Real Time Sign Language Recognition

Karthik S K

Mid Thesis Report

JUNE 2022

**Abstract**

Deaf people all over the world use sign languages to communicate visually. The most common ways to sign words and sentences are by waving fingers, arms, hands, and making motions with face. Sign languages are complete languages with vocabulary, grammar which has unique characteristics. Sign language recognition is a complex task. Several factors are employed to identify signs, including hand orientations with many forms, movement of hands, posture of body, and expressions in faces. Even with cutting-edge models, tackling the problem for a sign with larger vocabulary using a computer in real-world circumstances remains a difficulty. A complete language with unique grammar, and other distinguishing features called Indian Sign Language (ISL) is available. About 5 million hearing-impaired community in India use this language. There is presently no dataset is available on ISL in public domain to test Sign Language Recognition techniques (SLR). The Indian Lexicon Sign Language Dataset (INCLUDE) is considered in this study. The dataset has less than half million frames from over 4000 films and over 250 different word signs from more than 12 different word categories. INCLUDE is documented by getting assistance from signers who have expertise resembling natural conditions. Numerous deep neural networks with various methodologies to augment data, to extract features, to encode and decode have been evaluated. Due to the dataset's size and quality, these models can be tested on ISL for Sign Language Recognition. Various deep learning models is presented, and an accurate model is determined. The most accurate model has a 94.5 percent accuracy rate. This model just trains a decoder and employs feature extractor and encoder which is pre-trained. It improves on previous results and provides a quick way to support SLR in various languages.

**Table of Contents**

[LIST OF FIGURES 4](#_Toc107324778)

[LIST OF TABLES 5](#_Toc107324779)

[LIST OF ABBREVIATIONS 5](#_Toc107324780)

[CHAPTER 1 5](#_Toc107324781)

[INTRODUCTION 6](#_Toc107324782)

[1.1 Background 6](#_Toc107324783)

[1.2 Problem Statement 8](#_Toc107324784)

[1.3 Research Questions 9](#_Toc107324785)

[1.4 Aim and Objectives 9](#_Toc107324786)

[1.5 Significance of the Study 9](#_Toc107324787)

[1.6 Scope of the Study 10](#_Toc107324788)

[CHAPTER 2 11](#_Toc107324789)

[LITERATURE REVIEW 11](#_Toc107324790)

[2.1 Introduction 11](#_Toc107324791)

[2.2 Deep Learning Methods and Developments 12](#_Toc107324792)

[2.3 Sign Language Recognition and Feature Fusion 12](#_Toc107324793)

[2.4 Vision-based models 13](#_Toc107324794)

[2.5 Related Research 14](#_Toc107324795)

[2.6 Discussion 18](#_Toc107324796)

[2.7 Summary 19](#_Toc107324797)

[CHAPTER 3 20](#_Toc107324798)

[RESEARCH METHODOLOGY 20](#_Toc107324799)

[3.1 Introduction 20](#_Toc107324800)

[3.2 Research Approach 20](#_Toc107324801)

[3.3 Dataset Description 21](#_Toc107324802)

[3.4 Data Preprocessing 23](#_Toc107324803)

[3.5 Model Building 23](#_Toc107324804)

[3.5.1 Method 1: 3D CNN based Classification with Softmax in the last layer 24](#_Toc107324805)

[3.5.2 Method 2: VGG Feature Extraction + 2D CNN based Classification 25](#_Toc107324806)

[3.5.3 Method 3: Method 1/Method 2 with Attention mechanism 27](#_Toc107324807)

[3.5.4 Method 4: VGG Features + Transformer’s Encoder + GRU based Decoder 28](#_Toc107324808)

[3.5.5 Method 5: I3D Pretrained/Video Vision Transformer + Transformer’s Encoder + GRU based Decoder 28](#_Toc107324809)

[3.6 Evaluation Metrics 29](#_Toc107324810)

[3.7 Summary 30](#_Toc107324811)

[3.8 Required Resources 30](#_Toc107324812)

[3.8.1 Software Requirements 30](#_Toc107324813)

[3.8.2 Hardware Requirements 31](#_Toc107324814)

[References 31](#_Toc107324815)

[APPENDIX A: RESEARCH PLAN 32](#_Toc107324816)

[APPENDIX B: RESEARCH PROPOSAL 32](#_Toc107324817)

# LIST OF FIGURES

Figure 1: The training process for two stream I3D on Kinetics dataset 22

Figure 2: CNN architecture with Softmax output 23

Figure 3: VGG-16 Architecture 24

Figure 4: Confusion Matrix 28

# LIST OF TABLES

Table 1: INCLUDE: Category Size 20

Table 2: INCLUDE: Key statistics 21

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| SL | Sign Language |
| SLR | Sign Language Recognition |
| SLT | Sign Language Translation |
| CSLR | Continuous Sign Language Recognition |
| ISL | Indian Sign Language |
| EDA | Exploratory Data Analysis |
| CNN | Convolutional Neural Network |
| VGG | Visual Geometry Group |
| NMT | Neural Machine Translation |
| I3D | Inflated 3D |
| HMM | Hidden Markov Model |
| RNN | Recurrent Neural Network |
| LSTM | Long Short-Term Memory |
| WASL | Word-level American Sign Language |
| AUTSL | Ankara University Turkish Sign Language |

# CHAPTER 1

# INTRODUCTION

# 1.1 Background

Video and image creation and consumption are continually increasing. As a result, deep learning is increasingly being employed for classifying videos, detecting, and tracking objects, recognizing actions, answering questions visually, and encoding videos. These jobs have real-world applications, with significant accessibility implications. SLR (sign language recognition) is a form of action recognition problem (Sridhar et al., 2020). Sign languages are way of communicating techniques mostly where hearing impaired community uses throughout the world. The most frequent technique of signing words and sentences is by waving fingers, arms, hands, and making motions with face. Sign languages are fully formed languages with their own syntax and lexicon (Sridhar et al., 2020). Furthermore, they differ from one place to another and are often incomprehensible to one another, despite some commonality. It differs from languages spoken in a region with lexicon and articulation pace. The mode of communication in hearing-impaired community being sign languages have language frameworks which are unique. The purpose of sign language interpretation systems is to use techniques like vision to automatically translate sign languages. The two main goals in this technique are word-level sign language recognition (Sridhar et al., 2020) (also known as "isolated sign language recognition") and sentence-level sign language recognition (also known as "continuous sign language recognition").

There are various reasons as to why building models with computers for Sign language recognition can be difficult. Analysing numerous body components, such as the hand, arms, and face, is required. Although the hand motions for some pairs of signs appear to be extremely identical, they can be distinct in face expressions. Similar hand gestures when repeated can appear distinct. Another problem are the differences in how different signers perform a sign, such as body and position variances, duration variations of different parts of the signals, and so on. Variation in illumination and backdrop complicates the challenge, which is a problem in computer vision. When the number of signs in the corpus grows, these difficulties become more difficult to solve (Sincan and Keles, 2020)

Indian Sign Language (ISL) has its own linguistic characteristics (Sridhar et al., 2020). ISL differs from its western equivalents in several ways, the most notable of which being its lexicon. While other sign languages have few compound signs as features (signs that are made up of two or more signs), ISL has a widespread compounding system. Compounding signals for Male and Sibling, for example, are used to sign the word Brother, whereas compound signs for Female and Marry are used to sign the term Wife. This is also a deviation from Indian spoken languages, which are noted for their large number of intricate kinship phrases (Sridhar et al., 2020). Many possible compositions allow a wide lexicon on ISL. Another way in which ISL is different is in the amount of room required for signing. In ISL, signing from the top portion of the body represents distance and authority, whereas in many other Sign Languages, distance is indicated by a plane horizontal to front of the signer. These and other differences between ISL and other sign languages need the development and testing of novel methods for ISL Sign Language Recognition. Furthermore, India's vast population of deaf individuals – more than 5 million – emphasizes the significance of any solution that can be deployed in practice (Sridhar et al., 2020).

The resource availability such as big, standardized data is critical for the development and evaluation of models. Importantly, these datasets are not in the public domain, preventing additional research. These datasets also have two significant flaws. To begin with, the number of classes, or different word signs, is extremely minimal. The videos are frequently limited in some way. Only the hands are visible in the datasets because the photographs were cropped. Pictures or films with consistent backdrops, with the signer's clothing generally matching the background colour. The usefulness of training models on real time datasets is limited by these dataset restrictions. As a result, no ISL dataset of sufficient quality and size for further study in machine learning (Sridhar et al., 2020) is currently available.

The first publicly available ISL dataset, (Sridhar et al., 2020) Indian Lexicon Sign Language Dataset (INCLUDE), has been proposed to look at the absence of datasets in public domain and algorithms that can scale for recognizing ISL. Data augmentation, using networks already trained for extracting relevant features, and encoding and decoding were all part of the deep learning workflows for SLR on this dataset. The feature extraction was merged with already trained network for detecting important poses, and the encoding was done with already trained network, which is MobileNet, and the decoding was done with trained LSTMs which are bidirectional.

# 1.2 Problem Statement

A defined collection of hand gestures with specific meanings used by hearing-impaired people to communicate in daily life is referred to as a sign language. They communicate through body, face, and hand movements since they are visual languages. Around the world, there are more than 300 different sign languages. Even though there are numerous distinct sign languages, only a small portion of the community is conversant in any of them, making it challenging for people with special needs to freely interact with the public. SLR gives people a way to communicate in sign language even if they don't know it. It interprets a gesture into a language that is widely spoken, like English, after recognizing it.

SLR is a broad area of study, and while much work has been done in this area, there are still many issues that need to be resolved. The use of machine learning techniques enables computerized systems to make decisions based on data and experience. Two datasets are required for the classification algorithms: a training dataset and a testing dataset. The classifier learns from the training set's experiences, and the model is evaluated using the testing set. Numerous writers have created effective techniques for collecting data and classifying it. Previous research can be divided into two categories based on the data collecting method: direct measurement methods and vision-based approaches. The direct measurement techniques rely on sensors, motion-tracking devices, or gloves that collect motion data. The motion data retrieved enables precise tracking of hands, fingers, and other body parts, which supports the development of SLR techniques. The extraction of discriminative spatial and temporal information from RGB photographs is a key component of the vision based SLR techniques.

A variety of processing techniques have been employed to develop an SLR system. In SLR, Hidden Markov Models (HMM) are frequently employed. Multi Stream HMM (MSHMM), which is based on the two-common single-stream HMMs, Light-HMM and Tied-Mixture Density HMM, is one of the HMMs that have been employed. Neural networks, ANNs, Naive Bayes Classifiers (NBC), Multilayer Perceptron (MLP), unsupervised neural networks, Self-Organizing Maps (SOM), Self-Organizing Feature Maps (SOFM), Simple Recurrent Networks (SRN), Support Vector Machines (SVM), and 3D convolutional residual networks are some of the additional processing models that have been used. Additionally, researchers have developed their own techniques including the Eigen Value Euclidean Distance and the wavelet-based technique.

# 1.3 Research Questions

Below are the research questions which we will consider as part of the research study in Sign Language Recognition:

* Will try to evaluate if our model can be implemented for real time inference.
* Will try to check if it can recognize on videos outside the dataset, but similar domains.

# 1.4 Aim and Objectives

The main aim of this research is to propose a model that will perform Sign Language Recognition. Given a video with a person performing Sign Language for words, our model will recognize the video and will classify that video into one of the elements in vocab space.

The research objectives are formulated based on the aim of this study which is as follows:

* To Build a training pipeline that leads to a robust sign language recognition model.
* To build the system in real time
* To make an UI for the Sign Language Recognition Model

# 1.5 Significance of the Study

Sign languages are visual languages that use hand, face, and body motions to communicate. Learning a second language is not only beneficial to your brain, but it can also help you enhance your communication abilities. Early man utilized signs to communicate, according to some scholars, long before spoken language was established. Even though today's world progressed significantly, sign languages in their most basic form are still used such as pushing the index finger between lips to calm a noisy child, raising a hand to hail a cab, or pointing to an item on the menu.

Sign language is the link that connects us to the world of persons who are deaf or have difficulty speaking. A variety of hand, finger, arm, head, and face expressions are used to help the deaf with those around them and vice versa. It enables individuals to comprehend the world around them through visual descriptions and hence contribute to society. Some of the benefits of sign language include - assists persons who are deaf or hard of hearing aiding social inclusions. It provides deaf children an opportunity to educate themselves. Non-deaf volunteers who volunteer to learn sign language to communicate with the disabled are instilled with a sense of social responsibility as well as sensitivity.

# 1.6 Scope of the Study

The scope of the study has been outlined below:

* The scope of this project is limited to the INCLUDE dataset.
* We are not exploring Multi-Modal Sign Language and Very Deep recent Transformer Architectures due to lack of time and resource constraint.
* The study will also be limited to the evaluation of Accuracy as a metric.

# CHAPTER 2

# LITERATURE REVIEW

# 2.1 Introduction

Over 300 sign languages are used by 70 million deaf individuals worldwide, as per World Federation of the Deaf. Recognition of sign languages would aid in lowering social obstacles for sign language users. The vast majority of communication technology support either written or spoken language. Although tools and technologies for communication, such WhatsApp and Imo, have become integral parts of our lives, deaf individuals still face numerous challenges when utilizing these applications. These technologies can make it easier for the deaf community to communicate with the hearing majority community daily. As a result, the speech-impaired community, and deaf uses sign language as hand gestures in structural forms incorporating signs and visual motions to help with interaction daily (Rastgoo et al., 2021).

Palm orientation, handshape, movement, expression/non-manual messages, and placement are the five main components of sign language. It is important to carry out these characteristics appropriately for a reliable sign word. The advantages of sign language recognition are used in a wide range of applications, interpreting services, including translation tools, video remote human interpretation, multi person real time recognition systems, human hand monitoring of communication online for desktop settings, virtual reality settings, robotic controls, games, as well as everyday language interactions. Additionally, Red Green Blue Depth (RGBD) cameras, which have a large capacity to make depth maps, have become more affordable in recent years, making them a popular choice for lowering the price of hand pose recognition systems (Rastgoo et al., 2021).

In addition, most significant technological firms, have contributed to several initiatives including mixed reality (MR), virtual reality (VR), and augmented reality (AR) as well as new interactive computers. The applications of pose estimation and hand sign have been greatly expanded by this trend. Few research projects covering interactions between in Human and Computer (HCI) that regulate hand movement have been proposed. Therefore, the creation of hand sign language translation system that is automatic is essential to meet the demands of numerous applications, create equal opportunities for communication, and advance the welfare of the general population (Rastgoo et al., 2021).

The development of a sign language recognition system sentence translation or words into voice and text for allowing communication of deaf majority and hearing people is one of the fundamental problems. A system needs to be developed which can enable real conversation or chat between hearing people and deaf. These systems should also account for the problem of splitting videos, including sentences or sign words into separate words (Rastgoo et al., 2021).

# 2.2 Deep Learning Methods and Developments

Deep learning techniques have recently outpaced prior cutting-edge machine learning methods in a variety of applications, particularly in the field of Computer Vision and Natural Language Processing. To solve problems related to computer vision, deep learning models used that are significant are related to CNN, Auto Encoder, RNN which includes LSTM and GRU, GAN. The different models with many computations have enabled Deep learning to process various layers to learn from input and represent it at different levels of abstraction (Rastgoo et al., 2021). The initial attempt at simulating the human brain was made in 1943 by McCulloch and Pitts, who sought to comprehend the patterns which the brain can create utilizing simple cells that can interconnect, known as neurons.

Several feature learning algorithms be it supervised or unsupervised, probabilistic models, Neural networks are just a few of the many techniques that make up Deep Learning. The development of extensive, top-notch, and publicly accessible labelled datasets together with the capabilities of parallel GPU processing are two significant elements that have greatly contributed for the development of deep learning. The reduction of the gradient, which is vanishing, the development of powerful frameworks like MXNET, Theano, TensorFlow, and, as well as the proposal of techniques for regularization (augmenting data, normalizing of a batch, dropout), all played an important role for advancement of deep learning.

# 2.3 Sign Language Recognition and Feature Fusion

Different elements can be combined to increase the accuracy of sign language recognition. These features are grouped into three categories: utilizing only hand posture, utilizing both face and hand, and utilizing all of them. The following subsections provide more explanations.

With the development of reliable depth sensors, using hand posture features has become more important in recent past. Only hand features are used to recognize sign language in this category (Rastgoo et al., 2021). Following hand identification from input data, different deep learning architectures such as CNN, RNN, GAN, and others are used to extract hand features. In this area, having an effective hand detection model that can extract features is difficult. While CNN excels at dealing with still photos, it falls short when it comes to covering sequence information. To profit from the power of these models in extracting sequence information from visuals, CNN is paired with another deep learning model, such as RNN, LSTM, or GRU.

Combining hand features with face feature, sign language recognition accuracy improvement is possible. This is because the pose related to face has some important information on the grammar. There are difficulties to track faces of human from films, such as side movements and tilting head, only a few models have been presented in this category (Rastgoo et al., 2021). Recognizing the facial features with settings that are natural has become more important and hence it is critical for an accurate model for sign language recognition to be able to apply for real scenarios.

To further improve the accuracy, hand and face features are fused with other body parts. This helps to build models which are robust to variations in appearance, deformations, and occlusions. This feature fusion with body parts improve accuracy for situations where there are complex occlusions in face and hand.

# 2.4 Vision-based models

The input modalities for data that is visual in Sign Language Recognition have emerged in the recent past. Most frequent ones include are RGB and depth. RGB images/videos have content which are of high resolution, and depth inputs having precise information about the image object, it’s plane and the distance between them.

Another approach that is frequently employed by research community for combining features based on CNN is flow information, which is defined as the motion properties of every pixel in a sequence of video frames (Rastgoo et al., 2021). Optical Flow (OF) and Scene Flow (SF) are two types of flow information that have been employed in various models. There are various technologies for recording input data modalities, with cameras being the most prevalent. Cameras accept data in a variety of formats and quality levels. Microsoft Kinect has been very popular in producing videos with good quality. Flex sensors in the gloves can collect data on the movement of palms and fingers.

Another device for detecting and tracking hands is the Leap Motion Controller (LMC) system. The idea is that we define the targeted applications, as the device that is most suited to the application is highly dependent on it. To extract the necessary features, sign language recognition models use two types of input data: static and dynamic (Rastgoo et al., 2021). In recent years, deep learning models with inputs that are sequence based have been proposed. Dynamic inputs contain sequential information which helps increasing accuracy of sign language recognition. However certain obstacles such as cost of computation remains as a challenge.

# 2.5 Related Research

The research survey related to solving the problem of Sign Language Recognition (SLR) was carried out and is summarized in this section. The community having problems in hearing use sign languages for communication and hence it is meaningful and important for automatic translation of a sign language. The problems that are prevalent in the domain of SLR are mainly two - isolated SLR and continuous SLR (Huang et al., 2018). Word by word recognition is isolated SLR, whereas translating entire sentences is continuous SLR.

The current methods for continuous SLR while using isolated SLR as building blocks also add layers of pre-processing and post-processing which indicate temporal segmentation and sentence synthesis respectively. Most existing SLRs fall into the category of isolated SLR which deals with the recognition of words or expressions. Continuous SLR being more challenging involves reconstructing sentence structures, which divide the problem of recognizing sentences into three stages which are segmentation of videos with time, recognizing isolated word/expression, and sentence synthesis with a language model.

Most approaches in the recent past show great progress in SLR which are based on the Hidden Markov Model (HMM) with various features such as motion trajectory. The transitional movements between signs which are not correctly captured are addressed using the approaches of offline training and online recognition. The threshold matrix and rate thresholds are proposed in offline training, where each element of the matrix indicates the minimum probability when a segment belongs to a sign, and rate thresholds are defined as average probability for signs. If the evaluation of a certain segment is smaller than all the thresholds, then it is regarded as a transitional movement which should be removed. In the online recognition stage, Dynamic Time Warping (DTW) and Length-Root method is used to record the time intervals for fine segmentation and the endpoint for each candidate sign is determined (Zhang et al., 2014).

The temporal segmentation even though popular is difficult to implement due to below reasons:

● The transitional movement between hand gestures can be subtle and ambiguous.

● Inaccurate segmentation can incur significant performance penalty on subsequent steps.

● The isolated SLR step that deals with the recognition of words or expression requires per-video-frame labels which are extremely time consuming.

An end-to-end novel sequence-to-sequence model that can generate video captions is a relevant research area. The state-of-the-art performance demonstrated by Recurrent Neural Networks, specifically LSTMs in generating image captions, is trained on pairs of video-sentence and the model associates a sequence of video frames to sequence of words to describe an event in the video clip (Venugopalan et al., 2015). Attention mechanisms can also be incorporated into LSTM for automatic selection of most likely video frames (Yao et al., 2015).

Industry labs and research groups have created solutions which are open source for Neural Machine Translation (NMT). The solutions are based on platforms related to deep learning. However, these tools are targeted only towards research groups having a good understanding of deep learning architecture, and those who know how to handle code bases that are large.

Joey NMT (Kreutzer et al., 2019) designed a model which provides minimum code platform with quality that can be compared to more standard benchmarks with complex code bases. It includes standard network architectures (RNN, transformer, different attention mechanisms, input feeding, configurable encoder/decoder bridge), standard learning techniques (dropout, learning rate scheduling, weight tying, early stopping criteria), and visualization/monitoring tools.

RWTH-PHOENIX-Weather 2012 corpus for German sign language (DGS) recorded sign language footage in 2009 and 2010 by “PHOENIX”, a German public TV station. This corpus added more data with bilingual annotation in DGS glosses and written German, creating the RWTH-PHOENIX-Weather 2014 corpus. There are also annotations for spatial positions of face and hands of a signer for more than 40,000 video frames along with annotating hand shapes and on the frame level orientations (Schmidt et al., 2014).

The methods that were employed previously for continuous sign language recognition involved hidden Markov models (HMM) with limited capacity for capturing temporal information. Using the framework of deep networks for continuous SLR, transcribes videos of sign language sentences to sequences of ordered gloss labels. Deep CNN stacked with temporal fusion layers for extracting features, and bidirectional Recurrent Neural Networks (RNN) for sequence learning modules were proposed with an iterative optimization process to exploit the representational capability of deep neural networks with limited data. The end-to-end recognition model for alignment proposal was trained first, and then the proposal was used to tune the feature extraction module. This process can run iteratively to achieve improvements on the recognition performance (Cui et al., 2019).

The prior studies in Sign Language Translation (SLT) have shown that effectively recognizing individual signs improves the performance of translation to a great extent. A novel transformer architecture jointly learns Continuous Sign Language Recognition (CSLR) and Translation while being trainable in an end-to-end manner. The recognition and translation problems are addressed in a single unified architecture called Connectionist Temporal Classification (CTC) loss (Camgoz et al., 2020). The contributions can be summarized as:

● The novel formalization of Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT) which is multi-task uses the power of glosses without limiting the translation to spoken language.

● The state-of-the-art results achieved in successful application of transformers for Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT), outperforming all comparable previous approaches.

● Guiding future research in this field with baseline results which are new.

Indian Sign Language (ISL) which is in use by over 5 million hearing impaired community in India is a complete sign language with unique linguistic attributes with one grammar, syntax, and vocabulary. To carry out and assess Sign Language Recognition (SLR) methods, no dataset on ISL is available on public domain. Previously, ISL include video and image datasets with only one image indicating a sign. Since these datasets are not in the public domain, any further study is a challenge. Other limitations are with the word signs (classes) that are distinct being

very low, and reduced images with certain portions which are visible in the datasets. These constraints on the dataset pose challenges in limiting the models that can be built and trained for real world scenarios. The quality and size required for Machine Learning (ML) study are not available for ISL dataset (Sridhar et al., 2020).

The problems with datasets not in the public domain and models that cannot scale for ISL recognition has been addressed by proposing the first available ISL dataset in public domain called Indian Lexicon Sign Language Dataset (INCLUDE). It has less than half a million frames with over 250 classes from more than 12 different word categories. There are videos documented with the help of hearing-impaired community, with the resolution, lighting and background chosen to resemble real world scenarios. A subset that is small named INCLUDE-50 with 50 classes was also proposed for rapid evaluation. Numerous models for SLR on INCLUDE dataset included data augmentation, feature extraction using trained networks, with encoding and decoding. The feature extraction combined with trained pose network for detecting poses, encoding with MobileNet network that is trained, and decoding trained LSTMs which are bidirectional resulted with an accuracy of 94.5% on INCLUDE-50 and 85.6% on INCLUDE (Sridhar et al., 2020).

Deaf persons can use vision-based sign language recognition to communicate with others. Most available sign language databases, on the other hand, are confined to a small number of words. Models learned from those datasets cannot be used due to the low vocabulary size. Word-Level American Sign Language (WLASL) video dataset with more than 2000 words performed by more than 100 signers was proposed (Li et al., 2019). This dataset is handy to the researchers on the public domain. It is the largest publicly available ASL dataset for word-level sign recognition research. Various deep learning approaches for word-level sign identification evaluated the performance in scenarios which are large scale. Two different models were used and compared: (i) a holistic approach which is based on visual appearance and (ii) An approach that is based on human pose which is 2D. Both provide baselines which are valuable for method benchmarking in the community. Furthermore, the study offers a unique temporal graph convolutional networks (Posture-TGCN) which is pose based jointly to model spatial and temporal relationships in trajectories of human pose. Findings reveal that pose-based and appearance-based models perform similarly on 2,000 words/glosses, with top-10 accuracy up to 62.63 percent, showing the dataset's validity and difficulties.

Sign language identification is a difficult challenge in which number of sources are used to identify signs, such as shape and orientation of hand, movement of body, position, hand, and expressions of face. Even with state-of-the-art models, what remains a challenge is to solves these problems using a computer. A new large-scale multi-modal Turkish Sign Language dataset (AUTSL) with a benchmark and performance evaluation base models was proposed in this paper (Sincan and Keles, 2020). In total, there are over 38000 video samples in our collection, which includes over 200 signs by more than 40 individual signers. The samples include a wide range of backgrounds captured in both indoor and outdoor settings. Furthermore, throughout the recordings, signers' spatial placements and postures change. Each sample was captured using the Microsoft Kinect v2 and includes RGB colour, depth, and skeletal modalities. For user independent model evaluations, the benchmark training and test sets were created. Employed Convolutional Neural Networks (CNNs) to extract features and unidirectional and bidirectional Long Short-Term Memory (LSTM) models to characterize temporal information, and provided empirical evaluations using the benchmark. To boost the performance of our models, added feature pooling modules and temporal attention. On the AUTSL and Montalbano datasets, tested the baseline models.

On the Montalbano dataset, the models achieved competitive results with state-of-the-art approaches, achieving 96.11 percent accuracy. The models achieved 95.95 percent accuracy in AUTSL random train-test splits. The best baseline model achieved 62.02 percent accuracy in the proposed user-independent benchmark dataset. The disparities in the results of the identical baseline models demonstrate the difficulties in the benchmark dataset.

# 2.6 Discussion

For detecting hand from various sources of input, such as skeleton, images, flow features, videos and so on, most deep learning-based models in the past employed a CNN or a mix of a CNN with another approach. Even though an extensive research has been made in hand detection recently and several methods are suggested, still are many problems to overcome. Even accurate key point annotations are difficult to make manually due to high occlusions in hand key points. Because there are so many different hand forms and movements, even though detecting hand is mostly initial step in various tasks like action identification and sign language recognition, it is an extremely challenging assignment. Some of the most significant issues in this field include limited resolution, fluctuating lighting conditions, heavy occlusion, complicated interactions (Rastgoo et al., 2021). A hand's ability to hold things, at various scales as open or closed palms, with finger articulations, or holding other hands is another crucial consideration. We believe that using features from parts, like the face and body, combining different input, like flow information, image, skeleton, text, etc., and utilizing additional deep learning models combined with conventional methods could all help increasing the accuracy of detecting. Additionally, utilizing the human pose and face datasets, as well as fusing information from many domains with the datasets on hand detection, could be taken into consideration for further enhancement. There are many amazing models for detecting faces and people and combining their features with hand features may increase the detection's precision. The real hand detection needed in practical applications can be provided by utilizing new hardware breakthroughs and capabilities along with effective model implementations.

# 2.7 Summary

The development on models based on vision and deep learning approaches has been looked at for sign language recognition from the past. The succinct summary for sign language recognition models that are based on vision which correlate to obtained results. This also provides direction for other researchers on the advancements happened recently in the field of sign language recognition. Providing accurate models for sign language recognition involves lot of effort. Various visual cues which lead to multiple modalities needs to be integrated practically. New datasets, high computing power, multiple modalities and approaches, latest advancements in deep learning, all these put together can be used to address challenges.

# CHAPTER 3

# RESEARCH METHODOLOGY

# 3.1 Introduction

Sign language is a system of communication using visual gestures and signs as used by the deaf community and is the main mode of communication. Deaf and people who have problems hearing use sign language between their own community and other people. Computer recognition of sign languages covers sign gesture recognition and continues till text/speech generation. Sign gestures can be classified as static and dynamic with static gesture recognition being simpler than dynamic gesture recognition. While most of the approaches in recent times show progress in Sign Language Recognition (SLR) which are based on Hidden Markov Model (HMM), temporal segmentation techniques are also used to record time intervals for fine segmentation and the endpoint for each candidate sign is determined. Generating captions for videos using novel end-to-end sequence-to-sequence models is a relevant research area. Attention mechanisms are used in Recurrent Neural Networks (RNN), specifically LSTMs. In this research paper, as part of research methodology the following sections are covered which includes target dataset description, data pre-processing, different models and algorithms and evaluation metrics.

# 3.2 Research Approach

The primary means of communication for the deaf people is sign language, which employs visual cues and movements. Sign language is used by the deaf and hearing-impaired to communicate with others outside of their own community. Computer recognition of sign languages starts with the identification of sign gestures and continues through text and speech creation. Both static and dynamic sign motions can be recognized, with static gesture identification being the easier of the two. While Hidden Markov Model (HMM)-based approaches make up most recent advances in Sign Language Recognition (SLR), temporal segmentation techniques are also utilized to record time intervals for fine segmentation and to identify the endpoint for each potential sign. A pertinent study field is the creation of subtitles for films utilizing unique end-to-end sequence-based models. Recurrent Neural Networks (RNN), more notably LSTMs, employ attention processes. The target dataset description, data pre-processing, several models and algorithms, and assessment metrics are presented in this research article as part of the research methodology.

# 3.3 Dataset Description

Indian Lexicon Sign Language Dataset - INCLUDE - an ISL (Indian Sign Language) dataset contains 0.27 million frames across 4287 videos over 263-word signs from 15 different word categories. This dataset is recorded with the help of experienced signers. Each video is a recording of 1 ISL sign, signed by deaf students from St. Louis School for the Deaf, Adyar, Chennai (Sridhar et al., 2020).

Based on the 2 main principles - The videos resembling real life scenarios, and a dataset covering a diverse set of signs with multiple videos, the dataset was created by recording with the help 7 experienced signers from the school for the deaf. Each class is a sign in the dataset, which is signed by multiple signers, containing 2 to 6 videos per class.

The classes belong to 15 broad word categories which are popular words in ISL. This is shown in Table 1.1. The category-wise summary statistics are shown in Table 1.2.

Table 1: INCLUDE: Category Size

|  |  |  |
| --- | --- | --- |
| **Category** | **Number of Classes** | **Number of Videos** |
| Adjectives | 59 | 791 |
| Animals | 8 | 166 |
| Clothes | 10 | 198 |
| Colors | 11 | 222 |
| Days and Time | 22 | 306 |
| Electronics | 10 | 140 |
| Greetings | 9 | 185 |
| Means of Transport | 9 | 186 |
| Objects at Home | 27 | 379 |
| Occupations | 16 | 225 |
| People | 26 | 513 |
| Places | 19 | 399 |
| Pronouns | 8 | 168 |
| Seasons | 6 | 85 |
| Society | 23 | 324 |
| **Total** | **263** | **4287** |

Table 2: INCLUDE: Key statistics

|  |  |
| --- | --- |
| **Characteristic** | **INCLUDE** |
| Categories | 15 |
| Words | 263 |
| Videos | 4287 |
| Avg. Videos per Class | 16.3 |
| Avg. Video Length | 2.57s |
| Min. Video Length | 1.28s |
| Max. Video Length | 6.16s |
| Frame Rate | 25fps |
| Resolution | 1920x1080 |

# 3.4 Data Preprocessing

The preprocessing techniques on video data include - VGG (Visual Geometry Group) Feature Extraction, and I3D (Inflated 3D ConvNet) Feature Extraction & Vision Transformer. In this section each of the above feature extraction techniques are analyzed.

The temporal element is one of the key differences between information in a single image and information in a video. As a result, deep learning model architectures have been improved to include 3D processing to process temporal data as well. Through the I3D model, this article summarizes the architectural changes from photos to video. Researchers from DeepMind and the University of Oxford presented the I3D model in a paper (Carreira and Zisserman, 2017). The research examines prior approaches to the problem of video action detection while also offering a novel architecture, which is the emphasis of this paper. Their method starts with a two-dimensional design and then inflates all the filters and pooling kernels. They offer an additional dimension to be considered by inflating them, which in our instance is time. Filters in 2D models are square N x N, but when they are inflated, they become cubic N x N x N.

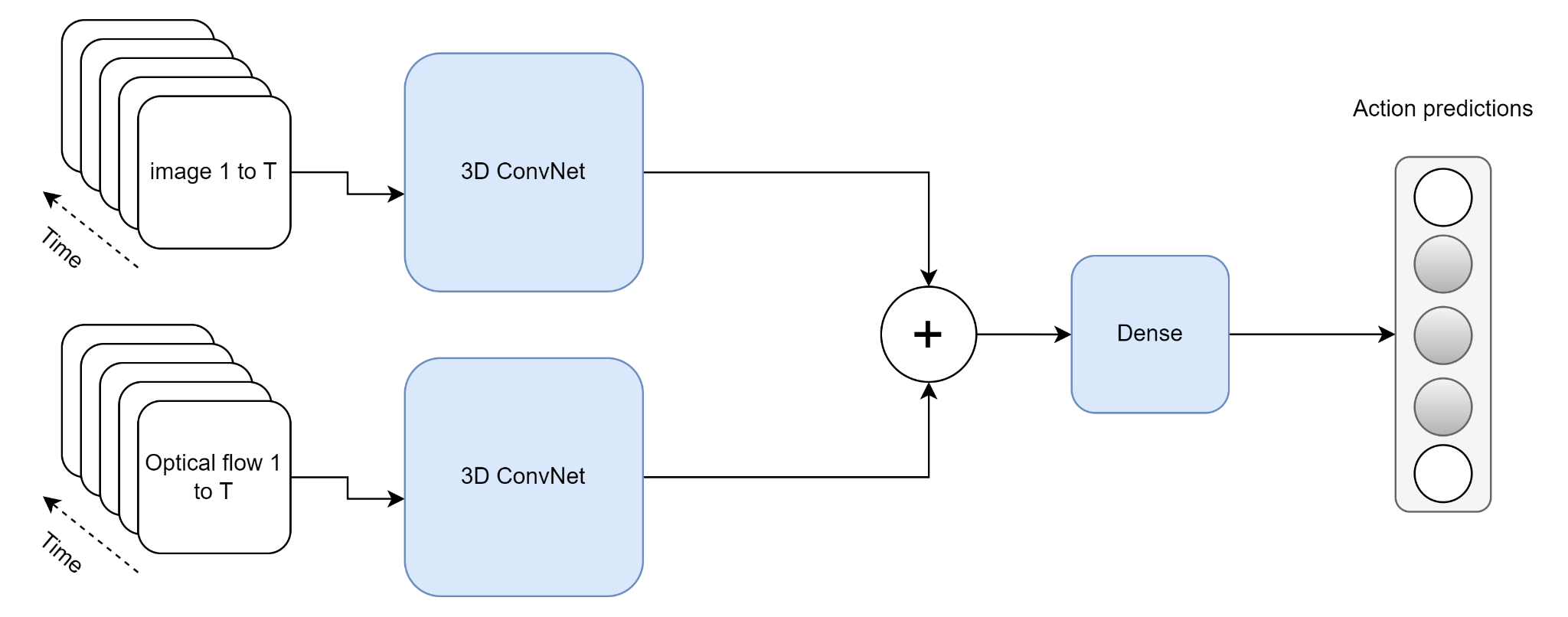


Figure 1: The training process for two stream I3D on Kinetics dataset

Since the dataset contains videos, some of the pre-processing techniques such as imputing missing values, outlier treatment is not applicable, and hence these are not mentioned in this section.

# 3.5 Model Building

In this section different models and algorithms that will be used for effective Sign Language Recognition are explained. The following methods are used:

* Method 1: 3D CNN based Classification with Softmax in the last layer
* Method 2: VGG Feature Extraction + 2D CNN/NN based Classification
* Method 3: Method 1/Method 2 with or without Attention Mechanism
* Method 4: VGG Features + Transformers Encoder + GRU based Decoder
* Method 5: I3D Pretrained/Video Vision Transformer + Transformers Encoder + GRU based Decoder

# 3.5.1 Method 1: 3D CNN based Classification with Softmax in the last layer

Convolutional Neural Networks (CNN) is used in a variety of applications. It is, without a doubt, the most widely used deep learning architecture. The enormous popularity and effectiveness of convnets has sparked a recent rise in interest in deep learning. AlexNet sparked interest in CNN in 2012, and it has grown rapidly since then. Researchers went from an 8-layer AlexNet to a 152-layer ResNet in just three years. CNN has become the go-to model for any image-related issue. They outperform the competitors in terms of accuracy. The fundamental advantage of CNN over its predecessors is that it discovers essential traits without the need for human intervention. In addition, CNN is computationally efficient. It performs parameter sharing and uses special convolution and pooling algorithms. CNN models can run on any device, making them globally appealing (Gu et al., 2015)

Fig 2 shows the architecture of CNN with Softmax output in the last layer.

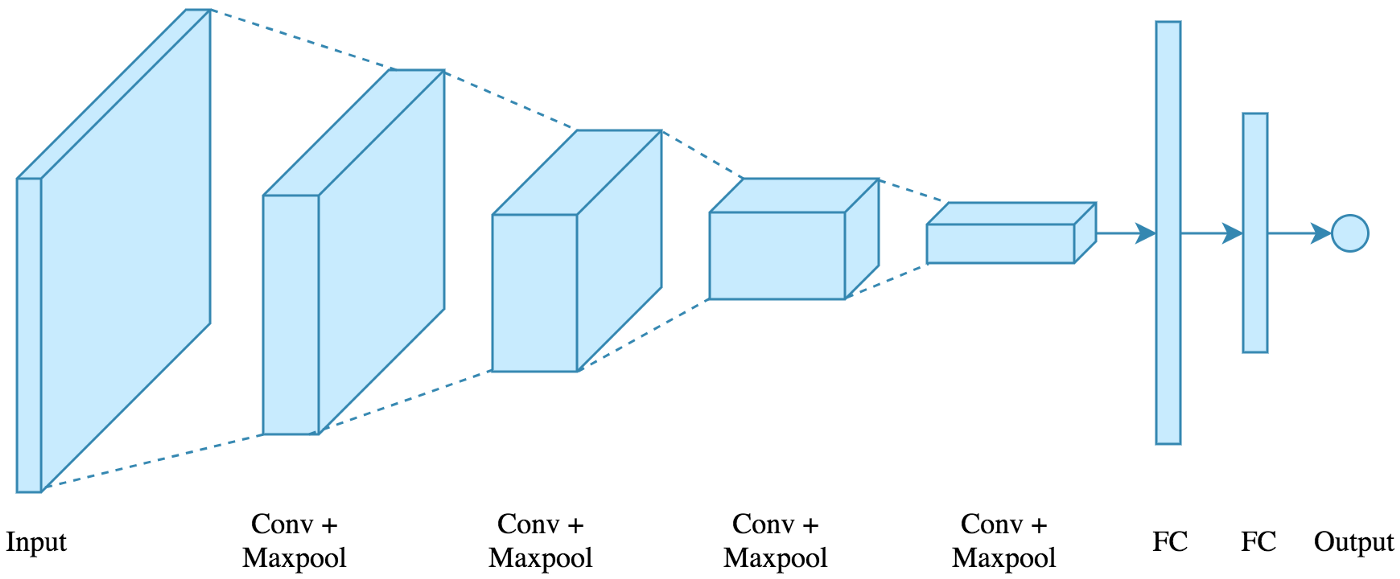


Figure 2: CNN architecture with Softmax output

The following architecture will be used: 4 convolutions + pooling layers will be followed by 2 fully connected layers. The input is an image, and the output is multi class. There are 4 methods which are used and are indicated below:

Conv2D: A convolutional layer is created using this method. The filter count is the first parameter, while the filter size is the second. As an activation function, relu non-linearity is used with Padding enabled. There are 2 options for Padding: same or valid. Same indicates padding with the number on the edge, while valid indicates no padding. For convolution layers, the stride is set to 1 by default.

MaxPooling2D: The only argument is the window size, which is used to create a max pooling layer. Because it is the most frequent, a 2x2 window is used. The default stride length is retained, which is equal to the window size 2.

Flatten: Flatten the output of the convolution + pooling layers before feeding it into the fully linked layers

Dropout: Dropout is a simple concept that is used to prevent overfitting. During training, a neuron is momentarily "dropped" or inhibited with probability p at each repetition. This signifies that at this iteration, all this neuron's inputs and outputs will be disabled. At each training step, the dropped-out neurons are resampled with probability p, so a dropped-out neuron at one step can become active at the next. The dropout-rate hyperparameter p is commonly a number around 0.5, which corresponds to 50 percent of the neurons being dropped out.

# 3.5.2 Method 2: VGG Feature Extraction + 2D CNN based Classification

VGG16 is a CNN (Convolutional Neural Network) that is widely regarded as one of the best computer vision models available today. The creators of this model analyzed the networks and enhanced the depth using architecture with very small (3x3) convolution filters, which outperformed previous-art setups significantly. The depth was increased to 16–19 weight layers, resulting in 138 trainable parameters. VGG16 is a 92.7 percent accurate object identification and classification system that can classify 1000 images into 1000 different categories. It's a common picture classification algorithm that's simple to utilize with transfer learning (Simonyan and Zisserman, 2014)

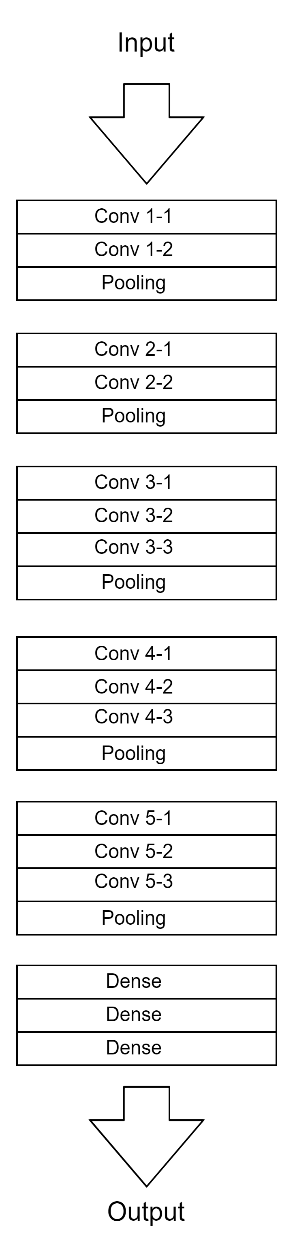


Figure 3: VGG-16 Architecture

Figure 3 shows the VGG -16 architecture. The key pointers to note:

The 16 in VGG16 stands for 16 weighted layers. VGG16 comprises thirteen convolutional layers, five Max Pooling layers, and three dense layers, for a total of twenty-one layers, but only sixteen weight layers, or learnable parameters layers.

VGG16 uses a 224, 244 input tensor size with 3 RGB channels.

The most distinctive feature of VGG16 is that, rather than having a huge number of hyper-parameters; it uses 3x3 filter convolution layers with stride 1 and always uses the same padding and maxpooling layer of 2x2 filter stride 2.

The convolution and maxpooling layers are placed in a regular pattern throughout the architecture.

* Conv-1 layer has 64 filters, Conv-2 layer has 128 filters, Conv-3 layer has 256 filters, and Conv 4 and Conv 5 layers have 512 filters.
* Following a stack of convolutional layers, three Fully Connected (FC) layers are added: the first two have 4096 channels each, while the third performs 1000-way image classification and so has 1000 channels (one for each class). The soft-max layer is the final layer.

Challenges:

* It takes a longer time to train (the initial VGG model took 2–3 weeks to train on the Nvidia Titan GPU).
* VGG-16 trained weights are 528 MB in size. As a result, it consumes a significant amount of storage space and bandwidth, making it inefficient.

# 3.5.3 Method 3: Method 1/Method 2 with Attention mechanism

This is the study (Bahdanau et al., 2014) that established the now-famous "Attention Mechanism". Even though the concept of attention has evolved, the mechanism described in this study is still recognized as "Bahdanau Attention". The study describes the following:

* The notion of employing Neural Networks to translate phrases from a source language to a target language is known as Neural Machine Translation (NMT). Until this study, such NMT models have relied on numerous networks, each of which had to be trained separately.
* The research proposes that a single, massive neural network be built and trained to comprehend a sentence and correctly translate it, which is the foundation for all current Sequence to Sequence models based on Encoder-Decoder architecture.
* Machine Translation is analogous to finding a target sentence y that maximizes the conditional probability of p(y|x), where x is the source sentence, from a probabilistic standpoint.
* The goal of an NMT task is to use a Parallel training corpus to maximize the Conditional Probability of Sentence Pairs. To simulate such a relationship, a parameterized model would be employed, with Backpropagation utilized to learn the parameter weights.

A source sentence is fed into an encoder, which converts it into a fixed-length vector. The translation (target sentence) from the Encoded Vector is output by a Decoder. For a given source-target sentence pair, the Encoder-Decoder system is jointly trained to maximize the conditional probability of an accurate translation. There are some limitations with encoder-decoder architecture. For information about the source sentence, the Decoder only uses the last encoded fixed-length vector. It's very difficult for the Encoder to compress all the information into a single vector when the source sentence is quite long. The performance of a basic encoder-decoder degrades significantly as the length of a source sentence increases, according to actual evidence.

The research proposes an Encoder-Decoder model extension that learns to 'align' and 'translate' together. When the NMT model generates a translated term, it does a soft search for a set of positions in the source sentence and looks for the positions with the highest concentration of relevant information. It's like selecting the words that make the most sense in the final translation. This is incompatible with the idea of storing the full source sentence into a single fixed-length context vector. The NMT model then predicts a target translation using context vectors associated with these source positions as well as previously generated translation outputs. The source text is encoded as a sequence of vectors, and the decoder selects a subset of these vectors to produce the translation. It allows the NMT model to interpret long words and do a selective search based on context importance rather than squashing all the information into a single vector.

# 3.5.4 Method 4: VGG Features + Transformer’s Encoder + GRU based Decoder

Due to advances in Sequence Modelling, such as the comeback of Long-Short Term Memory networks (LSTMs) and the development of Gated Recurrent Units (GRUs), generating captions in videos and summarizing them have been recently popular. Existing architectures use CNNs to extract spatio-temporal information and soft attention layers to model dependencies using GRUs or LSTMs. The layers which are attention based, help in paying to the important aspects where recurrent units are also improved; nonetheless, there are problems from recurrent units' intrinsic flaws. With Transformer model getting introduced, the Sequence Modeling domain has taken a new shape. 3D CNN architectures like C3D and Two-stream I3D for video extraction (Bilkhu et al., 2019) is used to construct a Transformer-based model for generating captions in videos. Techniques that can reduce dimensions are used to control the total size in a model.

# 3.5.5 Method 5: I3D Pretrained/Video Vision Transformer + Transformer’s Encoder + GRU based Decoder

Rather than employing frame-level feature extractors, networks to extract spatio-temporal information from videos are used directly. 3D convolutions are used in these structures to encode both spatial and temporal information in videos. Using 2D convolutions on an image or a video (series of frames) results in a single feature map. Using 3D convolutions on a set of frames, on the other hand, produces a set of feature maps. The size of the temporal kernel and the strides employed determine the number of feature mappings.

Recent advancements in the field of activity recognition have resulted in a variety of designs that can be used to extract spatio-temporal features. Instead of depending on a recurrent network to encode information from each time step, architectures that can directly offer temporal information are looked at. Features are extracted for the Transformer model using I3D (Inflated 3D) Convolutional Neural Networks for Activity Recognition (Bilkhu et al., 2019)

# 3.6 Evaluation Metrics

Accuracy is one of the metrics to describe the accuracy of an algorithm on a classification task. Since the dataset is balanced, Accuracy as the evaluation metric is used to measure the performance of the models. It is the number of samples that are paired divided by the number of samples.

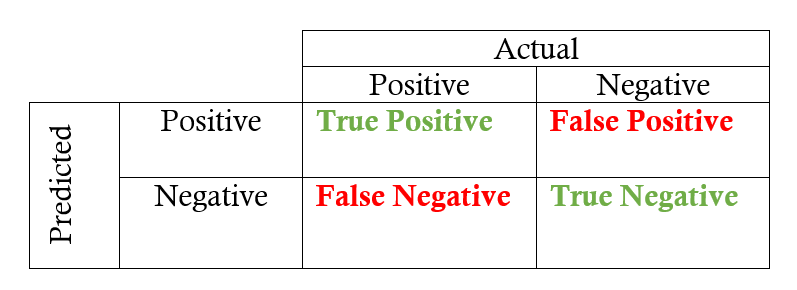


Figure 4: Confusion Matrix

Figure 4 shows the confusion matrix. Accuracy is the proportion of true results among the total number of cases examined.

Accuracy = (True Positive + True Negative) / (True Positive + False Positive + False Negative + True Negative)

Accuracy is a valid choice of evaluation for classification problems which are well balanced and not skewed or No class imbalance.

Top-1 Accuracy is the conventional accuracy; model prediction (the one with the highest probability) must be exactly the expected answer. It measures the proportion of examples for which the predicted label matches the single target label.

Top-5 Accuracy means any of the model’s top 5 highest probability answers matches with the expected answer. It considers a classification correct if any of the 5 predictions matches the target label.

# 3.7 Summary

Sign language, which makes use of gestures and visual signals, is the main form of communication for the deaf. Deaf and hearing-impaired people can communicate with others outside of their own community by using sign language. The process of creating text and speech during computer recognition of sign languages begins with the identification of sign motions. It is possible to identify both static and dynamic sign gestures, with static gesture identification being the simpler of the two. Temporal segmentation techniques are also used to record time periods for fine segmentation and to identify the endpoint for each potential sign, even though Hidden Markov Model (HMM)-based algorithms make up most recent developments in Sign Language Recognition (SLR). The development of movie subtitles using distinctive end-to-end sequence-to-sequence models is an important research area. Attention processes are used by Recurrent Neural Networks (RNN), particularly LSTMs. This research article presents the research methodology together with the target dataset description, data pre-processing, several models and algorithms, and assessment metrics.

# 3.8 Required Resources

The research will need below hardware and software resources throughout the implementation.

# 3.8.1 Software Requirements

Operating System: Ubuntu/Mac OS/Windows

* Programming Language: Python 3.9.1, Shell Script
* Package Manager: pip
  + Python Libraries:
    - OpenCV
    - NLTK
    - Matplotlib
    - Numpy
    - CSV

# 3.8.2 Hardware Requirements

A laptop with below configuration will be used:

* SSD: 512GB
* Ram: 40GB (depends on batch size)
* Graphics: NVIDIA 2080 RTI, 12GB

# References

Bahdanau, D., Cho, K. and Bengio, Y., (2014) Neural Machine Translation by Jointly Learning to Align and Translate. [online] Available at: http://arxiv.org/abs/1409.0473.

Bilkhu, M., Wang, S. and Dobhal, T., (2019) Attention is all you need for Videos: Self-attention based Video Summarization using Universal Transformers. [online] Available at: http://arxiv.org/abs/1906.02792.

Camgoz, N.C., Koller, O., Hadfield, S. and Bowden, R., (2020) Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. [online] Available at: http://arxiv.org/abs/2003.13830.

Carreira, J. and Zisserman, A., (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. [online] Available at: http://arxiv.org/abs/1705.07750.

Cui, R., Liu, H. and Zhang, C., (2019) A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia*, 217, pp.1880–1891.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G., Cai, J. and Chen, T., (2015) Recent Advances in Convolutional Neural Networks. [online] Available at: http://arxiv.org/abs/1512.07108.

Huang, J., Zhou, W., Zhang, Q., Li, H. and Li, W., (2018) Video-based Sign Language Recognition without Temporal Segmentation. [online] Available at: http://arxiv.org/abs/1801.10111.

Kreutzer, J., Bastings, J. and Riezler, S., (2019) Joey NMT: A Minimalist NMT Toolkit for Novices. [online] Available at: http://arxiv.org/abs/1907.12484.

Li, D., Opazo, C.R., Yu, X. and Li, H., (2019) Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. [online] Available at: http://arxiv.org/abs/1910.11006.

Rastgoo, R., Kiani, K. and Escalera, S., (2021) *Sign Language Recognition: A Deep Survey*. *Expert Systems with Applications*, .

Schmidt, C., Koller, O., Bellgardt, M., Forster, J. and Ney, H., (2014) *Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather*. [online] Available at: www.signspeak.eu.

Simonyan, K. and Zisserman, A., (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. [online] Available at: http://arxiv.org/abs/1409.1556.

Sincan, O.M. and Keles, H.Y., (2020) AUTSL: A Large Scale Multi-modal Turkish Sign Language Dataset and Baseline Methods. [online] Available at: http://arxiv.org/abs/2008.00932.

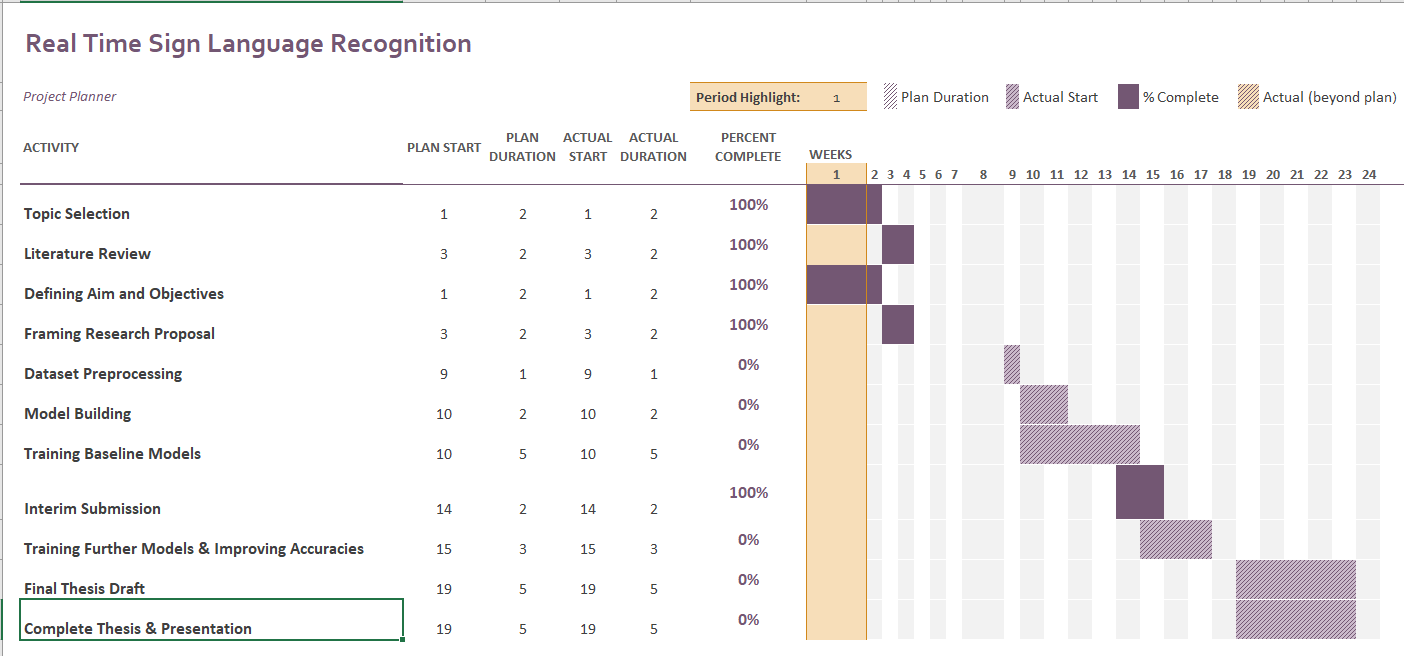
Sridhar, A., Ganesan, R.G., Kumar, P. and Khapra, M., (2020) INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition. In: *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, Inc, pp.1366–1375.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K., (2015) Sequence to Sequence -- Video to Text. [online] Available at: http://arxiv.org/abs/1505.00487.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H. and Courville, A., (2015) Describing Videos by Exploiting Temporal Structure. [online] Available at: http://arxiv.org/abs/1502.08029.

Zhang, J., Zhou, W. and Li, H., (2014) A Threshold-based HMM-DTW Approach for Continuous Sign Language Recognition. In: *Proceedings of International Conference on Internet Multimedia Computing and Service - ICIMCS ’14*. [online] New York, New York, USA: ACM Press, pp.237–240. Available at: http://dl.acm.org/citation.cfm?doid=2632856.2632931.

# APPENDIX A: RESEARCH PLAN



# APPENDIX B: RESEARCH PROPOSAL

Real Time Sign Language Recognition

Karthik S K

Research Proposal

MAY 2022

**Abstract**

Deaf people all over the world use sign languages to communicate visually. The most common ways to sign words and sentences are by waving fingers, arms, hands, and making motions with face. Sign languages are complete languages with vocabulary, grammar which has unique characteristics. Sign language recognition is a complex task. Several factors are employed to identify signs, including hand orientations with many forms, movement of hands, posture of body, and expressions in faces. Even with cutting-edge models, tackling the problem for a signs with larger vocabulary using a computer in real-world circumstances remains a difficulty. A complete language with unique grammar, and other distinguishing features called Indian Sign Language (ISL) is available. About 5 million hearing-impaired community in India use this language. There is presently no dataset is available on ISL in public domain to test Sign Language Recognition techniques (SLR). The Indian Lexicon Sign Language Dataset (INCLUDE) is considered in this study. The dataset has less than half million frames from over 4000 films and over 250 different word signs from more than 12 different word categories. INCLUDE is documented by getting assistance from signers who have expertise resembling natural conditions. Numerous deep neural networks with various methodologies to augment data, to extract features, to encode and decode have been evaluated. Due to the dataset's size and quality, these models can be tested on ISL for Sign Language Recognition. Various deep learning models is presented, and an accurate model is determined. The most accurate model has a 94.5 percent accuracy rate. This model just trains a decoder and employs feature extractor and encoder which is pre-trained. It improves on previous results and provides a quick way to support SLR in various languages.

**Table of Contents**

Abstract 2

[LIST OF FIGURES 4](#_Toc102555090)

[LIST OF TABLES 4](#_Toc102555091)

[LIST OF ABBREVIATIONS 4](#_Toc102555092)

[1. Background 4](#_Toc102555093)

[2. Related Research 6](#_Toc102555094)

[3. Research Questions (If any) 10](#_Toc102555095)

[4. Aim and Objectives 10](#_Toc102555096)

[5. Significance of the Study 10](#_Toc102555097)

[6. Scope of the Study 11](#_Toc102555098)

[7. Research Methodology 11](#_Toc102555099)

[8. Required Resources 20](#_Toc102555100)

[9. Research Plan 20](#_Toc102555101)

[References 21](#_Toc102555102)

LIST OF FIGURES

Figure 1: The training process for two stream I3D on Kinetics dataset 11

Figure 2: CNN architecture with Softmax output 12

Figure 3: VGG-16 Architecture 13

Figure 4: Confusion Matrix 19

LIST OF TABLES

Table 1: INCLUDE: Category Size 12

Table 2: INCLUDE: Key statistics 13

LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| SL | Sign Language |
| SLR | Sign Language Recognition |
| SLT | Sign Language Translation |
| CSLR | Continuous Sign Language Recognition |
| ISL | Indian Sign Language |
| EDA | Exploratory Data Analysis |
| CNN | Convolutional Neural Network |
| VGG | Visual Geometry Group |
| NMT | Neural Machine Translation |
| I3D | Inflated 3D |
| HMM | Hidden Markov Model |
| RNN | Recurrent Neural Network |
| LSTM | Long Short-Term Memory |
| WASL | Word-level American Sign Language |
| AUTSL | Ankara University Turkish Sign Language |
| GRU | Gated Recurrent Units |

# 

1. Background

Video and image creation and consumption are continually increasing. As a result, deep learning is increasingly being employed for classifying videos, detecting, and tracking objects, recognizing actions, answering questions visually, and encoding videos. These jobs have real-world applications, with significant accessibility implications. SLR (sign language recognition) is a form of action recognition problem (Sridhar et al., 2020). Sign languages are way of communicating techniques mostly where hearing impaired community uses throughout the world. The most frequent technique of signing words and sentences is by waving fingers, arms, hands, and making motions with face. Sign languages are fully formed languages with their own syntax and lexicon (Sridhar et al., 2020). Furthermore, they differ from one place to another and are often incomprehensible to one another, despite some commonality. It differs from languages spoken in a region with lexicon and articulation pace. The mode of communication in hearing-impaired community being sign languages have language frameworks which are unique. The purpose of sign language interpretation systems is to use techniques like vision to automatically translate sign languages. The two main goals in this technique are word-level sign language recognition (Sridhar et al., 2020) (also known as "isolated sign language recognition") and sentence-level sign language recognition (also known as "continuous sign language recognition").

There are various reasons as to why building models with computers for Sign language recognition can be difficult. Analyzing numerous body components, such as the hand, arms, and face, is required. Although the hand motions for some pairs of signs appear to be extremely identical, they can be distinct in face expressions. Similar hand gestures when repeated can appear distinct. Another problem are the differences in how different signers perform a sign, such as body and position variances, duration variations of different parts of the signals, and so on. Variation in illumination and backdrop complicates the challenge, which is a problem in computer vision. When the number of signs in the corpus grows, these difficulties become more difficult to solve (Sincan and Keles, 2020).

Indian Sign Language (ISL) has its own linguistic characteristics (Sridhar et al., 2020). ISL differs from its western equivalents in several ways, the most notable of which being its lexicon. While other sign languages have few compound signs as features (signs that are made up of two or more signs), ISL has a widespread compounding system. Compounding signals for Male and Sibling, for example, are used to sign the word Brother, whereas compound signs for Female and Marry are used to sign the term Wife. This is also a deviation from Indian spoken languages, which are noted for their large number of intricate kinship phrases (Sridhar et al., 2020) . Many possible compositions allow a wide lexicon on ISL. Another way in which ISL is different is in the amount of room required for signing. In ISL, signing from the top portion of the body represents distance and authority, whereas in many other Sign Languages, distance is indicated by a plane horizontal to front of the signer. These and other differences between ISL and other sign languages need the development and testing of novel methods for ISL Sign Language Recognition. Furthermore, India's vast population of deaf individuals – more than 5 million – emphasizes the significance of any solution that can be deployed in practice (Sridhar et al., 2020).

The resource availability such as big, standardized data is critical for the development and evaluation of models. Importantly, these datasets are not in the public domain, preventing additional research. These datasets also have two significant flaws. To begin with, the number of classes, or different word signs, is extremely minimal. The videos are frequently limited in some way. Only the hands are visible in the datasets because the photographs were cropped. Pictures or films with consistent backdrops, with the signer's clothing generally matching the background color. The usefulness of training models on real time datasets is limited by these dataset restrictions. As a result, no ISL dataset of sufficient quality and size for further study in machine learning (Sridhar et al., 2020) is currently available.

The first publicly available ISL dataset, (Sridhar et al., 2020) Indian Lexicon Sign Language Dataset (INCLUDE), has been proposed to look at the absence of datasets in public domain and algorithms that can scale for recognizing ISL. Data augmentation, using networks already trained for extracting relevant features, and encoding and decoding were all part of the deep learning workflows for SLR on this dataset. The feature extraction was merged with already trained network for detecting important poses, and the encoding was done with already trained network, which is MobileNet, and the decoding was done with trained LSTMs which are bidirectional.

2. Related Research

The research survey related to solving the problem of Sign Language Recognition (SLR) was carried out and is summarized in this section. The community having problems in hearing use sign languages for communication and hence it is meaningful and important for automatic translation of a sign language. The problems that are prevalent in the domain of SLR are mainly two - isolated SLR and continuous SLR (Huang et al., 2018). Word by word recognition is isolated SLR, whereas translating entire sentences is continuous SLR.

The current methods for continuous SLR while using isolated SLR as building blocks also add layers of pre-processing and post-processing which indicate temporal segmentation and sentence synthesis respectively. Most existing SLRs fall into the category of isolated SLR which deals with the recognition of words or expressions. Continuous SLR being more challenging involves reconstructing sentence structures, which divide the problem of recognizing sentences into three stages which are segmentation of videos with time, recognizing isolated word/expression, and sentence synthesis with a language model.

Most approaches in the recent past show great progress in SLR which are based on the Hidden Markov Model (HMM) with various features such as motion trajectory. The transitional movements between signs which are not correctly captured are addressed using the approaches of offline training and online recognition. The threshold matrix and rate thresholds are proposed in offline training, where each element of the matrix indicates the minimum probability when a segment belongs to a sign, and rate thresholds are defined as average probability for signs. If the evaluation of a certain segment is smaller than all the thresholds, then it is regarded as a transitional movement which should be removed. In the online recognition stage, Dynamic Time Warping (DTW) and Length-Root method is used to record the time intervals for fine segmentation and the endpoint for each candidate sign is determined (Zhang et al., 2014).

The temporal segmentation even though popular is difficult to implement due to below reasons:

* The transitional movement between hand gestures can be subtle and ambiguous.
* Inaccurate segmentation can incur significant performance penalty on subsequent steps.
* The isolated SLR step that deals with the recognition of words or expression requires per-video-frame labels which are extremely time consuming.

An end-to-end novel sequence-to-sequence model that can generate video captions is a relevant research area. The state-of-the-art performance demonstrated by Recurrent Neural Networks, specifically LSTMs in generating image captions, is trained on pairs of video-sentence and the model associates a sequence of video frames to sequence of words to describe an event in the video clip (Venugopalan et al., 2015). Attention mechanisms can also be incorporated into LSTM for automatic selection of most likely video frames (Yao et al., 2015).

Industry labs and research groups have created solutions which are open source for Neural Machine Translation (NMT). The solutions are based on platforms related to deep learning. However, these tools are targeted only towards research groups having a good understanding of deep learning architecture, and those who know how to handle code bases that are large.

Joey NMT (Kreutzer et al., 2019) designed a model which provides minimum code platform with quality that can be compared to more standard benchmarks with complex code bases. It includes standard network architectures (RNN, transformer, different attention mechanisms, input feeding, configurable encoder/decoder bridge), standard learning techniques (dropout, learning rate scheduling, weight tying, early stopping criteria), and visualization/monitoring tools.

RWTH-PHOENIX-Weather 2012 corpus for German sign language (DGS) recorded sign language footage in 2009 and 2010 by “PHOENIX”, a German public TV station. This corpus added more data with bilingual annotation in DGS glosses and written German, creating the RWTH-PHOENIX-Weather 2014 corpus. There are also annotations for spatial positions of face and hands of a signer for more than 40,000 video frames along with annotating hand shapes and on the frame level orientations (Schmidt et al., 2014).

The methods that were employed previously for continuous sign language recognition involved hidden Markov models (HMM) with limited capacity for capturing temporal information. Using the framework of deep neural networks for continuous sign language recognition, transcribes videos of sign language sentences to sequences of ordered gloss labels. Deep CNN stacked with temporal fusion layers for extracting features, and bidirectional Recurrent Neural Networks (RNN) for sequence learning modules were proposed with an iterative optimization process to exploit the representational capability of deep neural networks with limited data. The end-to-end recognition model for alignment proposal was trained first, and then the proposal was used to tune the feature extraction module. This process can run iteratively to achieve improvements on the recognition performance (Cui et al., 2019).

The prior studies in Sign Language Translation (SLT) have shown that effectively recognizing individual signs improves the performance of translation to a great extent. A novel transformer architecture jointly learns Continuous Sign Language Recognition (CSLR) and Translation while being trainable in an end-to-end manner. The recognition and translation problems are addressed in a single unified architecture called Connectionist Temporal Classification (CTC) loss (Camgoz et al., 2020). The contributions can be summarized as:

* The novel formalization of Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT) which is multi-task uses the power of glosses without limiting the translation to spoken language.
* The state-of-the-art results achieved in successful application of transformers for Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT), outperforming all comparable previous approaches.
* Guiding future research in this field with baseline results which are new.

Indian Sign Language (ISL) which is in use by over 5 million hearing impaired community in India is a complete sign language with unique linguistic attributes with one grammar, syntax, and vocabulary. To carry out and assess Sign Language Recognition (SLR) methods, no dataset on ISL is available on public domain. Previously, ISL include video and image datasets with only one image indicating a sign. Since these datasets are not in the public domain, any further study is a challenge. Other limitations are with the word signs (classes) that are distinct being very low, and reduced images with certain portions which are visible in the datasets. These constraints on the dataset pose challenges in limiting the models that can be built and trained for real world scenarios. The quality and size required for Machine Learning (ML) study are not available for ISL dataset (Sridhar et al., 2020) .

The problems with datasets not in the public domain and models that cannot scale for ISL recognition has been addressed by proposing the first available ISL dataset in public domain called Indian Lexicon Sign Language Dataset (INCLUDE). It has less than half a million frames with over 250 classes from more than 12 different word categories. There are videos documented with the help of hearing-impaired community, with the resolution, lighting and background chosen to resemble real world scenarios. A subset that is small named INCLUDE-50 with 50 classes was also proposed for rapid evaluation. Numerous models for SLR on INCLUDE dataset included data augmentation, feature extraction using trained networks, with encoding and decoding. The feature extraction combined with trained pose network for detecting poses, encoding with MobileNet network that is trained, and decoding trained LSTMs which are bidirectional resulted with an accuracy of 94.5% on INCLUDE-50 and 85.6% on INCLUDE (Sridhar et al., 2020).

Deaf persons can use vision-based sign language recognition to communicate with others. Most available sign language databases, on the other hand, are confined to a small number of words. Models learned from those datasets cannot be used due to the low vocabulary size. Word-Level American Sign Language (WLASL) video dataset with more than 2000 words performed by more than 100 signers was proposed (Li et al., 2019) . This dataset is handy to the researchers on the public domain. It is the largest publicly available ASL dataset for word-level sign recognition research. Various deep learning approaches for word-level sign identification evaluated the performance in scenarios which are large scale. Two different models were used and compared: (i) a holistic approach which is based on visual appearance and (ii) An approach that is based on human pose which is 2D. Both provide baselines which are valuable for method benchmarking in the community. Furthermore, the study offers a unique temporal graph convolutional networks (Posture-TGCN) which is pose based jointly modeling spatial and temporal relationships in trajectories of human pose. Findings reveal that pose-based and appearance-based models perform similarly on 2,000 words/glosses, with top-10 accuracy up to 62.63 percent, showing the dataset's validity and difficulties.

Sign language identification is a difficult challenge in which signs are identified by numerous sources of simultaneous local and global articulations, such as hand shape and orientation, hand movements, body position, and facial expressions. Even with state-of-the-art models, solving this problem computationally for a wide vocabulary of signs in real-world scenarios remains a challenge. A new large-scale multi-modal Turkish Sign Language dataset (AUTSL) with a benchmark and performance evaluation baseline models was proposed in this paper (Sincan and Keles, 2020) . In total, there are 38,336 isolated sign video samples in our collection, which includes 226 signs performed by 43 individual signers. The samples include a wide range of backgrounds captured in both indoor and outdoor settings. Furthermore, throughout the recordings, signers' spatial placements and postures change. Each sample was captured using the Microsoft Kinect v2 and includes RGB color, depth, and skeletal modalities. For user-independent model evaluations, the benchmark training and test sets were created. Employed Convolutional Neural Networks (CNNs) to extract features and unidirectional and bidirectional Long Short-Term Memory (LSTM) models to characterize temporal information, and provided empirical evaluations using the benchmark. To boost the performance of our models, added feature pooling modules and temporal attention. On the AUTSL and Montalbano datasets, tested the baseline models.

On the Montalbano dataset, the models achieved competitive results with state-of-the-art approaches, achieving 96.11 percent accuracy. The models achieved 95.95 percent accuracy in AUTSL random train-test splits. The best baseline model achieved 62.02 percent accuracy in the proposed user-independent benchmark dataset. The disparities in the results of the identical baseline models demonstrate the difficulties in the benchmark dataset.

3. Research Questions (If any)

The following research questions are suggested for each of the research objectives as highlighted as follows:

* To evaluate if the model can be implemented for real time inference
* To check if the model can recognize on videos outside the dataset, but similar domains

4. Aim and Objectives

The main aim of this research is to propose a model that will perform Sign Language Recognition. Given a video with person performing sign language for words, the model will recognize the video and will classify that video into one of the elements in vocab space.

The research objectives are formulated based on the aim of this study which are as follows:

* To build a training pipeline that leads to a robust sign language recognition model.
* To build the system in Real Time.
* To make an UI for the Sign Language Recognition Model.

5. Significance of the Study

Sign languages are visual languages that use hand, face, and body motions to communicate. Learning a second language is not only beneficial to your brain, but it can also help you enhance your communication abilities. Early man utilized signs to communicate, according to some scholars, long before spoken language was established. Even though today's world progressed significantly, sign languages in their most basic form are still used such as pushing index finger between lips to calm a noisy child, raising hand to hail a cab, or pointing to an item on the menu.

Sign language is the link that connects us to the world of persons who are deaf or have difficulty speaking. A variety of hand, finger, arm, head, and face expressions are used to help the deaf with those around them and vice versa. It enables individuals to comprehend the world around them through visual descriptions and hence contribute to society. Some of the benefits of sign language include - assists persons who are deaf or hard of hearing aiding social inclusions. It provides deaf children an opportunity to educate themselves. Non-deaf volunteers who volunteer to learn sign language to communicate with the disabled are instilled with a sense of social responsibility as well as sensitivity.

# 

6. Scope of the Study

The scope of the study covers the following:

* The scope of this study is limited to INCLUDE dataset.
* Not exploring MultiModal Sign Language and Very Deep recent Transformer architectures due to lack of time and resource constraint.
* The study will also be limited to the evaluation of Accuracy as a performance metric.

7. Research Methodology

Sign language is a system of communication using visual gestures and signs as used by the deaf community and is the main mode of communication. Deaf and people who have problems hearing use sign language between their own community and other people. Computer recognition of sign languages covers sign gesture recognition and continues till text/speech generation. Sign gestures can be classified as static and dynamic with static gesture recognition being simpler than dynamic gesture recognition. While most of the approaches in recent times show progress in Sign Language Recognition (SLR) which are based on Hidden Markov Model (HMM), temporal segmentation techniques are also used to record time intervals for fine segmentation and the endpoint for each candidate sign is determined. Generating captions for videos using novel end-to-end sequence-to-sequence models is a relevant research area. Attention mechanisms are used in Recurrent Neural Networks (RNN), specifically LSTMs. In this research paper, as part of research methodology the following sections are covered which includes target dataset description, data pre-processing, different models and algorithms and evaluation metrics.

7.1 Dataset Description

Indian Lexicon Sign Language Dataset - INCLUDE - an ISL (Indian Sign Language) dataset contains 0.27 million frames across 4287 videos over 263-word signs from 15 different word categories. This dataset is recorded with the help of experienced signers. Each video is a recording of 1 ISL sign, signed by deaf students from St. Louis School for the Deaf, Adyar, Chennai (Sridhar et al., 2020).

Based on the 2 main principles - The videos resembling real life scenarios, and a dataset covering a diverse set of signs with multiple videos, the dataset was created by recording with the help 7 experienced signers from the school for the deaf. Each class in the dataset is a sign, and multiple signers have signed them. It contains 2 to 6 videos per class (Sridhar et al., 2020) .

The classes which are word signs represent 15 broad word categories which are words that are popularity in ISL. This is shown in Table 1.1. The category-wise summary statistics are shown in Table 1.2 (Sridhar et al., 2020) .

Table 1: INCLUDE: Category Size (Sridhar et al., 2020)

|  |  |  |
| --- | --- | --- |
| **Category** | **Number of Classes** | **Number of Videos** |
| Adjectives | 59 | 791 |
| Animals | 8 | 166 |
| Clothes | 10 | 198 |
| Colours | 11 | 222 |
| Days and Time | 22 | 306 |
| Electronics | 10 | 140 |
| Greetings | 9 | 185 |
| Means of Transport | 9 | 186 |
| Objects at Home | 27 | 379 |
| Occupations | 16 | 225 |
| People | 26 | 513 |
| Places | 19 | 399 |
| Pronouns | 8 | 168 |
| Seasons | 6 | 85 |
| Society | 23 | 324 |
| **Total** | **263** | **4287** |

Table 2: INCLUDE: Key statistics (Sridhar et al., 2020)

|  |  |
| --- | --- |
| **Characteristic** | **INCLUDE** |
| Categories | 15 |
| Words | 263 |
| Videos | 4287 |
| Avg. Videos per Class | 16.3 |
| Avg. Video Length | 2.57s |
| Min. Video Length | 1.28s |
| Max. Video Length | 6.16s |
| Frame Rate | 25fps |
| Resolution | 1920x1080 |

7.2 Data Preprocessing

The preprocessing techniques on video data include - VGG (Visual Geometry Group) Feature Extraction, and I3D (Inflated 3D ConvNet) Feature Extraction & Vision Transformer. In this section each of the above feature extraction techniques are analyzed.

The temporal element being the key difference between image information and video information. As a result, deep learning models have been improved to include processing in 3D for temporal data as well. Researchers from DeepMind and the University of Oxford presented the I3D model in a paper (Carreira and Zisserman, 2017). The research examines prior methods to solve problems with detecting actions in videos, while also offering a novel method. Their method starts with a two-dimensional design and then all the filters and pooling kernels are inflated. It offers a dimension which is additional to be considered by inflating them, which is time. 2D model filters are N x N which is a square, but when they are inflated, becomes N x N x N which is cubic.

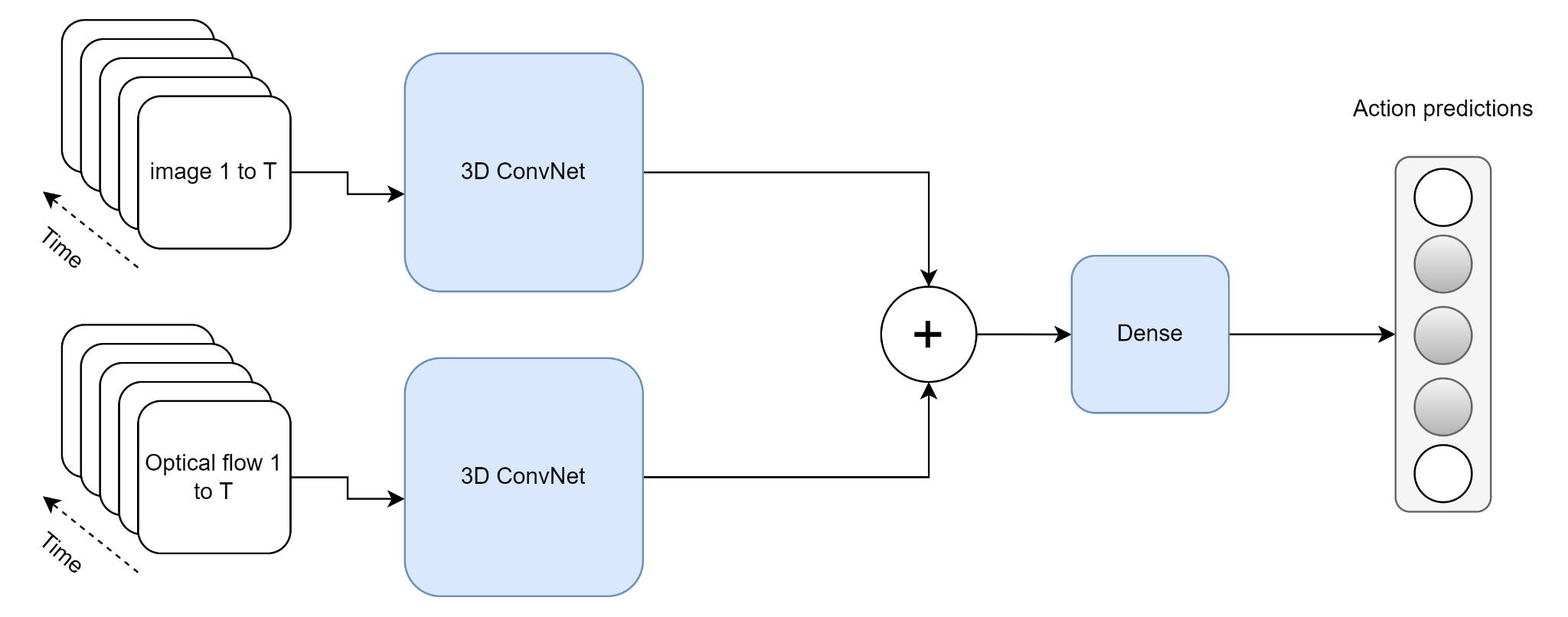


Figure 1: The training process for two stream I3D on Kinetics dataset

Since the dataset contains videos, some of the pre-processing techniques such as imputing missing values, outlier treatment is not applicable, and hence these are not mentioned in this section.

7.3 Model Building

In this section different models and algorithms that will be used for effective Sign Language Recognition are explained. The following methods are used:

* Method 1: 3D CNN based Classification with Softmax in the last layer
* Method 2: VGG Feature Extraction + 2D CNN/NN based Classification
* Method 3: Method 1/Method 2 with or without Attention Mechanism
* Method 4: VGG Features + Transformers Encoder + GRU based Decoder
* Method 5: I3D Pretrained/Video Vision Transformer + Transformers Encoder + GRU based Decoder

7.3.1 Method 1: 3D CNN based Classification with Softmax in the last layer

Convolutional Neural Networks (CNN) are used in a variety of applications. It is, without a doubt, the most widely used deep learning architecture. The enormous popularity and effectiveness of convnets has sparked a recent rise in interest in deep learning. AlexNet sparked interest in CNN in 2012, and it has grown rapidly since then. Researchers went from an 8-layer AlexNet to a 152-layer ResNet in just three years. CNN has become the go-to model for any image-related issue. They outperform the competitors in terms of accuracy. The fundamental advantage of CNN over its predecessors is that it discovers essential traits without the need for human intervention. In addition, CNN is computationally efficient. It performs parameter sharing and uses special convolution and pooling algorithms. CNN models can run on any device, making them globally appealing (Gu et al., 2015)

Fig 2 shows the architecture of CNN with Softmax output in the last layer.

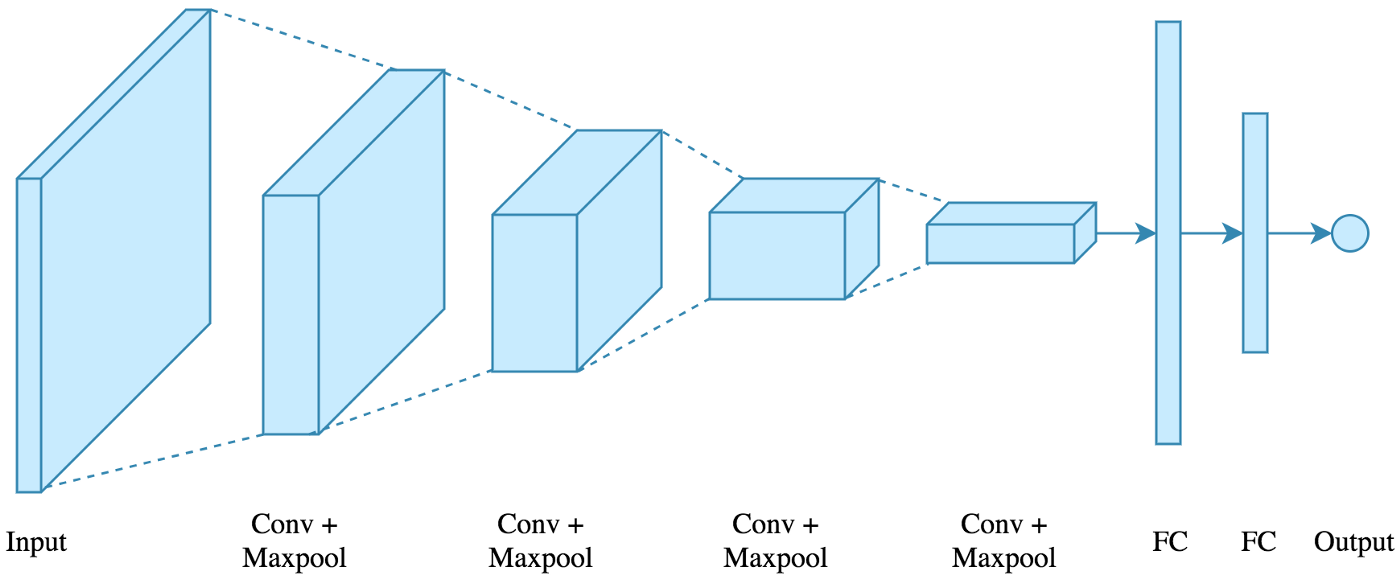


Figure 2: CNN architecture with Softmax output

The following architecture will be used: 4 convolutions + pooling layers with 2 fully connected layers. Image goes as an input, and the output is multi class. There are 4 methods which are used and are indicated below:

* Conv2D: A convolutional layer is created using this method. The filter count is the first parameter, while the filter size is the second. As an activation function, relu non-linearity is used with Padding enabled. There are 2 options for Padding: same or valid. Same indicate padding with the number on the edge, while valid indicates no padding. For convolution layers, the stride is set to 1 by default.
* MaxPooling2D: The only argument is the window size, which is used to create a maxpooling layer. Because it is the most frequent, a 2x2 window is used. The default stride length is retained, which is equal to the window size 2.
* Flatten: Flatten the output of the convolution + pooling layers before feeding it into the fully linked layers
* Dropout: Dropout is a simple concept that is used to prevent overfitting. During training, a neuron is momentarily "dropped" or inhibited with probability p at each repetition. This signifies that at this iteration, all this neuron's inputs and outputs will be disabled. At each training step, the dropped-out neurons are resampled with probability p, so a dropped-out neuron at one step can become active at the next. The dropout-rate hyperparameter p is commonly a number around 0.5, which corresponds to 50 percent of the neurons being dropped out.

7.3.2: Method 2: VGG Feature Extraction + 2D CNN based Classification

VGG16 is a CNN (Convolutional Neural Network) that is widely regarded as one of the best computer vision models available today. The creators of this model analyzed the networks and enhanced the depth using an architecture with very small (3x3) convolution filters, which outperformed previous-art setups significantly. The depth was increased to 16–19 weight layers, resulting in 138 trainable parameters. VGG16 is a 92.7 percent accurate object identification and classification system that can classify 1000 images into 1000 different categories. It's a common picture classification algorithm that's simple to utilize with transfer learning (Simonyan and Zisserman, 2014)

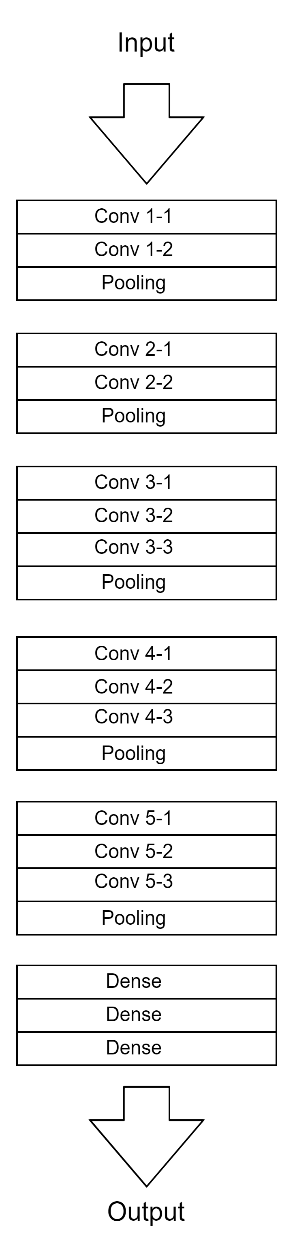


Figure 3: VGG-16 Architecture

Figure 3 shows the VGG -16 architecture. The key pointers to note:

* The 16 in VGG16 stands for 16 weighted layers. VGG16 comprises thirteen convolutional layers, five Max Pooling layers, and three Dense layers, for a total of twenty-one layers, but only sixteen weight layers, or learnable parameters layers.
* VGG16 uses a 224, 244 input tensor size with 3 RGB channels.
* The most distinctive feature of VGG16 is that, rather than having a huge number of hyper-parameters, it uses 3x3 filter convolution layers with stride 1 and always used the same padding and maxpooling layer of 2x2 filter stride 2.
* The convolution and maxpooling layers are placed in a regular pattern throughout the architecture.
* Conv-1 layer has 64 filters, Conv-2 layer has 128 filters, Conv-3 layer has 256 filters, and Conv 4 and Conv 5 layers have 512 filters.
* Following a stack of convolutional layers, three Fully Connected (FC) layers are added: the first two have 4096 channels each, while the third performs 1000-way image classification and so has 1000 channels (one for each class). The soft-max layer is the final layer.

Challenges:

* It takes a longer time to train (the initial VGG model took 2–3 weeks to train on the Nvidia Titan GPU).
* VGG-16 trained weights are 528 MB in size. As a result, it consumes a significant amount of storage space and bandwidth, making it inefficient.

7.3.3 Method 3: Method 1/Method 2 with Attention mechanism

This is the study (Bahdanau et al., 2014) that established the now-famous "Attention Mechanism". Even though the concept of attention has evolved, the mechanism described in this study is still recognized as "Bahdanau Attention". The study describes the following:

* The notion of employing Neural Networks to translate phrases from a source language to a target language is known as Neural Machine Translation (NMT). Until this study, such NMT models have relied on numerous networks, each of which had to be trained separately.
* The research proposes that a single, massive neural network be built and trained to comprehend a sentence and correctly translate it, which is the foundation for all current Sequence to Sequence models based on Encoder-Decoder architecture.
* Machine Translation is analogous to finding a target sentence y that maximizes the conditional probability of p(y|x), where x is the source sentence, from a probabilistic standpoint.
* The goal of an NMT task is to use a Parallel training corpus to maximize the Conditional Probability of Sentence Pairs. To simulate such a relationship, a parameterized model would be employed, with Backpropagation utilized to learn the parameter weights.

A source sentence is fed into an encoder, which converts it into a fixed-length vector. The translation (target sentence) from the Encoded Vector is output by a Decoder. For a given source-target sentence pair, the Encoder-Decoder system is jointly trained to maximize the conditional probability of an accurate translation. There are some limitations with encoder-decoder architecture. For information about the source sentence, the Decoder only uses the last encoded fixed-length vector. It's very difficult for the Encoder to compress all the information into a single vector when the source sentence is quite long. The performance of a basic encoder-decoder degrades significantly as the length of a source sentence increases, according to actual evidence.

The research proposes an Encoder-Decoder model extension that learns to 'align' and 'translate' together. When the NMT model generates a translated term, it does soft search for a set of positions in the source sentence and looks for the positions with the highest concentration of relevant information. It's like selecting the words that make the most sense in the final translation. This is incompatible with the idea of storing the full source sentence into a single fixed-length context vector. The NMT model then predicts a target translation using context vectors associated with these source positions as well as previously generated translation outputs. The source text is encoded as a sequence of vectors, and the decoder selects a subset of these vectors to produce the translation. It allows the NMT model to interpret long words and do a selective search based on context importance rather than squashing all the information into a single vector.

7.3.4 Method 4: VGG Features + Transformer’s Encoder + GRU based Decoder

Due to advances in Sequence Modelling, such as the comeback of Long-Short Term Memory networks (LSTMs) and the development of Gated Recurrent Units (GRUs), generating captions in videos and summarizing them have been recently popular. Existing architectures use CNNs to extract spatio-temporal information and soft attention layers to model dependencies using GRUs or LSTMs. The layers which are attention based, help in paying to the important aspects where recurrent units are also improved; nonetheless, there are problems from recurrent units' intrinsic flaws. With Transformer model getting introduced, the Sequence Modeling domain has taken a new shape. 3D CNN architectures like C3D and Two-stream I3D for video extraction (Bilkhu et al., 2019) is used to construct a Transformer-based model for generating captions in videos. Techniques that can reduce dimensions are used to control the total size in a model.

7.3.5 Method 5: I3D Pretrained/Video Vision Transformer + Transformer’s Encoder + GRU based Decoder

Rather than employing frame-level feature extractors, networks to extract spatio-temporal information from videos are used directly. 3D convolutions are used in these structures to encode both spatial and temporal information in videos. Using 2D convolutions on an image or a video (series of frames) results in a single feature map. Using 3D convolutions on a set of frames, on the other hand, produces a set of feature maps. The size of the temporal kernel and the strides employed determine the number of feature mappings.

Recent advancements in the field of activity recognition have resulted in a variety of designs that can be used to extract spatio-temporal features. Instead of depending on a recurrent network to encode information from each time step, architectures that can directly offer temporal information are looked at. Features are extracted for the Transformer model using I3D (Inflated 3D) Convolutional Neural Networks for Activity Recognition (Bilkhu et al., 2019)

7.4 Evaluation Metrics

Accuracy as the performance metric to measure how accurate an algorithm is on a classification task. Since the dataset is balanced, Accuracy as the evaluation metric is used to measure the performance of the models. It is the number of paired samples divided by total samples.

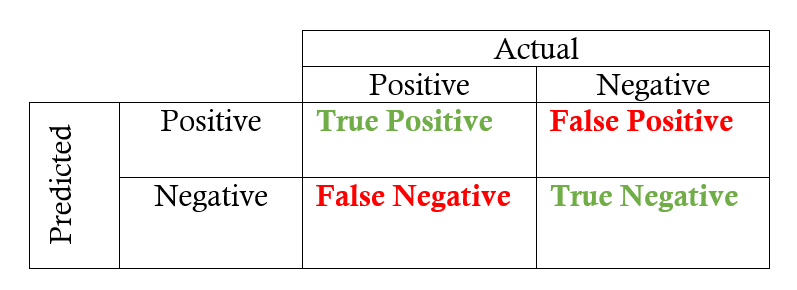


Figure 4: Confusion Matrix

Figure 4 shows the confusion matrix. Accuracy is the proportion of true results among the total number of cases examined.

Accuracy = (True Positive + True Negative) / (True Positive + False Positive + False Negative + True Negative)

Accuracy is a valid choice of evaluation for classification problems which are well balanced and not skewed or No class imbalance.

Top-1 Accuracy is the conventional accuracy, model prediction (the one with the highest probability) must be exactly the expected answer. It measures the proportion of examples for which the predicted label matches the single target label.

Top-5 Accuracy means any of the model’s top 5 highest probability answers matches with the expected answer. It considers a classification correct if any of the 5 predictions matches the target label.

8. Required Resources

The research will need below hardware and software resources throughout the implementation.

8.1 Software Requirements

Operating System: Ubuntu/Mac OS/Windows

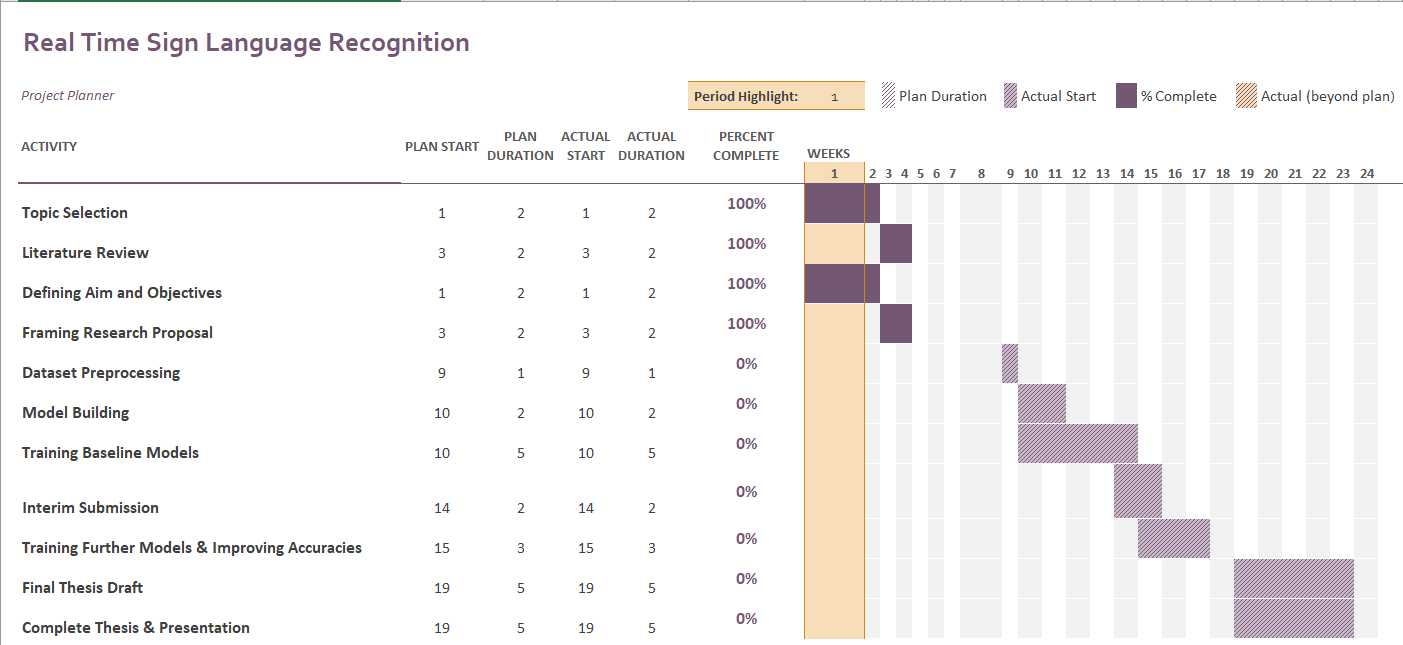
* Programming Language: Python 3.9.1, Shell Script
* Package Manager: pip
  + Python Libraries:
    - OpenCV
    - NLTK
    - Matplotlib
    - Numpy
    - CSV

#### Hardware Requirements

A laptop with below configuration will be used:

* SSD: 512GB
* Ram: 40GB (depends on batch size)
* Graphics: NVIDIA 2080 RTI, 12GB

9. Research Plan



References

Bahdanau, D., Cho, K. and Bengio, Y., (2014) Neural Machine Translation by Jointly Learning to Align and Translate. [online] Available at: http://arxiv.org/abs/1409.0473.

Bilkhu, M., Wang, S. and Dobhal, T., (2019) Attention is all you need for Videos: Self-attention based Video Summarization using Universal Transformers. [online] Available at: http://arxiv.org/abs/1906.02792.

Camgoz, N.C., Koller, O., Hadfield, S. and Bowden, R., (2020) Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. [online] Available at: http://arxiv.org/abs/2003.13830.

Carreira, J. and Zisserman, A., (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. [online] Available at: http://arxiv.org/abs/1705.07750.

Cui, R., Liu, H. and Zhang, C., (2019) A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. IEEE Transactions on Multimedia, 217, pp.1880–1891.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G., Cai, J. and Chen, T., (2015) Recent Advances in Convolutional Neural Networks. [online] Available at: http://arxiv.org/abs/1512.07108.

Huang, J., Zhou, W., Zhang, Q., Li, H. and Li, W., (2018) Video-based Sign Language Recognition without Temporal Segmentation. [online] Available at: http://arxiv.org/abs/1801.10111.

Kreutzer, J., Bastings, J. and Riezler, S., (2019) Joey NMT: A Minimalist NMT Toolkit for Novices. [online] Available at: http://arxiv.org/abs/1907.12484.

Li, D., Opazo, C.R., Yu, X. and Li, H., (2019) Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. [online] Available at: http://arxiv.org/abs/1910.11006.

Schmidt, C., Koller, O., Bellgardt, M., Forster, J. and Ney, H., (2014) Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. [online] Available at: www.signspeak.eu.

Simonyan, K. and Zisserman, A., (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. [online] Available at: http://arxiv.org/abs/1409.1556.

Sincan, O.M. and Keles, H.Y., (2020) AUTSL: A Large Scale Multi-modal Turkish Sign Language Dataset and Baseline Methods. [online] Available at: http://arxiv.org/abs/2008.00932.

Sridhar, A., Ganesan, R.G., Kumar, P. and Khapra, M., (2020) INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition. In: MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, Inc, pp.1366–1375.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K., (2015) Sequence to Sequence -- Video to Text. [online] Available at: http://arxiv.org/abs/1505.00487.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H. and Courville, A., (2015) Describing Videos by Exploiting Temporal Structure. [online] Available at: http://arxiv.org/abs/1502.08029.

Zhang, J., Zhou, W. and Li, H., (2014) A Threshold-based HMM-DTW Approach for Continuous Sign Language Recognition. In: Proceedings of International Conference on Internet Multimedia Computing and Service - ICIMCS ’14. [online] New York, New York, USA: ACM Press, pp.237–240. Available at: http://dl.acm.org/citation.cfm?doid=2632856.2632931.