Using Convolutional Neural Networks for Visual Sign Language Recognition

Towards a system that provides instant feedback to learners of sign language

Rami Aldahir
Department of Informatics, University of Sussex, United Kingdom
ra462@sussex.ac.uk

Ronald R. Grau*
Department of Informatics, University of Sussex, United Kingdom
r.r.grau@sussex.ac.uk

ABSTRACT

This paper is about a machine learning system that uses a convolutional neural network approach to recognizing finger-spelled letters in British sign language (BSL). Motivated by a desire to improve sign language instruction, the software allows users to practice their finger-spelling skills and receive instant feedback. A web cam captures footage of the sign language gestures made by the user, which is then processed to produce a feedback score that is displayed on screen. A small pilot study was carried out to test the basic feasibility of the approach and gauge its potential usefulness for sign language instruction. This article explains the steps involved in creating the solution and provides some details on how the machine learning model was trained. We share the results of initial user testing and reflect on the achievements and limitations of the system, the main challenges faced during development, the lessons we learnt, and future research that derives from this project.

CCS CONCEPTS

• **Human-centered computing**; • **Accessibility**; • **Machine learning**; • **Artificial intelligence**; • **Education**;

KEYWORDS

Sign language, deep neural networks, computer vision

ACM Reference Format:
Rami Aldahir and Ronald R. Grau. 2024. Using Convolutional Neural Networks for Visual Sign Language Recognition: Towards a system that provides instant feedback to learners of sign language. In *The 21st International Web for All Conference (W4A '24), May 13, 14, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3677846.3677848

1 INTRODUCTION

Sign language is the primary means of direct communication for many deaf and hearing-impaired individuals. While there is an obvious need to promote sign language proficiency within this group of people, the development of better learning tools would also make sign language more accessible to the general population and so, help create a more inclusive environment for deaf and hearing-impaired people in society. However, learning sign language requires the acquisition of specific psycho-motor skills to perform the related language gestures, and these are not easily learnt without help and practice. Despite the availability of helpful videos on the Internet and very recent progress in AI-based approaches, the bulk of sign language training today still rests on the availability of personal translators to instruct learners and give them feedback. Our motivation was to make the feedback element that forms part of the learning process more instantly accessible through computer vision techniques that use neural networks to facilitate the processing of visual footage of gestures. While much of the related work currently focuses on American Sign Language (ASL), our aim was to use British Sign Language instead, which uses both hands for fingerspelling, instead of just one (in ASL).

2 RELATED WORK

In this section, we provide some background on sign language instruction and pointers to related work on technological approaches to assist said instruction. Readers should note that this will be brief and is meant as a concise contextualization of the development that is then presented in the next section.

2.1 Sign language instruction

Sign language, such as British Sign Language (BSL), is traditionally taught through face-to-face instruction that is conducted by qualified trainers or deaf educators (e.g., [4]). This method emphasizes proper handshapes, movements, facial expressions, and body posture. Learners often mimic their instructors to grasp the nuances of sign language effectively and ensure correct signing [1]. While such an approach provides direct interaction and personalized feedback, it requires the availability of human translators. Within the realm of BSL instruction, one fundamental aspect is fingerspelling, which is the representation of English letters using specific handshapes and movements. Fingerspelling allows signers to spell out words, names, or concepts that do not have corresponding signs in the language. As for the learning of phrases, fingerspelling instructors make use of handshapes but with a larger focus on exaggerated movements and spacing between letters [3].

2.2 Existing technological approaches

Advancements in media and technology have made sign language more accessible through online resources such as videos and tutorials. Various examples can be found on websites dedicated to sign language education (e.g., [2] or [4]) or on video platforms such as YouTube.com. Learners can access a wide range of grammar and

[*] Corresponding author.

W4A '24, May 13, 14, 2024, Singapore, Singapore
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1030-8/24/05
https://doi.org/10.1145/3677846.3677848

vocabulary lessons. However, even with more material being available, there is still the major drawback of a lack of real-time feedback, which can greatly slow down learning and affect motivation. More recently, advanced AI image processing algorithms have become feasible that can recognize human hand gestures with acceptable accuracy to be potentially useful for training (e.g., [9, 10]). However, the wholesale translation of more complex sign language gestures (e.g., for sentences and phrases) into regular language remains difficult, because the grammatical structure of sign languages can be quite different from spoken languages [12], even for those that would translate back to the same spoken language (like American and British Sign Language). These differences can be explained in part by how these languages developed over time [1]. BSL is one of the oldest sign languages in Europe, with its early forms dating back many hundreds of years.

Currently, there is no solution for the automatic recognition and translation of British Sign Language, which emphasizes the need for making instruction and training more accessible and pervasive.

2.3 Convolutional neural networks

Our approach to obtaining data about individual sign language proficiency relies in large part on computer vision algorithms. Convolutional neural networks (CNNs) are useful for transforming image data of sign gestures into actionable insights. CNNs excel in learning image features and hierarchies, making them particularly well-suited for recognizing sign language gestures captured through video feeds. CNNs consist of multiple layers, each applying a filter to process input images. The filters convolve over the image, producing dot products and generating activation maps. CNNs are preferred for their efficiency in training, given their relatively low number of parameters [7].

Unlike traditional deep neural networks (DNNs), CNNs are optimized for image processing tasks, boasting efficiency in training due to their reduced parameter count (e.g., [13]). This efficiency gain is crucial for real-time applications like sign language detection, where the immediate recognition and processing of sign gestures is required to provide instant feedback to the learner.

CNNs have a structure that is similar to DNNs in that both consist of layers of neuronal nodes that can manipulate data in images to achieve a form of classification. However, CNNs have better precision at analyzing spatial information within images [8]. Through the utilization of convolutional layers, CNNs offer more control over how filters are applied to different regions of the sign gesture images, and so enhance the model's ability to separate intricate details that are crucial for detecting signs correctly. In the current approach, each input image, sourced from the video feed of the user's sign gestures, is resized to 416x416 pixels to strike a balance between resolution and computational efficiency.

3 SYSTEM DESCRIPTION

The gesture recognition system[1] described in this paper uses deep learning techniques, specifically a pre-trained model to carry out transfer learning on a custom dataset. In this section we will cover the process of collecting a dataset that can be used by the model to learn from, and the results that followed the training process.

Table 1: Augmentation techniques and related values

Augmentation Technique	Values
Rotation	From ±10deg to ±14 deg
Brightness	Factor chosen from 0.5, 0.75, 1.25, 1.5
Noise	Applied to up to 17% of Pixels
Exposure	Between −40% and +40%

3.1 Visual gesture detection in real-time

A pre-trained model was chosen to perform the image recognition. While it would have been possible to create a custom convolutional neural network, it would be more beneficial to the capabilities of the learning tool if transfer learning techniques[2] were used instead. This approach mirrors previous successful efforts for the recognition of American Sign Language [6]. The YOLOv5 (You Only Look Once version 5) model was chosen for its performance in real-time object detection [11]. It excels at detecting specific objects and can be trained on custom datasets, making it suitable for recognizing sign language gestures. YOLOv5 divides the image into different segments and applying a grid to predict object probabilities within each segment.

3.2 Collection of learning material

The dataset we created contains 934 training and validation images. Training images refer to the images that will be used for the model to learn which images represent which letter, while the validation images are used to allow the model to practice its skills on labelled unseen images.

In order for the tool to work with any user with any hand shape or size, and for any background, the images in the dataset were augmented to have a larger variation in their rotation, brightness, exposure, and noise properties. Table 1 below shows the different augmentation techniques that were carried out, as well as the values of the augmentation.

These techniques resulted in a larger variety of images being adjusted to represent different skin-tones, hand shapes and sizes, and environments. Using these new images in the model improved the performance of the learning tool, as the risk of overfitting the data onto just one representation was greatly reduced.

3.3 Model performance

Once the training process was completed, a file containing the weights of the training instance was generated. These weights represented the details of the associated classes (BSL signs).

The detection script will use these weights in conjunction with the users' webcam data to detect any gestures that the user may be signing. This process is carried out for every frame to ensure the user gets accurate real-time feedback. 1 1 shows the training results for the model, observing how loss decreases and converges to 0.0, while the accuracy (precision, recall, and accuracy metrics) increases to converge to 1.0. The loss in this graph represents the

[1] Accessible at https://github.com/TinyMushroom6/BSL-Fingerspelling-Learning-Tool

[2] Using a pre-trained model that has been trained with another large dataset (usually consisting of general items) to gather weights to apply to a new custom dataset.

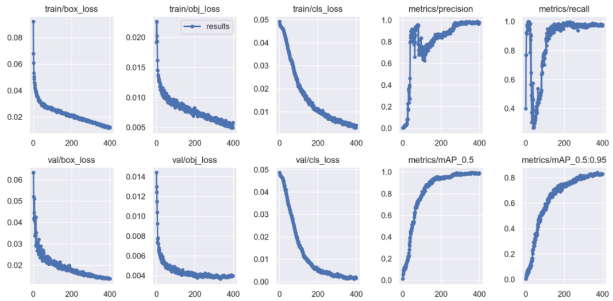

Figure 1: Results after training

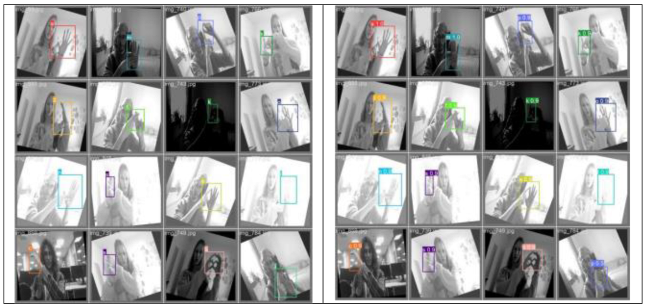

Figure 2: Class identification based on validation batches with ground truth batch (left) and prediction batch (right).

percentage of images in the dataset that have been classified with an error. In this case, seeing a reduction to a near 0.0 value is preferred as it means less images are being falsely classified. One challenge faced in getting results like these is to find a good balance for the number of epochs.

While more extensive training may have led to better results, there is a point of diminishing returns. There is also a limit to accuracy, given the dataset. Our dataset was quite small and so, any further training may have resulted in more overfitting: displaying better results on paper but not necessarily performing as well in practice when the tool is used with different people. 2 2 shows snapshots of visual processing done by the model using validation images. Here, we attempt to apply what the model has been trained to do.

In 2 2, the image on the left is the batch of images with correct label classes. The image on the right shows the model predictions, i.e. what class the model estimates an image belongs to. Accuracy shown is good, bar a few exceptions, such as the model detecting 'I' as 'P' in this instance.

3.4 Graphical user interface

Our first learning tool prototype has a very basic graphical user interface (3 3) that was realized with the TKinter Python library. Elements include a timer that starts a countdown from 10, so that a user gets 10 seconds to sign a randomly chosen letter that is prompted to them. The user can also use a dropdown menu to select any specific letter that they want to practice.

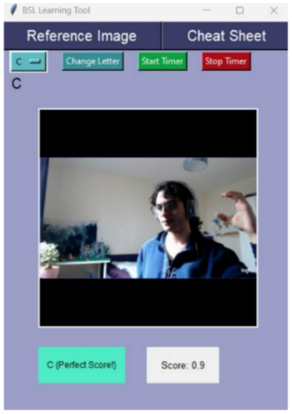

Figure 3: Screen snapshot of the GUI while a user is signing the letter "C"

The camera feed is displayed on screen so that users can relate the feedback they receive to the position they have adopted for a particular sign. The reference images allow users to refresh their memory about how a particular sign works or looks like [2], which is especially important for novices who have not internalized the shapes of signs yet. The software was subsequently tested to see if the gesture recognition mechanism would a) accurately recognize gestures that were correctly signed by an expert and award high feedback scores, and b) also recognize gestures that are not entirely correctly signed, and c) be of potential utility in improving a sign language novice's fingerspelling skill.

4 DISCUSSION

The system was evaluated in a small pilot test that involved a person who was an expert in British Sign Language with many years of experience (Person A), and another who had no prior sign language skills (Person B). The sign language proficiency of the participants was confirmed through a questionnaire before they engaged with the user tests. The main purpose of these tests was to confirm the correct function of the gesture recognition mechanism across different users. Having achieved a satisfactory technical gesture recognition accuracy, it was hypothesized that:

1. An expert user would achieve high feedback scores right from the start, as they already know how to sign individual letters in BSL. The scores were expected not to change a lot as the expert continues to use the tool to practice the signs, because their experience and acquired motor skills should be reflected by their ability to produce correct BSL gestures consistently.
2. A novice user would likely not score as highly as an expert initially, as they lack pre-learnt psycho-motor skills to perform gestures correctly at the first attempt, bar for some of the simplest gestures, perhaps. A lay person would likely need to adjust the position of their hands and fingers more often to get the correct sign. The scores were expected to improve as they continue practicing the same letters. However, we did not expect any dramatic increases in score as the successful acquisition of motor skills can take time [5].

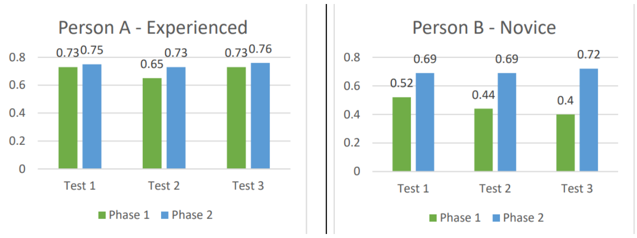

Figure 4: Results from phase 1 and phase 2 of the test for experienced and novice participants.

The tests were divided into two phases, with the first phase being designed to test the participants' knowledge of fingerspelling before using the tools and features of the program, while the second phase was designed to test their knowledge after learning with the program. The participants were asked to use the tool and their average scores for signing each letter were recorded. First, they were shown a reference image of what the signed letter looks like, and then they would try to sign that letter, and do so for every available letter from A to Z, with feedback scores recorded accordingly. A noteworthy detail is that the skin tone of the participants did not have any impact on the recognition performance because brightness augmentations were applied to the reference footage so as to compensate for potential color variations of the skin.

After phase 1, the participants were asked to spend 30 minutes using the learning tool to practice a range of different signs. They then repeated the test from phase 1. Each participant did three tests, and scores across these were averaged. The right-hand-side graph in Figure 4 shows an improvement of Person B's average score when signing a letter between the two phases. During phase 1, the novice performed less well with an average accuracy of 45%, which was then enhanced after practicing their skills with the software, to then achieve an average score of 70%. This was better than we had expected, and even close to the performance recorded for the experienced participant. As we had expected, the score for the expert signer did not change much throughout the experiment, indicating a proportionally lower improvement in the experienced signer's competency.

While the results give some positive indication of the tool's capabilities, it is important to point out that the number of participants who took part in this initial study was very small. The observed increase in proficiency for the novice user in particular will require confirmation by repeating the experiment with a larger number of participants.

One of the largest challenges was for the detection model to reliably carry out the recognition task across different users. This meant finding a good balance between speed and accuracy, to produce a well-optimized model that does not suffer from overfitting. While YOLOv5 performed well on custom data, an improved dataset would use more images. While a lot of current research targets American Sign Language, more should be done to also support other widely used sign languages in the world.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a prototype system that uses contemporary computer vision techniques and a convolutional neural network to recognize fingerspelled letters in British Sign Language. The system can currently provide basic performance feedback to the user. The system's initial evaluation provided the first indicators for its potential as a training aid in sign language instruction. The next step in further development and study will build on the initial system and findings to get more representative results about the usefulness of the approach for sign language training. Fundamental work is also required to explore the utility of systems like this one to support sign language learning independently and in hybrid learning arrangements and develop appropriate instructional methodologies that use different kinds of feedback to ensure the tool can effectively support psycho-motor skill acquisition. The next iteration of this study will involve more participants to confirm if the learning success that novices can experience with this system will really scale as well as we observed in our first tests. It will also provide better evidence for the recognition capabilities of CNN-based systems by using a panel of sign-language experts to perform a cross-evaluation of the performance displayed by novices and the scores awarded by the system. To further improve the reliability of the system's detection capabilities, it would be useful to revisit the dataset and add a larger variety of reference images. For example, the system might then compensate more effectively for visual background noise.

There is also potential for additional functionality, such as predicting the words that users are trying to sign based on the initial language gestures made (akin to the autocompletion feature in word processors). This could be done by combining the computer vision aspect of the current approach with Natural Language Processing algorithms and use of language transformers to identify the letters a user is signing in sequence and suggest possible words for them to complete. This could have the potential to make video-based sign language communication more efficient but also provide more insight into the internal structures of these languages, by getting a better understanding of the flow of interconnected signs that form different words and phrases.

REFERENCES

[1] M. Aronoff, Meir I., Sandler W. The Paradox of Sign Language Morphology. Language (Baltim). 2005 Jun;81(2):301-344. doi: 10.1353/lan.2005.0043. PMID: 22223926; PMCID: PMC3250214.
[2] British Sign Language Fingerspelling Alphabet Charts. Available at: https://www.british-sign.co.uk/fingerspelling-alphabet-charts/ [Accessed: 19/02/2024]
[3] City Lit.(2008). British Sign Language For Dummies. Publisher: For Dummies.
[4] deaf Action,(2011),https://www.deafaction.org. [Accessed: 12/02/2024]
[5] Alberto Casas-Ortiz, Jon Echeverria, Olga C. Santos. 2023. Intelligent Systems for Psychomotor Learning: A Systematic Review and two Cases of Study. In du Boulay, B., Mitrovic, A., & Yacef, K. (Eds.). Handbook of Artificial Intelligence in Education. Cheltenham, UK: Edward Elgar Publishing.
[6] T. F. Dima and M. E. Ahmed, "Using YOLOv5 Algorithm to Detect and Recognize American Sign Language," 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 2021, pp. 603-607, doi: 10.1109/ICIT52682.2021.9491672.
[7] LeCun Y, Bottou L, Bengio Y, Haffner P: Gradient- Based Learning Applied to Document Recognition, In Proceedings of the IEEE 1998.
[8] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. 2017. Deep Learning. Massachusetts: MIT Press.
[9] Papastratis, I., Chatzikonstantinou, C., Konstantinidis, D., Dimitropoulos, K., Daras, P. Artificial Intelligence Technologies for Sign Language. Sensors 2021, 21, 5843. https://doi.org/10.3390/s21175843

[10] Rung-Huei Liang and Ming Ouhyoung, "A real-time continuous gesture recognition system for sign language," Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 558-567, doi: 10.1109/AFGR.1998.671007.

[11] Tewari, A. (2021). YOLOv5: Object Detection on a Custom Dataset. Towards AI. Retrieved from https://pub.towardsai.net/yolo-v5-object-detection-on-a-custom-dataset-61d478bc08f9. [Accessed: 19/02/2024]

[12] W. Li, H. Pu and R. Wang, "Sign Language Recognition Based on Computer Vision", 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 919-922, 2021.

[13] Yamashita, R., Nishio, M., Do, R.K.G. *et al.*, Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018).