Title to be added

Karthik S K

**Research Proposal**

Liverpool John Moores University - Master’s in Data Science

MAY 2022

**Abstract**

Table of Contents

[LIST OF FIGURES 4](#_Toc102174208)

[LIST OF TABLES 4](#_Toc102174209)

[LIST OF ABBREVIATIONS 4](#_Toc102174210)

[1. Background 5](#_Toc102174211)

[2. Related Research 5](#_Toc102174212)

[3. Research Questions (If any) 8](#_Toc102174213)

[4. Aim and Objectives 8](#_Toc102174214)

[5. Significance of the Study 8](#_Toc102174215)

[6. Scope of the Study 8](#_Toc102174216)

[7. Research Methodology 8](#_Toc102174217)

[8. Requirements Resources 13](#_Toc102174218)

[8.2 Hardware Requirements 14](#_Toc102174219)

[9. Research Plan 14](#_Toc102174220)

[References 14](#_Toc102174221)

# LIST OF FIGURES

Figure 1.1: The training process for two stream I3D on Kinetics dataset 11

# LIST OF TABLES

Table 1: INCLUDE: Category Size 9

Table 2: INCLUDE: Key statistics 10

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| SL | Sign Language |
| SLR | Sign Language Recognition |
| SLT | Sign Language Translation |
| ISL | Indian Sign Language |
| EDA | Exploratory Data Analysis |
| CNN | Convolutional Neural Network |
| NN | Neural Network |

# 1. Background

# 2. Related Research

The research survey related to solving the problem of Sign Language Recognition (SLR) has been carried out and is summarized in this section. The hearing-impaired community around the world use some variants of 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). Isolated SLR recognizes word by word, whereas continuous SLR translates entire sentences.

The existing 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 sentence-to-sentence recognition problem into three stages - temporal segmentation of videos, isolated word/expression recognition, 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.

A novel end-to-end sequence-to-sequence model that can generate captions for videos is a relevant research area. Recurrent Neural Networks, specifically LSTMs which demonstrated state-of-the-art performance in image caption generation, is trained on video-sentence pairs and the model learns associating 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).

Various research groups and industry labs have developed open-source toolkits for Neural Machine Translation (NMT). These are based on open-source deep learning platforms. However, these tools are targeted only towards research groups with a solid background in machine translation and deep learning, along with experience in navigating, understanding, and handling large code bases. With all these, none of the existing NMT tools has been designed primarily for readability or accessibility for novices, which makes it challenging to understand how NMT is implemented, the features each toolkit implements and what toolkit can be chosen to code their own project as fast and simple as possible.

Joey NMT (Kreutzer et al., 2019) designed for novices provides clean, well-documented, minimum code which is of comparable quality to more complex codebases on standard benchmarks. The core features include 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.

Sign languages are under-resources with corpora recorded typically for linguistic research and this does not provide type/token ratios needed for statistical natural language processing. This data significantly differs from real language encountered outside the research lab. RWTH-PHOENIX-Weather 2012 corpus for German sign language (DGS) recorded and annotated real life sign language footage aired in 2009 and 2010 by the German public TV station “PHOENIX” in the context of weather forecasts. This corpus has been significantly extended by adding 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 the hands and face of a signer for over 40,000 video frames along with annotations for hand shapes and orientations on the frame level (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. With deep neural networks, a framework for continuous sign language recognition transcribes videos of sign language sentences to sequences of ordered gloss labels. Deep Convolutional Neural Networks (CNN) with stacked temporal fusion layers for feature extraction modules, 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.
* New baseline results to guide future research in this field.

Indian Sign Language (ISL) which is used by over 5 million deaf people in India is a complete sign language with unique linguistic attributes with one grammar, syntax, and vocabulary. To carry out and evaluate Sign Language Recognition (SLR) approaches, currently there is no publicly available dataset on ISL. Earlier efforts in creating datasets for ISL include video datasets and image datasets with a single image denoting a sign. Since none of these datasets are in the public domain, any further study is a challenge. Other limitations are with the number of classes i.e., the number of distinct word signs being significantly low, and images being cropped so that only certain portions (only hands) 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 size and quality expected for Machine Learning (ML) research are not available for ISL dataset.

The lack of public datasets and scalable models for ISL recognition has been addressed by proposing the first publicly available ISL dataset called INCLUDE - Indian Lexicon Sign Language Dataset. The dataset contains 0.27 million frames with 263 classes from 15 different word categories. There are videos recorded with the help of deaf community, with the background, resolution and lighting chosen to resemble real world scenarios. A smaller subset named INCLUDE-50 with 50 classes was also proposed for rapid evaluation. The deep learning pipelines for SLR on INCLUDE dataset included data augmentation, feature extraction with pre-trained networks, using encoding and decoding. The feature extraction combined with pre-trained pose detection network, encoding with pre-trained MobileNet network, and decoding with trained bidirectional LSTMs has an accuracy of 94.5% on INCLUDE-50 and 85.6% on INCLUDE (Sridhar et al., 2020).

# 3. Research Questions (If any)

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

# 4. Aim and Objectives

The main aim of this research is to propose a ………………………………. The identification of the breast cancer ………….

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

* To analyze the pattern and relationship between the risk factors ………….
* To suggest a suitable balancing technique …………..
* To compare between the predictive models ……………..
* To evaluate the performance of ……………….

# 5. Significance of the Study

# 6. Scope of the Study

# 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 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 |
| 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

|  |  |
| --- | --- |
| **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 & ViVit. 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.

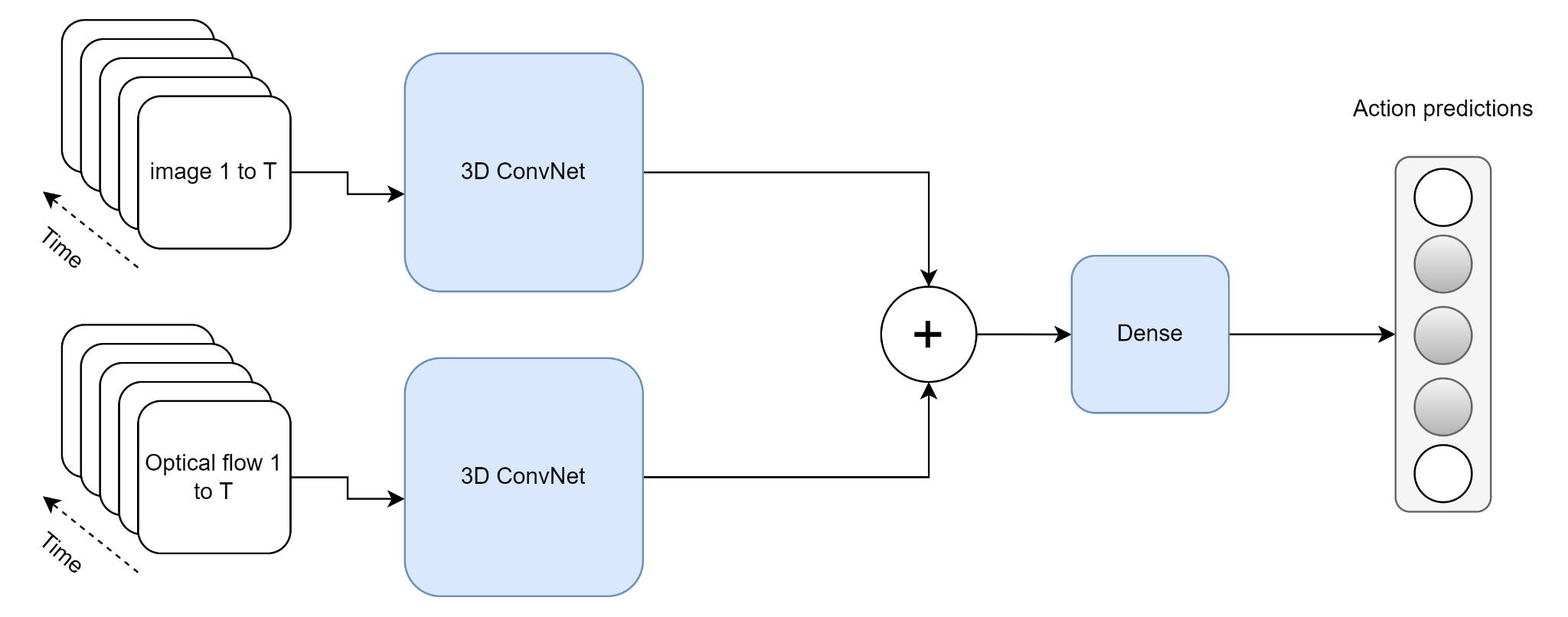


Figure 1.1: The training process for two stream I3D on Kinetics dataset

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 Models**

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

Fig 2.1 shows the architecture of CNN with Softmax output in the last layer.

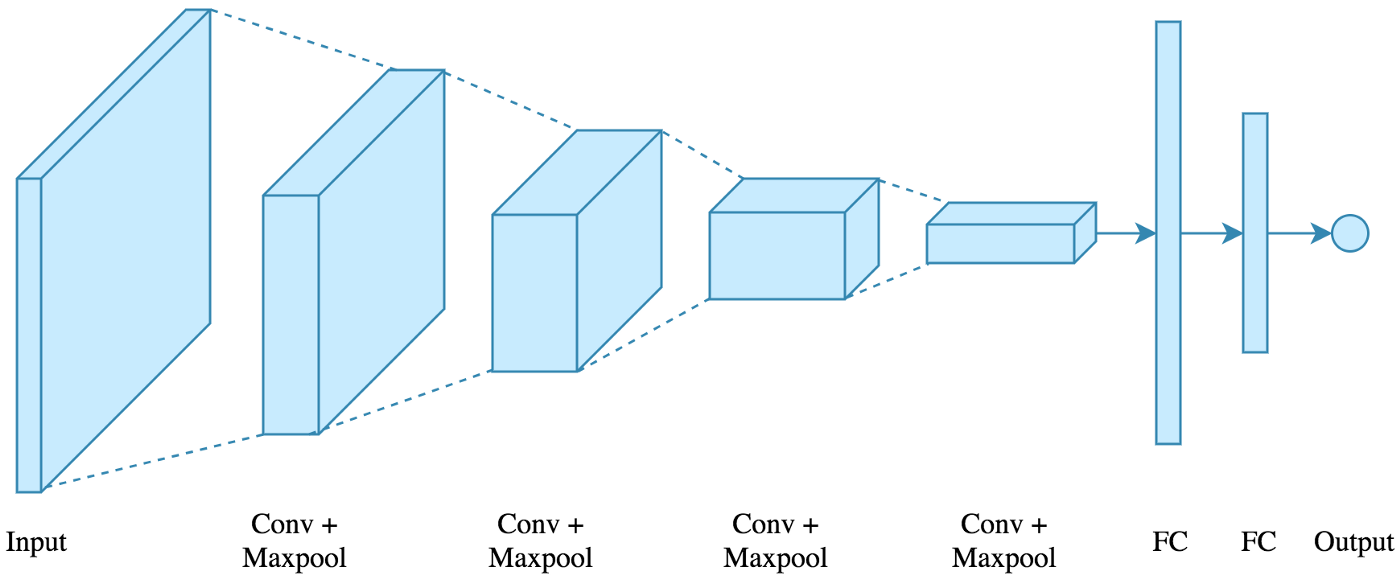


Figure 2.1: CNN architecture with Softmax output

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 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**

Data Preprocessing :

1. Video Data : VGG Feature Extraction, I3D Feature Extraction & ViVit
2. Text Data :

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 + Transformer’s Encoder + GRU based Decoder

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

Discuss Activation Function : Softmax, Leaky Relu

Write Short Notes around 1 or 2 Paragraph for each of the topics 3D CNN, VGG16, GRU, I3D, ViVIT.

# 8. Requirements 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

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.

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.

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.

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.

**Refer: Harvard Referencing Guide**