

# Towards an Interpretable Spanish Sign Language Recognizer

Itsaso Rodríguez-Moreno<sup>a</sup>, José María Martínez-Otzeta<sup>b</sup>, Izaro Goienetxea<sup>c</sup> and Basilio Sierra<sup>d</sup>

*Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU),  
Donostia-San Sebastián, Spain*

**Keywords:** Gesture Recognition, Spanish Sign Language, Interpretability.

**Abstract:** A significant part of the global population lives with hearing impairments, and the number of affected people is expected to increase in the coming decades. People with hearing problems experience daily difficulties in their interaction with non-deaf people, due to the lack of a widespread knowledge of sign languages by the general public. In this paper we present a blueprint for a sign language recognizer that takes advantage of the internal structure of the signs of the Spanish Sign Language (SSL). While the current dominant approaches are those based in deep learning and training with lot of recorded examples, we propose a system in which the signs are decomposed into constituents which are in turn recognized by a classical classifier and then assessed if their combination is congruent with a regular expression associated with a whole sign. While the deep learning with many examples approach works for every possible collection of signs, our suggestion is that we could leverage the known structure of the sign language in order to create simpler and more interpretable classifiers that could offer a good trade-off between accuracy and interpretability. This characteristic makes this approach adequate for using the system as part of a tutor or to gain insight into the inner workings of the recognizer.

## 1 INTRODUCTION

Sign languages are the main form of communication for a large proportion of people with hearing impairments. There is a great diversity of sign languages, because its evolution shares similar characteristics with spoken languages. While a significant number of non-deaf people learn non-native spoken languages out of necessity, or for professional or just intellectual reasons, deaf people tend to feel isolated even in its native communities due to the lack of interest of the general public for the sign languages.

Sign languages are quite complex, with rich grammatical structures and regional and international diversity, which makes the task of translating them into spoken languages very challenging. The signs are performed mainly with the hands, but the body position and the facial expression are also important. The hand which performs the more complex movements and moves the most is the dominant hand in the sign generation, which usually is also the dominant hand

in the everyday life of the sign speaker (left for left-handed, right for right-handed). A sign language recognizer should take into account hands, body and facial expression to perform its task correctly.

In order to favor the integration of sign language speakers, technological solutions have been devised to bridge the communication gap (Wadhawan and Kumar, 2021; Cheok et al., 2019; Er-Rady et al., 2017; Ong and Ranganath, 2005). The sign language recognition task can be divided in two main phases; the data acquisition and the classification. Regarding data acquisition there are two different approaches:

- Non-vision based, which make use of different sensors to get the information of the sign that is being performed, such us IMU (Inertial Measurement Unit) or WiFi.
- Vision based, where the acquired data are images recorded by a camera.

In addition, some of these data acquisition systems can be intrusive, for example when using data gloves, body trackers, or even colored gloves to perform hand segmentation. Depending on the captured data, different preprocessing and feature extraction methods are used (segmentation, dimensionality reduction,...). Concerning the classification, Hidden Markov Mod-

<sup>a</sup> <https://orcid.org/0000-0001-8471-9765>

<sup>b</sup> <https://orcid.org/0000-0001-5015-1315>

<sup>c</sup> <https://orcid.org/0000-0002-1959-131X>

<sup>d</sup> <https://orcid.org/0000-0001-8062-9332>

els (HMM) and Neural Networks (NN) are widely used. There is a difference between classifying static or dynamic signs; if the signs are static, a single frame has to be classified, while in dynamic signs, temporal information should also be considered.

The studies published so far have mostly focused on classifying isolated, static, one-handed signs captured by a camera and using a NN for classification, being American Sign Language (ASL) the most studied language.

In relation to Spanish Sign Language, in (Parcheta and Martínez-Hinarejos, 2017) the authors use HMMs to recognize 91 different signs captured by the Leap Motion sensor. The analyzed signs include dynamic gestures and sentences. Different HMM topologies are used, where the number of states is changed. In (Vazquez-Enriquez et al., 2021) the authors use two different architectures to perform isolated sign language recognition: a 3D Convolutional Neural Network (3D CNN) called S3D (Xie et al., 2018) for RGB data and a skeleton-based architecture called MS-G3D (Liu et al., 2020). In addition to two other datasets, they classify a subset of the LSE\_UVIGO (Docío-Fernández et al., 2020) SSL dataset. The authors of (Martínez-Martin and Morillas-Espejo, 2021) created a dataset with the Spanish alphabet which includes static and motion gestures, 18 letters and 12 letters respectively. The keypoints of the hands and arms extracted with OpenPose (Cao et al., 2019) are used to create the images which are used to perform the classification. They tried different CNN and Recurrent Neural Network (RNN) architectures to classify signs, taking into account the importance of temporal information in signs which require motion.

While many works have focused on providing some sort of feedback for spoken language learners (Pennington and Rogerson-Revell, 2019; Robertson et al., 2018), very few are dedicated to gestures in general (Banerjee et al., 2020), an even less to sign language (Paudyal et al., 2019). The aim of the system presented here is two-fold: to provide developers of machine learning models a visual way of testing and interpreting the predictions of their models, and to provide sign test students a visual and textual feedback about their performance. As a first step, only signs for which only a hand is needed are currently considered. The signs are formalized as sequences of hand configurations, where the sequence is defined as a regular expression, and the hand configurations have been learned from features derived from the spatial location of the different parts of the hand. The comparison between the intended and the recognized action is analyzed in two levels: hand configuration and sign.

The system is able to label the detected hand configuration and show the rationale of its prediction, and also the comparison with the intended sign, if they differ. With respect to the whole sign as a regular expression, where the underlying alphabet is the set of hand configurations, an explanation is also provided.

The rest of the paper is organized as follows. First, in Section 2 some basic concepts of Spanish Sign Language are explained in order to introduce the topic. In Section 3 the proposed approach is introduced, explaining the process that has been carried out. In Section 4 a discussion is presented and finally, in Section 5 the conclusions extracted from this work are presented and future work is pointed out.

## 2 SIGN LANGUAGE STRUCTURE

A sign is a combination of complex articulation positions and movements performed by a single hand (one-handed) or both hands (two-handed). In one-handed signs, the dominant or active hand is used to perform the sign. However, in two-handed signs, when the sign is symmetrical both hands act the same way, but in non-symmetrical signs the dominant hand moves while the passive hand serves as a base. Usually, the dominant or active hand is the right hand for right-handed people and the left hand for left-handed people.

Signs have four different elements (Blanco, 2009), which are equivalent to the phonemes of oral languages, and together they compose the articulation of the sign:

- Location (+ contact): the specific location where signs are performed. If a sign is performed in a corporal location, it can be in contact with that body part (+ contact) or not.
- Configuration (shape): the shape of the hand when performing the sign.
- Orientation: the orientation hands adopt when performing a sign.
- Movement: the movement usually done from the location when performing a sign.

In brief, to perform a sign, the dominant hand is placed in a location, it adopts a certain configuration and orientation in or on it, and usually performs a movement starting from that location. Nevertheless, in addition to these elements, there are some non-manual components which are fundamental to define a sign: the facial expression (eyebrows, eyes, cheeks, nose, lips, tongue) and the position of the head, shoulders or body.

As mentioned before, the shape of the hand when performing a sign is defined as a configuration. In SSL, there are three types of configurations: phonological (*queirema*), dactylogical and numerical. The phonological configurations obey a phonological system, as the distinctive sounds in oral languages, and can be classified according to different characteristics:

- Palm: extended or closed (fist).
- Fingers:
  - Extended, flexed or closed.
  - Glued or separated.
  - Which fingers are involved: index; thumb; index and thumb; middle; middle and thumb; index and middle; index, middle and thumb; pinky; pinky and index; pinky and thumb.
  - Thumb opposes the articulation of the others.

The dactylogical configurations of SSL represent the letters of the Spanish alphabet. These are used mostly when signing proper names. Lastly, the numerical configurations symbolize the natural numbers, both in isolation and incorporated in another sign.

In (Gutierrez-Sigut et al., 2016) a database of SSL is presented, where each sign is defined with the elements mentioned above, including the configurations. All the configuration and sign definitions in which this research is based have been obtained from this source.

### 3 PROPOSED APPROACH









In this section, the proposed approach and the followed pipeline are explained step by step.

**Data COLLECTION.** Although different elements as hand configuration, position, orientation or movement play a key role when recognizing a sign, as a first approach, we based the sign recognition in the recognition of different configurations and the movement from one configuration to another.

As a first approach, the eight different phonological configurations shown in Table 1 have been selected. These configurations are constituents of a wide variety of Spanish Signs as indicated in Table 1.

In the same vein, five different signs of the SSL have been chosen among the signs that use the previously selected configurations: well (*bien*), happy (*contento*), woman (*mujer*), man (*hombre*) and lis-

Table 1: Presence of selected configurations as constituents of SSL one-handed signs.

Configuration	#Signs	Configuration	#Signs
 4	124	 73	19
 50	189	 74	29
 58	55	 77	23
 59	235	 78	24

tener (*oyente*). The definitions of the mentioned signs are presented in Table 2.











A data set composed with images of the configurations which form those signs has been created. There are about 700 images for training each configuration. These values are shown in Table 3.

**Model GENERATION.** In Figure 1, the followed pipeline is shown graphically. Briefly, the method can be divided into two parts. The former is focused on the recognition of the configuration in static images, while the latter predicts the signs performed in a video using the previously trained configurations model as basis. In order to facilitate the whole process and make it easier to understand, a web app has been developed to both train new models and perform real time classification.

Since, as a first approach, it has been decided to use just the information of the hands to recognize the sign that is being performed, MediaPipe (Lugaresi et al., 2019) has been used to track the position of the hand in both images and videos. Specifically, MediaPipe Hands Tracking (Zhang et al., 2020) has been used, which offers a real-time hand tracking solution which includes 21 hand landmarks for each hand. Each hand landmark is composed of three values ( $x, y, z$ ), representing the coordinates of the key-point. In the case of the videos, these 21 landmarks are extracted for each frame.

After obtaining the landmarks for every image of the configurations data set, the features that are going to be used for training the model have to be selected. Apart from the already mentioned 21 hand-landmarks, there is the option to add the distance between finger tips or the distance from finger tips to thumb tip. These features can be selected all together,

Table 2: Definitions of the selected signs.

SIGN	INITIAL FACIAL/CORPORAL LOCATION	FINAL FACIAL/CORPORAL LOCATION	INITIAL HAND CONFIGURATION	FINAL HAND CONFIGURATION	MOVEMENT PATH
Well	Chin	High neutral space	 58	 59	Straight
Happy	Under the chin	Under the chin	 77	 78	
Woman	Right side of the neck	Under the right ear	 73	 74	Straight
Man	Close to the forehead	Close to the forehead	 4	 4	Straight
Listener	Chin	Chin	 50	 50	Circular

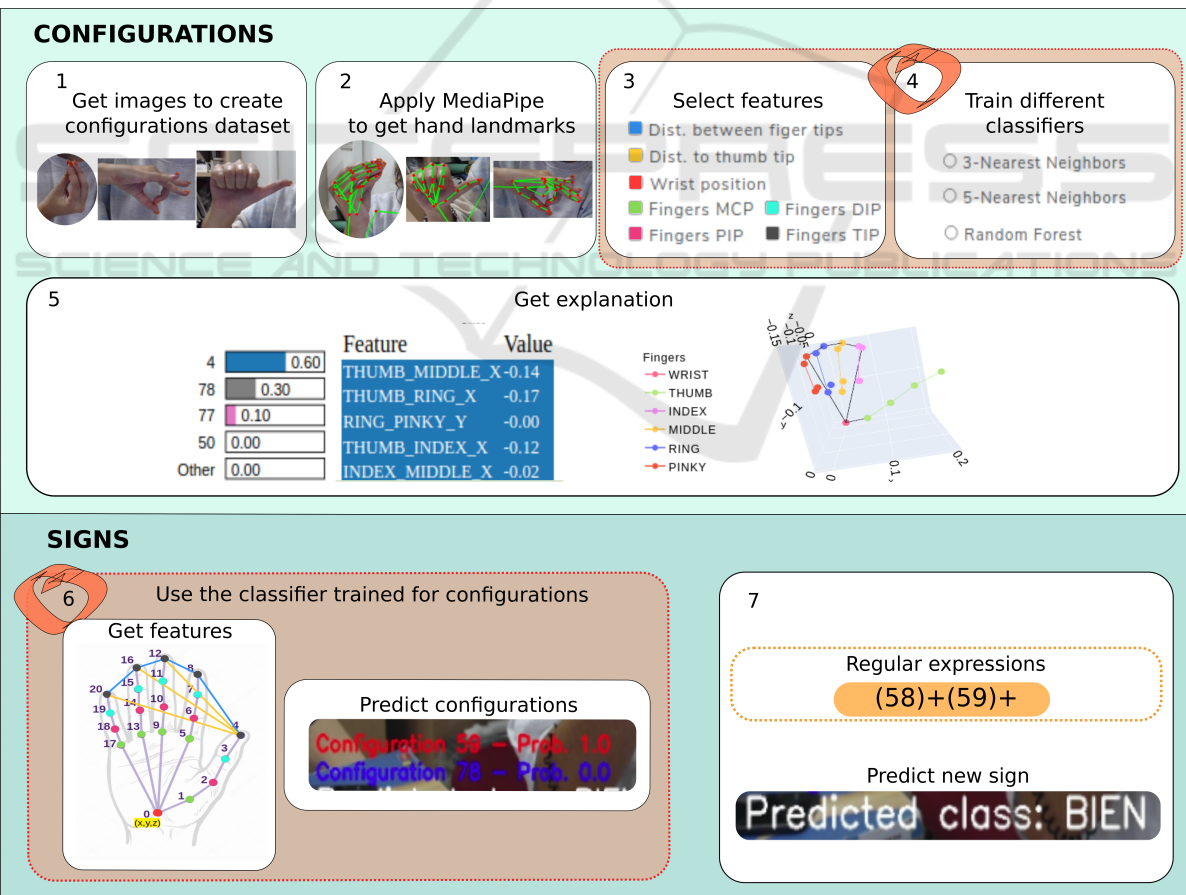


Figure 1: Pipeline. Colours in step 3 refer to positions in step 6 (MCP: Metacarpophalangeal joint; PIP: Proximal Interphalangeal joint; DIP: Distal Interphalangeal joint; TIP: Fingertip).

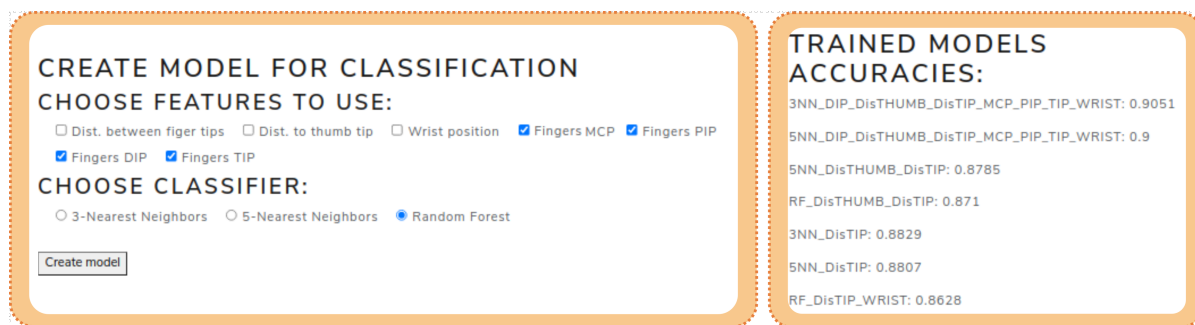


Figure 2: Training configuration model: choose features and classifier.

Table 3: Data-set.

Signs	Configurations	Number of images
Well	58	747
Man	4	700
Woman	73	732
Happy	77	668
Listener	50	585

one by one or in every possible combination. Apart from that, Random Forest or K-Nearest Neighbors ( $K = 3, 5$ ) classifiers can be trained. In Figure 2 it can be seen how the training process of the configurations model is done through the web app. The accuracy values obtained for the training models are displayed aside, which can be helpful when deciding which model to use for new case predictions.

**Prediction AND EXPLANATION.** Once a model is trained, the prediction of a new image of a configuration can be done as it can be seen in Figure 3. As the goal is to develop a tutor for SSL, there is the option to choose which configuration do you want to practice. This way, an image of the configuration is shown in order to guide the user. Among all the trained models, one has to be chosen to make the new predictions. So as to decide which one to select, the accuracy values shown above give a clue of the performance of each of the trained models. If the predicted configuration corresponds to the one selected to practice, the prediction text is displayed with green background. However, if it does not match, a red background is set.

Sometimes, it is quite difficult to understand the logic behind the predictions made by a model. If an explanation of the predicted configuration is required (*Explain results* button is pressed), the two graphical items shown in Figure 4 are added, giving an explanation for a frame prediction. On the one hand, a 3D-graph is created which shows the hand landmarks

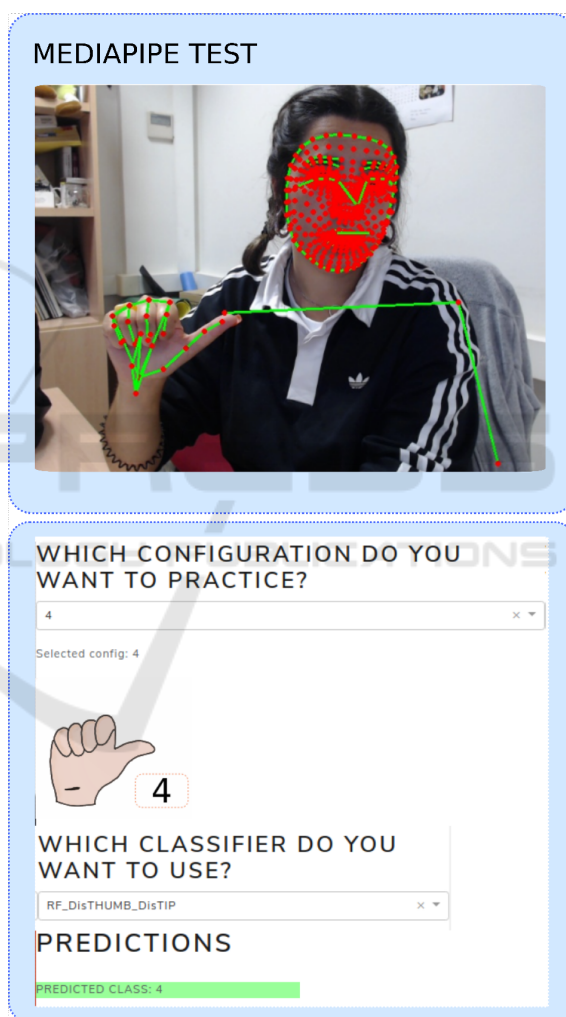


Figure 3: Real-time configuration prediction.

obtained by MediaPipe. Although the output of MediaPipe is also shown over the image the camera is recording (see top side of Figure 3), this 3D-graph mainly helps to verify if the obtained z-coordinates are correct, because they are estimated by MediaPipe from a 2D image. On the other hand, an explanation of the given prediction is obtained by LIME (Ribeiro

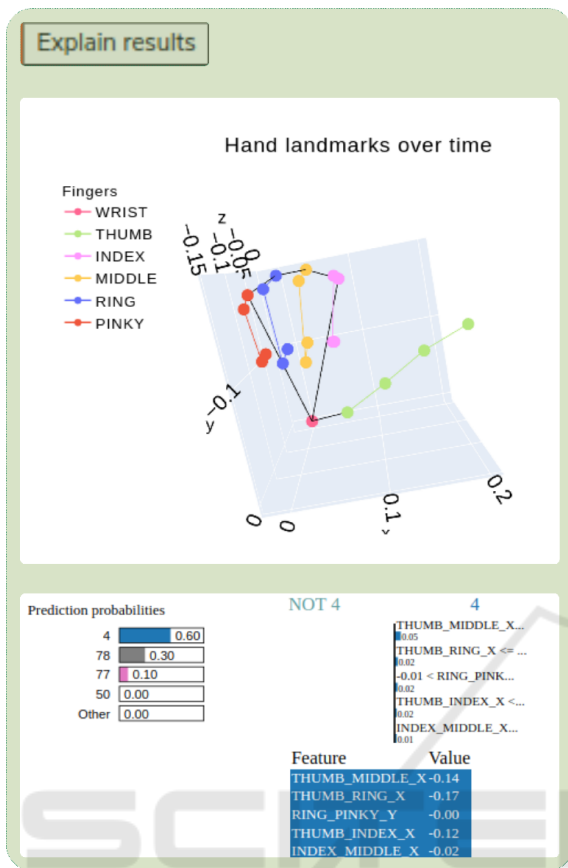


Figure 4: Explanation of the predicted configuration.

et al., 2016), a modular explanation technique which learns a local interpretable model around the prediction to give an explanation of predictions made with any classifier. As it can be seen at the bottom of the Figure 4, LIME offers several information. On the left side, the probability value of each label is indicated and, on the right side, the values of the most informative features are shown. These features might be the most informative either because they help to confirm the predicted class or because the values some of the features take indicate that the class can not be the predicted one. This way, it can be known which features have more impact when making a prediction.

Since each frame is labeled with a configuration by the classifier, a video can be summarized in a series of consecutive configuration names. Thus, a vector of configurations is obtained, a value for each frame of the video, and different regular expressions can be used to evaluate these vectors and decide which sign has been performed. The definition of the expressions can be seen in Table 4, which match with the definitions of the signs.

Using the definitions of the regular expressions, the prediction of new gestures can be performed in

Table 4: Regular expressions for each sign.

Sign	Regular expression
Well	'(58)+(59)+'
Happy	'(77)+(78)+'
Man	'(4)+(4)+'
Woman	'(73)+(74)+'
Listener	'(50)+(50)+'

real time. It has been decided to establish a sliding window of length 25 and step 1 to recognize a tentative sign, being the final prediction the mode of the last 10 predicted signs. In order to avoid the noise of incorrectly predicted configurations in between, it has been decided to establish another sliding window (within the sliding window of 25 frames) of length 10 and step 1. For each window the mode of the configurations belonging to that window is achieved, thus obtaining an array of 16 configurations ( $size\_gesture - size\_window + 1$ ) which will be the one evaluated with the regular expressions.

In the developed application, as with the configurations, it is requested to choose the gesture which is being performed to be able to indicate whether it is performed correctly or not. The models have to be chosen among the trained ones. As it can be seen in Figure 5, in addition to the predicted sign, the probability of the two most likely configurations are also indicated in order to understand the prediction. As long as a sign has not been performed (as mentioned before, the gesture length is set to 25) there is no prediction. Once a prediction can be made, a green background is established if the prediction coincides with the chosen sign and red if it does not match with the sign that was intended to reproduce.

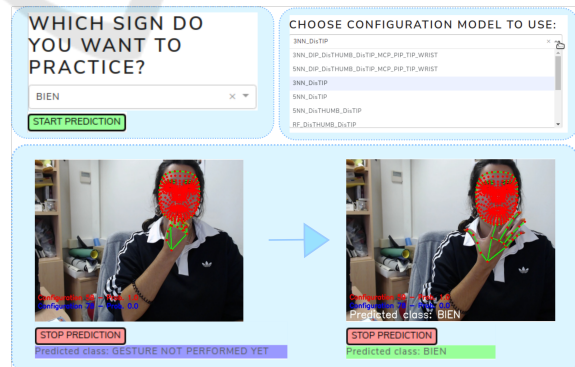


Figure 5: Real-time sign recognition.

## 4 DISCUSSION

The main goal behind the presented approach is to develop a tutor for learning Spanish Sign Language. Al-

though only the first steps are introduced, this system opens the door to many useful applications.

Since the goal is to support people who are learning sign language, improving the explanation module is crucial. The function of this module is to help to understand the results predicted by the classifier, as knowing what needs to be changed to get the desired answer can be really helpful. If the prediction is the one we expect, we can get the reason why the sign has been well performed. However, if we get an incorrect prediction, the explanation is used to indicate to the user what is being wrongly performed and, this way, the user can correct the aspects that make the sign an incorrect replica of the real sign.

This application can be approached from two different perspectives, one from the expert's side and the other from the user's side. In Table 5 the differences between both perspectives are indicated.

Table 5: Differences between the explanation given to an expert or a user.

Expert	<ul style="list-style-type: none"> <li>- <b>Knowledge:</b> the learning process of the classifier.</li> <li>- <b>Explanation:</b> LIME output.</li> <li>- <b>Action:</b> changes in the definition of the classifier.</li> </ul>
User	<ul style="list-style-type: none"> <li>- <b>Knowledge:</b> the sign.</li> <li>- <b>Explanation:</b> natural language.</li> <li>- <b>Action:</b> changes in the performance of the sign.</li> </ul>

While the expert has information about the learning process and the features that have been used to train the classification model, the user just has the visual information of the sign that he/she is learning. Hence, the information given by LIME has to be translated to natural language for the user to understand. Once the information is given, the user has the possibility to perform the sign again following the indications given by the explanation module. In the case of the expert, if the explanations received indicate that the performance of the classifier is poor (the wrong answers are due to a bad configuration of the model and not due to the performance of the user), some changes have to be done in the definition of the classification model.

As in the developed web application the sign or configuration to perform is indicated, it would be interesting if this explanation module gave information on both the chosen sign (or configuration) and the predicted one. Furthermore, although additional information apart from the hands is not considered yet, for information purposes a sentence could be added indicating the part of the body on which the sign should be performed (e.g. "Perform the sign under the chin").

## 5 CONCLUSION

In this paper the first steps towards a tutor application for learning Spanish Sign Language is presented. In the proposed approach the signs are decomposed in constituents which are in turn recognized by a classical classifier and then assessed if their combination is congruent with a regular expression associated with a whole sign. This way, unlike other systems based in deep learning, a simpler and more interpretable system is proposed, making it adequate to use for tutoring SSL and to better understand the performance of the recognizer.

As further work, we plan to extend the range of signs to recognize. Apart from the hand landmarks, specific body keypoints and the distance between them should be added as features too. Specifically in the signs used, presented in Table 2, the relevant locations are the chin, the ear and the forehead. For instance, adding the distances from the fingertips to them could be useful to distinguish between different signs. In another vein, the explanations LIME offers can be treated and displayed more clearly to the users. Taking as basis the information given for every feature, it can be translated to some sentences to inform the user what he/she should do to improve the performance of each sign (e.g. "Locate your thumb higher") as mentioned in Section 4.

## ACKNOWLEDGEMENTS

This work has been partially funded by the Basque Government, Spain, grant number IT900-16, and the Spanish Ministry of Science (MCIU), the State Research Agency (AEI), the European Regional Development Fund (FEDER), grant number RTI2018-093337-B-I00 (MCIU/AEI/FEDER, UE) and the Spanish Ministry of Science, Innovation and Universities (FPU18/04737 predoctoral grant). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

## REFERENCES

- Banerjee, A., Lamrani, I., Hossain, S., Paudyal, P., and Gupta, S. K. (2020). AI enabled tutor for accessible training. In *International Conference on Artificial Intelligence in Education*, pages 29–42. Springer.
- Blanco, Á. L. H. (2009). *Gramática didáctica de la lengua de signos española (LSE)*. Sm.
- Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh,

- Y. (2019). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186.
- Cheok, M. J., Omar, Z., and Jaward, M. H. (2019). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10(1):131–153.
- Docío-Fernández, L., Alba-Castro, J. L., Torres-Guijarro, S., Rodríguez-Banga, E., Rey-Area, M., Pérez-Pérez, A., Rico-Alonso, S., and Mateo, C. G. (2020). LSE\_UVIGO: A Multi-source Database for Spanish Sign Language Recognition. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 45–52.
- Er-Rady, A., Faizi, R., Thami, R. O. H., and Housni, H. (2017). Automatic sign language recognition: A survey. In *2017 International Conference on Advanced Technologies for Signal and Image Processing (AT-SIP)*, pages 1–7. IEEE.
- Gutiérrez-Sigut, E., Costello, B., Baus, C., and Carreiras, M. (2016). LSE-sign: A lexical database for Spanish sign language. *Behavior Research Methods*, 48(1):123–137.
- Liu, Z., Zhang, H., Chen, Z., Wang, Z., and Ouyang, W. (2020). Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152.
- Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.
- Martínez-Martin, E. and Morillas-Espejo, F. (2021). Deep Learning Techniques for Spanish Sign Language Interpretation. *Computational Intelligence and Neuroscience*, 2021.
- Ong, S. C. and Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(06):873–891.
- Parcheta, Z. and Martínez-Hinarejos, C.-D. (2017). Sign language gesture recognition using HMM. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 419–426. Springer.
- Paudyal, P., Lee, J., Kamzin, A., Soudki, M., Banerjee, A., and Gupta, S. K. (2019). Learn2Sign: Explainable AI for Sign Language Learning. In *IUI Workshops*.
- Pennington, M. C. and Rogerson-Revell, P. (2019). Using technology for pronunciation teaching, learning, and assessment. In *English pronunciation teaching and research*, pages 235–286. Springer.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.
- Robertson, S., Munteanu, C., and Penn, G. (2018). Designing pronunciation learning tools: The case for interactivity against over-engineering. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- Vazquez-Enriquez, M., Alba-Castro, J. L., Docío-Fernandez, L., and Rodríguez-Banga, E. (2021). Isolated Sign Language Recognition With Multi-Scale Spatial-Temporal Graph Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3462–3471.
- Wadhawan, A. and Kumar, P. (2021). Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, 28(3):785–813.
- Xie, S., Sun, C., Huang, J., Tu, Z., and Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321.
- 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*.