The Key Points: Using Feature Importance to Identify Shortcomings in Sign Language Recognition Models

Ruth Holmes, Ellen Rushe, Anthony Ventresque
School of Computer Science and Statistics, Trinity College Dublin & SFI Lero
{holmesru, ellen.rushe, anthony.ventresque}@tcd.ie

Abstract

Pose estimation keypoints are widely used in Sign Language Recognition (SLR) as a means of generalising to unseen signers. Despite the advantages of keypoints, SLR models struggle to achieve high recognition accuracy for many signed languages due to the large degree of variability between occurrences of the same signs, the lack of large datasets and the imbalanced nature of the data therein. In this paper we seek to provide a deeper analysis into the ways that these keypoints are used by models in order to determine which are most informative to SLR, identify potentially redundant ones and investigate whether keypoints that are central to differentiating signs in practice are being effectively used as expected by models.

Keywords: Sign Language Recognition, Feature Importance, Low-Resource Languages

1. Introduction

Training effective sign language recognition (SLR) models is a challenging task due to, not only the complex, multifaceted nature of signed languages, but also the lack of labelled data available to train models. Pose estimation models, such as MediaPipe (Lugaresi et al., 2019), have provided a reliable means of detecting the most salient parts of the body, termed *keypoints*, by training on large, diverse datasets allowing SLR models to be built with more generalised features than would be possible when training on a small dataset of images alone (Holmes et al., 2022). Another advantage of using pose estimation keypoints is that we can restrict the features that a model has access to if these features are redundant to the task at hand. Our ability to reduce the dimensionality of inputs is especially helpful when considering the small number of datapoints available relative to their high degree of variability. On the other hand, there are also some keypoints that *should* be integral to differentiating between signs, and we should expect that this be reflected by models. This means that feature importance can also potentially be used to identify issues in the underlying representations.

This paper seeks to explore the importance assigned to different keypoints by a recent state-of-the-art model for low resource signed languages. In particular, we explore the case of Flemish Sign Language (Vlaamse Gebarentaal, VGT). We seek to establish the keypoints that are assigned the least overall importance, whether these features are truly redundant or whether their apparent lack of importance indicates underlying issues with pose estimation, dataset construction or model architecture.

2. Related Work

Understanding the features that are more or less important is vital to our understanding of the underlying factors that lead to model classification performance. However, pose estimation keypoint importance is to-date largely unexplored.

Several works have performed permutation feature importance utilizing various forms of sensor data. For instance, Tateno et al. (2020) record surface electromyography signals from participants' forearms during signing and extract ten time and frequency-domain features for importance evaluation. Calado et al. (2021) utilize data collected via a sensory glove and inertial measurement units, and evaluate by combining permutation importance with a LOSO cross-validation on the training set. Sensors, however, are often invasive or require specialist equipment. In our work, we instead cast our attention on SLR from video using pose estimation.

Methods in feature selection are far more common in existing SLR works with studies such as (Bhuvan et al., 2016) and (Marin et al., 2016) employing F-score, Sequential, Random Forests and Extra Trees selection strategies. Alternatively, Bansal et al. (2022) use a hybrid mRMR-PSO (Minimum Redundancy and Maximum Relevance - Particle Swarm Optimization) algorithm for feature selection on several hand posture and finger-spelling datasets. Here mRMR is used initially to select informative features which are then further optimized using PSO. Feature selection, however, relies on the performance of the SLR model used. Performing feature selection using these models assumes they are correct in most cases. In this work, we do not make such an assumption and, instead, seek to use feature importance to understand model behaviour and identify deficits.

LREC-COLING 2024, pages 15970–15975
20-25 May, 2024.

3. Methodology

3.1. Keypoint Importance

Permutation Feature Importance (Breiman, 2001) is a model-agnostic procedure for determining the features that most contribute to the performance of a trained model. The procedure typically involves randomly shuffling the values of the feature of interest in order to remove any association between that independent feature and the target variable. If this feature contributes significantly to correct predictions, this shuffling operation should result in a marked decrease in performance. Conversely, if the decrease in performance is negligible, this suggests that this feature does not contribute significantly to classification decisions. We use this measure of importance with a slight modification which avoids the potentially costly shuffling operation. We, in place of shuffling a given feature, replace it with values drawn uniformly at random within the range of all features in the dataset. A more precise description of this procedure is given in Algorithm 1 and an example process is provided in our adjoining GitHub repository[1].

Algorithm 1 Permutation Feature Importance

```
Input: $f(.,\theta)$: Trained model with parameters $\theta$,
    $X^{(n\times m)}$: Testing keypoint matrix , $y$: Testing
    label vector, $k$: number of randomizations per
    feature.
$s_{baseline} = F1_{score}(y, f(X, \theta))$
for $d \leftarrow 1$ to $m$ do
    $\hat{X} = X$
    for $i \leftarrow 1$ to $k$ do
        $\hat{X}[*][d] = U \sim (X_{min}, X_{max})$
        $s_d^i = F1_{score}(y, f(\hat{X}, \theta))$
        $\delta_d^i = s_{baseline} - s_d^i$
    end for
    $\delta_d = \frac{1}{k}\sum_i^k \delta_d^i$
end for
```

4. Experimental Setup

In this section, we will discuss the data used in experiments, the model configuration and evaluation strategy.

4.1. Data

The dataset used for experiments was derived from Corpus VGT (Van Herreweghe et al., 2015). This dataset contains footage of individuals signing in Flemish Sign Language (Vlaamse Gebarentaal, VGT). The continuous nature of the signing in this dataset is of particular interest as this reflects the most realistic representation of real-world sign language use. These videos are broken down into word-level clips based on available gloss-tier annotations, resulting in a total 24,967 samples. A stratified split is performed to ensure a similar distribution of the 292 classes across the training, validation and test sets. Additionally, a grouped split is performed to ensure that the data configuration is signer-independent. There are 111 participants present in the data.

4.2. Keypoint Extraction

MediaPipe Holistic (Google, 2023), specifically the pose and hand solutions, are used to extract pose estimation keypoints from the aforementioned clips. MediaPipe uses the BlazePose (Bazarevsky et al., 2020) pose detection model to extract 'regions of interest'. Once these regions are extracted, face and hand-specific detectors are used to determine the remaining keypoints. Both the pose and facial model, BlazePose and BlazeFace (Bazarevsky et al., 2019) respectively, are lightweight deep convolutional networks designed for real-time applications. The hand-specific detector, MediaPipe Hands (Zhang et al., 2020), consists of both a palm detection model, which produces a crop of the hand, and a keypoint detection model, which infers whether a hand is present, the handedness and the actual keypoints of the hand present within this crop.

In addition, following keypoint extraction, all non-randomised keypoints are pre-processed in line with recent work (Holmes et al., 2023). Briefly, this entails the use of linear interpolation between frames in scenarios where MediaPipe fails to detect keypoints. The resulting keypoints are then scaled in such a way that the distance between shoulder co-ordinates for all samples is one, and keypoints are shifted to the centre of the chest.

4.3. Model

We use the deep SLR architecture and hyperparameters recently detailed in (Holmes et al., 2023), which are summarised in our adjoining GitHub repository[1]. Briefly, this model uses a combination of frame embeddings to learn frame-wise representations, learns local temporal information between these embeddings using convolutional layers and then encodes global temporal information using self-attention (Vaswani et al., 2017) across these latent representations. All models, including the baseline, use this configuration. For evaluation, we must consider that sign language data, like most language datasets, tend to be extremely imbalanced. In order to ensure that the performance of majority glosses do not conceal poor recognition

[1] https://github.com/hruth9/lrec-coling24

performance of minority glosses, we use a macro averaged F1-measure for evaluation. The difference in this measure is then computed for each feature as previously described in Algorithm 1.

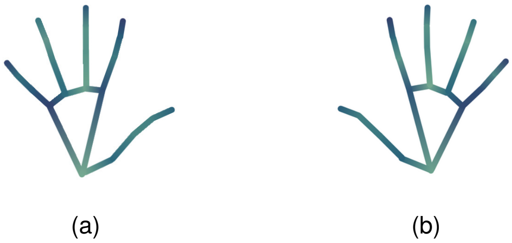

Figure 1: Right-hand 1a and left-hand 1b importance heatmap - darker colors indicate more important keypoints.

5. Results

In this section, we will detail the most salient features based on feature importance broken down by the primary keypoint clusters used in experiments: The body, right-hand and left-hand.

5.1. Hands

Figure 2 and 3 show the mean feature importance for the right and left hand respectively as defined in line 9 or Algorithm 1 over 30 random initialisations. The most notable observation from our analysis of the right-hand keypoints is that the fingers on the perimeter of the hand (i.e., the index and pinky finger) appear to be assigned more importance than the inner two fingers (i.e., the middle and ring finger). This could indicate that these fingers are under-utilised by the model due to a higher level of occlusion than the outer fingers, given their position. It is also evident that the base (“MCP") and top of the fingers are more important than the inner joints (“PIP" and “DIP"). This could not only indicate a high level of occlusion but is likely due to the lack of depth information. Another potential cause could be the varying importance of these fingers depending on the sign. If these fingers do not feature prominently in a large number of signs within the dataset, this could bias the model to under-utilise feature representations based on these fingers when making classifications, resulting in poor classification for signs where these keypoints are crucial. Upon inspecting per-gloss feature importance, we find that there are several cases where the inner fingers are assigned a large degree of importance, indicating that occlusion may not be the primary issue in many cases, but more so the varying degree of use for each digit. We find the same pattern for left-hand keypoint in Figure 3, with the majority of important keypoints being assigned to the base and tip of the fingers on the outer parts of the hand.

5.2. Body

The body keypoints used in this work contain facial, shoulder, torso and limbs along with a number of coarse-grained hand keypoints. Figure 5 shows the mean feature importance for these keypoints. Predictably, hand co-ordinates such as the pinky and index fingers are most important along with the wrists, with the co-ordinates on the right-hand side showing higher importance scores, presumably due to the majority of individuals being right-handed. Beyond these, most of the keypoints of lesser importance are below the waist with the notable exception of some of the facial keypoints. Though one mouth keypoint ("right mouth") is seemingly important to classification, the left-hand side does not appear to be as significant. This could be due to the heads of signers being slightly occluded due to their orientation in videos. The most noteworthy observation from this analysis is the lack of importance assigned to facial features, with only the right-hand side of the mouth ranking close to hand keypoints. Mouthing is an integral feature of signed languages across the world and, in this case, VGT. It is therefore surprising that these keypoints are not assigned as much importance as hand keypoints, especially considering that mouthing can be used to disambiguate signs (Mohr, 2014). We also observe that eye keypoints are minimally utilised by the model to compute classifications even though eye gaze is intrinsically linked to several linguistic features (Vermeerbergen et al., 2007). This shows that these keypoints only minimally assisted in discriminating between signs. This suggests that the keypoints used here may not provide a sufficient level of detail to effectively reflect facial movements.

5.3. Discussion

Based on our above analysis, we next provide a summary and discussion of our main observations and provide suggestions that may remedy some of the challenges raised.

Keypoint Distortion The lack of accurate depth information in pose estimation frameworks is a known issue (De Coster et al., 2023; Moryossef et al., 2023). In fact, the MediaPipe documentation itself states that the *z*-coordinate should be discarded as “the model is not fully trained to predict depth” (Google, 2023). This, along with the manner in which these joints are often occluded due to their position can contribute to keypoint distortion. Given the intricacies of continuous signing, it would clearly be beneficial to pay closer attention to the

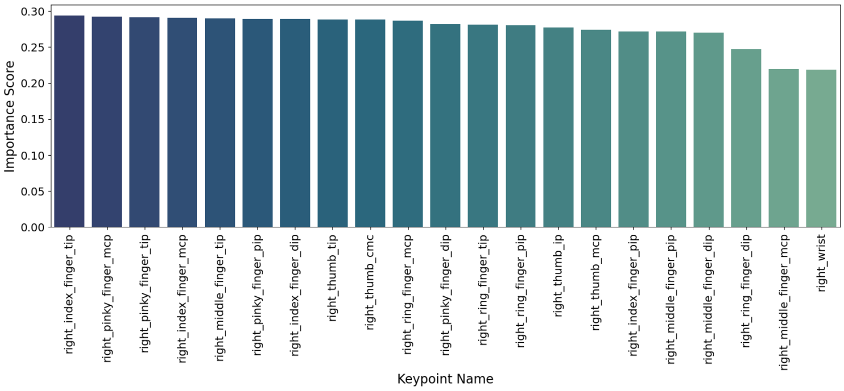

Figure 2: Importance of right hand keypoints on test set.

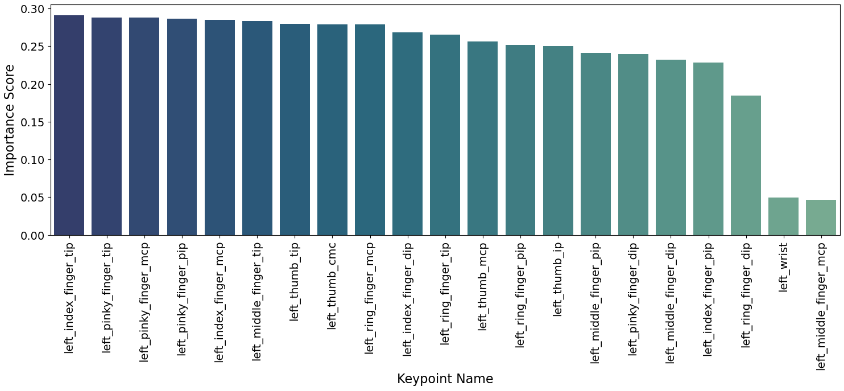

Figure 3: Importance of left hand keypoints on test set.

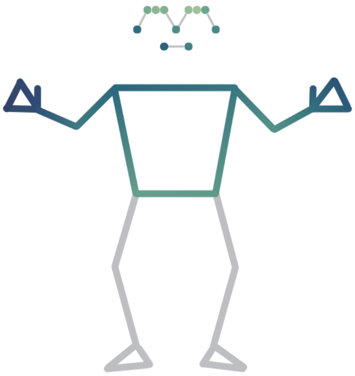

Figure 4: Importance of body keypoints on test set.

quality of pose estimation keypoints generated for middle and ring fingers on both hands, with special attention given to the joints in the fingers.

Gloss Imbalance We observe that due to the fact that some keypoints feature more prominently across the entire dataset, there is an imbalance in the number of signs that accentuate each keypoint. Given that deep learning models use an aggregation of the classification error across multiple examples to compute loss, feature imbalances will inevitably lead to an over-reliance on keypoints that are prominent in a large number of signs, at the expense of the classification performance on glosses centred around other keypoints. This effect could potentially be remedied by more focused data augmentation centred around less prominent keypoints or on signs where they are more dominant.

Facial Features The facial keypoints used in our analysis seem strikingly under-utilised. We hypothesize that this is due to the coarse grained nature of the facial features used (Mediapipe's Pose landmarker model[2]). Though detailed face meshes can have a reasonably large dimensionality relative to that of the body, we hypothesize that a more careful selection of facial keypoints, tailored specifically to SLR from a finer-grained mesh (such as MediaPipe's Face Mesh [3]) could provide crucial details that currently appear not to be represented.

Feature Importance for Feature Selection Though we have shown that feature importance is informative to identify potential deficiencies in

[2] MediaPipe Pose Landmarker
[3] MediaPipe Face Landmarker

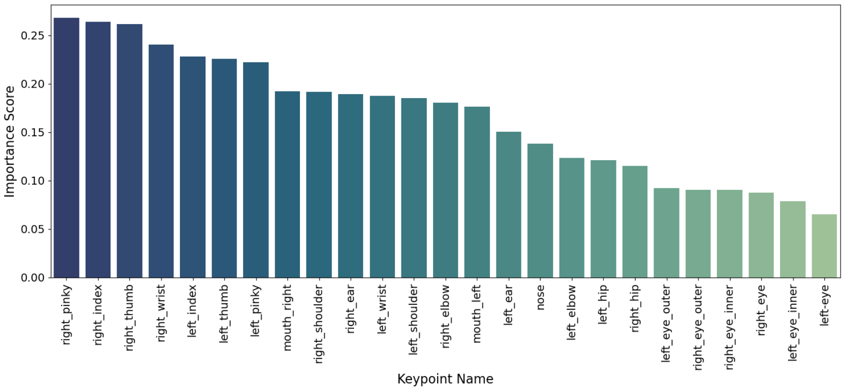

Figure 5: Importance of body keypoints on test set.

existing models, in doing so we have also shown that feature importance does not always translate to linguistic importance, with several keypoints being assigned low importance despite their linguistic utility. We therefore emphasize that we should ensure that models are sufficiently accurate, address issues of imbalance and interrogate the linguistic utility of keypoints that may not be captured by existing models before using this measure of importance as a feature selection tool.

6. Conclusion

In this paper we have analysed feature importance for a state-of-the-art SLR model on Corpus VGT in order to gain a better understanding of how this model uses the pose estimation keypoints provided. We have found that a low level of importance is assigned to a number of keypoints that are often crucial to differentiating signs in practice. We have outlined some potential causes for this lack of assigned importance and suggested avenues to potentially improve these representations to make them more helpful to SLR models. However, we acknowledge the limitation that is restricting our study to a single language in particular. In future, we aim to expand our work to include additional sign languages to provide a more language-agnostic evaluation of feature importance. It would also be of interest to include additional pose-based SLR architectures in such experiments to determine whether this reveals any overlapping trends in feature utilisation.

7. Acknowledgements

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).

8. Bibliography

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Sandhya Rani Bansal, Savita Wadhawan, and Rajeev Goel. 2022. mrmr-pso: A hybrid feature selection technique with a multiobjective approach for sign language recognition. *Arabian Journal for Science and Engineering*, 47(8):10365–10380.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. 2019. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*.

Malladihalli S Bhuvan, D Vinay Rao, Siddharth Jain, TS Ashwin, Ram Mohana Reddy Guddetti, and Sutej Pramod Kulgod. 2016. Detection and analysis model for grammatical facial expressions in sign language. In *2016 IEEE Region 10 Symposium (TENSYMP)*, pages 155–160. IEEE.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Alexandre Calado, Vito Errico, and Giovanni Saggio. 2021. Toward the minimum number of wearables to recognize signer-independent italian sign language with machine-learning algorithms.

IEEE Transactions on Instrumentation and Measurement, 70:1–9.

Mathieu De Coster, Ellen Rushe, Ruth Holmes, Anthony Ventresque, and Joni Dambre. 2023. Towards the extraction of robust sign embeddings for low resource sign language recognition. *arXiv preprint arXiv:2306.17558*.

Brigitte Garcia and Marie-Anne Sallandre. 2020. Contribution of the semiological approach to deixis–anaphora in sign language: the key role of eye-gaze. *Frontiers in Psychology*, 11:583763.

Google. 2023. Mediapipe holistic. https://github.com/google/mediapipe/blob/master/docs/solutions/holistic.md.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Ruth Holmes, Ellen Rushe, Mathieu De Coster, Maxim Bonnaerens, Shin'ichi Satoh, Akihiro Sugimoto, and Anthony Ventresque. 2023. From scarcity to understanding: Transfer learning for the extremely low resource irish sign language. In *Proceedings of the Eleventh International Workshop on Assistive Computer Vision and Robotics, in conjunction with IEEE/CVF International Conference on Computer Vision*, pages 2008–2017.

Ruth Holmes, Ellen Rushe, Frank Fowley, and Anthony Ventresque. 2022. Improving signer independent sign language recognition for low resource languages. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 45–52.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Giulio Marin, Fabio Dominio, and Pietro Zanuttigh. 2016. Hand gesture recognition with jointly calibrated leap motion and depth sensor. *Multimedia Tools and Applications*, 75:14991–15015.

Susanne Mohr. 2014. *Mouth actions in sign languages: An empirical study of Irish Sign Language*, volume 3. Walter de Gruyter GmbH & Co KG.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. *arXiv preprint arXiv:2310.13960*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Shigeyuki Tateno, Hongbin Liu, and Junhong Ou. 2020. Development of sign language motion recognition system for hearing-impaired people using electromyography signal. *Sensors*, 20(20):5807.

Mieke Van Herreweghe, Myriam Vermeerbergen, Eline Demey, Hannes De Durpel, Hilde Nyffels, and Sam Verstraete. 2015. Het Corpus VGT. Een digitaal open access corpus van videos and annotaties van Vlaamse Gebarentaal, ontwikkeld aan de Universiteit Gent ism KU Leuven. www.corpusvgt.be.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Myriam Vermeerbergen, Lorraine Leeson, and Onno Crasborn. 2007. Simultaneity in signed languages: A string of sequentially organised issues. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 281:1.

Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*.