LSU-DS: An Uruguayan Sign Language Public Dataset for Automatic Recognition

Ariel E. Stassi[1], Marcela Tancredi[2], Roberto Aguirre[4], Alvaro Gómez[6], Bruno Carballido[3,4], Andrés Méndez[3], Sergio Beheregaray[5], Alejandro Fojo[2], Víctor Koleszar[3] and Gregory Randall[6]

[1]*Centro Universitario Regional Litoral Norte, Universidad de la República, Paysandú, Uruguay*

[2]*Facultad de Humanidades y Ciencias de la Educación, Universidad de la República, Uruguay*

[3]*Centro Interdisciplinario en Cognición para la Enseñanza y el Aprendizaje, Universidad de la República, Uruguay*

[4]*Centro de Investigación Básica en Psicología, Facultad de Psicología, Universidad de la República, Uruguay*

[5]*Centro de Investigación y Desarrollo para la Persona Sorda, Uruguay*

[6]*Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, Uruguay*

Keywords: Uruguayan Sign Language, Sign Language Recognition, Public Dataset.

Abstract: The first Uruguayan Sign Language public dataset for automatic recognition (LSU-DS) is presented. The dataset can be used both for linguistic studies and for automatic recognition at different levels: alphabet, isolated signs, and sentences. LSU-DS consists of several repetitions of three linguistic tasks by 10 signers. The registers were acquired in an indoor context and with controlled lighting. The signers were freely dressed without gloves or specific markers for recognition. The recordings were acquired by 3 simultaneous cameras calibrated for stereo vision. The dataset is openly available to the community and includes gloss information as well as both the videos and the 3D models generated by OpenPose and MediaPipe for all acquired sequences.

1 INTRODUCTION

Sign languages are natural language systems which employ the space as material substrate and the visual perception of a combination of manual and non-manual parameters to convey meaning. Manual parameters refer to handshape, place of articulation, hand orientation and hand movement. Non manual parameters refer to lip patterns, gaze, facial expressions, and head and body posture of the signer (Von Agris et al., 2008a). The combination of these attributes gives rise to the sign, a basic element of this language. The association between the parameters of a sign and the meaning it carries is not universal, but specific to each language (e.g. American Sign Language, ASL; Argentinan Sign Language, LSA, or German Sign Language, DGS). The Uruguayan Sign Language (LSU) is the one commonly used by the deaf community in Uruguay.

With the advent of artificial intelligence, automatic recognition of manual gestures has gained great importance in the scientific community, with applications in various fields, including Automatic Sign Language Recognition (ASLR). Strictly speaking, ASLR must consider the dynamics of both manual and non-manual activity, i.e., recognition of body and facial activity. In addition to various technological applications, ASLR has allowed advances in the linguistic study of sign languages (Trettenbrein et al., 2021). Historically, ASLR was performed using different input data sources –data gloves, images, depth maps– and several approaches –combination of extracted features and the classification stage– focused on the classification in a subset of a particular sign language at different levels (Von Agris et al., 2008b; Cooper et al., 2011; Cheok et al., 2019). Recently, several efforts have been made to solve ASLR using deep learning approaches, learning spatial and temporal features from data and even employing weakly labeled learning techniques (Koller et al., 2016).

Naturally, ASLR requires a training stage, which is only possible from a properly labeled dataset. There are several datasets in the community available for ASLR research and development. Generally speaking, the available datasets can be classified in static and dynamic. The former are composed of still images or isolated data and are generally used for handshape or fingerspelling recognition. Among the

Stassi, A., Tancredi, M., Aguirre, R., Gómez, A., Carballido, B., Méndez, A., Beheregaray, S., Fojo, A., Koleszar, V. and Randall, G.
LSU-DS: An Uruguayan Sign Language Public Dataset for Automatic Recognition.
DOI: 10.5220/0010894200003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 697-705
ISBN: 978-989-758-549-4; ISSN: 2184-4313

open access datasets, we can mention the ASL Finger Spelling Dataset (Pugeault and Bowden, 2011), the NUS hand posture datasets I (Kumar et al., 2010), and II (Pisharady et al., 2013), LSA16 (Ronchetti et al., 2016a) from Argentina and the RWTH-PHOENIX-Weather MS Handshapes (Koller et al., 2016). On the other hand, dynamic datasets are composed of videos and are used for isolated sign recognition or continuous sign language recognition, depending on the linguistic content involved. Among these, we can mention the RWTH German Fingerspelling Database (Dreuw et al., 2006), SIGNUM (Von Agris and Kraiss, 2007), and LSA64 (Ronchetti et al., 2016b). For a review of sign language datasets please visit http://facundoq.github.io/guides/sign_language_datasets.

As for LSU, to the best of our knowledge, there are two direct antecedents to this work, the TRELSU Lexicon and the TRELSU-HS dataset. The TRELSU Lexicon is the first LSU monolingual dictionary and is publicly available[1]. Composed by 315 signs, this lexicon was built for LSU systematization and as a grammatization tool, and has only one repetition of each sign present in the corpus. Consequently it is not suited to train automatic sign recognition systems. In order to perform automatic recognition of the TRELSU Lexicon, TRELSU-HS[2] is an imbalanced dataset composed of more than 3000 static images for handshape recongnition with 30 classes sampled from 5 native signers (Stassi et al., 2020).

To enable the development of a system for the automatic recognition of LSU, it is necessary to build an appropriate dataset, including 3D information (by active or passive stereo, for example) in order to disambiguate hand occlusions and capture the spatial dynamics of the specific LSU signs. In this paper we introduce LSU-DS, the first dataset for automatic recognition of LSU at different levels: manual alphabet, isolated signs and sentences. LSU-DS is a dynamic public domain dataset which includes triplets of stereo videos and the 3D models for all the registers.

In the following sections we describe how the dataset was constructed, its characteristics, some ethical aspects involved and give elements about the dataset license of use. Finally, we comment on the obtained dataset and propose some future works.

[1]http://tuilsu.edu.uy/TRELSU/

[2]https://github.com/ariel-stassi/TRELSU-HS

2 METHODS

The acquisitions took place in Montevideo, Uruguay, during several sessions in November and December of 2019 in the Early Childhood Learning Laboratory of the Interdisciplinary Center on Cognition for Teaching and Learning (CICEA)[3], a perception laboratory with an instrumented room for the video registrations and a control room separated by a window, where the researchers directed the process.

Figure 1: LSU-DS participants: 10 signers, freely dressed without gloves or other markers. See text for more details.

Participants. The video clips were produced by 10 signers, one frame of the frontal camera video for each one are illustrated in Fig. 1. Five were female and five male (average age = 44.2 years with standard deviation = 13.9 years). Seven of them were born deaf. Nine signers have right dominant hand and 1 does not show preference for one particular hand. From the 10 participants, 3 have at least one other deaf person in the family nucleus. In one case several members of the family are deaf. Concerning educational level, 3 finished university level, 2 have some university studies, 2 finished the secondary studies, 2 are currently doing their secondary studies and 1 finished the basic school. Concerning the LSU language acquisition, 3 acquired the LSU language before they were 5 years old, 2 between 6 and 12 years old, and the rest after 12 years old.

Data Acquisition. Three Flir Backfly S cameras were used, connected in such a way that camera '0', located at the center of the 3-camera array, triggered the other two. Cameras '1' and '2' captured the left and right sides of the signer, respectively. All cameras were arranged at the same height and in such a way that the hands of the signer were always visible. The frame captured the signer, including his hands, at all times. Before the acquisition, the cameras were calibrated for stereo vision. The stereo calibration was estimated between camera '0' and camera '1' and between camera '0' and camera '2'. During the recordings, the signer stood still within a previously marked box on the floor, towards which the cameras were oriented.

[3]Interdisciplinary Space, Universidad de la República

Each subject produced several repetitions of a given letter, sign or sentence. The result was a triplet of raw videos. During the acquisition process timestamps from PsychoPy (Peirce, 2007) were used to mark the beginning and end of each letter, sign or sentence. Individual registers were segmented using timestamps and refined manually by visual inspection of each image sequence. Finally, 3 videos per register were conformed using 'ffmpeg', one per camera. Each video has a 1280×1024 pixels resolution and a 25 fps frame rate. The videos are saved in mp4 format by using codec H.264 (libx264). The pixel format used in the ffmpeg SW was the default one, i.e., *YUV420p*.

Interaction with the Signers During the Video Acquisition. In order to give instructions to the participants, the PsychoPy[4] software was used, which is a free development platform equipped with a wide range of tools for psychology and linguistics experiments. At the beginning of the registration session, a global explanation of the whole procedure was given and each participant filled the informed consent form. Then, each task was explained and it was verified the understanding by the participant. During the experiment, a set of video stimuli was presented through the PsychoPy, from which instructions were given and information on the occurrence of events (master cam timestamps) was collected to facilitate video segmentation. All communication with the signers were in LSU.

An initial and final position for each sign or sentence *resting position* was defined as the signer standing upright, looking straight ahead, with hands in front of the body, one on top of the other on the lower abdomen.

Data Curation. As is well known, data curation is a crucial and time consuming task that must be performed with great care. The acquired videos consisted of a set of frames in raw format. First, an automatic temporal segmentation of each task was conducted based on the timestamps registered with PsychoPy. The segmented frame sequences were then converted to mp4 videos. From that point on, the data curation was an iterative process with the following steps: (1) Refine the temporal segmentation by visual inspection, (2) Verify the linguistic content of each video, (3) Organize data in directories so that they are correctly interpreted by OpenPose and MediaPipe, (4) Process videos with OpenPose and MediaPipe, (5) Verify the OpenPose and MediaPipe detections on each video and (6) Produce the metadata files.

[4]PsychoPy is available at https://www.psychopy.org/

3 ETHICAL ASPECTS

The participant signers were volunteers and it was agreed that they could abandon the experiment at any moment without explanation. The acquisition of this dataset was approved by the Ethics Committee of the School of Psychology of the Universidad de la República. According to their recommendation, all participants signed an informed consent form that clearly explained the use of the registers for research purposes, including their public availability. The LSU-DS web page includes an empty copy of the informed consent document signed by each participant.

4 LINGUISTIC TASKS

LSU-DS includes recordings at three different levels of the LSU language: manual alphabet, isolated signs, and sentences. In order to best ensure independent repetitions of the linguistic tasks (LT), each letter and sign involved was performed by the signer only one time per complete task instance. In the case of sentences this was in general the procedure, with exceptions dully annotated in the metadata.

4.1 Manual Alphabet

Each signer produced two consecutive repetitions of the complete LSU alphabet: *A, B, C, D, E, F, G, H, I, J, K, L, LL, M, N, Ñ, O, P, Q, R, S, T, U, V, W, X, Y, Z.* An exception was signer 5 who made four repetitions. Signers were asked to execute the alphabet at a moderate speed and to return to the resting position after each letter.

4.2 Isolated Signs

The isolated signs are individual lexical pieces registered outside a sentence. Isolated here means that the recording of each sign begins and ends at the previously defined resting position. With some exceptions, three repetitions of each of 23 isolated signs were recorded for each signer, except for signer 10, who register four repetitions.

The 23 signs were selected by LSU specialists. Using a phonological complexity criterion, the signs were selected as examples of the following five categories:

a. Unimanual Signs without Movement The sign is defined by the activity of a single hand with constant parameters over time. The following 8 letters of the alphabet are also examples of this category: *A, B, C, D, E, F, I, K*.

b. Unimanual Signs with a Single Movement The sign is defined by the activity of a single hand, which presents the change of only one of the manual parameters, either the handshape, the place of articulation, or the hand orientation.

c. Unimanual Signs with More Than One Movement The sign is defined by a single hand activity, characterized by combining changes of more than one manual parameter, either handshape, place of articulation, or hand orientation.

d. Symmetrical Bimanual Signs: The sign is defined by both hands activity, which have a "mirror" behavior.

e. Asymmetrical Bimanual Signs: The sign is defined by both hands activity, which have independent or correlated behaviors, in the second case with time delays.

Table 1 includes a sign number, hereafter used to refer to that linguistic content on the dataset, as well as the Spanish translation and English meaning of the isolated sign.

Table 1: Isolated signs used in LT 2. 1st. column: LSU-DS ID of the sign. 2nd. column: Spanish translation. 3rd column: an example of English sentence using the sign (in italic). See text for category explanation.

ID	Spanish	Example in English	Category
sign_01	Gas	Oxygen is a *Gas*	*a*
sign_02	Poco	*Few* things	*a*
sign_03	Permiso	*Excuse me*, I need to pass	*a*
sign_04	Mate	She drinks *Mate* infusion	*b*
sign_05	Nieto	My *grandson* is handsom	*b*
sign_06	Perro	My *dog* is barking	*b*
sign_07	Mujer	Mary is a *woman*	*b*
sign_08	Rio	Nile *river*	*b*
sign_09	Ambulancia	The *ambulance* came fast	*c*
sign_10	Club	I do sports at the *club*	*c*
sign_11	Obsesión	He has an *obsession* with gambling	*c*
sign_12	Sucio	His clothes are *dirty*	*c*
sign_13	Playa	Cancún *beach*	*c*
sign_14	Colores	Rainbow *colors*	*d*
sign_15	Campamento	The *camping* has many tents	*d*
sign_16	Estudiar	I *study* maths	*d*
sign_17	Primo	I play with my *cousin*	*d*
sign_18	Imagen	The photo is an *image*	*d*
sign_19	Enseñar	To *teach* a language	*e*
sign_20	Mecánico	The *mechanic* repairs the car	*e*
sign_21	Responsable	She is a *responsible* person	*e*
sign_22	Teatro	The *theater* is an art	*e*
sign_23	Geometría	Pythagoras tought *geometry*	*e*

4.3 Sentences

In this task the signers were asked to perform seven sentences. Each sentence is composed by a subset of the isolated signs from the LT 2 in combination with some new signs (see table 2). The aim is to generate data in which the isolated signs are part of a continuous sentence and facilitate the elaboration of algorithms for automatic temporal segmentation, for example. This task was conceived as an initial step towards a system of continuous LSU automatic recognition.

The sentences includes different modes: affirmative, negative, interrogative, and exclamative. Table 2 shows the signs sequence of each of the used sentences as well as a possible rough translation[5]. The table includes also a number associated to each sentence, hereafter used to refer to that linguistic content on the dataset. Each signer perform between 2 and 6 repetitions. Subject 6 did not perform this task. More details can be seen in the LSU-DS website.

Table 3 includes, for each repetition and subject, both the cases where the signer performs the signs in same order as the proposal (marked as OK) and the variations introduced in the sentences. The table uses the following codes for variations: **[A]** Addition: the signer adds at least one sign; **[O]** Omission: the signer omits at least one sign; **[P]** Permutation: the signer changes the order of the signs; **[S]** Synonym: the signer uses at least one synonym sign; **[FV]** Phonetic variation: the signer introduces a phonetic variation of the sign (small differences in the sign parameters of the same sign); and **[ST]** Substitution: the signer changes at least one sign by another with different meaning.

The cases of multiple variations are also signaled. LSU-DS includes a metadata file for each register specifying the corresponding linguistic content with more detail. Note that some signers introduce a lot of variations, e.g. subject 10. This LT can be useful for linguistic studies. For ASLR purposes, the OK sentences as well as some variations (O, P, FV) can be used.

5 POSE DETECTION

LSU-DS includes the outputs of two methods for pose detection: OpenPose and MediaPipe.

OpenPose. OpenPose is a Convolutional Neural Network (CNN) based system used to jointly estimate body, hand, facial and foot postures (135 keypoints) of multiple persons on single images (Cao et al., 2019). LSU-DS includes the following 2D and 3D detections: hand, face and body outputs without

[5]The letter @ is used to refer to an undefined gender of a noun or an adjective.

Table 2: Sentences used in LT 3. Left: LSU-DS identifier used. Right: (above) signs sequence asked to perform and (below) a possible rough translation of the sentence.

ID	Signs sequence (above) and rough translation (below)
sent_01	MI PRIM@ RESPONSABLE MUY My cousin is too responsible
sent_02	IMAGEN PERR@ BEBÉ LIND@ There is a very cute puppy dog in the picture
sent_03	MI NIET@ GEOMETRÍA ESTUDIAR My granddaughter/grandson studies geometry
sent_04	MAESTR@ LLEVAR NIÑOS RÍO The teacher takes the children to the river
sent_05	¿CÓMO-ESTÁS? ¡VAMOS ESTUDIAR GEOMETRÍA! How are you? Let's go to study geometry!
sent_06	MI NIET@ MATE GUSTAR PERO IR RÍO NO My granddaughter/grandson likes mate but he/she does not like to go to the river
sent_07	ÉL/ELLA MECÁNICO PERSONA GUSTAR NO He/she does not like his/her mechanic

Table 3: LT 3. Variations introduced in the sentences by repetition and signer. See text for an explanation of the used codes. Empty cells are not performed instances.

Signer	Rep.	sent_01	sent_02	sent_03	sent_04	sent_05	sent_06	sent_07
s_1	r1	OK	A	OK	OK	OK	OK	OK
	r2	OK	OK	OK	OK	OK	OK	OK
s_2	r1	OK	P	OK	OK	P, S	O, ST, P	FV, P
	r2	A	A, P	O	OK	O	A, P	FV, A
	r3	P, S, A	O, S	OK	OK	A, P	O, P	FV, A, P
	r4	S	P, S	OK	P	OK	P	P
	r5	O, S	OK	OK	OK	P	S, P	P
	r6	S	OK	S	OK	P	S, P	ST, P
s_3	r1	OK	P	P	OK	OK	P	P
	r2	OK	P, A		OK	OK	P	P
s_4	r1	OK	OK	ST	OK	FV	ST	OK
	r2	OK	OK	OK	OK	OK		
s_5	r1	O	OK	ST	OK	ST	P	ST, S
	r2	FV	OK	ST	OK	ST	P	S
	r3	OK	OK	ST	OK			
	r4		OK	ST				
s_7	r1	OK	OK	OK	OK	OK	O	OK
	r2	OK	OK	OK	OK	OK	P	OK
	r3	OK	OK		OK		OK	OK
s_8	r1	OK	OK	OK	OK	OK	S, P	ST
	r2	OK	OK	P	OK	OK	S, P	OK
	r3	OK	OK	P	OK	OK	P	P
	r4							A
	r5							A
s_9	r1	S	S	FV	A	OK	S, P	ST
	r2	S	A, S	FV	A	OK	S, P	ST, P
	r3	S	S	FV	A	OK	S, P	ST
	r4				A	OK	S, P	ST
s_{10}	r1	S	A, S	O, A, S	FV, S, ST	S	S, A, O	O, S, P
	r2	S	A, S	O, A, S	FV, S, ST	S	S, A, O	ST, S, P
	r3	S	A, S	O, A, S	FV, S, ST	S	S, A, O	ST, S, P

post-processing.

OpenPose keypoints detection is performed on each of the 2D images acquired by the three cameras. The stereo calibration of the cameras allows to triangulate the keypoints and obtain their 3D position in meters. In this way, an articulated 3D model of the signers is produced, including the positions of hips, trunk, head, face, arms, hands and fingers as illustrated in Fig. 2, upper row. These 3D models can be useful to better learn the signing space and identify the performed signs, specially when a region of interest of the signer is occluded by another.

MediaPipe. MediaPipe is a machine learning based framework for building perception pipelines for multiple platforms (Lugaresi et al., 2019). The system estimates body, hand, facial and foot postures (543 keypoints) of multiple persons on single images. LSU-DS includes the following MediaPipe detections: hand, face and body outputs without any post-processing of detected landmarks. Figure 2 (lower row) illustrates the MediaPipe detections. The keypoints detected for each frame are provided in a text file associated to each video on the LSU-DS.

Comparison. A comparison was done between both systems for each register of LT 1 in order to evaluate their consistency. 2D keypoints were detected for each view. Then the euclidean distance was calculated between the corresponding OpenPose and MediaPipe relevant markers for ASLR (trunk, arms, and hands), whenever this comparison was possible (i.e. both methods detected all the keypoints to compare, which happens in 95% of the frames). For the hands, discrepancy was measured as the euclidean distance between the hand keypoints after subtraction of the centroid of the hand detected by the corresponding model. As a normalization, each difference was divided by the length of the signer's trunk given by OpenPose. The mean error between both methods was lower than 5% for the whole set of keypoints and for the hand the discrepancy was lower than 3%.

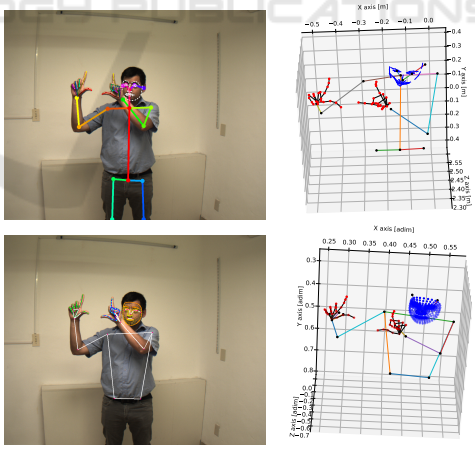

Figure 2: 2D detections and 3D reconstruction on a frame of an LSU-DS register with the frontal camera with pose detections. Upper row, OpenPose. Lower row, MediaPipe.

6 DATASET DESCRIPTION

The LSU-DS dataset is publicly available at https://iie.fing.edu.uy/proyectos/lsu-ds and can be down-

loaded after accepting the terms of use defined in the license (see Sec. 7).

Table 4 gives a general description of downloadable LSU-DS database. There are one json and one txt files per frame. The dataset includes some supplementary files (e.g. metadata files). The downloadable files are compressed by task. The table shows the dataset task sizes once decompressed.

Table 4: General characteristics of LSU-DS.

Name	Linguistic Task	Number of files		Size	
		video	json / txt	video (MB)	json / txt (GB)
LSU-DS-T1	1	1848	164.871	454	1.10 / 7.47
LSU-DS-T2	2	2148	277.158	800	1.83 / 12.56
LSU-DS-T3	3	570	118.185	385	0.79 / 5.36

The dataset is organized in a tree structure as shown in Fig. 3. Each task includes a directory per signer and subtask. For each signer and subtask there is a triplet of videos for a set of repetitions as explained in Section 4. For each task and signer the corresponding projection matrices are included. LT 3 includes a gloss file per video comprising: (a) temporal limits of each sign, (b) temporal limits of the whole sentence and (c) its Spanish translation. Each triplet of videos has the corresponding json and txt files per video frame, with the 2D keypoint detections on each camera view and the 3D keypoint detections, both produced by the OpenPose and MediaPipe SW.

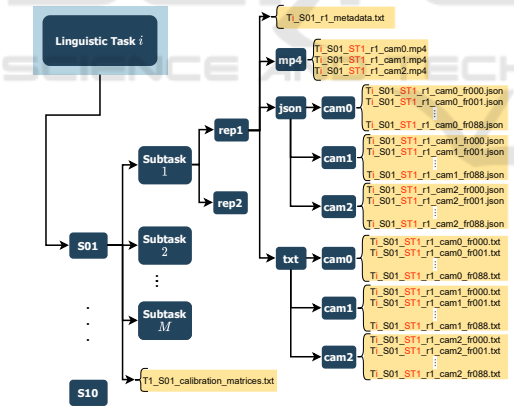

Figure 3: Dataset organization at different levels: linguistic task, subject, subtask, repetition. For each repetition there are three mp4 files corresponding to each camera, a json directory including one directory for each camera with the OpenPose json files corresponding to each frame and a similar structure with the MediaPipe txt files. The calibration matrices file is only available for OpenPose. Directories in blue and files in yellow. The position of task and subtask IDs in the filenames are in red.

As it can be observed in Section 4 the obtained dataset is approximately balanced in terms of samples per class. This aspect is particularly relevant for the training stage of a recognition system. Regarding LT 1, all the signers (with exception of signer 5) performed 2 repetitions. In the LT 2, all the subjects except the signer 1, performed at least 3 times the complete task. Finally, in the LT 3, all the signers (except signers 3 and 4) performed at least 2 repetitions of each sentence, in some cases perform 6 repetitions.

7 LICENSE

LSU-DS was developed in compliance with the Uruguayan personal data protection regulations. Data is available for research and educational purposes to qualified requesters only if the data are used and protected in accordance with the terms and conditions stated in the LSU dataset Restricted Use License, included in the web page.

8 USE RECOMENDATIONS

For a better LSU-DS use, some recommendations of the dataset and particularities encountered during data curation are given.

In the LSU-DS context, the term “margin” refers to the number of initial and final frames of an individual record, in which the signer stays in the resting position. During video segmentation, a margin of 15 to 20 frames was included around the actual sign in LT 1, and of 25 to 30 frames in LT 2 and 3. In LT 1 signers were asked to execute the manual alphabet at a moderate pace. Despite this, margins of less than 15 frames between letters were observed in some cases. In LT 2 and 3, the experimental design included a video stimulus before the execution of each sign. Therefore, it was expected that the execution of signs and sentences would have appropriate margins. However, some signers introduced movements that limit the “quality” of the sought margins.

Due to ethical considerations mentioned in Section 3, prior to the execution of the sign or sentence, signers were asked if they were ready and wished to continue. Some participants nodded with their head or with their hand, generating some artifacts in the desired margins.

During the execution of the letters, isolated signs or sentences themselves and without express request, some signers showed different behaviors compared to the rest. In LT 1, in some cases the signers change the resting position. Signer 5 adopt a different one, putting the hands at the sides and even behind the body. Signer 3 performed LT 1 and 2 changing the dominant hand between repetitions. There were also

slight variations in the phonetic configuration of the signs themselves. A particular case arises with subject 9 of sign *sucio* (*dirty* in English). The subject changed the sign to *olor* (*smell*). This is signaled in the metadata but illustrates the difficulty of this type of dataset, which combines the natural difficulties of any registration of repetitive gestures with the human introduced variability of LT.

Some signers showed clear lip activity during the execution of the signs or sentences. This illustrates the natural variability inherent to LSU, in the responses of the signers to the same stimuli. The metadata files specifies all this variations in detail.

As mentioned in Section 4 and illustrated in table 3, in LT 3 some subjects produced different signs than the claimed one (sign permutations, word association or word elicitation). In some cases the registers maintain their value for the recognition of signs in the context of a sentence. This can be useful for other purposes also, e.g. for psycholinguistic studies on the differences between the instruction and the execution of a sentence. These cases are labeled in the dataset, in order to be considered as special cases during the training of a continuous LSU speech recognition system.

The high degree of hands overlapping in the resting position, in some cases generates noisy OpenPose hand model fitting. It is suggested to take care of this consideration in the construction of detection or segmentation algorithms.

In LT 1, the relevant information is exclusively present in the handshapes, sometimes including hand movement. As the LT becomes more complex (isolated signs and sentences) the information is present not only in the handshapes but also in other manual parameters, the whole body and the relative position and dynamics of its parts. In LT 3 the facial activity is relevant during the execution of the signs, helping the signer to complete the grammar and semantics of the performed sentences. All these aspects of language must be taken into account in order to design suitable systems for processing and automatic recognition.

9 PRELIMINARY RESULTS

In order to explore the potentiality of the LSU-DS, the Video Transformer Network-Pose Flow (VTN-PF) method (De Coster et al., 2021) was used. VTN-PF is based on the VTN architecture which uses a CNN stage for feature extraction from RGB images, and multi-head self-attention and convolutional blocks for modeling the temporal interdependence between frames (Kozlov et al., 2019). VTN-PF also uses the OpenPose output for hand tracking and cropping and for pose flow estimation, improving the results for ASLR. For our experiment the LT 2 front view videos and their OpenPose models were fed to the network. We used the VTN-PF pretrained network[6] with the standard parameters (sequence length = 16 and temporal stride = 2). The linear layer was redefined and fitted (learning rate = 10^{-3}, early stopping with patience of 10 epochs) in order to classify over the 23 signs of LT2. A 10-fold signer independent cross-validation procedure was performed, by splitting the data in 3 subsets per fold: 1 signer for testing, and the other 9 for training and validation (one validation signer was randomly chosen and the rest used for training). The classification accuracy obtained with this approach was of 93.43% for top-1, 98.18% for top-2 and 99.58% for top-5.

10 CONCLUSIONS

This paper presents and describes the first dataset for automatic LSU recognition at different levels: manual alphabet, isolated signs, and sentences. The LSU-DS dataset is openly available to the scientific community and includes gloss information as well as pose models produced by OpenPose and MediaPipe approaches. A first experiment of training a state of the art ASLR network shows the potentiality of the LSU-DS dataset for machine learning applications.

The construction of this dataset was an interdisciplinary effort including linguists, engineers, psychologists and the active participation of the Uruguayan Deaf Community.

This effort is useful not only to improve the research activity both in the engineering and psycholinguistic fields but also to increase the visibility and inclusion of the deaf community. The LSU-DS is a new tool and a new ingredient to the already active field of LSU research that includes sociolinguistic studies, linguistic studies and cultural studies, among others.

In the LT 3 several signers showed some variations in the replication of the proposed sentences. This fact suggests the importance of a good understanding of the LSU complexity, and confirms the necessity of a more active involvement of the deaf community in the design and development of this type of tools.

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

11 FUTURE WORK

After more than one year of intense work in order to have the LSU-DS dataset, we can begin to use it in different fields. The dataset will be useful for psycholinguistic studies and also for the training and testing of automatic segmentation, detection and recognition algorithms.

This first experience was carried in the controlled conditions of the lab as other available datasets reported in the literature. In the future we plan to enrich the dataset with acquisitions outside the lab with varying scene conditions. This is a necessary trend that only recently has begun (a notable example is the AUTSL dataset (Sincan and Keles, 2020) in 2020).

The dataset will be enriched with new metadata and temporal markers to identify different linguistic units. In the future, these new data will include information such as oral language translation in English by sign and sentence, phonological description and grammar type of each sign as well as the syntactic function of the signs into sentences. A main challenge is to include labels referring to non-manual sign parameters such as body tilt, head motion/position, and facial expressions of the subjects done when signing.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contribution of María E. Rodino with the gloss, Federico Lecumberry and Gabriel Gómez for their help in the web site installation and to the reviewers' comments, which improved the article.

REFERENCES

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.

Cooper, H., Holt, B., and Bowden, R. (2011). Sign language recognition. In *Visual analysis of humans*, pages 539–562. Springer.

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 CVPR Workshops*, pages 3441–3450.

Dreuw, P., Deselaers, T., Keysers, D., and Ney, H. (2006). Modeling image variability in appearance-based gesture recognition. In *ECCV Workshop on Statistical Methods in Multi-Image and Video Processing*, pages 7–18.

Koller, O., Ney, H., and Bowden, R. (2016). Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3793–3802.

Kozlov, A., Andronov, V., and Gritsenko, Y. (2019). Lightweight network architecture for real-time action recognition.

Kumar, P. P., Vadakkepat, P., and Loh, A. P. (2010). Hand posture and face recognition using a fuzzy-rough approach. *International Journal of Humanoid Robotics*, 7(03):331–356.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C., Yong, M. G., Lee, J., Chang, W., Hua, W., Georg, M., and Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172.

Peirce, J. W. (2007). Psychopy-psychophysics software in python. *Journal of neuroscience methods*, 162(1-2):8–13.

Pisharady, P. K., Vadakkepat, P., and Loh, A. P. (2013). Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision*, 101(3):403–419.

Pugeault, N. and Bowden, R. (2011). Spelling it out: Real-time asl fingerspelling recognition. In *Computer Vision (ICCV Workshops), IEEE International Conference on*, pages 1114–1119. IEEE.

Ronchetti, F., Quiroga, F., Estrebou, C. A., and Lanzarini, L. C. (2016a). Handshape recognition for argentinian sign language using probsom. *Journal of Computer Science & Technology*, 16.

Ronchetti, F., Quiroga, F., Estrebou, C. A., Lanzarini, L. C., and Rosete, A. (2016b). Lsa64: an argentinian sign language dataset. In *XXII Congreso Argentino de Ciencias de la Computación*.

Sincan, O. M. and Keles, H. Y. (2020). AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. *CoRR*, abs/2008.00932.

Stassi, A. E., Delbracio, M., and Randall, G. (2020). TReLSU-HS: a new handshape dataset for uruguayan sign language recognition. In *1st International Virtual Conference in Sign Language Processing*.

Trettenbrein, P. C., Pendzich, N.-K., Cramer, J.-M., Steinbach, M., and Zaccarella, E. (2021). Psycholinguistic norms for more than 300 lexical signs in german sign language. *Behavior Research Methods*.

Von Agris, U., Knorr, M., and Kraiss, K.-F. (2008a). The significance of facial features for automatic sign language recognition. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE.

Von Agris, U. and Kraiss, K.-F. (2007). Towards a video corpus for signer-independent continuous sign language recognition. *Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May*.

Von Agris, U., Zieren, J., Canzler, U., Bauer, B., and Kraiss, K.-F. (2008b). Recent developments in visual sign language recognition. *Universal Access in the Information Society*, 6(4):323–362.