Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT 7), pages 45–52
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Improving Signer Independent Sign Language Recognition for Low Resource Languages

Ruth Holmes, Ellen Rushe, Frank Fowley, Anthony Ventresque
School of Computer Science, University College Dublin & SFI Lero
{ruth.holmes,frank.fowley}@ucdconnect.ie, {ellen.rushe,anthony.ventresque}@ucd.ie

Abstract

The reliance of deep learning algorithms on large scale datasets represents a significant challenge when learning from low resource sign language datasets. This challenge is compounded when we consider that, for a model to be effective in the real world, it must not only learn the variations of a given sign, but also learn to be invariant to the person signing. In this paper, we first illustrate the performance gap between signer-independent and signer-dependent models on Irish Sign Language manual hand shape data. We then evaluate the effect of transfer learning, with different levels of fine-tuning, on the generalisation of signer independent models, and show the effects of different input representations, namely variations in image data and pose estimation. We go on to investigate the sensitivity of current pose estimation models in order to establish their limitations and areas in need of improvement. The results show that accurate pose estimation outperforms raw RGB image data, even when relying on pre-trained image models. Following on from this, we investigate image texture as a potential contributing factor to the gap in performance between signer-dependent and signer-independent models using counterfactual testing images and discuss potential ramifications for low-resource sign languages.

Keywords: Sign language recognition, Transfer learning, Irish Sign Language, Low-resource languages

1. Introduction

Modern deep learning techniques rely heavily on large scale datasets. However this becomes a significantly limiting factor when such large datasets are unavailable or difficult to obtain, as is the case with many low-resource sign languages. This limitation is not unique to sign language recognition, with several techniques being proposed to perform image classification within this resource-constrained setting (Larochelle et al., 2008; Sharif Razavian et al., 2014). Sign language recognition adds an additional nuance to this challenge as models not only need to generalise to different variations of hand signs but also to new signers. Training models on a low number of signers causes them to learn the characteristics of particular individuals leading to significant levels of bias in the models and limited applicability in real-world settings (Kim et al., 2016).

In this work, we first quantify the disparity in performance between signer independent and signer-dependent models for Irish Sign Language (ISL) letter hand shape recognition. We show the effects of different input representations on the performance of signer-independent models trained on low-resource data and the tendency of raw image data to lead to significant bias, even when transfer learning is used. The experiments show that pose estimation alone may lead to increased performance in this scenario. To study the efficacy of pose estimation models, the effects of colour on existing pose estimation models are shown. Finally, we experiment with different levels of fine-tuning to assess whether this provides a regularisation effect.

The remainder of this paper is structured as follows. Section 2 describes current works studying the area of low resource datasets and signer-independent models along with the preprocessing techniques used; Section 3 describes our approach to evaluating transfer learning and input representations for low-resource sign language recognition; Section 4 describes the details of the dataset used in our experiments, models and evaluation techniques. We present and discuss the results of these experiments in Section 5. Finally, we conclude with a summary of our findings and a discussion on potential future work in Section 6.

2. Related Work

One of the over-arching issues associated with sign language recognition research is a distinct lack of large-scale, diverse datasets. In particular, less prevalent languages such as ISL and Italian sign language (LIS) whose users number approximately just 40,000-60,000 (Leeson et al., 2015; Branchini and Mantovan, 2020) experience this to a greater degree.

Due to the low-resource nature of these types of datasets, it is imperative that we consider the potential influences this has on real-world recognition scenarios. While differences in camera quality, lighting, and scenery are all valid and important considerations, it is also important that our methods properly account for diversity of signer. We must therefore also be considerate of gender, skin-tone, fluency, age, disability, etc (Bragg et al., 2019).

The lack of signer variety that comes with low-resource datasets is a recurring challenge in the literature. Specifically, there are many sign languages where data is extremely limited, both in availability and size. For example, (Nakjai and Katanyukul, 2019; Fagiani et al., 2015; Oliveira et al., 2017a; Oliveira et al., 2017c) experiment on datasets with fewer than 12 signers. This inevitably leads to bias in the models trained on these datasets. For example, the dataset used in our experi-

ments consist of just 6 signers (3 male, 3 female), all of whom are of similar skin-tone, dressed in dark long sleeves, and are recorded in extremely similar studio conditions. It is therefore clear that we need to address both preprocessing and training in a different way compared with scenarios where signers and data are in abundance.

In terms of preprocessing, several works have utilised raw images as input or a combination of images and auxillary features. Both Openpose (Cao et al., 1812) pose estimation and RGB values were used by (De Coster et al., 2021), optical flow and RGB values were combined by (Shi et al., 2018), data augmentation on raw images was performed by (Pigou et al., 2016) including rotation, stretching and shifting, while (Oliveira et al., 2017c) experiment with raw images alone. Other works take a more domain specific approach, using several image processing and feature selection techniques. (Nakjai and Katanyukul, 2019) perform thresholding and calculate the maximum contour area of each image before classification, (Fagiani et al., 2015) obtain the centroid coordinates of the hands with respect to the face, (Oyedotun and Khashman, 2017) convert images to binary and apply noise filtering while (Oliveira et al., 2017a; Oliveira et al., 2017c) also experiment with PCA and image blurring. (Fowley and Ventresque, 2021) create synthetic data for ISL fingerspelling recognition, achieving high performance in a signer independent setting. Though the authors are approaching a similar problem to that we aim to tackle, we instead focus on a less language specific-approach that does not require synthetic dataset design. Signer independent models are also addressed by (Kim et al., 2016) using neural network adaptation, however this assumes that a small number of examples from the test signer is available which we assume will be unavailable in our work.

While other works have studied signer independent models (Fowley and Ventresque, 2021; Kim et al., 2016), we do so explicitly in a low resource context. We experiment with the most effective preprocessing techniques in the literature and determine their contribution to classification performance in this context. Specifically, we examine the generalisability of different input representations in isolation, determine the most useful method of fine-tuning for pre-trained models, and discuss the impact of these experimental design choices on the overall classification performance.

3. Adapting to Low Resource Sign Languages

For languages where availability of data is limited, i.e. *low-resource languages*, training deep learning algorithms can be challenging due to their dependence on large-scale datasets (LeCun et al., 2015). Furthermore, as with other tasks that utilise bio-metric data, performance of subject-independent models tends to be distinctly lower than subject-dependent models (Kim et al., 2016; Lockhart and Weiss, 2014). This negative effect on performance tends to be amplified for low resource datasets as the number of subjects contained within them will naturally be lower.

3.1. Transfer Learning

An obvious choice for learning with limited image data is transfer learning (Sharif Razavian et al., 2014) due to the wide availability of pre-trained image models. However, the degree of fine-tuning needed to exploit the features learned from pre-trained models for sign language recognition is less obvious. In this paper we investigate the effect of fine-tuning an entire network on this domain-specific data versus fine-tuning only the final classifier. We assess whether it is necessary to adjust the parameters in the earlier layers of the network in order to adapt to this task or whether the potential regularising effects of simply training the final layers are more beneficial. We also assess this specifically in the signer-independent scenario compared to the signer-dependent scenario to determine whether signer-independent models benefit more from this regularising effect.

3.2. Input Representation

Though transfer learning alone vastly improves the ability of a network to learn image features with a small amount of data, there remains a question as to whether these are, in fact, the features the network *should* be learning in order to generalise to the largest number of signers possible. We seek to directly compare two of the most common input representations for sign-language recognition: raw image data with minimal pre-processing and pose estimation keypoints. Below is a discussion on the motivation for this comparison for low-resource sign language data.

3.2.1. Raw Image Data

The use of raw image data in deep learning models has become ubiquitous in computer vision. Raw color values, for instance, are vital in order to identify varying objects and textures. However, for low resource computer vision, there is a question as to whether color features are desirable to learn directly from the data relating to the task at hand. The role of incorrect white-balance, for instance, has been found to cause errors in deep learning models due to bias in datasets towards white-balanced data (Afifi and Brown, 2019). When we keep in mind that low-resource datasets have a low number of signers, the potential for the particular characteristics of signers such as skin tone, dress colour etc. to bias datasets is undeniable. We will show the sensitivity of sign language recognition models to colour by determining the disparity in performance between greyscale and RGB images.

3.2.2. Pose Estimation

Given the potential dependence of low resource computer vision models on less than optimal features, we

seek to determine whether extracted pose estimation could potentially outperform raw images (even with pre-trained models) and generalise better to signers not in the training data. Though many state-of-the-art pose estimation tools also use raw images as training data, they are typically trained on far more data than could ever be collected in a low-resource scenario. We hypothesise that using a highly accurate pose estimation model's output as sign language recognition model's input will allow for better generalisation, as the sign language recognition model is forced to learn only from the features that matter the most, i.e. the coordinates of body parts and their relationship to each other, with minimal dependence on the personal characteristics of the signer.

4. Experimental Setup

4.1. Dataset

The following section describes the dataset used for experiments. We describe the different dataset configurations we created to assess the affect of certain attributes on the overall performance and generalisation.

4.1.1. ISL Hand-shape Dataset

The dataset of Irish Sign Language Hand-shapes (ISL-HS) was originally curated by (Oliveira et al., 2017b) and is publicly available for download[1].

The dataset consists of 468 RGB24 videos of 3 male and 3 female signers performing the 26 ISL alphabet hand-shapes. Each hand-shape was recorded three times at 30 frames per second (fps) and resolution of 640 x 480 pixels. The curators of this dataset have also extracted the frames from these videos, converted them to greyscale and removed background features using a pixel-value threshold. The resulting frames include just the single hand and forearm used to perform the hand-shape. These hand-shapes can be further distinguished into two subcategories:

1. **Static hand-shapes**: All English letters with the exception of 'J', 'X' and 'Z' which include no dynamic movement in their action. These signs were performed using an arcing motion (vertical to horizontal) to better simulate real-world gestural permutations. There are on average 2291 grey-scale frames per hand-shape.

2. **Dynamic hand-shapes**: English letters 'J', 'X' and 'Z' which were performed only using the motion of the gesture itself thus resulting in relatively fewer frames on average (1809 frames per hand-shape) with 'X' having the least of all (1443).

4.1.2. Data Configurations

In order to ascertain the disparity in performance of signer-dependent versus signer-independent models, we create the following two dataset configurations.

[1] https://github.com/marlondcu/ISL

1. **Signer-dependent dataset**: Three trials of each letter are signed by each person in the dataset. The first trial is used for training, the second for validation and third for testing. This ensures that data from all signers present is available for training, validation and testing, while ensuring the frames used in each set are different. We also assess the effect of image colour composition on performance with the following variations.

   (a) The greyscale frames provided by (Oliveira et al., 2017b), see Figure 1a.

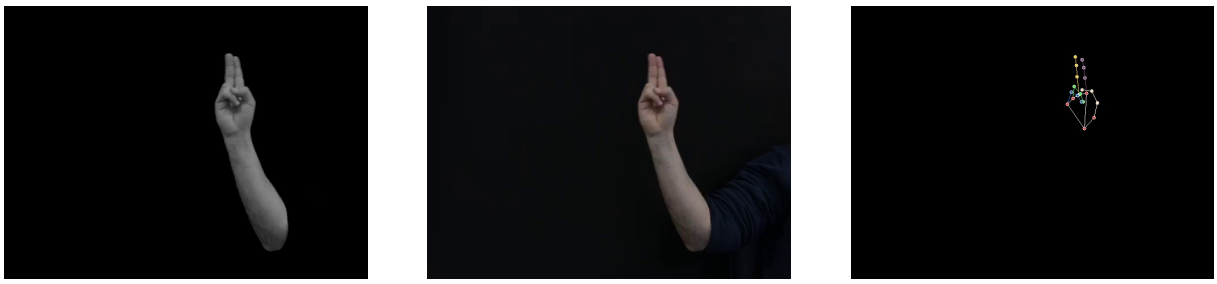
   (b) The RGB frames we extracted from the videos provided by (Oliveira et al., 2017b), see Figure 1b. We noted that this process lead to 143 fewer frames than the greyscale data provided in the public dataset. This is seemingly due to a small number of the original videos being very slightly longer than those provided in the public data.

2. **Signer-independent dataset**: To keep the signers in each set separate, data from *Person 1* and *2* is used for training, *Person 3* and *4* is used for validation and *Person 5* and *6* is used for testing. This also ensures that a similar number of examples are present in each set of this dataset as the signer-dependent dataset. Next we perform pose estimation on the signer-independent dataset to create a third data configuration. This is to assess the extent to which pose estimation can close the gap in performance between signer-dependent and signer-independent models. We use MediaPipe Hands (Zhang et al., 2020). Where the detection confidence surpasses a minimum threshold, we plot the pose estimation co-ordinates in 2D, modifying the default pose estimation plots to prevent landmarks from becoming visually overcrowded, see Figure 1c. Where the pose estimation confidence does not meet this minimum criteria, the raw frame is simply used. The minimum detection confidence set for our experiments was 0.5. We stress that though it is certainly possible to use the pose estimation co-ordinates directly as input features, this transformation into a 2D "image" allows a direct comparison of the same model architectures irrespective of the input and to hold all other algorithmic features and hyper-parameters constant. The following data configurations are used:

   (a) Greyscale frames provided by (Oliveira et al., 2017b)

   (b) RBG frames of from the videos provided by (Oliveira et al., 2017b). In the same way as the signer-dependent dataset, this lead to fewer RGB frames for each video than those provided in the grey-scale dataset.

   (c) Pose estimation images for greyscale frames.



(a) Greyscale. (b) RGB. (c) Pose estimation.

Figure 1: The letter U performed by *Person 2* in Greyscale, RGB and the corresponding pose estimation.

(d) Pose estimation images for RGB frames.

4.2. Models

For all experiments the same deep architecture and hyperparameters are used. This was done in order to ensure that all but the desired aspects of the data or model being tested were kept constant.

Table 1: Hyperparameters used across all VGG models.

Hyperparameter	Value
Normalisation	Standard for VGG16 [a]
Image resizing	(120, 160)
Optimiser	Adam
Initial learning rate	0.0001
Batch size	64
Number of epochs	50

[a] `https://pytorch.org/vision/stable/models.html`.

4.2.1. VGG network

For this network, we used an ImageNet pre-trained VGG network (Simonyan and Zisserman, 2014)[2]. An additional layer with 4000 unit, with ReLU (Nair and Hinton, 2010) activation and Dropout (Srivastava et al., 2014) of 0.5 was added along with and a classification layer with 26 outputs.

4.2.2. Fine-tuning

Fine-tuning was performed in two ways for each model:

1. The added layers of the network alone were fine-tuned on the ISL training set.

2. The entire network, including pre-trained layers, were fine-tuned.

This process was performed to determine whether a regularisation effect could be achieved by excluding the pre-trained layers from the fine-tuning process.

[2] `https://pytorch.org/hub/pytorch_vision_vgg/`

5. Results

This section first details the results of signer independent compared to signer dependent models in subsection 5.1. We then move on to compare raw images to a pose estimation representation in subsection 5.2. Additionally, we provide a discussion on our results and further analysis in subsection 5.3.

5.1. Signer-Independent Models

We can see in Table 2 that there is a sizable disparity between signer-independent and signer-dependent models, trained on greyscale images, even for a relatively homogeneous dataset. It is reasonable to expect that there would be an even larger gap in performance between these models for signers with significantly different characteristics to those in this dataset, highlighting the challenge with datasets of this size. One may expect that this drop in performance is an indicator of over-fitting however when we plot the validation accuracy over all 50 epochs in Figure 2, we can see that these models never perform anywhere near as well as their signer-dependent counterparts. This, once again, highlights the tendency of these models to learn characteristics of the training images not useful to generalisation. With respect to fine-tuning, interestingly, signer-independent models gain slightly more benefit from fine-tuning all layers in the network more than the signer-dependent models.

This disparity in performance is not unique to sign-language models with similar behaviour to be seen in fields like activity recognition (Lockhart and Weiss, 2014) and electroencephalography classification (Zhang et al., 2019). The performance of the signer-independent models shown here closely mirror that achieved by other authors (Fagiani et al., 2015; Shi et al., 2018) - diverting from the higher performance results achieved in the signer-dependent work of (Nakjai and Katanyukul, 2019; Oyedotun and Khashman, 2017; Oliveira et al., 2017a; Oliveira et al., 2017c).

All this indicates that models trained on raw images have a tendency to utilise signer-specific features when classifying hand shapes. Of course, a larger number of signers would likely help remedy this behaviour, though for low-resource languages such as ISL, this data tends not to be available. Therefore we con-

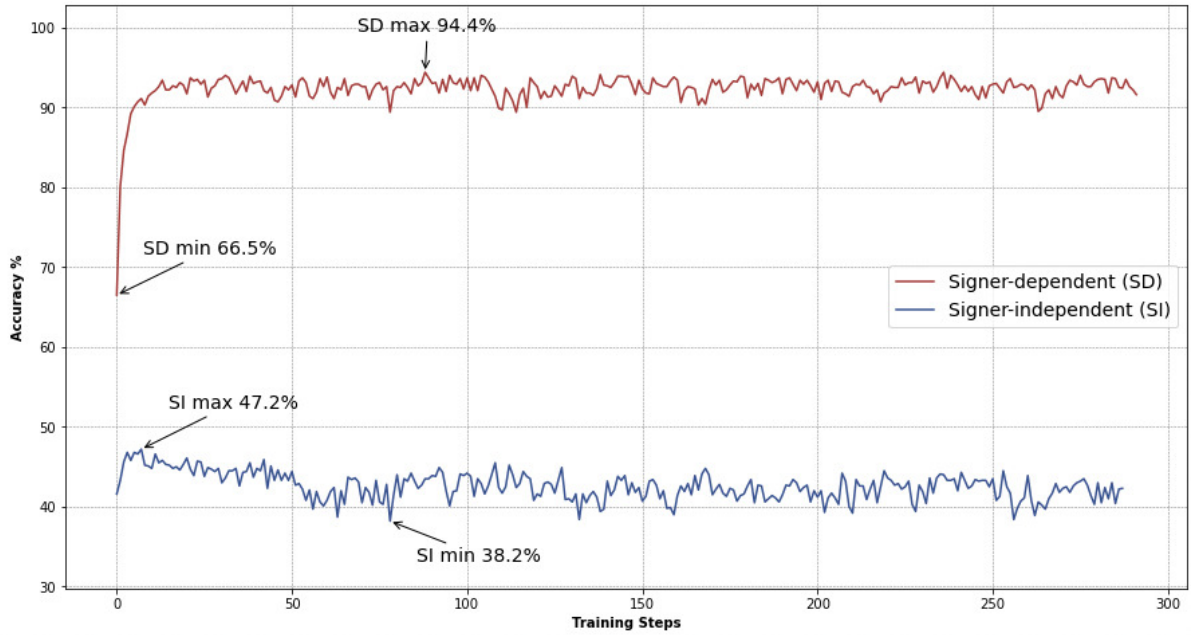


Figure 2: Validation accuracy for the signer-dependent and signer-independent models trained on greyscale images with only layers added to the pre-trained VGG network fine-tuned. Minimum and maximum values are labeled.

clude that signer-independent models using raw RGB images have limited generalisability for these low resource scenarios, even when pre-trained image classification models are used. This motivates a more generalisable input representation.

Table 2: Signer-dependent versus signer independent models on greyscale data. *Added layers* refers to models where only the layers added to the end of a pre-trained network were fine-tuned. *All layers* refers to models where all layers of a pre-trained model were fine-tuned.

Type	Fine-tuning	F1-Score
Signer-Dependent	Added layers	0.885
	All layers	0.882
Signer-Independent	Added layers	0.433
	All layers	0.463

5.2. Raw Images vs. Pose Estimation

Table 3: Pre-trained VGG network's performance on signer-independent data.

Fine-tuning	Input	F1-score
Added layers	Greyscale (∼48% MP frames)	0.486
	RGB	0.369
	RGB (∼99% MP frames)	**0.545**
All layers	Greyscale (∼48% MP frames)	0.468
	RGB	0.399
	RGB (∼99% MP frames)	0.542

For our main results in 3, we first look at the effect of converting greyscale images to MediaPipe landmarks, with roughly 48% of these images being successfully converted. We can see that these pose-estimation features increased the performance for greyscale images, especially when pre-trained weights are kept fixed, with this variation achieving a 4.8% increase over the best performing signer-independent model in Table 2. We also evaluated models trained on the corresponding RGB frames. Neither models trained on raw RGB images exceed the performance of the best model trained on greyscale images. Again, we can see in Figure 3 that validation accuracy for raw RGB data remains in this region of performance for the entire training period. This, at first, seems surprising given that pre-trained models are trained on colour images. However, we hypothesise that this is caused by features that are signer-specific, but irrelevant to the characteristics of a given sign, being more successfully learned by these models, leading to poor generalisation. This is despite the fact that regularisation is used in the form of Dropout in the second to last layer added to the VGG network. This behaviour is actually exacerbated when pre-trained weights are not fine-tuned. We can see that fine-tuning all layers leads to slightly increased performance for raw RGB images. In fact, we can see that both raw greyscale and RGB images show that a similar increase in performance can be gained from fine-tuning all layers of the network. Interestingly, we do not see such an increase when including pose images generated from MediaPipe.

Finally, we look at the effect of converting RBG images to MediaPipe pose estimation landmarks, with over 99% of images successfully converted. There is over

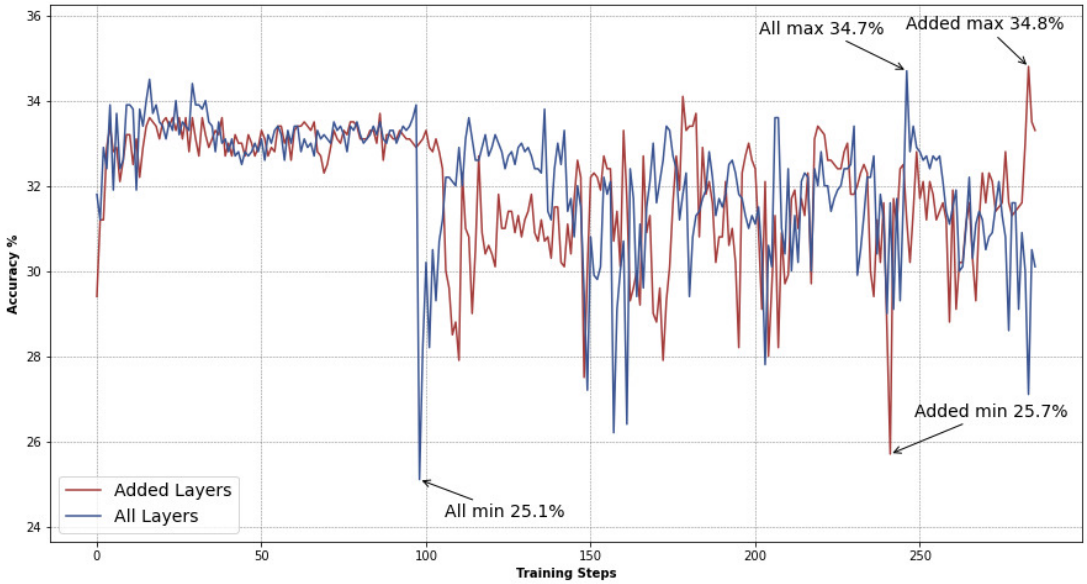


Figure 3: Validation accuracy for raw RGB frames with fine-tuning on added layers and for raw RGB frames with fine-tuning on all layers.

an 11% improvement compared to the next best models (greyscale images converted to MediaPipe, added layers fine-tuned), a pronounced boost in performance. It is fascinating that models trained on raw RGB images, in fact, come last in terms of performance. This provides evidence that pose estimation with minimal use of RGB images (less than 1% due to low pose estimation confidence) provides greater generalisation to unseen signers than utilising RGB images for a low resource dataset. We also observe greater performance when excluding pre-trained layers from fine-tuning.

5.3. Feature Analysis

We observed that some features had a significant effect on the overall performance of both pose estimation and sign recognition. We discuss some of our further analysis and observations below.

5.3.1. Pose Estimation

The effect of removing colour from images on pose estimation performance, even when using a popular pose estimation model trained on a large amount of data, illustrates the sensitivity of such models to colour in images. Table 4 compares the number of successfully converted greyscale images compared to RGB images. Figures 4a and 4b further break down this comparison by letter and signer ID respectively.

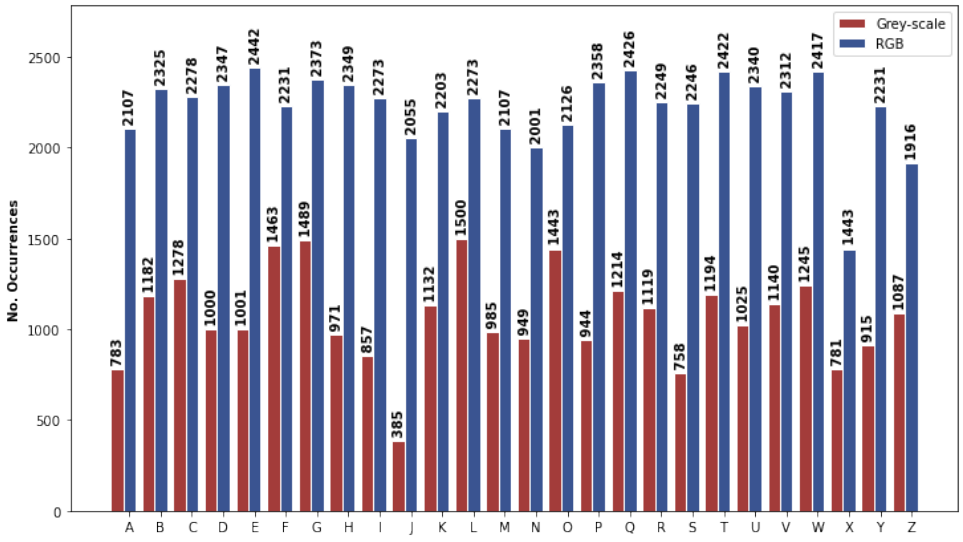
Table 4: Results of the Mediapipe conversion on both greyscale and RGB frames

Frame Type	# Frames Available	# Frames Converted	No. Frames Non-Converted	% Frames Converted
Grey-scale	58,114	27,840	30,274	47.9
RGB	57,971	57,850	121	99.7

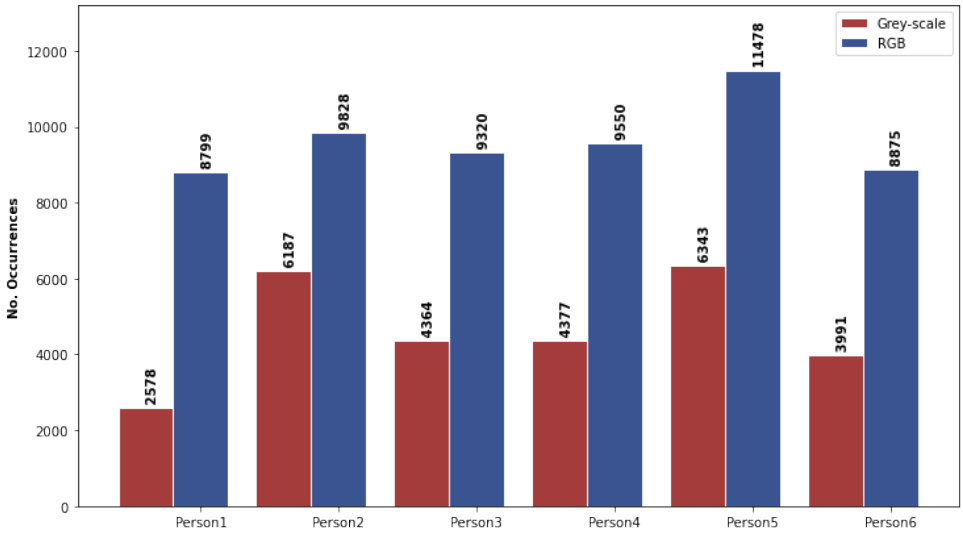
On the face of it, this may not appear to be a problem, due to the fact that modern images are unlikely to be in greyscale. However, the reliance on colour indicates that performance could also be greatly affected by lighting conditions and scenes not present in the training data. This type of behaviour has been observed in other computer vision tasks (Afifi and Brown, 2019). This should be taken into consideration if sign language recognition systems rely on such models and might motivate other types of feature pre-processing before pose estimation such as optical flow.

5.3.2. Signer-specific Characteristics

The large gap between signer-dependent and signer-independent models in the case of raw RGB images is challenging as it can be difficult to determine which of the characteristics specific to each signer is being learned when trained on low-resource data. Models pre-trained on ImageNet have indeed been found to be biased towards image texture over shapes of objects within images (Geirhos et al., 2018), which in this case translates to the texture of the clothes being worn by the signer and their skin texture. To evaluate whether this could be a large contributing factor, we create counterfactual examples of the signer-independent RGB test set described in 4.1.2 by applying a Gaussian blur to images to smooth the image texture. We do, in fact, see a decrease in performance for the model which was trained on raw RGB images which suggests that this is a contributing factor. This feature alone, however, is clearly just a single aspect that influences this gap in performance and further evaluations are needed to ascertain other contributing factors.



(a) Hand-shapes.



(b) Signers.

Figure 4: The number of frames converted to pose estimation for both hand-shapes and signers between the grey-scale and RGB images.

Table 5: VGG model with a Gaussian blur applied to test set

Fine-tuning	Input	F1-score
Added layers	RGB	0.359
All layers	RGB	0.384

6. Conclusion

In this work we have illustrated the large performance disparity between signer-independent and signer-dependent models in Irish Sign Language. We show that using accurate pose estimation as input when training on low-resource sign language datasets increases recognition performance. We have investigated the improvements needed for pose estimation models to become more effective and have used counterfactual examples to show the effect of texture on models using raw RGB data. It should be noted that these images account for just a small subset of ISL manual hand shapes. We also recognise that the resolution of the images used in these experiments and their distinct lack of background noise is often an overly optimistic representation of real-world finger spelling. However, this work is merely the beginning of a line of research that will perform more extensive analysis on the effects of input representation, the ways that this representation can be made more robust and the role of the network architecture in improving signer-independent generalisation.

7. Acknowledgements

We would like to warmly thank our colleague Thomas Laurent for his valuable feedback and comprehen-

sive proofreading assistance on several drafts of this manuscript. This work was supported, in part, by SignON, a project funded by the European Union's Horizon 2020 Research and Innovation programme under grant No. 101017255; and by Science Foundation Ireland grant 13/RC/2094 P2 to Lero - the Science Foundation Ireland Research Centre for Software (www.lero.ie); F. Fowley is funded by the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (D-REAL) under Grant No. 18/CRT/6224.

8. Bibliographical References

Afifi, M. and Brown, M. S. (2019). What else can fool deep learning? addressing color constancy errors on deep neural network performance. In *ICCV*, pages 243–252.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *ASSETS*, pages 16–31.

Branchini, C. and Mantovan, L. (2020). *A Grammar of Italian Sign Language (LIS)*. 12.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (1812). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. arxiv 2018. *arXiv preprint arXiv:1812.08008*.

De Coster, M., Van Herreweghe, M., and Dambre, J. (2021). Isolated sign recognition from rgb video using pose flow and self-attention. In *CVPR*, pages 3441–3450.

Fagiani, M., Principi, E., Squartini, S., and Piazza, F. (2015). Signer independent isolated italian sign recognition based on hidden markov models. *Pattern Analysis and Applications*, 18(2):385–402.

Fowley, F. and Ventresque, A. (2021). Sign language fingerspelling recognition using synthetic data. In *AICS*, pages 84–95.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Kim, T., Wang, W., Tang, H., and Livescu, K. (2016). Signer-independent fingerspelling recognition with deep neural network adaptation. In *ICASSP*, pages 6160–6164. IEEE.

Larochelle, H., Erhan, D., and Bengio, Y. (2008). Zero-data learning of new tasks. In *AAAI*, volume 1, page 3.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Leeson, L., Saeed, J. I., and Grehan, C. (2015). 18 irish sign language (isl). *Sign Languages of the World: A Comparative Handbook*, page 449.

Lockhart, J. W. and Weiss, G. M. (2014). Limitations with activity recognition methodology & data sets. In *Pervasive and Ubiquitous Computing*, pages 747–756.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *ICML*.

Nakjai, P. and Katanyukul, T. (2019). Hand sign recognition for thai finger spelling: An application of convolution neural network. *Journal of Signal Processing Systems*, 91(2):131–146.

Oliveira, M., Chatbri, H., Ferstl, Y., Farouk, M., Little, S., O'Connor, N. E., and Sutherland, A. (2017a). A dataset for irish sign language recognition. In *IMVIP*.

Oliveira, M., Chatbri, H., Ferstl, Y., Farouk, M., Little, S., O'Connor, N. E., and Sutherland, A. (2017b). A dataset for irish sign language recognition. In *IMVIP*.

Oliveira, M., Chatbri, H., Little, S., Ferstl, Y., O'Connor, N. E., and Sutherland, A. (2017c). Irish sign language recognition using principal component analysis and convolutional neural networks. In *DICTA*, pages 1–8. IEEE.

Oyedotun, O. K. and Khashman, A. (2017). Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications*, 28(12):3941–3951.

Pigou, L., Van Herreweghe, M., and Dambre, J. (2016). Sign classification in sign language corpora with deep neural networks. In *LREC*, pages 175–178.

Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR*, pages 806–813.

Shi, B., Del Rio, A. M., Keane, J., Michaux, J., Brentari, D., Shakhnarovich, G., and Livescu, K. (2018). American sign language fingerspelling recognition in the wild. In *SLT*, pages 145–152.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Zhang, D., Yao, L., Chen, K., and Monaghan, J. (2019). A convolutional recurrent attention model for subject-independent eeg signal analysis. *IEEE Signal Processing Letters*, 26(5):715–719.

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., and Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*.