



Ain Shams University
Faculty of Computer & Information Sciences
Artificial Intelligence Department

Intelligent Aided System to Support Children with Autism

By

Antony Nabil Naguib [AI]

Alzahraa Mofed Mohamed [AI]

Amira Barakat Ali [AI]

Aya Khaled Mohamed [AI]

Nagwa Mohamed Anwar [AI]

Under Supervision of

prof. Abeer Mahmoud

[professor],

CS Department,

Faculty of Computer and Information Sciences,

Ain Shams University.

Andrew Magdy

[Assistant Lecturer],

SC Department,

Faculty of Computer and Information Sciences,

Ain Shams University.

June 2023

Acknowledgements

All praise and thanks to ALLAH, who provided me with the ability to complete this work. I hope to accept this work from me.

I am grateful of *my parents, my family, my friends* and *my doctors* who are always providing help and support throughout the whole years of study. I hope I can give that back to them.

I also offer my sincerest gratitude to my supervisors, *Prof. Dr. Abeer Mahmoud, and T.A. Andrew Magdy*. who have supported me throughout my thesis with their patience, knowledge and experience.

Finally, I would like to thank all the people who gave me support and encouragement.

Abstract

Autism is a condition that affects how a person thinks, feels, interacts with others, and experiences their environment. It is a lifelong disability that starts when a person is born and stays with them into old age.

Every Autistic person is different to every other. This is why autism is described as a ‘spectrum’.

People with Autism Spectrum Disorder (ASD) may behave, communicate, interact, and learn in ways that are different from most other people. There is often nothing about how they look that sets them apart from other people.

Why we thought about this idea?

- Therapists are not always available.
- Cost efficient and more affordable.
- Autism spectrum disorders are common and some people lack the knowledge to properly deal with them.

Main Features:

- Face recognition
- Facial expression recognition
- Speech-to-Text
- Text-to-Speech
- Learning videos
- Games using face, hands, body pose landmarks and objects detection

The final result is a system that interacts with autistic people and allows them to do some exercises. By so doing this it will contribute in engaging them in significant activities instead of spending most of their time alone, therefore it could be a means of improving their condition by having such interaction.

التوحد هو حالة تؤثر على كيفية تفكير الشخص وشعوره وتفاعله مع الآخرين وتجاربه مع بيئته. إنها إعاقة مدى الحياة تبدأ عند ولادة الشخص وتستمر معه إلى الشيخوخة.

كل شخص توحدي مختلف عن الآخر. ولهذا السبب يتم وصف التوحد بأنه "طيف".

قد يتصرف الأشخاص الذين يعانون من اضطراب طيف التوحد (ASD) ويتواصلون ويتعلمون بطرق مختلفة عن معظم الأشخاص الآخرين، دون أن يكون هناك شيء يميزهم عن الآخرين من حيث الشكل.

لماذا فكرنا في هذه الفكرة؟

- ليس دائماً متاحاً لديهم العلاج اللازم.
- كفاءة التكلفة وأكثر اقتصادية.
- اضطرابات طيف التوحد شائعة وبعض الأشخاص يفتقرون إلى المعرفة اللازمة للتعامل معها بشكل صحيح.

الميزات الرئيسية:

- التعرف على الوجه
- التعرف على تعابير الوجه
- التحويل من الكلام إلى النص
- التحويل من النص إلى الكلام
- مقاطع فيديو تعليمية
- ألعاب تستخدم تقنية التعرف على الوجه واليدين ووضعيات الجسم وكشف الأجسام

النتيجة النهائية هي نظام يتفاعل مع الأشخاص ذوي التوحد ويسمح لهم بممارسة بعض التمارين. وبذلك، سيساهم في إشراكهم في أنشطة هامة بدلاً من قضاء معظم وقتهم بمفردهم، وبالتالي يمكن أن يكون وسيلة لتحسين حالتهم من خلال هذا التفاعل.

Table of Contents

Acknowledgements.....	2
Abstract.....	3
List of Figures.....	7
List of Tables.....	10
List of Abbreviations.....	11
Chapter 1: Introduction.....	12
1.1 Problem Definition.....	12
1.2 Motivation.....	13
1.3 Objectives.....	15
1.4 Methodology	15
1.5 Time plan.....	15
1.6 Thesis Outline.....	16
Chapter 2: Literature Review	17
Chapter 3: System Architecture and Methods	31
3.1 System Architecture	31
3.2 Description of methods and procedures used	33
Chapter 4: System Implementation and Results	55
4.1 Dataset	55
4.2 Description of Software Tools Used	56
4.3 Stepup Configuration (hardware).....	56
4.4 Experimental and Results	57
Chapter 5: Run the Application.....	60

Chapter 6: Conclusion and Future Work.....	69
6.1 Conclusion.....	69
6.2 Future Work.....	70
References.....	71

List of Figures

Figure 1.1: Autism prevalence rate through the years.....	13
Figure 1.2: Autism rates across countries and some US states.....	14
Figure 1.3: Time plan.....	15
Figure 2.1: Scene understanding cycle.....	23
Figure 2.2: CNN Architecture.....	24
Figure 2.3: YOLO v2 architecture.....	25
Figure 2.4: Pose estimation and skeletal structure.....	26
Figure 2.5: QTRobot.....	29
Figure 2.6: QTRobot Advantages and Disadvantages.....	30
Figure 3.1 System architecture.....	31
Figure 3.2 available videos and games.....	31
Figure 3.3: LBPH initial steps.....	34
.Figure 3.4: LBPH pixels calculations.....	34
Figure 3.5: LBPH histogram extraction.....	35
Figure 3.6: Haar features.....	36
Figure 3.7: Haar calculations.....	37
Figure 3.8: Haar integral image.....	38
Figure 3.9: Haar integral calculations.....	38
Figure 3.10: Attentional cascade.....	40
Figure 3.11: digital image.....	41
Figure 3.12: convolution.....	42
Figure 3.13: max pooling.....	43
Figure 3.14: YOLO object detection.....	44
Figure 3.15: YOLOv3 architecture.....	45
Figure 3.16: Non-Max Suppression.....	46

Figure 3.17: Pose detection.....	47
Figure 3.18: Pose detection landmarks.....	48
Figure 3.19: SSD bounding box.....	49
Figure 3.20: Estimator architecture.....	49
Figure 3.21: face landmark detection.....	50
Figure 3.22: BlazeFace inference result.....	51
Figure 3.23: BlazeBloch and double BlazeBlock.....	52
Figure 3.24: hand landmark model results.....	53
Figure 3.25: hand landmarks.....	54
Figure 4.1: dataset one.....	55
Figure 4.2: dataset two.....	55
Figure 4.3: dataset three.....	56
Figure 4.4: child trying virtual paint.....	58
Figure 4.5: child trying snake game.....	59
Figure 5.1: greeting screen.....	60
Figure 5.2: ready screen.....	60
Figure 5.3: face recognition.....	61
Figure 5.4: user chooses system mode.....	61
Figure 5.5: user chooses learn or play.....	62
Figure 5.6: educational video.....	62
Figure 5.7: alphabet video.....	63
Figure 5.8: colors video.....	63
Figure 5.9: numbers from 0 to 10 video.....	63
Figure 5.10: numbers from 10 to 100 video.....	63
Figure 5.11: emojis video.....	63
Figure 5.12: shapes video.....	63
Figure 5.13: ablution video.....	64

Figure 5.14: prayers video.....	64
Figure 5.15: fruits video.....	64
Figure 5.16: game list.....	65
Figure 5.17: spinball game.....	65
Figure 5.18: car game main menu.....	66
Figure 5.19: car game.....	66
Figure 5.20: snake game.....	66
Figure 5.21: paint game.....	67
Figure 5.22: facial expression (neutral).....	68
Figure 5.23: facial expression (happy).....	68

List of Tables

Table 2.1: summary of all applications investigated.....	27
Table 3.1: Inference speed across several mobile devices.....	52
Table 4.1: used datasets.....	55

List of Abbreviations

AI:	Artificial Intelligence
ASD:	Autism Spectrum Disorder
PDD-NOS:	Pervasive Developmental Disorder, Not Otherwise Specified
US:	United States
CDC:	Centers for Disease Control and Prevention
ADDM:	Autism and Developmental Disabilities Monitoring
CNN:	Convolutional Neural Network
YOLO:	"You Only Look Once" algorithm
ML:	Machine Learning
DSM-5:	Diagnostic and Statistical Manual of Mental Disorders, 5 th edit
ABA:	Applied Behavioral Analysis
MIT:	Massachusetts Institute of Technology
CEO:	Chief Executive Officer
CAL:	Computer Assisted Learning
DOF:	Degrees Of Freedom
LBPH:	Local Binary Pattern Histograms
IJRTE:	International Journal of Recent Technology and Engineering
ADHD:	Attention Deficit Hyperactivity Disorder

Chapter 1

Introduction

1.1 Preface

- The abilities of people with ASD can vary significantly. For example, some people with ASD may have advanced conversation skills whereas others may be nonverbal. Some people with ASD need a lot of help in their daily lives; others can work and live with little to no support.
- A child with ASD may:
 - Not want to be touched.
 - Want to play alone.
 - Not want to change routines.
- Until recently experts talked about different types of autism which all now fall under the umbrella of autism spectrum disorder. However, there is very little known about the causes of these disorders. these types are:
 - Asperger's syndrome. This is on the milder end of the autism spectrum. A person with Asperger's may be very intelligent and able to handle their daily life. They may be really focused on topics that interest them and discuss them nonstop. But they have a much harder time socially.
 - Pervasive developmental disorder, not otherwise specified (PDD-NOS). This diagnosis is a middle ground as it's given for those whose autism was more severe than Asperger's syndrome, but not as severe as autistic disorder. This older term is further along the autism spectrum than Asperger's and PDD-NOS. It includes the same types of symptoms, but at a more intense level.
 - Childhood disintegrative disorder. This was the rarest and most severe part of the spectrum. It described children who develop normally and

then quickly lose many social, language, and mental skills, usually between ages 2 and 4.

1.2 Motivation

Autism Spectrum Disorders (ASDs) are characterized by impairments in social interaction, communication and behavioral functioning that can affect the health-related quality-of-life outcomes of the affected child and the family. ASDs have increased in prevalence, leading to a demand for improved understanding of the comparative effectiveness of different pharmacologic, behavioral, medical, and alternative treatments for children as well as systems for providing services.

1 in 44 (or 2.3%) of children in the US were identified with ASD using estimates from CDC's Autism and Developmental Disabilities Monitoring (ADDN) Network. The 2021 prevalence estimate from data collected in 2018 is roughly 241% higher than estimates from 2000. The last estimate, reported in 2020, showed 1 in 54 kids identified with ASD. A mere year later the reported estimate increased to 1 in 44*

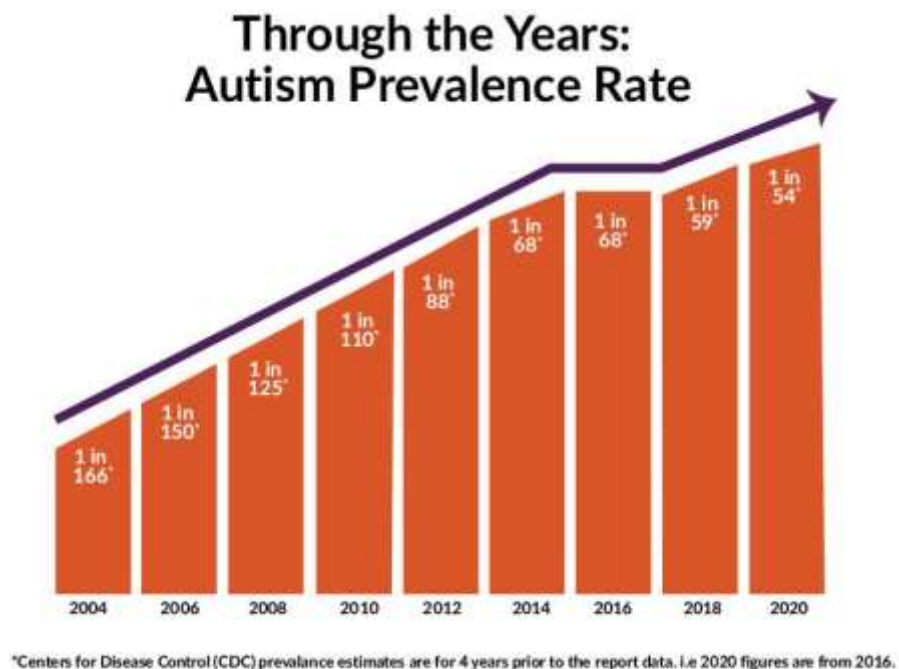


Figure 1.1: Autism prevalence rate through the years

Research suggests autism rates vary greatly across countries—and even among states in the US—cultural.

Examples include California at 3.9%, Missouri at 1.7%, South Korea at 2.84%, France at 0.36%, and Qatar at 1.14%. This variation is due to the fact that many developing countries do not have reliable autism statistics due to a lack of resources. This means most autism research is from affluent, English-speaking countries.

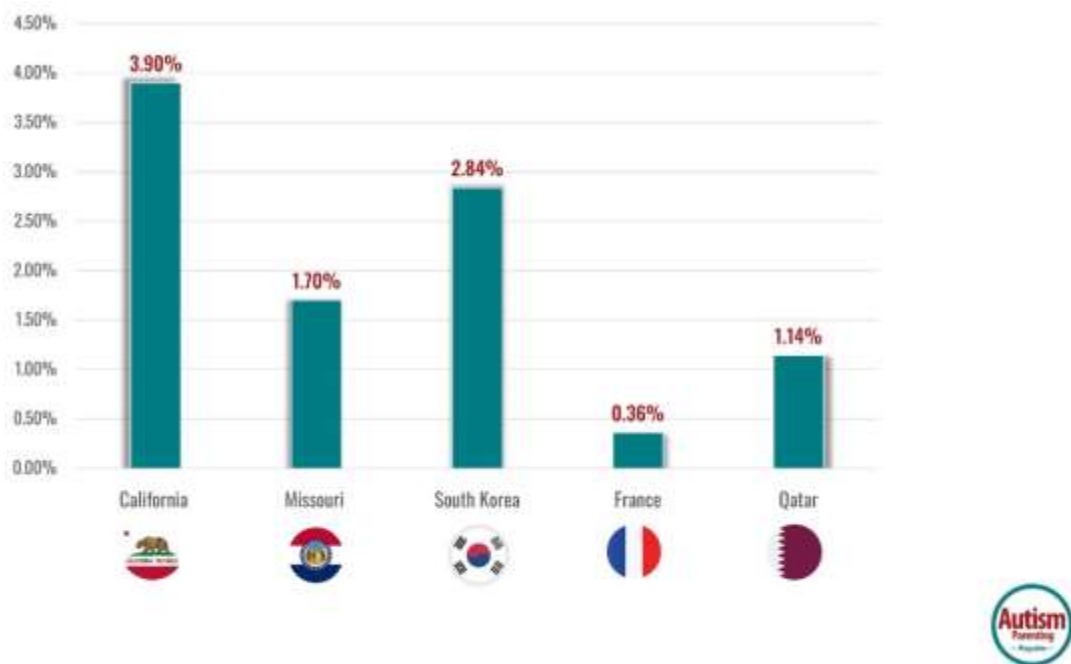


Figure 1.2: Autism rates across countries and some US states

Research suggests around 40% of autistic children and adolescents have at least one of anxiety disorders. Roughly half of all children with ASD may also experience symptoms of ADHD. Children on the spectrum are more likely than their peers to experience sleep, gastrointestinal, and weight management challenges. Children with ASD may also have an elevated risk of epilepsy. An estimated 90% of autistic individuals may have atypical sensory experiences. Individuals on the spectrum have a substantially heightened risk of dying from serious injury.

1.3 Objective

The aim of our Project is decreasing the percentage of autism for children and helping their families and people responsible for them by obtaining someone who sets and takes care of these children for certain periods and make these kids gain some skills. This could occur by designing an application showing some educational games. These games will allow children to do some exercises and improve their communication skills. Also, they aim to educate these children letters and some simple words in English. But they are not just any games, they are designed for kids who have autism and at the same time the application will be able to detect whether the kid is happy or sad and do actions based on that, also it will teach him the things in his surroundings and will try to communicate to him through voice. (Design requirements/constraints)

1.4 Methodology

The used scientific methods are machine learning such as Haar cascade classifier, deep learning such as CNN and YOLOv3 and computer vision tasks such as hand, face and pose landmarks task.

1.5 Time plan

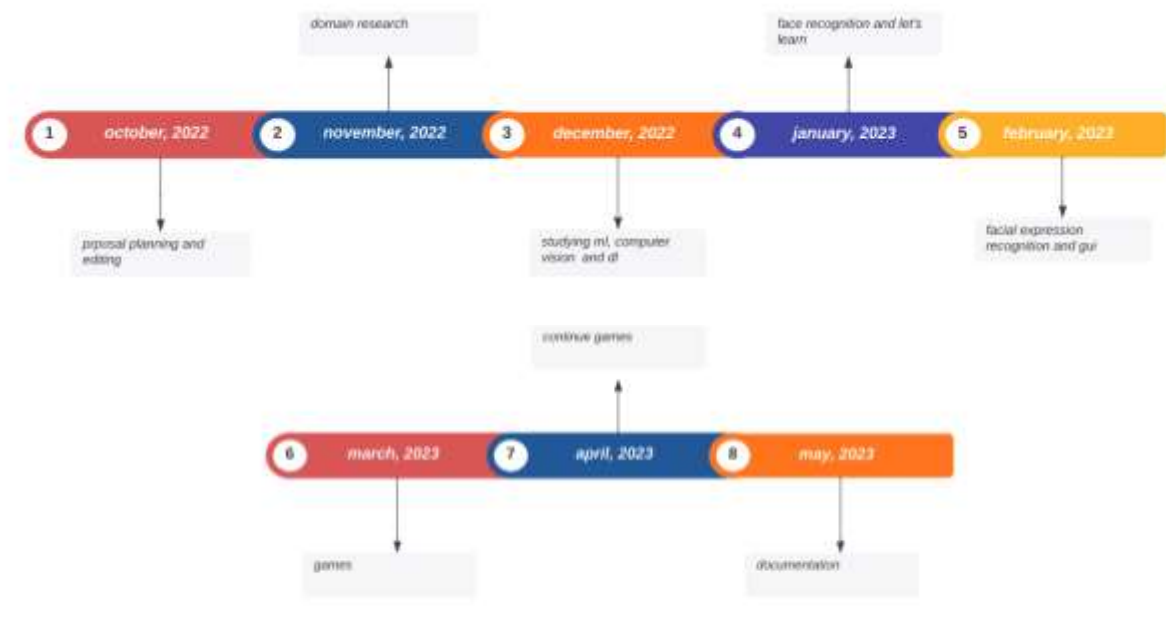


Figure 1.3: Time plan

1.6 Thesis outline

Chapter 2 includes a detailed description of the field of the project and will show you a survey of the work done in this field and a similar system.

Chapter 3 shows you a system architecture and analysis of the project and all the diagrams of our project like sequence diagram, use case and class diagram.

Chapter 4 will discuss the implementation of the project and all the algorithms and techniques used in order to accomplish our goal.

Chapter 5 shows you how to operate the project along with screen shots of the project representing all steps.

Chapter 6 it will be summary of the whole project along with the results obtained and What can be done in the future to improve the performance of the project and what additional functions could be added.

Chapter 2

Literature Review

Autism Spectrum Disorder (ASD), as defined by the Diagnostic and Statistical Manual Fifth Edition of the American Psychiatric Association (DSM 5) *, is a neurodevelopmental disorder associated with symptoms that include "persistent deficits in social communication and social interaction across multiple contexts" and "restricted, repetitive patterns of behavior, interests, or activities." The DSM 5 gives examples of these two broad categories:

Persistent deficits in social communication and social interaction across multiple contexts, as manifested by the following, currently or by history (examples are illustrative, not exhaustive):

- Deficits in social-emotional reciprocity, ranging, for example, from abnormal social approach and failure of normal back-and-forth conversation; to reduced sharing of interests, emotions, or affect; to failure to initiate or respond to social interactions.
- Deficits in nonverbal communicative behaviors used for social interaction, ranging, for example, from poorly integrated verbal and nonverbal communication; to abnormalities in eye contact and body language or deficits in understanding and use of gestures; to a total lack of facial expressions and nonverbal communication.
- Deficits in developing, maintaining, and understand relationships, ranging, for example, from difficulties adjusting behavior to suit various social contexts; to difficulties in sharing imaginative play or in making friends; to absence of interest in peers.

Restricted, repetitive patterns of behavior, interests, or activities, as manifested by at least two of the following, currently or by history (examples are illustrative, not exhaustive):

- Stereotyped or repetitive motor movements, use of objects, or speech (e.g., simple motor stereotypes, lining up toys or flipping objects, echolalia, idiosyncratic phrases).
- Insistence on sameness, inflexible adherence to routines, or ritualized patterns of verbal or nonverbal behavior (e.g., extreme distress at small changes, difficulties with transitions, rigid thinking patterns, greeting rituals, need to take same route or eat same food every day).
- Highly restricted, fixated interests that are abnormal in intensity or focus (e.g., strong attachment to or preoccupation with unusual objects, excessively circumscribed or perseverative interests).
- Hyper- or hypo reactivity to sensory input or unusual interest in sensory aspects of the environment (e.g. apparent indifference to pain/temperature, adverse response to specific sounds or textures, excessive smelling or touching of objects, visual fascination with lights or movement).

These symptoms result from underlying challenges in a child's ability to take in the world through their senses, and to use their body and thoughts to respond to it. When these challenges are significant, they interfere with a child's ability to grow and learn, and may lead to a diagnosis of autism.

Many parents are told autism is a behavioral disorder based on challenges in behavior. While children with autism do display behaviors that can be confusing, concerning, and even disruptive, the basis of these behaviors is a neurodevelopmental difference. Understanding autism based on behaviors is superficial at best. The behavioral perspective has dominated the "airwaves" for the past 15 years and Applied Behavioral Analysis (ABA) has become the most known intervention for autism as a result. However, clinical practice and research are creating a paradigm shift to more fully understanding autism from a neurodevelopmental perspective rather than simply behaviorally

The Market for Autism

When the challenges of autism are understood and appropriately addressed, and the autistic individual is accepted for who they are, their potential is no less than a neurotypical person. Too many professionals look at autism as something that needs to be controlled and contained. We look at autism as a neurodiversity that needs to be understood and the person needs to be supported in the right way.

Our system aims to do so as it's partially a neurological treatment where the system tries to mimic the role of a therapist to a certain extent. Through this system, steps are gradually taken to deal with this neurodevelopmental difference and decrease its percentage. Autism is not a psychological disorder. In fact, autism is not a mental illness and autistic persons do not choose to behave as they do. There is no known psychological factor shown to cause autism.

Researchers are exploring the idea that artificial intelligence (AI) could be used to diagnose autism and help people on the autism spectrum to improve social, communication, and emotional skills. Diagnosis of autism through the use of AI is now a reality and AI-based therapies in development show promise. Some AI-based apps are now downloadable for any smartphone user.

According to the publication Spectrum News, deep learning is sometimes better able than human beings to spot relevant patterns. These types of programs may be a good way to provide evaluators with confirmation of a diagnosis or suggest the need for further evaluation.

People with autism are often overwhelmed by the demands of human interaction. Social expectations, sensory challenges, difficulty with expressive and reception speech, and attentional issues can all interfere with optimal outcomes. To circumvent this problem, a number of innovative groups have started exploring ways to use AI to teach and engage people on the spectrum.

One of the most intriguing approaches to using AI in therapy involves creating and training robots to interact with autistic children. Their purpose is to give autistic children practice with identifying facial expressions, interacting socially, and responding appropriately to social cues.

Behavior Imaging^[14]

Behavior Imaging, a Boise, Idaho company, uses a system called the Naturalistic Observation Diagnostic Assessment. This tool is an app that allows parents to upload videos of their children for observation.

Clinicians watch the videos to make remote diagnoses. More recently, the company has started training AI-like algorithms to observe and categorize behaviors. The algorithms would not diagnose the children, but they might be used to point clinicians to specific behaviors that might otherwise have been missed.

Cognoa

Another use of AI-aided diagnosis is an autism screening tool created by Cognoa in Palo Alto California. This tool is a mobile app that parents can use without the involvement of a trained evaluator. It reviews answers to multiple-choice questions, as well as videos of the child.

SoftBank Robotics^[15]

SoftBank Robotics NAO humanoid robots are about two feet tall and look like science-fiction-style androids. They are capable of expressing emotions by changing the color of their eyes, moving their arms, and changing the tone of their voice.

Children with autism often respond more positively to NAO than to a human therapist and because NAO is a robot, it has unlimited patience and is able to repeat

the same cues in the same way over and over again without variation. Many children on the spectrum look forward to their time with and, in some cases, show NAO affection with hugs.

Massachusetts Institute of Technology^[16]

Researchers at MIT programmed a robot to integrate information about individual children using data from video, audio, and measurements of heart rate and skin sweat. Using this information, along with information about expected and appropriate behaviors, the robot can make sense of and respond to a child's behaviors.

Manatee^[18]

Manatee, a Denver startup specializing in AI apps for people with autism, is working with a company called Robauto to develop a robot called BiBli that can talk children through challenging interactions without judgment—at the child's own pace.

AI Apps for Autism^[17]

AI-based apps are less costly and easier to integrate into ordinary homes, schools, and therapists' offices than high-end robots. There are many autism apps on the market that support behavioral therapy and learning, but most are relatively simple logical tools for following a set of rules and earning points for doing so.

Assistive technology (AT) for autism includes a wide range of tools that can help someone learn, communicate, and carry out daily functions. These can range from simple picture boards and worry beads to sophisticated software, apps, and robots. AT tools can help people with many different areas of life including:

- Basic communication
- Reading, writing, and math

- Telling time and managing schedules
- Learning and using social skills
- Managing sensory challenges
- Staying safe
- Activities of daily living (managing household chores and self-care)

For some with autism, assistive technology can improve certain abilities. For others, it can enable them to do things they may not have been able to before.

How Deep Learning in CV is strengthening and improving AT^[4]

The task of automatically recognizing and locating objects is one of the primary tasks for humans in order to survive, work and communicate. As a consequence, the ability to automatically perform this task starting from images and videos is fundamental to build very powerful assistive devices able to understand and/or interact with their surroundings and, as a consequence, to help people with cognitive and/or physical limitations. On the other hand, deep architectures can learn more complex models than shallow ones, since they learn powerful representations of the objects without the need to perform hand design features.

The Deep Learning for Assistive Computer Vision learning frameworks for object recognition methods can mainly be categorized into two groups: one follows the traditional object detection pipeline, involving the generation of region proposals and the classification of each proposal into different object categories. The other regards object detection as a regression or classification problem, adopting a unified framework to achieve final results (categories and locations) directly.

A strictly related task is the one so called **scene understanding**, i.e. the ability not only to identify the targets (as object recognition does), but also to understand the other properties of the observed scene. In other words, this task entails recognizing the semantic constituents of a scene and the complex interactions that occur between them.



Figure 2.1: Scene understanding cycle

Convolutional Neural Networks (CNNs) can be really useful also to solve this task. A recent approach to address the aforementioned challenge consists in using the convolutional patch networks, which are CNNs trained to distinguish different image patches giving the possibility to perform pixel-wise labeling.

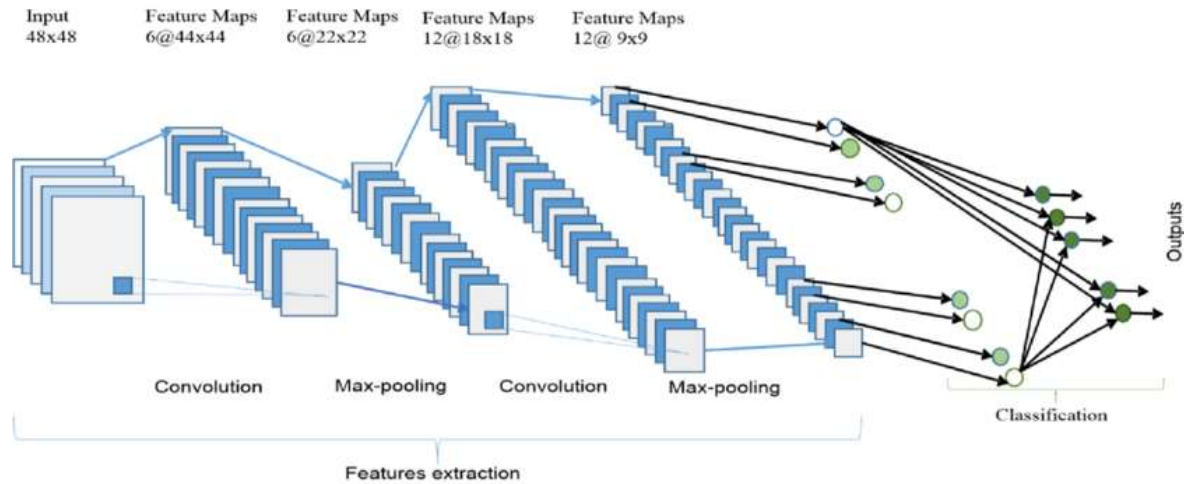


Figure 2.2: CNN Architecture

Different approaches using deep learning have been recently proposed and, among all, the one based on ensemble of models, each of which is optimized for a limited variety of poses, is capable of modeling a large variety of human body configurations.

Object localization and recognition^[6]:

Object Localization and Recognition is one of the areas of computer vision that is maturing very rapidly thanks to deep learning. Nowadays, there is a plethora of pre-trained deep learning models which can be used for this task, so it only takes a small amount of effort to build a system able to detect most of the objects in an image or video even in the presence of multiple overlapping objects and different backgrounds. In addition to detecting even multiple objects in a scene, recent deep learning based architectures are also able to precisely identify their boundaries and relations to one another.

This is achieved by deep structured learning which, for example, can learn relationship by using both feature, geometry, label and even physics and inferences about the abstract properties of the whole system. The recent advantages in object localization and recognition have been already employed in different applications. For instance, CNNs are effectively exploited to improve the performance in the autonomous navigation. In the YOLOv2 engine, which is one of the fastest strategies for object detection.

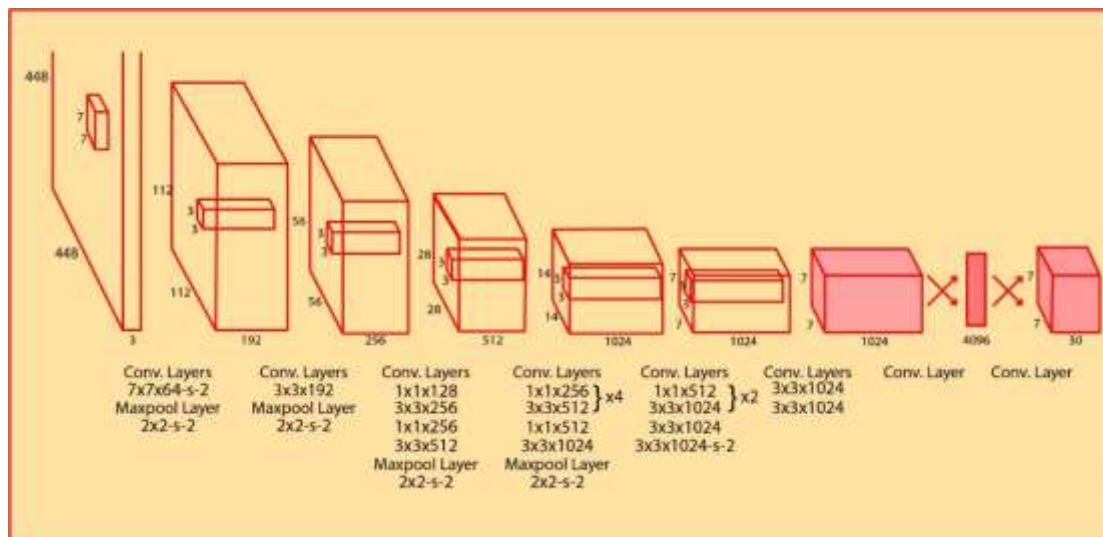


Figure 2.3: YOLO v2 architecture

The YOLOv2 engine was recently used also in to build a multimodal computer vision framework for human assistive robotics with the purpose of giving accessibility to persons with disabilities.

Human pose estimation and tracking^[19]

The estimation of the articulated motion of the human body is useful for a number of real world applications including medical rehabilitation, human-robot interaction and in general to create smart environments suitable to understand people behaviors.

Pose estimation is generally pursued by detecting and extracting the positions of the joints of the human body from different sources such as a single image, a sequence of images, and RGB-D data.

The main goal is to reconstruct the skeletal structures of the people in the scene and hence provide information about their body posture, the motion of the body, and human gestures.



Figure 2.4: Pose estimation and skeletal structure

In the context of assistive technologies, monitoring the pose of a child over time could reveal important information both during clinical trials or natural behaviors.

Survey on the Roles and Benefits of Social Robots in the Therapy of Children with Autism

Computers enable each and every child to actively participate in the learning process, because computer aided education provides individualized education. Computer Assisted Learning (CAL) has been used to support traditional academic learning (e.g. vocabulary, grammar, reading and mathematics); social skill development, life skill development and to reduce challenging behaviors. CAL provides a lot more than just ‘bells and whistles’ to the education and treatment of people with ASD. The approach seems to foster increased concentration and motivation, by appealing to the common preferences and skill sets of a majority of

people with ASD. This survey is aim at surveying some researched based Computer Aided tools that are developed for the therapy of children in the autism Spectrum.

Applications developed for helping the children with ASD^[1]

This section presents some of the applications that are developed for helping the children with autism disorder for treatment and educational purpose, the type of therapy they provide and the Technique used in developing the application. Some of them help in monitoring the progress of the treatment as well as some for monitoring their whereabouts.

S (NO)	Name of Application	Type of Therapy to Provide	Technique Used
1	Mocotons: Mobile Communications Tools for Children with Special Needs[6]	An assistive technology for children with special needs to help them communicate using cards that are loaded in the library. It is also flexible enough to handle multiple functions currently supported by different devices inside the classroom[6].	Picture Based, Touch Screen and Activity player.
2	Educational app for children with Autism Spectrum Disorders (ASDnet)[7]	This App is created to help children with ASDs to improve their communication skills with the people around them. By using the Application, children with ASDs can learn by listening to the audio sound after interacting with the picture of the object. The app also helps their parents and care givers to understand the needs of their children because it can avoid any unwanted miscommunication which will lead to tantrums and so on[7].	Picture based with sounds, Touch Screen and Activity Player.
3	Fill Me App: An Interactive Mobile Game Application for Children with Autism [8]	This is a Game application, which focused mainly on Science for identifying the human's body parts. Accumulating the best time for focus monitoring, eye-catching graphics, simple level of exercises, video tutorial and background music. The teacher or the parent can view or check the progress of the child's performance through the web application via internet [8].	A scoring system for focus monitoring, eye-catching graphics, simple level of exercises, video tutorial and background music[8].
4	An Interactive App for Learning Numeracy and Calculation for Children [9]	The application was built for Children with mild category of autism, and its specific task is to teach numeracy and necessary calculation like addition and subtraction. Real life data was collected[9].	Augmented Reality used, which makes the application elements, looks like a real world environment. Elements: The application is colorful and attractive so that it gets the attention of every autistic child. It is an interactive application, which makes it under the HCI (Human-computer interaction)[9].
5	AutismAid: A learning mobile application for autistic children [10]	An Application that will provide interoperability between an autistic person and caregivers to contribute on the treatment and monitoring in persons with ASD. The solution was implemented by creating a mobile application and tested with a population in an Autism center in the United Arab Emirates. The results show that the application made a positive treatment contribution to a person with ASD [10].	A system, which includes components, deployed in the mobile mode and cloud [10]. All components except for the data monitoring system components are deployed in the mobile mode. The mobile mode contains six components: User interface, account system, gaming system, Scheduling system, child-monitoring system and the database manager.
6	Autism Children's App using PECS [2]	An Android based mobile application (app) providing better learning environment with inclusion of graphical representation in a cost effective manner[2].	Basic demographic information needed creation of username etc. are required to be registered using the system. Activities involved Single Word learning, PECS's book, Differentiate, Question & Answer, Caregivers or Teachers can monitor progress [2].

Table 2.1: summary of all applications investigated

Existing similar systems:

QTRobot^[20]

QTRobot was developed by LuxAI a startup spun out of the University of Luxembourg, founded by Aida Nazarikhorrām and Pouyan Ziafati. The goal was designing an expressive social robot for use in therapy, education, and research. Applications included learning support of children with autism and experiments in human-robot interaction. LuxAI started building QTRobot in 2016, finished a final prototype in mid-2017, and the following year began trials at various centers in Luxembourg, France, Belgium, and Germany. LuxAI plans to conduct longer-term trials, studying the robot's impact on social competence, emotional well-being, and interaction with people.

It is an expressive little humanoid designed as a tool for therapists and educators. It uses facial expressions, gestures, and games to teach children with autism spectrum disorder about communication, emotions, and social skills.

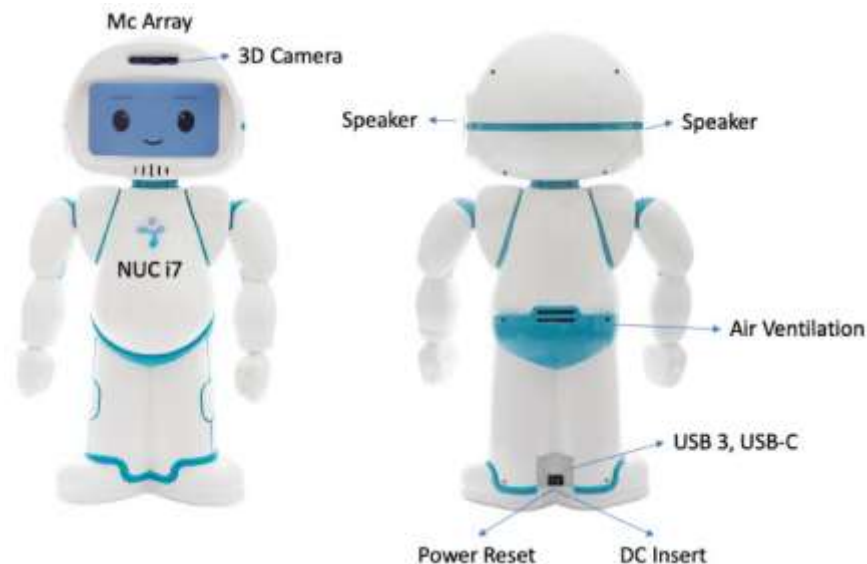


Figure 2.5: QTRobot

This is equipped with:

- (a) a face display that can show movies, thus emulating basic emotional expressions;
- (b) a 3D intel RealSense camera that enables vision and gesture recognition in space, as well as excellent resolution for facial recognition; and
- (c) microphones to recognize where the sound is coming from and speakers which allow the robot to produce verbal communication or play sounds.

QTrobot is a recently developed social robot. Until now, there are only two studies demonstrating its effectiveness as a mediator of behavioral interventions on children with ASD. Costa et al. have examined the use of QTrobot in long emotional-ability training for ASD, providing restricted evidence of the positive effects of the robot-mediated intervention. In another study, Costa et al. have evaluated the usefulness of QTrobot by assessing children's attention, imitation, and presence of repetitive and stereotyped behaviors. They obtained significant positive results in all considered parameters.

Advantages vs. Disadvantages

The most significant advantages of this device are the physical appearance QTrobot has more closely related human features, with different levels of motion which allow for an easier identification of social actions and expressions, facilitating the transfer of skills learned in the human–robot context to a human–human interaction. QTrobot is built precisely to a child's physical dimensions; it moves its arms with multiple DOF. Its display allows the presentation of animated faces and emotional facial expressions combined with arm movements and voice. Concerning technological features, the architecture of QTrobot is characterized by simple programming using Internal software, easy to customize with different behaviors (RealSense) useful for robot-assisted applications in the ASD domain. Furthermore, QTrobot has been developed to be employed in both homes and therapy settings.

The most significant disadvantages are that it has few sensory features and effective usage only with digital tablets. QTrobot is only equipped with RealSense which does not allow this kind of interactive spatial evaluation. Moreover, the child–robot interaction is mediated by the use of a digital tablet that could create an overstimulation for the child. Another pitfall is the lack of applications in clinical trials.

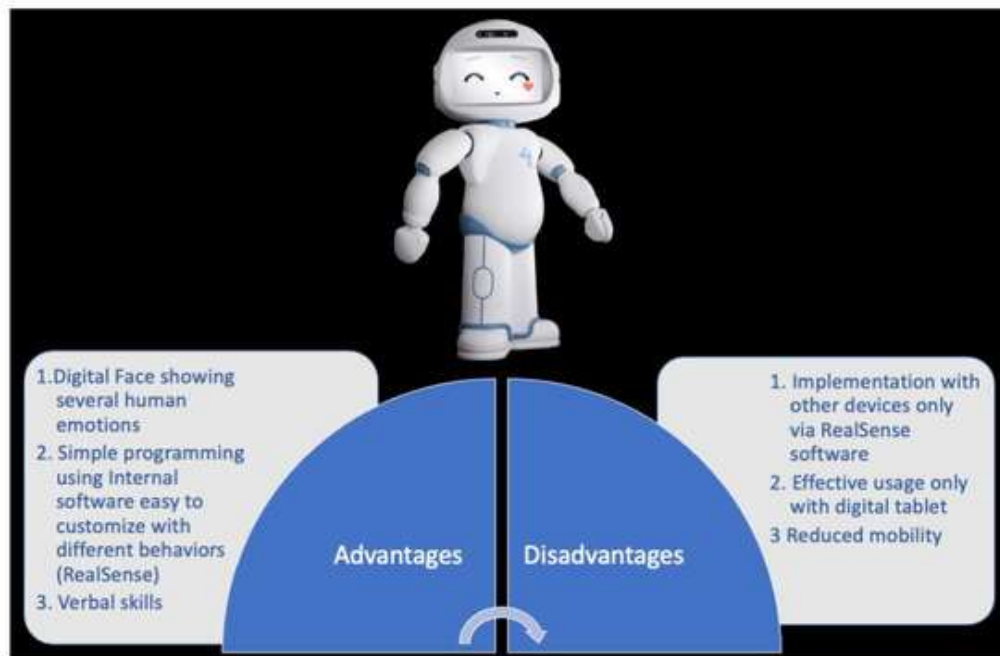


Figure 2.6: QTrobot Advantages and Disadvantage

Chapter 3

System Architecture and Methods

3.1 System architecture



Figure 3.1 System architecture



Figure 3.2 available videos and games

3.1.1 Face recognition

Identify the user and display a hello message using their name.

If user not found in the dataset, ask them to state their name in order to display the hello message

3.1.2 User chooses "let's go together" or "choose by yourself"

If user chooses "choose by yourself", two buttons appear allowing the user to either choose "let's learn" or "let's play" and depending on what they choose, a list of available educational videos/ games are displayed for them to choose from.

If user chooses "let's go together", the system moves on to the next step, which is facial expression recognition.

3.1.3 Let's learn

If the user chooses "choose by yourself" in the previous step:

A list of available educational videos, is displayed for the user to choose from.

The list includes the following topics: "Alphabets", "Numbers from 0 to 10", "Number from 10 to 100", "colors", "Emojis", "Abultion", "Praying", "Shapes" and "Fruit".

3.1.4 Let's play

If the user chooses "choose by yourself" in the previous step:

A list of available games, is displayed for the user to choose from. The list includes the following games: "Spin Ball", "Snake", "Car", "Card" and "Virtual Paint".

3.1.5 Facial expression recognition

If the user chooses "let's play" in the previous step:

In this phase, several images of the user are captured which are then used to determine whether they are "happy", "sad", "angry", "neutral", "surprised"

3.1.6 Pre-determined sequence

In this phase, the sequence of games/ videos played are determined based on the emotions of the user detected in the previous step.

If "**happy**", the sequence is Snake game, Spin ball game, play all educational videos, virtual paint and Car game.

If "**sad**", the sequence is Card game, Snake game, play all educational videos, Spin game, Car game and Virtual paint.

If "**angry**", the sequence is Car game, Virtual paint, play all educational videos, Snake game and Card game.

If "**neutral**", the sequence is Car game, Snake game, play all educational videos, Virtual paint and Card game.

If "**surprised**", the sequence is Virtual Paint, Card game, Car game, play all educational videos and Snake game.

3.2 Description of methods and procedures used

3.2.1 LBPH Recognizer^{[21][8]}

The Local Binary Pattern Histogram (LBPH) algorithm is a face recognition algorithm based on a local binary operator, designed to recognize both the side and front face of a human.

Training the algorithm:

To train the algorithm, we need to use a dataset with the facial images of the people we want to recognize. Each image has a unique ID and if there exists more than one image for the same user, all of them are given the same ID.

Applying the LBP operation

The first computational step of the LBPH is to create an intermediate image that describes the original image in a better way, by highlighting the facial characteristics using the concept of a sliding window, based on the parameters radius and neighbors.

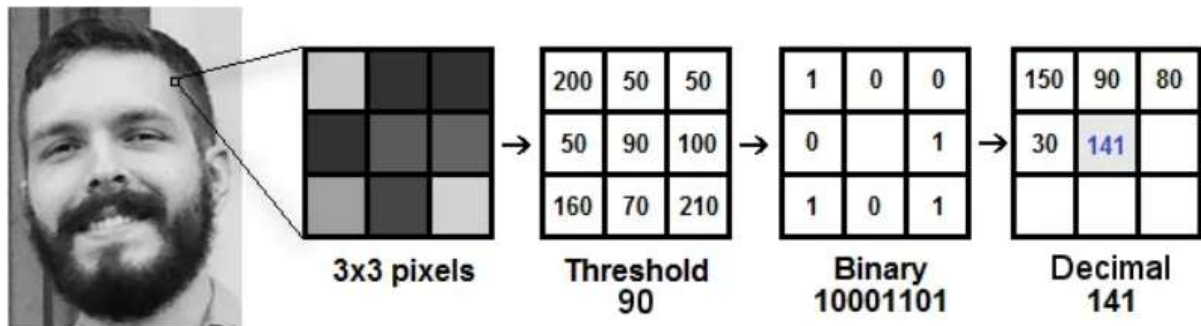


Figure 3.3: LBPH initial steps

Suppose we have a facial image in grayscale:

- We start by dividing the image into several windows of size 3x3 pixels where each 3x3 matrix contains the intensity of each pixel (0~255).
- Then, we need to take the central value of the matrix to be used as the threshold.
- If the value of neighbor is greater than or equal to the central value it is set as 1 otherwise it is set as 0.
- Thus, we obtain a total of 8 binary values from the 8 neighbors.
- After combining these values we get a 8 bit binary number which is translated to decimal number for our convenience.
- This decimal number is called the pixel LBP value and its range is 0-255.

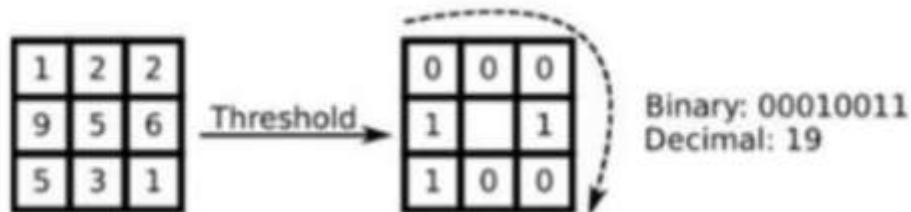


Figure 3.4: LBPH pixels calculations

- The decimal value is set to the central value of the matrix.
- At the end of this procedure (LBP procedure), we have a new image which represents better the characteristics of the original image.

Extracting the Histograms:

Using the image generated in the last step, we can use the Grid X and Grid Y parameters to divide the image into multiple grids.

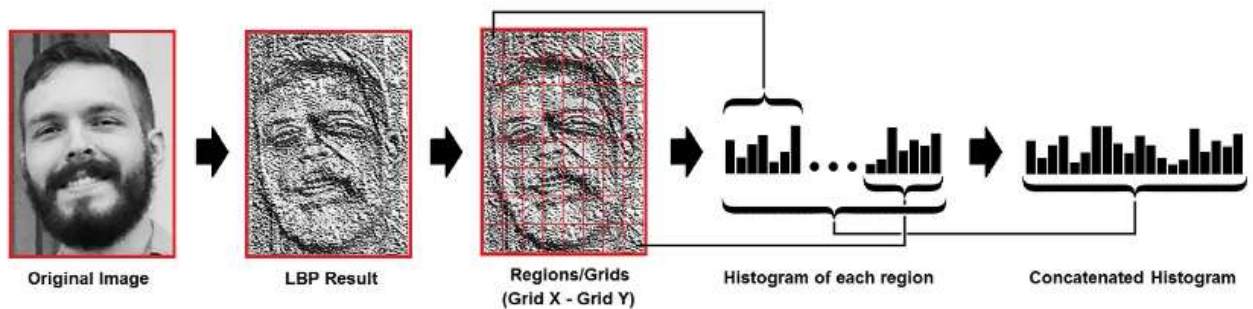


Figure 3.5: LBPH histogram extraction

As we have an image in grayscale, each histogram (from each grid) will contain only 256 positions (0~255) representing the occurrences of each pixel intensity.

Then, we need to concatenate each histogram to create a new and bigger histogram. Supposing we have 8x8 grids, we will have $8 \times 8 \times 256 = 16.384$ positions in the final histogram. The final histogram represents the characteristics of the image original image.

Performing the face recognition:

After the algorithm is trained, each histogram created is used to represent each image from the training dataset. Given an input image, we perform the steps again for this new image and creates a histogram which represents the image.

To find the image that matches the input image, compare two histograms and return the image with the closest histogram which can be done using various approaches, such as: euclidean distance, chi-square, absolute value, etc.

The output is the ID from the image with the closest histogram and the calculated distance, which is used as a confidence measurement. Where the lower confidences are better because it means the distance between the two histograms is closer.

3.2.2 Haar Cascade Classifie^[10]

Haar Cascade is an Object Detection Algorithm used to identify faces in an image or a real time video. The algorithm uses edge or line detection features. The algorithm is given a large number of positive images consisting of faces, and a large number of negative images not consisting of any face to train on them.

Haar-feature selection:

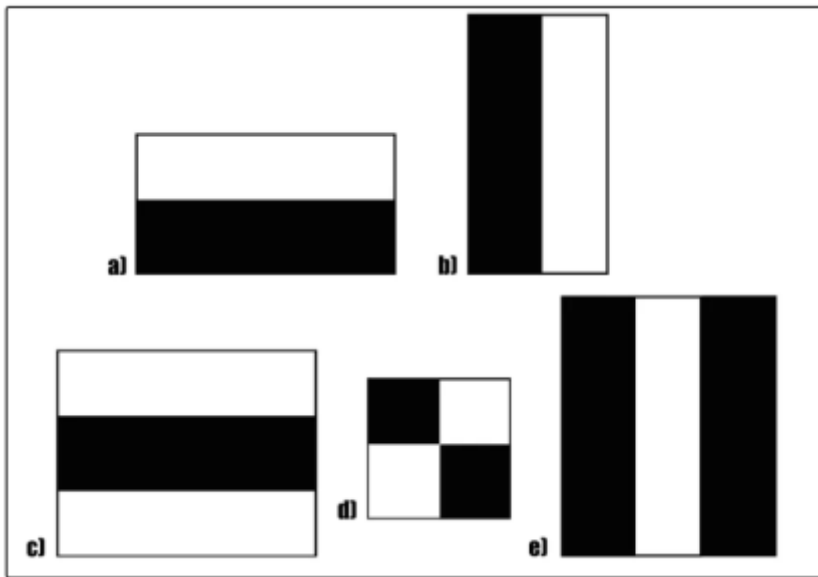


Figure 3.6: Haar features

Haar features consists of dark regions and light regions. It produces a single value by taking the difference of the sum of the intensities of the dark regions and the sum of the intensities of light regions. It is done to extract useful elements necessary for identifying an object and to find out the edges or the lines in the

image, or to pick areas where there is a sudden change in the intensities of the pixels.

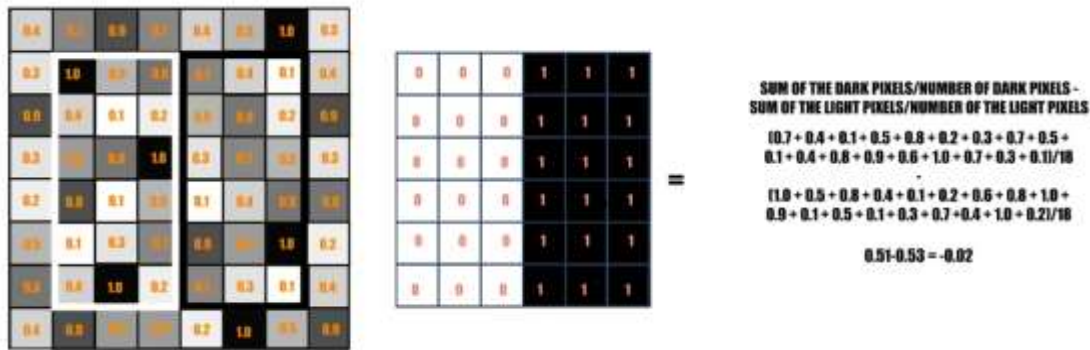


Figure 3.7: Haar calculations

A sample calculation of Haar value from a rectangular image section has been shown here. In the image above, the Haar feature can detect a vertical edge

To calculate the Haar value, the sum of the pixel intensities in the dark and light regions is computed, and their difference is taken. If there is an edge present in the image, the Haar value will be closer to 1.

Haar features traverse the entire image to detect edges and structures in different directions and depending on the feature each one is looking for, these are broadly classified into three categories. The first set of two rectangle features are responsible for finding out the edges in a horizontal or in a vertical direction (as shown above). The second set of three rectangle features are responsible for finding out if there is a lighter region surrounded by darker regions on either side or vice-versa. The third set of four rectangle features are responsible for finding out change of pixel intensities across diagonals.

However, calculating Haar values for the entire image involves a large number of mathematical calculations. To address this, the Integral Image concept is used, which involves calculating the sum of all pixels to the left and above a given pixel in the image.

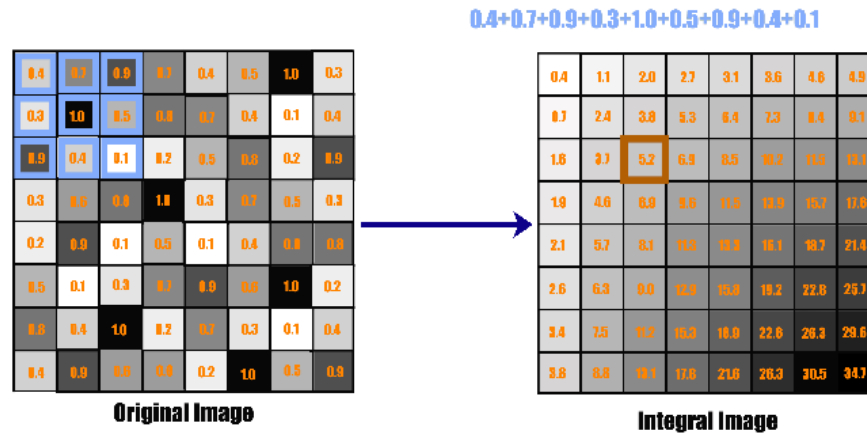


Figure 3.8: Haar integral image

The Integral Image approach reduces the time complexity of calculations and makes the process more efficient. For example, with the Integral Image, only four constant value additions are needed for any feature size, compared to 18 pixel value additions previously needed. To tackle this, they introduced another concept known as The Integral Image to perform the same operation.

An Integral Image is calculated from the Original Image in such a way that each pixel in this is the sum of all the pixels lying in its left and above in the Original Image. The calculation of a pixel in the Integral Image can be seen in the above GIF. The last pixel at the bottom right corner of the Integral Image will be the sum of all the pixels in the Original Image.

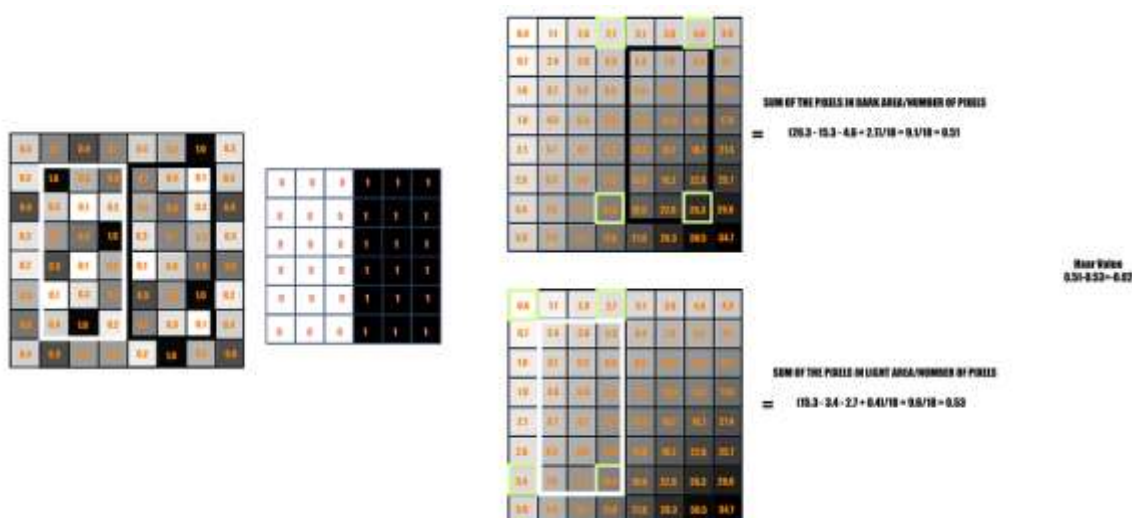


Figure 3.9: Haar integral calculations

In the above image, there is no edge in the vertical direction as the haar value is - 0.02, which is very far from 1.

AdaBoost

The Haar Cascade research involves a set of features that can capture facial structures such as eyebrows, eyes, and lips. Originally, there were approximately 180,000 features, which were reduced to 6,000 using a feature selection technique called AdaBoost.

A Boosting Technique called AdaBoost is used, in which each of these 180,000 features were applied to the images separately to create weak learners. Some weak learners produced lower error rates than others, and these were selected while irrelevant ones were eliminated. Weak learners are designed to misclassify only a minimum number of images and perform better than random guessing. The final set of 6,000 features is selected through this technique.

Attentional Cascade

After selecting the final set of 6,000 features, the Viola-Jones method uses a standard window size of 24x24 to detect facial features in training images. To simplify this process, they proposed a technique called The Attentional Cascade.

The idea behind this is, not all the features need to run on each and every window. If a feature fails on a particular window, then we can say that the facial features are not present there. Hence, we can move to the next windows where there can be facial features present.

Features are applied on the images in stages. The stages in the beginning contain simpler features, in comparison to the features in a later stage which are complex. If a window fails to detect facial features in an early stage, it is discarded, saving processing time in subsequent stages.

The second stage processing would start, only when the features in the first stage are detected in the image.

The first stage consists of two simpler features, and the second one consists of a single complex feature.

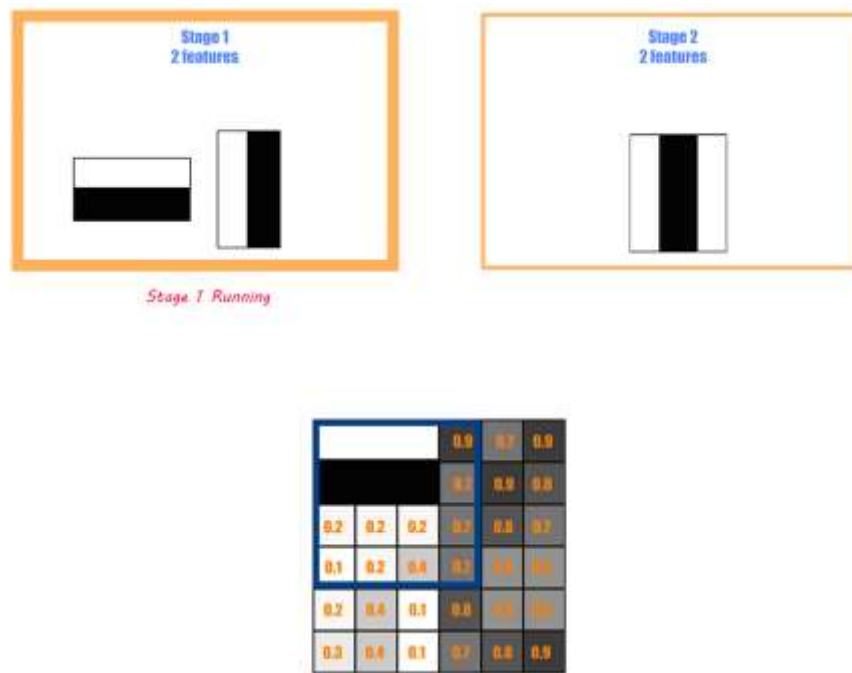


Figure 3.10: Attentional cascade

With a huge set of features in a number of stages, this technique would reduce the workload on the later stages, as most of the windows will get rejected in the initial stages only.

The Viola-Jones method uses 38 stages with varying numbers of features in each stage. The initial stages with simpler and fewer features remove windows without facial features, reducing the false negative ratio. Later stages with more complex features focus on reducing the error detection rate, achieving a low false positive ratio.

3.2.3 CNN^[11]

A Convolutional Neural Network, also known as CNN or ConvNet, is a class of neural networks that specializes in processing data that has a grid-like topology, such as an image. A digital image contains a series of pixels arranged in a grid-like that contains pixel values to denote how bright and what color each pixel are.

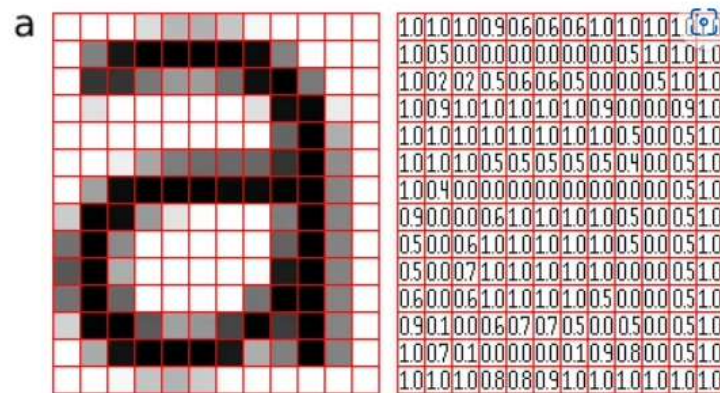


Figure 3.11: digital image

Convolutional Neural Networks (CNNs) process image data by breaking it down into smaller regions, where each neuron in the network is responsible for processing data within its own receptive field. This mimics the design of the human vision system, which also uses receptive fields to process visual information.

CNNs arrange layers hierarchically, with simpler patterns detected in the earlier layers and more complex patterns detected in later layers. This hierarchical arrangement allows computers to "see" and process visual information in a way that is similar to humans.

Architecture

A CNN typically has three layers: a convolutional layer, a pooling layer, and a fully connected layer.

Convolution Layer:

- The convolution layer is the primary building block of a CNN, responsible for the majority of the network's computation. It performs a dot product between two matrices: a set of learnable parameters (kernel) and the receptive field of the image.

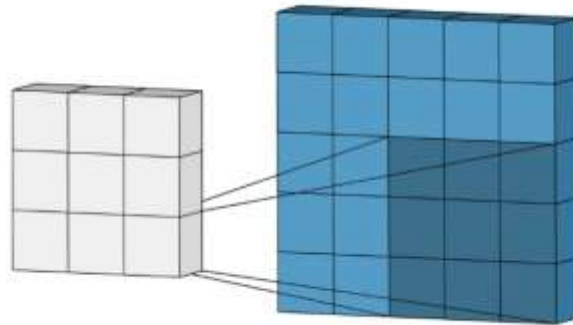


Figure 3.12: convolution

- During the forward pass, the kernel slides across the height and width of the image, producing an activation map that displays the response of the kernel at each spatial position. The sliding size of the kernel is called a stride, resulting in a two-dimensional representation of the image.

Pooling Layer:

- The pooling layer in a CNN replaces the output of the network at certain locations by summarizing nearby outputs. This reduces the spatial size of the representation, decreasing the computation and weights required. The operation is processed on each slice individually.
- Several pooling functions exist, including averaging over a rectangular neighborhood, L2 norm, and weighted averages based on distance. However, the most popular is max pooling, which reports the maximum output from the neighborhood.

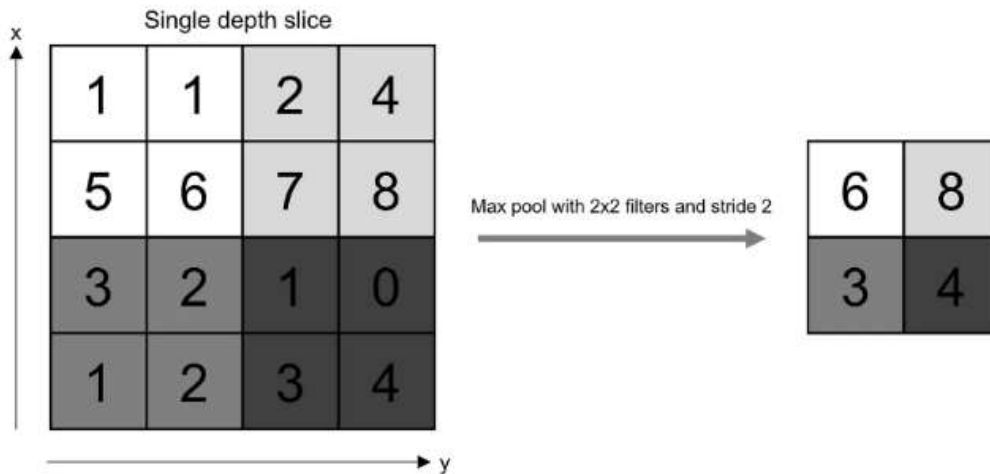


Figure 3.13: max pooling

Fully Connected Layer:

- Neurons in this layer have full connectivity with all neurons in the preceding and succeeding layer. This is why it can be computed as usual by a matrix multiplication followed by a bias effect.
- The FC layer helps to map the representation between the input and the output.

Designing the network:

We designed a CNN for facial expression recognition using the Keras deep learning library and the ImageDataGenerator class to preprocess and augment the image data.

The architecture consists of several convolutional layers, each followed by an activation function, batch normalization, and max pooling. The final layers are fully connected layers that perform classification. The model is trained using the categorical cross-entropy loss function and the Adam optimizer.

To prevent overfitting, the code uses early stopping, model checkpointing, and learning rate reduction techniques. The model is trained for 25 epochs, and the performance is evaluated using accuracy as the metric.

3.2.4 YOLO v3^[9]

YOLOv3 (You Only Look Once, Version 3) is a real-time object detection algorithm that identifies specific objects in videos, live feeds, or images. The YOLO machine learning algorithm uses features learned by a deep convolutional neural network to detect an object.

YOLOv3 is an improved version of YOLO and YOLOv2. YOLO is implemented using the Keras or OpenCV deep learning libraries.

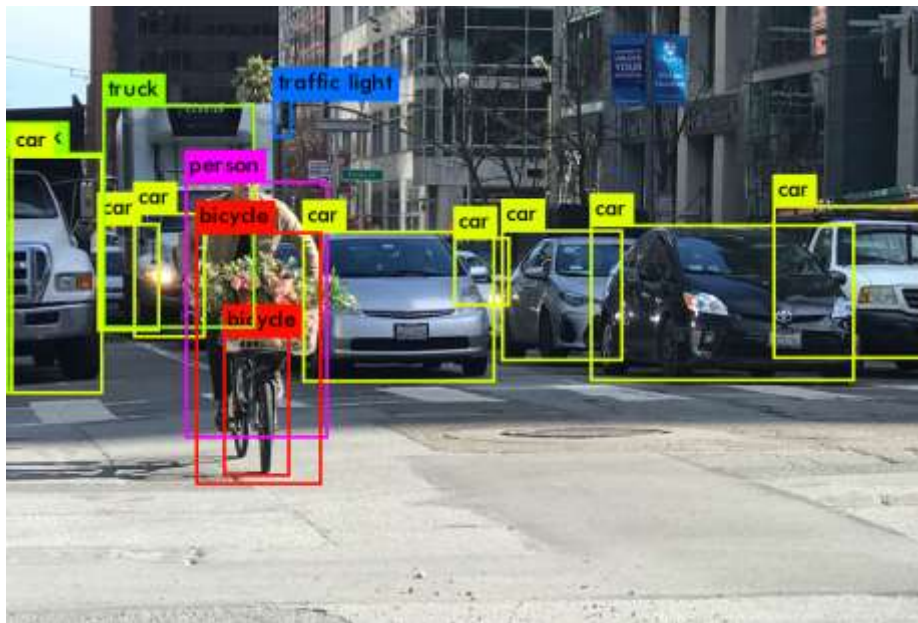


Figure 3.14: YOLO object detection

As typical for object detectors, the features learned by the convolutional layers are passed onto a classifier which makes the detection prediction. In YOLO, the prediction is based on a convolutional layer that uses 1×1 convolutions. The YOLO algorithm is named “you only look once” because its prediction uses 1×1 convolutions; this means that the size of the prediction map is exactly the size of the feature map before it.

How does YOLOv3 work?

YOLO is a Convolutional Neural Network (CNN) for performing object detection in real-time. CNNs are classifier-based systems that can process input images as structured arrays of data and recognize patterns between them (view image below). YOLO has the advantage of being much faster than other networks and still maintains accuracy. It allows the model to look at the whole image at test time, so its predictions are informed by the global context in the image. YOLO and other convolutional neural network algorithms “score” regions based on their similarities to predefined classes. High-scoring regions are noted as positive detections of whatever class they most closely identify with. For example, in a live feed of traffic, YOLO can be used to detect different kinds of vehicles depending on which regions of the video score highly in comparison to predefined classes of vehicles.

YOLOv3 Algorithm Steps:

1. The YOLOv3 architecture is based on the architecture of feature extraction model, Darknet-53. For deducing the detection of objects, 53 layers are stacked together, therefore, resulting in a fully convolutional architecture of 106 layers.

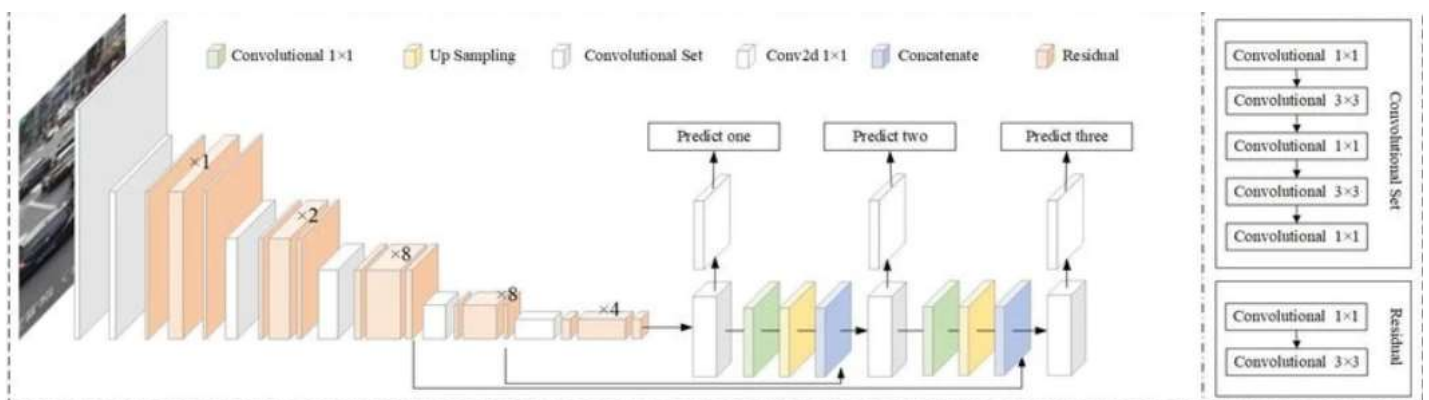


Figure 3.15: YOLOv3 architecture

2. The objects are identified at three layers of the architecture- 82nd, 94th, and 106th.

3. Each layer of the 53-layered architecture is followed by batch normalization layer and the implementation of Leaky ReLU activation function.
4. The basic idea of the YOLOv3 architecture is to divide the image into cells of size $S \times S$. One grid cell per object is responsible for the object's prediction.
5. Each grid cell has five parameters that come in handy to specify the location of bounding box. These are $(x, y, w, h, \text{confidence})$ where x and y represent the coordinates of the box's center, w and h are used to specify the width and height of the box, confidence denotes the absence/presence of any object. Besides that, the image also has class probabilities corresponding to each grid cell.
6. Since the probabilities are evaluated for each grid cell, there are chances that the algorithm can predict multiple bounding boxes for same objects. To avoid this situation, the Non-Max Suppression method is used.

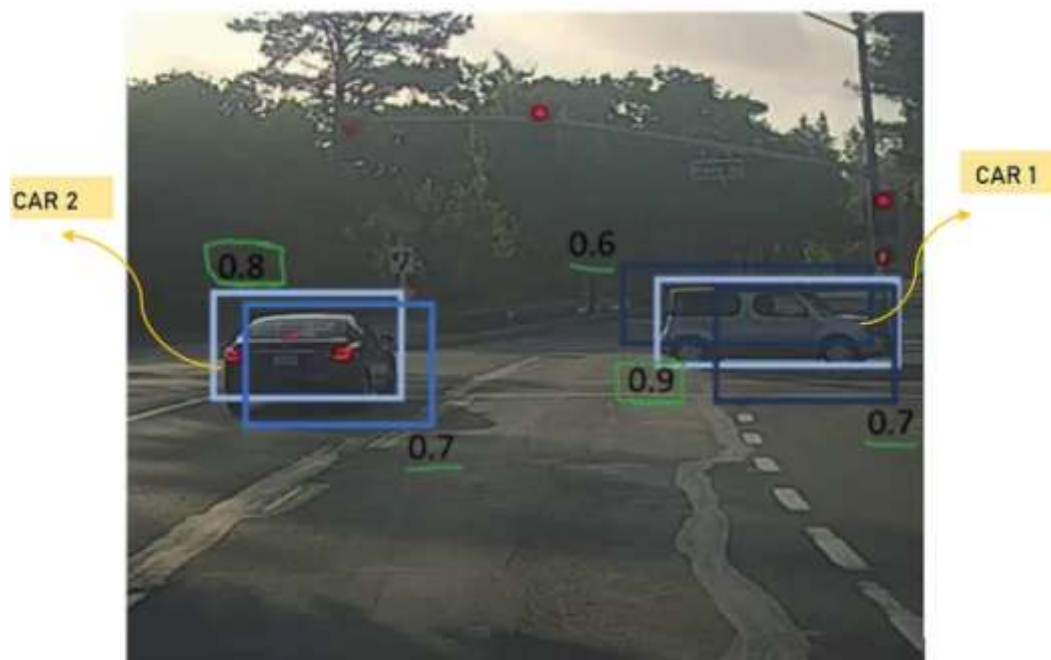


Figure 3.16: Non-Max Suppression

7. The input to the first layer of the model is a batch of n images of the shape $(n, 416, 416, 3)$ where $(416, 416)$ represents the width and height of the image and 3 is for the number of channels: Red, Green, Blue.
8. The architecture of YOLOv3 model is free from any type of pooling layers and for the convolutional layers, the stride is 2 for down sampling the feature maps.

9. Multiple filters are convolved in a convolutional layer for generating multiple feature maps.
10. The YOLOv3 model makes predictions at three different positions in the network. The three different positions are at 82nd, 94th, and 106th layers. The stride for these three layers is 32, 16, and 8 respectively.
11. Thus, the output of the 82nd layer is 13×13 ($416/32 = 13$) and it is responsible for detecting large objects as the stride is 32. Similarly, for the 94th layer, the output is of the size 26×26 ($416/16 = 26$) as the stride is 16 and it is responsible for detecting medium-sized objects. For the last layer (106th), the stride is 8 and it detects large objects. So the output is of the size 52×52 ($416/8 = 52$).

3.2.5 Blaze pose^[13]

BlazePose is a human pose estimation model developed by Google that uses machine learning to estimate the 2D and 3D poses of a person from a single RGB image or video frame. It is part of the MediaPipe framework provides a pre-trained, open BlazePose model that can compute (x,y,z) coordinates of 33 skeleton keypoints.



Figure 3.17: Pose detection

Input and output:

BlazePose consists of two machine learning models: a Detector and an Estimator.

- The Detector cuts out the human region from the input image.
- The Estimator takes a 256x256 resolution image of the detected person as input and outputs the 33 keypoints according the following ordering convention

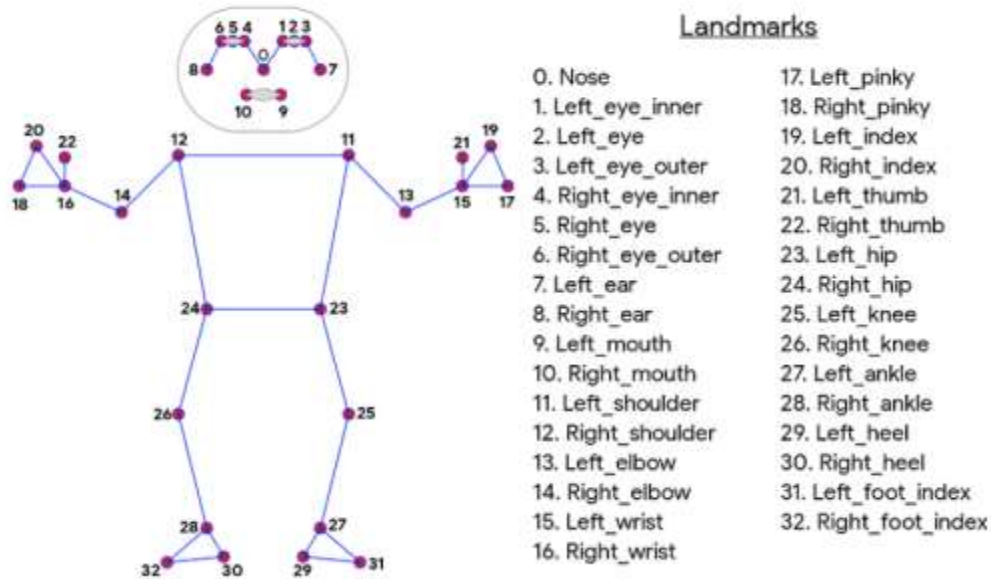


Figure 3.18: Pose detection landmarks

Architecture:

The Detector is an Single-Shot Detector(SSD) based architecture. Given an input image (1,224,224,3), it outputs a bounding box (1,2254,12) and a confidence score (1,2254,1). The 12 elements of the bounding box are of the form (x,y,w,h,kp1x,kp1y,...,kp4x,kp4y), where kp1x to kp4y are additional keypoints. Each one of the 2254 elements has its own anchor, anchor scale and offset need to be applied.

the bounding box is determined from its position (x,y) and size (w,h) . In alignment mode, the scale and angle are determined from $(kp1x,kp1y)$ and $(kp2x,kp2y)$, and bounding box including rotation can be predicted.

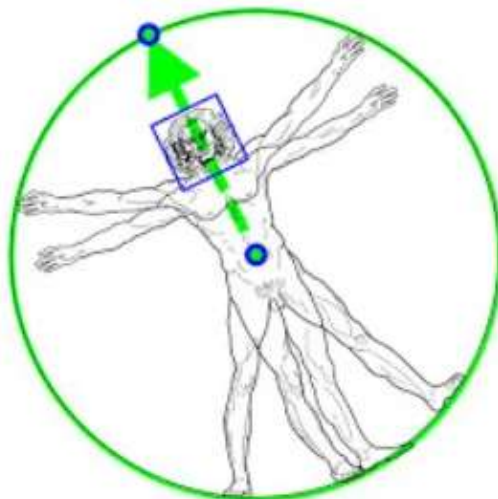


Figure 3.19: SSD bounding box

The *Estimator* uses heatmap for training, but computes keypoints directly without using heatmap for faster inference.

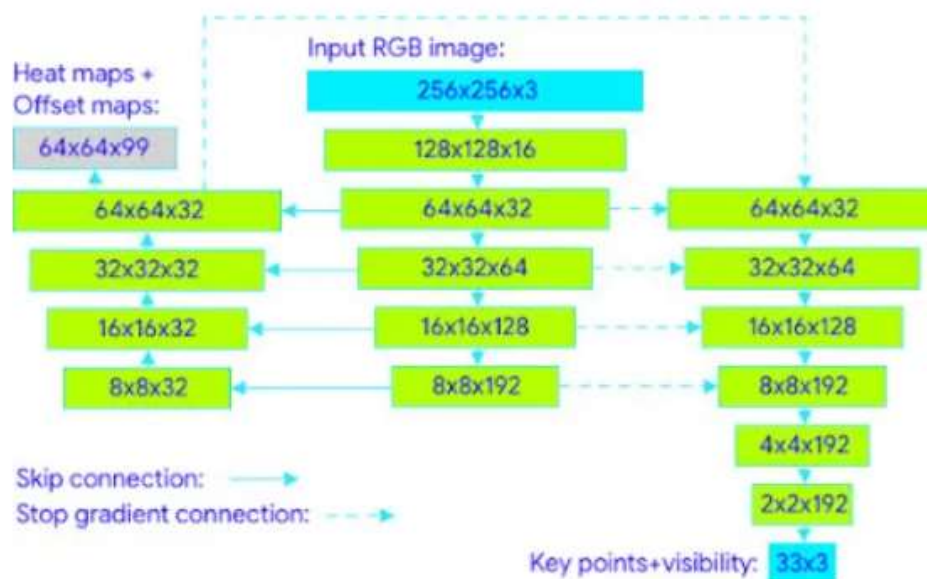


Figure 3.20: Estimator architecture

The first output of the estimator is a set of 195 landmarks with dimensions of $(1,195)$, where each landmark consists of 5 values $(x, y, z, \text{visibility}, \text{presence})$. The

landmarks represent 33 key body parts, and for each body part, there are 5 landmarks representing different aspects of the body part's position and appearance.

The z-values of the landmarks are calculated based on the position of the person's hips. If the value is negative, the corresponding keypoint is between the hips and the camera, while if the value is positive, the keypoint is behind the hips.

The visibility and presence values are stored in a range between a minimum and maximum floating-point value, which are then converted to probabilities using a sigmoid function. The visibility value represents the probability that a keypoint exists in the frame and is not occluded by other objects, while the presence value represents the probability that a keypoint exists in the frame. This information can be used to filter out unreliable landmarks and improve the accuracy of the pose estimation.

3.2.6 Face landmark detection models^[22]

MediaPipe Face Landmark Model is an openCV machine learning model developed by Google that can detect and track multiple facial landmarks in real-time video streams or static images.

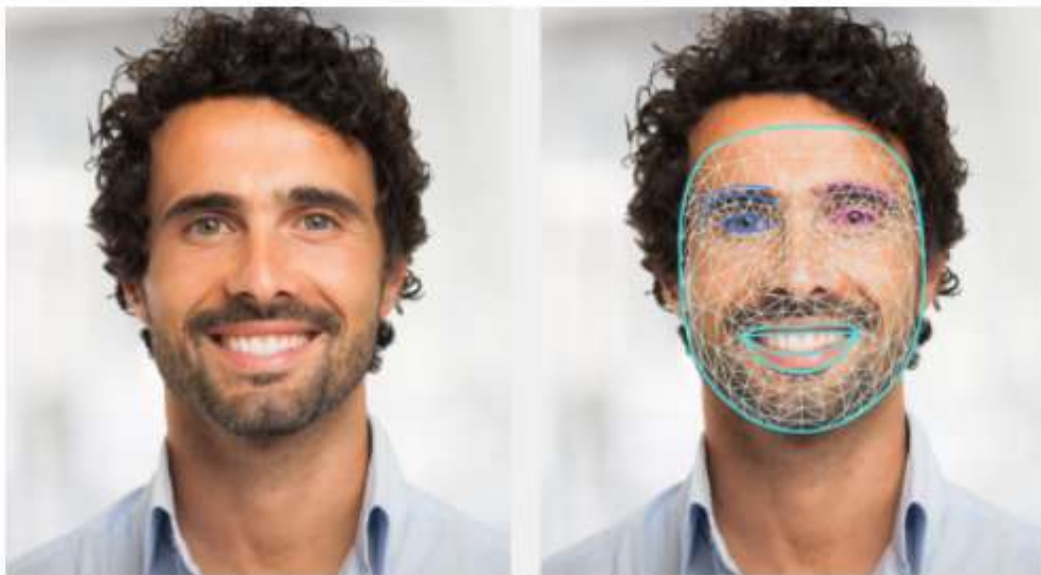


Figure 3.21: face landmark detection

The Face Landmarker uses a series of models to predict face landmarks. The first model detects faces, a second model locates landmarks on the detected faces, and a third model uses those landmarks to identify facial features and expressions.

- **Face detection model:** detects the presence of faces with a few key facial landmarks. The face detection model is the BlazeFace short-range model, a lightweight and accurate face detector.
- **Face mesh model:** adds a complete mapping of the face. The model outputs an estimate of 478 3-dimensional face landmarks.
- **Blendshape prediction model:** receives output from the face mesh model predicts 52 blendshape scores, which are coefficients representing facial different expressions.

BlazeFace:

BlazeFace is a machine learning model developed by Google to rapidly detect the location and keypoints of faces.

The position of the face and the keypoints of the face can be obtained simultaneously. There are six key points: eyes, nose, ears, and mouth. It is also possible to detect multiple people at the same time.



Figure 3.22: BlazeFace inference result

BlazeFace is designed to perform very fast inference on mobile GPUs. Specifically, it runs nearly 2.3 times faster than MobileNetV2-SSD.

Device	MobileNetV2-SSD, ms	Ours, ms
Apple iPhone 7	4.2	1.8
Apple iPhone XS	2.1	0.6
Google Pixel 3	7.2	3.4
Huawei P20	21.3	5.8
Samsung Galaxy S9+ (SM-G965U1)	7.2	3.7

Table 3.1: Inference speed across several mobile devices

BlazeFace uses an improved network based on MobileNet.

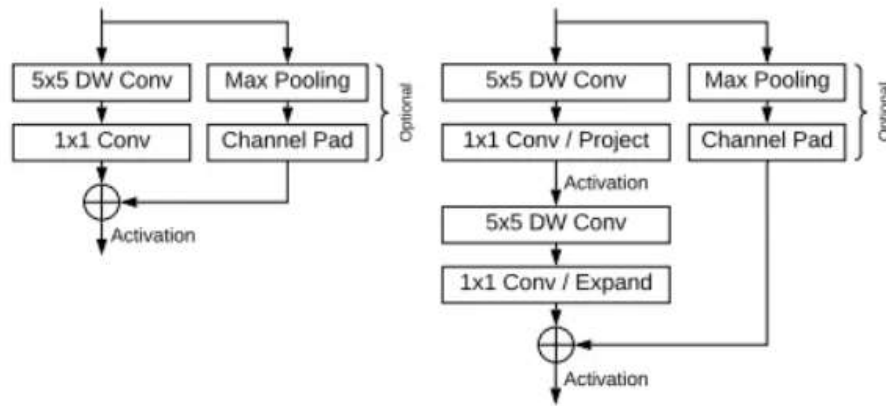


Figure 3.23: BlazeBlock and double BlazeBlock

3.2.7 Hand landmark model^[12]

HandLandmarker is a computer vision model that is designed to detect and localize key points on a human hand in real-time. The model is based on deep learning and uses a convolutional neural network (CNN) architecture to analyze images and identify the location of various landmarks on the hand.

The HandLandmarker model has been trained on a large dataset of annotated hand images, which allows it to accurately detect and locate key points on the hand, such

as the fingertips, knuckles, and wrist. The model is designed to work in real-time, which makes it suitable for applications such as gesture recognition, virtual and augmented reality, and robotics.

The quality of the annotations is critical for the performance of the computer vision model. Poorly annotated images can lead to inaccurate detection and localization of the hand landmarks, which can negatively impact the performance of the model. Therefore, it is important to ensure that the annotations are accurate and consistent across the dataset.

Annotated hand image datasets are typically created by collecting images of hands under various conditions, such as different lighting conditions, hand poses, and backgrounds. The dataset is then split into training, validation, and testing sets, which are used to train, tune, and evaluate the performance of the computer vision model.

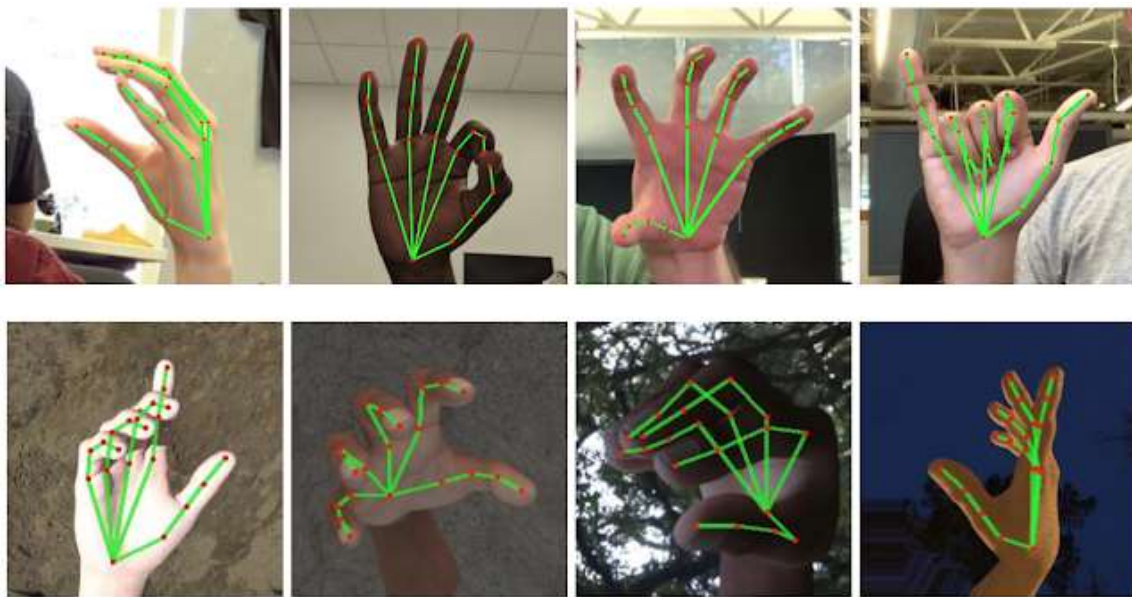


Figure 3.24: hand landmark model results

The HandLandmarker model uses a single-shot detector (SSD) approach, which means that it can detect multiple keypoints in a single pass through the image. This approach makes the model efficient and suitable for real-time applications.

Features

- Input image processing - Processing includes image rotation, resizing, normalization, and color space conversion.
- Score threshold - Filter results based on prediction scores.

The Hand Landmarker accepts an input of one of the following data types:

- Still images
- Decoded video frames
- Live video feed

The Hand Landmarker outputs the following results:

- Handedness of detected hands
- Landmarks of detected hands in image coordinates
- Landmarks of detected hands in world coordinates

The hand landmark model bundle detects the keypoint localization of 21 hand-knuckle coordinates within the detected hand regions. The model was trained on approximately 30K real-world images, as well as several rendered synthetic hand models imposed over various backgrounds.



Figure 3.25: hand landmarks

Chapter 4

System Implementation and Results

3.1 Dataset

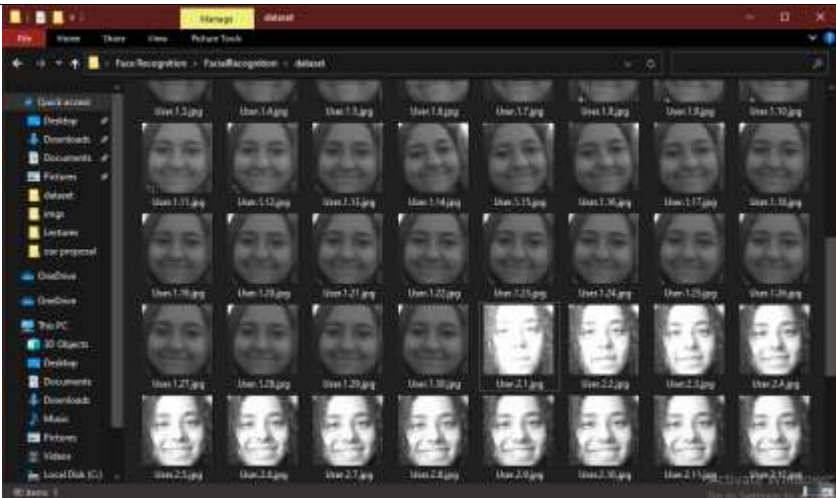
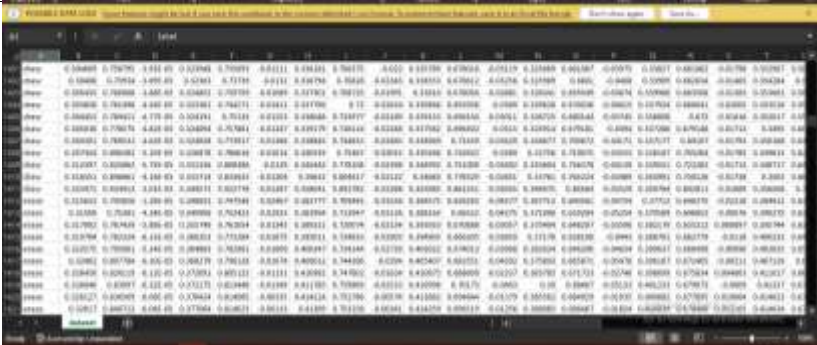
Data Set	Model used in	Description	Simulation
Faces of software users (we create it ourselves)	Face recognition	We run the face recognition code to input the new user ID and the camera captures 30 images for the person's face to add them to the dataset folder and pass them to the face training code	
Hand landmark positions for drawing and erasing (csv)	Virtual paint game	The train file which includes NNS takes the file to learn hand positions in which to draw or erase	

Figure 4.1: dataset one

Figure 4.2: dataset two

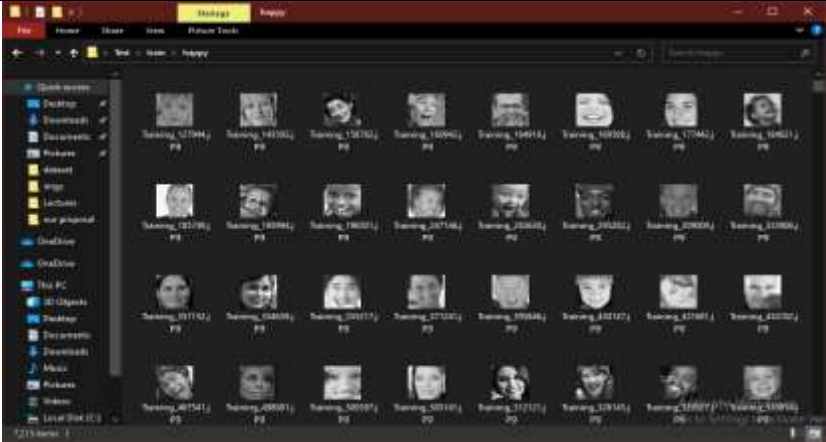
Random Images for different expressions	Facial expression	The train file takes these images to output each label reaction to later on recognize what's the child's facial expression	
---	-------------------	--	--

Figure 4.3: dataset three

Table 4.1: used datasets

The rest of the models are pre-trained using media pipe and cv2

3.2 Description of programs used:

Pycharm community edition:

- can be used offline
- offers a range of powerful features and tools which help developer productivity and efficiency
- offers intelligent code completion and error highlighting to help you write code faster and with fewer mistakes.

Google colab:

- Free access to computing resources such as CPU, GPU
- Easy access to pre-installed libraries of commonly used Python libraries

3.3 Stepup Configuration (hardware):

The hardware components used in the project are 2 Intel Xeon virtual CPUs, 13 GB of RAM, Nvidia T4 GPU with 12 GB of VRAM, all provided by Google

Colab, in addition to local hardware components which are: Intel core I7-1065G7 and 12 GB of RAM.

3.4 Experimental and Results

Face recognition:

Face recognition is trained using a dataset consisting of 30 pictures for each current system user and as an accuracy metric we used a confidence score.

The confidence score is the difference between the image of the user in the dataset and a frame from the video.

- It's calculated using distance between histograms
- The lower the number, the better because it means the distance between the two histograms is closer.
- We used a threshold of 100 where if the number is higher the user is considered unknown.
- We then converted the confidence into accuracy to represent how similar the person in the front of the camera is to the identified user from the dataset.

The system overall works well and identifies the user existing in the data set however the accuracy is sometimes lower than 50. We have also found that the model works a lot better when the user's pictures in the dataset includes a variety of poses and movement.

A paper using a similar algorithm was published by a student in IJRTE on march 2020, they achieved an average accuracy of 77%

Facial expression recognition:

Facial expression recognition is trained using a dataset obtained from a kaggle competition. We chose to split the dataset into training and validation to obtain better overall accuracy.

The validation accuracy: 68.19 %

Overall system performance:

The system runs smoothly and all features are implemented and work well.

The system is targeted at autistic children as an end user so, we tried the system on a 5 years old autistic child and was also diagnosed with ADHD

- He enjoyed the colorful gui and was invested the entire time while using the system
- He was focused during watching the educational videos and interacted with the questions at the end of the videos.
- He found it hard at first to play the spinball game and the car game, but after a while of practicing, he learned how to play.
- He really enjoyed the snake game and paint game especially the fact that he can change the drawing color while drawing.

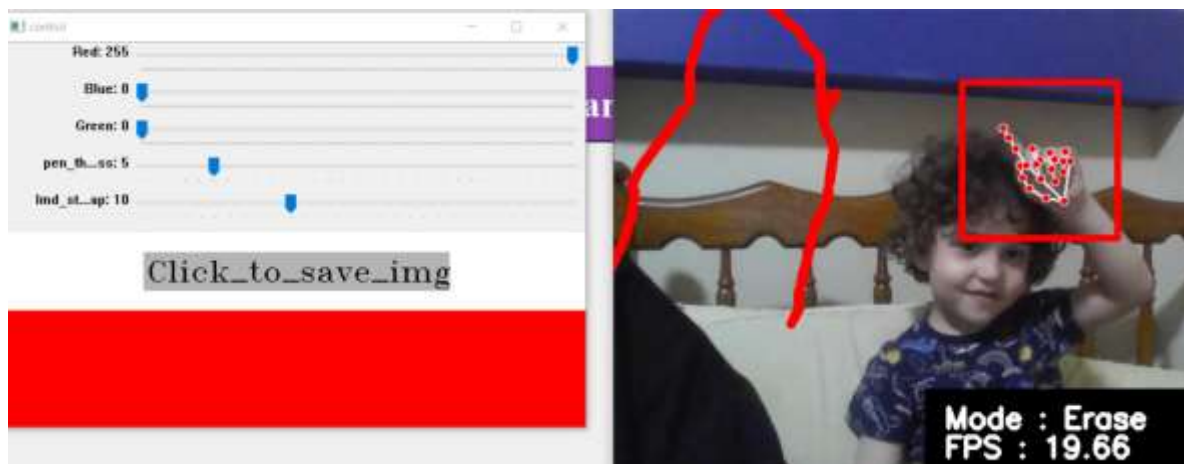


Figure 4.1: child trying virtual paint

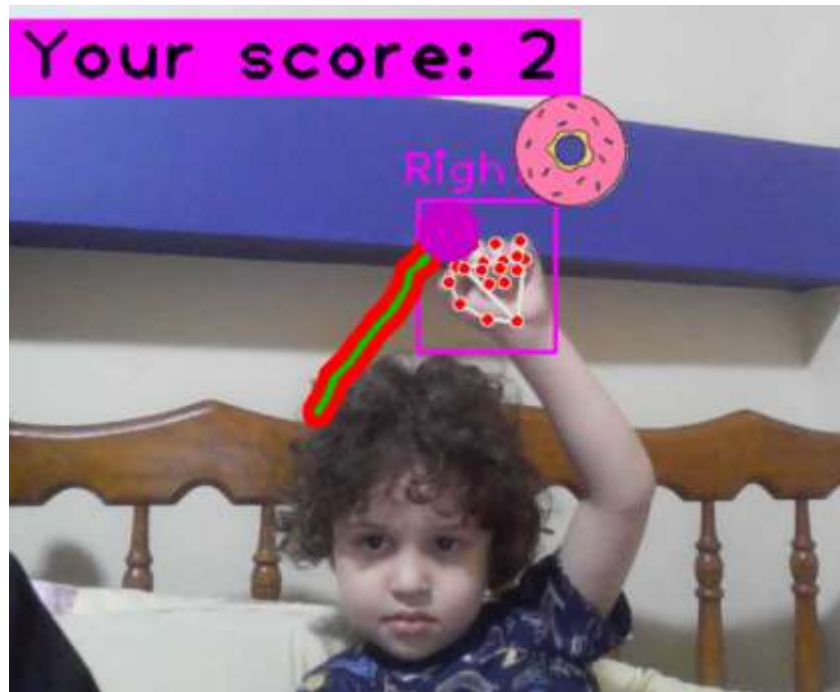


Figure 4.2: child trying snake game

- We also tried the happy and neutral sequences:
 - If "**happy**", the sequence is Snake game, Spin ball game, play all educational videos, virtual paint and Car game.
 - If "**neutral**", the sequence is Car game, Snake game, play all educational videos, Virtual paint and Card game.

He was responsive to both and the system was overall successful in educating and entertaining him.

Chapter 5

Run

When running the system, The user is met with the following greeting screens:

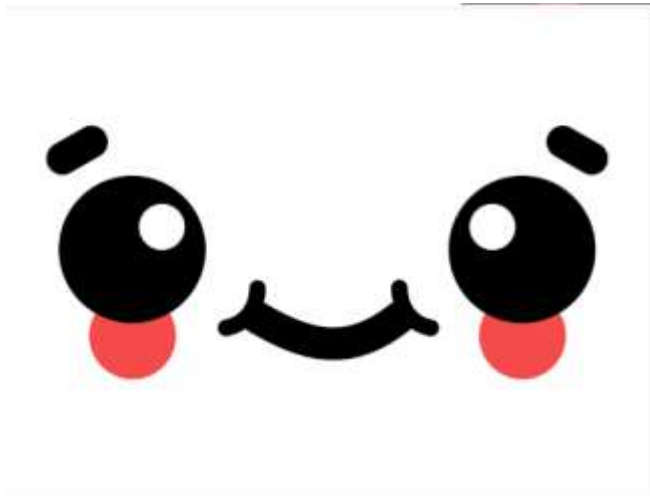


Figure 5.1: greeting screen

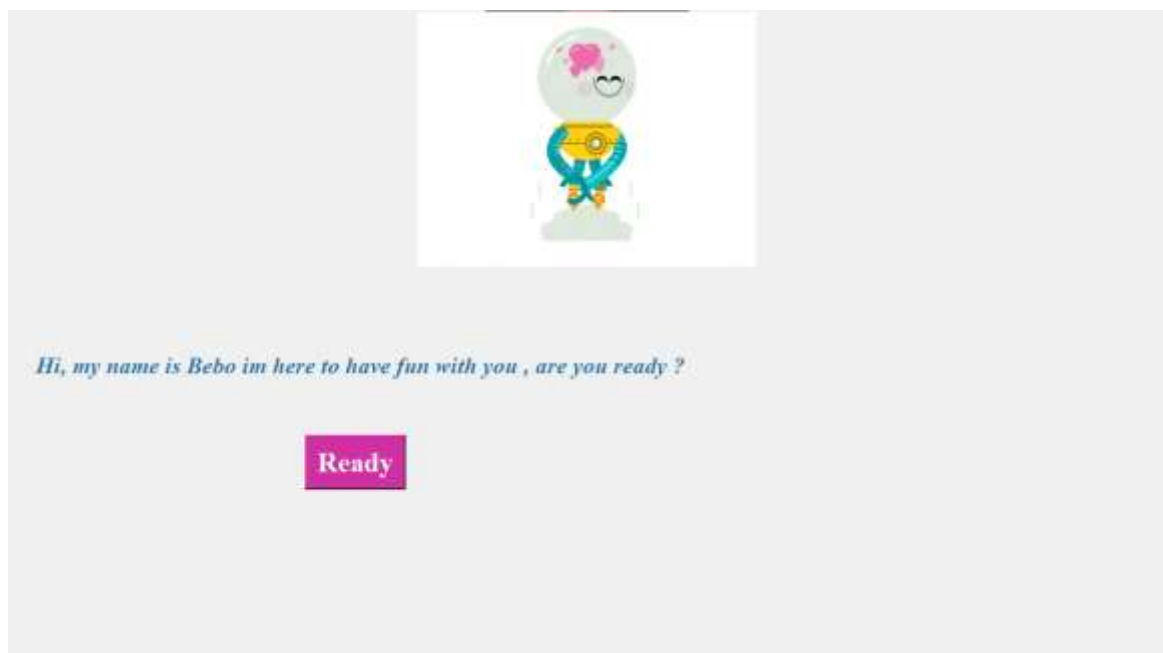


Figure 5.2: ready screen

The system then tries to id the user using face recognition



Figure 5.3: face recognition

After clicking ready, the user is provided two options "let's go together" and "choose by yourself"

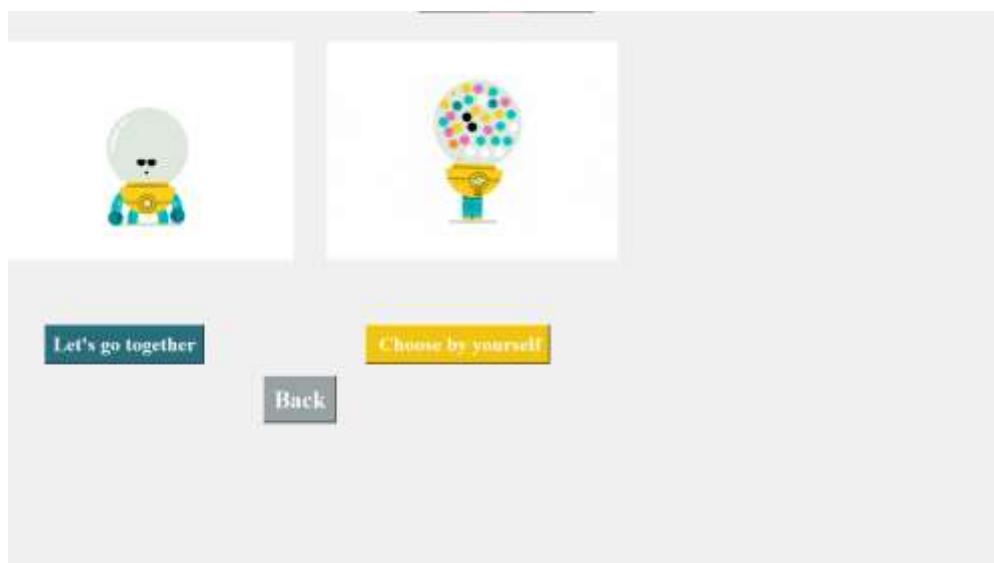


Figure 5.4: user chooses system mode

Choose by yourself

If the user chooses "Choose by yourself", they are asked to choose between "let's play" and "let's learn"

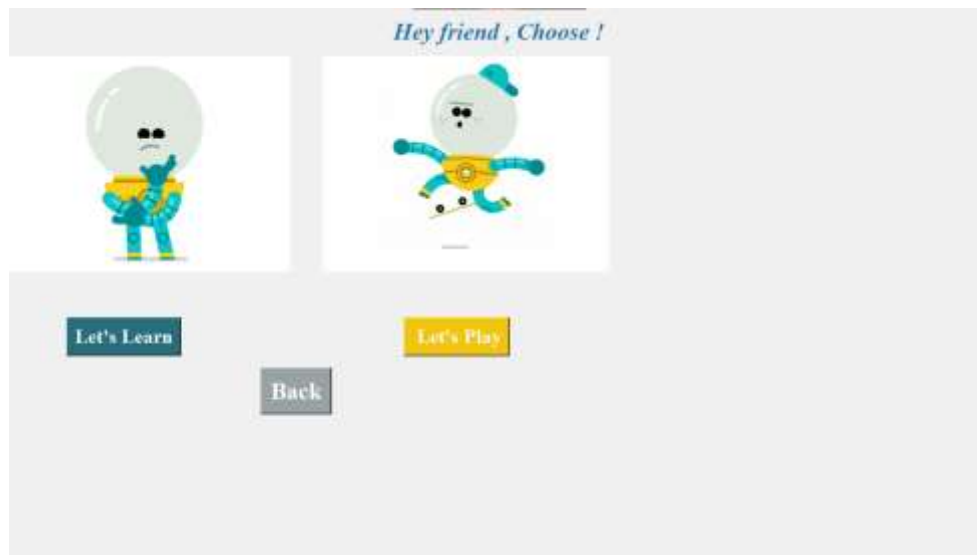


Figure 5.5: user chooses learn or play

If the user you chooses "let's learn", a list of available educational videos is displayed. The list includes the following topics: "Alphabets", "Numbers from 0 to 10", "Number from 10 to 100", "colors", "Emojis", "Abultion", "Praying", "Shapes" and "Fruit".



Figure 5.6: educational videos

Alphabet



Figure 5.7: alphabet video

Colors

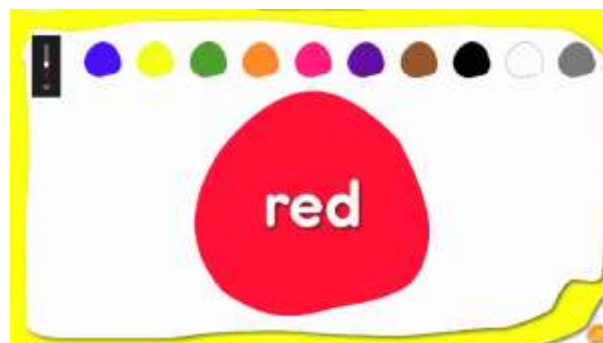


Figure 5.8: colors video

Numbers from 0 to 10



Figure 5.9: numbers from 0 to 10 video

Numbers from 10 to 100



Figure 5.10: numbers from 10 to 100 video

Emojis

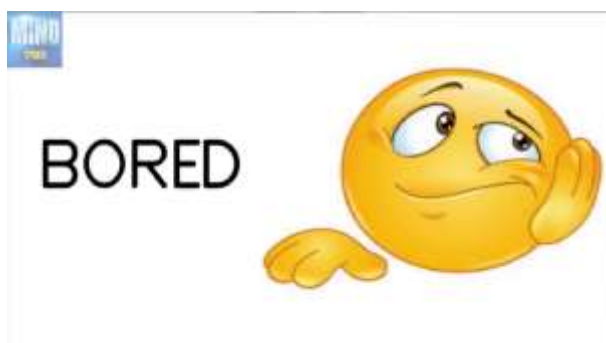


Figure 5.11: emojis video

Shapes



Figure 5.12: shapes video

Ablution



Figure 5.13: ablution video

Prayers



Figure 5.14: prayers video

Fruits



Figure 5.15: fruits video

If the user chooses "let's play", a list of available games is displayed.



Figure 5.16: games list

Spinball game

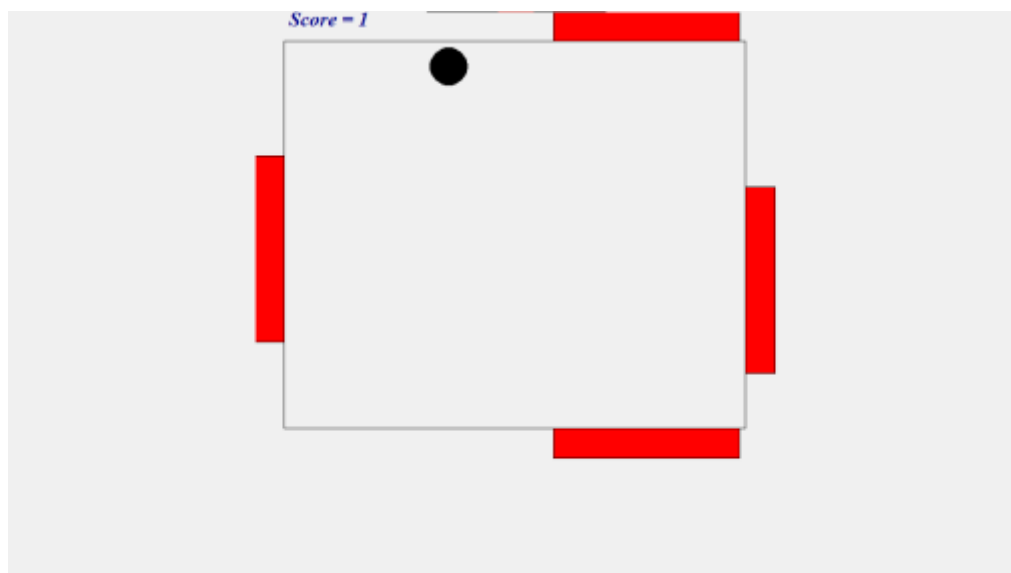


Figure 5.17: spinball game

Car game

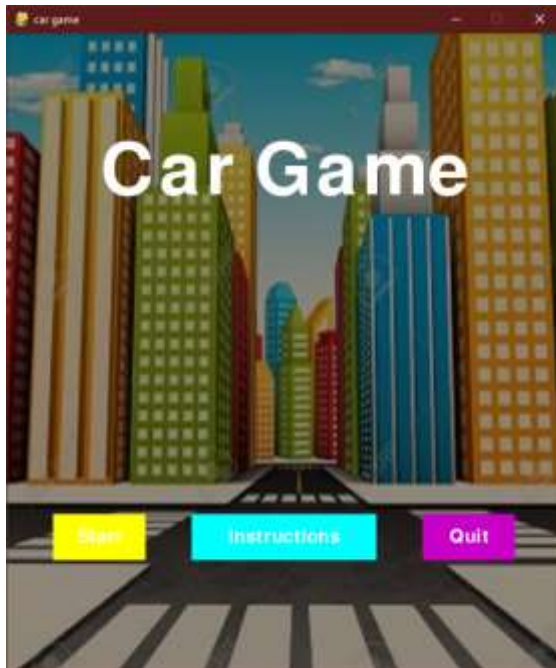


Figure 5.18: car game main menu

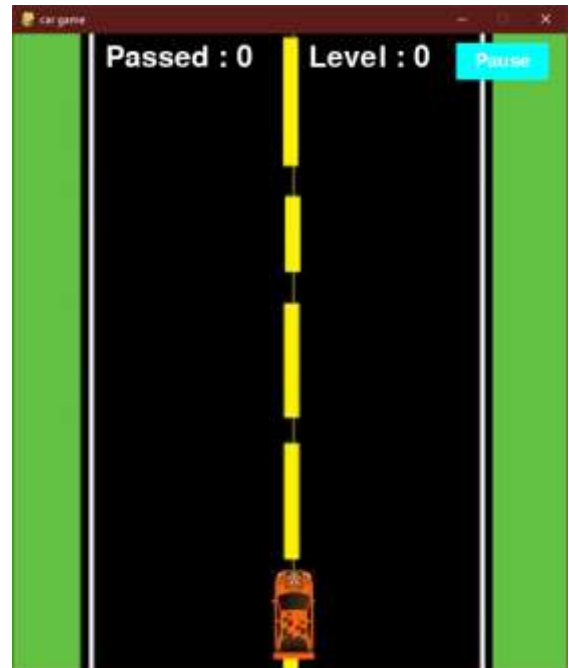


Figure 5.19: car game

Snake game

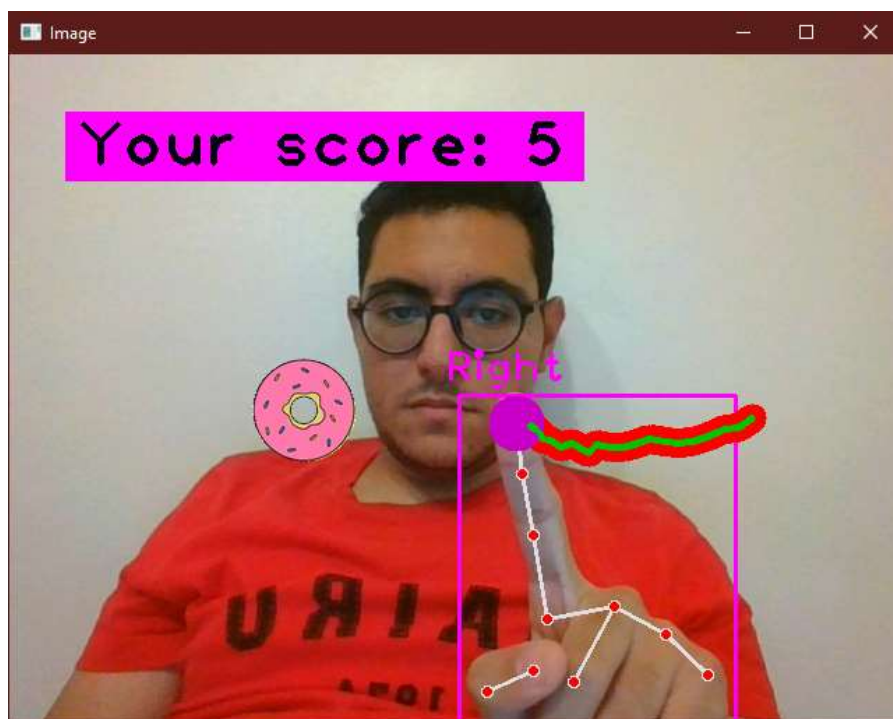


Figure 5.20: snake game

paint game



Figure 5.21: paint game

Card game : the system gives the user an object to draw vocally and the user holds the card to the camera for the system to verify.

Let's go together

If the user chooses "let's go together", the system identifies the user's emotions to determine the sequence of videos and games to be played. The sequence consists of videos and games displayed above but ordered differently depending on the user's emotions.

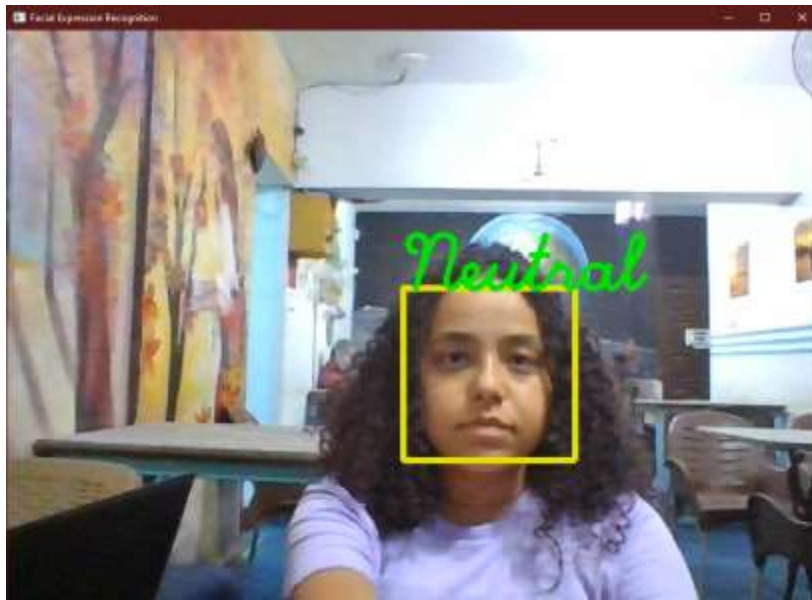


Figure 5.22: face expression (neutral)



Figure 5.23: facial expression (happy)

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The system is an interactive system that aims to educate children with ASD in a fun way and help manage the disorder and negate its symptoms. The system includes face recognition implemented using LBPH recognizer with confidence score as an accuracy metric, the system also includes facial expression recognition implemented using a custom CNN using the Keras deep learning library and the ImageDataGenerator class to preprocess and augment the image data. To prevent overfitting, the code uses early stopping, model checkpointing, and learning rate reduction techniques. The model is trained for 25 epochs, and the performance is evaluated using accuracy as the metric. The data was split into training and validation to achieve better accuracy. The proposed method has 68.19 % validation accuracy.

The rest of the system consists of educational videos on various topics and several interactive games including car game, snake game, spinball game, virtual game and card game. The games are played using hand and head movement and aims to help children gain better motor functions and a better grasp on directions and communication skills. The system was tested on one child with ASD and ADHD and he was responding well to all features and his performance and grasp of the concepts mentioned above improved over the testing time indicating that the system succeeded in achieving the desired benefits.

6.1 Future work

- The system can be used to help children with ASD adapt with the disorder and negate it's symptoms.
- Overtime, the system should improve their communication skills and motor functions and coordination.
- The system can be added to a robot or an online server to be distributed on a wider range.
- The system can be improved by turning the videos into interactive programs to reap the full benefits of the videos.
- Testing the program on a large number of end users and make necessary modifications.
- Thoroughly test each sequence of games and videos for different modes and adjust them accordingly.

REFERENCES

- [1] Amina Adamu and Abdullahi Saleh. “A Survey on Software Applications use in Therapy for Autistic Children”. Conference: 2019 15th International Conference on Electronics, Computer and Computation (ICECCO), December 2019.

- [2] Alfio Puglisi, Tindara Caprì, Loris Pignolo, Stefania Gismondo, Paola Chilà, Roberta Minutoli, Flavia Marino, Chiara Failla, Antonino Andrea Arnao, Gennaro Tartarisco,,Antonio Cerasa and Giovanni Pioggia. “Social Humanoid Robots for Children with Autism Spectrum Disorders: A Review of Modalities, Indications, and Pitfalls”. Institute for Biomedical Research and Innovation (IRIB), National Research Council of Italy (CNR), 98164 Messina, Italy, Department of Life and Health Sciences, Link Campus University, Via del Casale di S. Pio V, 44, 00165 Rome, Italy, S’Anna Institute, 88900 Crotona, Italy, Pharmacotechnology Documentation and Transfer Unit, Preclinical and Translational Pharmacology, Department of Pharmacy, Health Science and Nutrition, University of Calabria, 87036 Arcavacata, Italy

- [3] Angela Bollin, BS TR, CTRS Julia VanderMolen, Ph.D Grand Valley State University Taylor Bierwagen. “The Impact of Assistive Technology on Autism Spectrum Disorder: A Systematic Review”.

- [4] Meylinda Mari, Faaizah Shahbodin, Ibrahim Ahmad. “Assistive Technology for Autism Spectrum Disorder: A Review of Literature”. October 2018, Conference: International MEDLIT Conference 2018, At: Palace of the Golden Horses Hotel, Seri Kembangan, Selangor.

- [5] Claire A G J Huijnen, Monique A S Lexis, Rianne Jansens, Luc P de Witte. “How to Implement Robots in Interventions for Children with Autism? A Co-Creation Study Involving People with Autism, Parents and Professionals”. October 2017.

- [6] "<https://towardsdatascience.com/localization-and-object-detection-with-deep-learning-67b5aca67f22>", march 2019
- [7] Marlyn Maseri, Mazlina Mamat, Hoe Tung Yew, and Ali Chekima. "The Implementation of Application Software to Improve Verbal Communication in Children with Autism Spectrum Disorder: A Review" November 2021.
- [8] Swati Sucharita Barik and Sasmita Nayak. "Human Face Recognition using LBPH". International Journal of Recent Technology and Engineering, March 2020.
- [9] Liquan Zhao *ORCID and Shuaiyang Li. "Object Detection Algorithm Based on Improved YOLOv3" . Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education (Northeast Electric Power University), Jilin 132012, China.
- [10] M.S. Minu, Kshitij Arun, Anmol Tiwari, Priyansh Rampuria Assistant Professor. "Face Recognition System Based on Haar Cascade Classifier". Department of Computer Science Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu.
- [11] Shima Alizadeh Azar and Fazel. "Convolutional Neural Networks for Facial Expression Recognition". Stanford University.
- [12] "https://developers.google.com/mediapipe/solutions/vision/hand_landmarker". March 2023.
- [13] "https://developers.google.com/mediapipe/solutions/vision/pose_landmarker". May 2023

- [14] J. M. Rehg, A. Rozga, G. D. Abowd and M. S. Goodwin, "Behavioral Imaging and Autism," in IEEE Pervasive Computing, vol. 13, no. 2, pp. 84-87, Apr.-June. 2014, doi: 10.1109/MPRV.2014.23.
- [15] "<https://unitedrobotics.group/en/robots/nao/documentation>"
- [16] "<https://www.media.mit.edu/projects/engageme/overview/>", december 2019
- [17] " <https://carmenbpingree.com/blog/9-best-autism-apps-for-skill-development-and-confidence/>"
- [18]"<https://www.kickstarter.com/projects/jhartman/baby-bibli-the-artificial-autism-robot/posts>"
- [19] " <https://www.analyticsvidhya.com/blog/2022/01/a-comprehensive-guide-on-human-pose-estimation/>", february 2020
- [20] "https://www.theautismpage.com/qt-robot/". 2023
- [21]<https://towardsdatascience.com/face-recognition-how-lbph-works-90ec258c3d6b>, november 2017
- [22] https://developers.google.com/mediapipe/solutions/vision/face_landmarker, june 2023