventions and treatments for individuals. The data, for example, can be utilized to adapt therapy for people with depression in order to boost their engagement and treatment effectiveness [48].

Accelerating progress in this paradigm involves scalable data collection infrastructure that addresses equity and privacy concerns, machine learning-based data processing and analysis methodologies, and the development of data quality validation tools that address bias and noise concerns.

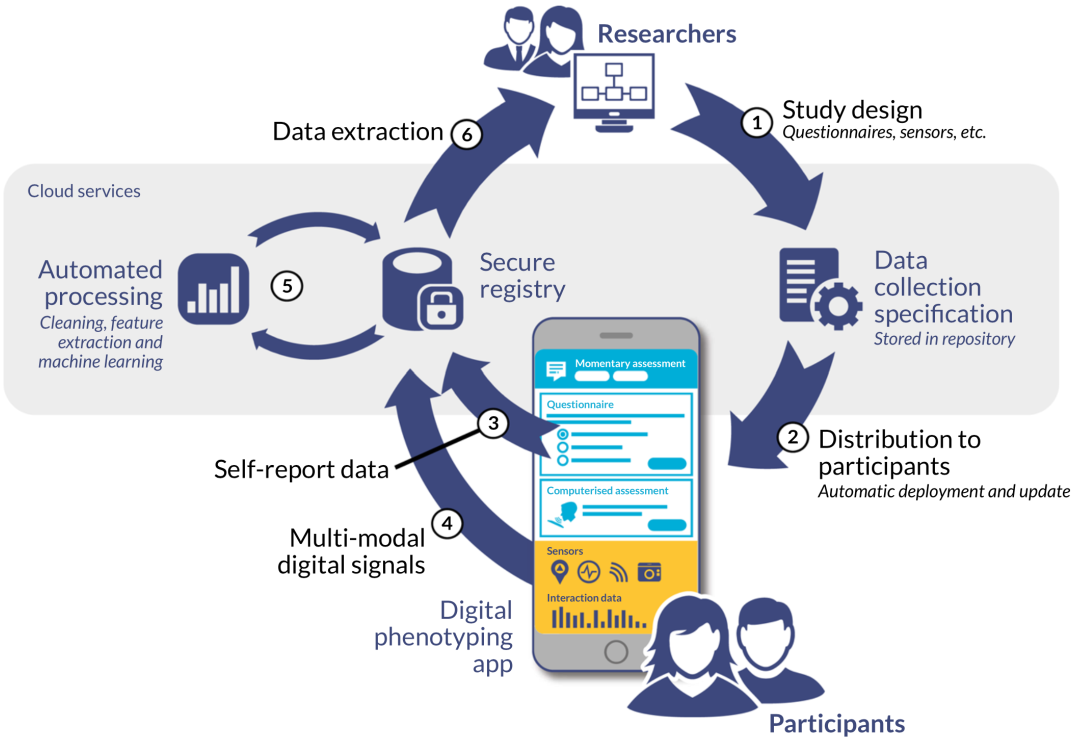


Figure 1.1: An integrated multi user platform for digital phenotyping adopted from [48]. the multi-modal data is recorded from participants based on some studies defined by experts. Then this data is analyzed automatically and the resulting insight would guide experts to recommend diagnosis or further data collection.

1.2 Cognition

Cognition is defined as the mental actions or process of acquiring knowledge and understanding through thought, experience and senses [24, 27, 27]. There are many different types of cognitive processes that include:

Attention: The cognitive process of selectively focusing on one piece of information while dismissing other perceptible data.

Language: The ability to comprehend and express oneself through spoken and written language, allowing humans to communicate with one another.

Learning: The cognitive processes involved in receiving new data, synthesizing information, and integrating it with prior knowledge.

Memory: The faculty of the brain by which data or information is encoded, stored, and retrieved when needed. It is the systematic collection of data through time with the goal of influencing future behavior. [103]

Thought: An essential part of every cognitive process that lets people to engage in decision-making, problem-solving, and higher reasoning. [24]

Everything in a person's daily life, including their general health, is influenced by cognitive processes. Every piece of information that humans pick up from their surroundings has to be transformed into signals that their brain can comprehend. It is critical that the world's experience be simplified to the fundamentals in order for the brain to comprehend all of this incoming information. Furthermore, cognition encompasses not just what happens within people's heads, but also how their thoughts and mental processes influence their behaviors. Humans' behavior and interactions with the environment are influenced by their awareness of the world around them, their recollections of past events, and their understanding of language [24]. Physical activities can be used to both measure and teach cognitive skills, as they are a significant expression of cognitive functions. This will be described further in the next subsection.

1.2.1 Embodied Cognition

Embodied cognition is the theory that many characteristics of cognition, are shaped by the entire body of the organism [1]. High-level mental constructs such as concepts, as well as performance on various cognitive tasks such as reasoning and judgment, are among the cognition aspects. The motor system, the perceptual system, physical interactions with the environment, and the world assumptions built into the organism's structure are all examples of physical features [1].

Embodied cognition has a short history, having been presented by philosophers Martin Heidegger, Maurice Merleau Ponty, and John Dewey in the early twentieth century. Embodied cognition has also been empirically researched in recent decades. Early scientists suggested that cognition could be represented using formal logic, and that the brain might be viewed as a digital computing unit. To put it another way, the mind was viewed as a separate computer program from the body, with the brain serving as general-purpose hardware [81]. Chomsky's theory of language as a series of meaningless symbols fit this paradigm [25].

Cognitive linguists, such as Lakoff, argued that semantics evolved from the nature of the body [66]. Their research looked into how, when, and why people use metaphors. Humans, for example, understand having control as ″Up″ and being subject to control as ″Down″ by using sentences such as ″I have control over him,″ ″I am on top of the situation,″ and so on. Love is also described as a physical force: ″I could feel the electricity between us″ and ″They gravitated to each other immediately″ [66, 81].

This concept not only demonstrated how common metaphors are in ordinary language, but it also asserted that many key foundations of Western philosophy, such as the notions of reason and language being separate from the body, were erroneous. To summarize, it was proposed that the ordinary conceptual framework of humans is basically metaphorical [81].

This notion was expanded by claiming that philosophical concepts are also formed metaphorically. They asserted that the mind is inherently embodied, that abstract concepts are mostly metaphorical, and that the majority of thought is unconscious. As a result, because cognition is grounded in bodily experience, reason does not rest on abstract laws. For example, thinking about the future causes people to lean forward slightly, whereas thinking about the past causes them to lean back slightly because the future is ahead [67]. This research focuses primarily on executive functions, which coordinate, integrate, and control cognition processes. The following subsection will provide a brief overview of the Executive Function.

Despite the evidence for the functional importance of embodied cognition, neuro-cognitive evaluation methodologies have remained largely unchanged. Neuro-psychologists use seated activities to measure attention, memory, reasoning, and other executive function (EF) tests, whether they are trained on The Halstead-Reitan battery or the clinically more flexible Boston Process Approach (e.g., Edith Kaplan) [12].

1.3 Executive Functions and ADHD

Executive functions are necessary for high-order problem solving and goal-directed behavior [13]. Inhibitory control, cognitive flexibility, and working memory are the three primary areas of executive functions. Executive functions disorders in working memory, cognitive flexibility, response inhibition, multiple simultaneous attention, and planning and sequencing, can cause lifetime issues in academic success, employment, relationship development, and community participation [13, 49, 102].

Attention Deficiency Hyperactivity disorder (ADHD) is a psychiatric neurodevelopmental illness characterized by cognitive deficits, particularly in executive functions. ADHD is common in children and young adolescents, beginning around the age of six, and affects boys three times more than girls. Low academic success, grade retention, school suspensions and expulsions, poor peer and family relations, anxiety and depression, aggression, early substance addiction, driving accidents, and marital and work challenges are all linked to ADHD.

Researchers discovered in 2009 that adolescents with ADHD had slower brain maturation than their classmates, and that the area of the brain that allows pupils to engage on dull repetitive tasks, such as schoolwork, has a lower amount of dopamine receptors and transporters. It explains why children can play video games for long periods of time but struggle to finish their schoolwork on time. As a result, it's critical to increase our understanding of executive functions and create better techniques for assessing children's cognitive abilities. [11].

1.4 Cognitive Assessment

Many studies have been undertaken to better understand the relationship between cognitive deficiencies and various psychiatric neuro-developmental diseases, and a number of diagnostic and management approaches have been recommended. Traditionally, a diagnosis begins with obtaining comprehensive background information from children, parents, and school teachers, followed by trained psychologists administering standardized tests and a feedback session on performance to explain the findings and make recommendations for possible treatments or interventions.

One of the most popular paper-based cognitive assessment test is The Swanson, Nolan and Pelham Teacher and Parent Rating Scale (SNAP). It 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 [6]. Each question essentially counts the number of times a particular symptom or behavior occurs. The survey is intended for use by children and young adults. The findings shed light on inattention, hyperactivity, impulsivity, and other factors.

1.4.1 Computerized Tests

Computerized tests for cognitive assessment provide the advantages of speed, accuracy, and low cost. Computerized tests have several advantages over paper-based tests, including standardized test administration over a wide range of participants, automatic scoring and reporting, and self-paced instructions.

Also Computerized assessments provide a consistent quantitative measurement of performance, allowing for more regular evaluations of cognitive function. In some circumstances, computerized examinations can be conducted at home for a low cost [125, 107, 54].

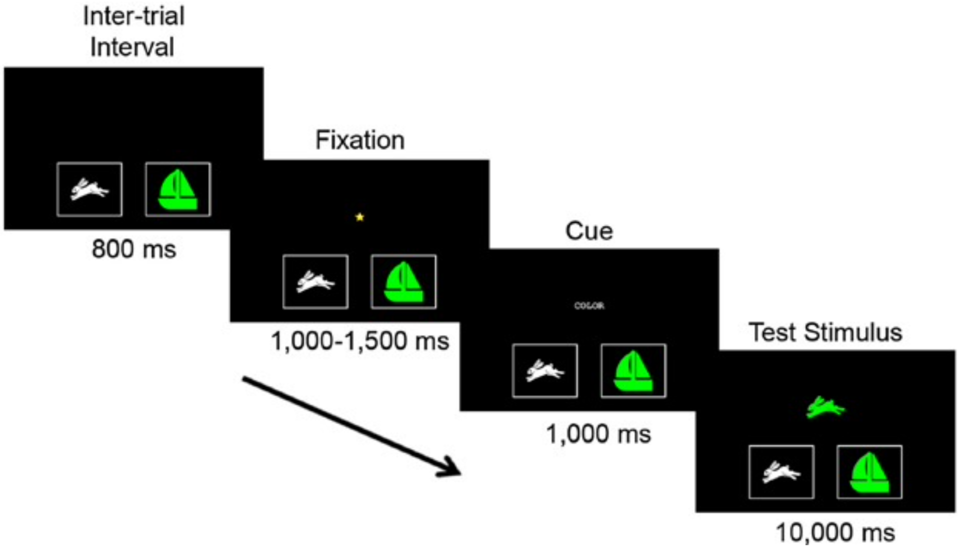


Figure 1.2: Trial sequence for the Dimensional Change Card Sort test [125].

One commonly used computerized test for cognitive assessment is the Dimensional Change Card Sort test [125]. The purpose of this exam is to assess cognitive flexibility, often known as task switching or set shifting. This test, created by Zelazo and colleagues and based on Luria's work on rule use, has been widely used to investigate the development of cognition in children. Children are shown two target cards (e.g., a blue rabbit and a red boat) and instructed to sort a sequence of bivalent test cards (e.g., red rabbits and blue boats) first according to one dimension (e.g., color), then according to the other (e.g., shape) in the standard version of the DCCS. Both the standard and more difficult versions of this task have excellent test–retest reliability in children [125] (Figure 1.2).

1.4.2 Physical Tests

Because physical activities are an important manifestation of cognitive functions [31], physical activities can be employed to both assess cognitive skills and to train such skills [32]. Physical activities should be incorporated into cognitive training since research shows that physical fitness and exercise in children leads to quantifiable increases in cognitive skills and academic achievement [26]. Understanding the relationship between physical manifestations of cognitive skills and other sorts of manifestations, such as response to computer-based problem-solving activities, is still an open problem [46] which tackling it is the main focus of this work.

The difficulty of assessing performance in physical activities is a fundamental impediment to advancing this understanding. The tasks should be designed in such a way that their cognitive demands correspond to those imposed by computer-based training activities. As a result, these physical activities can be used to improve sustained attention, self-control, working memory, and cognitive flexibility [118].

1.5 The Role of Machine Learning

It's difficult to overestimate the impact of machine learning technologies on modern society. Machine learning algorithms are used to recognize objects in photos, convert speech to text, match posts or products to users' interests, and select appropriate search results. Increasingly, these applications make use of a brand of methods called deep learning [68, 64].

Supervised learning is the most widely used approach in machine learning. A massive dataset of photos of buildings, cars, and people, each labeled with its category, is produced in order to design a system that can classify images. The machine learning model is presented an image during training and produces a vector of scores, one for each category. The true category should have the highest score of all the categories, however this rarely occurs prior to training. As a result, an objective function is created that calculates the difference between the output scores and the intended score pattern. To mitigate this error, the model adjusts its internal adjustable parameters (weights).

In practice, a procedure called stochastic gradient descent (SGD) is used for adjusting the model weights.It entails feeding the model the input vector and computing the outputs and errors. The weights are then modified based on the average gradients computed. This process is repeated for many small sets of examples (mini batches) from the training set until it converges. It is called stochastic because each mini batch gives a noisy estimate of the average gradient over all examples. When compared to significantly more advanced optimization techniques, this simple procedure frequently succeeds in obtaining a satisfying set of weights remarkably quickly. After training, the system's performance is evaluated using a different collection of examples known as a test set to determine the machine's generalization capabilities [68].

The ability of traditional machine learning approaches to analyze natural data in its raw form was limited. As a result, designing a feature extractor (encoder) that turned raw data such as pixel values of an image into an appropriate internal representation or feature vector required rigorous engineering and extensive domain experience. The resulting feature vector would then be processed by a learning module (classifier) to detect or classify patterns in the input.

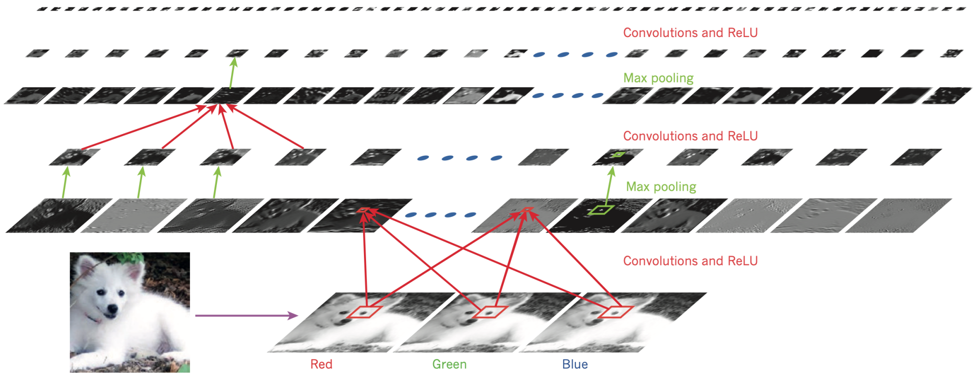


Figure 1.3: Inside a convolutional network: Each rectangular image is a feature map corresponding to the output for one of the learned features, detected at each of the image positions [68].

Representation learning is a group of techniques that enable a model to be fed raw data and automatically encode the representations required for detection or classification. Deep learning methods are special case of representation learning employing multiple levels of representation, obtained by composing simple but non-linear layers that each transform the representation at one level (starting with the raw input) into a representation at a higher and more abstract level. Very complex functions can be learned by conforming to a sufficient number of such transformations. Higher layers of representation accentuate characteristics of the input that are important for discriminating while suppressing irrelevant variations in classification tasks [68].

Convolutional neural net (CNN) [64] is an example of deep learning methods frequently used for acquiring representation of images (Figure 1.3). After an image is fed into a CNN, the first layer of representation's learnt features typically represent the presence or absence of edges at specific orientations and positions in the image. The second layer recognizes motifs by recognizing specific edge configurations, despite of minor differences in edge placements. TThe third layer can put motifs together in larger groups that resemble portions of recognized items. Finally, further layers recognize objects as assemblages of these pieces. Deep learning is distinguished by the fact that these layers of representations are learned from data using a general-purpose learning technique rather than by human engineers [68].

1.5.1 Self-supervised Learning

Due to the immense effort required in manually annotating millions of data samples, the supervised technique to learning features from annotated data has virtually hit its limit. This is because most modern supervised computer vision systems attempt to learn some type of image representation by searching big datasets for a pattern between data points and their annotations.

Despite the abundance of data available on the internet, the lack of annotations has compelled researchers to seek for alternate methods for utilizing it. Self-supervised methods are at the forefront of efforts to adapt deep learning methods to learn feature representations without costly annotations [60]. In other words, the data itself provides the supervision in self-supervised learning.

In order to learn the underlying representations from unlabeled data, self-supervised learning algorithms have included both generative and contrastive approaches. Creating numerous pretext assignments that aid in learning features using pseudo labels has been a common strategy. Tasks such as context prediction [30], image inpainting [90], colorizing grey-scale images [128], solving jigsaw puzzles [86, 111], counting objects [87] and video frame prediction [89, 70, 69] have proven to be effective for learning effective representations.

Generative models gained a great deal of popularity after the introduction of Generative Adversarial Networks (GANs) [39, 93]. The work later became the foundation for many successful architectures such as DiscimNet [4], Self-Attention GAN [127], Rot-GAN [22] and FutureGAN [5]. These methods inspired more researchers to switch to training deep learning models with unlabeled data in a self-supervised setting.

Contrastive learning (CL) is a discriminative method for grouping similar samples together and separating dissimilar samples. A contrastive loss is calculated for computer vision problems using feature representations retrieved from an encoder network. For example, one sample from the training dataset is taken and a transformed version of the sample is obtained by applying appropriate data augmentation techniques. During training, the augmented version of the original sample is considered as a positive sample, and the rest of the samples in the batch/dataset (depends on the method being used) are considered negative samples. Following that, the model is trained to learn to distinguish positive from negative samples. As a result, the model learns input representations that can be employed in downstream tasks like activity recognition or object detection as shown in Figure 1.4 [52, 73, 113, 109, 92].

In order to pull similar instances closer and push away dissimilar instances from each other, a similarity metric that measures the closeness between the representations of two instances is employed. The most common similarity metric used is cosine similarity that acts as a basis for different contrastive loss functions. The cosine similarity of two variables (vectors) is the cosine of the angle between them [52].

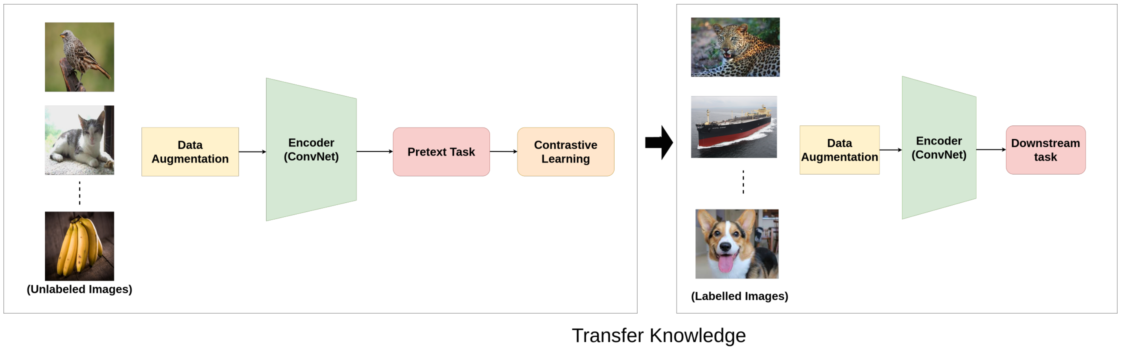


Figure 1.4: An overview of contrastive learning in practice [52].

Contrastive learning focuses on comparing the representations with a variant of Noise Contrastive Estimation function [41] called InfoNCE [113] that is defined as follows:

$$L = -log \frac{exp(sim(q, k_+)/\tau)}{exp(sim(q, k_+)/\tau) + \sum_{i=0}^{K} exp(sim(q, k_i)/\tau)} \tag{1.1}$$

where q is the original sample, k_+ represents a positive sample, and k_i represents a negative sample. τ is a hyper-parameter used in most of the recent methods and is called temperature coefficient. The *sim* function can be any similarity function, but generally a cosine similarity is used as mentioned earlier. The initial idea behind Noise Contrastive Estimation was to perform a non-linear logistic regression that discriminates between observed data and some artificially generated noise [52].

CHAPTER 2

ATEC: ACTIVATE TEST FOR EMBODIED COGNITION

The ATEC system has been proposed in order to examine multiple cognitive parameters of children, such as working memory, reaction inhibition, and coordination, using physical exercises [12, 29, 8]. This system was created to allow both professionals and non-experts to handle it with ease. The ATEC system features a recording and administration interface that were created to keep the assessments running smoothly. Because sensor-based data collection is more expensive and impractical with children, this system simply records video data.

The participants' front and side views were recorded using two Microsoft Kinect V2 cameras. RGB, depth, audio, and skeleton data are all recorded. The recording modules are linked to an Android-based administrative interface that manages the assessment's flow and allows the administrator to choose between tasks. Figure 2.1 represents the data collection setup. To guarantee that the subjects understand the rules of the activity, each task comprises an instructional film as well as practice videos [12, 29, 8]. The goal is to develop intelligent software that will allow instructors to view automated system prediction and performance visualizations alongside recorded video of participants, as shown in Figure 3.

The instructional video provide a brief demonstration of how the current task should be executed. The recording modules will activate once the evaluation is started, and Aliza, the on-screen instructor, will assist the students through each task. Annotation software was also created to allow computer scientists and cognitive experts to visualize and annotate the data. An expert examines each recording of the assessment based on a set of task-specific criteria.

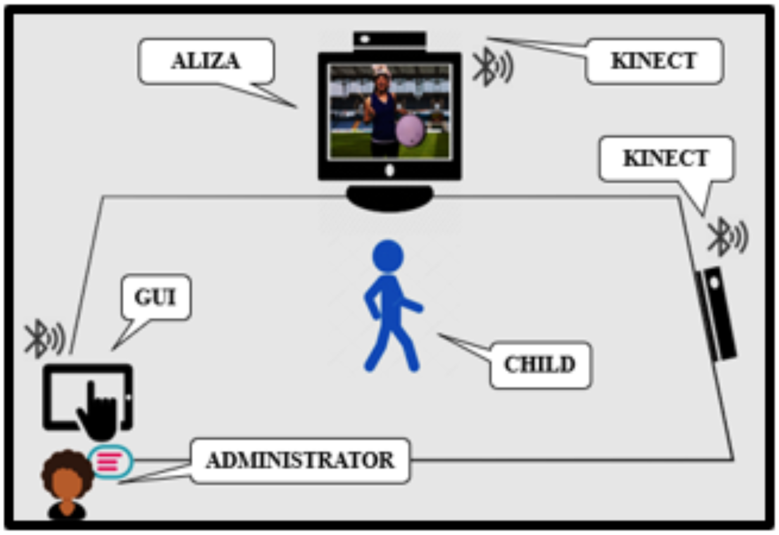


Figure 2.1: The ATEC data collection setup [29]. The participant performs the tasks based on instructions while being recorded by two Kinect cameras. The administrator monitor the whole task by an intelligent GUI.

The automated scoring system is then evaluated using this expert annotation as a baseline [12, 29, 8].

Children aged 5 to 11 (mean (sd) = 8.04 (1.36)) were recruited from the community (N = 55) through local public schools and through fliers displayed on bulletin boards. In accordance with procedures approved by the University IRB, their parents supplied written informed consent, and the children offered verbal assent. Although all of the children were in regular classes, 9 (16.4 percent) of them received additional services through a 504-plan approved by the school. The population was ethnically diverse (56.4 percent Caucasian) and 58.2% male [12].

Before the testing procedure, the parents are required to complete pre-screening paperwork which collects information about the history of children and the family. This pre-screening is followed by paper based assessment

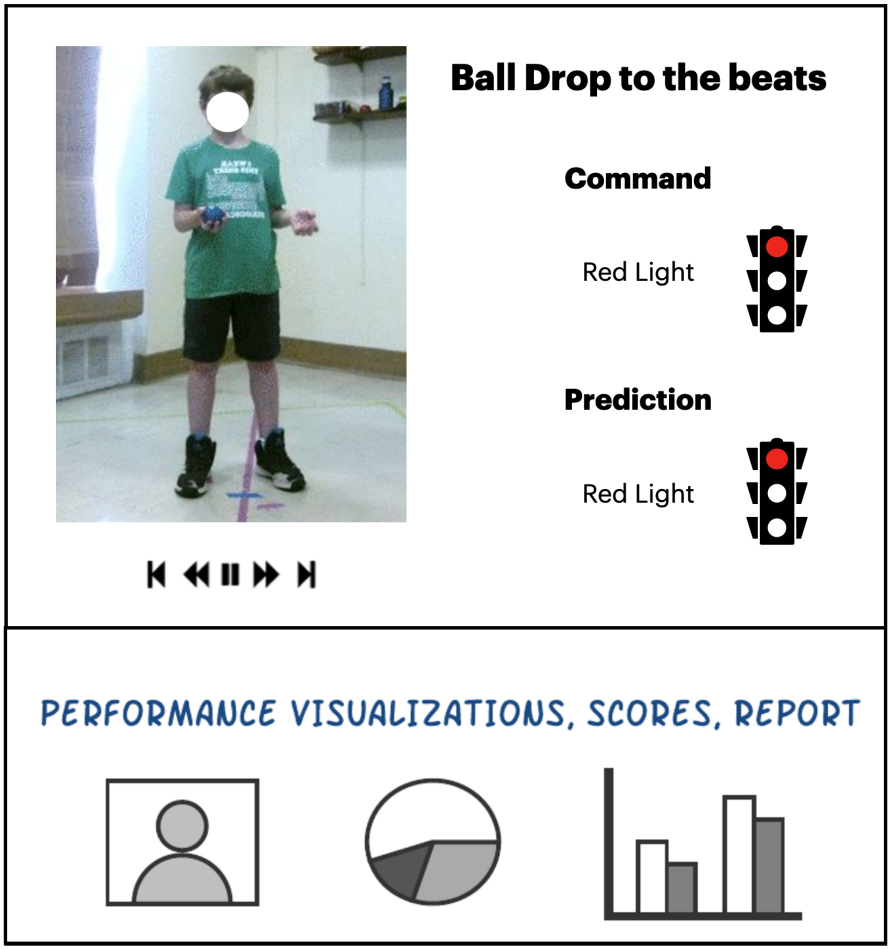


Figure 2.2: ATEC system GUI.

tests such as Child Behavior Checklist (CBCL) [3], Social Responsiveness Scale, Swanson, Nolan and Pelham questionnaire [6], etc. Then participants are requested to take part in standard computer tests from the NIH toolbox such as Flanker and Working Memory Test (WMT) [125] to gauge various cognitive measures such as attention, response inhibition, etc. Finally, the children will perform all the tasks from the ATEC program in two trials one week apart.

ATEC was created to concretely quantify embodied cognition, a concept with broad acceptance but limited consensus on its precise meaning. The ATEC results show that it has contemporaneous validity with traditional neuro-psychological and parent-reported Executive Functioning (EF) measures. ATEC demonstrated discriminant validity between children at risk for EF-related impairments and children who were not, and it had a strong association with the CBCL parent-rated assessment of real-world functioning [12].

ATEC Total Score alone explains a large amount of variance in CBCL parent-rated functioning, according to multiple regression analysis, and no other neuro-psychological measure contributed significantly to the model. None of the other measures entered the model when ATEC Total Score was not included. This conclusion could imply that a measure of cognition in action is more closely related to functioning than traditional assessments that do not entail movement [12]. The researchers discovered high test–retest reliability with acceptable practice effects, as well as an expected moderate connection with age, implying that embodied cognition is linked to normal development. Bell et al. [12], provides a more detailed examination of the ATEC system.

ATEC system consists of 17 physical exercises with different variations and difficulty levels, designed to provide assessment of executive and motor functions including sustained attention, self-regulation, working memory, response inhibition, rhythm and coordination as well as motor speed and balance. The measurements are converted to ATEC scores which describes the level of development (early, middle, full development). Table 2.1 represents the list of all ATEC tasks that has been devised for variety of cognitive measures. Description of the ATEC tasks that have been incorporated in this thesis are provided

Category	Test
Gross Motor, Gait and Balance	Natural Walk, Gait on Toes, Tandem Gait, Stand Arms Outstretched, Stand on One Foot
Synchronous Movements	March Slow, March Fast
Bilateral Coordination and Response Inhibition	Bi-Manual Ball Pass with Green, Red 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, Finger Tap, Appose Finger Succession

Table 2.1: ATEC tasks to assess various cognitive measures.

in following subsections [12].

2.1 Bi-manual Ball Pass

Bilateral coordination is defined as the ability to coordinate both sides of the body at the same time in a controlled and orderly manner.Bilateral coordination suggests that both sides of the brain are working together successfully. Fine motor skills such as buttoning shirts, visual motor tasks such as writing, and gross motor activities such as walking, climbing stairs, and so on will be challenging for children who are unable to coordinate both sides of their bodies appropriately [12].

In Bi-manual Ball Pass task the participants are required to pass the ball from one hand to another hand in rhythm with the beats for a total of 8 times. This task has two trials. In the first (slow) trial, the beats are played every 1.5 seconds but in the second (fast) trial, the beats are provided every 1 seconds [29].

2.2 Ball Drop to the Beat

Another important component of cognitive functions is Attention. The ability to focus on a certain aspect of information while ignoring other perceptible information is characterized as attention. Similarly, response inhibition (inhibitory control) is an executive function that allows an individual to regulate their natural or habitual dominant behavioral responses to stimuli by inhibiting their impulses. It will enable them to adopt a more appropriate conduct that is compatible with their objectives. The following are two ATEC tasks for assessing bilateral coordination and response inhibition [12, 29].

Ball Drop to the Beat is a core ATEC task devised to evaluate both audio and visual cue processing while performing upper-body movements. In this task, the participant is required to pass a ball from one hand to the other while following verbal and visual instructions. According to the rules, the participants are required to pass the ball for Green Light (Pass), keep the ball still in their hand for Red Light (No Pass) and move the ball up and down with the same hand for Yellow Light (Raise). The light colors are presented both audibly and visually to gauge both audio and visual accuracy and response inhibition. The task is assessed at 60 beats per minute (slow trial) and 100 beats per minute (fast trial) for a total of 16 counts for each trial [29].

Apart from accuracy and response inhibition, ATEC exercises also measure rhythm. During the test, the ATEC on-screen host, Aliza, announce the stimuli in a rhythmic manner by saying Green/Red/Yellow Light in two beats; First beat for the color and the second one for the word light. Thus the subjects are required to perform the actions in two beats. For instance, for Green Light (Pass)

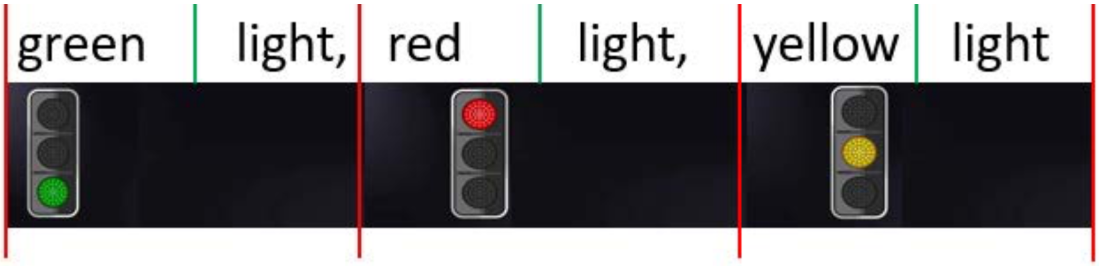


Figure 2.3: Audio-visual stimuli during the Ball Drop to the Beat task [29].

and Yellow Light (Raise) commands, the ball is raised on the first beat and either passed or lowered on the second beat. Figure 2.3 illustrates both audio and visual stimuli. Each segment (activity) in this diagram is divided by red lines, and each segment contains two beats separated by green lines [29]. Figure 2.4 represents a sample for each class of action performed by the children.

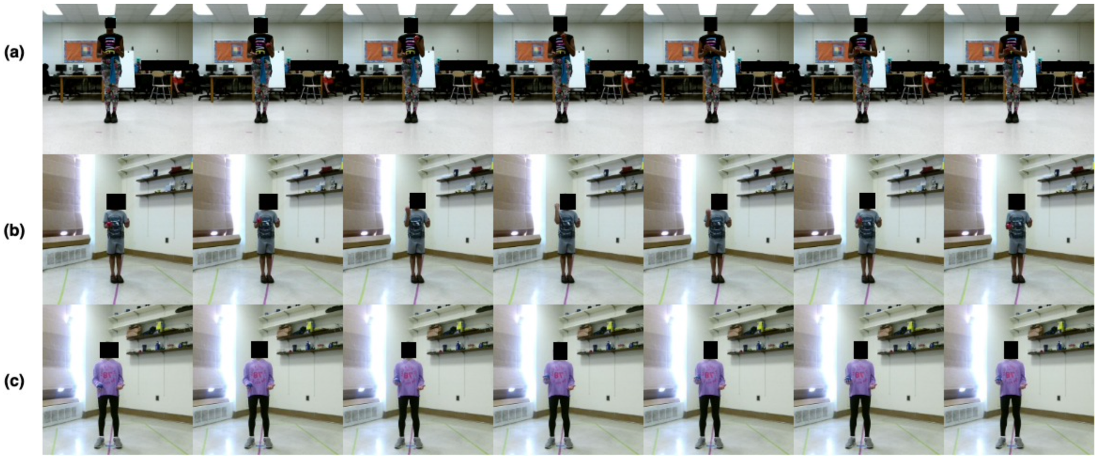


Figure 2.4: Samples actions from the dataset. Row (a): ball pass, (b): hand raise, (c) no pass [95].

In this task, the score for Response inhibition (RI) is determined by dividing the number of correct Red Light (No Pass) actions by total number of Red Light commands. Similarly the score for attention is defined as number of correct

Green Light (Pass) and Yellow Light (Raise) actions divided by total number of Green Light and Yellow Light commands [12].

2.3 Finger Opposition

One of the important executive function tasks in ATEC system that assesses the sensori-motor function is the well established Finger Opposition test. The Finger Opposition test is an exercise where the subjects are instructed to sequentially tap their index, middle, ring and little finger against their thumb 2.5. The subjects are expected to perform the sequential movement for every count/beat provided by the therapist [20, 19].

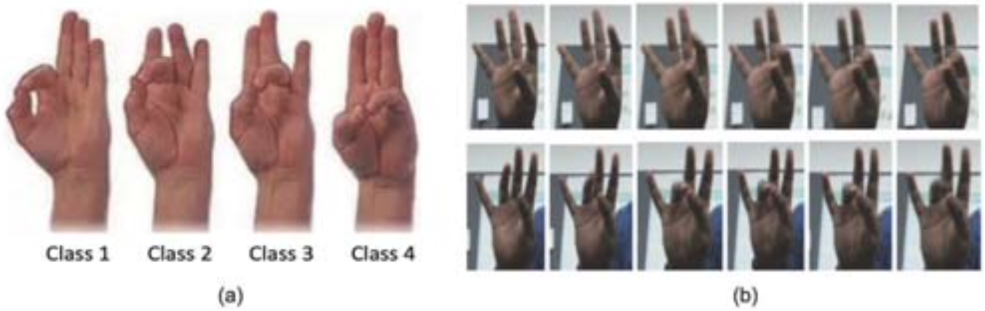


Figure 2.5: Examples of Finger Opposition task; (a) 4 different classes. (b) Sample frame sequence for class 1 (top) and class 3 (bottom) [8].

Finger Opposition as a task, has been used in multiple conditions like Parkinsonism and cerebellar diseases [104]. Authors in [105] used Finger Opposition task to identify brain activity related to cognitive behavior. Authors in [112] has shown in their work, that learning sequential finger movements helps in evolving reorganization within primary motor cortex through fMRI.

2.4 Tandem Gait Forward

In this task, the participants are asked to walk in a straight line where for every step, the heal of the foot moving froward is expected to touch the toes of the leg behind. The subject's score is calculated as the total number of correct steps performed out of the total number of 8 expected steps [12, 121]. An example of a valid and an invalid step are presented in Figure 2.6. In these figures, children's body are covered by their estimated SMPL body mesh [61] in order to protect their privacy.

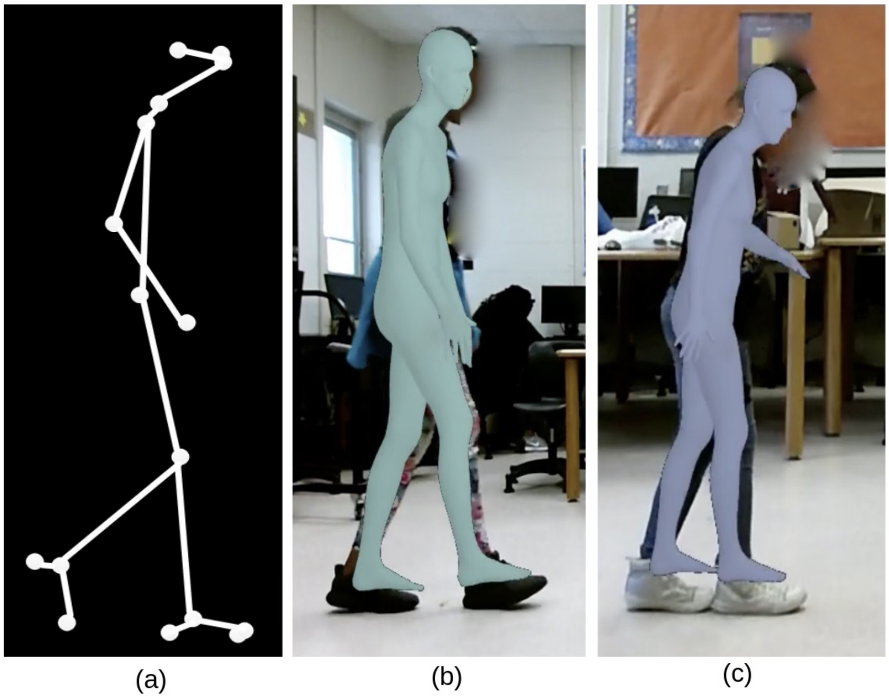


Figure 2.6: Example of tandem gait task; (a) Skeleton key points, (b) An invalid step, (c) A valid step.

There has been a plethora of research in recent years that tackle the problem of analyzing body gait for prediction and diagnosis of multiple disorders. In [79], machine learning methods have been widely used for gait assessment through the estimation of spatio-temporal parameters. The proposed methodology was tested on gait data recorded on two pathological populations (Huntington's disease and post-stroke subjects) and healthy elderly controls. They used data from inertial measurement units placed at shank and waist. In [80], wearable sensor technologies were employed for development of new methods for monitoring parameters that characterize mobility impairment such as gait speed outside the clinic. In their work, authors try to extend these methods that are often validated using normal gait patterns to subjects with gait impairments [121].

The focus of the work described in [58] was on diagnosis of Vascular Dementia during or prior to vascular cognitive impairment. They explored gait analysis which include stride length, lateral balance, or effort exerted for a particular class of activity. Although gait has clear links to motor activities, they investigate an interesting link to visual processing since the visual system is strongly correlated with balance. Various gait metrics have been investigated, and their potential to identify vascular cognitive impairment has been evaluated [121].

In [108], the issue of support for diabetic neuropathy (DN) recognition is addressed. In this research, gait biomarkers of subjects is used to identify people suffering from DN. To achieve this, a home-made body sensor network was employed to capture raw data of the walking pattern of individuals with and without DN. The information was then processed using three sampling criteria and 23 assembled classifiers in combination with a deep learning algorithm [121].

In [57], the effects of human fatigue due to repetitive and physically challenging jobs that cause Work-related Musculoskeletal Disorder (WMSD) was investigated. This study was designed to monitor fatigue through the development of a methodology that objectively classifies an individual's level of fatigue in the workplace by utilizing the motion sensors embedded in smartphones. Using Borg's Ratings of Perceived Exertion (RPE) to label gait data, a machine learning algorithms was developed to classify each individual's gait into different levels of fatigue. Finally, in [84], the aim of the study was to determine whether gait and balance variables obtained with wearable sensors could be utilized to differentiate between Parkinson's disease and essential tremor [121].

2.5 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 capability to sustain for a given period of time. Scores are determined based on the number of seconds, the participant can withstand without stopping. In the first round, subjects stand on their left foot and in the second round, they stand on their right foot [12, 29]. An example of a participant standing on her right foot is depicted in Figure 2.7.

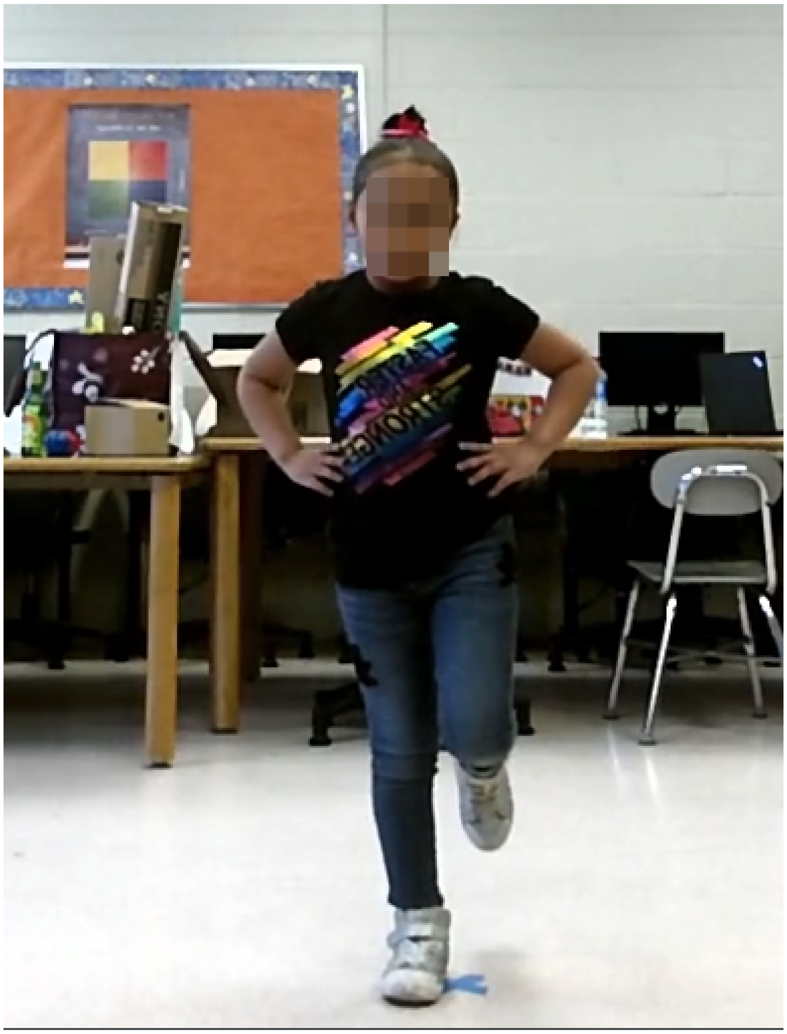


Figure 2.7: An example of a participant standing on their right foot.

CHAPTER 3

ATEC: SUPERVISED METHODS

In this chapter, three different approaches for designing an automated assessment system for ATEC tasks are presented. All these approaches employ supervised learning, i.e., they need all the training data to be manually annotated by experts [68].

3.1 Deep Learning based approach

This study presents a prototype of an intelligent automated system for evaluating participants' performance on the Finger Opposition task. This system has a graphical user interface (GUI) that allows you to view the subjects' performance statistics based on how accurately they complete the task.(Figure 2.2). Deep learning techniques for hand detection [72] and action classification [110] are used to build the proposed prediction system. Convolutional Neural Networks (CNN) are used in both methods, and have been shown to be extremely effective in image and sequence classification. The hand detector extracts the hand from the scene, which is then classified by the action recognition system. Figure 2.5(a) depicts four classes based on the Finger Opposition task (class 1: thumb against index finger, class 2: thumb against middle finger, class 3: thumb against ring finger, and class 4: thumb against little finger).

Most state-of-the-art methods perform well only for specific datasets, so action recognition remains an open problem. For the methods to work with other datasets, extensive fine tuning of the hyper parameters is required. A dataset for the Finger Opposition task was created and combined with an existing dataset built by Srujana et. al. [37]. The combined dataset is made up of RGB image frames collected from subjects while they were performing the exercise. The hands were cropped manually and saved as the most useful information for predicting the task.

This dataset consists of data from 10 subjects (approximately 4500 images) with various hand angles to increase the system's robustness. The image frames were manually divided into sequences (a group of image frames that determines a class) and annotated as shown in Figure 2.5(b). When one of the four fingers touches the thumb and returns to its original position, the sequence is complete. A total of 200 dollars in training sequences were annotated, with sequence lengths ranging from 10 to 28 dollars. Similarly, there were a total of 50 validation sequences, and the system was tested with 30 real-time sequences.

3.1.1 Proposed Method

The methodology used is explained in detail in this section. Figure 3.1 depicts the proposed system's pipeline. This will be used to evaluate and assess physical activities that may reveal executive function deficits. It is made up of several parts that work together to achieve the desired result.

Hand Detector

The Hand Detector is the first component in the computer vision pipeline, and its goal is to detect the active hand in the scene, for which we use an approach

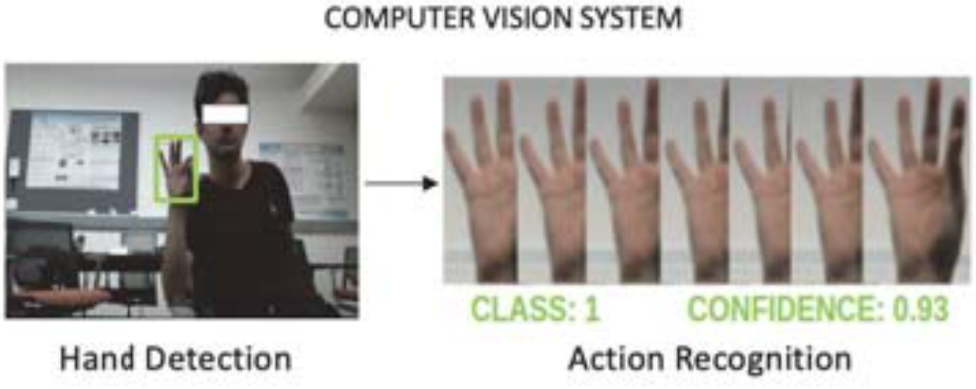


Figure 3.1: Deep Leaning based method architecture [8]

called Single Shot Multi-Box Detector (SSD) [72]. SSD distinguishes itself by employing only a single deep neural network for the entire process of detecting the hands, whereas other methods, such as Faster-RCNN [98], employ multiple elements in their pipeline, making SSD more time efficient. Experiments also revealed that SSD strikes a fine balance between detecting smaller objects, speed, and mAP.

The algorithm divides the given input frame into a grid of size $N \times N$, with a set of default boxes with different ratios and scales generated for each cell in the grid. The network generates scores for the presence of objects (hand) in each of the default bounding boxes during prediction. If the score exceeds a certain threshold, the system assumes there is a hand in the generated default box. Finally, non-maximum suppression is used to remove duplicate predictions. Furthermore, this procedure is carried out at various scales of the feature map in order to capture hands of various sizes. The system was pre-trained with Ego-Hands dataset [9, 115] which contains more than 4800 image frames with approximately 15000 ground-truth labeled hands. This dataset was chosen because it contains images of people performing different activities that involved hand movements (playing chess, playing cards, solving puzzles etc.).

The dataset was divided into train (80%), validation (10%), and test (10%). Mean Average Precision was used to evaluate the detector (mAP). When tested at the 0.5 threshold, the system's mAP was 96%. On a single GPU, the system can produce 15 frames per second. Because other parts of the captured frames are irrelevant for classification, only the detected hand is advanced to the next stage.

Action Recognition System

The proposed action recognition system is built on 3D Convolutional Neural Networks (CNNs) [110], which are the natural successors to standard 2D CNNs [64]. They are a type of artificial neural network in which the weights of spatial filters in each layer are shared across the entire image. Instead of 2D spatial filters, 3D spatio-temporal filters are used in 3D CNNs, which means they extract features from both the spatial and temporal dimensions by performing 3D convolutions, capturing motion information encoded in multiple adjacent frames.

Residual Deep Neural Networks (ResNet) [44] were used as a special variant of CNNs for this work. To bypass some layers, the ResNet employs skip connections or short-cuts. The goal of skipping layers is to avoid vanishing gradient problems, which can make it easier to build deeper networks that are easy to train and optimize.

One of the major challenges of 3D CNNs is that they have a large number of learnable parameters, requiring a large amount of data for training. Training such deep networks with a small amount of data results in overfitting of the model. Thus, in this experiment, a relatively shallow network (3D-ResNet10) [42], which is essentially a ResNet with 10 layers, was used. Each block of 3D-ResNet10 is comprised of convolutional layers with 3D kernel of size 3 × 3 × 3, Batch Normalization (BN) [50] and Rectifier Linear activation units (ReLU) [85].

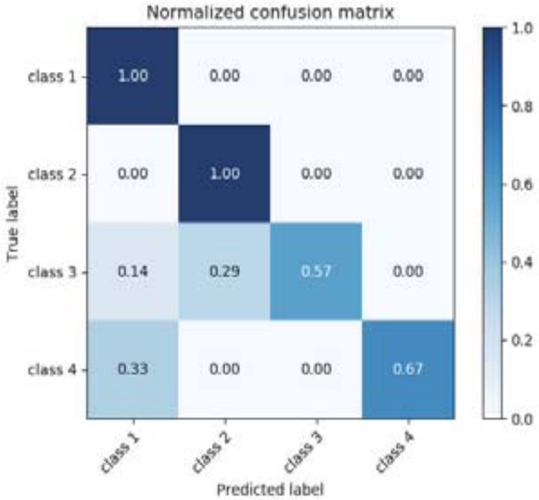


Figure 3.2: Confusion matrix for the action recognition system, demonstrating that all actions in classes 1 and 2 have been correctly classified, whereas actions in classes 3 and 4 are more difficult [8].

We divided our dataset into training and validation with a 4 to 1 ratio during training. During our training process, we used K-fold cross validation. We used 30 real-time sequences for testing. The length of the sequence during training was 8, implying that we trained our network with 8 images for each sample. The image frames were RGB images with a resolution of 64 × 64. Our network was optimized using the Adam optimizer [59] with a learning rate of 0.1, which was divided by 10 when the validation loss became saturated. The confusion matrix generated for the test dataset with the best performing model (ResNet10) is shown in Figure 3.2. According to the confusion matrix, the system correctly classified classes 1 and 2, but incorrectly classified classes 3 and 4. The Pytorch

[88] framework was used for training.

Scoring System

Based on the task rules and guidelines, the scoring system computes the scores for the task performed. The predictions are pre-processed in order to smooth them before calculating the scores. We use a smoothing operation similar to the moving average technique to avoid duplicate predictions of the same class when subjects switch from one finger to another and to correct any errors in right classification. A prediction output example might be (1,1,1,1,2,2,3,2,3,3,4,4,4,4), where the number represents the class predictions of a window and a confidence score is associated with each prediction.

If the confidence score for a prediction is less than a certain threshold (i.e. the system is not confident in its prediction) and the current prediction differs from its neighbors, the current prediction will be updated. The duplicate predictions are combined into one prediction after the smoothing process by averaging their confidences. When the subject performs all four sub-sequences (thumb to index finger, thumb to middle finger, thumb to ring finger, and thumb to little finger), he or she has completed a sequence. When a complete sequence is achieved, the subject receives a full point. The score ranges from zero to the total number of times the entire sequence is performed (The maximum is considered as five in this case). Subjects are expected to switch from one sub sequence to another only after hearing the system's beat.

Method	Accuracy
2D-CNN+LSTM	0.655
2D-CNN + GRU	0.60
Multi-Stream Network	0.76
3D-CNN (ResNet10)	0.89

Table 3.1: Experimental results for Deep Learning based method [8].

3.1.2 Results and Discussion

Multiple attempts were made to select the best method for the Finger Opposition task. Based on recent surveys on action recognition [45, 63] we developed and trained methods that have been shown to work best on public datasets. Table 3.1 shows the methods that were tried, and the best method was chosen based on the validation results. Based on the validation results, we discovered that 3D Convolutional Networks performed the best on our dataset.

Because residual networks perform better for action recognition [42] ResNet-18 was built and trained first. However, the model performed poorly, with validation and test accuracy of less than 40%. The model's inability to generalize could be one of the reasons. As a result, an attempt was made to vary the network depth by using ResNet-50 and ResNet-10. While ResNet-50 performed poorly as predicted, ResNet-10 achieved both validation and testing accuracy, as well as precision and recall greater than 80%, which was the best of all approaches tested.

3.2 Multimodal Approach

Human Activity Recognition (HAR) is still an active area of research in the field of computer vision. In recent years, HAR has taken giant leaps with deep learning-based approaches [116]. Several approaches have been proposed recently that have taken advantage of various feature extraction techniques such as video-based features [56] and skeleton-based features [97] to recognize human activities. However, the nature of the data and the problem to be solved have a high influence on system performance.

To address this, attempts have been made to fuse various combinations of features such as body poses, optical flow, objects in the scene, hand pose, and so on [35, 55]. However, an optimal combination of features and an effective approach to fusing them is still a work in progress in HAR [126]. In this work [95] a combination of the following three modalities for HAR is proposed: optical flow, objects in the scene, and human poses (skeletal information). In addition, an attention-based approach is proposed to combine the aforementioned modalities.

A quick survey of literature shows that the fusion of multiple modalities can improve HAR performance. Franco et al. fused skeleton-based features with video-based such as Histogram of oriented gradients (HOG) [35]. The results showed that the fusion of the two modalities improved the performance of the HAR tasks. Object detection is an important computer vision modality, which has been extensively researched in the last decades [129]. Kapidis et. al. combined hand and object detection to recognize human actions from an egocentric view camera [55]. The combination of the two modalities improved the performance of the HAR system.

To the best of our knowledge, there is no work proposing to combine the following three modalities for HAR tasks; Optical flow (video-based HAR), human poses (skeleton-based HAR), and object detection. The proposed system is evaluated on a real-world dataset built for the ATEC task: Ball Drop to the beat. The results show that the proposed method outperforms state of the art approaches when applied to the Ball-Drop dataset.

Data from 25 children between the age of 6 and 10 were collected for two sessions (two weeks apart), providing a dataset of 50 sessions. Each session had multiple trials with different pace (1 sec, 1.5 sec) and cues (visual, auditory). The distribution of the scores of the dataset was $\mu = 14.4$ and $var = 2.03$ with 0 being the lowest score and 16 being the maximum score for every trial. Each video recording was broken down into multiple segments with each of them annotated based on their action and rhythm by psychologists. Hence, a total of 3300 annotated video segments were extracted from the recordings. Figure 2.4 represents a sample for each class of action performed by the children.

3.2.1 Proposed Method: Multimodal

In this work, human action recognition (HAR) is performed based on a multimodal approach. Each modality is discussed with its architectural details and the approach used to fuse the output from the modalities is addressed as represented in Figure 3.4.

Modality 1: Optical Flow

Optical flow is one of the most widely used features to represent motion for HAR. It is often formulated as a problem of estimating the 2D projection of a true 3D motion. For a given segment of video, optical flow aims at capturing the motion information between consecutive frames [100]. In this work, optical flow is computed using the off the shelf implementation of from the OpenCV toolbox [15].

With optical flow being computed, a deep neural network based architecture inspired from [42] was used to extract useful information from the optical flow segment. The architecture is based on 3D Convolutional Neural Networks (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 ($3 \times 3 \times 3$). Figure 3.3(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) [85]. Down-sampling of the inputs is performed at $conv3\_x$, $conv4\_x$ and $conv5\_x$ while increasing the feature size. A comparatively shallow network with 18 layers was empirically selected as represented in Figure 3.3.(a). After training, the features were extracted from pre-logit layer which was used during fusion.

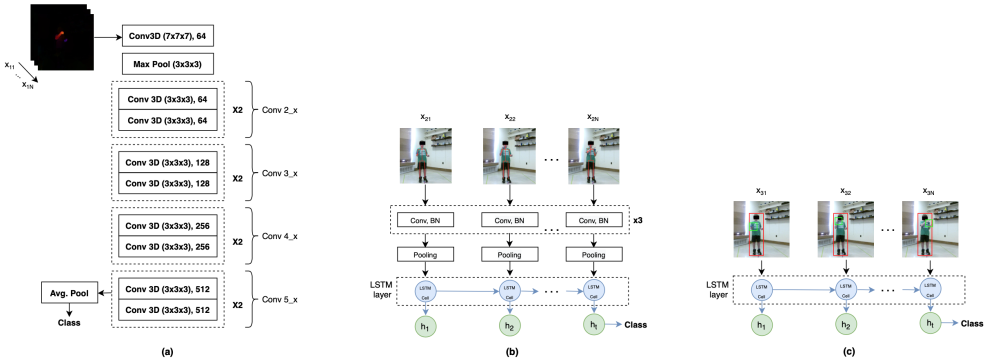


Figure 3.3: (a) Optical flow based encoder network. (b) Action prediction using sequence of body key-points. (c) Prediction of objects coordinates in the scene [95].

Modality 2: Human Body Pose

Recent research on human detection and pose estimation in RGB image frames shows that deep learning-based methods [34, 16] have achieved better results in any complex scene, paving the way for human action feature learning. These 2D/3D human poses acting as trajectories of skeleton joints, is one of the most effective representations for characterizing the dynamics of human actions. Each coordinate in the skeleton is known as a joint or a key-point and a valid connection between two key-points is referred to as a limb or a pair. An open-source pose estimation framework is used for this purpose [61]. During pose estimation, any missing key-points in a given image frame is 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, 18 key-points are extracted from each frame that represent various body joint positions including facial key-points such as eyes, ears, and nose. In this work, only 9 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 key-point is represented as a 3D coordinate (x, y, z) on the image plane. Hence, a given frame P at time t is represented by the coordinates of the 9 key-points as shown in the following equation:

$$P_t = [(z_{1,t}, y_{1,t}, v_{1,t}), (z_{2,t}, y_{2,t}, v_{2,t}), ..., (z_{9,t}, y_{9,t}, v_{9,t})] \tag{3.1}$$

where z denotes the coordinate extending from left to right and y extending top to bottom and v representing the depth for each key-point. 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 is extracted. The subnet is represented in Figure 3.3(b).

Modality 3: Object detection

This modality aims at detecting objects in the scene. It is essential to identify the objects being interacted along with its 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. The sequence of coordinates

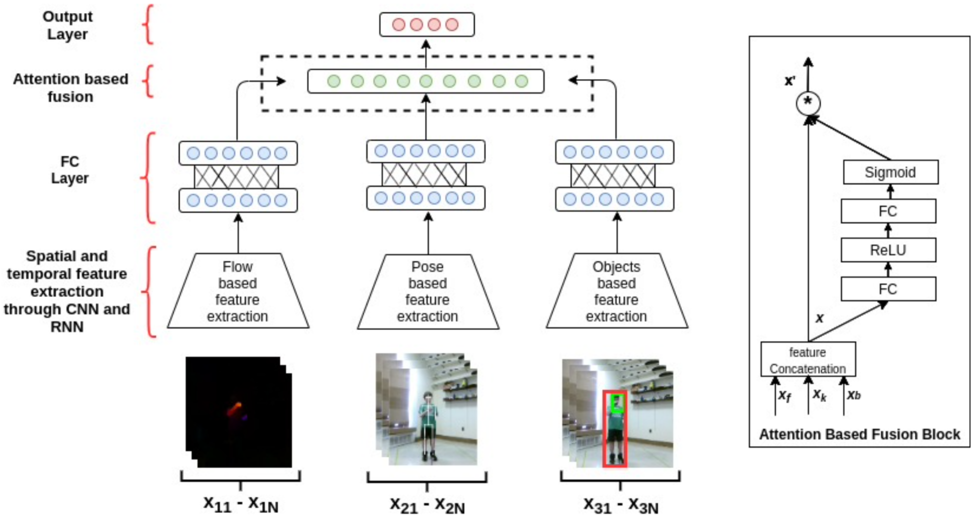

Output Layer
Attention based fusion
FC Layer
Spatial and temporal feature extraction through CNN and RNN
Flow based feature extraction
Pose based feature extraction
Objects based feature extraction
x_{11} - x_{1N}
x_{21} - x_{2N}
x_{31} - x_{3N}
x'
*
Sigmoid
FC
ReLU
FC
x
feature Concatenation
x_f
x_k
x_b
Attention Based Fusion Block

Figure 3.4: Multi-modal fusion [95].

are fed into a subnet to identify the trajectories of the objects being interacted with leading to identification of the actions [55]. 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 [96] 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 was 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 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 is extracted. The subnet is represented in Figure 3.3(c) represents the architecture to predict actions through objects in the scene.

Multi-Modal Fusion

In a multi-modal action recognition problem, not all modalities or features within a modality equally contribute towards the 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 [114, 117] based fusion approach is proposed inspired from [47]. 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. The overall architecture, including the attention-based fusion module is represented in the Figure 3.4. In order to calculate the weights of features of each modality, first all features are concatenated into one vector as follows:

$$x = [x_f, x_k, x_b] \tag{3.2}$$

where $x_f \in R^{C_f}$ is the feature vector obtained from optical flow subnet, Figure 3.3(a), $x_k \in R^{C_k}$ is the feature vector from the pose subnet, Figure 3.3(b), $x_b \in R^{C_b}$ is the feature vector from objects position based subnet, Figure 3.3(c) and finally $x \in R^C(C = C_f + C_k + C_b)$ comprising of features from all modalities. Further, to calculate attention weights for features of x, function F_w is introduced as represented in equation 3. 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.

$$\alpha = F_w(x, W) = \sigma(g(x, W)) = \sigma(W_2\delta(W_1 x)) \tag{3.3}$$

where *delta* refers to the ReLU [85] function, $W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$. In order to generalize, the gating mechanism is parameterized by forming a bottleneck with two Fully-Connected (FC) layers (W_1 and 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:

$$x' = F_a(x, \alpha) = \alpha x \tag{3.4}$$

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.

3.2.2 Results and Discussion

Prior to the development of the proposed multi-modal approach, several methodologies that have achieved the state of the art results on existing popular action recognition datasets were attempted. Table 3.2 shows the different attempted methods with their results. All of the mentioned methods used different features such as optical flow, RGB images, Pose information etc. as

Method	Accuracy	Features
3D CNN [42]	73.0%	RGB
Two Stream I3D [17]	82.5%	RGB + Flow
CNN+LSTM [36]	69.0%	RGB
DeepGRU [77]	61.0%	KP
Dillhoff et. al. [29]	78.0%	KP
Attnsense [76]	81.0%	RGB
Multimodal [95]	89.8%	KP + Object Pose + Flow

Table 3.2: A comparison of the proposed Multimodal method's performance to that of other supervised methods [95].

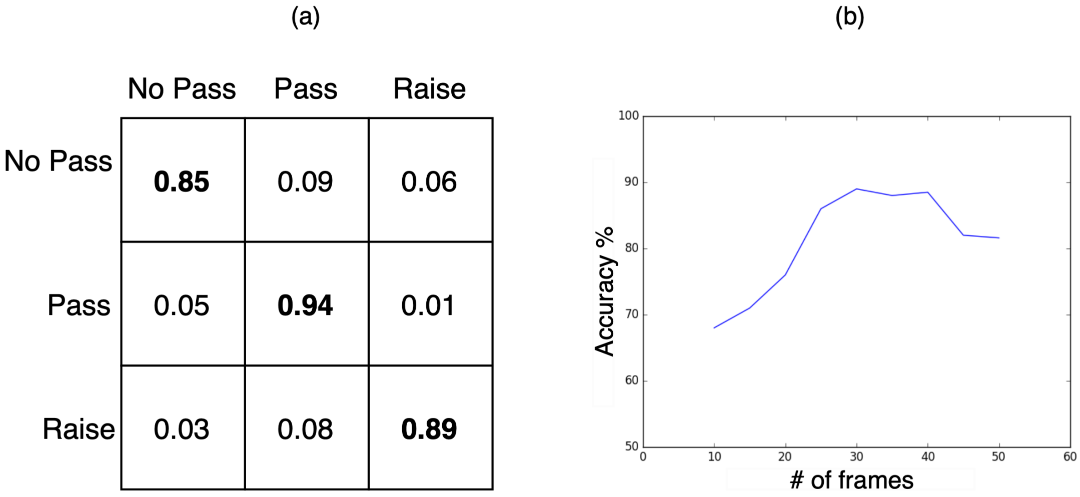
mentioned in the modalities column. For the modalities column, RGB represents RGB based video segments, flow represents optical flow based sequence of frames, object pose represents objects in the scene based action recognition and body key-points (KP) represents human pose-based action recognition.

In Table 3.2, for 3D convolution-based approach, ResNet with variable depth sizes (18, 34, 51) and inception model were trained. Although, it was observed that as the depth of the model increased, the model started to overfit. Hence, the results shown is only for resNet-18. For the two Stream I3D, inception-based model was trained for RGB based sequences and optical flow based sequences. During testing, the outcome of both the models were combined for final prediction. The hyper-parameters for initial training were used as recommended by the authors of the papers followed by fine tuning in the later trials.

For training, the dataset was split into training, validation and testing set based on participants. This was done to ensure that the training set did not include video segments from the participants that were in validation and testing set as it might influence the results. The validation was performed after every epoch of training in order to identify the right epoch to stop the training to avoid overfitting. At the end of training, the model was evaluated on the test set. Stochastic Gradient Descent based optimization with momentum was used during the training. Since the dataset is comparatively smaller than the other publicly available datasets, extensive temporal and spatial augmentation was performed during the training. A video clip of size *t* is generated with a randomly selected temporal position as the starting frame. If the video is shorter than *t* frames, then its looped through until it matches the size *t*. For spatial augmentation, a spot is randomly chosen between four corners and center of the image and multi-scale cropping was performed after which the images were spatially resized. Cross-entropy loss was used during training with starting learning rate set to 0.0001 and divide by 10 every time the validation loss saturates, a weight decay of 0.001 and 0.9 for momentum. To train the models, four NVIDIA GTX 1080 Ti GPUs were used whereas for testing, one GPU was used.

As mentioned before, several features can be extracted from RGB image sequences that include body pose, optical flow, objects in the scene, and an effective combination of modalities is still unsolved and depends on the problem. To identify the effective combination, training was performed extensively on different combinations and with different fusion approaches. Table 3.3 contains the results of the experiments. All results were averaged over 5-folds.

Similarly, to fuse the features from individual modalities, in addition to the approach mentioned in Figure 3.4, other approaches were also attempted. The natural concatenation (nat. Concat) is a vanilla approach where output features of different modalities were directly concatenated followed by a softmax layer to classify the actions. On the other hand, balanced concatenation (bal. Concat)



	No Pass	Pass	Raise
No Pass	**0.85**	0.09	0.06
Pass	0.05	**0.94**	0.01
Raise	0.03	0.08	**0.89**

Figure 3.5: (a) Confusion matrix of proposed method in [95]. (b) Graph representing model accuracy as a function of number of frames.

aims to convert the feature vectors from different modalities into same dimensional size followed by concatenation and a softmax layer. As the goal was to deploy the proposed system for future data collection, it was important to measure the execution time of the model which is presented in Table 3.3, especially when multi-modal approaches are used. Figure 3.5(a) presents the normalized confusion matrix of the proposed method on the test data to predict actions.

As mentioned in section 4.2, for a given segment, the system detects at what point of time in a given segment, an action takes place, but does not provide any information about what the action is. The approach compares the prediction from HAR system with the command fired, if the predicted action does not match with the command fired, the child gets ”0” points assuming that the child did not either complete the task or did a different action for the command. If the command matches with the prediction from the HAR system, the rhythm scores were calculated with approach mentioned in section 4.2. The rhythm

Method	Accuracy	Time (sec.)
Optical Flow (Flow)	72.0%	0.229
Body Key-points (KP)	76.0%	0.106
Objects Trajectories (Obj)	68.0%	0.103
Flow+KP (natural-concat)	82.0%	0.236
Flow+KP (balanced-concat)	83.9%	0.239
Flow+KP (self-Attn)	84.6%	0.240
Flow+Obj (natural-concat)	84.1%	0.232
Flow+Obj (balanced-concat)	83.9%	0.236
Flow+Obj (self attn)	84.0%	0.241
KP+Obj (natural-concat)	79.0%	0.118
KP+Obj (balanced-concat)	76.3%	0.123
KP+Obj (self-Attn)	79.5%	0.139
KP+Obj+Flow (natural-concat)	89.0%	0.254
KP+Obj+Flow (balanced-concat)	87.5%	0.259
KP+Obj+Flow (self-attn)	89.8%	0.260

Table 3.3: Ablation study for Multimodal method [95].

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 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.

Table 3.2 conveys that many existing 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 3.2 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 3.3 it can be observed that the body key-point based model has achieved the highest accuracy as a single modality. However, the accuracy is not as high as needed for the assessment system. 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 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 Multimodal method. Adding the object detection as an additional modality has improved the accuracy by 5.2% for attention-based fusion. Furthermore, the combination of optical flow and object position, as well as the combination of optical flow and body key-points, provide comparable accuracy, which is higher than the combination of key-points and 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 3.3 proves the same. Irrespective of what the modalities are being combined, the attention based fusion produce slightly better results. It was necessary to investigate the time taken for each of the models to process one segment. As expected, the attempt with 3 modalities has the highest execution time of 0.2603 seconds. Since the proposed system does not have the requirement to process the frames in real-time, the execution time is acceptable. Tests were done in order to identify the optimal number of frames/time steps that can be considered for the model. Figure 3.5(b) shows that initially, as the number of frames increases, the model performance increases, but then saturates and drops beyond a certain point. This could be because the model sampled the same set of frames during training while performing temporal augmentation, resulting in over-fitting of the model.

CHAPTER 4

ATEC: SELF-SUPERVISED METHODS

Recent advances in Deep Learning [68] and the challenge of collecting massive amounts of labeled data have sparked interest in unsupervised or self-supervised learning research. Because unlabeled image and video sequences can be gathered automatically without human intervention, successful models that learned abstract low-dimensional features of images and videos without supervision could greatly benefit Computer Vision tasks [63, 18].

As a result, much research effort has been directed toward methods that can adapt to new conditions without requiring costly human supervision. This chapter's main focus is on using self-supervised visual representation learning to recognize human activity in ATEC system recorded videos. Self-supervised learning techniques that include generative [39] and contrastive [52] approaches have produced state-of-the-art low-dimensional representations on the majority of computer vision benchmarks [30, 22, 86, 111, 21]. The video representation obtained from self-supervised methods can be used to obtain participants' digital phenotype [51, 48, 2, 7].

4.1 Generative Approach

The method proposed in this work (Figure 4.1) is inspired by [39, 93, 22], which augments Generative Adversarial Networks (GAN) with self-supervised rotation loss to improve discriminator network representation capability. However, the proposed work differs significantly from the existing methods. First and foremost, the goal of this work is to develop a low-dimensional representation of videos rather than still images. Second, an auxiliary loss is added to the discriminator network in [22] to detect random rotation angles on still images. However, in this work, the discriminator distinguishes between three different spatial transformations, such as rotation, translation, or shearing, as well as a temporal transformation that shuffles the temporal order of frames. All of the aforementioned transformations are applied at random to video frames. Furthermore, a thorough ablation study is carried out to investigate the impact of each different transformation.

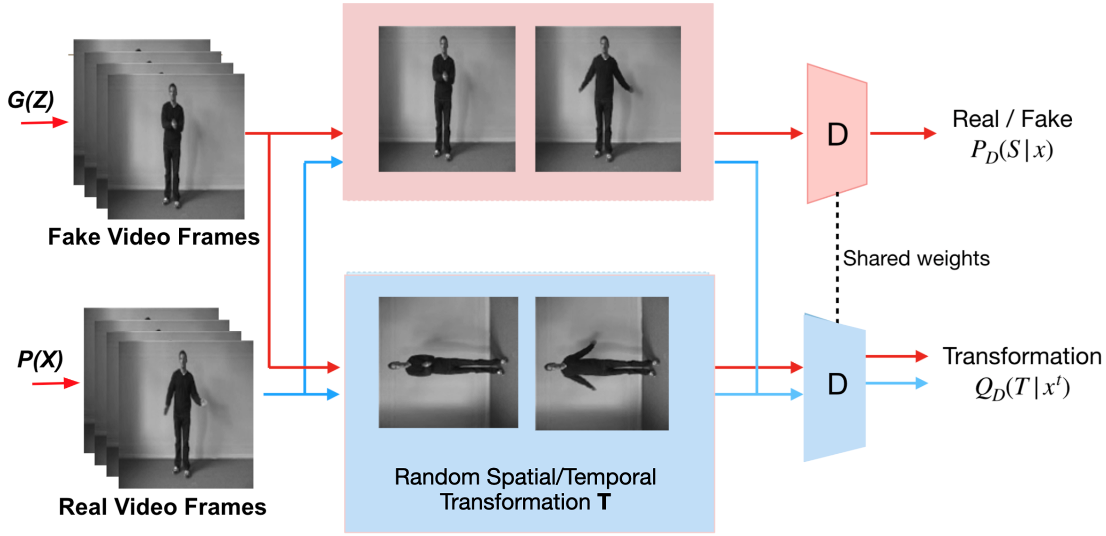


Figure 4.1: Architecture of proposed Augmented GAN method [122].

4.1.1 Proposed Method: Augmented GAN

This section begins by introducing GAN, which serves as the foundation for the methods used in this work. The proposed Augmented GAN is then described in detail, as well as how video representation (features) are extracted from it.

These features are fed into a simple two-layer multi-layer perceptrons (MLP) network for subsequent classification tasks like human activity recognition.

Generative Adversarial Networks (GAN)

GAN [39, 93] is a framework for producing a model distribution that mimics a given target distribution, and it consists of a generator $G(z; \theta_g)$ that produces the model distribution and a discriminator $D(x; \theta_d)$ that distinguishes the model distribution from the target. Training data is denoted by x and input noise is z with probability distribution of $P_z(z)$.

In practice, differentiable CNNs with parameters are used to implement both the generator and the discriminator: θ_g and θ_d. D is trained to maximize the probability of assigning the correct label to both training examples and samples from G. At the same time G is trained to minimize $log(1 - D(G(z)))$. In other words, D and G play the following two-player minimax game with value function $V(D, G)$:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{data}(x)}[logD(x)] + \mathbb{E}_{z \sim P_z(z)}[log(1 - D(G(z)))]. \tag{4.1}$$

But using GAN in practice is challenging because of instability occurring in training, mode collapsing, etc., as shown in [75, 65]. However, in recent years a variety of novel techniques such as gradient penalty [40] or spectral normalization [83] have been proposed to solve some of the challenges.

Self-supervised Learning

Discriminator forgetting is one of the main issues with GANs that limits their ability to provide good representation [22]. Because in practice, as the parameters of the generator G change, so does the distribution P_G, causing the discriminator's learning process to be non-stationary. In other words, the discriminator is not encouraged to maintain a useful data representation as long as the current representation is useful for class discrimination.

To address the aforementioned issue, the discriminator network is augmented with a self-supervised task such as predicting rotation angle [38] or counting objects in an image [87] to encourage GAN to learn useful compact representations. This work proposes a method for spatial and temporal transformation of video frames. One transformation is chosen at random and applied to frames of input video in this method (Figure 4.1). The self-supervised task after that is to predict the transformation used on video frames. As a result, both the generator and discriminator's loss functions are modified as follows:

$$L_G = -V(D, G) - \alpha \mathbb{E}_{x \sim P_G} \mathbb{E}_{t \sim T} [log Q_D(T = t | x^t)] \tag{4.2}$$

$$L_D = V(D, G) - \mathbb{E}_{x \sim P_{data}} \mathbb{E}_{t \sim T} [log Q_D(T = t | x^t)] \tag{4.3}$$

where $V(D, G)$ is the value function from Equation 1 and $t \in T$ is a transformation selected from a set of possible spatial and temporal transformations. x^t is input x transformed by transformation t, $Q_D(T|x^t)$ is discriminator distribution over possible transformations and α is self-supervised loss weight. Three different spatial affine transformations, such as rotation, translation, and shearing, as well as a temporal transformation, in which the temporal order of video frames is shuffled, are chosen for this method. Figures 4.2 and 4.3 show examples of spatial and temporal transformations, respectively.

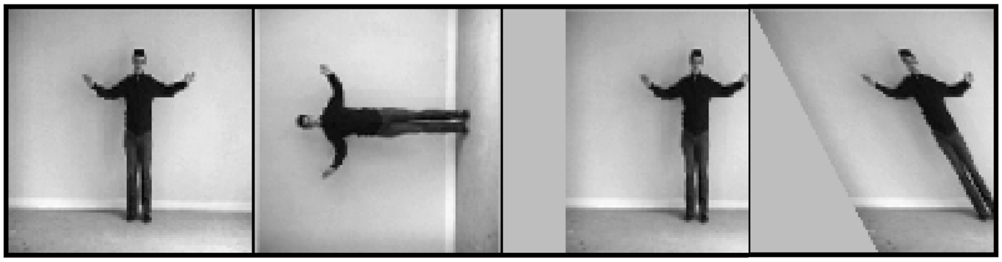


Figure 4.2: Examples of spatial transformation used. From left to right: Original Image, rotation, translation, shear.

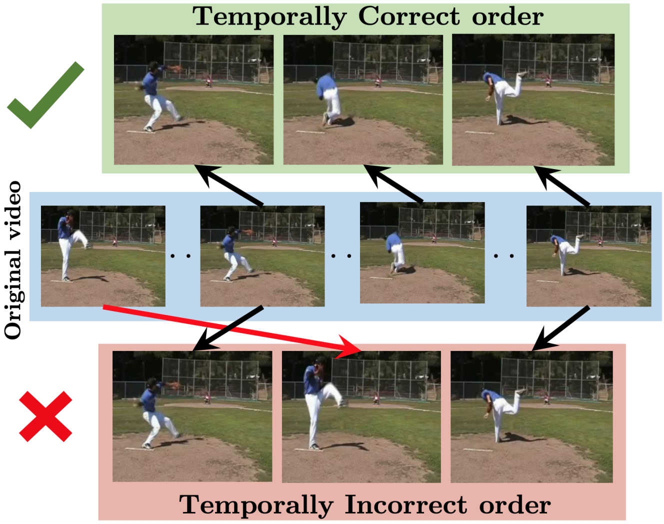


Figure 4.3: Examples of temporal transformations used in the Augmented GAN method; the classifier attempts to determine whether the temporal order of video frames has been shuffled or not. (Adopted from [82])

Only four rotation classes corresponding to rotation angles were considered: 0°, 90°, 180° and 270°. Respectively, three classes for translation (vertical, horizontal and both), three for shearing (vertical, horizontal and both) and one class for temporal transformation (shuffled or not) were chosen. So in total eleven different transformation classes were selected.

As explained in [22], the generator and discriminator work together to predict the transformation task. The discriminator is trained only on true data to detect transformations. This means that the generator is motivated to generate images that are easy to detect by the discriminator. The discriminator has two heads, as shown in Figure 4.1, with the former, like normal GANs, predicting whether non-transformed video frames are real or fake. In contrast, the latter head predicts the transformation class of transformed inputs.

After training is completed, output of the last layer before the heads is extracted as a compact representation of the input video. Then a simple 2-layer feed forward MLP is trained on extracted video representations for human activity recognition.

4.1.2 Results and Discussion

In this section, the datasets used in this experiment are introduced. This is followed by a discussion of how the neural network models are used and how they are trained. Finally, results of both baseline and proposed method are presented. It should be noted that in this work the focus is on providing compact representation of videos that can be exploited for activity recognition, thus evaluating fidelity of generated image frames is not pursued.

Datasets

Three different video datasets were used in this work to evaluate the performance of the proposed method for providing video representation useful for activity recognition. The first two are publicly accessible video datasets such as KTH [99] and UCF101 [106], which contain short video clips of humans performing various activities. The third dataset, dubbed Ball-Drop (Ball Drop to the Beat) for simplicity, is based on one of the tasks designed for the ATEC system to assess both audio and visual cue processing of children while performing upper-body movements [29, 8, 95]. One of the primary reasons for pursuing self-supervised learning is that manually annotating this dataset proved to be time-consuming and error-prone.

All of the datasets used in this article were divided into three groups. First, 80% of each dataset was considered unlabeled and was only used to train self-supervised GANs. The remaining 20% (labeled data) were fed into a trained discriminator network to extract video representations after training (features). The features were then divided into train and test sets in a 4 to 1 ratio for activity recognition.

Models

A 6 layer convolutional neural net (CNN) was used in self-supervised GANs for both the generator and the discriminator. Since the input is video, in discriminator the first 2 layer and for generator the last 2 layers employ 3D convolutional nets [42, 110]. As discussed by [75, 65] performance of GANs depends on many different hyper-parameters and there is no set of hyper-parameters that guarantee superior performance on all datasets and finding one require massive computational budget. Due to our limited computational budget, very deep complex networks such as densenet and resnet101 [44] were avoided and a small grid search was performed for tuning the hyper-parameters.

All models, including the baseline GAN and the proposed self-supervised GAN, were trained for 100 epochs using the PyTorch framework [88] with ADAM [59] as the optimizer and the following empirically selected parameters. The generator learning rate is 0.0001, the generator learning rate is 0.0004, beta1 is 0.5, and beta2 is 0.999. To stabilize the training process, all methods used spectral normalization. In addition, the self-supervised GAN parameter $alpha$ in equation 4.2 was set to 0.25. Finally, a two-layer MLP was trained with the ADAM optimizer with similar hyper-parameters for classification on extracted features.

Results

After training all the baseline and proposed methods including GAN, features (representation) of labeled video were extracted. Then, a supervised (MLP-based) human activity recognition method was trained on features and the average top-1 classification accuracy on test set was calculated by using 5-fold cross validation and presented in Table 4.1. Baseline methods include GAN [39] and self-supervised GAN with only rotation as learning task (GAN+Rotation) [22] and proposed methods are self-supervised GAN with three different spatial transformations such as rotation, translation and shearing (GAN+Spatial), self-supervised GAN with only temporal transformation (shuffling) of video frames (GAN+Temporal) and finally self-supervised GAN with both spatial and tem-

Method	KTH	UCF101	Ball-Drop
GAN	71.46 ± 2.5	64.68 ± 0.4	77.93 ± 2.7
GAN+Rotation	74.47 ± 2.5	66.86 ± 0.6	80.47 ± 2.5
GAN+Spatial	76.41 ± 2.0	66.95 ± 1.6	81.99 ± 4.5
GAN+Temporal	76.09 ± 3.2	70.88 ± 0.7	80.69 ± 3.7
GAN+SpatioTemporal	77.13 ± 3.6	69.17 ± 1.8	84.53 ± 3.0

Table 4.1: Experimental results for Augmented GAN method [122].

poral transformations (GAN-SpatioTemporal).

The experimental results prove superiority of the proposed Augmented GAN method over baseline GAN and GAN+Rotation for providing a useful representation of videos, specially for Ball-Drop dataset which is the focus of this paper. It is also interesting to see that in UCF101 dataset, GAN+Temporal outperforms GAN+Spatial and even GAN-SpatioTemporal.

Following that, an ablation study is carried out to investigate the effect of various spatial transformations used in the proposed method on downstream classification accuracy. The Augmented GAN method was trained in the first step using only one spatial transformation (rotation, translation or shearing). Then two transformations were used, and finally all three. Table 4.2 shows the top 1 classification accuracy of using features extracted from these methods applied to the Ball-Drop dataset. Although rotation outperforms other transformations like translation and shearing when used alone, combining different spatial transformations yields the best results.

Method	Ball-Drop
GAN	77.93 ± 2.7
GAN+Rotate	80.47 ± 2.5
GAN+Translate	80.04 ± 3.3
GAN+Shear	79.52 ± 3.3
GAN+Rotate+Translate	81.32 ± 5.1
GAN+Translate+Shear	80.33 ± 3.1
GAN+Rotate+Shear	81.01 ± 4.6
GAN+Rotate+Translate+Shear	81.99 ± 4.5

Table 4.2: Impacts of different transformation combinations on classification accuracy for the Ball-Drop dataset in [122]. It demonstrates that no spatial transformation is redundant.

4.2 Contrastive Approach

Contrastive learning (CL) is a self-supervised learning approach for grouping similar samples together and separating dissimilar samples. Its goal is to train a model to distinguish between positive and negative samples. As a result, the model learns input representations that it can use in downstream tasks such as activity recognition or object detection [52, 73, 43]. Along with state-of-the-art contrastive methods, this work employs a new family of self-supervised methods that do not require a large number of negative samples and are thus easier to train [43, 23, 10]. This thesis proposes a novel Self-supervised architecture (Figure 1) for Human Activity Modeling (SelfHAM). It is the first time that a self-supervised approach is used to improve the accuracy of computer vision models used in Embodied Cognition assessment by utilizing publicly available unlabeled data.

for this method, the focus of the automated assessment system is on three core tasks: bi-manual ball pass, ball drop to the beat, and tandem gait forward.

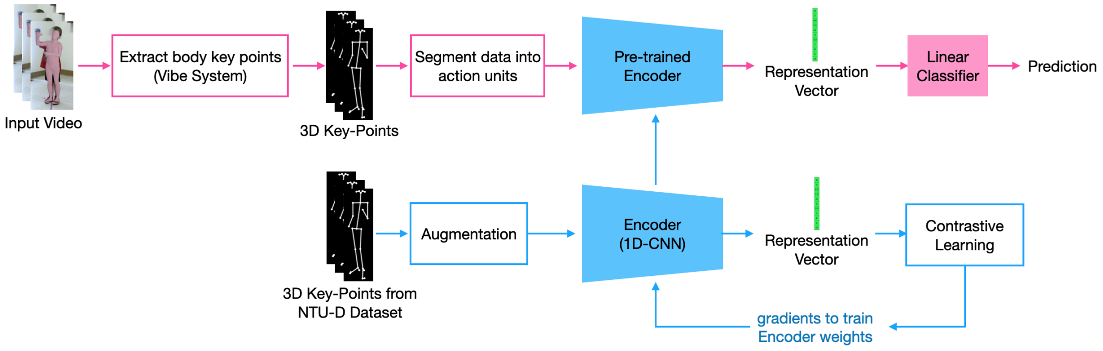


Figure 4.4: Proposed architecture for SelfHAM: **top** (pink)—supervised classification; **bottom** (blue)—self-supervised pre-training [123].

These tasks are part of a larger system called ATEC [12, 29, 95], which is described in detail in Chapter 2. In order to automatically evaluate a subject's performance, first the VIBE [61] human pose estimation system was used to extract 3D-body key-points. Then, a deep learning-based model was trained to classify subject actions. Furthermore, in order to improve the accuracy of the system, the model was pre-trained on the NTU-RGB+D 120 dataset [101, 71] and then fine-tuned on our ATEC dataset [12]. Three different state-of-the-art self-supervised learning methods, including those from MoCo [43], SimSiam [23], and VICReg [10], were employed to pre-train the model in a self-supervised manner, and their performances were compared to a supervised learning approach. The results show that a pre-trained model can outperform a supervised learning approach when a small amount of annotated data is available for training. It should be mentioned that all of the self-supervised methods used in this work were originally designed to extract features from still images, so we adapted them to extract representations from a sequence of human body key-points.

4.2.1 Proposed Method: SelfHAM

Methodology

Contrastive learning (CL) [52] tries to group similar (positive) samples closer and diverse (negative) samples further from each other. Representations are obtained by feeding input data into an encoder network. Contrastive learning focuses on comparing the representations with a variant of the noise contrastive estimation function [41] called InfoNCE [113], which is defined as follows:

$$L = -log \frac{exp(sim(q, k_+)/\tau)}{exp(sim(q, k_+)/\tau) + \sum_{i=0}^{K} exp(sim(q, k_i)/\tau)} \tag{4.4}$$

where q is the original sample (query), k_+ represents a positive sample, and k_i represents a negative sample. τ is a hyper-parameter used in most of the recent methods and is called the temperature coefficient. The *sim* function can be any similarity function, but generally a cosine similarity is used. The cosine similarity of two vectors is defined as the cosine of the angle between them. The initial idea behind noise contrastive estimation was to perform a non-linear logistic regression that discriminates between observed data and some artificially generated noise [52].

Since the number of negative samples affects the performance of CL methods [52], different strategies are used for selecting a large number of negative samples. In the first contrastive learning methods, a large batch size is used and all the samples in the batch except for the query and one positive sample are considered as negative. Because large batch sizes inversely affect optimization during training, one possible solution is to maintain a separate dictionary known as a memory bank containing representations of negative samples. However, since maintaining a memory bank during training is complicated, the memory bank can be replaced by a momentum encoder. The momentum encoder (MoCo) (Figure 4.5 left) generates a dictionary as a queue of encoded samples, with the current mini-batch enqueued and the oldest mini-batch dequeued [43]. The momentum encoder shares the same parameters as the query encoder (θ_q) and its parameters (θ_k) are updated based on the parameters of the query encoder ($\theta_k = m\theta_k + (1 - m)\theta_q, m \in [0, 1)$: momentum coefficient).

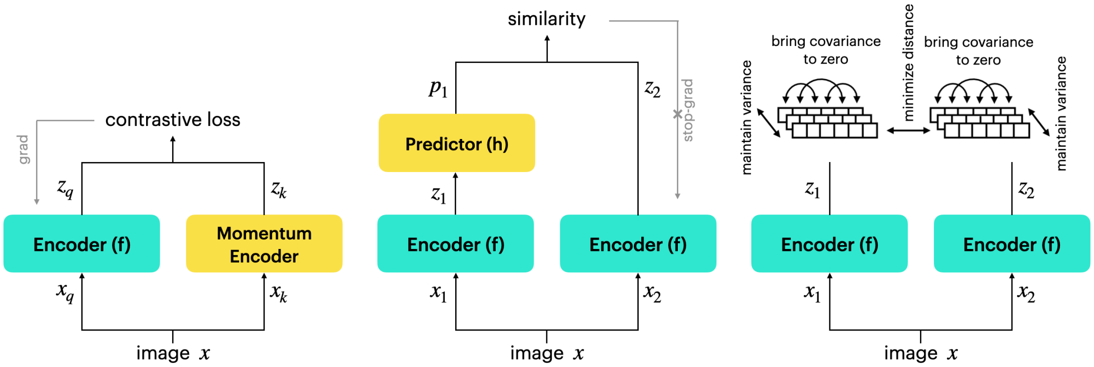


Figure 4.5: Different self-supervised learning architectures (x_1 and x_2 stand for different augmentations of image x): left—MoCo [43]; middle—SimSiam [23]; right—VICReg [10]

Most self-supervised methods involve specific forms of Siamese networks [14]. Weight-sharing neural networks with two or more inputs are known as Siamese networks. All outputs collapsing to a constant are undesirable trivial solutions to Siamese networks. There have been various general solutions for preventing the collapse of Siamese networks. The SimSiam method [23] proposed by Chen et al. prevents collapsing solutions by directly maximizing the similarity of an image's two views, using neither negative pairs [21], nor a momentum encoder [43]. The authors argue that stop-gradient operation is critical in preventing collapsing solutions. The SimSiam method architecture is depicted in Figure 4.5 (middle).

SimSiam methods take as input two randomly augmented views x_1 and x_2 from an image x. The two views are processed by an encoder network f. The encoder f shares weights between the two views. A prediction MLP network h matches the output of one view to the output of another view. The negative cosine similarity of two output vectors $p_1 = h(f(x_1))$ and $z_2 = f(x_2)$ is defined as follows:

$$D(p_1, z_2) = -\frac{p_1}{||p_1||_2} . \frac{z_2}{||z_2||_2} \tag{4.5}$$

where $||.||_2$ is l2-norm. Finally, the final loss function is defined below:

$$L = \frac{1}{2} D(p1, stopgrad(z2)) + \frac{1}{2} D(p2, stopgrad(z1)) \tag{4.6}$$

In the first term, the encoder on x_2 receives no gradient from z_2, but in the second term, it receives gradients from p_2 (and vice versa for x_1) [23]. The loss is calculated for each sample, and the total loss is averaged across all samples.

VICReg (Variance-Invariance-Covariance Regularization) proposed by Bardes et al. [10] is another self-supervised method tackling the collapsing solution problem. The VICReg architecture illustrated in 4.5 (right) is symmetric and is based on three simple principles: variance, invariance, and covariance. The variance principle is a simple but efficient strategy for preventing collapse by constraining the variance of the representations along each dimension independently. Without requiring any negative pairs, the invariance principle learns invariance to various views of an image employing a standard mean-squared Euclidean distance. Finally, the covariance principle uses the Barlow twins' covariance criterion [124], which decorrelates the different dimensions of learned representations with the goal of spreading information across dimensions and avoiding dimension collapse [10].

In the VICReg method, given an image x, two augmented views x_1 and x_2 are encoded using the encoder network f into representations $z_1 = f(x_1)$ and $z_2 = f(x_2)$. The overall loss function is a weighted average of the invariance, variance, and covariance terms [10]:

$$L(z_1, z_2) = S(z_1, z_2) + \lambda(V(z_1) + V(z_2)) + \gamma(C(z_1) + C(z_2)) \tag{4.7}$$

where λ and γ are hyper-parameters that regulate how important each phrase in the loss is. The overall objective function is computed as the sum of the loss function taken on all samples in the dataset. The variance, invariance, and covariance terms that make up the loss function are described here. First the variance term is defined as follows:

$$V(z) = \frac{1}{d} \sum_{i=1}^{d} max(0, 1 - \sqrt{Var(z) + \epsilon}) \tag{4.8}$$

where d is the dimension of feature vector z, ϵ is a small scalar for avoiding numerical instabilities, and $Var(x)$ is the unbiased variance estimator. Inspired by the Barlow twins [124], the covariance regularization term C is defined as the sum of the squared off-diagonal coefficients of covariance matrix of z ($Cov(z)$), with a factor $1/d$ that scales the criterion as a function of the dimension:

$$C(z) = \frac{1}{d} \sum_{i \neq j} Cov(z)^2_{i,j} \tag{4.9}$$

This term makes the off-diagonal coefficients close to 0 in order to decorrelate the different dimensions of the projections and prevent them from encoding the same information. Finally, the invariance criteria S between Z_1 and Z_2 is determined as the mean-squared Euclidean distance between each pair of vectors:

$$S(z_1, z_2) = \frac{1}{n} \sum_{i} ||z_{1i} - z_{2i}||_2^2 \tag{4.10}$$

Proposed System

The architecture of the proposed computer vision system is depicted in Figure 4.4. First, the subject's 3D body key-points were extracted using the VIBE system [61]. VIBE (Video Inference for Body Pose and Shape Estimation) is a video pose and shape estimation method that predicts the parameters of SMPL body models for each frame of an input video. From these key-points, 17 of them including head, hands, hip, feet, and toes were selected. Finally, the extracted data were divided into equal segments (with overlap), with each segment corresponding to an action. The numbers of segments for each task were as follows: ball drop to the beat—16; tandem gait forward—8; and stand on one foot—10. Each segment ($X \in \mathbb{R}^{32\times51}$) included 32 samples with 51 features. The features were x,y,z coordinates for each of the 17 key-points rasterized into one vector. Then, the input was fed into an encoder network to obtain the compact representation $z \in \mathbb{R}^{256}$. Finally, a linear classifier was used to classify input segments into action classes according to each task (Figure 4.4).

In order to pre-train the classifier model, the publicly available NTU-RGB+D 120 [101, 71] was used. This dataset contains 120 action classes and 114,480 video samples. In this work, only 3D skeletal data were employed. Similar to the gait dataset, 17 equivalent key-points (head, hands, hip, feet, and toes) were selected. In this work, a four-layer 1D convolutional neural network (CNN) with one penultimate transformer layer [114] was used as the encoder network. For all self-supervised methods, a projector network consisting of three fully connected layers was used. The projector network mapped the representations to a higher dimension of 1024. The SimSiam method also incorporates a two-layer fully connected predictor network that acts as a bottleneck by decreasing dimension of feature vectors to 256 and increasing it back to 1024. All networks used in this work employed batch normalization, except for the last layer.

4.2.2 Results and Discussion

All of the models used in this work were trained using the Pytorch framework [88] for 200 epochs. Stochastic gradient descent (SGD) was employed as an optimizer with a batch size of 512, learning rate of 0.1, and weight decay of 1×10^{-4}. The learning rate followed a cosine decay schedule [74] with 10 warm-up epochs. Furthermore, for the contrastive learning method MoCo, the temperature hyper-parameter τ and momentum coefficient μ were chosen as 0.1 and 0.999, respectively. For the VICReg method parameters, λ and γ were chosen as 1.0 and 0.04, respectively.

For evaluating the performance of the proposed methods in the case of a small amount of annotated data, three scenarios were defined. In the first scenario, 50% of the data were used for training and the other 50% for testing. In the second scenario, 25% of the data were used for training and the remaining 75% for testing. Finally, for the third scenario, 10% of data was used for training and the remaining 90% was used for testing. The average classification accuracy was calculated by cross-validation. The results for the baseline supervised method are shown in the first row of Table 4.3.

It is clear from the results that the baseline method classification accuracy decreases as training set becomes smaller, whereas the self-supervised methods maintain their performance and even outperform the supervised methods. We also compared the proposed methods to a supervised multi-modal ap-

Table 4.3: Different methods' top-1 classification accuracy for different training/test splits of ATEC tasks: 50%— using 50% of the dataset for training and the remaining 50% for testing; 25%— using 25% of the dataset for training and the remaining 75% for testing; 10%—using 10% of the dataset for training and the remaining 90% for testing.

Approach	**Ball Drop to the Beat**			**Tandem Gait**			**Stand on One Foot**		
	50%	**25%**	**10%**	**50%**	**25%**	**10%**	**50%**	**25%**	**10%**
Supervised	**77.61**	61.98	54.79	74.81	60.21	51.72	88.98	79.36	76.93
Multimodal	75.29	52.77	48.13	69.52	55.46	51.11	87.82	75.55	71.96
MoCo	74.70	71.87	70.63	75.52	73.70	72.36	89.63	88.07	85.59
SimSiam	75.44	74.08	71.91	75.81	74.42	73.46	89.76	88.05	87.24
VICReg	77.11	**74.65**	**72.59**	**75.96**	**74.45**	**73.51**	**90.54**	**89.86**	**89.59**

proach that had previously been used successfully on the ball-drop-to-the-beat task [95]. Results show that among self-supervised methods, VICReg attains the highest overall accuracy. When the size of the training set was reduced from 50% of the total dataset to 10% in the ball-drop-to-the-beat task, the accuracy of the supervised approach dropped by about 23%, whereas the accuracy of the VICReg method dropped by about 5%. In the tandem gait task, when the size of the training set decreased from 50% to 10% of the total dataset, the accuracy of the supervised approach declined by around 23%, while the VICReg technique only dropped by about 3%. In the stand-on-one-foot task, the supervised approach's accuracy was reduced by roughly 12% when the training set was decreased from 50% to 10% of the whole dataset, whereas the VICReg technique's performance was lowered by just around 1%. One reason for the multimodal approach's poor performance is that it uses a more complicated model that includes optical flow and object location in addition to the human body skeleton, making it prone to over-fitting in cases with small training datasets.

To summarize, an integrated self-supervised system using publicly available unlabeled data was proposed to improve the accuracy of computer vision models used in Embodied Cognition assessment. It was able to achieve acceptable performance despite only being trained on 10% of the data. Furthermore, new neural network models that excel at extracting representations from a sequence of human body key-points were proposed for all of the self-supervised methods used in this work.

CHAPTER 5

PHYSIOLOGICAL DATA PROCESSING

5.1 Cognitive Fatigue detection from fMRI

Cognitive fatigue is defined as subjective lack of mental energy that will interfere with habitual and required activities of an individual. Although cognitive fatigue has been known as a symptom of neurological damage for over a century, but a precise and thorough model for studying cognitive fatigue remains elusive. One major hurdle is the absence of consistent correlation between objective measures of cognitive fatigue such as response time (RT), error rate (ER) or even brain lesions and subjective reports of fatigue [119]. As a result when participants are asked to perform a cognitive task repeatedly, they are likely to experience cognitive fatigue while their performance would not be affected.

Furthermore, another limiting factor in cognitive fatigue research has been the fact that historically, researchers have evaluated subjective fatigue using scales which assess trait fatigue, rather than state fatigue. For example, one commonly used fatigue questionnaire is the Fatigue Severity Scale (FSS), which asks subjects to rate their fatigue over the past week. Thus, the FSS is a measure of trait fatigue, or the degree of fatigue that subject are vulnerable to experience over an extended period of time. On the contrary, assessment of state fatigue, or the extent to which subjects are experiencing fatigue at the moment of assessment, may more consistently correlate with behavior because it is measured at the same time the behavior is measured. [119]

Although there are plenty of approaches to evaluate cognitive fatigue through various tasks and assessment tests [28, 94], using objective measures, such as Response Time (RT) and Error Rate (ER), have certain limitations. During an assessment, objective measures of cognitive fatigue do not suffer even if the self-reported fatigue increases. Therefore, self-reported fatigue scores cannot be used as a reliable assessment of cognitive fatigue. In this work [120], response in brain activity was investigated through Functional Magnetic Resonance Imaging (fMRI) data for healthy individuals when they experience cognitive fatigue. The main hypothesis is that, as the cognitive fatigue increases, there would be an increase in the brain activation even if the subjects performance, such as RT and ER, does not change.

In order to induce cognitive fatigue, a cognitive task called N-back task is used. In the N-Back task, participants are presented a sequence of stimuli one-by-one. For each stimulus, they need to decide if the current stimulus is the same as the one presented N trials ago, where $N = 0, 1, 2, ...$. The higher the number, the more difficult the task is. The factors that seem to impact the performance are not only the N, but also the speed of presentation and the size of the stimuli set. For our experiments, the N is chosen to be 0 and 2.

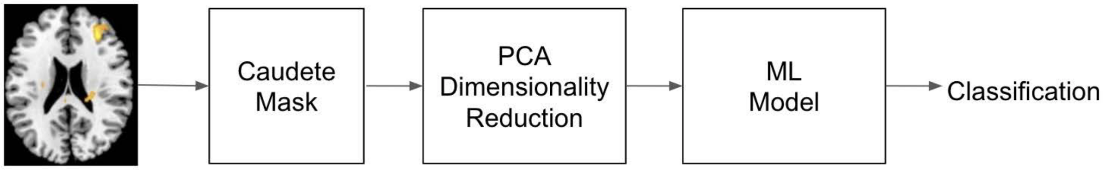


Figure 5.1: Overview of the proposed System for detection of CF from fMRI images [120].

Figure 5.1 presents an overview of the proposed system for cognitive fatigue assessment. Based on prior research [119, 62], it was identified that the Caudate nucleus of the basal ganglia in the brain is the Region of Interest (RoI) when it comes to assessment of cognitive fatigue. Thus, input fMRI images were multiplied by Caudate mask, which is a spatial mask with binary values. Since the dimension of Caudate mask used in this experiment was higher than input image, it was down-sampled. Down-sampling the mask directly, totally erased some parts of mask, so the mask was first dilated and then down-sampled. Applying the mask, allows to extract a more focused region of the scan.

In order to further extract the most contributing region, dimensionality reduction was performed with Principle Component Analysis (PCA) [91] and the most relevant principle components were selected. Two Machine Learning models were trained on the data including; Logistic Regression, and Convolutional Neural networks (CNN). For CNN model, 3D Convolutional Neural network was applied [42] directly on input images and the spatial dimensionality reduction by PCA was skipped.

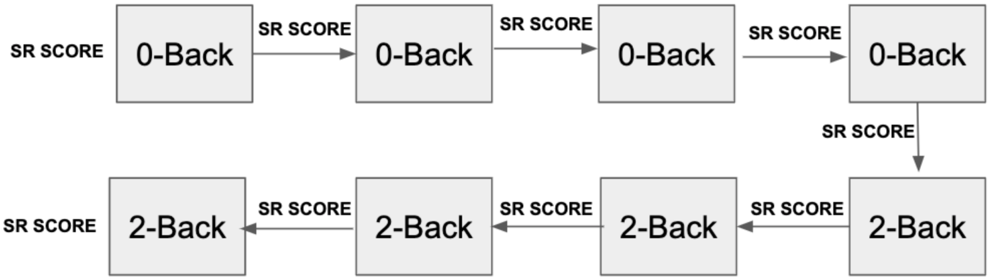


Figure 5.2: Data collection setup [120]. Subject are required to participants in eight N-back tasks (four 0-backs and four 2-backs) while fMRI data being recorded. After each trial a subjective fatigue score (SR) was self reported.

For the data collection, 22 participants participated. Their average age was 41 years, and 12 of them were female. Each participant performed four repetition of each task with either first four trials being the 0-back task or 2-back task

Method	Test Accuracy
Logistic Regression	73%
Convolutional Neural Networks	34%

Table 5.1: Experimental results for the proposed methods for detection of CF from fMRI images [120].

and the following four trials being the other, while fMRI data being recorded. Figure 5.2 represents the data collection procedure. As represented in the figure, fatigue scores (SR) were self-reported by the participants after every trial. The participants were asked to provide a fatigue SR between 0 to 100 after every trial and the scores were converted into 5 classes for the classification with the fatigue ranging from no fatigue to severe fatigue. Each fMRI data block is a 4D tensor with the dimension $135 \times 54 \times 64 \times 50$ where 135 represents the number of frames in time.

After applying the Caudate mask, both temporal and spatial dimensionality reduction was performed. For spatial dimensionality reduction, multiple methods were attempted with PCA working the best for the given scenario. For each sample, the block was rasterized into a vector of size, 2592000 ($54 \times 64 \times 50$) and the top 72 principal components were chosen after applying PCA. In order to perform temporal augmentation on the data, chunks of size equal to 72 were extracted for every epoch. Hence, the final dimension of the data being fed to the model were of dimension $72x72$. All the hyper-parameters were chosen empirically by employing grid search.

The data was split into train and validation set with the ratio of 4 to 1 with k-fold cross validation (k=5). Table 5.1 presents the average validation accuracy of different machine learning methods on predicting the fatigue score. From the preliminary results it can be observed that, the logistic regression method achieves the highest accuracy while CNN model does not work well. One of the reasons could be because of the small size of the dataset, the model would over-fit and not be able to learn any relevant features from the training set. One possible avenue for future research would be to perform similar experiments on patients affected with Traumatic Brain Injury (TBI) and Multiple Sclerosis (MS) to evaluate the robustness of the proposed model.

5.2 Assessment of Fatigue using wearable sensors

The development of an experimental apparatus for causing cognitive fatigue (CF) through a variety of cognitive and physical tasks while simultaneously capturing physiological data is the main goal of this work. Furthermore, the self-reported visual analog scores (VAS) of participants are reported following each activity to validate the induction of fatigue. Last but not least, an assessment system is created that uses machine learning (ML) models to detect CF conditions from sensor data, offering an objective evaluation [53].

As shown in Figure 5.3, an experimental setup was constructed around a custom-built wearable t-shirt (Pneumon) that was used to record physiological data using the attached sensors and a MUSE 2.0 headband. In two separate study sessions, data from 32 healthy people (18-33 years old, average age 24 years, 28/72% female/male) were collected. A MUSE 2.0 headband sensor was used to collect brain EEG data. Throughout the experiment, participants were required to wear the t-shirt and the MUSE headband. The researchers began by taking a baseline reading from the sensors for one minute while the subject

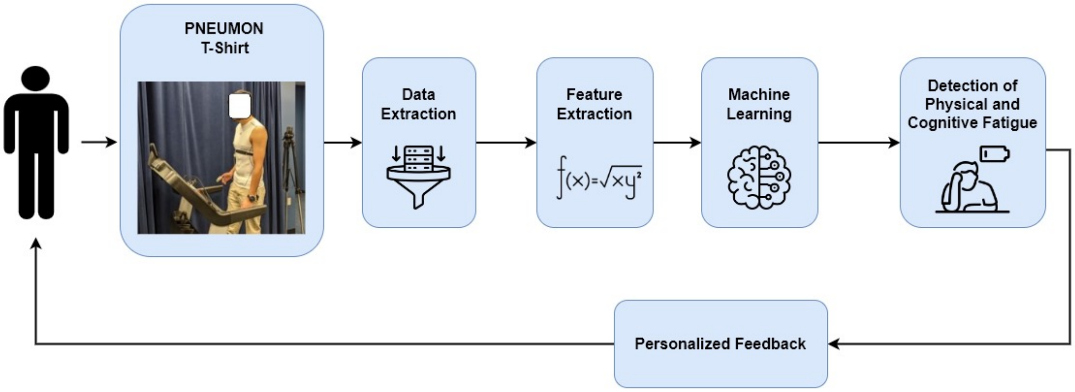


Figure 5.3: Flow diagram of the tasks performed by a participant [53].

stood still. The experiment's goal was to induce Cognitive Fatigue (CF) while simultaneously collecting sensory data. Following that, the participants were asked to complete several sets of N-Back tasks to induce CF, as shown in Figure 5.4. In these tasks, the subject is shown a series of letters one after the other. The subject's goal is to see if the current letter matches the letter shown N steps back. If it does, the subject must carry out the specified action (pressing the space bar on the keyboard). After the subjects stood still for 90 seconds after completing the physical task, additional data was collected (sensor reading 3 in Figure 5.4 [53].

The study was split into two separate sessions on different days for each participant. Both a morning and an evening session were required of the subjects. The impact of the time of day on the data was eliminated. The order in which the tasks were finished was the only distinction between the two sessions. In contrast to the second session, which gave preference to the intellectually demanding 2-Back game over the physically demanding activity, the first session followed the flow depicted in Figure 5.4. Participants were required to answer

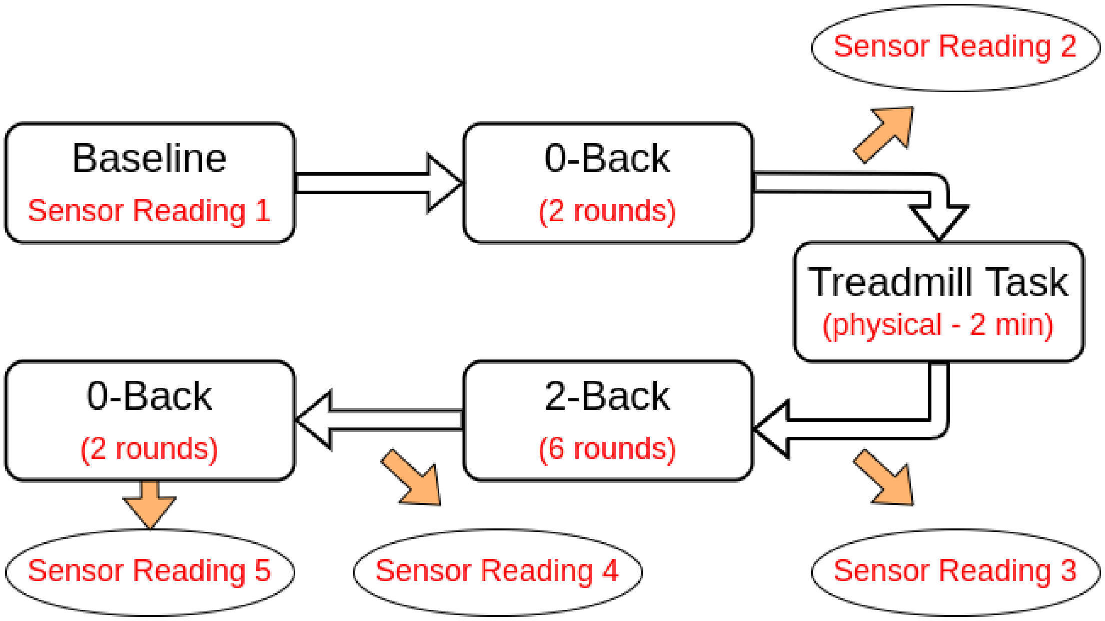


Figure 5.4: System Flow Diagram: Data collection using the sensors attached on the PNEUMON t-shirt and MUSE 2.0 worn by one of the participants while performing tasks presented in Fig. 5.3. Features extracted from the recorded signals were used to train ML models for detection of state of CF [53].

a brief survey detailing their current degrees of physical and mental exhaustion after completing each assignment. Additionally, they offered VAS scores, which ranged from 1 to 10. According to the survey results, CF appears to be induced in more than 80% of the subjects after the fourth block, supporting the hypothesis on which our experimental setup was based [53].

We collected EEG signals using the MUSE 2.0 headset while the subjects performed various tasks during the experiment. It is made up of four electrodes (FP1, FP2, TP9, TP10) that make contact with different parts of the head, as shown in Figure 5.5 (d). EEG signals are used to measure electrical activity in the brain. The wearable t-shirt contained physiological sensors that collected ECG, EDA, and EMG signals simultaneously throughout the experiment, as

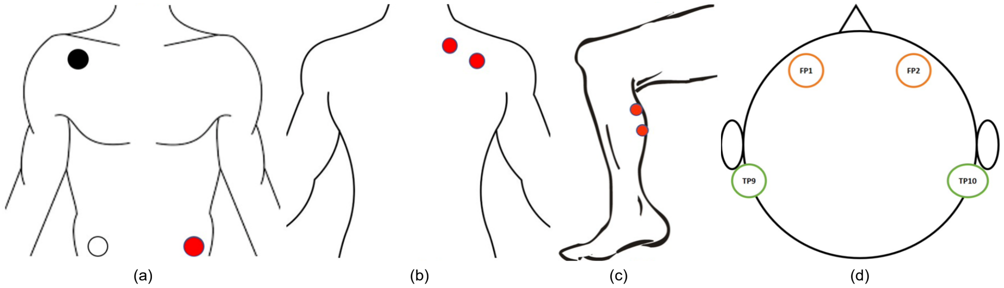


Figure 5.5: Sensor placements on the human body (a) **ECG**: right shoulder, the left and the right hip forming Einthoven's triangle [33] (b) **EDA/GSR** electrodes on the left shoulder to record the skin conductivity, (c) **EMG** electrodes recording muscle twitches from the right calf, (d) **EEG** sensor positions in the 10-10 electrode system used by MUSE. It records data from the TP9, FP1, FP2, and TP10 positions in the system [53].

shown in Figure 5.5 (a-c). Fatigue can harm the cardiovascular, endocrine, and nervous systems. As a result, these multi-modal signals aid in the tracking of the subject's physical state and can provide accurate information on whether the person is fatigued or not. ECG signals, which reflect the electrical activity of the heart, reveal changes in the cardiovascular system. EDA, on the other hand (also known as galvanic skin response, or GSR) measures the skin conductivity of the body to reflect the activity of the sympathetic nervous system, which is dependent on physiological and emotional activation. Finally, EMG signals measure the voltage difference between two electrodes as muscles contract and relax [53].

For training the ML models, we extracted 100 domain and frequency features from EEG signals and 169 combined features from ECG, EDA, and EMG using the neurokit2 framework [78]. Data collected in sensor readings 1, 2, and 3 (before the 2-Back tasks in Figure 5.4) were labeled as ”No CF” condition based on participant responses. The data collected during the final two readings (4 and 5, following the 2-Back rounds) were labeled as ”CF” conditions. Instead of processing the entire signal for a task as a single input, we divide the time signal into multiple slices based on different window sizes for training (5 seconds, 10 seconds, 20 seconds). Each signal slice received the same label as the original signal, and features were extracted. During training, it also increased the number of input data points to the ML models. Similarly, the input signal was divided into smaller slices during inference based on the window size chosen during training, and each slice was classified as one of the classes by the model. Finally, the entire signal block was classified based on the class with the highest count among the classified slices. Because noise in some of the slices may not contribute significantly to the final classification result, this technique makes the model more robust to noise or outliers in the signals [53].

The dataset was randomly divided into three sets: train (70%, 22 subjects), validation (15%, 5 subjects), and test (15%, 5 subjects). In addition, each model was subjected to 5-fold cross-validation. In the analysis, three different ML models were used: Logistic Regression (Log Reg.), Support Vector Machines (SVM) and Random Forest (RF). The features extracted from the signals were used to train the first three models. Finally, for the detection of CF, features from all data modalities were combined and normalized, as shown in Table 5.2 [53].

The average recall (Avg. Recall) shown in all four tables is the average recall obtained through 5-fold cross-validation for each model. The best value obtained for each model among different window sizes was considered. Because our primary goal is to detect true fatigue conditions in subjects while avoiding false negatives, recall is critical. Table 5.2 shows that RF performs the best, with

Table 5.2: Detection of CF with EEG + ECG + EDA + EMG Features [53].

Model	Accuracy (Window Size)				Avg.
	5s	10s	20s	Full Block	Recall
Log Reg.	64.0%	66.9%	66.1%	60.4%	0.69
SVM	70.3%	74.6%	74.5%	70.3%	0.79
RF	67.9%	**77.2%**	76.8%	74.5%	**0.81**

a CF prediction accuracy of 77.2% and an average recall of 81%. Furthermore, we can confirm that a window size of 10s appears to be the most effective.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this work, the technical description of ATEC, an integrated computer-vision system for assessing embodied cognition in children with executive function disorder, is presented. The ATEC system includes both recording and administrative interfaces, which were designed to streamline the assessments without any interruptions. This system only records video data, since sensor-based data collection can be more expensive and impractical with child participants. The ATEC system consists of variety of physical exercises with different variations and difficulty levels designed to provide assessment of executive and motor functions. The main focus of this work was applying self-supervised visual representation learning for human activity recognition in ATEC system recorded videos. Finding an effective human activity representation will help us to improve the accuracy of the automated computer-vision system with much less annotated training data.

In order to improve the performance of the proposed system, we tried to pre-train the encoder network on large public dataset (NTU) by using self-supervised learning. Different self-supervised methods were investigated to obtain the best representations. The results supported our claim that pre-trained models can outperform supervised learning approaches when small amounts of annotated data are available for training. When the size of the training set was reduced from 50% of the total dataset to 10% in the ATEC task, the accuracy of the supervised approach dropped by about 20%, whereas the accuracy of the self-supervised methods dropped by less than 5%. Improving the efficacy of the proposed approach in order to deploy it in real-world applications, as well as applying it to the remaining ATEC tasks, such as sailor step [12, 29], finger-oppose [8], etc., will be the focus of future works. The ultimate goal of this work is to create a comprehensive digital phenotyping framework capable of collecting multimodal data from a variety of sensors such as cameras, wearables, and so on, in order to monitor human behavior.

BIBLIOGRAPHY

[1] Embodied cognition. `https://en.wikipedia.org/wiki/Embodied_cognition`.

[2] A survey on big data-driven digital phenotyping of mental health. *Information Fusion*, 52:290 – 307, 2019.

[3] T. Achenbach and T. M. Ruffle. The child behavior checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatrics in review*, 21 8:265–71, 2000.

[4] Unaiza Ahsan, Chen Sun, and Irfan Essa. Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks, 2018.

[5] Sandra Aigner and Marco Körner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans, 2018.

[6] 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*, 13(1):155–167, 1985.

[7] Dennis Ausiello and Stanley Shaw. Quantitative human phenotyping: The next frontier in medicine. *Transactions of the American Clinical and Climatological Association*, 125:219–28, 08 2014.

[8] 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)*, pages 164–169, Aug 2019.

[9] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1949–1957, 2015.

[10] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2021.

[11] R. A. Barkley. Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of adhd. *Psychological bulletin*, 121(1):65, 1997.

[12] Morris D. Bell, Andrea J. Weinstein, Brian Pittman, Richard M. Gorman, and Maher Abujelala. The activate test of embodied cognition (atec): Reliability, concurrent validity and discriminant validity in a community sample of children using cognitively demanding physical tasks related to executive functioning. *Child Neuropsychology*, pages 1–11, 2021. PMID: 33985422.

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

[14] Jane Bromley, James Bentz, Leon Bottou, Isabelle Guyon, Yann Lecun, Cliff Moore, Eduard Sackinger, and Rookpak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:25, 08 1993.

[15] Thomas Brox, Andres Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 3024, pages 25–36, 01 2004.

[16] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019.

[17] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

[18] Javier Selva Castelló. A comprehensive survey on deep future frame video prediction. Master's thesis, Universitat de Barcelona, The address of the publisher, 1 2018. An optional note.

[19] R. Chan, D. Shum, T. Toulopoulou, and E. Y. Chen. Clinical test of apposition and counter-apposition of the thumb. *Annales de chirurgie de la main: organe officiel des societes de chirurgie de la main*, 5:67–73, 1986.

[20] R. Chan, D. Shum, T. Toulopoulou, and E. Y. Chen. Assessment of executive functions: review of instruments and identification of critical is-

sues. *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists*, 23 2:201–16, 2008.

[21] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[22] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss, 2018.

[23] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020.

[24] Kendra Cherry. What is cognition? `https://www.verywellmind.com/what-is-cognition-2794982`, 2020.

[25] Noam Chomsky. *Language and Mind*. Cambridge University Press, 2006.

[26] 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*, 52:65–69, 2011.

[27] Emma Davis, Nicola Pitchford, and Ellie 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 (London, England : 1953)*, 102:569–84, 08 2011.

[28] John DeLuca, Helen M. Genova, Frank G Hillary, , and Glenn Wylie. Neural correlates of cognitive fatigue in multiple sclerosis using functional mri. *Journal of the neurological sciences*, 270(1):28–39, 2008.

[29] Alex Dillhoff, Konstantinos Tsiakas, Ashwin Ramesh Babu, Mohammad Zakizadehghariehali, Benjamin Buchanan, Morris Bell, Vassilis Athitsos, and Fillia 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*, iWOAR '19. Association for Computing Machinery, 2019.

[30] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction, 2015.

[31] J. E. Donnelly and K. Lambourne. Classroom-based physical activity, cognition, and academic achievement. *Preventive medicine*, 52:36–42, 2011.

[32] 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*, 81(12):733–740, 2011.

[33] Willem Einthoven, G Fahr, and A De Waart. On the direction and manifest size of the variations of potential in the human heart and on the influence of the position of the heart on the form of the electrocardiogram. *American heart journal*, 40(2):163–211, 1950.

[34] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation, 2018.

[35] Annalisa Franco, Antonio Magnani, and Dario Maio. A multimodal approach for human activity recognition based on skeleton and rgb data. *Pattern Recognition Letters*, 131, 03 2020.

[36] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Two stream lstm: A deep fusion framework for human action recognition, 2017.

[37] Srujana Gattupalli, Ashwin Ramesh Babu, James Robert Brady, Fillia Makedon, and Vassilis Athitsos. Towards deep learning based hand keypoints detection for rapid sequential movements from rgb images, 2018.

[38] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018.

[39] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[40] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017.

[41] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

[42] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, 2018.

[43] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[45] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey, 2017.

[46] 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*, 215:59–68, 2012.

[47] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.

[48] Kit Huckvale, Svetha Venkatesh, and Helen Christensen. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *npj Digital Medicine*, 2:1–11, 09 2019.

[49] C. Hughes and A. Graham. Measuring executive functions in childhood: Problems and solutions? *Child and adolescent mental health*, 7(3):131—142, 2002.

[50] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[51] S.H. Jain, B.W. Powers, Hawkins J.B., and Brownstein J.S. The digital phenotype. *Nature Biotechnol*, 33:462–463, 5 2015.

[52] Ashish Jaiswal, Ashwin ramesh babu, Mohammad Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9:2, 12 2020.

[53] Ashish Jaiswal, Mohammad Zaki Zadeh, Aref Hebri, and Fillia Makedon. Assessing fatigue with multimodal wearable sensors and machine learning, 2022.

[54] Eya Kalanthroff, Eddy Davelaar, Avishai Henik, Liat Goldfarb, and Marius Usher. Task conflict and proactive control: A computational theory of the stroop task. *Psychological Review*, 125, 10 2017.

[55] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas P. J. J. Noldus, and Remco C. Veltkamp. Egocentric hand track and object-based human action recognition, 2019.

[56] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[57] Swapnali Karvekar. Smartphone-based human fatigue detection in an industrial environment using gait analysis. 2019.

[58] Arshia Khan, Janna Madden, and Kristine Snyder. Framework utilizing machine learning to facilitate gait analysis as an indicator of vascular dementia. *International Journal of Advanced Computer Science and Applications*, 9, 01 2018.

[59] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[60] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[61] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation, 2019.

[62] Alexander D. Kohl, Glenn R. Wylie, H.M. Genova, Frank G. Hillary, and J. Deluca. The neural correlates of cognitive fatigue in traumatic brain injury using functional mri. *Brain injury*, 23(5):420–432, 2009.

[63] Yu Kong and Yun Fu. Human action recognition and prediction: A survey, 2018.

[64] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[65] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in gans, 2018.

[66] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

[67] George Lakoff and Mark Johnson. *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. Basic Books, 2009.

[68] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(10):436–444, 2015.

[69] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction, 2018.

[70] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences, 2017.

[71] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[72] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016.

[73] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive, 2020.

[74] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[75] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study, 2017.

[76] HaoJie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. Attnsense: Multi-level attention mechanism for multimodal human activity recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3109–3115. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[77] Mehran Maghoumi and Joseph J. LaViola Jr au2. Deepgru: Deep gesture recognition utility, 2019.

[78] Dominique Makowski, Tam Pham, Zen Juen Lau, Jan Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53, 02 2021.

[79] Andrea Mannini, Diana Trojaniello, Andrea Cereatti, and Angelo Sabatini. A machine learning framework for gait classification using inertial sensors: Application to elderly, post-stroke and huntington's disease patients. *Sensors*, 16:1–14, 01 2016.

[80] Ryan McGinnis, Nikhil Mahadevan, Yaejin Moon, Kirsten Seagers, Nirav Sheth, John Wright, Steve Dicristofaro, Ikaro Silva, Elise Jortberg, Melissa Ceruolo, Jesus Pindado, Jacob Sosnoff, Roozbeh Ghaffari, and Shyamal Patel. A machine learning approach for gait speed estimation using skin-mounted wearable sensors: From healthy controls to individuals with multiple sclerosis. *PLOS ONE*, 12:e0178366, 06 2017.

[81] Samuel McNerney. A brief guide to embodied cognition: Why you are not your brain. `https://blogs.scientificamerican.com/guest-blog/a-brief-guide-to-embodied-cognition-/why-you-are-not-your-brain/`, 2011.

[82] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification, 2016.

[83] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks, 2018.

[84] Sanghee Moon, Hyun-Je Song, Vibhash Sharma, Kelly Lyons, Rajesh Pahwa, Abiodun Akinwuntan, and Hannes Devos. Classification of parkinson's disease and essential tremor based on gait and balance characteristics from wearable motion sensors: A data-driven approach, 04 2020.

[85] V. Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[86] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles, 2016.

[87] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count, 2017.

[88] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[89] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move, 2016.

[90] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting, 2016.

[91] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space, November 1901.

[92] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning, 2020.

[93] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015.

[94] Akilesh Rajavenkatanarayanan, Varun Kanal, Maria Kyrarini, and Fillia 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*, page 398–400, 2020.

[95] Ashwin ramesh babu, Mohammad Zadeh, Ashish Jaiswal, Alexis Lueckenhoff, Maria Kyrarini, and Fillia Makedon. A multi-modal system to assess cognition in children from their physical movements. 10 2020.

[96] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.

[97] Bin Ren, Mengyuan Liu, Runwei Ding, and Hong Liu. A survey on 3d skeleton-based action recognition using learning method, 2020.

[98] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[99] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. volume 3, pages 32 – 36 Vol.3, 09 2004.

[100] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J. Black. On the integration of optical flow and action recognition. In *German Conference on Pattern Recognition (GCPR)*, volume LNCS 11269, pages 281–297. Springer, Cham, October 2018.

[101] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[102] T. Shallice and P. W. Burgess. Deficits in strategy application following frontal lobe damage in man. *Brain*, 114(2):727–741, 1991.

[103] Lauralee Sherwood. *Human Physiology: From Cells to Systems*. Cengage Learning, 2015.

[104] I. Shimoyama, T. Ninchoji, , and K. Uemura. The finger-tapping test: a quantitative analysis. *Archives of neurology,*, 47(6):681– 684, 1990.

[105] J.F. Smith, K. Chen, S. Johnson, J. Morrone-Strupinsky, E. M. Reiman, A. Nelson, J. R. Moeller, and G. E. Alexander. Network analysis of single-subject fmri during a finger opposition task. *Neuroimage*, 32(1):325–332, 2006.

[106] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

[107] Gijsbert Stoet. Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44, 11 2016.

[108] Eddy Sánchez-DelaCruz, Roberto Weber, Rajesh Biswal, Jose Mejia, Gandhi Hernández-Chan, and Heberto Gómez-Pozos. Gait biomarkers classification by combining assembled algorithms and deep learning: Results of a local study. *Computational and Mathematical Methods in Medicine*, 2019:1–14, 12 2019.

[109] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2019.

[110] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition, 2018.

[111] Trieu H. Trinh, Minh-Thang Luong, and Quoc V. Le. Selfie: Self-supervised pretraining for image embedding, 2019.

[112] L. G. Ungerleider, J. Doyon, , and A. Karni. Imaging brain plasticity during motor skill learning,. *Neurobiology of learning and memory*, 78(3):553–564, 2002.

[113] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

[114] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

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

[116] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, Mar 2019.

[117] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks, 2017.

[118] B. E. Wexler. Integrated brain and body exercises for adhd and related prob- lems with attention and executive function. *nternational Journal of Gaming and Computer-Mediated Simulations*, 5(3):10–26, 2013.

[119] G.R. Wylie, E. Dobryakova, J. DeLuca, N. Chiaravalloti, K Essad, and H Genova. Cognitive fatigue in individuals with traumatic brain injury is associated with caudate activation. *Scientific reports*, 7(1):1–12, 2017.

[120] Mohammad Zaki Zadeh, Ashwin Ramesh Babu, Jason Bernard Lim, Maria Kyrarini, Glenn Wylie, and Fillia Makedon. Towards cognitive fatigue detection from functional magnetic resonance imaging data. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA ’20, New York, NY, USA, 2020. Association for Computing Machinery.

[121] Mohammad Zaki Zadeh, Ashwin Ramesh Babu, Ashish Jaiswal, Maria Kyrarini, Morris Bell, and Fillia Makedon. Automated system to measure tandem gait to assess executive functions in children. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, PETRA 2021, page 167–170, New York, NY, USA, 2021. Association for Computing Machinery.

[122] 'Mohammad' 'Zaki Zadeh', Ashwin Ramesh Babu, Ashish Jaiswal, Maria Kyrarini, and Fillia Makedon. Self-supervised human activity recognition by augmenting generative adversarial networks. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, PETRA 2021, page 171–176, New York, NY, USA, 2021. Association for Computing Machinery.

[123] Mohammad Zaki Zadeh, Ashwin Ramesh Babu, Ashish Jaiswal, and Fillia Makedon. Self-supervised human activity representation for embodied cognition assessment. *Technologies*, 10(1), 2022.

[124] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

[125] P. D. Zelazo, J. E. Anderson, J. Richler, K. Wallner-Allen, J. L. Beaumont, and S. Weintraub. Nih toolbox cognition battery (cb): Measuring executive function and attentionk. *Monographs of the Society for Research in Child Development*, 78(4):16–33, 2013.

[126] H. Zhang, Yi-Xiang Zhang, B. Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. A comprehensive survey of vision-based human action recognition methods. *Sensors (Basel, Switzerland)*, 19, 2019.

[127] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2018.

[128] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization, 2016.

[129] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey, 2019.