Warwick University

CS907 Dissertation Project

Interim Report

Mar Galiana Fernández

Supervised by: Dr Ahir Bhalerao

Abstract

This report presents the progress of the sign language recognition system developed for the dissertation project. The goal of the system is to convert American Sign Language gestures into written English text using videos of sign language interpreters as input. The report includes an overview of previous research in this field, details about the methodology used, and information about the completed experiments. It also outlines the planned next steps for the project.

12th July 2023

Contents

1 Introduction

This chapter gives an overview of the techniques used for Sign Language Recognition. It describes the aim of the thesis and the motivation behind it.

1.1 Overview

Sign Language is the communication tool used for those having a speech or hearing impairment. There is no universal sign language, to be precise, there are almost 140 according to the Ethnologue [1], and nearly every country has its own national sign language and finger-spelling alphabet. Hence there is a need for systems capable of recognising the sign gestures and conveying the message to the population that does have no knowledge of it. These systems are called Sign Language Recognition (SLR). In this report, the main objective will be to evaluate the state-of-the-art, compare it with our results and identify the next steps.

1.2 Motivation

Sign languages are not studied in school and only a tiny part of the world's population is proficient in them. This causes a problem when a person with a speech or hearing impairment tries to communicate an emergency or an everyday task. Nowadays, people are more aware of this situation and attempt to solve it by learning about them. However, at the moment, it is not enough to build a society that includes people living with these disabilities. [2, 3]

This research project aims to accelerate this process in order to reach this state of inclusivity sooner.

1.3 Aim of the Thesis

The aim of this thesis is to develop sign language recognition based on American Sign Language (ASL). We have chosen ASL as it is the one for which the most data is available. The inputs of the system to implement will be videos of an interpreter and the texts corresponding to each sign will be the outputs. We have decided to implement this system using videos as input data due to the fact that these can be obtained using a camera, an easy and economical

device. As we will see in section 2, most of the existing research has been done using expensive and complicated to use sensors. Our mission is to help with the communication process for those people who cannot communicate with verbal language, so it is important to try to minimise and facilitate as much as possible the use of the system to be developed.

In this project, existing Sign Language Recognition systems will be studied and evaluated in order to find the most promising one according to our requirements, which are explained in section 3. We will test several data pre-processing techniques, algorithms, optimisations and evaluation metrics, in order to come up with a proposal for an improved SLR system.

The report is organized as follows: Section 2 provides an overview of past research and advancements made in the same fields. In Section 3, the progress made so far and the results of the conducted experiments are presented. Section 4 outlines the upcoming steps and future directions of the project. Subsequent sections discuss the limitations, ethical considerations, and project management aspects. The report concludes with Section 8, which presents the achieved conclusions.

2 Background and research

Various techniques have been used in Sign Language Recognition (SLR) based on system requirements. Wadhawan and Kumar [4] summarised the rates of different papers on American Sign Language (ASL) training models. They compared the output rates based on five characteristics:

- Data Acquisition: Different methods were used to acquire data, including cameras and sensors. Cameras provided either images or videos, while sensors included gloves, Kinect, arm sensors, electroencephalogram, and leap motion.
- Type of Signs: Signs can be static (no movement, e.g., ASL alphabet) or dynamic (requiring movement for interpretation). Static signs often used camera-acquired images, while dynamic signs used videos.
- Modelling Algorithm: Various algorithms and techniques were employed. Previous papers using cameras and dynamic signs used Dynamic Time Warping (DTW) and Support Vector Machine (SVM).

Wadhawan and Kumar [4] also discuss other aspects regarding the dataset used to train the model. The main characteristics are:

- Interpreter mode. Comparing whether it is either isolated, continuous or both. Isolated signing refers to independent signs without any connections to preceding or succeeding signs. Continuous signing involves several signs without distinct pauses between them.
- Number of hands needed to interpret the signs. It can be single- or double-handed.

Table 1 summarises the above information. It is taken from the article by Wadhawan and Kumar [4], although only the information relevant to this article has been retained. Only the articles training the model with data containing dynamic signs or both (dynamic and static) are presented in the table.

Paper	**Data Acquisition**	**Gestures**	**Technique**	**Rate**
Oz and leu [5]	Gloves	Dynamic	NN	95%
Sun et al [6]	Kinect	Dynamic	Latent SVM	86%
Sun et al. [7]	Kinect	Dynamic	Adaboost	86.8%
Jangyodsuk et al. [8]	Camera	Both	DTW	93.38%
	Kinect	Both	DTW	92.54%
Wu et al. [9]	Arm sensors	Dynamic	Decision tree	81.88%
			SVM	99.09%
			NN	98.56%
			Naïve Bayes	84.11%
Usachokcharoen et al. [10]	Kinect	Dynamic	SVM	95%
Savur and Sahin [11]	Arm band	Both	SVM	82.3%
Sun et al. [12]	Kinect	Both	Latent SVM	86%
Kumar et al. [13, 14]	Camera	Static	SVM	93%
		Dynamic	SVM	100%
Savur and Sahin [15]	Armband	Dynamic	SVM and ensemble learner	60.85%
AlQattan and Sepulveda [16]	Electroence-phalogram	Dynamic	LDA	75%
			SVM	76%

Table 1: Summarised review of ASL recognition systems comparing the Reported Rate obtained. This table is located in the Sign Language Recognition Systems: A Decade Systematic Literature Review article [4].

From table 1, the most remarkable papers for this report are the ones from Jangyodsuk et al. [8] and Kumar et al. [13, 14]. Both of these researchers have used cameras as the data acquisition method and the dataset used to train the model uses dynamic signs.

Kumar et al. [13, 14] developed a sign language recognition system for recognising both static and dynamic signs in American Sign Language. They focused on predicting the letters a-z (where only the letters j and z are dynamic). The system was able to perform dynamic backgrounds with minimal decorations, as it relies on skin colour segmentation to identify gestures. Since the signs they wanted to predict do not use a facial expression, they removed this section of the video using Viola-Jones face detection followed by subtraction of the detected region. Once the data was processed, they ex-

tracted a curved feature vector, following previous work from Bhuyan et al., [17]. Afterwards, these feature vectors were classified using pre-trained SVM classifiers. Static and dynamic gestures were differentiated by measuring the distance travelled by the hand in subsequent frames. Dynamic gesture recognition was performed using four gestures for testing, which are: "j", "z", "no", and "goodbye", achieving an accuracy of 100%.

Jangyodsuk et al. [8] employed a dataset consisting of videos taken with a standard RGB camera, with three signers each making 1,113 signs, for a total of 3,339 different signs. They followed previous work by Dalal et al. [18], who applied the Dynamic Time Warping (DTW) method to compare the hand trajectory using a Histogram of Oriented Gradient (HoG) to represent hand shape. However, they standardised the features to have a mean value of 0 and a standard deviation of 1. With this improvement, their accuracy increased, on average, by about 10%.

Using Hand trajectory matching with hand shape distance using HoG features as shape representation, they archived an accuracy of 93.38

There exist other papers published after the Wadhawan and Kumar article [4] was released which have other interesting methods applied.

Borg and Camilleri [19] implemented, in 2020, a two-stage system, which has the hand key points features obtained via OpenPose as the input to the model. They use Hidden Markov models (HMMs) to obtain the subunits of sign concatenation (SU). The SU descriptors are employed to train the SU Recurrent Neural Networks (RNNs), using a Connectionist Temporal Classification (CTC) framework to handle the temporal sequence. Once the SU RNNs are trained, a second-level RNN is added for sign recognition. RWTH-Phoenix Weather [20] was the dataset used which contained 1230 unique signs with 9 signers. With this implementation, they archived an accuracy of 71.9%.

Zheng et al. [21] proposed a model that reduced the training size data by 9.3%, compared with the state-of-the-art of 2020. This model used the Frame Stream Density Compression (FSDC) algorithm to detect and reduce redundant similar frames, which shortens long sign sentences without losing information. They further implemented a temporal convolution (T-Conv) connected to a dynamic hierarchical bidirectional GRU (DH-BiGRU). The RWTH-Phoenix Weather Dataset [20] was used, which is the same as the one utilised by Borg and Camilleri [19].

Al-Hammadi et al. [22] used a 3D Convolutional Neural Network (3DCNN) where three instances of the 3DCNN structure were trained to extract the hand gesture features from the beginning, middle, and end of the video sample. Afterwards, they studied three techniques for feature fusion: multilayer perceptron (MLP) neural network, long short-term memory (LSTM) network, and stacked autoencoder. Using the 3DCNN and MLP in a 40 classes dataset called KSU-SSL dataset, they reached a recognition rate of 84.38%.

Finally, Hassan et al. [23] conducted an experiment using various models, with the most successful one being the SlowFast Neural Network. They utilized a pre-trained model called *SLOWFAST_8×8_R50* obtained from the PySlowFast GitHub repository [24]. The experiment utilized the WLASL [25] dataset and achieved predictions for 300 different labels, resulting in a TOP 1 accuracy of 79.34%, which implied an improvement of 23.2% over the previous state-of-the-art performance. The researchers mentioned that the limitations they encountered were related to the time-consuming nature of model training, which required a total of twenty-four hours spread across multiple days, as well as some hardware limitations.

Even though there is intensive research conducted on gesture recognition, the majority of them have been executed using data coming from complex hardware and inconvenient for the user to carry on a daily basis. Research so far indicates that the architectures being used are vastly differing, therefore it is necessary to bring all the information together and try to get the best out of each one of them.

3 Progress

3.1 Proposed Idea

This project aims to develop a sign language recognition system based on ASL. The system will take videos of individual interpreters performing ASL gestures as input and provide translations of these signs into English text as output. Throughout the project, a series of experiments will be conducted, incorporating existing studies and novel approaches to enhance the system. These experiments can be categorised into four phases: data selection, data processing, model selection, and optimisation.

Two experiments have been conducted, utilising the same dataset but employing different models and data processing techniques. The first experiment utilises a 3D Convolutional Neural Network (3DCNN), while the second experiment employs a SlowFast Neural Network. Each experiment will be elaborated upon in the following sections, aligned with the established four phases.

The 3DCNN model was chosen due to its proven high performance in the state-of-the-art literature. In contrast, the SlowFast Neural Network incorporates two channels: one dedicated to detecting the slow frequencies of the video, and the other channel focuses on capturing the high frequencies. This approach has demonstrated significant utility in action recognition [24] as well as Sign Language Recognition [23]. In the context of sign language interpretation, videos of interpreters primarily exhibit movement in specific regions, which can be efficiently processed through the fast channel. Therefore, we have high expectations for the promising results that can be achieved through the implementation of this technique.

3.1.1 Dataset selection

The model will use already existing labelled datasets. The data acquisition is out of the scope of this project. We are going to focus on building the predictor model system, not on the process of data acquisition and data labelling. The videos from the datasets that are going to be used will contain dynamic signs from the ASL.

The following datasets have been examined in order to determine which one is going to be the most accurate to the project requirements.

- WLASL [25]: it is the largest video dataset for Word-Level American Sign Language (ASL) recognition.
- The American Sign Language Lexicon VideoDataset [26].
- RWTH-Phoenix Weather Dataset [27]
- Datasets from the Kaggle community.[1]

Among these options, the WLASL dataset provides a higher variety of backgrounds, speaker speeds, physique, and camera orientations. Additionally, it features 2,000 common words in ASL. Therefore, the WLASL dataset is the most suitable choice for this project.

The experiments aim to gain valuable insights into model behaviour and determine effective preprocessing methods. To achieve this, the initial experiment focuses on a subset of the dataset. Both experiments used a selection of ten random labels from the WLASL dataset was chosen. These labels include words such as drink, trade, before, bowling, computer, cool, go, thin, help, and tall. To ensure a sufficient sample size for each label, eight videos were allocated for training, while two videos were reserved for validation and two for testing purposes. By examining this subset of labels, we were able to analyse the model performance and evaluate the impact of the chosen preprocessing techniques.

3.1.2 Data processing

Different types of data processing will be tested to help the proposed model generalise and predict signs more precisely.

In the first experiment, which utilized a 3DCNN model, several tests were conducted to determine the relevant portion of the sign interpretation within the video. For each video, 20 consecutive frames were extracted, with variations in the starting point. These frames were obtained from the beginning, middle, and end of the video, as well as randomly selected positions. This approach helped reduce the data size by 10 frames per video while capturing different sections of the sign interpretation.

[1] https://www.kaggle.com

In the second experiment, which utilised a SlowFast Neural Network model, the essential features of the images were enhanced to prioritise the interpretation of Sign Language within the interpreter's body. As stated in section 2, other studies, such as Jangyodsuk et al. [8], Borg and Camilleri [19], and Al-Hammadi et al. [22], have also adopted a similar approach to detect the hand's position in the images.

As explained in section 1, the face is also used when interpreting ASL. This is the reason why the three enhanced features of each video have been the face, the left hand, and the right hand. Each frame of every video was further processed, as depicted in figure 1, which illustrates an interpreter interpreting the ASL sign for the word "secretary". The left picture displays the original frame, while the one on the right showcases the processed frame that will be fed into the model. The resulting image consists of four sections: the top left section displays a zoom-in of the left hand, and the top right section does the same for the right hand. The face is shown in the bottom-left part of the image, and in the bottom-right part, the entire image is displayed with a subtle blur applied. This blur aims to emphasise the other features without losing the overall context.

Figure 1: Representation of how the videos are processed in Experiment 2.

Several libraries were employed to detect the hands and faces in the original images. The initial library used was OpenPose, as also utilized by Borg and Camilleri [19]. However, the results obtained from OpenPose were not accurate enough, necessitating the search for an alternative library. Eventually, Mediapipe was chosen for hand detection and distinguishing between

the left and right hands. Although there may still be some errors in certain frames, there has been a significant improvement compared to OpenPose.

For face detection, a library called face_recognition, built using dlib's state-of-the-art techniques, is being utilized. The accuracy of this library is remarkable, delivering reliable results in face-detection tasks.

To ensure diversity in our dataset and capture a wide range of real-life scenarios, future experiments will incorporate the following data processing techniques. By introducing greater variety, we can mitigate the risk of overfitting our model. Some of the insights we aim to explore include:

- Crop the image in such a way that the performers are placed in distinct locations within the image.
- Modify the speed of the video to have different speeds in the way the performers gesticulate. Speed modification in a video can be archived by removing certain frames for each constant number of frames, for example, removing two frames every six frames. Another option could be not to use a linear function when deciding which frames to remove. Using the first experiment, we will determine where the most important part of the gesture is located in an ASL sign, so if it is at the beginning, we could use an exponential function to remove more frames from the end of the video, or, otherwise, use a logarithmic function. We can also combine the different frequencies to have a larger variance in the speed of all the videos in the dataset.

Reducing the memory footprint of our models is an additional objective when processing data. Video datasets consume a substantial amount of memory in our system. The following two techniques can help address this challenge:

- Removing the initial and final frames of videos to reduce memory requirements during model training.
- Selectively removing frames within the videos using linear, logarithmic, or exponential functions, as mentioned earlier, to further alleviate memory costs.

Furthermore, we aim to establish a signer-independent mode by ensuring that the training and testing sets do not contain the same interpreter. Random splitting of the data is referred to as signer-dependent mode [22].

3.1.3 Model selection

After analyzing the experiments mentioned in the previous paper (Section 2), several promising techniques have emerged, including:

- Support Vector Machine [13, 14].
- Dynamic Time Warping using Histogram of Oriented Gradient [8].
- Hidden Markov models combined with Recurrent Neural Network [19].
- Temporal convolution network connected to a dynamic hierarchical bi-directional GRU. [21].
- 3D Convolutional Neural Network using a multilayer perceptron for feature fusion. [22].
- Slow Fast Neural Network [23].

Based on the results obtained from these techniques, we have chosen to employ a 3DCNN model for the first experiment. Specifically, we have utilized the pre-trained MoViNet-A0-Base model from TensorFlow [2]. This model has been trained on the kinetics dataset, achieving a TOP 1 accuracy of 72.28% with an input size of 50x172x172 pixels.

The second experiment used a SlowFast Neural Network pre-trained in the paper from Damen et. al. [24]. This model archived a TOP 1 accuracy of 77% with the K400 dataset using a frame length x sample rate of 8x8.

3.1.4 Optimisation

Upon optimising the hyper-parameters of the first experiment, we obtained the results presented in Table 2.

Test	**Frames position**	**Accuracy**
1	Beginning	40%
2	Middle	40%
3	End	20%
4	Random Start	40%

Table 2: Results of the second experiment after optimising the model

[2]https://github.com/Atze00/MoViNet-pytorch

The model's accuracy decreased by 20% when utilising the final frames of the video, suggesting that the crucial part of the video is found in both the beginning and middle sections. These sections are equally important in capturing the essential information required for accurate interpretation.

As the second experiment is currently under development, the results are not yet available. However, we can leverage the findings from the first experiment to optimise the dataset size and reduce the training time by removing the last frames of each video. This approach will help mitigate the training timing cost and streamline the experimentation process.

The progress of the experiments is being tracked in a GitHub Repository. This repository will be made publicly available at the conclusion of the dissertation project, providing transparency and insight into the training processes of each model. By sharing this information, others will have the opportunity to understand and replicate the training procedures employed throughout the project.

4 Plan

Our objective is to develop a system that can accurately recognize the dynamic gestures of American Sign Language while optimizing computational costs. Throughout the development process of this dissertation project, we have the following required submissions:

- Dissertation specification: Submitted on the 2nd of March. The current document serves as the dissertation specification.
- Dissertation presentation: Scheduled for the 5th of May.

Moving forward, two reports need to be submitted, including the current report and the final dissertation report:

- Interim report: Due on the 13th of June.
- Dissertation report: Due on the 7th of September.

From now until the final submissions, our focus is on implementing the following key points:

- Obtain results from the SlowFast Neural Network: Increase the number of labels and data samples while optimising them to compare with the state-of-the-art approaches.
- Identify the most influential features: Through the second experiment discussed in section 3, we aim to determine which features contribute the most to accuracy and which ones do not provide significant information to the model. For instance, if the face is found to be less relevant, we can de-prioritise facial recognition to allocate more importance to other features. Different combinations of data processing will be evaluated to validate this.
- Apply data augmentation techniques: As described in section 3.1.2, data augmentation will be employed to help the model generalize better by introducing variations in the training dataset.
- Evaluate the trained model with other Sign Languages: We will retrain the model using appropriate datasets from different sign languages to assess if the same model architecture can be applied across multiple languages. This approach will prevent us from being limited to a single sign language.

- Discuss the potential of Video Generative Adversarial Networks (VGANs): We will explore how VGANs, based on the current state-of-the-art and considering the model's requirements and constraints, can contribute to data augmentation and further enhance the model's performance.

By addressing these points, we aim to improve the accuracy, generalisation, and language adaptability of the developed system. In addition to the ongoing work, we will also prioritize the final report to document all the conducted experiments and the corresponding results. This comprehensive report will provide a detailed account of the methodology employed, the findings obtained, and the conclusions derived from our research and experimentation.

5 Appraisal and reflection

During the course of the mentioned experiments (see 3), several challenges have been encountered. Initially, hardware limitations necessitated the use of university-provided servers for implementing the experiments. However, each student had limited space allocated, and since the dataset primarily consisted of videos (which occupy more storage compared to images), it posed a challenge. After multiple communications with the university's IT team, they increased the storage capacity assigned to me, enabling the training of models on the remote server.

Another challenge arose while working with pre-trained models, as not all research papers provided comprehensive details on the model architecture and expected input data format. This made it challenging to retrain the models and accurately understand the required input data format. It required multiple attempts and extensive investigation to achieve successful training. However, debugging code on the remote server is complex due to limited accessibility.

In addition, it is often challenging to locate pre-trained models as they are not always readily available. Many research papers do not provide specific details or direct links to the pre-trained models they utilised. Luckily, when a paper employs a pre-trained model, they typically cite the source from which it was obtained. This citation serves as a helpful reference, facilitating the search for suitable pre-trained models that are accessible for download.

Despite these challenges, proactive measures were taken to overcome them and ensure the smooth progression of the experiments.

6 Ethics

The WLASL dataset [25] employed in this study involves secondary analysis of publicly available data. The dataset creators explicitly state that "all the WLASL data is intended for academic and computational use only" [25]. Since this report utilises the dataset for academic purposes and properly cites it, there is no violation of ethical consent.

7 Project Management

The project has been managed and monitored using a dedicated GitHub repository, which will be made accessible for public viewing upon the final submission of the dissertation project report. For implementation purposes, I have utilised my personal computer as well as the servers provided by the university to leverage additional computational capacity.

Supervision of the project is being conducted by Dr Ahir Bhalerao, with whom I have regular meetings every two weeks. These meetings serve as opportunities to review the progress, discuss new developments, analyse results, and plan the next steps of the project. These regular meetings with my supervisor ensure effective guidance and facilitate a smooth progression of the research work.

8 Conclusion

In conclusion, this dissertation project aims to develop an effective American Sign Language recognition system that utilises video input data in a signer-independent mode. The primary objective is to alleviate communication challenges faced by individuals with speech and hearing impairments.

Two experiments have been conducted thus far using the WLASL dataset. The first experiment involved utilising a 3DCNN model and testing the most relevant frames to understand where the crucial information lies within the signing process. The second experiment incorporates a SlowFast neural network, which detects hand movements and performs facial recognition to assign greater importance to the relevant features during model training.

Looking ahead, several further steps will be taken to enhance the system. These steps involve implementing various approaches, including testing different models to identify the optimal performance and experimenting with different types of data augmentation techniques to improve the model's generalisation capabilities.

By continuously refining and expanding the system, we strive to develop a robust and inclusive solution that empowers individuals with speech and hearing impairments to communicate more effectively through American Sign Language recognition.

References

[1] Alexey Karpov, Irina Kipyatkova, and Milos Zelezny. Automatic technologies for processing spoken sign languages. *Procedia Computer Science*, 81:201–207, 2016. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

[2] Barbara R Schirmer. *Psychological, social, and educational dimensions of deafness.* Allyn & Bacon, 2001.

[3] Annalene Van Staden, Gerhard Badenhorst, and Elaine Ridge. The benefits of sign language for deaf learners with language challenges. *Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer*, 25(1):44–60, 2009.

[4] Ankita Wadhawan and Parteek Kumar. Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, 28:785 – 813, 2019.

[5] Cemil Oz and Ming C. Leu. Linguistic properties based on american sign language isolated word recognition with artificial neural networks using a sensory glove and motion tracker. *Neurocomputing*, 70(16):2891–2901, 2007. Neural Network Applications in Electrical Engineering Selected papers from the 3rd International Work-Conference on Artificial Neural Networks (IWANN 2005).

[6] Chao Sun, Tianzhu Zhang, Bingkun Bao, and Changsheng Xu. Latent support vector machine for sign language recognition with kinect. *2013 IEEE International Conference on Image Processing*, pages 4190–4194, 2013.

[7] Chao Sun, Tianzhu Zhang, Bing-Kun Bao, Changsheng Xu, and Tao Mei. Discriminative exemplar coding for sign language recognition with kinect. *IEEE Transactions on Cybernetics*, 43(5):1418–1428, 2013.

[8] Pat Jangyodsuk, Christopher Conly, and Vassilis Athitsos. Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '14, New York, NY, USA, 2014. Association for Computing Machinery.

[9] Jian Wu, Zhongjun Tian, Lu Sun, Leonardo Estevez, and Roozbeh Jafari. Real-time american sign language recognition using wrist-worn motion and surface emg sensors. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–6, 2015.

[10] Panupon Usachokcharoen, Yoshikazu Washizawa, and Kitsuchart Pasupa. Sign language recognition with microsoft kinect's depth and colour sensors. *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 186–190, 2015.

[11] Celal Savur and Ferat Sahin. Real-time american sign language recognition system using surface emg signal. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 497–502, 2015.

[12] Chao Sun, Tianzhu Zhang, and Changsheng Xu. Latent support vector machine modeling for sign language recognition with kinect. *ACM Trans. Intell. Syst. Technol.*, 6(2), mar 2015.

[13] Anup Kumar, Karun Thankachan, and Mevin M. Dominic. Sign language recognition. *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, pages 422–428, 2016.

[14] D. Anil Kumar, Polurie Venkata Vijay Kishore, A. S. Chandrasekhara Sastry, and P. Reddy Gurunatha Swamy. Selfie continuous sign language recognition using neural network. *2016 IEEE Annual India Conference (INDICON)*, pages 1–6, 2016.

[15] Celal Savur and Ferat Sahin. American sign language recognition system by using surface emg signal. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 002872–002877, 2016.

[16] Duaa AlQattan and Francisco Sepulveda. Towards sign language recognition using eeg-based motor imagery brain computer interface. *2017 5th International Winter Conference on Brain-Computer Interface (BCI)*, pages 5–8, 2017.

[17] M.K. Bhuyan, D. Ghosh, and P.K. Bora. Feature extraction from 2d gesture trajectory in dynamic hand gesture recognition. In *2006 IEEE Conference on Cybernetics and Intelligent Systems*, pages 1–6, 2006.

[18] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer*

Vision and Pattern Recognition (CVPR'05), volume 1, pages 886–893 vol. 1, 2005.

[19] Mark Borg and Kenneth P. Camilleri. Phonologically-meaningful subunits for deep learning-based sign language recognition. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 199–217, Cham, 2020. Springer International Publishing.

[20] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

[21] Jiangbin Zheng, Zheng Zhao, Min Chen, Jing Chen, Chong Wu, Yidong Chen, Xiaodong Shi, and Yiqi Tong. An improved sign language translation model with explainable adaptations for processing long sign sentences. In *Computational Intelligence and Neuroscience*, volume 2020, page 11, 2020.

[22] Muneer Al-Hammadi, Ghulam Muhammad, Wadood Abdul, Mansour Alsulaiman, Mohamed A. Bencherif, and Mohamed Amine Mekhtiche. Hand gesture recognition for sign language using 3dcnn. *IEEE Access*, 8:79491–79509, 2020.

[23] Ahmed Hassan, Ahmed Elgabry, and Elsayed Hemayed. Enhanced dynamic sign language recognition using slowfast networks. In *2021 17th International Computer Engineering Conference (ICENCO)*, pages 124–128. IEEE, 2021.

[24] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020.

[25] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020.

[26] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The american sign language lexicon video dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

[27] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, volume 9, pages 3785–3789, 2012.