

LSA64: An Argentinian Sign Language Dataset

Franco Ronchetti^{*1}, Facundo Quiroga^{*1}, César Estrebou¹, Laura Lanzarini¹,
and Alejandro Rosete²

¹ Instituto de Investigación en Informática LIDI, Facultad de informática,
Universidad Nacional de La Plata

{fronchetti,fquiroga,cesarest,laural}@lidi.unlp.edu.ar

² Instituto Superior Politécnico Jose Antonio Echeverría
{rosete}@ceis.cujae.edu.cu

Abstract. Automatic sign language recognition is a research area that encompasses human-computer interaction, computer vision and machine learning. Robust automatic recognition of sign language could assist in the translation process and the integration of hearing-impaired people, as well as the teaching of sign language to the hearing population.

Sign languages differ significantly in different countries and even regions, and their syntax and semantics are different as well from those of written languages. While the techniques for automatic sign language recognition are mostly the same for different languages, training a recognition system for a new language requires having an entire dataset for that language. This paper presents a dataset of 64 signs from the Argentinian Sign Language (LSA). The dataset, called LSA64, contains 3200 videos of 64 different LSA signs recorded by 10 subjects, and is a first step towards building a comprehensive research-level dataset of Argentinian signs, specifically tailored to sign language recognition or other machine learning tasks. The subjects that performed the signs wore colored gloves to ease the hand tracking and segmentation steps, allowing experiments on the dataset to focus specifically on the recognition of signs.

We also present a pre-processed version of the dataset, from which we computed statistics of movement, position and handshape of the signs.

Keywords: sign language recognition, handshape recognition, lexicon, corpus, automatic recognition.

1 Introduction

Sign language (SL) recognition is a complex multidisciplinary problem. It bears many similarities to speech recognition, but presents some additional difficulties [8]:

1. There is little formal standardization in most sign languages, even within a region.

* Contributed equally

2. Sign language specification languages themselves are not well standardized, and there is no consensus on which type of specification is more appropriate.
3. Signs are intrinsically multimodal: they are formed by a combination of hand shapes, positions, movements, body pose, face expression, and lip movements. In contrast, speech recognition usually requires only sound input.
4. Creating datasets for training sign language recognition systems requires being able to capture and model all of these signals.
5. There are relatively few sign language users with respect to the general population, and therefore finding expert subjects to record signs is more difficult.

For these reasons, it is generally considered that we are still far from robust sign language recognition systems.

There are numerous publications dealing with the automatic recognition of sign languages, and [1,8] present reviews of the state of the art in sign language recognition. The full task of recognizing a sign language involves a multi-step process. In the context of video-based recognition, and considering only manual information, this process can be simplified as:

1. Tracking and segmenting the hands of the interpreter in every frame of the video.
2. Recognizing the shapes of the hands, the movements they made and their positions.
3. Recognizing the sign as a syntactic entity (a visual word).
4. Assigning semantics to a sequence of signs (a visual sentence).
5. Translating the semantics of the signs to the written language.

These tasks are mostly independent from each other, and involve different techniques. Tracking, segmentation and modeling of the hand are mostly signal processing and 3D modeling tasks, while assigning semantics to a sequence of signs and translating from sign to written languages are more related to natural language processing.

1.1 Datasets

There are many datasets for sign language recognition. We distinguish three types, depending on the problem they target: handshape recognition, sign recognition or sentence recognition. Each type of dataset presents a greater challenge than the previous one, and allows experiments with more steps of the recognition pipeline.

Table 1 presents the most prominent video-based, research-level datasets for recognition. Since the dataset described in this paper focuses on sign-level recognition, we only list sign and sentence-level datasets ³.

³ A more detailed reference about sign language datasets can be found at http://facundoq.github.io/unlp/sign_language_datasets/index.html

Table 1. Recognition-oriented, video based, sign language datasets used in recent papers.

Name	Classes	Subjects	Samples	Language level	Availability
DGS Kinect 40 [1]	40	15	3000	Word	Contact Author
DGS RWTH-Weather [8]	1200	9	45760	Sentence	Public Website
DGS SIGNUM [8]	450	25	33210	Sentence	Contact Author
GSL 20 [1]	20	6	840	Word	Contact Author
Boston ASL LVD [3]	3300+	6	9800	Word	Public Website
PSL Kinect 30 [2]	30	1	300	Word	Public Website
PSL ToF 84 [2]	84	1	1680	Word	Public Website

In general, the datasets that are video-based must rely on skin color tracking and segmentation, and are therefore not robust for background variations or interpreter clothes, as well as hand-hand or hand-face occlusions [6]. Usually, to perform features extraction these datasets require the addition of morphological information as a subsequent step to the color filtering to identify the position and shape of the hand, which can be extracted using depth cameras or other sensors, but these limit the application of the methods with respect to using normal video cameras, given the availability of each type of devices.

The largest sign language dataset available (in terms of number of classes), the *American Sign Language Lexicon Video Dataset* (ASLLVD) [3], contains more than 3300 signs from the American Sign Language, but near-perfect hand tracking and segmentation on this dataset is difficult [3], making it hard to use it to evaluate a sign recognizer that focuses on the syntactic and semantic recognition. The situation is similar for the SIGNUM and RWTH-Weather datasets. Moreover, the dataset ASLLVD has only 6 subjects, and an average of 3 samples per class.

1.2 Argentinian Sign Language Dataset (LSA64)

Sign languages are different in each region of the world, and each has their own lexicon and group of signs. Thus, sign language recognition is a problem that needs to be tackled differently in each region, since new movements or handshapes or combinations thereof require new training data, and possibly involve new challenges that were not considered before [8,1].

To the best of our knowledge, there are no available datasets for the Argentinian Sign Language (LSA). There are only a few dictionaries that focus on teaching the language. Since they were recorded with this aim in mind, they have only one sample for each sign, low image quality, and poor annotations, thus making them unsuitable for training automatic recognition systems. There

is need for a research-level dataset that represents the group of signs used in LSA.

This paper presents a sign dataset called LSA64. The dataset consists of 64 signs from the LSA, and was recorded with normal RGB cameras. It is publicly available⁴, and we also provide a preprocessed version of the dataset to facilitate experiments and reproducibility.

The subjects wore colored gloves for the recording (single colored gloves, with different colors for each hand). This methodology allows researchers to easily bypass the tracking and segmentation steps, and focus on the subsequent steps of the recognition [9].

While the dataset has less classes than ASLLVD, RWTH-PHOENIX-Weather or SIGNUM, it has more samples and subjects than many other datasets (Table 1), and it is publicly available with a preprocessed version.

The document is organized as follows: Section 2 describes the LSA64 dataset and the recording conditions. Section 3 presents statistics and information of the signs recorded, to aid in the understanding of the dataset. Section 4 details an experiment carried out to establish a baseline on this dataset, and Section 5 presents the general conclusions.

2 Dataset

The sign database for the Argentinian Sign Language, created with the goal of producing a dictionary for LSA and training an automatic sign recognizer, includes 3200 videos where 10 non-expert subjects executed 5 repetitions of 64 different types of signs. Signs were selected among the most common in the LSA lexicon, and include both verbs and nouns. Some examples can be seen in Figure 1.

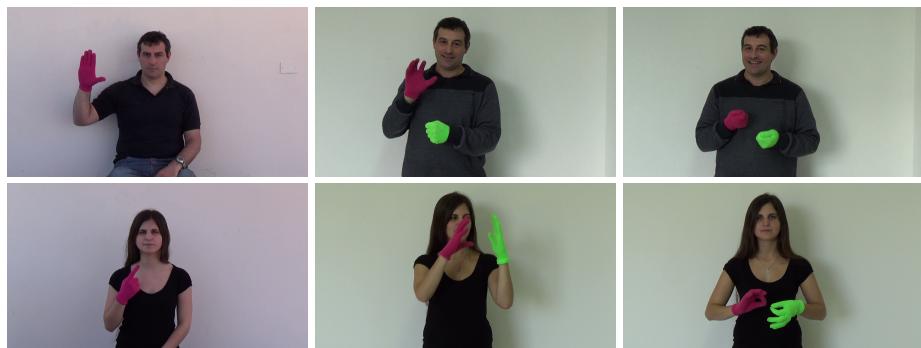


Fig. 1. Snapshots of six different signs of the LSA64 database. There are overlaps in positions and handshapes. The images on the left are from the first set of recordings.

⁴ The dataset and relevant information can be found at <http://facundoq.github.io/unlp/lsa64/>.

2.1 Recording

The database was recorded in two sets. In the first one, 23 one-handed signs were recorded. The second added 41 signs, 22 two-handed and 19 one-handed. The final dataset then contains 42 one-handed signs and 22 two-handed ones.

The first recording was done in an outdoors environment, with natural lightning, while the second took place indoors, with artificial lightning (Figure 1). Subject 10 from the first recordings was unavailable for the second set of recordings, and was replaced by another subject. This change in no way diminishes the utility of the dataset, since the set of classes recorded in the first session is disjoint from the ones recorded in the second session.

In both sets of recordings, subjects wore black clothes and executed the signs standing or sitting, with a white wall as a background. To simplify the problem of hand segmentation within an image, subjects wore fluorescent-colored gloves. These substantially simplify the problem of recognizing the position of the hand and performing its segmentation, and remove all issues associated to skin color variations, while fully retaining the difficulty of recognizing the handshape. Additionally, each sign was executed imposing few constraints on the subjects to increase diversity and realism in the database. The camera employed was a Sony HDR-CX240. The tripod was placed 2m away from the wall at a height of 1.5m.

In the following subsections we show statistics and information of the signs to better understand the nature and challenges of the dataset. These statistics show that signs in this dataset possess significant overlap in terms of types of movements, initial and final positions and handshapes, producing non-trivial experiment settings to test new sign language recognition models. All the information has been computed from the pre-processed version of the dataset described in Section 3.

2.2 Handshapes

In Figures 2 and 3 we can observe the different handshapes of the right and left hand respectively for each class of sign. There is plenty of repetition between handshapes, although their 2D projection may be different depending on the rotation of the hand.

2.3 Positions

Figure 4 presents the mean initial and final positions for each hand, along with the covariance. While a few signs can be identified by their positions, they overlap significantly in most cases.

2.4 Trajectories

Figure 5 shows sample trajectories of each sign, as performed by subject 2. There is much overlap in movements for both one-handed (for example, signs 1, 5, 7, 13 and 19) and two-handed signs (for example, signs 31, 32 and 61).

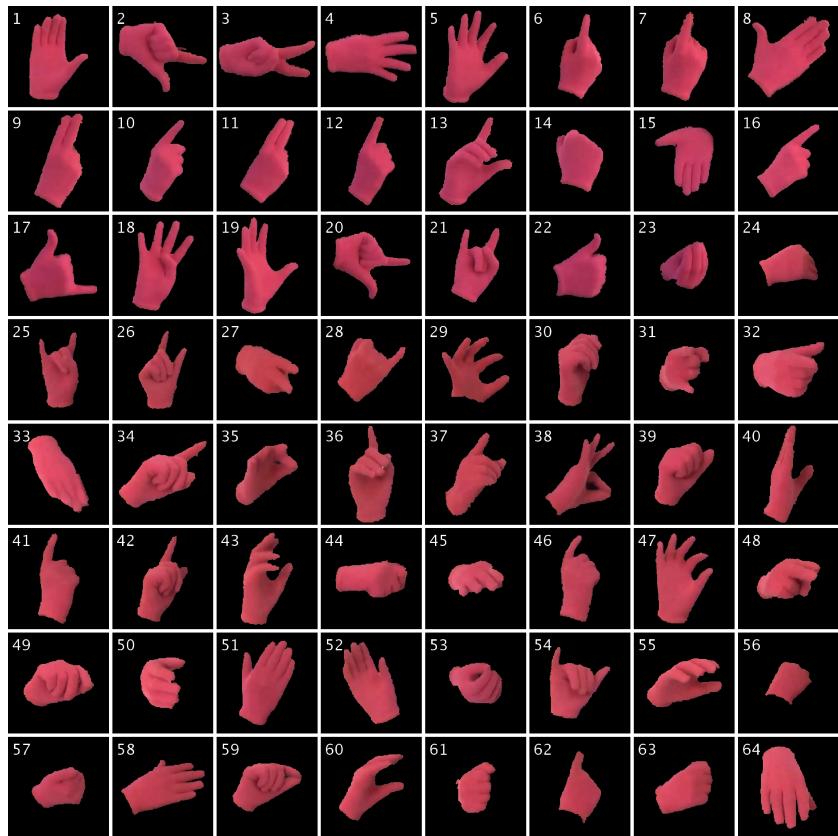


Fig. 2. Images of segmented hands as captured in the LSA64 database. Each image shows the initial handshape of the right hand for each sign in the dataset.

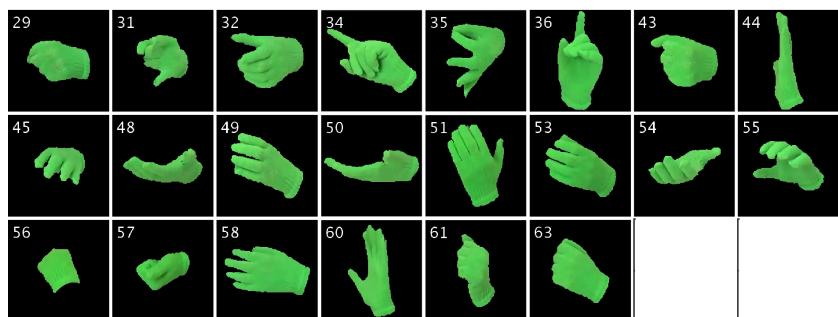


Fig. 3. Images of segmented hands as captured in the LSA64 database. Each image shows the initial handshape of the left hand for the two-handed signs of the dataset.

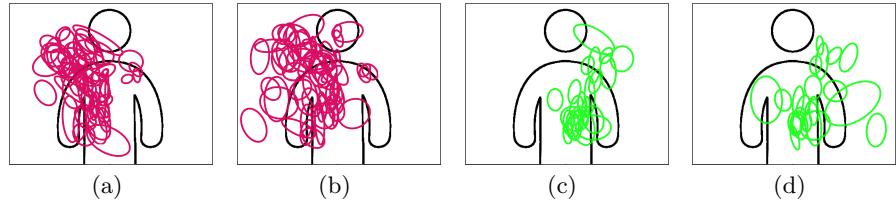


Fig. 4. Means for the initial and final positions of the right hand for each sign (a and b), and also for the left hand (c and d). Circles around means represent the covariance of the samples.

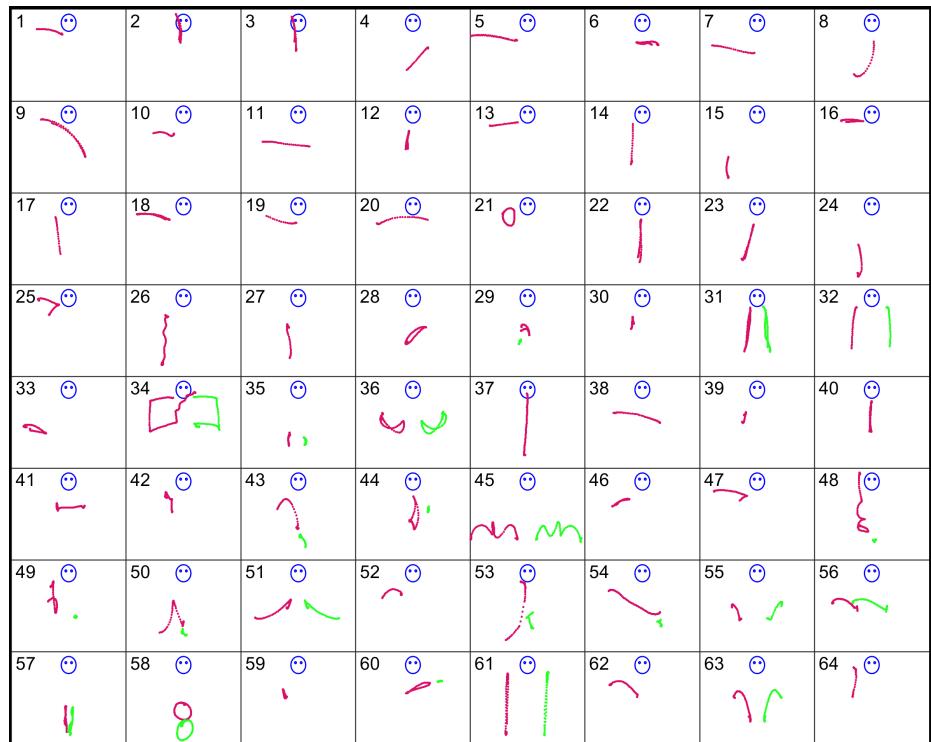


Fig. 5. Sample trajectories for each sign in LSA64. The left-hand trajectory is shown in light green, the right-hand one in red, and the head position as a blue circle.

2.5 Amount of movement

Figure 6 shows the amount of movement for each hand, measured as the maximum distance between two points in the trajectory of the hand. The movement in the left hand is significantly smaller than that of the right hand in many signs, consistent with the fact that the right hand is the dominant one for all the signers.

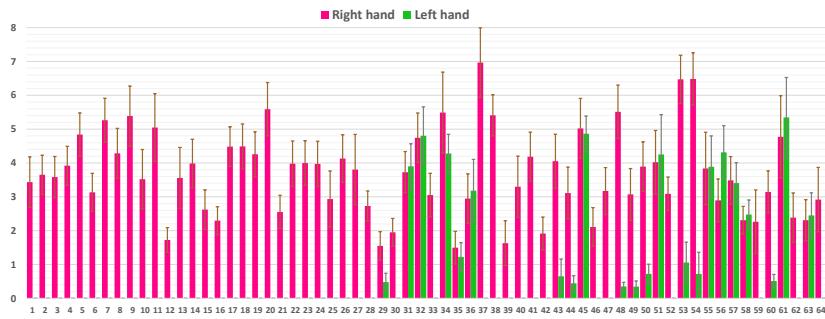


Fig. 6. Amount of movement for each class of sign. Red bars show the amount of movement of the right hand, and green bars the movement of the left hand (in two-handed signs).

3 Preprocessed version

We provide a pre-processed version of the dataset to alleviate the overhead of performing experiments with the data.

From the dataset we extracted the hand and head positions for each frame, along with images of each hand, segmented and with a black background, as shown in Figure 7.

The tracking and segmentation of the hands uses the techniques described in [5]. Additionally, the head of the subject is tracked via the Viola-Jones's face detector [7]. The 2D position of each hand is translated so that the head is at the origin. The positions are then normalized by dividing by the arm's length of the subject, measured in centimeters/pixels. In this way, the transformed positions represent distances from the head, in units of centimeters.

The result of this process is a sequence of frame information, where for each frame we calculate the position of both hands, and we extract an image of each hand with the background segmented.

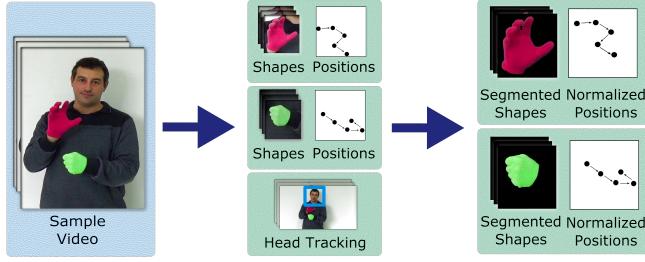


Fig. 7. Feature extraction steps for the preprocessed version of LSA64. From the sample video, we track the position of both hands and head in all frames. The shapes of each hand are segmented and the positions of them are re-centered with respect to those of the head.

4 Baseline Experiments

In this section, we briefly describe the model and results obtained in signer-dependent and independent experiments with the dataset, to establish a baseline performance. The model we used is described fully in [4].

The model we used to get a baseline performance on the dataset classifies the information for each hand separately and then multiplies the probabilities output by the classifiers, per class. The model for each hand includes three classifiers, each processing position, movement or handshape information.

The movement classifier contains one left-to-right Hidden Markov Model (HMM) per class, with skip transitions. All the models have 4 states. The output probabilities are modeled with a Gaussian Mixture Model (GMM) in each state. The models are trained with EM with the trajectories computed in the pre-processed version of the dataset.

The handshape classifier also employs HMM-GMMs, but uses as input the output of the static handshape classifier described in [5]. The static classifier is run on the segmented hand for each frame of the video, and the sequence of probabilities is fed into the handshape HMM-GMM to obtain the probability of each class for the whole sequence of frames.

The position classifier models the initial and final positions of the signs of each class with a set of gaussian distributions. There are two gaussian per class, one for the initial and another one for the final position.

For each class, the model outputs the product of the probabilities given by the position, movement and handshape information of both hands. For one-handed gestures, the information of the left hand is ignored, and so the probabilities output by the left hand model are not multiplied.

We performed a subject-dependent classification experiment with the model, using stratified repeated random sub-sampling validation as the cross-validation scheme, with 30 runs and an 80-20 training-test split. For each run, we measured the classification accuracy of the model on the test dataset. The mean accuracy obtained was 95.95% (standard deviation $\sigma = 0.954$).

5 Conclusion

We have presented a dataset of signs from the Argentinian Sign Language. To the best of our knowledge, there are currently no research-oriented datasets of this language created or available.

The subjects used colored gloves in the recording of the signs to significantly ease the tracking and segmentation steps. Nonetheless, we also provide a pre-processed version of the dataset to facilitate experimentation and reproducibility.

We have also presented a set of statistics and extra information to characterize the dataset and allow researchers to easily understand its nature. The signs in this dataset possess significant overlap in terms of types of movements, initial and final positions and handshapes, producing non-trivial experiment settings to test new sign language recognition models.

We intend to expand the dataset with both new signs and a set of annotated LSA sentences to provide a complete basic working vocabulary for Argentinian Sign Language.

References

1. Cooper, H., Holt, B., Bowden, R.: Sign language recognition. In: Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (eds.) *Visual Analysis of Humans: Looking at People*, chap. 27, pp. 539 – 562. Springer (Oct 2011)
2. Kapuscinski, T., Oszust, M., Wysocki, M., Warchol, D.: Recognition of hand gestures observed by depth cameras. *International Journal of Advanced Robotic Systems* 12 (2015)
3. Neidle, C., Thangali, A., Solaroff, S.: Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In: Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Citeseer (2012)
4. Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., Rosete-Suárez, A.: Sign language recognition without frame-sequencing constraints: A proof of concept on the argentinian sign language. IBERAMIA: Iberoamerican Society of Artificial Intelligence (2016)
5. Ronchetti, F., Quiroga, F., Lanzarini, L., Estrebou, C.: Handshape recognition for argentinian sign language using probsom. *Journal of Computer Science and Technology* 16(1), 1–5 (2016)
6. Roussos, A., Theodorakis, S., Pitsikalis, V., Maragos, P.: Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. In: Trends and Topics in Computer Vision - ECCV 2010 Workshops, Heraklion, Crete, Greece, Revised Selected Papers, PartI. pp. 258–272 (2010)
7. Viola, P., Jones, M.J.: Robust real-time face detection. *International journal of computer vision* 57(2), 137–154 (2004)
8. Von Agris, U., Zieren, J., Canzler, U., Bauer, B., Kraiss, K.F.: Recent developments in visual sign language recognition. *Universal Access in the Information Society* 6(4), 323–362 (2008)
9. Wang, R.Y., Popović, J.: Real-time hand-tracking with a color glove. *ACM transactions on graphics (TOG)* 28(3), 63 (2009)