Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC 2022), pages 1–8
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

PeruSIL: A Framework to Build a Continuous Peruvian Sign Language Interpretation Dataset

Gissella Bejarano[1,4], **Joe Huamani-Malca**[2], **Francisco Cerna-Herrera**[2],
Fernando Alva-Manchego[3], **Pablo Rivas**[1]
[1]Baylor University, [2]Pontificia Universidad Católica del Perú,
[3]Cardiff University,[4]Universidad Peruana Cayetano Heredia
{gissella_bejaranonic, Pablo_Rivas}@baylor.edu,
{huamani.jn, francisco.cerna}@pucp.edu.pe,
alvamanchegof@cardiff.ac.uk

Abstract

Video-based datasets for Continuous Sign Language are scarce due to the challenging task of recording videos from native signers and the reduced number of people who can annotate sign language. COVID-19 has evidenced the key role of sign language interpreters in delivering nationwide health messages to deaf communities. In this paper, we present a framework for creating a multi-modal sign language interpretation dataset based on videos and we use it to create the first dataset for Peruvian Sign Language (LSP) interpretation annotated by hearing volunteers who have intermediate knowledge of PSL guided by the video audio. We rely on hearing people to produce a first version of the annotations, which should be reviewed by native signers in the future. Our contributions: i) we design a framework to annotate a sign Language dataset; ii) we release the first annotated LSP multi-modal interpretation dataset (AEC); iii) we evaluate the annotation done by hearing people by training a sign language recognition model. Our model reaches up to 80.3% of accuracy among a minimum of five classes (signs) AEC dataset, and 52.4% in a second dataset. Nevertheless, analysis by subject in the second dataset show variations worth to discuss.

Keywords: Continuous Sign Language, Peruvian Sign Language, multi-modal dataset

1. Introduction

An increasing number of calls highlight the need to research sign language, and develop technologies with a multidisciplinary approach. For instance, Bragg et al. (2019) introduced the term Sign Language Processing (SLP) to refer to the task of building models that are able to perform a complete translation process. Similarly, Yin et al. (2021) urges the inclusion of SLP in the more-developed research area of Natural Language Processing (NLP). This can bring enormous benefits in inheriting and adapting the advancements reached in machine translation to sign language translation. For example, several annotation systems have been developed for sign language research with focus in linguistics. However, these annotation systems might not be suitable when working with machine learning models.

Recent advancements in SLP based on computer-vision rely on datasets of *continuous* sign language interpretation. For example, several work is addressing sign language temporal segmentation (Renz et al., 2021a; Renz et al., 2021b) and even aligning subtitles to perform this task (Bull et al., 2020; Bull et al., 2021). On the other hand, (Camgoz et al., 2020b; Camgoz et al., 2018; Camgoz et al., 2020a) focus on sign language recognition and translation. All of these works use at least one dataset of sign language interpreters, such as RWTH-PHOENIX-Weather (Forster et al., 2012) or BSL-1k (Albanie et al., 2020). In that sense, Continuous Sign Language performed by interpreters, and properly reviewed by deaf people, can contribute to the development of more resources for the task. However, to really develop sign language technology and to include deaf people in its design and construction, we need more standardized annotation conventions, less background-controlled videos, and more variations in the topics covered in the datasets.

2. Peruvian Sign Language (LSP)

Peruvian Sign Language (LSP by its acronym in Spanish) is the aboriginal sign language from Peru. There are around half a million deaf people in the country (INEI and CONADIS, 2012), and at least 10,000 people have LSP as their mother tongue (INEI, 2018). LSP is an understudied language that has only recently been officially recognized by the government (MIMP, 2017). Although both public and private institutions are required to provide sign language interpretation in Peru, not many do so since the law is not properly enforced. There are almost no resources in LSP, and the few existing research has primarily studied its grammatical properties (Madrid Vega, 2018) and aspects of their users (Elizabeth and Parks, 2010), or has built a dictionary from a partially annotated dataset (PUCP-DGI) (Rodríguez Mondoñedo and Arnaiz, 2015). On the other hand, computational approaches have only focused on isolated sign language alphabet recognition (Lazo et al., 2019; Mejía Gamarra et al., 2020; Berrú-Novoa et al., 2018; Nureña-Jara et al., 2020).

Although more technological tools are accessible to people with hearing disabilities, they are still based on the written version of a spoken language and not on the main language in which deaf people communicate, for instance. To contribute to bridging that gap, we introduce PeruSIL – a framework for building multi-modal datasets for Continuous Peruvian Sign Language interpretation. Our framework proposes an annotation convention based on the glossing system but simplified, and a pipeline to combine manual and automatic multimodal annotations (Section 2). In addition, we use our framework to create the first multi-modal dataset for Peruvian Sign Language interpretation (Section 3), which includes videos, unaligned audio, transcripts, text, and keypoint landmarks (pose, hands, and facial). For this dataset, original videos were acquired from a Peruvian government's TV program developed for remote school education during the COVID-19 pandemic. We also evaluate the usefulness of our annotated dataset as training data for a sign recognition model, tested in both in-domain and out-of-domain settings (Section 4). Our framework and dataset are part of a larger project aiming to create a larger and online Peruvian Sign language/Spanish dictionary, and an automatic Peruvian Sign Language Translation framework.

3. PeruSIL Framework

In this section, we detail our proposed annotation convention and the pipeline used to build a multi-modal sign language interpretation dataset. We highlight the challenge of collecting sign language datasets due to the lack of videos from native signers, and the limited availability of experts for their annotation, as mentioned by Dreuw and Ney (2008).

3.1. Convention for Annotation

We used two levels of annotation: one for a Spanish word representing the sign, and another for the sentences in Spanish. Sentence annotation can contribute to future sentence segmentation and translation models. As mentioned in Cormier et al. (2012), the annotation of a sign language corpus should be machine-readable through a systematic annotation. Even when to a large extent, a sign could be easily related to a unique English word, this is not always straightforward. It is the case that sometimes there are several options of glosses for just one sign. This is particularly sensitive when a sign can be interpreted both as a verb or a noun. In that sense, it is usually necessary to rely on grammatical knowledge of the sign language being annotated and also to establish a particular criteria for the annotation based on the needs of the investigation. Due to the few LSP users that know an annotation convention such as the glossing system, we simplify it to a convention that is more suitable to use in a machine learning approach. We expect that machine learning models extract and learn the more specific nuances from the visual information rather than from costly annotation of variants and classifiers in the glosses. Some of the criteria that we simplified are as follows. We relate one sign with only one token or Spanish word in lowercase. We use infinite forms for verbs and singular masculine forms for nouns. We expect to use uppercase only for entities for future identification. For the sentence level annotation, we keep the modifiers of the words (i.e. time, number) and expect the machine learning models to learn from them to match it to a final translation statement. Sign is related to more than one word. In those cases, we assign the closest word related to the sign as if it was seen isolated. For example, this is the case for "helado" ("ice cream") in LSP and "comer helado" ("to eat ice cream"), whose sign are the same. Table 1 shows and explain all the conventions that we considered when instructing the volunteers to annotate the signs and sign sentences.

3.2. Multi-modal Pipeline

In this section we describe the pipeline or information process that we followed to combine the manual annotations and the addition of unaligned audio and keypoint landmarks annotation in an automatic manner. Figure 1 shows more details about the process and structure of the final files of a multi-modal dataset generated by our framework. We provide code of our scripts in our GitHub repository.[1]

3.2.1. Manual Annotation

To obtained the two levels of annotations defined in the annotation convention in two stages. First, we asked a group of volunteers to transcript the video in text files with proper punctuation. Then, we merged them with the YouTube automatically-generated transcripts and time boundaries of subtitles. Given that we are creating an interpretation dataset, the hearing volunteers performed oral-based punctuation by listening to the audio. Note that sign language sentence segmentation might need more understanding such as the one shown in Fenlon et al. (2007) that analyses visual markers as boundaries in intonational phrase of British Sign Language. After the merge of the files, we obtained corrected and punctuated SRT files that can be used as a raw approach to automatic segmentation based on audio that is unaligned to the signing. Second, considering the audio from the video, other group of three volunteers identified repetitive spoken words and aligned them with repetitive signs in the videos. In other words, they used the unaligned audio and punctuated transcript as a reference to identify temporal boundaries of the two levels of annotations, described in 3.1, by signs and by sign sentences. For this second part, these three volunteers used ELAN (Wittenburg et al., 2006), an annotation tool for audio and video, as shown in Figure 2. These volunteers had intermediate knowledge of LSP and rely also in the audio to identify vocabulary in for this task.

3.2.2. Automatic Multi-modal Annotation

After the video segmentation is done manually using ELAN for each tier, we cut each original video using

[1] https://github.com/gissemari/PeruvianSignLanguage

Criteria	Explanation	Example
Lower case	The gloss at the 1st tier, except proper nouns should be written with lower case. In that way, it could be easier to identify them in future works.	1st tier:“Peru”,“yo”,“vivir”(Peru, I, live) 2nd tier: “Yo vivo en Peru” (I live in Peru)
Uppercase	Entities and following convention of writing sentences in Spanish in 2nd tier	
Fingerspelling	They should be annotated separated by a hyphen	1st tier: “yo”, “P-A-T-R-I-C-I-A” (I, Patricia) 2nd tier: “Yo soy Patricia” (I am Patricia)
One sign - several words	Assign the closest single word to the sign in the 1st tier level, as if it was isolated. Use both words in the 2nd tier	1st tier:“helado” (ice cream) 2nd tier: “Comer helado” (to eat ice cream)
Gender & Number	If a sign could have both genres, prefer the male genre in the 1st tier, and the correct reference in the 2nd tier	1st tier: “niño”,“niño” (boy, boy) or “dos”,“niño” (two, boy) 2nd tier: “Dos niños” (Two boys or boys)
Verbs	Annotate the verb in present and the sign of the time	1st tier: “antes”,“yo”,“ir”,“Cusco” (past, I, go, Cusco) 2nd tier: “Yo fui a Cusco” (I went to Cusco)
Unknown sign	When a sign is not identifiable, “NNN” should be used in both tiers	1st tier:“antes”,“comí”,“NNN” (past, eat, NNN) 2nd tier: “Ayer comí NNN” (yesterday I ate NNN)

Table 1: Annotation Convention

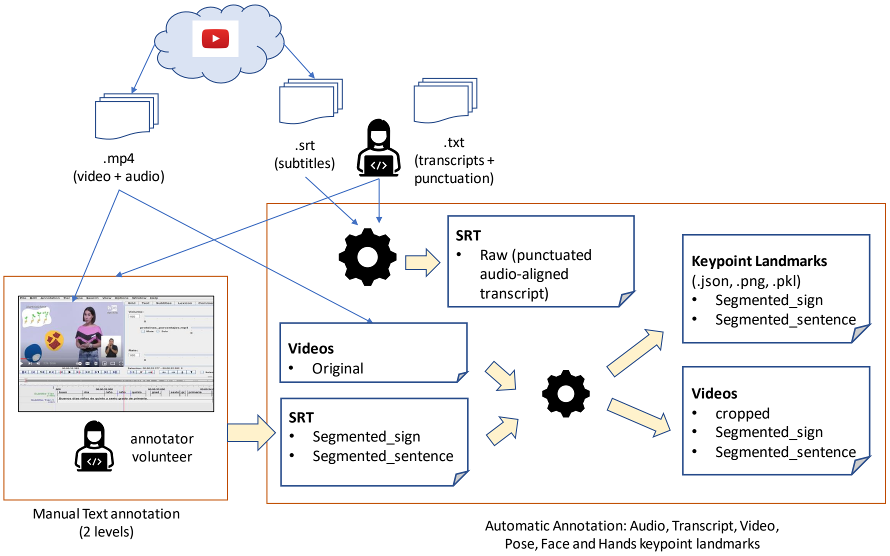
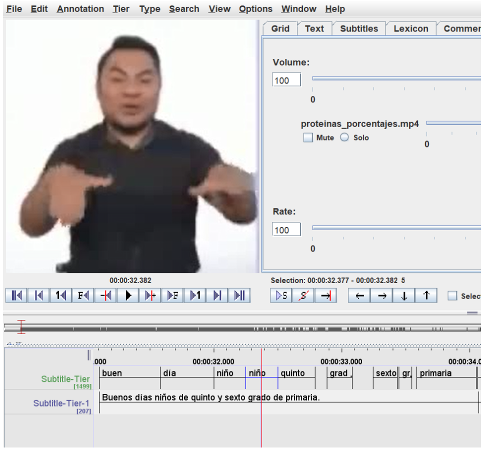


Figure 1: Pipeline for the manual and automatic annotation of PeruSIL


File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Commer
Volume:
100
0
proteinas_porcentajes.mp4
Mute Solo
Rate:
100
00:00:32.382
Selection: 00:00:32.377 - 00:00:32.382 5
Selec
.000
00:00:32.000
00:00:33.000
00:00:34.0
Subtitle-Tier [1499]
Subtitle-Tier-1 [207]

buen | día | niño | niño | quinto | grad | sexto | gr | primaria

Buenos días niños de quinto y sexto grado de primaria.

Figure 2: Tiers for the two-level annotation: individual sign y sign sentence (AEC dataset)

the exported two SRT files and save new videos of isolated signs and of sentences in different folders. We use the opencv and pysrt libraries to work with the videos and the SRTs files. Then, we use the MediaPipe open-source platform from Google to annotate the keynote landmarks for each frame at every new video (Lugaresi et al., 2019). The MediaPipe platform provides different sets of landmarks around a body: face, pose, hands, and a set called holistic to retrieve all the previous sets. Our framework generates this annotation in three different types of files for every frame of the segmented videos: visualization in images (.png), data interchange format (.json), and object structure serialization or intermediate storage (.pkl). These two steps are executed as part of the process implemented in our repository to generate the three types of files. In our repository we also provide the link to our dataset. In Figure 3 we show the keypoint landmarks annotation for the sequence of frames of two video instances for the same sign *IDEA*.

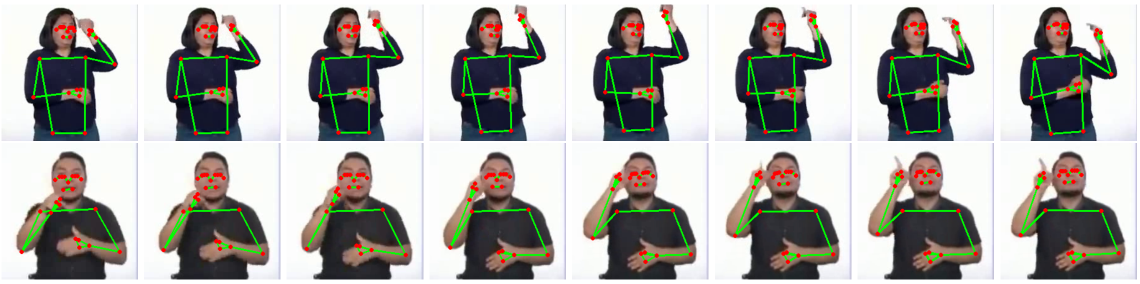


Figure 3: Sequence of frames for the sign IDEA performed by two subjects of AEC

4. Peruvian Sign Language Interpretation: AEC Dataset

During the 2020 and 2021 lockdowns, the Peruvian Government offered remote public school education through a show called “Aprendo en Casa” (AEC, or *I learn at home* in English). Episodes of this show were released on TV channels, social media, and the YouTube channel PeruEduca. In this section, we detail how we leveraged the publicly-available videos from AEC to build a multi-modal dataset for Peruvian SLP using the proposed framework in the previous section.

We selected two subject-videos (of 20 to 30 minutes each). We processed two videos where interpreters translate audio-visual content in the right-most bottom white square, using black clothing and a white background. The rate of a sample of the original downloaded video and the segmented videos is 29.97 fps (frames per second), and the interpretation part had a size of 220 x 220 pixels.

We created a dataset consisting of >500 unique signs, >2000 instances, and >150 sentences. In Figure 4, we show the histogram of instances per unique sign. More than 400 signs have less than 10 instances, and only a few signs, like TO-EAT and PERCENTAGE, have more than 50 instances. This is due to the topics related to the two selected videos, one about knowing our emotions in the subject of socio-cultural development,[2] and the other about healthy food, proportions, and percentages in the Math course.[3] On the other hand, the average number of words in a sentence is 8.80, with a minimum of 1 and a maximum of 34 words. Figure 5 shows the frequency of sentences with a different number of signs/words, and most of them have less than 15 words. We provide direct access to the dataset generated by our pipeline in our github repository.

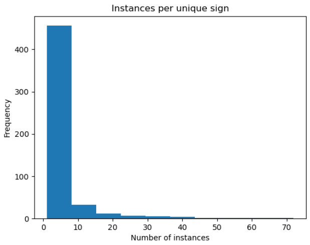


Figure 4: Distribution of instances of signs

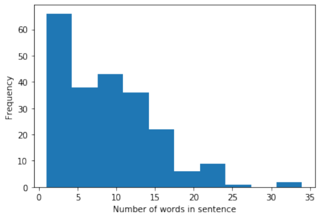


Figure 5: Distribution of sentences lengths

5. Evaluation through a Sign Language Recognition Model

To assess the usefulness of the annotations produced with the PeruSIL framework, we trained and tested a machine learning model in our interpreter-based dataset, AEC, annotated by hearing people for sign language recognition. We also evaluated its performance on common signs found in an additional LSP dataset, PUCP-DGI (Rodriguez Mondoñedo and Arnaiz Fernandez-Concha, 2022), which provides annotations by sign with a gloss convention that lacks standardization for computational processing. In other words, they assign variations of gloss depending on conjugation, number and gender of the translation in context. We identified some of those variations and modified their gloss manually in order to obtain a few more instances.

5.1. ChaLearn Model

The selected machine learning architecture corresponds to one of the 26 participants in 2021 Looking at People Large Scale Signer Independent Isolated SLR CVPR Challenge (De Coster et al., 2021).[4] The challenge consists of performing sign language recognition in a dataset of 226 classes and 36,302 videos.[5] In this subsection we explain the preprocessing needed to perform the feature extraction process to feed the model.

[2] https://youtu.be/7fGAIL2dtk8

[3] https://youtu.be/P4IckOY9P3w

[4] https://github.com/m-decoster/ChaLearn-2021-LAP

[5] https://chalearnlap.cvc.uab.es/dataset/40/description/

5.1.1. Preprocessing

The preprocessing step of this existing model consists of two main phases. The first phase occurs before training, when the keypoints from each video are used to obtain the pose flow data. Then, the original authors estimated keypoints of every frame using the OpenPose library, and substract the keypoints landmarks of the previous frame to calculate the direction of the signer's pose's movement. Due to technical restrictions in our server, we used MediaPipe instead of OpenPose to provide similar landmark estimations. In case the frame count number is lower than the defined sequence length, the missing data is filled by repeating the last video frame, as in a padding form. The second phase occurs during training. The process begins by cropping both hands in a timestep (frame). This cropped area is calculated based on the direction from the elbow to the wrist, where the size of the crop is defined by half the sum of the distance between the shoulders and the distance between the center of the hand and their respective shoulder. The final result is two cropped frames from each signer's hand, which is called RGB data.

5.1.2. Feature Extraction and Training

The model starts by creating batches of preprocessed data: pose flow data and RGB data. The starting RGB data dimension is $B * T * X * C * H * W$ where B is the batch size, T is the number of timesteps, X the number of hands (always 2), C is the channel size, H the height, and W the width of each the frames. This RGB data is modified to $B * (T.X) * C * H * W$ to have both hands sequentially in order from their respective timestep. Then, the dimensions are modified again to $(B.T.X) * C * H * W$ to convert each cropped hand at each time step in an instance to be processed by a pretrained ResNet (He et al., 2016). In that way, the ResNet processes all sequential data in parallel, reducing time for feature extraction. The ResNet output is transformed using a 2D convolutional network to an embedding of certain size (default 512), and then matched to the batch size $(B.T.X) * (FeatureSize)$. Each feature embedding is concatenated with a pose flow data according to timestep-and-batch order, and then normalized. The result of this union is passed to a linear neural network with a Relu function activation, and used as input to a positional encoding that feeds a self-attention model that works similarly to a recurrent neural network whose inner parts consist of a series of a Multi head attention model and position wise Feedforward, using a hidden layer size of 2*embedding size and 2 heads. The model is trained with an Adam optimizer. Lastly, the result is used to learn a cross-entropy layer to do sign recognition of the classes defined. Figure 6 shows the Feature extraction process together with how the model calculates the output through the neural network.

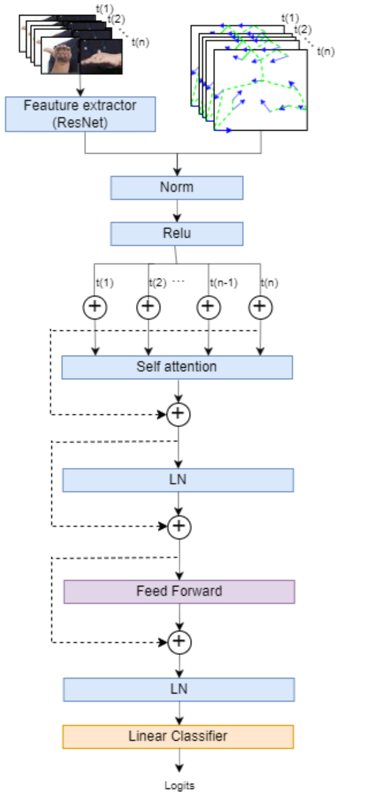


Figure 6: Feature extraction and model

5.2. Experiments

We tested the ChaLearn model in a subset of 5 signs with at least 10 instances per sign: think, see, feel, say, do/make. Figure 7 shows the distribution of the number of frames or length of the video signs for both datasets. The PUCP-DGI dataset shows a broader range of lengths compared to AEC. We hypothesize that this reflects that the pace in native LSP is lower than the interpretation of the LSP. To perform hyperparametrization tuning, we run 3 experiments for each combination in GPUs.

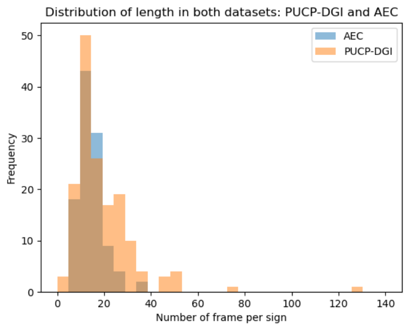


Figure 7: Distribution of length in both datasets: AEC and PUCP-DGI

5.2.1. In-domain Dataset

We did not experiment with hyper-parameter values such as number of *hidden units* or *number of attention heads* due to the good results the ChaLearn model already provided. However, we tested the ResNet architecture used for feature extraction (*rn18* or *rn34*) and *length* of the videos or number of frames as: 10, 15, 20. Also, the *stride* hyperparameter influences how this frames are selected within the set of original frames in a sign video: 1 or 2. Another hyperparameter that we tested was *learning rate* with values 1e-4, 1e-5 and 1e-6.

In Table 2, we show the results of our experiments for the set of 5 classes. In this section we analyze our F1 micro for AEC, which can be interpreted as a direct value to the accuracy through sign classes. In general, the hyperparameter that seems to impact the most in the results is the learning rate. In addition, more consistent results in AEC are reached by 1e-4. For example, for both feature extraction architectures (rn18 and rn34), we found higher F1 results with 1e-4, such as 80.3%. Additionally, higher values remain accross different values of sequence length and stride.

5.2.2. Out-of-Domain Dataset

The PUCP-DGI dataset was created by the Linguistics Department at the Pontificia Universidad Católica del Perú. This dataset includes video recordings of 27 deaf individuals in different classrooms, and each lasts between 1 to 9 minutes approximately. This dataset contains three tiers or levels of annotation: *gloss*, *description* and *classification*, as can be seen in Figure 8. The importance of this dataset is that it was annotated by one deaf person who is an expert in LSP and Spanish. However, we consider this dataset to be partially annotated because the gloss used in the first tier is not completely standardized.

For instance, for this dataset, we identify that variations of a gloss can correspond to the same sign, such as plural and feminine of a certain word. Considering that, we identify 1,382 different tokens, which can include modified tokens of the same sign. However, in order to balance the common classes, we modify these gloss variations and standardize the annotation of some of the instances of this dataset. Additionally, to deal with the problem of some camera movement in videos, frame-by-frame processing has been done to keep the signer in focus.

Feature extractor	Sequence Length	Learning Rate	Stride	F1 micro AEC (%)	F1 micro PUCP-DGI (%)
rn18	10	1.0E-04	1	**80.3 ± 2.6**	20.2 ± 6
rn18	10	1.0E-05	1	59.1 ± 4.5	46.4 ± 4.9
rn18	10	1.0E-06	1	30.3 ± 5.2	49.6 ± 2.7
rn18	10	1.0E-04	2	78.8 ± 2.6	20.8 ± 7.9
rn18	10	1.0E-05	2	53 ± 6.9	48.8 ± 7.3
rn18	10	1.0E-06	2	30.3 ± 6.9	**51.8 ± 3.6**
rn18	15	1.0E-04	1	78.8 ± 2.6	22.2 ± 4.5
rn18	15	1.0E-05	1	57.6 ± 6.9	50.6 ± 9.8
rn18	15	1.0E-06	1	31.8 ± 4.5	49.8 ± 4.1
rn18	15	1.0E-04	2	74.2 ± 5.2	30 ± 11.5
rn18	15	1.0E-05	2	50 ± 4.5	50.4 ± 9.5
rn18	15	1.0E-06	2	31.8 ± 0	51 ± 1.5
rn34	10	1.0E-04	1	80.3 ± 2.6	27.6 ± 10.5
rn34	10	1.0E-05	1	56.1 ± 2.6	39.9 ± 5.2
rn34	10	1.0E-06	1	38.3 ± 11.4	34.7 ± 6.9
rn34	10	1.0E-04	2	74.2 ± 2.6	27.2 ± 7.5
rn34	10	1.0E-05	2	57.6 ± 2.6	40.7 ± 2.3
rn34	10	1.0E-06	2	38.3 ± 11.4	36.5 ± 6.2
rn34	15	1.0E-04	1	78.8 ± 2.6	27.2 ± 3.1
rn34	15	1.0E-05	1	53 ± 9.5	51.4 ± 4.2
rn34	15	1.0E-06	1	30.3 ± 2.6	**52.4 ± 1.2**
rn34	15	1.0E-04	2	75.8 ± 6.9	23.6 ± 4.3
rn34	15	1.0E-05	2	42.4 ± 6.9	50.6 ± 3.1
rn34	15	1.0E-06	2	33.3 ± 2.6	49.6 ± 2.5

Table 2: Comparison of F1 micro and macro for both datasets

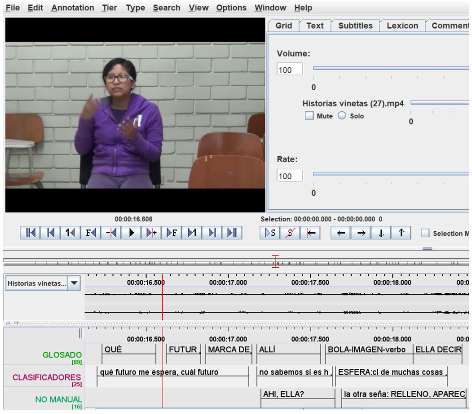

File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Comment
Volume:
100
0
Historias vinetas (27).mp4
Mute Solo
0
Rate:
100
0
00:00:16.606
Selection: 00:00:00.000 - 00:00:00.000 0
Selection M
Historias vinetas...
00:00:16.500
00:00:17.000
00:00:17.500
00:00:18.000
GLOSADO [89]
QUÉ
FUTUR
MARCA DE
ALLÍ
BOLA-IMAGEN-verbo
ELLA DECIR
CLASIFICADORES [25]
qué futuro me espera, cuál futuro
no sabemos si es h
ESFERA:cl de muchas cosas
NO MANUAL [16]
AHI, ELLA?
la otra seña: RELLENO, APAREC

Figure 8: PUCP-DGI tiers for three-level annotation: gloss, description and classification

As shown in Table 2, the best-averaged F1 for PUCP-DGI dataset for rn18 is 51.8%, reached by a sequence length of 10, stride 1 and learning rate of 1e-6 for the group of 5 classes, and best averaged F1 for rn34 is with length 15 and learning rate of 1e-6 as well. We focus on the F1 micro metric due to class imbalance in PUCP-DGI dataset. We also experiment with sets of 10 and 15 classes and they had worse results than random guessing in the test set. It is interesting to notice that while 1e-4 result in higher F1 values for AEC, 1e-6 was a better hyperparameter value for PUCP-DGI results across feature extractor architectures, 51.8% and 52.4% respectively. We hypothesize the reasons can be the class imbalance in this second dataset, the quality of pose estimation, the difference in the pace of recording, and clothing or background of the signers. These results equate to a baseline of 53.2% calculated using the DummyClassifier in scikit-learn with a stratified strategy . We disaggregate PUCP-DGI F1 values by each signer. Table 3 shows the setting with the higher value of accuracy for the PUCP-DGI dataset, and analyzes the accuracy by subject in the two settings where the best results were achieved (rn18 and rn34). Using rn18, we found that subjects that represent 59% of the instances reach individual accuracy of more than 50%.

6. Conclusions

In this paper, we have presented PERUSIL, a framework to annotate sign language. We use this framework to annotate a continuous Peruvian Sign Language inter-

Subject	Number of instances	F1 in rn18	F1 in rn34
Subject15	1	0 %	0 %
Subject4	5	0 %	20 %
Subject8	1	0 %	0 %
Subject9	9	22 %	22 %
Subject13	4	25 %	100 %
Subject14	6	33 %	33 %
Subject2	3	33 %	0 %
Subject17	5	40 %	20 %
Subject19	5	40 %	60 %
Subject23	5	40 %	40 %
Subject27	8	50 %	38 %
Subject30	2	50 %	50 %
Subject6	14	50 %	50 %
Subject3	37	62 %	68 %
Subject12	3	67 %	33 %
Subject24	5	80 %	100 %
Subject18	15	87 %	87 %
Subject22	16	88 %	44 %
Subject20	7	100 %	57 %
Subject21	5	100 %	80 %
Subject25	3	100 %	100 %
Subject29	2	100 %	0 %
Subject37	5	100 %	20 %
Subject42	1	100 %	100 %
Subject5	1	100 %	0 %

Table 3: Frequency of instances within the set of 5 selected classes for training in PUCP-DGI dataset and accuracy by subject

pretation dataset of >500 unique signs and >150 sign sentences. We share publicly a multi-modal Sign Language interpretation resources. For the Peruvian LSP research community, this dataset becomes the first one to provide not only annotated isolated signs but annotation of continuous sign sentences. Our work can trigger the development of other sign language processing stages such as sign segmentation, sign classification (recognition), machine translation, language generation, human computing interaction, etc. Moreover, our proposed framework can help reduce the need of several experts by allowing hearing volunteers to annotate sign language interpretation videos based on audio. Further analysis in the inter-rater reliability of volunteer annotations needs to be tested. Nevertheless, it is highly recommended that the deaf community gets involved in the annotation task. This approach can help automate and scale annotation, as well as to build resources for low-resource sign language.

We demonstrated that a sign language recognition model trained on our dataset achieves moderate results when evaluated in a second dataset of native signers which was partially annotated by one deaf person. We expect that the dataset helps build better sign language processing models that a manage other challenges, such as non-controlled video environments and different rates and settings of recordings. We reached an accuracy of 80.3% in the same dataset, and an accuracy of 52.4% by testing the model in a second dataset. Although these results correspond to a reduced number of classes. Our future work will explore the development of sign language recognition models based on transfer learning, and data augmentation, which can allow working with a higher number of classes.

7. Acknowledgments

We thank all the volunteers who gathered to follow our proposed annotation framework, specially to Jessica Bornas, Edwin Diaz and Sandra Arana; our partners for volunteers training, José Ortiz (Capacitadero) and Alexandra Arnaiz; and Miguel Rodriguez-Mondoñedo for the review. The GPUs we used were provided by Chameleon Resource provided by NSF Keahey et al. (2020) and the Artificial Intelligence Research Group at PUCP (IAPUCP).

8. Bibliographical References

Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N., and Zisserman, A. (2020). Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In Andrea Vedaldi, et al., editors, *Computer Vision – ECCV 2020*, pages 35–53, Cham. Springer International Publishing.

Berrú-Novoa, B., González-Valenzuela, R., and Shiguihara-Juárez, P. (2018). Peruvian sign language recognition using low resolution cameras. In *2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–4.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Ringel Morris, M. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

Bull, H., Gouiffès, M., and Braffort, A. (2020). Automatic segmentation of sign language into subtitle-units. In *European Conference on Computer Vision*, pages 186–198. Springer.

Bull, H., Afouras, T., Varol, G., Albanie, S., Momeni, L., and Zisserman, A. (2021). Aligning subtitles in sign language videos. *CoRR*, abs/2105.02877.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020a). Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319. Springer.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020b). Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Cormier, K., Fenlon, J., Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R., and Woll,

B. (2012). From corpus to lexical database to online dictionary: Issues in annotation of the BSL corpus and the development of BSL SignBank. In Onno Crasborn, et al., editors, *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 7–12, Istanbul, Turkey, May. European Language Resources Association (ELRA).

De Coster, M., Van Herreweghe, M., and Dambre, J. (2021). Isolated sign recognition from rgb video using pose flow and self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3441–3450.

Dreuw, P. and Ney, H. (2008). Towards automatic sign language annotation for the elan tool. In *Workshop Programme*, volume 50.

Elizabeth and Parks, J. (2010). A sociolinguistic profile of the peruvian deaf community. *Sign Language Studies*, 10(4):409–441.

Fenlon, J., Denmark, T., Campbell, R., and Woll, B. (2007). Seeing sentence boundaries. *Sign Language & Linguistics*, 10(2):177–200.

Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., and Ney, H. (2012). RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus. In Nicoletta Calzolari, et al., editors, *8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3785–3789, Istanbul, Turkey, May. European Language Resources Association (ELRA).

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

INEI and CONADIS. (2012). Primera encuesta nacional especializada sobre discapacidad 2012.

INEI. (2018). Resultados definitivos de los censos nacionales 2017.

Keahey, K., Anderson, J., Zhen, Z., Riteau, P., Ruth, P., Stanzione, D., Cevik, M., Colleran, J., Gunawi, H. S., Hammock, C., Mambretti, J., Barnes, A., Halbach, F., Rocha, A., and Stubbs, J. (2020). Lessons learned from the chameleon testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association, July.

Lazo, C., Sanchez, Z., and del Carpio, C. (2019). A static hand gesture recognition for peruvian sign language using digital image processing and deep learning. In Yuzo Iano, et al., editors, *Proceedings of the 4th Brazilian Technology Symposium (BTSym'18)*, pages 281–290, Cham. Springer International Publishing.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., and Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines.

Madrid Vega, R. M. (2018). Clasificadores en la lengua de señas peruana (LSP). BSc dissertation, Pontificia Universidad Católica del Perú, Lima, Perú.

Mejía Gamarra, J. E., Alonso Salazar Cubas, M., Sosa Silupú, J. D., and Enrique Córdova Chirinos, C. (2020). Prototype for peruvian sign language translation based on an artificial neural network approach. In *2020 IEEE XXVII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–4.

MIMP. (2017). Decreto supremo que aprueba el reglamento de la ley n° 29535, ley que otorga reconocimiento oficial a la lengua de señas peruana.

Nureña-Jara, R., Ramos-Carrión, C., and Shiguihara-Juárez, P. (2020). Data collection of 3d spatial features of gestures from static peruvian sign language alphabet for sign language recognition. In *2020 IEEE Engineering International Research Conference (EIRCON)*, pages 1–4.

Renz, K., Stache, N. C., Albanie, S., and Varol, G. (2021a). Sign language segmentation with temporal convolutional networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139. IEEE.

Renz, K., Stache, N. C., Fox, N., Varol, G., and Albanie, S. (2021b). Sign segmentation with changepoint-modulated pseudo-labelling. *CoRR*, abs/2104.13817.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In Nicoletta Calzolari, et al., editors, *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559, Genoa, Italy, May. European Language Resources Association (ELRA).

Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., and Alikhani, M. (2021). Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*.

9. Language Resource References

Rodriguez Mondoñedo, Miguel and Arnaiz Fernandez-Concha, Alexandra. (2022). *Archivos de videos en mp4*. Pontificia Universidad Católica del Perú.

Rodríguez Mondoñedo, Miguel and Arnaiz, Alexandra. (2015). *Repositorio Lengua de Señas Peruana*.