UQ Winter Research Project 2021
Supervisor: Jessica Korte

AUSLAN Detection in Video of One or More Persons Signing

By

Aravind Punugu

Introduction

AUSLAN is the sign language used by the Australian Deaf Community. The goal of this project is to be able to detect in a video stream when someone is using AUSLAN. The ability to detect signing allows for a sign language recognizer (SLR) to determine the start and end points of when it needs to translate. Furthermore, detection can also assist people to have a *voice* in video calls (Moryossef et al., 2020, pp. 1). This problem of sign language detection comes under Human Activity Detection (HAD). In HAD, the end goal is to be able to detect one or more types of activities in a video stream of humans.

Literature Review

According to (Xu et al., 2017), there have been quite a few approaches to HAD. (Xu et al., 2017, pp. 2) list a number of papers that have used CNNs for 2D spatial feature extraction while using RNNs to encode a sequence of frames and predict a label. They've also listed a number of attempts made using 3D ConvNets, which can perform convolutions on spatial and temporal features (Xu et al., 2017, pp. 1). So, there are at least two ways of achieving HAD, more specifically sign language detection.

The paper by (Borg & Camilleri, 2019) uses the CNN-RNN method of detection. They use a VGG-16 CNN, pre-trained on the ImageNet dataset to extract visual features from the frames of the video. They also used an identical network to extract features from optical flow data, and other types of motion data (MHI, and multi frame differencing). By performing late fusion of both CNNs, the extracted features can be passed to the RNN as time sequences for classification. Both LSTM and GRU networks are used instead of traditional/baseline RNNs to minimise training problems such as vanishing gradients. The results of the experiments suggest that using only optical flow/motion data can achieve an accuracy of approximately 83% with a loss of 0.54. They also show that using only spatial features can achieve an accuracy of 85% with a loss of 0.51. Combining motion and spatial features improved accuracy by 2% to 87%. The dataset consists of youtube videos of all sorts of things from sports, to people signing, to motivational speeches. The dataset fits the title of the paper "Sign Language Detection in the wild".

Inspired by (Borg & Camilleri, 2019), (Moryossef et al., 2020) use only optical flow data to solve almost the same problem. In (Borg & Camilleri, 2019) they attempt to detect sign language in any scenario, hence the name "in the wild". However, (Moryossef et al., 2020) use data from the DGS Corpus. The videos used from the DGS Corpus were professionally developed, with simple backgrounds, and the signer's clothing was uniform. They are much "cleaner'' than the data used by (Borg & Camilleri, 2019). This is a major factor to consider in the area of sign language detection and recognition. (Moryossef et al., 2020) explore more deeply the use of body pose estimation, to calculate optical flow, and how effective the method is when using both the signer's body and hands, only the signer's hands, and only the signer's body. The other difference between (Moryossef et al., 2020) and (Borg & Camilleri, 2019) is that (Moryossef et al., 2020) are trying to implement the detection model on a lightweight device such as a mobile phone, to be used in video conferencing (Moryossef et al., 2020, pp. 10). Within the constraints of mobile phones, it's understandable why the DGS Corpus was used as their dataset, since most video conferences and video calls have participants with clean backgrounds.

Another approach proposed by (Shi et al., 2021), suggests that combining detection with recognition would improve the overall accuracy, including detection precision. The paper is more specific to finger spelling segments within a signing segment. Their approach looks at modifying the cost function of the detection model, to also include the cost of the sign recognition model, which is called a multi-task model (Crawshaw, 2020). The sign recognition model depends on the accuracy of detecting the start and end points, since inaccurate start and end points will lead to the sign recognition model missing certain parts of the finger-spelling segment, which inaccurately labels certain signs. Therefore, by combining sign recognition, and detection of start and end points, it's possible to get better performance, since inaccurate recognition is reflected in the model's precision to detect finger-spelling segments The model used by (Shi et al., 2021) is based on the R-C3D model from (Xu et al., 2017). This model is an end-to-end model (Appendix 1). (Shi et al., 2021) also compare their model with the CNN-RNN model with recognition built into it, which was less accurate than the R-C3D model.

Data

A relatively large repository of “AUSLAN in the wild” data are news press conferences. 36 videos from both the ABC Australia and 7News YouTube channels have been labelled. In total there are approximately 1.8 million frames with a resolution of 1280x720, with the exception of one video at 352x232. Since these are press conferences, the speaker is the main focus of the video, while the AUSLAN interpreter is usually on the left or right side. In many instances the interpreter’s full body is in view, they're wearing simple colours; usually black, and in most of the videos, the background is constant, where the background is usually a wall of tiled state government, or organisational logos. There are some exceptions to background consistency, when there are complex and variable backgrounds, with the signer wearing clothes with complex colours.

Since these are press conferences, the vocabulary is usually formal. The typical end of a long signing segment is the “thank you” sign. During a non-sign segment almost every signer has the same neutral position. Furthermore, a significant percentage of the videos are related to the Coronavirus, which further restricts vocabulary.

Some attempts at recognition like (Karabas et al., 2013) and (Pansare et al., 2012) use image processing techniques to filter out a person's hand for better recognition. Thinking along similar lines, canny edge detection was used on the press conference videos. Canny is a popular edge detection algorithm that attempts to highlight edges for a given image (OpenCV, 2021). Although the exact sign couldn't be made out, it is possible to tell signing from non-signing segments. However there was too much noise when the background or the signer's clothes got too complex, or when the video bit rate was low. So this idea was quickly abandoned in favour of using the original RGB image.

During the process of labelling the video frames; small scripts were developed for tedious tasks, and can be found at (Punugu, 2021). The data labelling is intuitively formatted, with the sign and non-sign segments separated, and the start and end points separated by a colon. Each video's labelling is put into separate files, with the file's name resembling that of the video's. A limitation of the video labelling, was that it was done by one person, therefore the video labelling is subject to bias.

Conclusion

Due to time constraints a model could not be created. However, as outlined before, there are many effective methods to approach this problem. With the data labeled, the future of this project could be more focused on developing a model, and perhaps processing the data further.

Appendix

1. An End to End model is defined as a singular deep neural network that tries to solve a complex problem, without using multiple models for different aspects of the problem and pipe-lining their results (Glasmachers, 2017).

References

Borg, M., & Camilleri, K. P. (2019). SIGN LANGUAGE DETECTION "IN THE WILD" WITH RECURRENT NEURAL NETWORKS. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6836575&tag=1

Crawshaw, M. (2020, september 10). *Multi-Task Learning with Deep Neural Networks: A Survey*. https://arxiv.org/abs/2009.09796

Glasmachers, T. (2017). Limits of End-to-End Learning. http://proceedings.mlr.press/v77/glasmachers17a/glasmachers17a.pdf

Karabas, M., Bhatti, Z., & Shah, A. (2013). A Model for Real-Time Recognition and Textual Representation of Malaysian Sign Language through Image Processing. https://ieeexplore.ieee.org/document/6836575

Moryossef, A., Tsochantaridis, I., Aharoni, R., Ebling, S., & Narayanan, S. (2020). Real-Time Sign Language Detection using Human Pose Estimation. Retrieved 06 22, 2021, from https://slrtp.com/papers/full_papers/SLRTP.FP.04.017.paper.pdf

OpenCV. (2021). *Canny Edge Detection*. https://docs.opencv.org/master/da/d22/tutorial_py_canny.html

Pansare, J. R., Gawande, S. H., & Ingle, M. (2012). Real-Time Static Hand Gesture Recognition for American Sign Language (ASL) in Complex Background. https://www.scirp.org/pdf/JSIP20120300010_94521693.pdf

Punugu, A. (2021). *video frame labeling tools*. GitHub. https://github.com/tannishpage/video-frame-labeling-tools

Shi, B., Brentari, D., Shakhnarovich, G., Toyota Technological Institute at Chicago, & University of Chicago. (2021). Fingerspelling Detection in American Sign Language. https://openaccess.thecvf.com/content/CVPR2021/papers/Shi_Fingerspelling_Detection_in_American_Sign_Language_CVPR_2021_paper.pdf

TensorFlow. (2021, 03 25). *Recurrent Neural Networks (RNN) with Keras*. https://www.tensorflow.org/guide/keras/rnn

Xu, H., Das, A., & Saenko, K. (2017, August 17). R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. https://arxiv.org/abs/1703.07814