Real Time Implementation of a Machine Learning Model Sign Language Recognition System Using Human Pose Estimation

Kristian Yair González Márquez

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics



September 2023

Supervisor: David McQuaid

**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Capstone Project |
| **Assessment Title:** | Real Time Implementation of a Machine Learning Model Sign Language Recognition System Using Human Pose Estimation |
| **Lecturer Name:** | David McQuaid |
| **Student Full Name:** | Kristian Yair González Márquez |
| **Student Number:** | 2022421 |
| **Assessment Due Date:** | 22/09/2023 |
| **Date of Submission:** | 22/09/2023 |
| **Word Count:** | 7,028 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**INDEX**

[FIGURES INDEX 3](#_Toc1)

[TABLES INDEX 3](#_Toc2)

[Real Time Implementation of a Machine Learning Model Sign Language Recognition System Using Human Pose Estimation 5](#_Toc3)

[Objectives 6](#_Toc4)

[Possible Elements of Validity to Apply in this Project 6](#_Toc5)

[CHAPTER 1 LITERATURE REVIEW 7](#_Toc6)

[1.1 Sign Language 9](#_Toc7)

[1.2 Object detection and object segmentation 10](#_Toc8)

[1.3 Human Pose Estimation 13](#_Toc9)

[1.4 Deep Neural Networks 15](#_Toc10)

[1.5 Sign Language Recognition Systems 18](#_Toc11)

[1.6 Challenges 21](#_Toc12)

[1.7 Conclusion 22](#_Toc13)

[CHAPTER 2 METHODOLOGY 23](#_Toc14)

[2.1 Ethical Considerations 23](#_Toc15)

[2.2 Primary Research Method 23](#_Toc16)

[2.3 Sampling 24](#_Toc17)

[2.4 Dataset Description 25](#_Toc18)

[2.5 Exploratory Data Analysis 26](#_Toc19)

[2.6 Keypoint Collection 30](#_Toc20)

[REFERENCES. 33](#_Toc21)

FIGURES INDEX

[Figure 1.1 Machine Translation methods distribution between 2004 and 2022. Taken from (De Coster et al., 2023) 20](#_Toc1)

[Figure 2.1 Frequency Data Distribution in Training Dataframe Top 50 Words 28](#_Toc2)

[Figure 2.2 Frames per Video in Training and Testing Dataframes Top 50 Words 30](#_Toc3)

TABLES INDEX

[Table 2.1 Top 50 Glosses with more videos in training ASL Citizen Dataset 28](#_Toc1)

[Table 2.2 Descriptive statistics of Training and Testing Datasets combined 29](#_Toc2)

Real Time Implementation of a Machine Learning Model Sign Language Recognition System Using Human Pose Estimation

*Machine Learning algorithms for Signal Language Recognition have been receiving more interest in recent years because they will provide more opportunities and improve communication in case of success to the deaf and hard of hearing population.* *Machine learning and deep learning algorithms can be trained on labelled datasets of sign language gestures, allowing them to learn the patterns and variations in hand movements, gestures and signs*.

*Human Interpreters are used as a way for traditional sign language translation but this is time consuming, expensive, and there is not always someone available to carry it out. Because of that finding an accurate and efficient machine learning algorithm has started to be a big point of interest, also it is essential to mention that there are some systems that require to have a device attached to help with the interpretation, however, this is expensive and we fall again in the same situation as Human Interpreters, they are not always available.*

*By developing this system considering only a camera as input, capturing the human joints position prior to the classification, is an interesting approach that has been spoken for a few years now, and working on this there is a good chance to make a meaningful contribution in the data analytics area while helping in the inclusion of people in the deaf, mute, and hard of hearing community. Furthermore, this area of study requires a diverse skill set, including computer vision, object recognition, data pre-processing, machine learning, and deep learning which could help to improve significantly my knowledge, experience and expertise in data analytics.*

Objectives

* Determine the impact of different factors on the performance of Sign Language recognition models, such as lighting conditions, camera angles, and variations in hand gestures, and develop techniques to mitigate these factors.
* Implement a machine learning model testing different techniques to recognize and translate American Sign Language.
* Evaluate the usability and effectiveness of the interface in real-world settings considering speed and accuracy of the classification when using commodity hardware.

Possible Elements of Validity to Apply in this Project

Current: A lot of research has been done on Sign Language Recognition using machine learning algorithms, however because of the nature of technology constantly evolving, a lot of that research is now outdated due to recent developments in visual recognition systems as well as Deep Learning algorithms, so with the rapid advancements in these areas, it is important to ensure that the research project is based on the most current and up-to-date information available, this can be done verifying multiple sources of information even including grey materials but with caution.

Reliable: Plenty of information about the latest developments in visual recognition with the newest techniques is done in GitHub repositories or social network media like YouTube without strong documentation to back up their ideas. Because of this verifying the trustworthiness of the source of information is essential to minimize the risk of falling into misinformation, also, documenting the sources used in this project can also help ensure the transparency of the results, backing up the ideas on those grey materials with other sources can also be a good way to mitigate this risk.

# LITERATURE REVIEW

Millions of deaf and people with difficulties hearing use sign language to communicate between each other and also with other people, communication with signs could be difficult for individuals who are not fluent in this complicated expressive language.

Developing a sign language recognition system has the potential to shorten the communication barrier by translating automatically sign language to spoken/written language. This can be done using computer vision techniques capturing and analysing the visual features of sign language, passing it through a machine learning algorithm to classify it and translate the signs.

Sign language recognition systems have started to be a special research subject in recent years due to the recent advancements in computer hardware and software, as well as sign language data sets that have been developed for training and testing these systems.

The current state of research focused on the previously used algorithms and techniques will be treated in this literature review to identify what can be done in this project differently in order to have the most accurate and efficient model possible taking in consideration that the final goal of this research project is the implementation in a real time situation.

The chosen topics for this literature review were selected carefully, considering their importance as background knowledge prior to starting the Data Analysis Project, the literature review begins with a summary of Sign Language, providing an essential understanding of this unique form of communication. By exploring the fundamental aspects of Sign Language, to have an overview about it before jumping into the creation of models without a clear direction on how they should work.

Next, the literature review encompasses object detection, a crucial concept to explore Human Pose Estimation as once the person and their body parts are accurately detected, the model can focus solely on analysing and interpreting those specific elements, effectively disregarding the background information present in the image.

Deep Learning was also touch as it is behind the frameworks for Human Pose Estimation and will be also the responsible to perform the classification of the Sign Language.

And finally, a quick review on challenges that people working on Sign Language Recognition systems has encountered in the past because by exploring these challenges, such as limited datasets or variability in sign language gestures, it can be considered while implementing various techniques and models in their own projects.

It is important to mention that as the Data Analysis Project progresses, the literature review may be subject to change in subsequent submissions to incorporate new findings, methodologies or insights obtained during the testing and implementation phases. This adaptability ensures that the project remains updated, aligned with current research trends, and ultimately enhances the overall quality and effectiveness of the project.

* 1. Sign Language

According to (WHO, 2023) 5% of the world population or 430 million people have a certain degree of hearing loss and nearly 2.5 billion people by 2050 will have this condition. Sign language is a structured set of hand gestures with specific meanings used by hearing impaired people to communicate in everyday life (Konstantinidis et al., 2018) that allows them to communicate between each other and with the rest of people in a non-verbal way, a deaf person is a person with a disability to hear and someone mute is who has a disability to communicate verbally. It's challenging to communicate with others as they can't speak or listen and there is where sign languages come in useful because allow people to communicate without using spoken language (Sharvani Srivastava, 2021b). Sign Language is a structured form of hand gestures involving visual motions and signs to achieve the communication involving head, facial expressions, arms, hands, body, and fingers to represent the words (Cheok et al., 2019). There are several categories of hand gestures, including conversational gestures, controlling gestures, manipulative gestures, and communicative gestures (Ying Wu and Huang, 1999). Sign Language has its own vocabulary and grammar that is completely different from spoken and written languages, for transmitting information spoken languages employ the oratory abilities to create sounds that are mapped against certain words and grammatical combinations. The oratory parts are next absorbed and processed by the auditory faculties using visual senses (Sahoo et al., 2014).

Since studies in the 1960s and 1970s established that linguistic processes were not limited to the spoken modality, sign language study has evolved at an exponential rate (Stokoe, 2005), (Klima and Bellugi, 1979), over the course of four decades, studies of sign language have made substantial contributions to our knowledge of language, thought, and social interaction. Sign language is utilised in different regions of the world in the same way that spoken languages are, there exist different sign languages over the world some examples of these are the Japanese, British, Indian, Arabic, and American sign language (Zeshan, 2006). The most common visual language used by the deaf community in North America was the focus of several of these investigations which is American Sign Language (ASL) (Hauser et al., 2016).

ASL is a complete natural language with a number of linguistic features that set it apart from English (Liddell, 2003), some have speculated that ASL developed more than 200 years ago as a result of the blending of indigenous sign languages with French Sign Language (LSF, or Langue des Signes Française) (Cagle, 2010). Some features of LSF and the original local sign languages have made their way into modern American Sign Language, which is a rich, complex, and mature language. The current versions of ASL and LSF are two separate languages. Although they share some common indicators, neither set of users is capable of understanding the other's meanings (NIDCD, 2021), (Hosain et al., 2020) noted that about 6000 hand gestures are used to represent common words in American Sign Language while finger spelling is used to represent less common words and proper nouns.

American Sign Language is one of the most influential, as a result of their introduction in newly emerging educational systems, some sign languages have a high percentage of (old) ASL (-influenced) lexicon (Kusters, 2020). Not just in situations where the use of a national sign language was not yet institutionalized ASL was imported, the status of ASL and/or the accessibility of resources in ASL are likely significant factors in these shifts. Teachers in some deaf schools initially started out using a local or national sign language before switching to ASL (Kusters, 2021).

Signs can be written down in textual form using glosses. Glossing means choosing an appropriate English word for signs in order to write them down, although it is not the same as translation, it is conceptually very similar. A gloss on a signed story can be a series of English words written in small capital letters that correspond to the signs in an American Sign Language story (Othman and Jemni, 2012), for example, children who are fluent in American Sign Language and have a need to go through the process of learning to read English can benefit from using ASL gloss, which can be thought as the "elusive" intermediary system (Supalla et al., 2017), this SL gloss annotation form is crucial to help machines understanding and processing the sign language (Johnston, 2008).

* 1. Object detection and object segmentation

(Brownlee, 2019) refers to Object detection as a set of computer vision tasks related between each other involving the identification of objects in digital photographs. Basically, these tasks, can be divided in two major tasks, image classification, that is the class prediction of an object in an image, and object localization that is, as it name defines, locating one or more objects in an image, usually drawing a bounding box around them.

These object detection models can be classified in two major types: a two stage detector that locates the regions in which objects are present in the first one, and then on the second stage the object classification is done, in the other hand, for the one stage detector, it feeds the entire image to a CNN in one single step to locate the regions of interests and classify all the objects in the image (Zhao et al., 2019). These models produced highly accurate results in less computational time and outperformed traditional machine learning algorithms due to the recent development of high-performance GPUs for computation, and large benchmark datasets availability for training models (Alzahrani and Al-Baity, 2023).

Image segmentation is the association pixels process in an image matching them with their respective object class labels. This process is used in a lot of industries such as healthcare, transportation, robotics, fashion, home, and tourism. (Wang et al., 2022), it is also defined as the process of dividing images into regions with different features and extracting the regions of interest (ROIs). Image segmentation presents two challenges: how to define "meaningful regions" because the uncertainty of visual perception and the diversity of human comprehension lead to a lack of a clear definition of the objects, making image segmentation an ill-posed problem trying to effectively represent the objects in an image (Yu et al., 2023).

Object detection is essential in many computer vision tasks for analysing and comprehending visual data due to its ability to identify and locate specific objects within a given image or video, object detection has a wide range of applications thanks to this. Classification algorithms benefit from object detection because it allows them to focus on the information they need while ignoring irrelevant background. In Sign Language Recognition for example, once the hands are identified, the algorithm can ignore the rest of the scene and focus solely on the hands, leading to more precise interpretations of the signs being made. This focused approach not only improves accuracy but also speeds up the training process, as the algorithm does not require data from as many diverse situations. By leveraging object recognition, the algorithm can target specific objects of interest, resulting in faster and more effective training (Sunmok Kim, 2018).

Object detection also extends beyond SLR systems. Object detection is crucial for human pose estimation because it plays a vital role localizing where the body joints area (Amadi and Agam, 2023). By utilizing robust person detectors and concentrating on joint detection within bounding box regions, the algorithm can bypass the challenges of dealing with large-scale changes in images and unnecessary information. This focused method makes it easier to estimate human poses, which improves accuracy and performance. The algorithm can save time and effort by focusing its resources and computational power on the relevant objects of interest and discarding everything else (Wang et al., 2020).

Overall, object detection is important because it provides algorithms with valuable context for making decisions based on what they see in an image, SLR systems benefit greatly from object detection because it helps them better understand and interpret visual data in a variety of ways, including by increasing classification accuracy, boosting training efficiency, and allowing for precise pose estimation.

* 1. Human Pose Estimation

Human Pose estimation is the process of finding human figures in pictures and videos and figuring out which joints (keypoints) are present (Moryossef et al., 2021). It is essential for giving machines the ability to see and understand people and their interactions (Cao et al., 2021), 3D human pose estimation is widely increasingly used for lots of applications such as sports instruction, limb rehabilitation training, augmented reality, and intelligent security (Meng and Gao, 2021).

Deep learning based top-down and bottom-up pose estimation are the two categories in which human estimation pose algorithms are divided (Gui and Luo, 2022).

Traditional top-down human pose estimation detects people in an image or video and draws a bounding box around them using object detection, and then a pose estimator extracts body keypoints from the bounding box. This simple method has drawbacks like high computational cost and runtime proportional to the number of people. Bottom up is the opposite to top down as it draws keypoints on the image and then uses part affinity maps to map it to different people in the image (Gojariya et al., 2021), the most significant benefits of utilizing these second method are the lightweight network and the accelerated processing times (Martinez et al., 2017).

Nowadays, there exist multiple pose estimation techniques, in this literature review, the actual state of the art will be reviewed. The first one to be reviewed is MediaPipe, a framework developed by Google for creating pipelines that carry out inference over any type of sensory data building a perception pipeline. A perception pipeline processes data from the real world using sensors and algorithms to create digital representations that can be analysed and understood by a computer system, face detection and segmentation, hands detection, pose detection, and any kind of object recognition are examples of perception pipelines (Lugaresi et al., 2019). For tasks like object detection, face detection, hand tracking, and pose estimation, it offers a wide variety of pre-built models which are trained on sizeable, varied datasets that are relevant to the task at hand like the COCO dataset which contains over 200,000 images of people in various poses and activities. They serve as the skeleton of nodes, edges, or landmarks, tracking important points on various body parts and each coordinate point is normalized in three dimensions (Halder and Tayade, 2021).

For this project four MediaPipe models listed below are of particular interest.

* Hand Landmark Model: locates 21 hand-knuckle coordinates as keypoints within the identified hand regions. About 30K real-world images and several rendered synthetic hand models imposed over different backgrounds served as the model’s training data. A palm detection model and a hand landmarks detection model are included in the hand landmarker model package. First, the palm identification model locates hands inside the input image, and then the hand landmarks recognition model recognizes specific hand landmarks on the palm detection model's cropped hand image.
* Face Mesh Model: similar to the hand landmark model, it detects face landmarks and facial expressions in selfie-like images and videos, producing a total of 468 3D keypoints. It performs the task using two deep neural network models, a face location detector that operates on the entire image, and a 3D face landmark model that operates on those locations and uses regression to predict the 3D surface.
* Pose Landmark Model: locates human body landmarks in images using machine learning, and they can process either a single image or an ongoing stream of images. The model produces 33 body pose landmarks in image coordinates and three-dimensional (x, y, z) world coordinates.
* Holistic Landmark Model: it is no more than the combination of the three models described above allowing to analyse full body gestures, poses, and actions giving a total output of 543 landmarks (33 pose landmarks, 468 face landmarks, and 21 hand landmarks per hand) in real time.

More MediaPipe documentation is available at the official Google’s Developer MediaPipe site (Google, 2023).

Another method is using OpenPose, OpenPose was developed by researchers from Carnegie Mellon University and maintained working by Yaadhav Raaj and Ginés Hidalgo and was the first known method to do a whole body pose estimation consisting of three different blocks, body-foot detection, hand detection, and face detection which combined together output a total of 135 keypoints divided as follows: 25 body keypoints, 70 face keypoints, and 40 hand keypoints (Martinez, 2019). Openpose's CNN architecture is multi-stage. A CNN analyses the picture, producing a set of feature maps that are sent into the first stage. The first stage generates a collection of PAFs by concatenating the prediction with the original image features repeatedly to get improved predictions. Using the same iterative approach as the first stage, the second stage predicts confidence maps. Part association is assisted by PAFs (Part Affinity Fields), while part detection is assisted by confidence maps. Each stage is composed of up of numerous convolution blocks that are generated by three 3x3 convolutional kernels at the same time (Badiola-Bengoa and Mendez-Zorrilla, 2021).

And the last one that will be reviewed is another object detection model which includes a module for pose estimation as MediaPipe, its name is YOLOv8, YOLO states for “You Only Look Once” and was first released by (Redmon et al., 2016), since its creation, as is an open-source framework different people and companies such as Meituan and Ultralytics have been involved in improving and developing these algorithms with different versions until the latest release of YOLOv8 (Terven and Cordova-Esparza, 2023). 2016 saw the release of YOLOv2, which added batch normalization, anchor boxes, and dimension clusters to the original model, with the help of a more effective backbone network, multiple anchors, and spatial pyramid pooling, YOLOv3 was released in 2018, further improved the model's performance. Innovators like Mosaic data augmentation, a new anchor-free detection head, and a new loss function were included in the 2020 release of YOLOv4. The performance of the model was further enhanced by YOLOv5, which also added fresh features like automatic export to well-known export formats, integrated experiment tracking, and hyperparameter optimization. Many of Meituan's autonomous delivery robots use YOLOv6, which the company open-sourced in 2022, Pose estimation on the COCO keypoints dataset is one of the extra tasks that YOLOv7 added, and finally, YOLOv8 builds on the success of earlier versions as a cutting-edge, state of the art model by adding new features and enhancements for improved performance, flexibility, and efficiency. The full spectrum of vision AI tasks, such as detection, segmentation, pose estimation, tracking, and classification, are supported by YOLOv8. Because of its adaptability, YOLOv8 can be used in a variety of contexts and applications YOLOv8 pose module is the model that will be useful to test on this project as similar to OpenPose and MediaPipe, it produces an array with 17 keypoints exporting to express the human position (Ultralytics, 2023).

Human Pose Estimation can be used for action recognition, the idea behind this is to extract body joint locations (keypoints) and then use this as input to another algorithm/neural network to select visual features in space and time to perform the action classification (Luvizon et al., 2018).

* 1. Deep Neural Networks

Artificial Neural Networks (ANNs) often known as neural networks are innovative systems and computational approaches for machine learning, knowledge demonstration, and eventually the application of acquired information to maximize the output responses of complex systems (Chen et al., 2019). Artificial neural networks are built in the same way as the human brain which has billions of neuron nodes connected between each other, each neuron has a cell body that processes information by transporting it to and from the brain (van Gerven and Bohte, 2017).

The system is formed of a lot number of highly linked processing components known as neurons (Walczak and Cerpa, 2003), which work together to solve problems and communicate information via synapses (electromagnetic connections in the human brain). The neurons are interconnected and organized into layers. The data is received by the input layer, and the final outcome is generated by the output layer with one or more secret layers placed between the two of them (Dastres and Soori, 2021).

Deep Neural Networks are neural networks that including the input and output layers have more than three layers of neurons. These layered representations are using models referred to as "neural networks," which are arranged into literal layers and stacked one on top of the other (Schmidhuber, 2015).

In the 21st century, deep learning emerged and offered a neural network language that was superior to Natural Language Processing (NLP) due to its ability to vectorize multiple features against a sequential dataset in contrast to NLP models significantly enhancing unsupervised learning (Md Asif Jalal, 2018). This area of machine learning makes use of several high level abstraction models, deep learning and computer vision advancements have already produced valuable products for audio manipulation, face recognition, and vehicle detection.

The Deep Learning Framework provides a huge advancement for designing, training, and validating neural networks. Since signs are static, we can concentrate on hand posture and placements using real-time recognition technology, which can operate more quickly and provide an accurate translation. Different approaches had been used in previous works to do the SL classification, two of which are Convolutional Neural Networks and Recurrent Neural Networks.

The CNN model is an important Neural Network component used for image recognition and classification during sign and face detection or recognition (Pathak and Maheshwari, 2019). CNN models are composed of neurons with weights and biases. Specific neurons receive input data in response to actions, and weighted sums take over, activating certain functions and producing certain outputs. The CNN models are frequently used in multi-channel images (Dhulipala, 2022). CNNs as ANNs are inspired by human nature, CNNs emerged from the study of human’s visual cortex and thanks to the advancements in computation and data availability in the last years, these Deep Neural Networks have achieved an amazing performance on image recognition tasks, the most basic architectures stack a few Convolutional Layers, Pooling Layers, and so on making the image smaller and smaller between layers, finally passing through a common neural network which layers are the fully connected layers, these ones are the ones in charge of generating the output (Géron, 2019), the convolutional layer determines the output of neurons connected to important local regions, the pooling layer reduces the number of parameters by performing down sampling along spatial dimensionality, and the fully connected layers perform the same tasks as any standard Artificial Neural Network, producing class scores from the activations to be used for classification (O’Shea and Nash, 2015).

Recurrent Neural Networks (RNNs) are supervised machine learning models that are composed of artificial neurons with one or more feedback loops, feedback loops are recurring cycles across time or sequence optimizing the network weights with the goal of minimizing the difference between the output and target pairs (i.e., the loss value), the simplest RNN has three layers which are input, recurrent hidden layer (containing the feedback loops), and output layer (Salehinejad et al., 2017).

Long short-term memory (LSTM) is a kind of RNN architecture that can remember values at arbitrary intervals. They are used to classify, process, and predict time series with known time lags and unknown durations. The LSTM is known as the cell state, and its recursive nature is shown by a looping arrow. As a result, the prior interval’s data is saved in the cell state. The cell state is adjusted by a remember vector situated beneath it, while the input modification gates adjust it. The gates also teach the network what to save, what to forget, what to remember, what to focus on, and what to output. The cell and hidden states are used to collect data for processing in the following state trying to resolve the gradient vanishing problem (Yamak and P. K. G, 2019) allowing to solve complex and artificial long time lag tasks (Hochreiter and Schmidhuber, 1997), additional to the LSTM units, there is another approach introduced a couple of years ago by (Cho et al., 2014) which name is Gated Recurrent Unit (GRU) consisting of an Update gate and a Reset Gate, being the Update gate an assistant to the model in determining how much past knowledge (from earlier time steps) should be passed on to the future, and the Reset Gate helps the model to decide how much past information has to be forgotten (Chung et al., 2014), LSTM and GRU can also be combined to achieve higher accuracies and this has been demonstrated by (Kothadiya et al., 2022) where six different architectures were tested LSTM, GRU, LSTM-LSTM, GRU-GRU, LSTM-GRU, and GRU-LSTM being LSTM-GRU the one achieving the highest accuracy predicting Indian Sign Language video frames over 11 signs.

* 1. Sign Language Recognition Systems

Recent years have seen remarkable development in AI technologies related to Natural Spoken Languages, however, sign languages haven’t seen similar progress, especially in terms of word and sentence recognition even when sign language recognition technology has seen some progress over the years, with the advent of machine learning, particularly deep learning, more advanced recognition models have emerged but still not at the same level than spoken languages (Selvaraj et al., 2022). These models can learn and extract features automatically from large datasets, improving recognition accuracy, the use of machine learning techniques enables computerized systems to make decisions based on data and experience. Two datasets, a training dataset and a testing dataset, are required by the classification algorithms. The classifier learns from the training sets experiences, and the model is evaluated using the testing set (Rosero-Montalvo, 2018). Numerous authors have created effective methods for data collection and classification which can be divided into two categories based on the data acquisition method, these categories are direct measurement methods and vision-based approaches (Sarma and Bhuyan, 2021).

Direct measurement methods that use devices such as motion data gloves, motion capturing systems, or sensors (Kudrinko et al., 2021), basically any electronic system that can convert the signal, in this case Sign Language, into an electric or audible signal acting as a bridge for communication. However, there are some issues with sensor data acquisition, such as noise, poor human manipulation, and a faulty ground connection, as well as making the system more expensive because it requires specialized devices and training to use them adequately (Elakkiya, 2020). The extracted motion data can be used to track fingers, hands, and other body parts accurately, leading to the development of robust SLR methodologies (Sharvani Srivastava, 2021a), one of the most recognized devices of this kind is the Microsoft Kinect sensor, which captures a Red-Green-Blue (RGB) image and a depth map thanks to its infrared project, infrared sensor and RGB camera integration in one single device (Zhang, 2012), one example of success using the Kinect for American Sign Language recognition that was presented by (Cao Dong et al., 2015) where a Random Forest Classifier is fed with the joint angles in order classify 24 static signs reaching a 90% of accuracy.

The vision based SLR systems can work with a device as simple as a laptop webcam or phone camera getting RGB images to extract discriminative spatial and temporal information. Because it is not necessary to physically attach sensors to humans, vision based systems have recently gained popularity in recent years despite being prone to reliability issues in the past due to background noise, colours, and lighting variations in the real world (Kin Yun Lum, 2020). But nowadays with the latest developments in vision and machine learning, the classification of images has considerably improved, and with this, the vision based Signal Language Recognition algorithms. Sign Language recognition is not a new problem in computer vision, researchers have used classifiers from a variety of categories over the last two decades, which we can be roughly divided into linear classifiers, neural networks, and Bayesian networks (Anagha.G, 2022).

(De Coster et al., 2023) did recently an interesting article analysing the state of the art in Sign Language Recognition systems over the last years, the below table shows a selection of 57 research papers dedicated to SLR, and can be seen the evolution of them, from 2004 to 2018, all of them were using Statistical Machine Translation, but then, since 2019 all of them have started to use Neural Machine Translation, except for one article (Luqman and Mahmoud, 2020), where they used Rule-based Machine Translation.

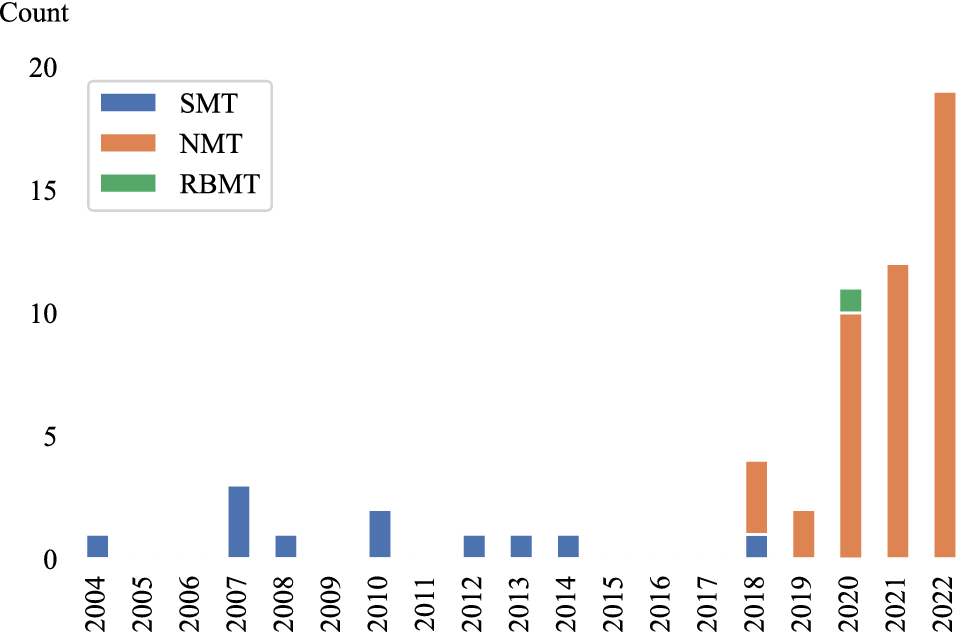


Figure 1.1 Machine Translation methods distribution between 2004 and 2022. Taken from (De Coster et al., 2023)

Another point to highlight is how the research for sign language, has increased since 2018 passing from only 10 since 2004, to 47 in the last years which corresponds with the trend of Neural Machine Translation systems.

Since OpenPose launched in 2016, different algorithms for pose estimation have been widely used trying to solve action recognition problems due to its simplicity and high accuracy to implement the different methods to do this, OpenPose and similar frameworks feed the vision based systems with that extra information about coordinates that the direct measurement systems have but with any kind of camera.

Human Body Pose Estimation is one of the fundamental tools used in the study of human behaviour, particularly in the areas of recognizing actions and sign languages. When there are a lot of occlusions or severe deformations, using the body features can help improve recognition accuracy (Rastgoo et al., 2021), this technique of human keypoints extraction allowed (Ko et al., 2018) to develop a novel Korean Sign Language recognition system that reached 89.5% accuracy feeding the keypoints to a neural network in 100 complete sentences that can be used in emergencies, another approach using keypoint extraction also was done recently by (Kim and Baek, 2023) where its model executes four principal tasks: extraction of keypoints, keypoint normalization, and skip sampling (SASS) to then achieve the Sign Language Translation, it is important to mention that they removed the lower body keypoints and face keypoints to end with a 55 keypoints extract which is after that normalized before feeding the algorithm. (Khan, 2022) after creating a system that detects finger spelling using MediaPipe Hand solution to extract hand only joints and OpenPose for a system able to recognize complete words extracting the full body keypoints concluded as a future work it might be interesting to try a combination between OpenPose and MediaPipe to improve the results being OpenPose the responsible to detect the body position and MediaPipe Hands module to extract the hand gestures keypoints.

* 1. Challenges

Sign Language Recognition systems faces the same challenges than every computer vision problem. The environment's structure, such as light and movement speed can have an influence with the predictive ability. The change in views causes the gesture to seem different in 2D space (Cheok et al., 2019), high computational cost is also an important point to mention even when Graphic Process Units in the last years have try to address this problem (Thompson et al., 2020), achieving high speed video real-time processing while maintaining accuracy is a difficult challenge which involves optimizing algorithms, hardware acceleration, and efficient memory usage. However, as stated by (Subramanian et al., 2022), finding a prototype that acquires the sign gesture and its corresponding text is the primary difficulty in creating a sign language recognition system and even when there are different techniques available, the challenges of hand tracking, occlusion of hand movements, high computational cost, feature selection and lower learning efficiency still exist.

(Singh, 2022) noted in three points, important facts that make sign language recognition systems challenging to be deployed nowadays which are “Limited number of datasets available”, “Domain restricted data”, “Lack of variety in datasets”, so data is an important challenge on this task, the author points that the few datasets available don’t have a wide variety of signers (10-20 average), lots of them are collected in the same environment, and have limited vocabulary what makes the model after training not being able to generalize.

The lack of standardization and availability of proper linguistic and grammar rules was also highlighted by (Attar et al., 2022) as an important obstacle to the development of effective sign language translation systems. Because of sign languages lack linguistic and grammar rules, researchers have been forced to use data-driven translation approaches in Sign Language Translation systems .

* 1. Conclusion

Machine Learning and deep neural networks are leading a significant improvement in the performance of sign language recognition systems. In this literature review the potential of deep learning to be implemented in an SLR has been explored. The ability of these networks to learn and extract features from large datasets leads to improved classification accuracy. It has been demonstrated that CNNs and RNNs can successfully be applied to the task of sign language recognition in several works before after extracting the human joints present on the image.

Object identification and segmentation techniques are really important for sign language recognition systems, these strategies aid in the identification and location of relevant items in the video stream, such as hands and body parts and further, estimating the signers pose is another crucial component of sign language recognition, it consists on finding and keeping track of the positions of key body landmarks within images or videos of the subject in question. Frameworks such as MediaPipe, OpenPose, and YOLOv8 are examples of some of the most popular options for pose estimation.

Direct measurement methods and vision based approaches are the two primary classifications that can be applied to get information about sign language recognition systems. When it comes to capturing motion data, direct measurement methods involve the utilization of specialized hardware in the form of data gloves or motion capture systems. Vision based approaches rely on computer vision techniques and use RGB images obtained from cameras, this literature review has been helpful in identifying that the project will be done in a vision based method as it is more suitable in a real-time situation due to non-additional equipment being required other than a camera and a device to run the algorithm. When deciding the best approach, time and computational costs must also be considered to fit commodity hardware.

# METHODOLOGY

* 1. Ethical Considerations

It is important nowadays to consider ethics in every Artificial Intelligence project that will be developed as this is evolving rapidly and changing our lifestyles in positive and negative ways, (Green, 2018) made an important statement to take in consideration researching this science area, “AI, as the externalization of human intelligence, offers us in amplified form everything that humanity already is, both good and evil”, also one of the concerns the author points as the most important when dealing in this kind of applications is privacy, and privacy is a point that can be easily addressed on this project even if data is collected, as the final training dataset won’t be images or videos, the final dataset will consist of human pose joints coordinates only which will be just seen as numbers.

Despite the fact that machine translation systems have improved greatly in recent years, they are still sometimes not trustworthy enough for use in areas where lexical and conceptual precision are particularly important, such as in sectors dealing with cultural expression and literature or medical field. It is anticipated that AI-based Machine Translation will be developed principally for the primary world languages, especially English, because these languages have access to the massive datasets necessary for the technology's success. This is harmful to the preservation of different languages, because of this is always important to note that AI interpretations at this stage, can not substitute human interpreters and should be used with caution and in noncritical situations.

Considering these ethical challenges in sign language recognition projects not only safeguards individual privacy and fosters trust in AI technologies but also recognizes the importance of human expertise and cultural diversity. By incorporating ethical considerations into the development process, it is possible to create systems that respect user privacy, promote inclusivity, and enhance communication accessibility for individuals with hearing disabilities, while acknowledging the limitations and potential risks associated with AI-based translations.

* 1. Primary Research Method

The selected primary research method to be employed in this Sign Language Recognition project is experimentation. The experimentation will be focused on exploring three different frameworks for obtaining the pose estimation keypoints which are OpenPose, MediaPipe, and YOLOv8, also, different neural network architectures including CNN, LSTM-GRU, AND GRU-LSTM will be evaluated to compare them and see which one can achieve the highest accuracy in real world settings.

The experimentation will be made in a systematically manner varying the combination of these techniques and architectures taking in consideration their influence on the accuracy of the sign language recognition system. By applying the principle of Concomitant Variation, which involves the manipulation of a variable and observing the resulting changes, the goal is to identify the optimal combination that yields the highest accuracy.

In each step of the experimentation, it is expected that temporal sequence will be present, the temporal sequence in this sign language recognition system will be related to the accuracy, as it will change accordingly, while different techniques of human pose extraction and architectures are tested and refined. The theoretical support will be provided by the literature review that will be the guidance while the experimentation process is conducted.

It is relevant to mention that this experimentation stage on the project won’t be restricted only to the findings mentioned on this literature review. If new relevant materials are presented while this project is worked, this new material may be included in the literature review, and in the experimentation. By doing this, it is ensured that the research remains current by incorporating the latest advancements and insights in the field.

If when the experimentation on this project is finalized, non-spurious association is not achieved, additional research methods may be incorporated, one of the possibilities could be conducting some depth interviews with experts or individuals with expertise in sign language recognition. Doing these interviews could provide insights to understand the factors influencing accuracy, explore further research needed, and identify different strategies or potential considerations that need to be addressed for this project.

* 1. Sampling

The populations of interest in this project are potential users of sign language translation and direct American Sign Language users which means people who interacts with mute, deaf and hard hearing community, and the mute, deaf and hard hearing community who sometimes suffer a gap of communication because there is not many people out of the deaf community that knows how to interpret sign language and an interpreter is not always available.

Collecting data directly from the deaf and hard of hearing community raises ethical considerations and requires careful attention to ensure the protection of their rights and privacy. Given the vulnerable nature of this population, alternative approaches will be adopted to address these concerns. Rather than getting data collected directly from individuals within the community, existing datasets available on the internet can be utilized for training and testing the sign language recognition models. These datasets, obtained from reliable sources, can provide valuable insights and information for developing and evaluating the models without directly involving vulnerable individuals.

For the experimentation phase, instead of sampling, each combination of the identified data acquisition technique (OpenPose, MediaPipe, YOLOv8) and neural network architecture (CNN, LSTM-GRU, GRU-LSTM) will be implemented and tested. This approach allows for a thorough evaluation of each combination performance and accuracy in real world settings. By systematically trying out each combination, the project aims to gather accurate and reliable data on the effectiveness of different techniques and architectures.

When executing the experimentation process, the achieved accuracies in real world settings for each of the combinations will be noted down by doing it in this way it is ensured that the project captures the practical implications and effectiveness of each combination, providing valuable insights for optimizing the sign language recognition models. By focusing on real world performance, the project aims to develop a robust system that can effectively bridge the communication gap and improve the overall accessibility and inclusivity for individuals using sign language.

* 1. Dataset Description

The data that have been used along the project to train the different models and testing is a portion of the just released “ASL Citizen” by (Desai et al., 2023) , it is a conjunction work of Microsoft Research, University of Washington, Boston University, and University of Maryland in an effort to support machine learning models for Isolated Sign Language Recognition (ISLR), it is composed of 83,399 videos for 2,731 different signs performed by 52 signers, 35 signers for training, 6 for validation, 11 for testing. It is important to mention that these videos were collected by people in everyday settings recording themselves with a webcam available, this is a great advantage comparing it to other previous ISLR datasets where the videos are recorded by professionals in laboratories or classrooms with certain characteristics as it gives the dataset a rich variety of conditions that allows the algorithm training to be more robust. The videos in the dataset where already cleaned by the research team removing the ones where there was not a person detected by the YOLOV3 algorithm, and also blurred the background in videos where additional people was also detected to protect their privacy.

The given dataset includes 3 different CSV files, each containing the videos for training, testing, and validation, the features contained are as below:

* Participant ID: Which of the 52 signers is performing.
* Video File: Name of the file.
* Gloss: The actual meaning of the file based on a previous released dataset ASL-LEX (Caselli et al., 2017).
* ASL-LEX Code: Gloss identifier in the ASL-LEX dataset.
  1. Exploratory Data Analysis

Taking in consideration the amount of time and resources available for this project only fifty words of the complete ASL Citizen catalogue are considered, and the decision on words to be used was based on the fifty glosses with more samples on the training dataset which can be found in Table 2.1.

| **Gloss** |
| --- |
| DOG1 |
| HURDLE/TRIP1 |
| BITE1 |
| BREAKFAST1 |
| DEMAND1 |
| DARK1 |
| MECHANIC1 |
| PARTY1 |
| DECIDE1 |
| WHATFOR1 |
| ROCKINGCHAIR1 |
| DEAF1 |
| EDIT1 |
| DEVELOP1 |
| RIVER1 |
| FINE1 |
| ELEVATOR1 |
| BELT1 |
| AXE1 |
| BACKPACK1 |
| SHAVE1 |
| CHRISTMAS1 |
| BEE1 |
| PATIENT2 |
| BASKETBALL1 |
| NOON1 |
| HALLOWEEN1 |
| LUNCH1 |
| EAT1 |
| TWINS1 |
| CANCER1 |
| DINNER1 |
| CONFUSED1 |
| RAZOR2 |
| MEAT1 |
| SQUEEZE |
| MICROSCOPE1 |
| DRAG1 |
| THEY1 |
| SINK |
| MOVIE1 |
| FLOAT1 |
| LOCK1 |
| DOWNSIZE1 |
| GUESS1 |
| KNIGHT1 |
| JEWELRY |
| MAPLE |
| FOREIGNER1 |
| HOSPITAL1 |

Table 2.1 Top 50 Glosses with more videos in training ASL Citizen Dataset

The frequency distribution on the training dataset once this filter was applied ends with 950 videos in total for 50 words, having a minimum of 18 videos and maximum of 24 videos per word falling 75% of them in the range of 18-20 samples.

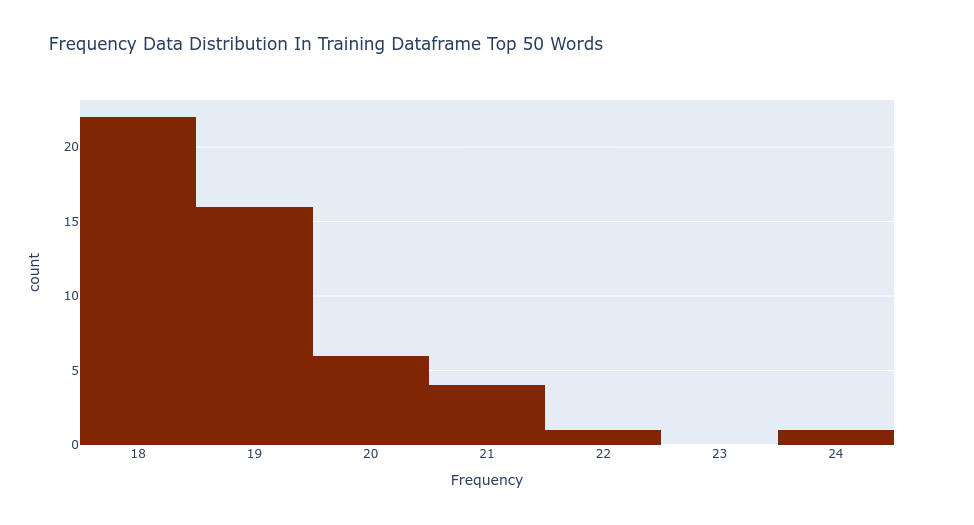


Figure 2.1 Frequency Data Distribution in Training Dataframe Top 50 Words

These same glosses where collected from the validation and test dataset ending with a total of 191 and 729 videos respectively. Taking in consideration that the end goal of this research is to train a neural network and the nature of the human pose estimation techniques used, some extra pieces of information not contained on the dataframes were gathered and analysed using the OpenCV library, such information becomes more important while the process of this work progresses, below is the list of columns added to the dataframes including a short description:

* Frames: Total number of the individual images that compose the video.
* FPS: Rate at which consecutive frames were recorded.
* Length: Total duration in seconds.
* Width: Number of pixels from the left edge to the right edge of the frame.
* Height: Number of pixels from the top edge to the bottom edge of the frame.

After this information was obtained the below statistics describe the combination of training and testing dataframes.

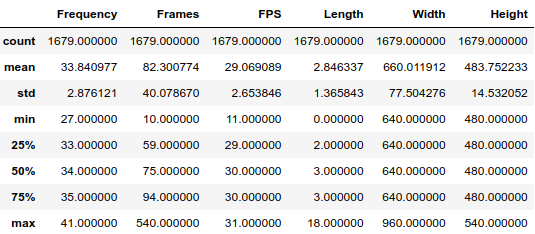


Table 2.2 Descriptive statistics of Training and Testing Datasets combined

From these, it can be seen how the videos were captured with different devices varying the FPS rate, which was expected from the start as Microsoft described the dataset was collected individually by the signers themselves without involving any kind of support other than the seed signer who indicated how the signs should be performed, which is helpful in general for the training, but a few more steps were required to be performed prior to starting collecting the keypoints data as the neural network input layer needs to have a specific input shape and there are videos with different lengths, meaning different quantity of frames.

The idea behind the keypoint collection is to have an array per video containing the keypoints per frame obtained from the Human Pose Estimation algorithm, because of this all the videos were filtered considering the ones lasting more than 1 second, with an FPS rate equal/greater than 29 and containing maximum 130 frames, this decision was made with the idea that the final arrays will be of 130 x number of keypoints, meaning that all the videos lasting less than 130 i.e. 120, the missing 10 keypoint arrays will be padded with zeroes. 130 is the number chosen as this is exact point between the 3rd quantile and the upper fence, ending with videos lasting more than a second and maximum an approximate of 4 seconds which is enough to capture the movement sequences to perform any sign, there are videos like “4256498922014438-SHAVE.mp4” lasting more than that where the signer performs a diversity of signs for the same action which will just provoke the input layer having to be larger, arrays for training bigger on memory containing zeroes only, and confusing the algorithm.

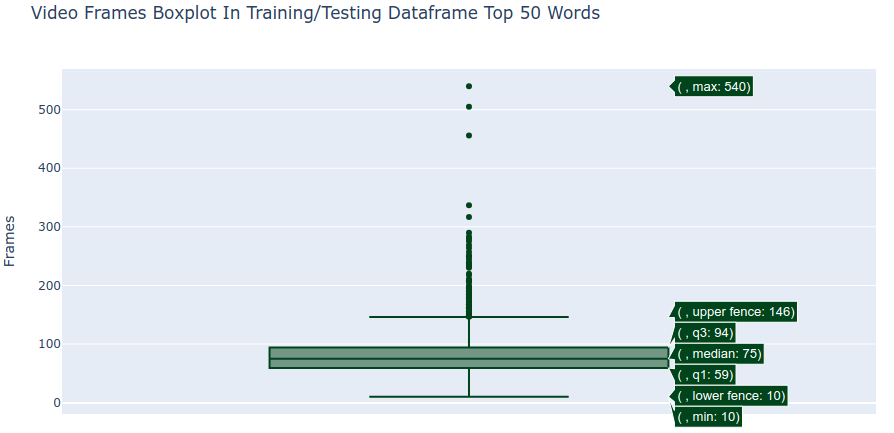


Figure 2.2 Frames per Video in Training and Testing Dataframes Top 50 Words

After the filtering has been applied, for training there were glosses with just 13 videos per gloss which was considered too low to train the network, because of this, the decision to join the training and testing datasets was taken ending with 1,338 videos, having a frequency distribution from 20 to 32 samples per gloss having a minimum of 45 frames and a maximum of 130 frames per observation, for validation the validation dataset was used applying the same rules ending with 161 videos having 2 to 5 samples per gloss.

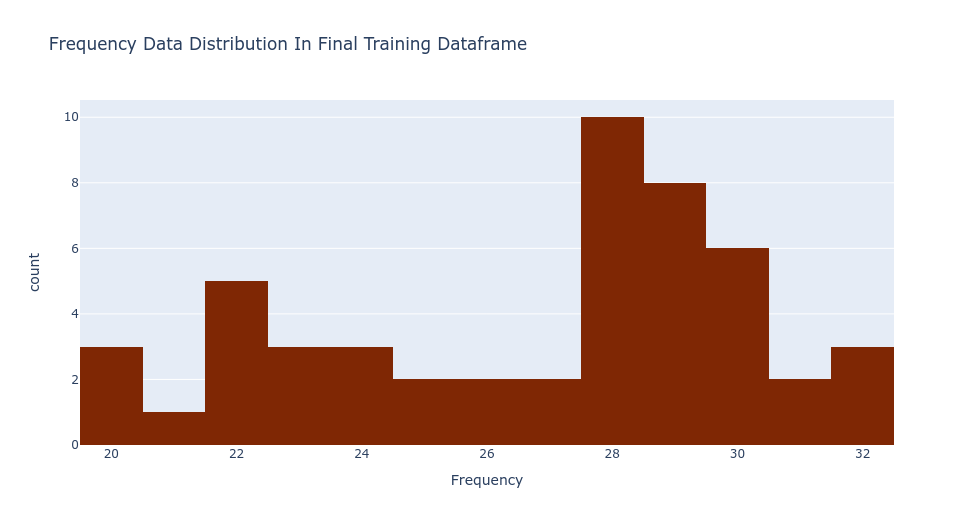


Figure 2.3 Frequency Data Distribution in Final Training Dataframe

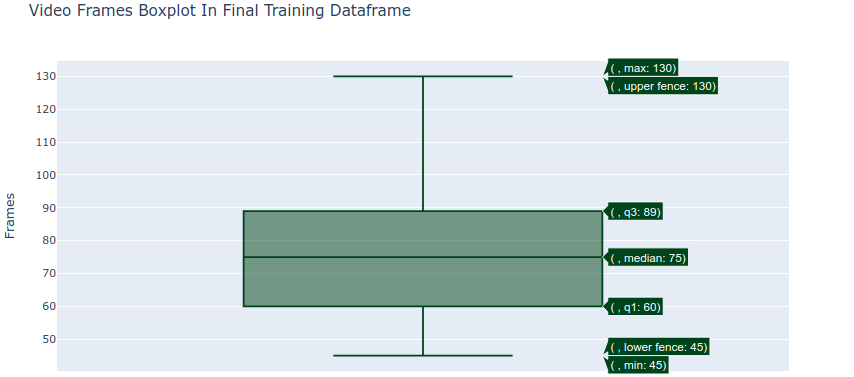


Figure 2.4 Frames per Video in Final Training Dataframe

* 1. Keypoint Collection

For this section, 2 main frameworks were used, Mediapipe and MMPose. Both of them were discussed in section 1.3, but briefly will be explained to have a better understanding.

Mediapipe, developed by Google is an open-source framework that allows to easily implement tasks like object detection, face detection, hand tracking, and pose estimation with pre-built models being the Holistic Model the one used on this project, the hands detection model was also tested, however it was discarded as the model at the moment this project was developed, the Mediapipe-hands instead of having a property identifying the left hand and right hand separately as in the Holistic model, it just creates an array with the keypoints visualized in the image, even though the right/left hands are identified, these keypoints are not always arranged in the same way pointing to additional challenges and noise for the algorithm to learn.

This being said, the Holistic model outputs 543 landmarks in total, 33 pose landmarks, 468 face landmarks, and 21 hand landmarks per hand, giving normalized x,y,z coordinates according to the width and height of the given image. In the project only the 42 hand keypoints are considered, as using the whole given landmarks reaches a total of 1,662 keypoints (543 x 3 coordinates) increasing model complexity and noise induced by the proposed zero padding to standardize the inputs to 130, so 42 hand keypoints are flatten to get a total of 126 points and whenever a hand is not detected, all the points for the hand are automatically filled with zeroes ensuring that each frame has consistently 126 points. A loop running through each of the videos collects an array containing an array of 126 x number of frames per video, however this does not ensure the homogeneous shape required to feed a neural network and as mentioned previously pre-padding was the technique decided to feel the missing frames in videos shorter than 130, pre-padding was chosen over other techniques like post-padding or truncating because as mentioned by (Dwarampudi and Reddy, 2019), truncating drives to information loss and post-padding performs worse than pre-padding on LSTMs due to LSTMs nature to remember, so as in a human brain, it’s easier to remember recent information than older information, ending finally with a 3 dimension numpy array composed of (#videos, 130, 126), this process was also done flipping the frames 180 degrees prior to the keypoint collection using this as a data augmentation process.

REFERENCES.