Human Activity Monitoring For Mental Health Assessment

-Final Report-

Authors:

Erik Low
Adi Widjaja
Matthew Johnson
Abdullah Al Lawati
Christopher Patterson

Project Advisor:

Nikolaos Papanikolopoulos

Table of Contents:

1. Executive summary:

The Human Activity Monitoring for Mental Health Assessment group has been developing a software solution to assist doctors and mental health professionals in diagnosing neurodevelopmental disorders. This will be especially useful for reaching patients in areas with limited access to health care. The goal of this solution is to use software to analyze video and provide vital statistics for professionals in the mental health field.

The typical method for neurodevelopmental disorder detection is regular screenings throughout a patient's childhood. Specifically, with Autism Spectrum disorder (ASD), screenings include interaction with a child to elicit telltale behaviors and actions. How the child interacts in the screening is used to determine the likelihood of ASD. Although these methods are effective, they can be time-consuming, hard to administer for some patients, and must occur at specific points in a child's development. The proposed solution to aid in this process uses video capture and processing to identify the actions of humans and detect abnormalities present.

The software solution consists of three main steps to provide professionals with useful information. First, the video is processed by an open source library called OpenPose. For a given video input, OpenPose outputs JSON files for each frame containing key-point data. The key-points represent points on a human body such as the elbows, hips, or knees. These key points allow us to detect people within a video frame and extract information about the person's pose, or body stance. This extracted information can be used with the statistical analysis performed in the second step. The second step includes analysis of the vectors we obtain from OpenPose. Using an algorithm to calculate the angle of a person's arm relative to their spine, we perform a Fast Fourier Transform (FFT) to find the dominant frequencies of the motion. As an extra analysis in step two, we are using Machine Learning using a Convolutional Neural Network (CNN) to take the OpenPose data and learn different human actions within a video. The third step is writing the FFT information calculated to the input video and allows a Medical Professional analyzing the data to view it easily, helping them in making a diagnosis.

Testing for the software was performed using a train and test method for machine learning on the KTH video dataset, a dataset of videos for specific actions. This is a proof of concept for identifying actions associated with neurodevelopmental disorders. The CNN was tested and yielded an 80% success rate for identifying activities within the KTH dataset. Factors that limit the success rate of this algorithm include videos that have more than one person in the frame and when the person of interest is moving quickly or out of frame. Another limiting factor, specifically hindering the performance of our Neural Network, is the limited training dataset to classify neurodevelopmental disorders. Currently, we are using the KTH actions dataset as a proof of concept for using neurodevelopmental videos.

With the success of this software concept, the project can be expanded to be more successful and useful. To improve the results of the project, further improvement can be made by filtering out the people in a frame to focus on just one person, even if multiple people are present in a video. Working with medical professionals to develop Machine Learning training for neurodevelopmental disorders would improve the Neural Network performance. Further research can be done to develop an intuitive interface to connect medical professionals and patients using smart devices or personal computers so the project can reach people with limited medical resources. Lastly, to improve the accuracy of the results a hardware solution should be explored to provide a consistent video setup that could also track and follow a person of interest so the frame never loses them.

2. Problem definition

2.1 Introduction

Recently in mental health, a lot of attention has been focused on the monitoring of human behaviors. Assessing human behavior can provide significant information on mental health, especially for children during development. Neurodevelopmental disorders are mental disorders that primarily affect the function of the brain and neurological system. Common disorders include ADHD, OCD, autism, and schizophrenia. These disorders often cause difficulties with language, speech, learning, behavior, motor skills, memory, and more. Neurodevelopmental disorders usually develop during early stages of development during childhood. Based on parental responses to survey questions from 2006 - 2008, ~15% of children of ages 3 to 17 were affected by neurodevelopmental disorders in the United States [1]. According to multiple studies, researchers have stated that the percentage of children affected by neurodevelopmental disorders have been steadily increasing over the past few decades [1]. Although the question of why neurodevelopmental disorders occur remains unclear, early detection can help lower the chance of long-term effects.

Unless parents or caregivers are aware of all the symptoms of these disorders, the current and most obvious solution in most cases is for a clinician to observe the patient’s behaviors. This solution is often difficult and time-consuming. With the roles of engineering, sensor technology, cognitive science, neuroscience, kinesiology, and computational theory, modern technology will be able to detect the most subtle movements and behaviors. Software will be developed to be used as a tool to aid clinicians monitor human activities and behaviors for assessing mental health. This proposed solution will be developed to process real-time or non-real-time video and insert an overlay that will identify possible symptoms associated with neurodevelopmental disorders. Early versions of the software shall be able to access human behavior to identify unusual hand flapping, related to autism. The patient under observation will be able to be examined behind a non-uniform background in real-time. Future versions of the software shall be able to identify a wide range of symptoms associated with multiple disorders behind a cluttered environment, such as in public. It will be designed to be easily accessible, such as with smartphone cameras or webcams, which are readily available to most areas of the world. With this newly proposed solution, monitoring human activity and behavior for mental health assessment becomes more accessible and more reliable. Parents and caregivers will have options when it comes to assessing mental health. Clinicians will be able to examine patients with higher precision in a less amount of time. Not only would this be a valuable solution to communities with limited access to resources, but this solution will be beneficial worldwide.

2.2 Prior Work

The study of determining human behaviors from a video source is known as human action recognition. The accurate recognition of human actions in real-world video sources can be a highly challenging task due to viewpoint variations, cluttered backgrounds, and object occlusions [2]. While human action recognition has been studied for decades, rapid advances in computer vision and machine learning have brought a renewed interest in this technology and its applications [3]. Common applications of human action recognition include video surveillance, video retrieval, augmented reality entertainment, and autonomous driving vehicles [3].

Early approaches to action recognition often only analyzed a single frame at a time to classify the current action [2]. This approach, however, fails to incorporate the motion information present in contiguous frames. For some simple actions, a single frame is all that is needed to properly classify an action, which is why approaches that only classify individual frames can be quite successful for certain actions, such as running [4]. More complex actions, however, may require more frames in order to correctly classify the action. The work of [4] attempted to determine the ideal number of frames necessary to correctly recognize most human actions. They argued that a large look-ahead, or window, is not necessary to classify human action since human vision can recognize actions almost instantly. In their experiments, they found 5-7 frames, or 0.3-0.5 seconds of video, to be the ideal number for action recognition. Using more frames was found to have little impact on the success of the classifier, and only served to lower the efficiency of the classifier.

Using multiple frames, however, does complicate the classification process. Often, when only a single frame is processed at a time, a convolutional neural network is used to identify the action in each frame. Convolutional neural networks, or CNNs, have been shown to achieve superior performance on visual object recognition tasks when trained with appropriate regularization [2]. Another form of neural network, a time delay neural network or TDNN, is able to process temporal data using shift register like inputs. TDNNs are often used for applications like speech processing where temporal information must be included in the classification processes. Interestingly, time delay neural networks and convolutional both work in the same way with the only difference being in how the input data is prepared. The work of [2] showed that a 2D CNN model could be combined with the TDNN model to create a 3D convolutional network which could discriminate features in a video along both spatial and temporal dimensions.

The authors of [2] claimed their model was a single step approach to human action recognition. This is in contrast to the two-step approach of first extracting features from the video frames, and then in the second step train classifiers based on the extracted features. Their experimental methodology, however, contradicts this claim. When testing their 3D CNN model on surveillance video from the London Airport, they first used a human detector on each frame and a detection tracker between frames. This was done in order to isolate the video of each individual in the frame, and then the isolated video of each person was fed into the 3D CNN separately. One could argue that using human detection and tracking constitutes a feature extraction step, and their approach is really a two-step approach contrary to their claim. This is not to say that the work of [2] was unsuccessful or didn't offer a valuable contribution to the field. On the contrary, their model showed quite promising results on the standardized data sets it was tested it on, and in some ways, it is the basis of the approach proposed in this paper. Pointing out that [2] really used a two-step approach is only meant to illustrate that human action recognition is really a two step problem. The first step is determining the action representation, or how to best represent the action [3]. In [2], the action representation was the cropped video of individual people extracted from the surveillance video containing multiple people. The second step is the action classification, which is how to determine the action from the representation used [3]. In the [2], the 3D CNN constituted their solution to the action classification problem.

To solve the action representation problem, early human action recognition attempts often focused on tracking human body parts as input features. However, alternative methods were pursued because of the challenge of correctly and efficiently tracking human body parts [4]. Since then, however, the field of human pose recognition, determining the locations of body parts, has advanced significantly. The current state-of-the-art pose recognition library, OpenPose, is capable of performing 2D multi-person pose keypoint detection in real-time and can estimate the location 25 body keypoints, 2x21 hand keypoints, and 70 face keypoints [5]. The OpenPose API, written in C++, is based on the work of [6], [7] and [8]. This work uses a non-parametric representation, which the authors term Part Affinity Fields, to associate body parts with the individuals within the image. The previous state-of-the-art solution to multiple pose estimation by [9] was not able to achieve real-time performance since the last step used a computational expensive linear programming technique to associate part detection candidates to individuals. Because of the real-time performance provided by the OpenPose API, this project proposal suggests using the motion vectors of the motion of keypoints between frames relative to other keypoints as the action representation, and this will be used as input features to a CNN for the action classification.

2.3 Problem Statement

Many neurodevelopmental disorders are accompanied by various behavioral traits. Autism spectrum disorder, for example, is typically accompanied by repetitive physical behaviors such as hand flapping or rocking [1]. The physical nature of these behavioral attributes allows us to see and detect abnormalities. This task, however, can be difficult in rural areas with little access to health resources. Computer Vision could be used to bridge this gap, allowing software aided diagnosis or recommendations for patients who may express physical traits of a neurodevelopmental disorder. The goal of this project is to develop a system to process real-time video for tracking human movement and behavior. This system shall detect human hand flapping movements and classify it as different speeds. If time permits, this project could be extended to connect medical professionals to patients via smartphone or computer video and could be improved to analyze and detect the physical attributes associated with common neuro-disorders.

2.4 Product Design Specification

2.4.1 Customer/Client Needs

#	Need	Importance (5 is high)
1	Program shall be able to provide user feedback in real time	4
2	Program shall be able to monitor human behaviors through common smartphone video.	3
3	Program must be able to distinguish a human in an uncluttered environment	5
4	Program must be robust to identify both stationary and moving targets	5
5	Program shall track humans fully visible and near a fixed camera	5

6	System shall identify symptoms associated with disorders to aid clinicians in monitoring human behavior	5
7	System shall provide a user interface that is accessible to medical professionals without a computer science background.	3

2.4.2 Design Specifications

#	Needs	Metric	Units	Ideal Value	Acceptable Range	Importance (5 is high)
1	3,4,5	Tracking accuracy	pixels	10	±2	3
2	2,3,4,5	Tracking distance	meters	3	3-4	4
3	2	Image quality	Megapixels	12	8-12	5
4	2	Camera focal length	millimeters	4.3	4-6	5
5	1	Response time	milliseconds	4	3-5	4

6	4	Movement speed	meters/s	2	±10%	4
7	3,4,5,6,7	Success Rate	%	80	75-100	5

3. Design Description

3.1 Overview

The designed solution to solve the problem statement was to develop a software algorithm which can analyze video within a reasonable time frame and extract the poses of a human through an entire video. Once extracted, these poses are used to analyze the movements and patterns of the human throughout the video. Data and information related to any actions recognized will be displayed onto the original video. Because of the interactive nature of this program in aiding mental health assessment, the design solution will interface with a common smartphone and computer hardware to obtain video and display results to the user for further analysis.

3.2 Design Description

The software algorithm can be broken down into essentially four distinct components which work together to meet design requirements:

1. Video processing
2. OpenPose Keypoint Extraction
3. Keypoint Analysis and Action Classification
4. Presentation of Data

Together, these functional components will allow identification, tracking, and analysis of human actions within video frames and provide data of a patient’s actions to licensed clinicians.

3.2.1 Video Processing

To meet the video requirements of the design specification, the software solution will import smartphone video but can work from single images, webcams, and Flir/Point Grey IP cameras. The video will be converted into a series of individual frames, which will be done using the OpenPose library. With current specifications, we expect to process video with one person in the frame per test. This simplifies tracking due to the fact the all keypoints recorded belong to a single person. This may also increase the processing speed due to the runtime depending on the number of detected people when processing hands and facial features. However, for future applications, multi-person videos may require extra video pre-processing to distinguish between individuals and track them accordingly, which OpenPose is capable of.

3.2.2 OpenPose Keypoint Extraction

Once a video is processed into individual frames, each frame will be analyzed to detect human figures present. OpenPose provides an API library which will analyze frames and return keypoint vectors. The OpenPose library can track up to 25 body/foot keypoints, 2x21 hand keypoints, and 70 face keypoints on a single image. For simplification and quicker runtime, only the 25 keypoints of the body will be detected by OpenPose. The results are outputted in the form of JSON files, which is used to store the data structures containing the x and y keypoint values and confidence values. A JSON file is produced for every frame of a video, so the JSON files act as a time reference for a video. If a JSON file does not contain or has an incomplete set of keypoints, the JSON file will be removed and replaced with the next valid JSON file to increase the performance of upcoming procedures. A video will also be outputted with each original frame overlaid with its keypoints and keypoint vectors. An example of this is shown in Figure 3.2.2.

Note that the OpenPose keypoint extraction process requires extremely high-speed computer processing in order for this step to reach a close-to-real-time response time. Remote access was granted to use a computer consisting of 4 NVIDIA TitanX GPUs, but physical access was not due to sensitive data on the machine.

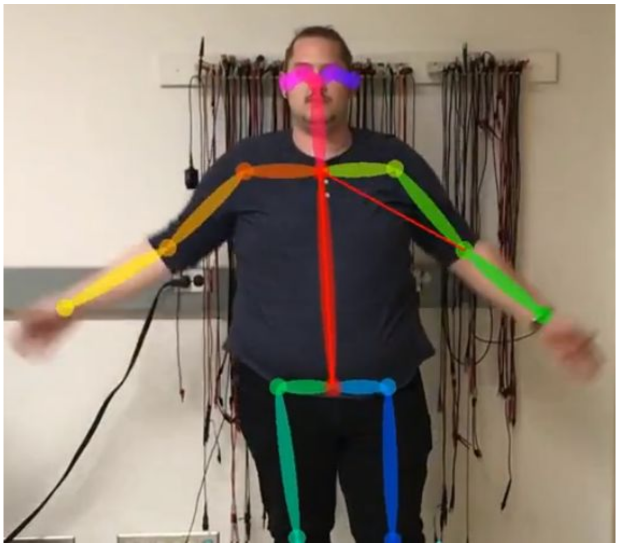

Figure 3.2.2. The output of a frame with an overlay of the location of the keypoints and keypoint vectors. Note that the two additional thin, red vectors shown between the middle hip and neck and between the neck and elbow were not created by OpenPose.

3.2.3 Keypoint Analysis and Action Classification

With keypoints on a frame acting as points on an x-y plane, vector analysis is able to be performed on the keypoints of each frame throughout a video. Vectors between these keypoints are created that will help achieve detection of movement. In this case, the angle relative to the spine and left arm is calculated and monitored for waving actions throughout the video. Fast Fourier Transforms are also performed to calculate the frequency of motion of the arm relative to the spine for waving actions.

The keypoint data structure of the imported video is also propagated through a Convolutional Neural Network (CNN). The CNN will process the keypoint data of a video and recognize certain actions 12 frames at a time throughout the video. These actions include: running, jogging, waving, clapping, and boxing. For this to be done, the CNN must first be trained. This is done by feeding in hundreds of videos from the KTH dataset, and the CNN algorithm will run through the data thousands of times to learn the distinction between the actions. 80% of the videos containing these actions are used for training the CNN and the rest are reserved for testing. The CNN will not learn to classify specific symptoms associated with ASD due to the lack of data.

3.2.4 Presentation of Data

OpenPose outputs a modified version of the original imported video. This video consists of keypoints and keypoint vectors are overlaid onto each frame, as mentioned previously. Any recognized actions listed above that are detected by the CNN throughout the video will also be displayed onto the video. The hand-waving action most closely resembles the symptom hand flapping, so further analysis will be done by the software during this action.

During the hand-waving action, two vectors will be created and analyzed to determine the angle and frequency. These two vectors will also be displayed as thin, red lines, both starting from the neck keypoint, but one ending at the middle hip keypoint and the other ending at the left elbow keypoint. The angle and frequency of motion relative to the two vectors will be displayed onto the video. If the angle surpasses a certain threshold value, the text corresponding to the angle will turn red, acting as a notification system for the user. An example of CNN detecting hand-waving and a non-hand-waving action is shown in Figure 3.2.4a and 3.2.4b.

As mentioned before, there is not enough data for the CNN to recognize specific symptoms of ASD. Therefore, this design is a proof of concept for the idea of detecting ASD symptoms.

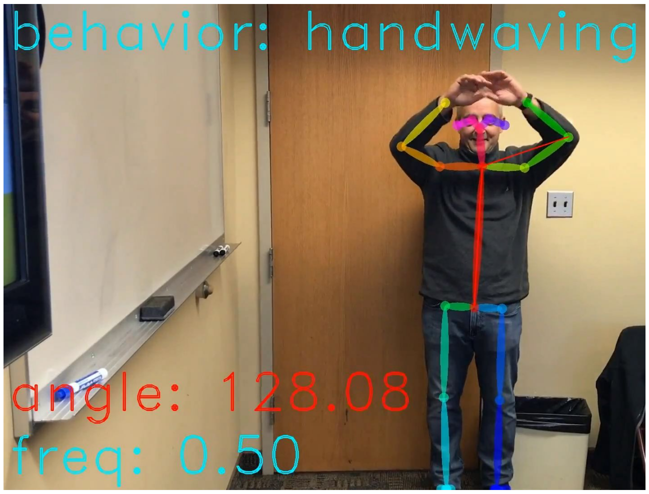

Figure 3.2.4a. An output of a video during hand-waving which includes angle and frequency calculations along with the two vectors associated with them.

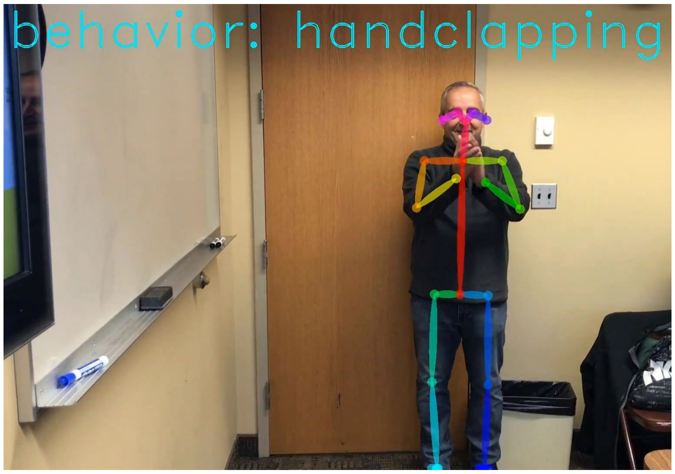

Figure 3.2.4b. An output of a video during a non-hand-waving action. The calculations of angle and frequency are not displayed. The vectors associated with these parameters are not displayed.

4. Evaluation:

In this section, the product designed by the Human Health and Activity Monitoring group will be evaluated by comparing the product to its design specifications set at the start of the project, and through the process of customer evaluation. This aims to ensure that the product has accomplished and met its purpose and the customer's needs as well.

4.1. Evaluation by Analysis:

Table 4.1 shows the specifications that will be analyzed in this section and what were the desired values to be met by this product. It also shows the methods to be used in order to confirm whether the product was able to meet the specification requirements or not.

Specification	Ideal Value	Test Method
Tracking Distance	3 meters	Taking videos at different distances ranging from 1 to 3 meters, and checking whether the key-points were visible and able to track the person's pose Extra: taking videos at distances ranging from 1 to 20
Camera Focal Length	4.3 mm	Take videos with cameras that have a focal length of 4.3 mm and checking whether the product was able to detect, process, and classify the recorded video
Image Quality	12 Megapixels	Taking videos with 12 Megapixels camera, and checking whether openpose tracked and classified the results accurately. Extra: taking videos with lower quality cameras
Movement Speed	2 m/s	Record videos of people running at speeds greater than 2m/s and check if the keypoints were displayed with a high level of accuracy

Table 4.1: Specifications that will be analyzed, their ideal values, and the test methods to verify if they were met

4.1.1. Tracking Distance

In order for this product to be effective, it has to interpret physical behaviours within a reasonable range. The specified range was to be capable of detecting activities within a 3 meters range. 3 meters was chosen to be an ideal value here, since the designing team of this product aimed to detect Autistic children's behaviour as they interact with objects, such as toys and electronic entertainment devices at a close proximity to the recording device.

In order to test whether the specifications were met or not, videos from distances ranging from 1 to 20 meters were recorded and processed. The results are shown in figures 4.1.1. and 4.1.2.

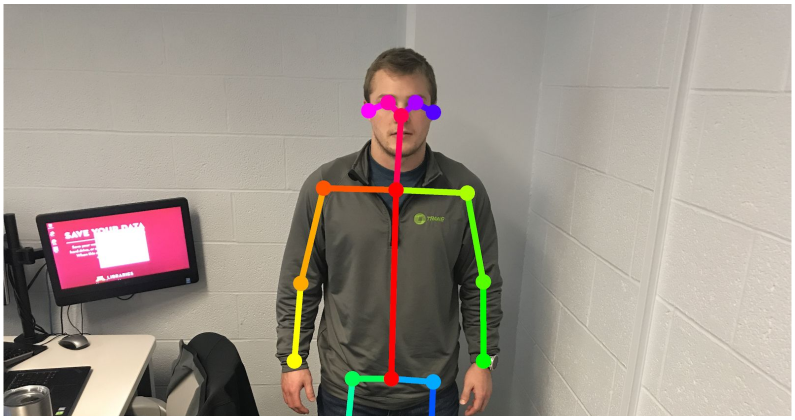

Figure 4.1.1. video recorded at a distance of 1 meter

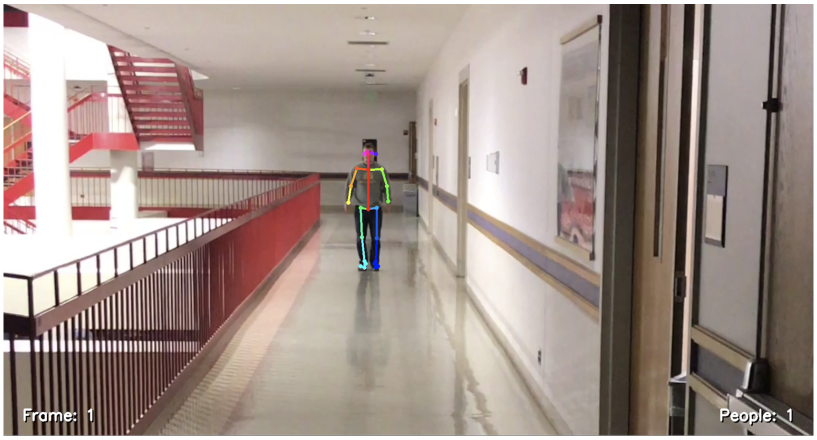

Figure 4.1.2. video recorded at a distance of 20 meters

Figures 4.1.1. and 4.1.2. both show that the product was capable of detecting movements at distances ranging from 1 to 20 meters. This adds an additional 17 meters into the required specification which might not be necessary. However, having an extended range of tracking ability expands on the capabilities of this product. The product is now not limited in detecting autistic behaviours when the child is interacting with a physical object that is 3 meters away from the recording device, but the child has more freedom now to move around and interact with many other objects within a 20 meter radius. The testing performed when designing the product was done between 1 and 20 meters, which means that the product may be able to track distances greater than 20 meters, this, however, requires further testing and validation.

4.1.2. Camera Focal Length & Image Quality:

The reason for choosing a focal length of 4.3 mm and a resolution of 12 megapixels for the ideal values of camera focal length and image quality is because almost all the smartphones in the market right now have these camera specifications. The reason for referencing the specifications to a smartphone is based on the intention of designing a mobile phone application that can encompass the capabilities of this product.

In order to test whether the specifications were met or not, videos were taken via an iPhone 7 Plus which has the specifications listed in Table 4.1.

To expand on the range of mobile phones or any other video recording device that can benefit from this product's capabilities, videos with a lower resolution were also tested to check whether the product is capable of tracking and displaying all the keypoints accurately. Figures 4.1.3. and 4.1.4. show the results of videos representing different resolutions.

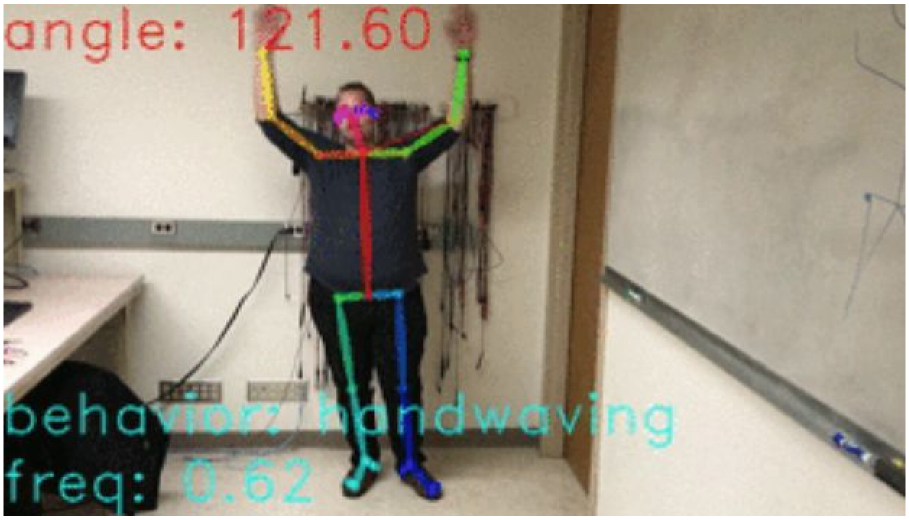

Figure 4.1.3. Video taken with an iPhone 7 Plus

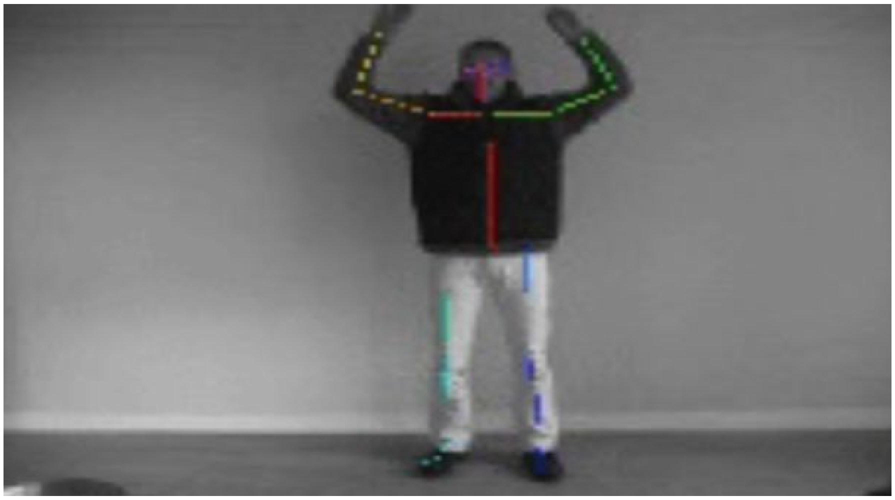

Figure 4.1.4. Low-quality video

From figure 4.1.3 it can be clearly seen that all the keypoints were displayed accurately on the pose of the person performing a hand waving behavior when the video was taken with an iPhone 7 Plus. The keypoints were also displayed accurately on the person waving In figure 4.1.4, which shows a video with low quality and resolution. This now verifies that the product is capable of detecting a wide range of physical behaviors regardless of the quality of the video.

4.1.3. Movement Speed:

A motion speed of 2 m/s was set in order to track the fast hand movements which are symptoms common amongst autistic children. In order to verify whether the product meets this specification, a video of a person running was recorded and processed to see whether the product was capable of accurately tracking the person's motion and displaying all the 25 keypoints. This test, however, will not accurately detect the speed of motion but will verify that the product can detect speeds higher than 2 m/s since a person running has an average speed of 3.5 m/s [10]. The results are shown in figure 4.1.5.

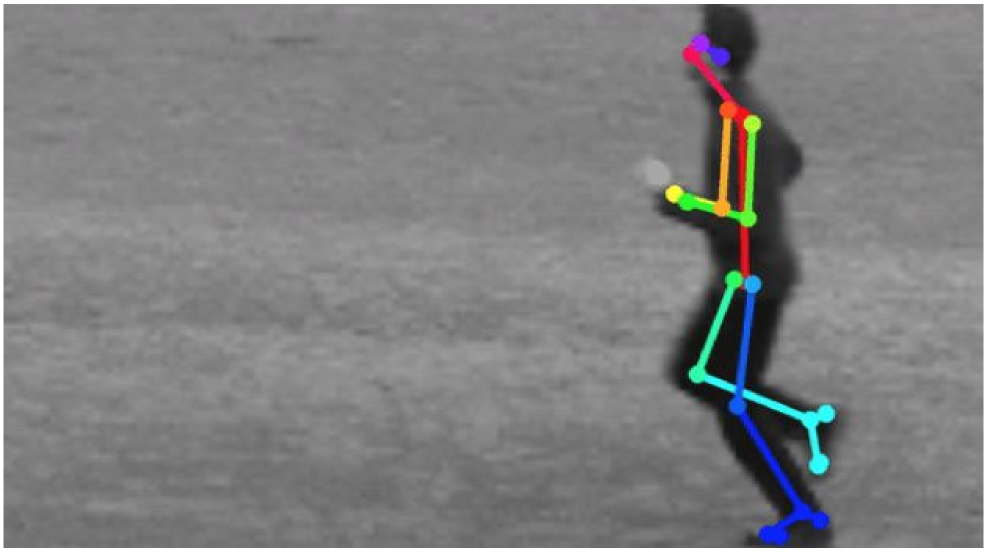

Figure 4.1.5. A person running

Figure 4.1.5 clearly shows that the product was able to detect all the keypoints accurately. This, therefore, verifies that the product can detect motions equal to 2 m/s.

4.2. Evaluation by a customer:

Abdullah Yousef Al Ghailani is a resident at Sultan Qaboos University Hospital in Muscat, who has been working in the field of behavioral medicine for the last 2 years. Al Ghailani viewed all of the demos and results generated by the Human Activity and Health Monitoring group to check and see if further improvements to this product can indeed be capable of tracking and aiding medical professionals in detecting autism in its early stages. After thorough review and analysis by himself and other consultants at the hospital as well, Al Ghailani gave the remarks shown in table 4.2.

Evaluated part of the product	Al Ghailani's feedback
The capability of this product in tracking autistic behaviors now or in the future	The product will not be able to specifically track autistic behaviors. There are many other mental disorders such as epilepsy and seizures that are accompanied by repetitive and frequent behavioral movements. Autistic behaviors are also not limited to hand flapping, many other children do different things or "ritualistic" behaviors to calm themselves down and feel much safer. These include rapid hand and leg movements and repetitive wall headbanging. However, the product has the potential of detecting general mental disorders, since it can track the frequency of certain behaviors, which is one of the greatest strengths of this product, since all of these mental disorders are accompanied by repetition.

Table 4.2. Medical student feedback on the product capabilities

5. Conclusion:

In the section, how well the design has met the initial specifications will be discussed. In addition, design strengths and weaknesses along with future recommendations will be covered. Finally, marketing opportunities for this product will be presented.

5.1. How Well Design Specifications Have Been Met:

As mentioned in section 4.1, The specifications either met or exceeded the specifications set at the beginning of the designing process. However, there is one specification not evaluated in section 4.1 which has not met the requirement. This specification was for the designed product to have a response time of 4 milliseconds. The time available for the designing team was not sufficient enough to integrate all of the steps which include; keypoint extraction, forming the extracted values in a vector, passing the extracted values through a fast fourier transform, displaying the angle and the frequency, and classifying the behavior at the end. The product does all of these steps sequentially, but further work can be done in the future to bring the response time down to the desired specification.

5.2. Design Strengths and Weaknesses:

5.2.1. Design Strengths

Many of the design strengths of the product can be seen in section 4 (evaluation) of this report. With the versatility and tracking options allowed within the OpenPose software, this product will allow for a wide range of tracking environments, movement patterns and speeds, and multiperson tracking to name a few. The filtering of the pose keypoints will eliminate any noise that is associated with the background of the environment of the subject. Therefore, this technology can be used in any environment as long as the subject is fully visible within the frame. Also, tracking can be done at a distance of 20+ meters away. With this, options for tracking methods and environments can be increased since the subject is not limited to a small radius of the previously defined 3 meter tracking distance.

A threshold for the speed at which a human can be accurately tracked was not tested. However, videos of subjects running showed the ability to track up to a speed of 4+ meters per second. This exceeded the initial specification given. Again, this is a strength due to the fact that accurate tracking can be done with a broader range of activity and movement patterns.

Another strength of this product is the field of activity recognition is in the early stages of research and development. This means that there is a large amount of area for growth. This product provides just the beginning edge of activity recognition within a small and specific application. However, there is an innumerable number of future applications that this technology can be applied to. More of these will be discussed in the following section (5.3.).

5.2.2. Design Weaknesses

With the product in the early stages of development, some weaknesses have still not been addressed. First, not much has been done for testing with new input videos. Therefore we are not sure how the system will react to videos that differ much from the KTH dataset used to train the network. This can be mitigated with additional testing and training of the neural network, however, time has not permitted to do so. Another weakness is the limited amount of data that was properly classified to be given to the system for training. This only allowed for the recognition of 6 different actions. Also, the training set used a stationary camera with the subject always facing the camera. Again with limited testing, we are not sure how the system will react to videos that differ from the dataset. Finally, the only specification that was not exceeded during the design phase of the project was real-time processing with a delay of 4ms. In its current state, this is not a big issue as it takes under 3 minutes to complete the entire processing of the video. However, steps will be taken with further development to minimize the delay of the processing.

5.3. Marketing Opportunities:

To begin, an overview of the current process for diagnosing ASD will be given. This will help to better understand how this product will improve the process. There is no medical tests, such as a blood test, to diagnose ASD. Therefore it has proven to be difficult to diagnose at an early age based on a child's behavior and development. Diagnosing occurs in two steps. First a developmental screening. If this visit shows signs of a problem a comprehensive diagnostic evaluation is needed. A comprehensive diagnostic evaluation is given by a highly trained medical specialist and is a much more thorough examination of the child's behaviour. By the age of 2, a diagnosis from an experienced medical professional can be considered very reliable. However, most children do not receive a diagnosis until much later. This is where a product like ours will benefit the most. This product will not make diagnostics but rather provide statistics and pattern recognition for the physician to analyze. This could greatly improve the process in which a diagnosis occurs. First, a system like this could provide preliminary data to the physician. This data could be gained from videos taken in the home where the child is in a more comfortable setting. If this product is not used in the home setting, it could provide a second set of eyes that will track and analyze the child's movement during all visits, not just the specified visits used to initially detect ASD. Second, video records of the child's movement could be kept and comparisons of the child's development can be made throughout all check-ups. Also, with the system being familiar with all patterns associated with ASD and possibly other disorders, it could detect any one of the unusual patterns that a physician may miss. All of these would aid in the process, in hopes that a diagnosis can be made sooner and more efficiently. The end goal would be to greatly improve the chance of early detection which will generally lead to an improved long-term prognosis of the child [11].

6. References

[1] Health: Neurodevelopmental disorders. In America’s Children and the Environment. EPA, 3rd edition, 2015.

[2] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):221–231, Jan 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.59. URL https://ieeexplore.ieee.org/document/6165309.

[3] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. CoRR, abs/1806.11230, 2018. URL https://arxiv.org/pdf/1806.11230.pdf.

[4] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008. doi: 10.1109/CVPR.2008.4587730. URL https://ieeexplore.ieee.org/abstract/document/4587730.

[5] Gines Hidalgo, Zhe Cao, Tomas Simon, Shih-En Wei, Hanbyul Joo, Yaser Sheikh, and Yaadhav Raaj. Openpose. GitHub Repository, 2018. URL https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. CoRR, abs/1611.08050, 2016. URL http://arxiv.org/abs/1611.08050.

[7] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. CoRR, abs/1602.00134, 2016. URL http://arxiv.org/abs/1602.00134.

[8] Tomas Simon, Hanbyul Joo, Iain A. Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. CoRR, abs/1704.07809, 2017. URL http://arxiv.org/abs/1704.07809.

[9] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. CoRR, abs/1511.06645, 2015. URL http://arxiv.org/abs/1511.06645.

[10] Long III, L. L., & Srinivasan, M. (2013). Walking, running, and resting under time, distance, and average speed constraints: optimality of walk–run–rest mixtures. Journal of The Royal Society Interface, 10(81).

[11] Screening and Diagnosis. Autism Spectrum Disorder, U.S. Department of Health and Human Services. https://www.cdc.gov/ncbddd/autism/screening.html