Automatic Irish Sign Language Recognition

Irene Hernández B.Sc.

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Augmented and Virtual Reality)

Supervisor: Aljoša Smolić

August 2018

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Irene Hernández

August 26, 2018

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Irene Hernández

August 26, 2018

Acknowledgments

I would like to express my gratitude to Prof. Aljoša Smolić, Tejo Chalasani, Sebastian Lutz, and Koustav Ghosal, for their time, wisdom and dedication. Thank you to my friends and classmates: it has been a pleasure. Honourable mention to the people on the fourth floor at Phoenix House, for respecting my little corner. A big thank you to my family, for their constant support and dad jokes. I feel I must also thank The Smiths and The National, for providing a soul-wrenching soundtrack.

Irene Hernández

University of Dublin, Trinity College
August 2018

Automatic Irish Sign Language Recognition

Irene Hernández, Master of Science in Computer Science

University of Dublin, Trinity College, 2018

Supervisor: Aljoša Smolić

A deep learning approach to automatic Irish sign language recognition has not been thoroughly examined before. Two frameworks are introduced in this work: a convolutional neural network to extract spatial features and classify the 23 letters of the Irish sign language alphabet that do not incorporate motion, and a convolutional neural network with a multi-stream input channel and three-dimensional filters in order to detect spatio-temporal features and classify 8 dynamic gestures in Irish Sign Language.

Summary

This dissertation presents a framework for automatic Irish sign language recognition. The objective is the evaluation of performance of deep learning algorithms on Irish Sign Language data towards automatic classification of signs. Automatic sign language recognition is an area of research with a wide scope that encompasses topics such as hand posture recognition, motion trajectory modelling, or sign sequence segmentation. This work addresses sign recognition from single frames through the design of a neural network model that achieves an accuracy of 99.87% for hand gesture recognition on Irish sign language data. Sequences of signs that extend over time to complete a trajectory are also studied: the proposed architecture produces an accuracy value of 79.66% in spatio-temporal Irish sign classification on a small set of Irish Sign Language videos that are not subject to concrete conditions. This document contributes to the field of sign language recognition with a comprehensive exploration of convolutional neural networks applied to Irish Sign Language data, an attempt at which has only been documented for non-spatio-temporal signs. In particular, this analysis aims attention at the type and administration of input data, especially when dealing with a small dataset, as well as the implications and handling of temporal information.

Glossary

ASL American Sign Language. 2, 4, 5, 9–11

AUC Area Under the Curve. ix, 30, 31

CNN Convolution Neural Network. 7, 8, 10, 11, 15, 16, 19, 34, 37

EAF ELAN Annotation Format. 13

HMM Hidden Markov Model. 36, 37

ISL Irish Sign Language. 1, 2, 5, 6, 11–13, 15, 28–31, 33–37

LSTM Long Term Short Memory. 17, 19

PCA Principal Component Analysis. 5

ReLU Rectified Linear Units. 11, 16, 27

RNN Recurrent Neural Network. 17, 37

ROC Receiver Operating Curve. ix, 30, 31

ROI Region Of Interest. 34

SGD Stochastic Gradient Descent. 27

SVM Support Vector Machine. 5

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The present dissertation aims to explore gesture recognition models so as to evaluate deep learning algorithms applied to Irish Sign Language media. Before delving into the proposed solutions, the relevance and potential impact of the work hereby presented is laid out.

1.1 Motivation for Research Topic

Sign languages have their own linguistic structure, grammar and characteristics, and are independent of the rules that govern spoken languages. They are visual languages that rely on hand gestures as well as on bodily and facial expressions. Sign languages in different countries are vastly different from one another, so enabling easy communication is important: not just to break the barrier between hearing and deaf individuals, but also between people who do not sign in the same language. In Ireland, there are 5000 Deaf people and it is estimated that around 50000 people regularly communicate using ISL [2]. The relevance of automatic sign language recognition is significant: access to interpreters is not available under all circumstances, and societal inclusion is critical for all people, regardless of their hearing ability. ISL has a vital cultural component for the Deaf community, which was acknowledged on September 2017, when legal recognition was granted as the Irish Sign Language Bill 2016 [3] was passed. After President Michael D. Higgins's approval, ISL is now regarded as a native and independent language [4]. This brings about more rights and easier access to public services

for deaf people.
Automatic sign recognition is a widely researched area, especially for American Sign Language. CNNs for gesture recognition have been extensively applied to ASL data ([5], [6], [7]). However, this is not the case for ISL. While automatic ISL detection has been undertaken by making use of techniques like Principal Component Analysis [8], Support Vector Machines, or Hidden Markov Models [9], research on ISL recognition within the field of deep learning is scarcely documented [10], despite the successful outcome for this kind of architectures in similar endeavours ([11], [12]).
The applications of automatic sign language recognition are diverse: sign-to-text translation, assistive technology for hearing-impared individuals, teaching system for sign language students, among others.

1.2 Dissertation Overview

An outline of this dissertation's structure follows. Chapter 2 presents a review of related work in order to frame this work in the context of two areas of research: hand posture recognition, which relates to static sign recognition, and spatio-temporal gesture recognition, which is linked to signs that require movement. Said chapter contains an overview of deep learning concepts and applications in the realm of image recognition, ultimately focusing on state-of-the-art gesture recognition investigation.
Chapter 3 offers a detailed description of the frameworks assembled to tackle static and dynamic Irish sign classification. The results yielded by the experiments in the aforementioned chapter will be discussed and measured against the state of the art in ISL recognition in Chapter 4.
Chapter 5 analyses the findings of this project, based on which Chapter 6 concludes with a proposal of potential lines of work.

Chapter 2

State of the Art in Sign Language Recognition

2.1 Related Work

Computer vision is the science that aims to obtain meaningful information from digital data, image or video, and comprises applications like object detection, 3D reconstruction and augmented reality, among others. This chapter will describe visual recognition, zeroing in on hand posture recognition, both at single-frame level and over a short temporal period.

2.1.1 Hand Posture Recognition

There are two general categories to sort hand gesture recognition methods, depending on whether they are vision-based or involve data gloves. The latter entail the presence of sensors that transduce physical movement into electrical signals that describe the posture of the hand, or simply wearable assets that help track the subject's hands through colour segmentation. The former approach to hand gesture recognition can be subdivided in two more specific groups [13]: methodologies based on 3D modelling of the hand, which are applicable in many contexts but time prohibitive and overly complex for most cases, and appearance-based systems that work well for communicative gestures and are generally less computationally expensive [14]. The structure of said systems can be broadly illustrated in a few stages: image pre-processing and segmen-

tation, feature extraction, and gesture classification. Hand segmentation is primarily concerned with skin detection and normalization, commonly tackled with parametric (Gaussian modelling) and non-parametric (histograms) techniques. However, its simplicity comes at the cost of little robustness, as non-dynamically adapting models are sensitive to lightning conditions and skin tones [15]. State-of-the-art techniques include HSV colour segmentation with thresholding and center of mass normalization [16], and orientation histograms that combine edge detection with pixel-by-pixel comparison [17]. Feature extraction deals with data representation to maximise performance. Past works have implemented solutions that use the contour of the hand, the location of fingers, the palm center coordinates, etcetera. In [18], a feature vector is created to hold the mean values of brightness pixels and the ratio aspect of the bounding box that encases the hand. Other researchers opt for features like the center of gravity of the segmented hand and the distance from this location to the most distant point in the fingers [19], to then construct a circle accordingly and extract a binary signal to count the number of active fingers. The selected feature in [20] is the geometric central moment given by fitting the input hand to a Gaussian distribution. Dynamic hand gestures call for some sort of tracking, either on a frame-by-frame basis, or making use of tracking information. Given the motivation for this research, neural networks –which have been utilised with the purpose of extracting the hand shape [21], and for hand gesture recognition [22][23]– shall be delved into later on.

In the realm of sign language applications, there is a variety of pertinent works such as [24], where a system is developed for open finger detection applied to ASL alphabet signs, with boundary tracing and cusp detection (to locate the fingertip) techniques, prefaced by edge detection and clipping. Additionally, Clynch et al. [25] focus on letter gestures from English sign language and build a pre-classification framework by using histograms to reflect the distribution of distances between random pairs of points in the object. Pansare et al. [26] prioritise the creation of a real-time, robust system for static hand gestures in ASL, whose structure follows a traditional scheme of pre-processing to select the region of interest, followed by feature extraction (centroid and area of hand edge region) along with feature matching to recognise the sign according to the least Euclidean distance. Rheka et al. [27] introduce the concept of derived features from the available feature set in order to render a framework resistant to viewpoint variations, and come up with a hybrid feature set that couple Speeded Up Robust Features (SURF,

[28]) and Hu Moment Invariant features [29]. For the specific problem of ISL, Daniel Kelly contributes with a new user independent hand shape feature [30]: the graph for the hand contour gives the size function, which is an integer-valued function that in this case is modified to incorporate eigenspace information. This makes the feature representation more robust to noise and changes in shape associated with interpersonal differences. PCA, a statistical process that describes observations in terms of linearly uncorrelated values known as principal components, is used to reduce dimensionality and accentuate relevant parts. They report a ROC AUC score of 0.973 for 23 hand signs in ISL, for their own dataset with coloured gloves-wearing individuals, classified employing a set of SVMs trained on the features extracted from labelled data. Another solution proposed for ISL recognition is described in [8]. PCA is also suggested, so an eigenspace depiction of the data, where each dimension is represented by an eigenvector of the covariance matrix, is the focal point of their experimentation. The recognition accuracy for the same 23 ISL classes was shown to be directly proportional to the number of eigenvectors considered, reaching a value of 95% for 29 dimensions.

2.1.2 Spatio-Temporal Gesture Recognition

Two areas may be distinguished within the field of spatio-temporal gesture recognition. On the one hand, it might refer to extracting gestures of interest in a continuous stream; i.e., identifying the start and end point of the sign. This is commonly modelled making use of HMM. For instance, in [31] an HMM-based system that takes into consideration both hand postures and facial cues is presented, reporting an accuracy of 88.5% for a bimodal face and body database. Likewise, a multi-modal implementation is introduced in [32], where palm orientation and place of articulation besides hand shape and motion are included in the framework, which recognises 38 ASL signs with a rate of 93.9%, applying a tree-based classifier where each feature corroborates the similarity between signs or lack thereof. This category also includes [13], where spatial sign segmentation is extended to obtain a spatio-temporal representation that consists of a sequence of feature vectors. These vectors have motion and shape components to characterise the changes in appearance throughout the sequence. In [33], Lee et al. implement the modelling of the likelihood threshold that determines whether a sequence should be classified as a significant gesture. This approach is extended in [34]

by optimizing the combination of the number of states and the number of potential next states in the model. The parameters are iteratively recomputed and the gesture sequence is segmented into subsequences to match the Viterbi algorithm [35] at each iteration until the likelihood converges.

On the other hand, the alternative interpretation, which is the predominant idea that concerns this research, understands spatio-temporal gesture recognition as the classification of frame sequences, each of which presents an isolated gesture with a certain trajectory over time. With regards to hand gesture classification for sign language recognition, attention will be drawn to the following papers. [36] describes a user-independent system for German sign language sequences, combining maximum Likelihood Linear Regression and maximum a posteriori estimation. Spatio-temporal feature extraction techniques based on forward, backward and bi-directional predictions, whose performance is tested on Arabic sign language gestures, are detailed in [37]: the prediction errors are accumulated according to a threshold to represent the trajectory of the sign. Another approach to spatio-temporal sign language recognition is suggested in [38], creating a view-invariant way of measuring the similarity between two sign sequences (one is the current observation and the other belongs to a set of known sequences). This converts the recognition process into a stereo vision-based verification task. Under the necessary assumptions that the fundamental matrix associated with two views should be unique when the observation and template signs are obtained synchronously under virtual stereo vision and vice versa, an accuracy of 92% was reached for Chinese sign language data. Cooper et al. [39] achieve 74.3% for a vocabulary of 164 signs by engineering each sequence as a succession of smaller sub-units that were then assembled into word level by Markov chains. The strength of this paper is its expandable potential. Contrastingly, [40] makes use of an accelerometer and electromyography (EMG); the fusion of features for both channels proved ambiguous and user-dependent. State-of-the-art results for spatio-temporal recognition of ISL are given by [41]. Kelly's work presents techniques for automatic training of models applied to ISL unsegmented videos, offering accurate sign identification in unseen videos. The contribution expands over several fields of study, rendering a system capable of classifying hand postures independently of the hands subject, that also detects epenthesis in temporal sequences. Explicit labelling is not required, as the multiple instance learning algorithm extracts isolated samples of signs then used to train the models. The result is real-time classi-

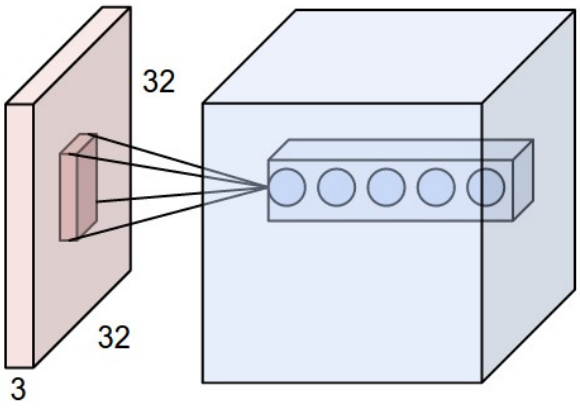


Figure 2.1: Example of 32x32 input image with 3 channels (red) and convolutional layer with 5 neurons (blue) [1].

fication of hand shapes and a framework for motion gesture classification. The main drawbacks are related to the limited conditions under which data was collected and the unpragmatic computational power that a larger vocabulary would entail.

2.1.3 Deep Learning

Deep learning is a subfield of machine learning that refers to deep neural networks, which combine neurons (units whose output corresponds to the result of an activation function that is fed the sum of the linear combination of the input according to a set of weights) grouped into multiple layers [42]. Setting up a network with fully connected neurons for an image-related task is not viable for different reasons. One being dimensionality, for if pixel intensity was the input, the number of weights would be large, and the model would be too complex and overfit. It would also be inefficient due to variance within the same class, since fully connected neurons do not register spatial information, the network would not be translation invariant.
A convolutional layer is composed of stacked local feature detectors that learn local image data because the input of the neuron is limited to the size of the filter. Therefore, the dimensions of the output, which may be regarded as a volume, depend on the size and number of said filters (Figure 2.1). CNNs are advantageous over classic image classification algorithms and fully connected neural networks because they are robust to variation (shift, scale, distortion) in images and fewer parameters are needed. Besides,

they reproduce the way biological neural networks detect increasingly abstract features thanks to the hierarchy of convolutional filters, whose layers gradually transition from low level to high level detectors.
Another relevant theoretical concept motivated by the similarities between image and video data is the extension of 2D convolutions to accommodate a third dimension, thus having a volume as input and output. The receptive field of a 3D convolution extends over both spatial dimensions as well as the temporal one. Therefore, motion information is also captured.

Due to the convenience of modelling spatio-temporal features, 3D CNNs have been used for action recognition [43] and, to a lesser extent, for action detection [44]. 3D CNN models have shown superior performance to 2D CNNs in video analysis, in works such as [45], on the topic of skeleton-based action recognition, or [46], which reinforces the usefulness of volumetric convolutions in the extraction of spatio-temporal features.

Overview of Image Recognition

Image classification is a widely researched task in the domain of computer vision algorithms. Deep learning procedures have stimulated progress, showing unparalleled results in related challenges, such as the ImageNet competition. AlexNet [47] was the turning point back in 2012, due to the vast improvement achieved. Said paper established several foundations that manifested the value of neural networks, such as the popular baseline block that combines convolution and pooling layers, finishing with fully connected layers, as well as ubiquitous design techniques. Since then, over the last few years, more outstanding progress has taken place. In 2014, VGGNet [48] was introduced as a network that proposed increasing depth, and consequently the amount of discriminative features, whilst reducing the number of parameters as a result of more intensive spatial pooling. The next pivotal work looked into the computational aspect that became crucial as networks got deeper. The GoogLeNet [49] architecture introduced the Inception module, which decoupled cross-channel and spatial correlations, rendering the convolution operation as a concatenation of filters, so as to learn features on multiple levels. The resource demands were addressed with 1x1 convolutions that would reduce the number of feature maps. In 2015, ResNet [50] allowed for networks with far more depth through the implementation of skip connections that link layers,

whereby information is propagated and better backpropagation is attained. DenseNet [51] extended this idea by connecting all layers and propagating raw feature maps through all of them.

The main challenges faced with deep learning-based image recognition have to do with the need for vast amounts of labelled data, even if a related dataset is used to pre-train the network and take advantage of transfer learning, along other issues like design optimisation, which can be challenging and time-intensive. Semi-supervised and one-shot learning, as well as having neural architectures learn their own optimal parameters and stack blocks to maximise accuracy, are steps towards solving such problems through automating the design process and making the most of features learnt for other labelled data.

State of the Art in Gesture Recognition

Neural networks present well-known benefits in a broad sense and for hand gesture recognition [52]: they are flexible and capable of handling complex data and inconstant environments since they are able to generalise and learn from existing patterns in the dataset. Deep learning applied to gesture classification has produced a series of pertinent publications, a significant share of which require depth information. Many of these use data captured with Microsoft Kinect, which features an RGB camera, a depth sensor, and an array of microphones. Out of the multiple works that employ the Kinect sensor, [53] is excerpted because of its focus on sub-units to encode information of different nature (location, motion, and hand shape) present in the frame sequence for a sign. The features extracted for each sub-unit are used independently to then perform classification at sign level. This approach is conceptually similar to the architecture presented later on, when multi-input networks are inspected. [54] builds on previous depth-based studies by testing on unseen subjects and environments, and deals with the temporal scope of videos by performing 3D max pooling to reduce the time dimension. Hand and upper body features are considered separately, and the results are concatenated. In [55], intensity and depth, as gathered from images from the ASL alphabet, are kept separate during the first layers of the network, because the authors recognise the intrinsically distinct nature of these two kinds of information. Molchanov et al. [56] employ multiple spatial scales to tackle hand gesture recognition

for depth and intensity information, using 3D convolutions on spatio-temporal volumes composed of interleaved depth and image gradient data, observing a test accuracy of 74.4%.

The need for depth data creates capturing prerequisites such as tracking gloves or other devices, which evidently impairs the scalability and ease of implementation. Solutions that do not rely on such input information are reviewed next. Pre-processing tasks take up the initial stages of the pipeline for the framework developed in [12]. The convolutional neural network in this human hand gesture recognition system is preceded by skin colour modelling, using Gaussian Mixture Models to prevent sensitivity to light conditions, and calibration of the hand to a neutral position. Transitive hand motion is also considered through the inclusion of a post-processing step whereby a classifier assigns the few frames corresponding to transition gestures to either the previous or next gesture. [57] offers a real-time (although slow) implementation that fine-tunes a pre-trained GoogleNet and takes into account the temporal succession of frames by keeping a running cache that evaluates the cost of similar letters and probabilities at a position in a word, for a sequence delimited by the user. In [6], Bheda et al. attempt to create a fairly traditional CNN with no transfer learning involved to classify the handshapes that correspond to the alphabet and numbers in ASL. This course of action echoes a segment of the work undertaken in Chapter 3. Similarly, [14] proposes a system for ASL alphabet recognition. In this case, the neural network is preceded by pre-processing procedures, which encompasses re-sizing, change of colour space (RGB to grey), Canny edge detection, and feature extraction, with a comparison of techniques: segmentation proves preferable to histogram when the set of gestures is small and diverse. Some of the latest, noteworthy deep learning efforts in our area of interest include [21], whose CNN learns seven gestures. Transitive gestures are dealt with by interpolating from adjacent frames. However, it is not entirely generalizable, as there is disparity in performance depending on the sign and it shows sensitivity to hand characteristics. The feature extraction focuses on hand morphology (with prior skin filtering and segmentation), obtained using Self-Growing and Self-Organized Neural Gas network, whereby data is clustered aiming for small distance between points in the same class and large distance between clusters. This method achieves high performance accuracy at the expense of a very limited system, as strict hand and background conditions are required. In [58], a time delay neural network classifies signs from ASL

data with prior extraction of motion trajectories from continuous frames. [59] delivers a scalable solution for ASL recognition that works with a regular camera but is dependant on the capturing conditions. As previously pointed out, the exploration of deep learning techniques to recognise ISL gestures is very limited. The state-of-the-art accuracy is 99.87% [60] for 23 static ISL signs with minimal pre-processing, for a CNN comprised of four convolutional layers with ReLU non-linearity and two fully connected layers with 128 and 23 neurons respectively.

Chapter 3

Design and Implementation

This section provides a detailed analysis of deep learning techniques applied to ISL data. Despite the complexity of sign language communication, two main channels are identified: the hand posture channel, given by position and finger orientation, and the spatio-temporal channel, which defines the trajectory and spatial component of the hand [41]. Consequently, two particular areas of interest are considered:

- Static sign recognition: each sign is conveyed by a single frame.
- Dynamic sign recognition: the target signs involve motion and therefore have a temporal dimension.

3.1 Datasets

Two datasets were employed to account for each of the aforementioned fields of experimentation.

3.1.1 The Irish Sign Language hand-shape (ISL-HS) dataset

The ISL-HS dataset [8] contains 468 videos that capture 6 subjects performing with rotation the 26 hand-shapes that correspond to the letters of the English alphabet, 3 times each. Only the arm and hand are in frame (Figure 3.1). Extracted frames from the videos whose background is removed by thresholding are also provided. For the

developed framework, 23 labels are used, excluding the letters *j*, *x*, and *z*, which entail hand motion and are out of the scope of the first area of research.

Figure 3.1: Frame 1 for Person 1 performing sign for 'A'.

3.1.2 Signs of Ireland Corpus

Compiled by the Center for Deaf Studies (School of Linguistic, Speech and Communication Services, Trinity College Dublin), this dataset is composed of 55 videos featuring 40 human subjects from 5 different locations in Ireland. Out of the 55 videos, 40 depict personal stories that are specific to each of the 40 subjects. The ISL content of the rest of the videos is common to all 15 of them: subjects sign the same story. The videos capture the subjects sitting down (their upper body is in frame) against a background that varies, although in most cases it is a neutral-coloured wall. ELAN is a software tool that allows for visualization and editing of complex annotations, and EAF is an XML file format registers information in several categories, such as eyebrow and body movement, eye gaze direction or the person's attitude. The annotations are supplied as EAF files. For the purposes of this research, only lexical gloss, which gives the meaning of the sign in English, is considered. The 8 most common signs among all videos are selected as target labels. Each sample consists of a set of equally spaced frames extracted from the time interval defined by the timecodes in the annotations. This configuration of input data is defined as trajectory stacking [61]. An example is shown in Figure 3.2. Predictably, even the most repeated signs over 55 videos do not

constitute a large dataset. The fact that poor generalization is foreseen will influence some design decisions, as described later on.

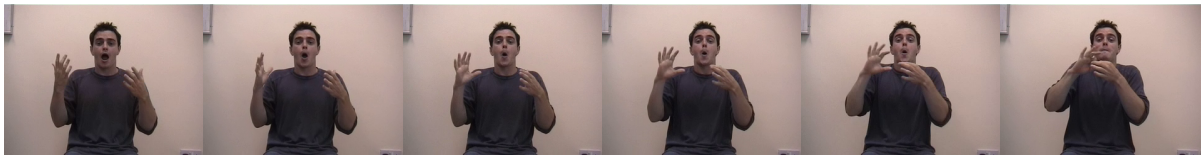


Figure 3.2: Frame sequence featuring Subject 13 (Sean, Dublin) signing 'Home'.

Data from the Signs of Ireland corpus are used with permission of the Centre for Deaf Studies, Trinity College Dublin.

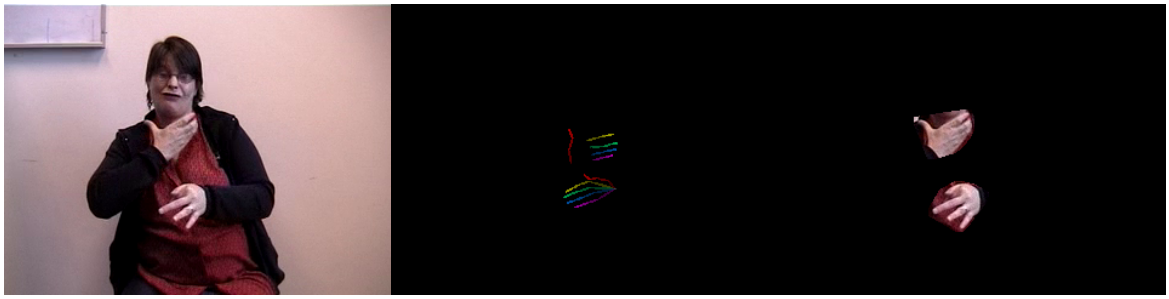
3.2 Framework pipeline

3.2.1 Pre-processing

Hand Segmentation

The Signs of Ireland Corpus presents complex conditions (e.g., low resolution, varying lightning and environment) that obstruct the task at hand, rendering it necessary to prepare the data. Processing the video frames to detect the person's hands with conventional image processing techniques, which included background subtraction based on running average, skin colour thresholding, upper body and face cascading classifiers, and finger recognition based on contour detection produced subpar results. Contrarily, OpenPose [62] trained on the CMU Panoptic Studio dataset provided the best results for hand segmentation among the evaluated detectors (HandSegNet from [63] and hand detector from [64]). The right hand is detected by forward feeding a network based on VGG-19 that produces 2D confidence maps of keypoints and their connections (see Figure 3.3b). Given the elbow and wrist positions, the hand location is estimated under assumptions derived from general human physicality. The left hand is identified by simply flipping the input image and repeating the process. This enables straightforward segmentation, by creating a mask according to the obtained keypoints (Figure 3.3c).

The ISL-HS dataset is cropped around the hands to extract the region of interest and



(a) Original frame (b) OpenPose output (c) Masked frame

Figure 3.3: Frame featuring Subject 05 (Michelle, Dublin) signing 'Frog'.

highlight the relevant information with a traditional image processing approach.

It has been established that several components play into the interpretation of ISL (or any sign language for that matter). It could be essential to include all channels of information for the recognition to be successful. For instance, certain words are signed with the same hand posture but at different start and end positions, and vice versa. For this framework, given that it deals with limited vocabulary, it is preferable to center and focalise the hands in the input images, consequently removing relative spatial information, as CNNs benefit from centered target close-ups.

Data Augmentation

The data obtained from the Signs of Ireland Corpus was artificially augmented to prevent overfitting given the limited amount of samples (5349 sign sequences after said augmentation). The small size will influence some design decisions, as explained further into the chapter. Affine transformations are applied to extend the data through transformations in the data-space. Label preservation is evaluated by observing that the sign can still be recognised, which ensures that the data augmentation techniques are creating plausible images from existing frames [65]. The training data is rotated by -5, -10, 10 and 5 degrees. Salt-and-pepper noise is added with a probability of 0.05 in 5 instances.

In terms of normalisation, both scaling and centering are introduced. Images are normalised between values -1 and 1, which improves convergence speed. Mean subtraction and standard deviation division are also applied to achieve a data distribution

that resembles a zero-centered Gaussian, which promotes uniform gradients. For both datasets, the subjects' hands are clearly singled out prior to being fed to the network. Therefore, colour does not provide any information that would be helpful to localise the object of interest or distinguish between classes. This is the reason why input channels that depict natural data (hands) are converted to greyscale.

3.2.2 Network Architecture

Static Sign Recognition Model

The proposed architecture (Figure 3.5) for the single-frame signs classifier consists of 2-dimensional convolutions that extract features spatially. The appeal of CNNs for this task is linked to their capability to learn local spatial features.

The first layer tends to learn about basic pixel features like lines and corners. This is more evident when observing one of the input frames convolved with the filters in the first layer, as in Figure 3.4.
The number of filters increases to make the most of the networks's scope, which is narrower at the beginning: there are fewer generic features that the model is concerned with during the early layers, and more high level elements that need to be learnt by deeper layers. The size of the filters, on the other hand, gets smaller at successive layers, as this is known to be advantageous [47]. The CNN follows the long-established structure of convolution with ReLU activation followed by max pooling.

Dynamic Sign Recognition Model

The spatio-temporal model follows the general structure depicted in 3.2.2, with two key incorporations. First, since the input information extends over the temporal dimension, three-dimensional convolutional layers are employed. As previously stated, they extend the convolution kernels found in 2D CNNs with an extra dimension that allows for the computation of temporal features. They preserve the principle of local constraint along the new temporal dimension. This fact is useful for sign language recognition, because the aim is to model local variations that describe the trajectory of the gesture over time.

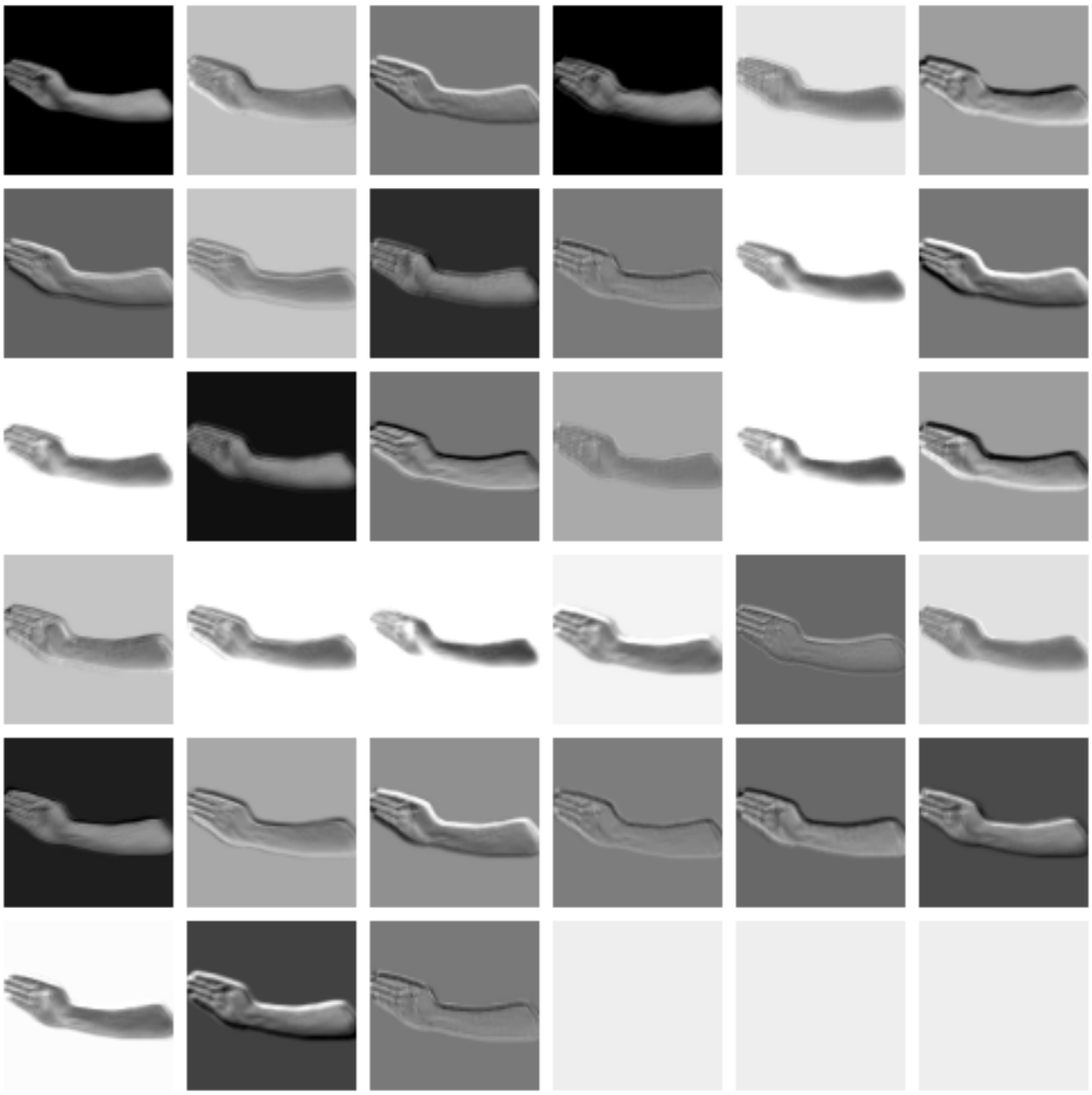


Figure 3.4: Convolution of sample frame with first 32 filters.

Downsampling is performed only on the spatial dimension, since convolutions are enough to reduce dimensionality along the depth axis. The architecture for a single stream is depicted in Figure 3.6.

LSTMs are a variant of RNNs. RNNs, unlike feedforward networks, have memory with regard to the input in time. Their feedback loop signifies that past predictions affect new ones. LSTMs expand the time domain that the neural network has influence over,

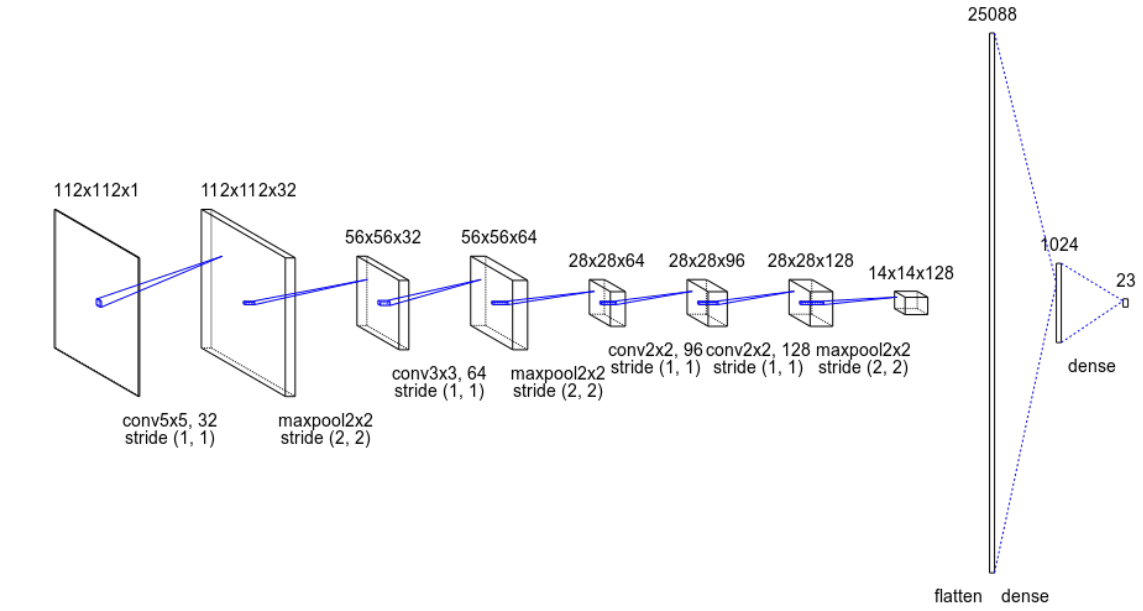


Figure 3.5: Static Sign Recognition Architecture.

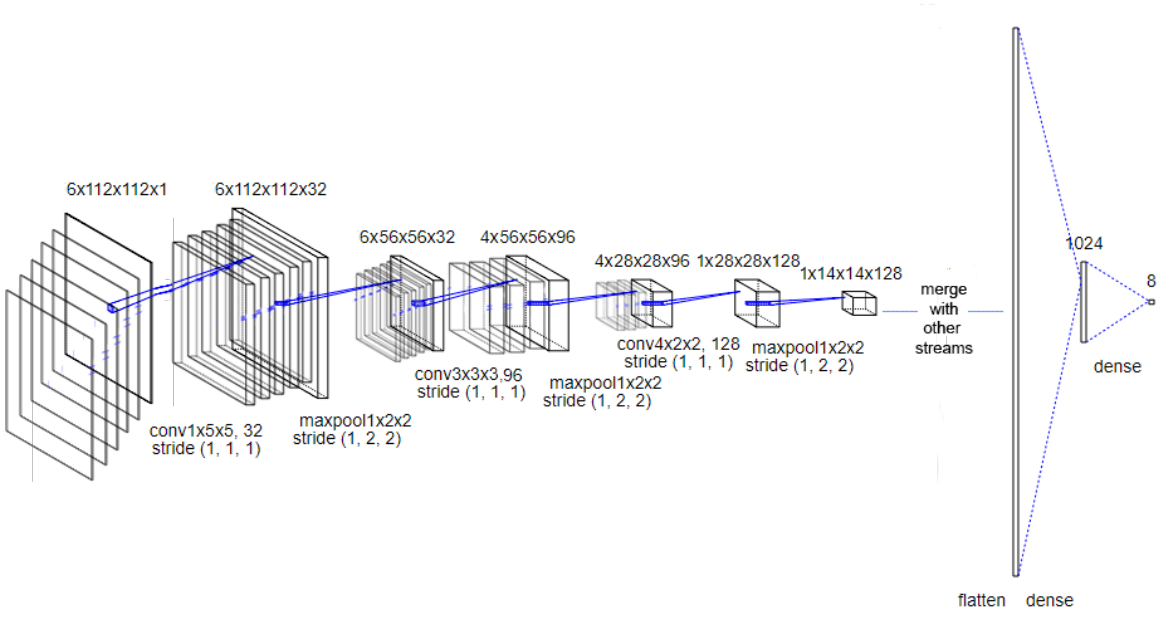


Figure 3.6: Dynamic Sign Recognition Architecture, for one stream.

by virtue of a gated cell that stores certain information, according to what it progressively learns about when and how much to let through. An LSTM-based architecture on its own would not perform well because both spatial and temporal information is not preserved due to flattening. In some instances, notably when the sequences have variable length, a combination of LSTM with 3D CNNs is arranged as a two-step classifier wherein the 3D CNN acts as an encoder for sub-gestures, which are composed of a few frames, and the LSTM unit makes a prediction on the full, longer sequence [66]. Because there are no long term dependencies, LSTMs were discarded during the design process.

The other crucial addition is a multi-stream set-up. Multi-input networks refer to models where each input is fed to independent convolutional streams and trained separately. Such streams are combined at a later point, thus integrating information from different sources. This type of architectures has been explored in past works such as [61] or [45], mostly within the area of action recognition. A variable number of streams has been analysed, together with the type of input suited for the classification of hand signs. Past works suggest a temporal and a spatial stream; i.e., a static frame to learn spatial features, and a series of frames (original data or some other information that adequately typifies temporal characteristics). Nevertheless, this is not applicable in this context since a dynamic sign is not fully characterised with one frame, especially considering the subjective variations between people.

In summary, the dynamic sign classifier consists of the following phases: after temporal segmentation according to dataset annotations, hand detection, region extraction and feature extraction (if used), are implemented on a single frame basis, followed by gesture recognition for each sequence.

Chapter 4

Results

This chapter pertains the results obtained for the experiments described in the following section.

4.1 Experiments

The experiments were implemented using the TFLearn library for TensorFlow and ran on an Nvidia GeForce GTX 1080 graphics card with 8 GB of dedicated memory, using CUDA V9.1.85 and cuDNN V7.0.5.

4.1.1 Nature of Input Data

Optical flow can be defined as the "pattern of apparent motion of objects, edges and surface in a visual scene caused by the relative motion between an observer (an eye or a camera) and the scene" [67]. In the field of action recognition, training on optical flow information rather than raw stacked frames has proven beneficial ([68], [69]). These antecedents suggest potential for the experiments carried out on this basis. Two implementations are tested: dense and sparse optical flow. The difference between them is that sparse techniques compute the flow for specific pixels according to tracked features, whereas dense methodologies process all of them. The particular OpenCV implementation used in the project makes use of the pyramidal implementation [70] to compute sparse optical flow, whereas the Gunner Farneback's algorithm [71] was put

to work in pursuance of dense optical flow. For an in-depth review of both methods, please refer to Appendix A.

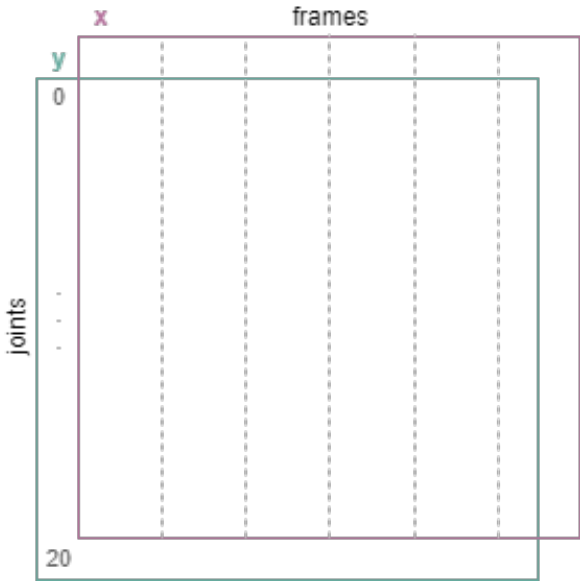
For our case, optical flow did not grant any advantages over raw image data. This can be explained by the fact that optical flow manifests movement at a scale that does not properly differentiate the motion for the different sign language trajectories. The kind of movement that sets ISL signs apart suggests the need for an approach that captures the subtleties of hand gestures. In other words, the location of the hands over the sequence is seemingly well modelled by optical flow, but details like finger position and orientation are overlooked. Additionally, there is natural disparity between subjects, who might place their hands at a different height or distance from body for the same sign, either because of individual styles, or due to the organic transition from other signs: the articulation of a sign can be influenced by the preceding and following signs, an occurrence known as co-articulation [72].
This realization led to the execution of experiments with the methodology of performing feature extraction to compute key points distributed along the hand. This was achieved by obtaining data via OpenPose, as mentioned in 3.2.1. The aforemenetioned tool outputs an RGB skeleton of the detected hand, with colour-coded fingers. This specificity is valuable, as proven by upcoming results, because it does carry the precise information that was noted to be missing in previous undertakings. A visualization of input information is depicted in Figure 4.1.

(a) Right hand (b) Right skeleton (c) Left hand (d) Left skeleton

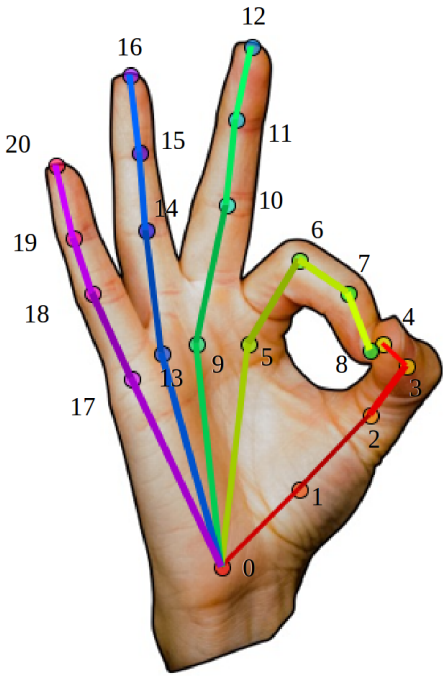
Figure 4.1: Sample of input data streams, featuring Subject 05 (Michelle, Dublin) signing 'Frog'.

Ultimately, the best performing framework reflects an earlier vision: spatio-temporal trajectories can be construed as a set of 2D or 3D points depicting human joints [73]. A different approach to introducing the keypoints as a stream was explored: a four di-

mensional matrix that contains the skeleton information for every frame (Figure 4.2a). Its rows correspond to the joints of the hand (Figure 4.2b), and each of its columns contains the coordinates for one frame in the sequence. Its third dimension has two components (x and y points). The 2D convolutions in the network this matrix was fed through were able to gather temporal information thanks to the layout of such information over spatial coordinates.



(a) Matrix of hand keypoints for a given sign sequence.



(b) OpenPose hand keypoints [74].

Figure 4.2: Hand keypoints data.

It should be noted that left and right hands were considered separately: the baseline network had two streams, corresponding to each of the hands. In all cases, the supplementary data was fed as streams that were parallel to the segmented hands streams, and always keeping the separation between data pertaining the right or the left extremity, as displayed in Figure 4.3.

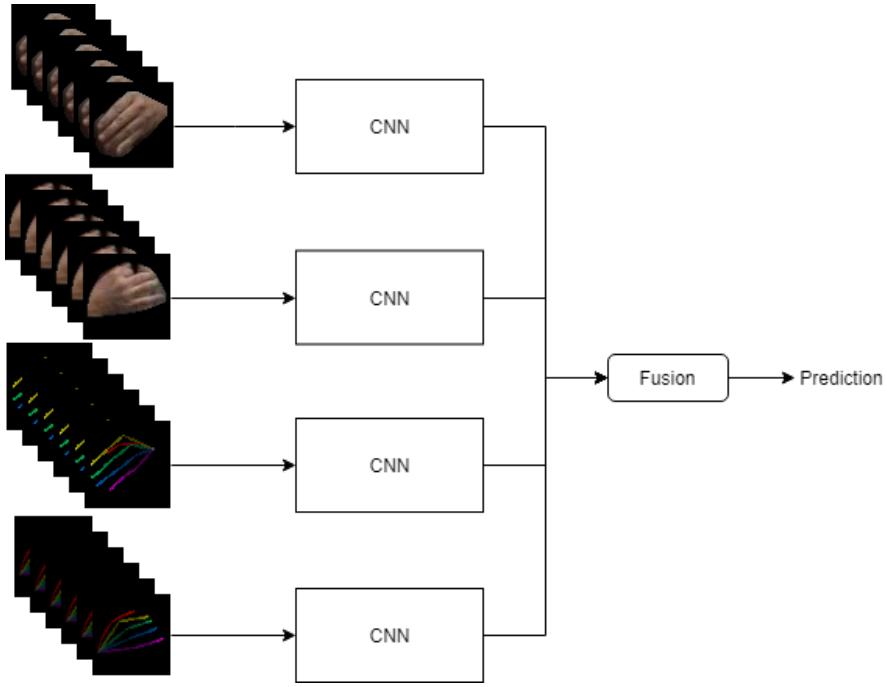


Figure 4.3: Multi-stream model.

4.1.2 Effect of Temporal Filters

In this section, the handling of temporal data through the network is analysed, which is very pertinent when extending the preceding static sign classifier to a dynamic sign model that deals with sequential images. The fusion of time information is thoroughly inspected in [75], whose conclusion supports that maintaining throughout the layers temporal information that is gradually fused grants, to higher layers, access to global spatio-temporal features. This is the rationale behind the increasing depth of the convolutional filters, until the time dimension has no depth and the features are expressed in a 2D map.

It is also interesting to look at the amount of frames that appear necessary to infer the represented gesture. There is a restriction imposed by the Signs of Ireland Corpus dataset: given the frame rate of the videos, the duration of some signs is not enough to capture a large number of frames. Memory and computational power are added factors that favour a lower sequence depth. Ergo, the experiments were conducted within the range of 4 to 8 frames per sign, modifying the temporal extent of the filters to ensure steady filtering of the time dimension whilst preserving the slow fusion architecture.

This manifested that more data per sequence does not entail better performance. Six-frame long trajectories were enough to fully capture the sign, since they trace smooth paths. To further evaluate the effect of the filters (and nature of data) for the best performing model, the deconvolution method introduced in [76] is employed. Figure 4.4 shows the outcome for a deconvolution after the first convolutional block for randomly selected OpenPose samples, which are part of one of the streams. It is appreciable that the fingers are the focal point, which is coherent.

Figure 4.4: Right-hand skeleton samples from the deconvolutional layers.

The deconvoluted image of a dense optical flow sample (Figure 4.5) reinforces the perception that this data is too coarse to be valuable in the context of hand sign recognition. Reconstructing frames from the input stream of segmented hands shows that the network first learns the general appearance of the extremity (Figure 4.6).

Figure 4.5: Dense optical flow sample from deconvolutional layer.

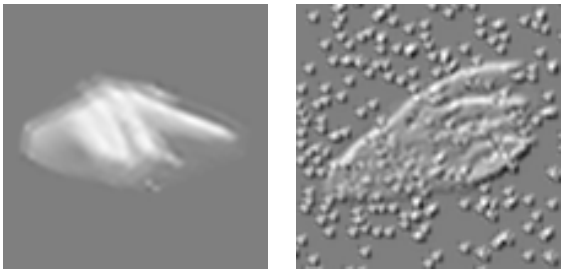


Figure 4.6: Segmented hand samples from deconvolutional layer.

4.1.3 Fusion Techniques

In the sphere of multi-input networks, the matter of combining streams appropriately is a subject of research. Several paths were explored in the experiments. Generally, two categories are recognised: early and late fusion. The former refers to merging models that combine the input data, as raw data or feature descriptors, whereas the latter fuses information after classification. Early fusion was briefly analysed adding the segmented hand as the fourth channel to its skeleton representation. While this improved the accuracy results when compared to the bare-bones network with only the detected hands as inputs, it did not yield any improvement over the four-stream version of the architecture. As per conventional multi-stream architectures, we focus on late fusion. Generally, the final prediction that results from late fusion can be formally expressed as in 4.1 [77].

$$y = f(s^1, ..., s^M) \tag{4.1}$$

where y is the final predicted label, s^m are the scores for the mth stream, and f represents a transition function.

In relation to said function, averaging (which mimics the merging technique in [61]), as well as concatenating the outputs of the fully connected layers, were examined. Late fusion assumes all streams to be expressly complementary, which is, as a matter of fact, not detrimental in this situation. However, this fusion methodology can cause an imbalance in the interpretation of input information, depending on the magnitude of the activations for each stream, and might lead to overfitting [61].

Concatenating the outputs of the convolutions along the depth dimension (number of filters) at some layer was the next fusion mode put into action. Although Park et al. [78] suggest merging at this stage, their proposal includes element-wise multiplica-

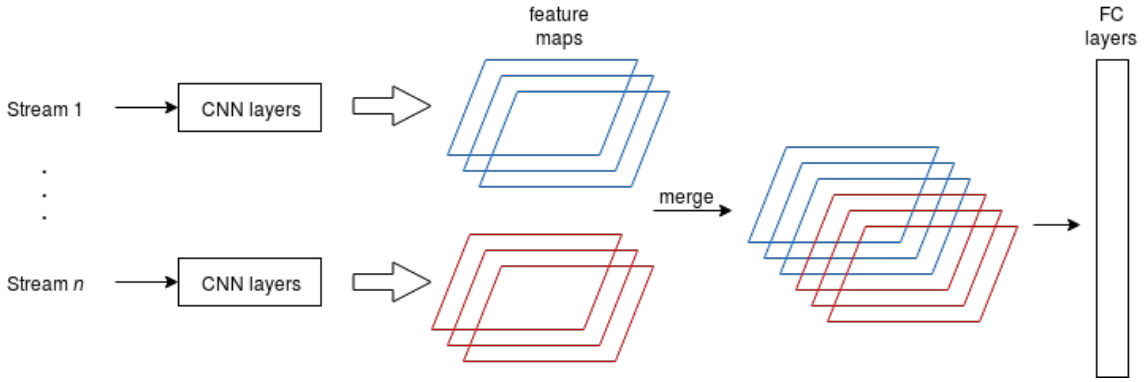


Figure 4.7: Concatenation of feature maps after last convolutional block.

tion, which amplifies or suppresses feature activations, preceded by 1x1 convolutions. For our architecture, a slightly different approach worked better. The feature maps were simply stacked, experimenting with and without a temporal component; that is, fusing before and after the last convolutional block. The superior alternative, whose schematic representation is displayed in Figure 4.7, was found to be the latter. One of the experiments mentioned in 4.1.1, which included a keypoint matrix instead of the colour-coded hand skeletons, required a modified architecture for the keypoint streams, since the dimensions of the input data were completely different. Conceptually, the ideas for gradual temporal fusion and late feature fusion were perpetuated, but the disparate shapes of the frames and keypoint data made it unfeasible to keep the merging technique that stacks feature maps. Fusing the output of the fully-connected layers was a variable alternative, with either element-wise sum or multiplication providing the lowest error.

4.1.4 Hyperparameters

The optimization of hyperparameters is a fundamental part in the construction of neural networks, for the quality of results is dependant on them. Several models have been proposed toward finding the right set of values in the hyperparameter space, but it still remains a computationally expensive and time consuming task. The present section reviews the hyperparameter choices for the frameworks, whose tuning was fulfilled mostly via grid search (sampling values over a specified range) or coordinate descent (fixing all hyperparameters but one, which is adjusted).

- Optimizer: SGD performs a parameter update according to the approximate gradient based on a few traning samples. This optimizer is extensively used in landmark papers ([50], [51], [79]). Better test accuracy for dynamic sign recognition was attained with SGD than with other optimizers; namely Adam, RMSProp and AdaGrad. Adaptive optimizers have been found, in fact, to generalise poorly, especially when there are few data points and many parameters [80]. For the static sign model, as the available data is sufficient, the Adam optimizer was chosen.

- Learning rate: the initial rate at which weights are updated was assessed over the range from 10^{-6} to 10^{-1} and the optimum value was found to be 0.01 for SGD and 0.0001 for Adam.

- Dropout: This is a regularization technique introduced during the training deep neural neurals in order to reduce overfitting by randomising the activation or omission of neurons in accordance an input probability [81]. It prevents co-adaptation of neurons so that they do not rely too heavily on certain units. A high dropout ratio (0.7) is chosen before the last fully connected layer, as it helps with generalisation when training from scratch [61].

- Activation Function: the final neuron output is determined by ReLU (4.2), a popular option that solves issues like exploding gradients in the case of the sigmoid function, and is more time-effective.

$$\varphi(s) = max(0, s) \tag{4.2}$$

  The last fully connected layer has softmax activation (4.3), which assigns a probability of each of the labels, such that the sum over all classes is 1.

$$\varphi(s) = exp(s_k) / \sum_{K} exp(s_k) \tag{4.3}$$

- Depth: In regards to the spatio-temporal model, the number of blocks in the network was inspected for a small range, as the performance quickly showed no improvements for deeper versions of the model. Three convolutional blocks

provided higher test accuracy than deeper models. This kind of performance is likely a product of how small the Signs of Ireland Corpus dataset is: a more complicated model leads to overparametrization, whose complexity motivates an overfitting problem. A similar design methodology was followed for the static framework: progressively deepening the model until no further improvement was observed.

- Width: Reducing the number of filters per layer, thus creating a narrower network, showed a slight decrease in performance, since the problem was not accurately represented when lowering the amount of feature detectors.

4.2 Implementation Results

The frameworks for the experiments that have been previously explained so as to perform spatio-temporal ISL recognition, including the baseline architectures with dual input, are summed up in Table 4.1.

Model	Description
2 streams (I)	Masked left and right hand in greyscale
2 streams (II)	Skeleton for left and right hand in RGB colour space.
4 streams (I)	Masked left and right hands in greyscale + right and left skeletons in RGB.
4 streams (II)	Masked left and right hands in greyscale + greyscale right and left skeletons.
4 streams (III)	Masked left and right hands in greyscale + matrix of keypoints for each hand.
4 streams (IV)	Masked left and right hands in greyscale + dense optical flow.
4 streams (V)	Masked left and right hands in greyscale + sparse optical flow.

Table 4.1: Model descriptions for the dynamic sign recognition multi-stream architecture.

Each model will be referred to as the title given to the row, for simplicity and clarity in the upcoming presentation of results. The metrics for the numerical values found for the devised models were selected to facilitate direct comparison with the latest work in the ISL recognition.

Model	Accuracy
Original Frames	0.9979
Cropped Frames	0.9998

Table 4.2: Accuracy results for the static sign recognition model.

Model	Accuracy
2 streams (I)	0.6610
2 streams (II)	0.6271
4 streams (I)	0.7966
4 streams (II)	0.7288
4 streams (III)	0.7627
4 streams (IV)	0.6780
4 streams (V)	0.6830

Table 4.3: Accuracy results for the dynamic sign recognition model.

4.2.1 Accuracy

Accuracy is the ratio of correct predictions over the number of predictions. In this multi-class context, categorical accuracy is computed for each class. The mean of categorical accuracies is what is referred to as accuracy in this chapter. Various results are included in Table 4.3 for the dynamic ISL classifier model trained and tested on 8 classes, in order to gain perspective on the most notable experiments of the proposed solutions.

These accuracy values lead to some observations: we can infer the presence of colour-sensitive units [82] because of the difference in result for 4 streams (I) and 4 streams (II); the network does learn from the information associated to the colour coding of fingers, since the greyscale depiction of the hand joints rendered an accuracy lower by 7%. With regards to employing optical flow algorithms to convey the temporal aspect of the signs, the lack of success was already considered in Section 4.1.1, because it dictated the course of action for the design process. Those findings echo the discussion in [83], which argues that optical flow is successful in increasing the robustness of the network to the general appearance of the scene, regardless of temporal coherence. This is not useful in this context, since the environment is the same for all classes, and the hands are segmented a priori in any case. Moreover, the small frame dimensions of the available videos prevent greater flow accuracy, which has greater importance

Model	ROC AUC Score
2 streams (I)	0.8762
2 streams (II)	0.8525
4 streams (I)	0.9526
4 streams (II)	0.9231
4 streams (III)	0.9352
4 streams (IV)	0.9147
4 streams (V)	0.9212

Table 4.4: ROC AUC scores for the dynamic sign recognition model.

for small displacements [83]. In all cases, the implementation benefits from additional multi-domain streams of data.

As for the static sign recognition framework, we can refer to the state-of-the-art accuracy for static ISL recognition as given by [60], where a value of 0.99875 is attained for the 23 motionless letters of the alphabet. The model developed in this research reaches an accuracy of 0.9998, thus improving the cited result. The structure of the network is fairly similar to the characteristics of that presented in [60], in a general sense. The same dataset is used: a reduced version of the ISL-HS dataset, where redundant frames have been are removed. The differences in the preparation of aforesaid input data are more pronounced: Oliveira et al. avoid any image pre-processing, whereas in our case the hand is cropped so that the background is less substantial and the hand is the focal point. It does, indeed, make a difference – the accuracy is noticeably worse without it (Table 4.2).

4.2.2 ROC AUC Score

The ROC curve graphically conveys the ratio of correctly classified samples versus data erroneously assigned to that same class; i.e., it plots true positive rate or recall (correct positive predictions over the number of actual positive samples) against false positive rate or inverse recall (samples incorrectly classified as negatives over the total number of negatives). Having defined the binary case, we can extend this metric to encompass of multi-label scenarios such as this one by simply taking the average. The state-of-the-art ROC AUC metric for ISL recognition of spatio-temporal signs

is that of [41]. The ROC AUC score for isolated dynamic gestures assessed on a proprietary ISL dataset of similar conditions for 8 classes is 0.9750. Said data is comparable to the Signs of Ireland Corpus handled in this dissertation: their dataset captures individuals, with their upper body in frame, performing natural sign language sentences, wearing coloured gloves. This is the most notable difference as it simplifies the hand tracking aspect. For the experiments, eight different signs are extracted from these videos, and two channels are considered for parallel models (right and left hand), like the implementation here presented. However, it should be noted that different labels are considered, and there are intrinsic dissimilarities like the labelling format and the changing environments. As seen in Table 4.4, the best-performing suggested framework, which takes both segmented hands and RGB OpenPose-rendered skeletons, does not surpass the state of the art.

For completeness, although it does not stand as the state of the art, it should be acknowledged that the above-mentioned work presents a ROC AUC score of 0.977 for static ISL recognition, considering the 23 non-moving alphabet signs as well. Given the successful outcome observed in Section 4.2.1, our static model logically outperforms that implementation, achieving a ROC AUC equal to 0.9999.

4.3 Research Question Results

This research posed the question of how ISL recognition performs using deep learning algorithms. The preceding section displayed quantifiable results that cover a significant in the scope of sign language recognition. First, the static sign framework dealt with the alphabet signs, which are used for spelling out proper nouns and other words for which there is no sign (this is known as fingerspelling). The outcome shows that ISL letter detection is perfectly suited for deep learning architectures: even a simple network improves the performance of other approaches, with nearly perfect recognition rate for all classes, as seen in the confusion matrix in Figure 4.8.

The second model looked at signs for specific words with translation over time, and a certain hand posture at each instant. Whilst the performance is not superior to the state of the art, the result for dynamic sign recognition is competitive and promising. The confusion matrix (Figure 4.9) manifests recognition errors particularly significant for the *see* sign, since it is very similar to the gesture for *but*: both require raising the

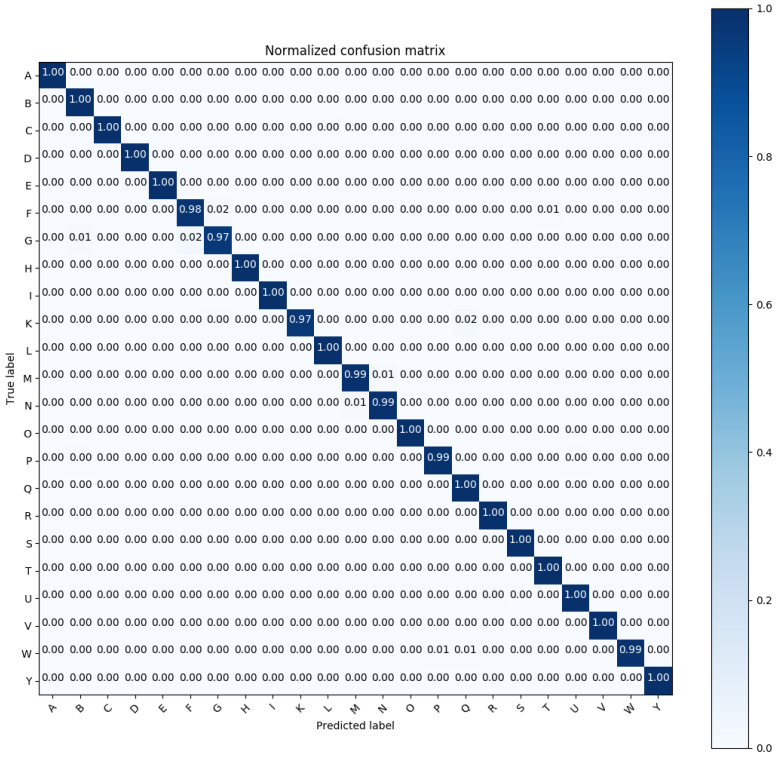


Figure 4.8: Confusion matrix for static sign model.

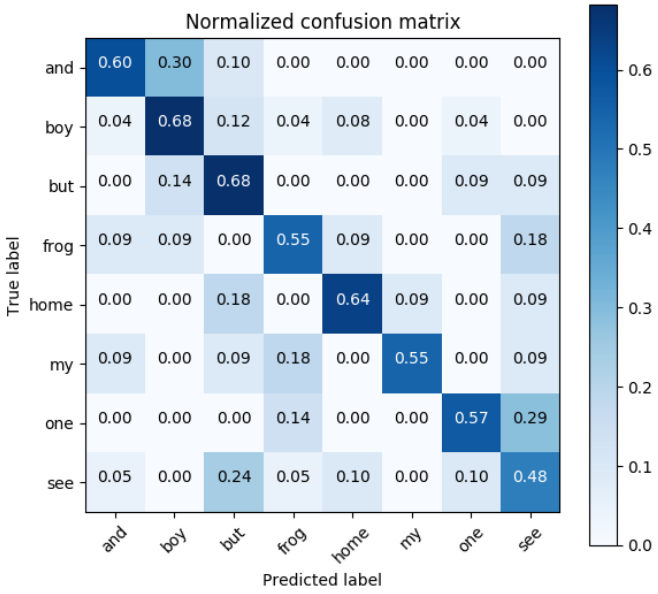


Figure 4.9: Confusion matrix for dynamic sign model.

index finger, the main difference being the position of the hand with respect to the face. The best performance is observed for the most populated classes (*boy* and *but*), thusly supporting the usefulness of a greater amount of samples. One of the strengths of this classifier is that regular videos were used as input: low resolution, irregular capturing conditions, diverse subjects, and independent of electromechanical devices (no gloves or depth sensors). As stated by Vijay et al. [84], a system of maximum efficiency with low cost and good results against complex backgrounds should be preferred. This suggests potential for deep learning applied to ISL, which would benefit from data curated more extensively and precisely.

Chapter 5

Conclusion

With the intention of offering a thorough study of neural networks that learn from ISL images and videos, it has been found that some theoretical notions that work well in related tasks are not as adequately suited for our purpose. Delving into the types of input information constituted an interesting exploration: whereas keypoint coordinates for the joints and fingers proved successful in the learning process, optical flow did not provide an insight into spatio-temporal manual signs. 3D convolutional layers, on the other hand, were positively validated in the modelling of short temporal correspondences. In terms of managing the depth or time dimension through the CNN, it was found that gradual fusion meant that the network learnt about the spatial features first and progressively incorporated the temporal features.

Another point to review is the importance of the size of the dataset. The multi-input set-up provided a means to making the most of the available data, by creating streams that contained alternative representations or extracted features of the information. It seems apt to assume the concept of mutual reinforcement, as articulated by Liu at al. in [45] when asserting the productivity of a two-stream CNN. Pre-processing was a consequential aid, particularly ROI extraction, although previous works have noted that ignoring the features outside of the extracted region can hinder the understanding of the context of the information within the ISL [85]. This has been shown to be true, but it is not a substantial concern for the given framework. In any case, the relative position of the hands in the frame would definitely provide relevant contextual information for a larger dictionary. Some fusion methods for getting a final score

from the multiple input streams were analysed, and retaining spatial information when merging gave the optimum result.

The framework meant for static sign recognition manifested the capability of deep learning systems, even with straight-forward implementations, to perform image recognition tasks.

Many applications can be derived from these results, with some refinements: a real-time fingerspelling captioning application, a tool to automatically subtitle ISL videos given the signs' timecodes, or a virtual ISL teaching assistant that evaluates the student by comparing to the learnt signs.

Ultimately, answers in different domains were found for the research question, which might propel new lines of work, as detailed next.

Chapter 6

Future Work

Parallelisms to Daniel Kelly's published works have been established throughout the document, since they stand as a major study of automatic learning and recognition of ISL. In keeping with these correspondences, future courses of action can be suggested as follow-up lines of work to this dissertation.
An interesting area is continuous sign recognition, which aims to detect and isolate signs from sequences. This involves recognising movement epenthesis, a term that refers to the transitional motion between signs. Regarding ISL, Kelly [34] models each gesture as a succession of sub-patterns. Each of these segments is an HMM state. The HMM is trained to classify each sequence as a sign or epenthesis, according to a probability distribution computed for a two-hand gesture. It would be interesting to apply a different approach to ISL data, such as identifying lip movement patterns: Pfister et al. [86] focus on the openness of the mouth to distinguish between sign and silence, because it is common to mouth the word that is being signed, so there is co-occurrence.

Furthermore, this research could be extended to encompass more factors of expression besides hands. A multi-modal system that incorporates facial and body language would be able to better convey the full meaning and connotations of what is being transmitted through signs. In [87], these non-manual descriptors (facial expressions, head motion and eyebrow gestures) are studied for ISL. Said complementary elements of communication have been analysed applying deep learning algorithms for other sign languages, whose performance in respect to ISL could be assessed in the future. Koller

et al. [88] look at mouth shape modelling in the context of sign language and propose a weakly supervised method that relies on out-of-task data to tackle the lack of labelled data. The solution is a combination of a pre-trained CNN with an HMM framework that determines the most likely sequence according to the network's softmax outputs. Further widening the scope, Caridakis et al. [89] include head pose, eye gaze and expressivity (with tracking methods) and facial expressions (with a RNN). The Signs of Ireland Corpus contains some of this information in the video annotations. However, the scarcity of the available data hints at the need for further data collection and labelling, in order to employ techniques from the works cited above. In any case, the impact on including these modifiers into automatic sign language recognition is still indefinite [89].

Throughout the development of this project, the lack of a great amount of precisely annotated data has posed some issues. Clearly, gathering more samples and strongly supervised data would assist the progress of developing automatic ISL recognition systems. Weak supervision of sign language information, an idea backed by previous works such as [90], [91] or [92], is also an option to compensate for the lack of data and the noisy and weak labelling that is frequent with subtitles, whose alignment may not be exact.

Bibliography

[1] A. Karpathy, "Stanford University CS231n: Convolutional Neural Networks for Visual Recognition,"

[2] L. Leeson and J. Saeed, *Irish Sign Language: A Cognitive Linguistic Account*. Edinburgh University Press Series, Edinburgh University Press, 2012.

[3] "Irish sign language act 2017 (act 40 of 2017) irish sign language bill 2016 (bill 78 of 2016) @Houses of the Oireachtas," Dec. 2017.

[4] "Irish sign language given official legal recognition @The Irish Times," Dec. 2017.

[5] T. Starner and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," in *Proceedings of International Symposium on Computer Vision - ISCV*, pp. 265–270, Nov 1995.

[6] V. Bheda and D. Radpour, "Using deep convolutional networks for gesture recognition in american sign language," *CoRR*, vol. abs/1710.06836, 2017.

[7] P. Mekala, Y. Gao, J. Fan, and A. Davari, "Real-time sign language recognition based on neural network architecture," 04 2011.

[8] M. Oliveira, H. Chatbri, Y. Ferstl, M. Farouk, S. Little, N. OConnor, and A. Sutherland, "A dataset for irish sign language recognition," in *Proceedings of the Irish Machine Vision and Image Processing Conference (IMVIP)*, 08 2017.

[9] D. Kelly, J. McDonald, and C. Markham, "Weakly supervised training of a sign language recognition system using multiple instance learning density matrices," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 2, pp. 526–541, 2011.

[10] S. L. N. E. O. Marlon Oliveira, Houssem Chatbri and A. Sutherland, "A comparison between end-to-end approaches and feature extraction based approaches for sign language recognition," 12 2017.

[11] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Computer Vision - ECCV 2014 Workshops* (L. Agapito, M. M. Bronstein, and C. Rother, eds.), (Cham), pp. 572–578, Springer International Publishing, 2015.

[12] H. I. Lin, M. H. Hsu, and W. K. Chen, "Human hand gesture recognition using a convolution neural network," in *2014 IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 1038–1043, Aug 2014.

[13] Y. Zhu, G. Xu, and D. J. Kriegman, "A real-time approach to the spotting, representation, and recognition of hand gestures for humancomputer interaction," *Computer Vision and Image Understanding*, vol. 85, no. 3, pp. 189 – 208, 2002.

[14] Y. Cui and J. Weng, "Appearance-based hand sign recognition from intensity image sequences," *Computer Vision and Image Understanding*, vol. 78, no. 2, pp. 157 – 176, 2000.

[15] M. Soriano, B. Martinkauppi, S. Huovinen, and M. Laaksonen, "Skin detection in video under changing illumination conditions," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 1, pp. 839–842 vol.1, Sept 2000.

[16] M. M. Hasan and P. K. Mishra, "Hsv brightness factor matching for gesture recognition system," *International Journal of Image Processing (IJIP)*, vol. 4, no. 5, pp. 456–467, 2010.

[17] H. Kaur and J. Rani, "A review: Study of various techniques of hand gesture recognition," in *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, pp. 1–5, July 2016.

[18] X. Li, "Gesture recognition based on fuzzy c-means clustering algorithm," *Department of Computer Science. The University of Tennessee Knoxville*, 2003.

[19] Malima, Ozgur, and Cetin, "A fast algorithm for vision-based hand gesture recognition for robot control," in *2006 IEEE 14th Signal Processing and Communications Applications*, pp. 1–4, April 2006.

[20] M. M. Hasan and P. K. Mishra, "Features fitting using multivariate gaussian distribution for hand gesture recognition," *International Journal of Computer Science and Emerging Technologies*, vol. 3, 04 2012.

[21] E. Stergiopoulou and N. Papamarkos, "Hand gesture recognition using a neural network shape fitting technique," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 8, pp. 1141 – 1158, 2009.

[22] M. Maraqa and R. Zitar, "Recognition of arabic sign language (arsl) using recurrent neural networks," 09 2008.

[23] G. Murthy and R. Jadon, "Hand gesture recognition using neural networks," in *Advance Computing Conference (IACC), 2010 IEEE 2nd International*, pp. 134–138, IEEE, 2010.

[24] J. Ravikiran, K. Mahesh, M. Suhas, D. Dheeraj, S. Sudheender, and N. Pujari, "Finger detection for sign language recognition," vol. 2174, 03 2009.

[25] G. Clynch and E. Young, "Hand posture recognition in sign language using shape distributions," in *6th Annual Information Technology and Telecommunications (ITT)*, 2006.

[26] J. R. Pansare, S. Gawande, and M. Ingle, "Real-time static hand gesture recognition for american sign language (asl) in complex background," vol. 03, pp. 364–367, 01 2012.

[27] J. Rekha, J. Bhattacharya, and S. Majumder, "Hand gesture recognition for sign language: A new hybrid approach,"

[28] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008. Similarity Matching in Computer Vision and Multimedia.

[29] M.-K. Hu, "Hu, m.k.: Visual pattern recognition by moment invariants. ire transaction of information theory it-8," vol. 8, pp. 179 – 187, 03 1962.

[30] D. Kelly, J. McDonald, and C. Markham, "A person independent system for recognition of hand postures used in sign language," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1359–1368, 2010.

[31] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, pp. 64–84, Feb 2009.

[32] L. Ding and A. M. Martinez, "Modelling and recognition of the linguistic components in american sign language," *Image and Vision Computing*, vol. 27, no. 12, pp. 1826 – 1844, 2009. Visual and multimodal analysis of human spontaneous behaviour:.

[33] H.-K. Lee and J. H. Kim, "An hmm-based threshold model approach for gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 961–973, Oct 1999.

[34] D. Kelly, J. McDonald, and C. Markham, "Recognizing spatiotemporal gestures and movement epenthesis in sign language," in *2009 13th International Machine Vision and Image Processing Conference*, pp. 145–150, Sept 2009.

[35] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, pp. 268–278, March 1973.

[36] U. von Agris, D. Schneider, J. Zieren, and K. . Kraiss, "Rapid signer adaptation for isolated sign language recognition," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pp. 159–159, June 2006.

[37] T. Shanableh, K. Assaleh, and M. Al-Rousan, "Spatio-temporal feature-extraction techniques for isolated gesture recognition in arabic sign language," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, pp. 641–650, June 2007.

[38] Q. Wang, X. Chen, L.-G. Zhang, C. Wang, and W. Gao, "Viewpoint invariant sign language recognition," *Comput. Vis. Image Underst.*, vol. 108, pp. 87–97, Oct. 2007.

[39] H. Cooper and R. Bowden, "Large lexicon detection of sign language," in *Human–Computer Interaction* (M. Lew, N. Sebe, T. S. Huang, and E. M. Bakker, eds.), (Berlin, Heidelberg), pp. 88–97, Springer Berlin Heidelberg, 2007.

[40] J. Kim, J. Wagner, M. Rehm, and E. Andre, "Bi-channel sensor fusion for automatic sign language recognition," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–6, Sept 2008.

[41] D. Kelly, "Computational models for the automatic learning and recognition of irish sign language," 01 2010.

[42] Y. B. Ian Goodfellow and A. Courville, "Deep learning." Book in preparation for MIT Press, 2016.

[43] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 221–231, Jan 2013.

[44] R. Hou, C. Chen, and M. Shah, "An end-to-end 3d convolutional neural network for action detection and segmentation in videos," *CoRR*, vol. abs/1712.01111, 2017.

[45] H. Liu, J. Tu, and M. Liu, "Two-stream 3d convolutional neural network for skeleton-based action recognition," *CoRR*, vol. abs/1705.08106, 2017.

[46] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: generic features for video analysis," *CoRR*, vol. abs/1412.0767, 2014.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, p. 2012.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[51] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016.

[52] T. Hninn Hninn Maung, "Real-time hand tracking and gesture recognition system using neural networks," vol. 38, 02 2009.

[53] E. Ong, H. Cooper, N. Pugeault, and R. Bowden, "Sign language recognition using sequential pattern trees," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2200–2207, June 2012.

[54] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3d convolutional neural networks," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, June 2015.

[55] S. Ameen and S. Vadera, "A convolutional neural network to classify american sign language fingerspelling from depth and colour images," *Expert Systems*, vol. 34, 2017.

[56] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–7, June 2015.

[57] B. Garcia and S. Alarcon Viesca, "Real-time american sign language recognition with convolutional neural networks,"

[58] M.-H. Yang, N. Ahuja, and M. Tabb, "Extraction of 2d motion trajectories and its application to hand gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1061–1074, Aug 2002.

[59] V. Bheda and D. Radpour, "Using deep convolutional networks for gesture recognition in american sign language," *CoRR*, vol. abs/1710.06836, 2017.

[60] M. Oliveira, H. Chatbri, S. Little, Y. Ferstl, N. E. O'Connor, and A. Sutherland, "Irish sign language recognition using principal component analysis and convolutional neural networks," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8, Nov 2017.

[61] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *CoRR*, vol. abs/1406.2199, 2014.

[62] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[63] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4913–4921, 2017.

[64] D. Victor, "Real-time hand tracking using ssd on tensorflow," 2017.

[65] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: When to warp?," 11 2016.

[66] K. Mullick and A. M. Namboodiri, "Learning deep and compact models for gesture recognition," *CoRR*, vol. abs/1712.10136, 2017.

[67] D. J. Patel, S. Upadhyay, B. D. Lucas, and M. Zhariy, "Optical flow measurement using lucas kanade method," 2013.

[68] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1932–1939, June 2009.

[69] K. Lertniphonphan, S. Aramvith, and T. H. Chalidabhongse, "Human action recognition using direction histograms of optical flow," in *2011 11th International Symposium on Communications Information Technologies (ISCIT)*, pp. 574–579, Oct 2011.

[70] J. yves Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.

[71] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *SCIA*, 2003.

[72] N. Michael, C. Neidle, and D. Metaxas, "Computer-based recognition of facial expressions in asl : From face tracking to linguistic interpretation," 2010.

[73] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109 – 125, 2016.

[74] P. C. Laboratory, "Openpose." `https://github.com/CMU-Perceptual-Computing-Lab/openpose`.

[75] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, June 2014.

[76] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR*, vol. abs/1311.2901, 2013.

[77] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, and X. Xue, "Multi-stream multi-class fusion of deep networks for video classification," in *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, (New York, NY, USA), pp. 791–800, ACM, 2016.

[78] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–8, March 2016.

[79] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *CoRR*, vol. abs/1611.05431, 2016.

[80] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," 05 2017.

[81] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[82] M. Engilberge, E. Collins, and S. Ssstrunk, "Color representation in deep neural networks," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2786–2790, Sept 2017.

[83] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," *CoRR*, vol. abs/1712.08416, 2017.

[84] P. K. Vijay, N. N. Suhas, C. S. Chandrashekhar, and D. K. Dhananjay, "Recent developments in sign language recognition: A review," *Int J Adv Comput Eng Commun Technol*, vol. 1, pp. 21–26, 2012.

[85] S. Eppel, "Setting an attention region for convolutional neural networks using region selective features, for recognition of materials within glass vessels," *CoRR*, vol. abs/1708.08711, 2017.

[86] T. Pfister, J. Charles, and A. Zisserman, "Large-scale learning of sign language by watching TV (using co-occurrences)," in *British Machine Vision Conference*, 2013.

[87] D. Kelly, "Computational models for the automatic learning and recognition of irish sign language," 01 2010.

[88] O. Koller, H. Ney, and R. Bowden, "Deep learning of mouth shapes for sign language," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 477–483, 2015.

[89] G. Caridakis, S. Asteriadis, and K. Karpouzis, "Non-manual cues in automatic sign language recognition," *Personal Ubiquitous Comput.*, vol. 18, pp. 37–46, Jan. 2014.

[90] D. Kelly, J. M. Donald, and C. Markham, "Weakly supervised training of a sign language recognition system using multiple instance learning density matrices,"

IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, pp. 526–541, April 2011.

[91] H. Cooper and R. Bowden, "Learning signs from subtitles: A weakly supervised approach to sign language recognition," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2568–2574, June 2009.

[92] P. Buehler, A. Zisserman, and M. Everingham, "Learning sign language by watching tv (using weakly aligned subtitles)," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2961–2968, June 2009.

[93] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision (darpa)," in *Proceedings of the 1981 DARPA Image Understanding Workshop*, pp. 121–130, April 1981.

Appendix A: Optical Flow

The computation of the sparse optical flow is based on the Lucas-Kanade algorithm [93], which was proposed in 1981 as a method for image alignment in stereo vision. The correspondence problem is closely liked to motion estimation: its extension to optical flow calculation is a derived from performing template matching on small pixel neighbourhoods.

This method associates a movement vector with features of interest by comparing successive images, under the assumption that the difference is small enough for the approximation to be adequate. Changes in intensity (image gradients) are registered for sub-regions of pixels in order to deduce or discard displacement. There are just two variables per pixel: u and v, which stand for motion along each coordinate. As a result, taking into consideration a set of adjacent pixels leads to more equations than unknowns, which defines an overdetermined system. The Least Squares solution is therefore appropriate. The motion vector for one pixel satisfies equation 1.

$$I_x * u + I_y * v = -I_t \tag{1}$$

where I is the intensity increment in each direction (x,y), which then gives the total intensity difference for that pixel (I_t).
For a pixel neighbourhood, the least square solution for the equations can be expressed in matrix form as follows.

$$V = (A^T * A)^{-1} * A * b \tag{2}$$

where A is the matrix of brightness change in both axes for all pixels in the region, b represents the vector of total local brightness variation , and V is the movement vector.
The particular OpenCV implementation used in the project makes use of the pyramidal

implementation [70]. This is an iterative version of Lukas-Kanade that operates on different image levels, meaning recursive representations of the frame,starting from the original, with decreasing resolutions. Conceptually, this algorithm computes the optical flow at the last level (i.e., the lowest resolution) and propagates this estimation to higher levels. This is done for every level in the pyramid. The propagated output works as an initial guess that enables refinement through the levels. The optical flow at each level is the outcome of an optimization step that intends to minimise the matching error function between two comparable images (I_1 and I_2), as seen in 3.

$$e_L(\mathbf{d}) = \sum_x \sum_y a_n I_1^L(x, y) - I_2^L(x + g_x^L + d_x^L, y + g_y^L + d_y^L) \tag{3}$$

where $\mathbf{g}$ is the initial guess, and $\mathbf{d}$ is the residual pixel displacement vector, whose minimisation is the objective. Towards this goal, standard Lucas-Kanade is applied successively, until the pixel residual is below a certain threshold, or until a maximum number of iterations is reached.

Gunner Farneback's algorithm [71] was put to work in pursuance of dense optical flow. Conceptually, this method begins by approximating the two patches from contiguous frames by quadratic polynomials. The idea is that the global displacement can be inferred for any two polynomials under ideal translation, by equating the coefficients. This perfect relationship between two signals is not realistic, which is why local polynomial approximations constitute a more suitable approach, as is the notion of global displacement not being constant over the entire image; it is instead a function with slow spatial variation. The aforementioned approximations concede a primary constraint. Prior knowledge about the displacement is also incorporated so as to reduce potential errors produced by larger displacements, due to the assumption of local models. This means that introducing an a priori estimate might help create a smaller relative displacement, which is less prone to errors.