Warwick University

CS908 Research Methods

Dissertation Specification

Mar Galiana Fernández

Supervised by
Dr Ahir Bhalerao

Abstract

This report describes the specifications of the sign language recognition system to be implemented for the thesis project. The system to be developed will translate American Sign Language gestures into English text, using videos of an interpreter as input data. Includes previous research performed in this field along with the methodology that is being applied, together with details of the work plan to be followed.

2nd March 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 project, the main objective will be to study the ones that already exist and to try to enhance the most promising among them.

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 economic

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 organised as follows: section 2 describes the past research and advances made in the same fields. In Section 3 we describe the methodology that is going to be used, explaining the proposed idea and the work plan. In the last section, section 4, we give the conclusion of our proposal.

2 Related Work

Many techniques have been developed to recognise sign language. The decision of which technique to use is made depending on the requirements of the system to build. Wadhawan and Kumar [4] presented a summary table in 2019 of the reported rate of different papers which trained the model with American Sign Language (ASL). They compared the output rate of each modelling depending on five characteristics:

- The first important feature that Wadhawan and Kumar [4] used to distinguish between the architectures of each recognition system was the form in which the data was acquired. There are several options and we are not going to go through all of them, however, the most common techniques involved the use of cameras or sensors. When a camera was used, the input data could either be images or videos. Regarding sensors, there are a variety of different possibilities, in the article, they compare the models using: gloves, Kinect, arm sensors, electroencephalogram, and leap motion...

- Two types of signs can be considered: static and dynamic. When the prediction is done using static signs and a camera as a data acquisition tool, images are often the source data, for instance in the case of the article *Sign Language Recognition Using Convolutional Neural Networks* [5]. However, when the model is predicting dynamic signs, the input data used tends to consist of videos, as they have used in the article *Phonologically-Meaningful Subunits for Deep Learning-Based Sign Language Recognition* [6]. Wanting to differentiate this feature in the modelling architecture is relevant as the model that will be proposed in this project will use videos as input data to predict dynamic gestures. We will have to compare the reporting rate we are going to obtain with articles that have used the same technique as ours.

- What we are really interested in is the modelling algorithm and technique used in each previous report. We need to understand which one is producing the best results given a model that has been trained over the same type of data as the one we are going to use. From the list of papers analysed by Wadhawan and Kumar [4], the ones that use the camera and train the models with dynamic gestures used Dynamic Time Warping (DTW) and Support Vector Machine (SVM).

Wadhawan and Kumar [4] also discuss other aspects in previous articles, such as the interpreter mode, comparing whether it is isolated, continuous

or both and if the signs in the dataset are single- or double-handed. This information is not relevant to us at this point, as there is no single dataset that has been specified to be used.

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 [7]	Gloves	Dynamic	NN	95%
Sun et al [8]	Kinect	Dynamic	Latent SVM	86%
Sun et al. [9]	Kinect	Dynamic	Adaboost	86.8%
Jangyodsuk et al. [10]	Camera	Both	DTW	93.38%
	Kinect	Both	DTW	92.54%
Wu et al. [11]	Arm sensors	Dynamic	Decision tree	81.88%
			SVM	99.09%
			NN	98.56%
			Naïve Bayes	84.11%
Usachokcharoen et al. [12]	Kinect	Dynamic	SVM	95%
Savur and Sahin [13]	Arm band	Both	SVM	82.3%
Sun et al. [14]	Kinect	Both	Latent SVM	86%
Kumar et al. [15, 16]	Camera	Static	SVM	93%
		Dynamic	SVM	100%
Savur and Sahin [17]	Armband	Dynamic	SVM and ensemble learner	60.85%
AlQattan and Sepulveda [18]	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].

The papers we are most interested in from table 1 are the ones from Jangyodsuk et al. [10] and Kumar et al. [15, 16]. Both of these researchers

have used cameras as the data acquisition method and the data used to train the model consists of dynamic signs.

Kumar et al. [15, 16] 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 extracted a curved feature vector, following previous work from Bhuyan et al., [19]. 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 (j, z, no, goodbye) achieving an accuracy of 100%.

Jangyodsuk et al. [10] 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. [20], 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 [21] implemented, in 2020, a two-stage system, which has as the input the hand key points features obtained via OpenPose. Then, 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 [22] 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. [23] 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 [22] was used, which is the same as the one utilised by Borg and Camilleri [21].

Finally, Al-Hammadi et al. [24] 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%.

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 of them.

3 Methodology

3.1 Proposed Idea

The aim of this project is to devise a sign language recognition system based on American Sign Language. The input data will be videos of individual interpreters performing an ASL gesture and the output will be the translation of these signs into English text. During the course of the project, we will conduct several experiments using previous studies, combinations of them, and new approaches that we believe can bring improvement to the system. These tests could be divided into the following four phases: data selection, data processing, model selection and optimisation.

3.1.1 Dataset selection

The choice of the datasets. We are going to 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 we are going to use will contain dynamic signs from the ASL.

Various datasets have been used in some of the above-mentioned works that we will try to include in our tests:

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

Our idea is to combine several datasets to obtain the most realistic cases and test our model with different situations. We will seek to have a variety of image backgrounds, the speaker's speed and physique, and camera orientation... We will be able to perform all these tests by selecting the right datasets and joining them together.

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

3.1.2 Data processing

Different types of data processing will be tested to help the proposed model predict signs more precisely. Below are the methods we have already discussed and which we expect will improve our model.

To begin with, we want to have diversity in our dataset in order to be able to capture most of the cases that will be encountered in reality. Besides, the more variety we archive, the less likely we are to overfit our model. Some of the insights we wanted to develop are:

- We will also try to crop the image in such a way that the performers are placed in distinct locations within the image. We will look for variety so that the performer is placed randomly between the left, right and centre of the video.

- Modify the speed of the video to have different speeds in the way the performers gesticulate. We can use several techniques and apply them to certain videos in the dataset, aiming to have more variety in the dataset to prevent overfitting. 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. We would have to 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.

An additional goal we aim to archive when processing data is to reduce the memory cost of our models. Datasets containing videos occupy a large amount of data in our system, the application of the following two techniques could help reduce this problem.

- Removing the initial and final parts of the videos will reduce the amount of memory needed to train the model. These parts of the video do not contain important information about the sign, as the performer has not yet started to gesture.

- Removing some of the frames within the video using either a linear, logarithmic or exponential function, as mentioned above, would also reduce the memory cost of our proposal.

In addition, we want to have a signer-independent mode, which consists of splitting the training and testing sets without including the same interpreter in both sets. If the splitting is made randomly it is called signer-dependent mode. [24]

3.1.3 Model selection

We will implement our model using different techniques in order to test which ones give us the best performance given our requirements defined in section 3.1.

Once we have studied the related work, explained in section 2, we know that the most promising techniques are:

- Support Vector Machine. [15, 16]
- Dynamic Time Warping using Histogram of Oriented Gradient. [10]
- Hidden Markov models combined with Recurrent Neural Network. [21]
- Temporal convolution network connected to a dynamic hierarchical bi-directional GRU. [23]
- 3D Convolutional Neural Network using a multilayer perceptron for feature fusion. [24]

After testing these techniques, we should be able to identify which ones appear to be the most promising.

3.1.4 Optimisation

After selecting the most promising techniques, we will focus on optimising them in order to assess whether we can improve their current results. One way we could optimise them is by testing different data processing and combining the different experiments performed in this research work.

In addition, we also want to test the SlowFast network technique, which has been used in several video recognition systems with very promising results. But it does not seem to have been used yet in sign language recognition systems. We are interested in discovering whether this method could improve the existing ones.

3.2 Work Plan

We aim to develop a system capable of recognising the dynamic gestures of American Sign Language while optimising the computational cost. During the development process of this dissertation project we will have the following required submissions:

- Dissertation specification on the 2nd of March. This submission is the current document.
- Dissertation presentation on the 5th of May.
- Interim report on the 13th of June.
- Dissertation report on the 7th of September.

For this purpose, we have drawn up a work plan divided into three phases.

3.2.1 First Phase

The first phase has already started and is intended to finalise with the dissertation presentation. The aim of this phase is to experiment with the different pre-existing models with the main purpose of being able to introduce which models seem to be the most promising at the presentation.

We will start working with smaller versions of datasets and pre-trained classifiers to avoid spending an excessive amount of time training the models.

3.2.2 Second Phase

In the second phase, we intend to increase the dataset, as we will be focusing on the most promising models. With the additional data, we will identify the best settings, configurations and architectures for the selected models.

We plan to start this phase after the exam period, as we are aware that combining exams with the development of this phase is complicated. Our intention is to finish this stage by the end of July, in order to allow enough time for the implementation of phase three.

3.2.3 Third Phase

In the third phase of this project, we plan to investigate and experiment with the SlowFast network model to discover whether it performs better than the current implementations in the research mentioned above under the section 2. We plan to finish this section by mid-August so that we allow ourselves time to refine the dissertation report which is due in September.

We realise what an ambitious plan we have, however, if we are capable of pursuing it, we will develop an improved sign language recognition system.

4 Conclusion

In this dissertation project, we will develop an American Sign Language recognition system using videos as input data in a signer-independent mode. The system we will implement is an approach to overcome the communication difficulty faced by those who suffer from speech and/or hearing impairments.

Most of the current proposals use high-tech sensors to implement this kind of system. This creates a disadvantage when it comes to using these systems in the user's daily life, apart from the fact that they are expensive, they lack practicality. For this reason, we are focusing on implementing a system which obtains data through a camera, a more practical method for the user.

We will test different approaches to determine which model offers the best performance while trying to reduce the computational cost of it.

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] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign language recognition using convolutional neural networks. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, pages 572–578, Cham, 2015. Springer International Publishing.

[6] Mark Borg and Kenneth P. Camilleri. Phonologically-meaningful subunits for deep learning-based sign language recognition. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, page 199–217, Berlin, Heidelberg, 2020. Springer-Verlag.

[7] 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).

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.