Experimental Framework Design for Sign Language Automatic Recognition

Darío Tilves Santiago,[1] *Ian Benderitter,*[2] *Carmen García Mateo*[1]

[1]AtlanTTic Research Center – University of Vigo
[2]Polytech Nantes

dtilves@gts.uvigo.es, ian.benderitter@etu.univ-nantes.fr, carmen.garcia@uvigo.es

Abstract

Automatic sign language recognition (ASLR) is quite a complex task, not only for the intrinsic difficulty of automatic video information retrieval, but also because almost every sign language (SL) can be considered as an under-resourced language when it comes to language technology. Spanish sign language (SSL) is one of those under-resourced languages. Developing technology for SSL implies a number of technical challenges that must be tackled down in a structured and sequential manner.

In this paper, the problem of how to design an experimental framework for machine-learning-based ASLR is addressed. In our review of existing datasets, our main conclusion is that there is a need for high-quality data. We therefore propose some guidelines on how to conduct the acquisition and annotation of an SSL dataset. These guidelines were developed after conducting some preliminary ASLR experiments with small and limited subsets of existing datasets.

Index Terms: hearing impaired, dataset recommendations, automatic recognition, convolutional neural networks, ASL

1. Introduction

Sign language (SL) is one of the preferred ways of communication for hearing-impaired people and their families; indeed, often it is the only communication mechanism.

Automatic language recognition is highly developed for written and oral languages, but not for SLs [1]. In recent years, with the emerging interest in both neural networks (NNs) and graphical processing units (GPUs) and ongoing improvements in their efficiency, there has been significant advances in oral and written speech recognition. However, this is not the case for automatic SL recognition (ASLR). This is due to differences in recognition problems, and also to the lack of available images and videos tailored to the development of ASLR systems using machine-learning techniques.

Therefore, an important drawback to be solved is the lack, both in quantity and quality terms, of SL data for processing by deep learning algorithms. It is possible to find SL datasets, but their purpose is to be used as learning material for SL courses, rather than for automatic image recognition. The main drawback is that there are not enough repetitions of each gesture performed by one signer, and more signers are needed in order to ensure variability.

Many researchers have to find creative solutions to this problem. Some apply data augmentation to increase the volume of their data. Data augmentation [2] transforms the data in case of images by enlarging them, trimming them, changing the angle, altering the background and changing the illumination. Other possibilities are to artificially create hand gestures [3] or to record hand gestures with a device like Leap Motion which instead of acquiring images, captures the 3 dimensional movement of the hands and each finger [4], [5].

The study of SL is also constrained by its complexity. While there has been some improvement with isolated gesture recognition despite the dataset situation, there has been no advance whatsoever in continuous gesture recognition. The reason is twofold: datasets are available for isolated but not for continuous gestures, and the technology is not advanced enough to identify the latter. Nevertheless, there have been some discussions about how to proceed in continuous gesture recognition as described in [6], [7], [8] and [9].

In this paper, several datasets are described in terms of their characteristics in Section 2. Section 3 reports details of an initial study of the American SL (ASL) alphabet using transfer learning and provides a set of recommendations on how to create a robust dataset. Finally, some conclusions are drawn in Section 4.

2. Overview of existing datasets

Table 1 lists hand sign datasets (#1-#19) found in the literature, along with relevant information for its use in ASLR. The purposes of these datasets are different: some were designed for automatic recognition problems and others were created for pedagogical reasons. Shown are the dataset name and lexicon, along with the SL if any. Also provided is information on size and on variety in terms of number of signers, of repetitions per signer and type of data, with indications as to whether the dataset is labelled or not. The following will explain how Table 1 is organized.

Items #1-4 are datasets which were used in our research as training and testing material for preliminary experiments. The second column indicates the language of each dataset: items #1-11 are ASL, #12 and #13 are German SL (GSL), #14 and #15 are Spanish SL (SSL), #16 is Argentinian SL (ArSL), and #17-19 are hand gestures which do not belong to any SL. The datasets are sorted by their lexicon: from alphabets and numbers to words and sentences. All the datasets are publicly available, except for Grades Online (#14) which requires contact with the creators [10]. All the datasets are monolingual, except for SpreadTheSign (#15), which collects suggestions for signs from different SLs around the world. Some of the datasets, due to their size and/or insufficient labelling, are not completely characterized in Table 1 (indicated NA for 'not available').

10.21437/IberSPEECH.2018-16

Table 1: *Summary of datasets found in the literature.*

#	Name	Language and Lexicon	Size	Data Type	# Repetitions	# Signers	Labelled
1	Superpixel [11]	ASL: 24 alphabet signs	131688 images	RGB	500	5	Yes
2	Fingerspelling [12]	ASL: 24 alphabet signs and digits 1-9	31000 images	Depth	200	5	Yes
3	Massey University [13]	ASL: 24 alphabet signs and numbers 0-9	2524 images	RGB	5	5	Yes
4	Padova Senz3D [3]	ASL: letters B, D, I, S, and digits 2, 3, 4, 5, 9, 10	2640 images	RGB + Depth	30 + 30	4	Yes
5	Padova Kinect [14]	ASL: letters A, D, I, L, W, Y, and digits 2, 5, 7	2800 images	RGB + Depth	10 + 10	14	Yes
6	HKU Kinect Gesture [15]	ASL: letters A, L, Y and digits 1-5	3000 images	RGB	60	5	Yes
7	NTU Microsoft Kinect [16]	ASL: letters A, L, Y and digits 1-5	2000 images	RGB + Depth	10 + 10	10	Yes
8	Cvpr15 [17]	ASL: 7 words and digits 1-9	68000 images	Depth	500	8	Yes
9	RWTH-50 [18]	ASL: 83 words	8844 videos	Grey scale	2	3	Yes
10	ASLLVD [19]	ASL: words	992 videos	RGB	2	5	Yes
11	RWTH-104 [20]	ASL: 201 sentences	201 videos	Grey scale	1	3	Yes
12	German Spelling [21]	GSL:30 alphabet signs and digits 1-5	3080 videos	Grey scale	2	20	Yes
13	RWTH PHOENIX [22]	GSL: sentences	592383 images	RGB	1	NA	No
14	Grades Online [10]	SSL: 750 words	750 videos	RGB	1	2	Yes
15	SpreadTheSign [23]	SSL: words and sentences	+2000 videos	RGB	1	NA	Yes
16	LSA64 [24]	ArSL: 64 words	3200 videos	RGB	5	10	Yes
17	SKIG [25]	Hand gestures: 10	2160 videos	RGB + Depth	18 + 18	6	Yes
18	Microsoft Gesture-RC [26]	Hand gestures: 12	594 videos	RGB + Depth	NA	NA	No
19	ChaLearn 2016 [27]	Hand gestures	140945 videos	RGB + Depth	1 + 1	NA	No

3. ASLR: preliminary experiments

Learning a language is a step by step process: we first learn the alphabet and numbers, then basic words (objects, adjectives, etc.), then how to relate words (subjects, verbs, adverbs, etc.) and how to contextualize them, and finally correct grammar and syntax, eventually managing to make coherent conversation.

That same learning process applies to automatic speech recognition systems, in stepping up the interaction versatility with machines. And that same process ought to be applied to SL recognition systems.

We trained different NNs using some of the listed datasets in order to compare their performance. The lexicon used to train the NNs was composed of 33 ASL isolated signs: the alphabet 24 gestures and digits from 1-9. Isolated gestures were selected because they were easier to identify, as commented above. We used only one hand image as input feature to identify the isolated sign.

This experiment helped devise a proposed set of guidelines on creating a robust dataset. Also demonstrated was the usefulness of combining both depth and RGB images in datasets in our training with both types of pictures.

3.1. NN architecture

An NN was first selected for the experiment. We used the Convolutional Neural Network (CNN), a deep learning approach to object identification. A CNN is based on relating a set of features to an object definition. It is designed as a set of layers containing neurons that extract features, from specific to general as the data goes through the layers. In training the CNN, this process is repeated, with the importance given to each feature changing until the CNN is capable of recognizing all the different objects provided as inputs.

We used transfer learning to demonstrate the efficiency of using trained CNNs to recognize hand signs. Transfer learning takes advantage of NNs trained for specific applications, retraining them for another application, meaning that the features extracted for the first problem are used for the new one. This is very useful to reduce the time consumed and the images needed in training because there is no need to learn all the features from scratch.

It was selected the VGG16 CNN pre-trained on the ImageNet Database [28]. Its last layer was modified in order to serve as a classifier of 33-sign ASLR experiment. The input features consist of 224x224-pixel image of one hand. So depending on the dataset, hand segmentation and resizing have to be performed.

3.2. Data selection

In order to experiment with recording and segmentation techniques, an in-house dataset was acquired. This dataset is denoted UVIGO in further explanations and results. It consists of recordings of the 33 ASL signs of the lexicon described in Section 3. Kinect2 sensor was used acquiring both RGB and depth streams of the upper body. It was recorded by two signers over four days to ensure variable in recording environments (clothing, lighting conditions, etc.). 100 repetitions of each sign were used for training and 10 repetitions for testing.

Of the available datasets described in Section 2, two were selected for NN training and testing: Superpixel (#1), composed of alphabetical RGB images of one hand, and Fingerspelling (#2), containing both alphabetical and numerical depth pictures

of one hand. 100 randomly-selected repetitions of each sign were used for training and 10 for testing.

Two further datasets were selected only as test datasets: the RGB images of Massey University dataset (#3), and the depth images of Padova Senz3D dataset (#4).

3.3. Data preprocessing

Image preprocessing in machine learning typically consists of cropping, resizing and feature extraction operations. Next we describe the preprocessing tasks needed for each dataset.

As said before, the used CNN required 224x224-pixel images as input features, so resizing had to be performed for Superpixel, Fingerspelling and Massey University datasets. Cropping (hand segmentation) is not needed for these datasets.

Regarding UVIGO dataset, hand segmentation is made using the information provided by the Kinect2 sensors. Kinect2 gives access to RGB and depth image streams, along with 25 body joint coordinates. In the RGB stream our segmentation algorithm uses 4 body joints to easily locate a hand. Then it crops a square around the detected hand. For the depth stream we use again 4 joints to locate the hand. This hand is segmented in the xy plane as the RGB images. The algorithm applies another cropping in the z axis using the depth values of the hand to set a threshold (Figure 1). This eliminates the background from the depth images.

Finally, we also performed a segmentation in Padova dataset. It was applied the depth stream cropping described before, but in this case instead of using the joints provided by Kinect2, we directly used the depth values to distinguish the hand from its surroundings.

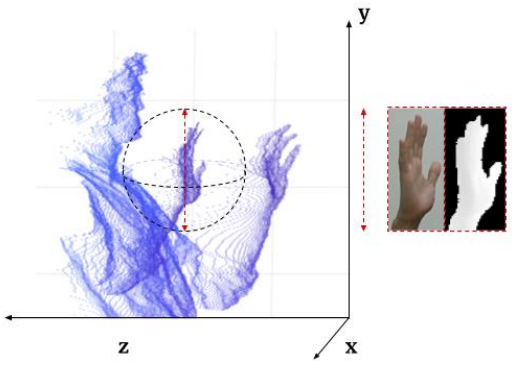

Figure 1: *Segmentation representation.*

3.4. Experiments and results

The results presented in this section are accuracy scores calculated from CNN predictions. All the CNNs were trained in the same way, fixing the number of epochs to five, and with a low initial learning rate, which is divided by four every epoch. With such a reduced number of epochs, and taking advantage of the higher learning slope that characterize transfer learning, the goal was to reduce overfitting.

In a first experiment, each training dataset was separately used to train a CNN. Table 2 shows the results in terms of accuracy over the testing data.

The highest scores in Table 2 are from testing each CNN with its testing material (main diagonal). However, despite UVIGO CNN having the lower score of these three, it performed better when testing it with the #1 and #2 datatests (second row in Table 2). Counterwise, Fingerspelling CNN had the highest accuracy (99.61%) but the lowest accuracy values over the other datatests (fourth row in Table 2).

UVIGO CNN (which was trained using both RGB and depth images) tested over the other datasets gave better results, which it seems that generalizes better. Therefore, it resulted that training with RGB pictures or depth pictures separately did not perform well in recognizing depth and RGB pictures, respectively.

As seen in Table 1, #1 and #2 both have five signers unlike UVIGO which has two signers. More signers mean more variation, and therefore more generalization. However, these results seem to indicate that using RGB and depth images is as important as the number of signers.

Table 2: *Test comparison for the trained CNN in terms of accuracy. The labels of the first column refer to the training material used to train the CNN*

Training data/Testing data	UVIGO	Superpixel	Fingerspelling
UVIGO	**89.76%**	27.97%	20.96%
Superpixel	17.39%	**92.48%**	7.81%
Fingerspelling	10.00%	6.38%	**99.61%**

In a second experiment, the three training datasets were jointly used to train three CNNs.

Table 3 summarizes results for the training of these three CNNs: “All RGB” label means a CNN trained with only RGB images from UVIGO and Superpixel datasets, “All depth” label means a CNN trained with only depth images from UVIGO and Fingerspelling datasets, and “All” label means a CNN trained with both types of images. They are also compare to Massey (#3) and Padova (#4) datasets.

It can be seen that the CNNs based on both RGB and depth images outperformed the CNNs trained with only one kind of image. All (Table 3) exceeded the 50% precision mark on external datasets. All depth also did so for depth, but only for the Padova dataset because it consists of depth images.

Table 3: *Comparison between different type of images. The labels of the first column refer to the training material used to train the CNN*

Training data/Testing data	All RGB	All depth	Massey	Padova
All RGB	91.62%	6.56%	25.53%	7.31%
All depth	8.53%	95.33%	16.73%	59.70%
All	95.21%	96.77%	52.51%	69.63%

It can be concluded that mixing depth and RGB images is an effective way to create a robust and flexible dataset to train systems for SL recognition. However, if only one type of image must be chosen, it is preferable to work with depth datasets because overall have better performance than RGB datasets This may be due to the absence of background noise which was filtered while preprocessing.

3.4.1. Model application

The best trained network was incorporated in a hand sign recognition application, resulting in segmentation speeds of 0.001s for RGB images and 0.007s for depth images (both

hands at the same time), and a prediction time of 0.054s. The whole process was therefore performed in 0.116s.

3.5. Dataset acquisition guidelines

As mentioned, training a sign recognition NN requires significant amounts of data, with sufficient diversity in sign execution and in recording environments to ensure robustness. Considering the results, dataset acquisition guidelines are described next to meet these requirements.

It is therefore recommended to record data from at least 10 different signers and to ensure a balance between male and female and native and non-native signers. It is also recommended to spread recording sessions over several days to ensure a variety of conditions (lighting, clothes, etc.) and settings.

It is highly recommended to record RGB and depth at the same time, preferably with a resolution of at least 640x480 pixels. Depth images are useful because a more precise segmentation is possible. Another advantage is that different lightning conditions, clothes and settings do not have any influence because they are not perceived by the sensors. Nonetheless, recording sex, height and anatomical differences is important.

To train the system for each sign the same way, it is highly recommended to formalize the number of images recorded per sign, signer and recording setting. 100 images per isolated gesture and 20 videos per continuous gesture for each signer is proposed as a good compromise. For a sufficient number of signers, this represents a sufficiently large and sufficiently heterogeneous set of empirical data to avoid training the network with images too similar to each other.

Formalization also requires efficient and convenient organization and annotation. It is therefore recommended to do the following: organize the dataset in a folder tree, with a single folder per signer, and with each containing a sub-folder per sign and sub-sub-folders for each repetition of that sign; identify signer folders using an identification code, and name sign folders after the sign name; identify each image with the signer identification code, the recording date (or code), the repetition number and the sign name; and finally, if feasible, assign each recorded sign a code and use that instead of its name.

It is recommended describe files in terms of three sections: with the lexicon and the associated code; with the signer's name and respective code with information on date of birth, why they use SL and their dominant hand; and with details on the recording session, equipment used, recorder name, recording date, file name and recording conditions. Recommended is to use a spreadsheet (or CSV) because it is very easy to filter out and retrieve key statistics from the data.

4. Conclusions and further work

Although a few studies exist on ASLR, progress is still limited by the lack of datasets. The fact that the few available lack reliability and convenience makes the creation of new datasets mandatory in order to progress in this field.

We have described an experimental framework designed to study the reliability of existing datasets and the combination of RGB and depth data and experimentally trained CNNs. The system achieved over 50% precision for challenging datasets from both natural and artificial recording environments. This method can be used to create a complete and straightforward dataset suitable for research.

Future research will focus on improving system training to match or surpass accuracy rates for both depth and RGB images separately with all datasets included and on extending recognition to the Spanish SL. An expansion of the acquired dataset is also planned, as well as the creation of two computer applications, one for recording and another for real-time gesture recognition.

5. Acknowledgements

This work has received financial support from the Xunta de Galicia (Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019, Galician Research Network TecAnDaLi ED431D 2016/011 and Grupos de Referencia Competitiva GRC2014/024) and the European Union (FEDER/ERDF).

6. References

[1] J. Isaacs and S. Foo, "Hand pose estimation for American sign language recognition," *IEEE Thirty-Sixth Southeastern Symposium on System Theory*, 2004, pp. 132-136.

[2] D. Guo, W. Zhou, M. Wang and H. Li, "Sign language recognition based on adaptive HMMS with data augmentation," *IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, 2016*, pp. 2876-2880.

[3] A. Memo, L. Minto and P. Zanuttigh, "Exploiting Silhouette Descriptors and Synthetic Data for Hand Gesture Recognition", *Eurographics Italian Chapter Conference, Eurographics Association*, pp. 15-23, 2015.

[4] Zuzanna Parcheta and Carlos-D. Martínez-Hinarejos, "Sign Language Gesture Recognition Using HMM", *L.A. Alexandre et al. (Eds.): IbPRIA 2017, LNCS 10255*, pp. 419-426, 2017

[5] Carlos D. Martínez-Hinarejos and Zuzanna Parcheta, "Spanish Sign Language Recognition with Different Topology Hidden Markov Models", *Proceedings of the Interspeech 2017*, Stockholm, Sweden, pp. 3349-3353, 2017.

[6] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677-691, April 1 2017.

[7] E. Tsironi, P. Barros, C. Weber and S. Wermter, "An Analysis of Convolutional Long Short-Term Memory Recurrent Neural Networks for Gesture Recognition", *Neurocomputing*, vol. 268, pp. 76-86, 2016.

[8] X. Shi, Z. Chen, H. Wang and D. Y. Yeung, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting", *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15),* vol. 1, pp. 802-810, 2015.

[9] O. Koller, H. Ney, and R. Bowden, "Automatic Alignment of HamNoSys Subunits for Continuous Sign Language Recognition", *LREC Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, Portorož, Slovenia, pp. 121-128, 2016.

[10] "Grades," [Online]. Available: http://grades.uvigo.es/. [Accessed 12 07 2018].

[11] C. Wang, Z. Liu and S. C. Chan, "Superpixel-Based Hand Gesture Recognition With Kinect Depth Camera", *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 29-39, Jan. 2015.

[12] B. Kang, S. Tripathi, T. Q. Nguyen, "Real-time Sign Language Fingerspelling Recognition using Convolutional Neural Networks from Depth Map", *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, 2015, pp. 136-140.

[13] A. L. C. Barczak, N. H. Reyes, M. Abastillas, A. Piccio and T. Susnjak, "A New 2D Static Hand Gesture Colour Image Dataset for ASL Gestures", *Research Letters in the Information and Mathematical Sciences*, vol. 15, pp. 12-20, 2011.

[14] G. Marin, F. Dominio, P. Zanuttigh, “Hand Gesture Recognition with Leap Motion and Kinect Devices”, *IEEE International Conference on Image Processing (ICIP)*, Paris, 2014, pp. 1565-1569.

[15] Z. Ren, J. Yuan, J. Meng and Z. Zhang, “Robust Part-Based Hand Gesture Recognition Using Kinect Sensor”, *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110-1120, Aug. 2013.

[16] S. C. Chan, C. Wang and Z. Liu, “Hand gesture recognition based on canonical formed superpixel earth mover's distance”, *2016 IEEE International Conference on Multimedia and Expo (ICME)*, Seattle, WA, 2016, pp. 1-6.

[17] X. Sun, Y. Wei, S. Liang, X. Tang and J. Sun, “Cascaded Hand Pose Regression”, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 824-832.

[18] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, “Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition”, *Deutsche Arbeitsgemeinschaft für Mustererkennung Symposium (DAGM), Lecture Notes in Computer Science*, Vienna, Austria, pp. 401-408, Aug. 2005.

[19] C. Neidle, A. Thangali and S. Sclaroff, “Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD)”, *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012*, Istanbul, Turkey, May 27, 2012.

[20] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, “Speech Recognition Techniques for a Sign Language Recognition System”, *INTERSPEECH 2007*, Antwerp, Belgium, pp. 2513-2516, Aug. 2007.

[21] P. Dreuw, T. Deselaers, D. Keysers, and H. Ney, “Modeling Image Variability in Appearance-Based Gesture Recognition”, *ECCV Workshop on Statistical Methods in Multi-Image and Video Processing (ECCV-SMVP)*, Graz, Austria, pp. 7-18, May 2006.

[22] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney, “RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus”, *Language Resources and Evaluation (LREC)*, Istanbul, Turkey, pp. 3785-3789, 2012.

[23] European Sign Language Center, “Spread The Sign,” 2006. [Online]. Available: www.spreadthesign.com. [Accessed 11 07 2018].

[24] F. Ronchetti, F. Quiroga, C. Estrebiu, L. Lanzarini and A. Rosete, “LSA64: An Argentinian Sign Language Dataset”, *XX II Congreso Argentino de Ciencias de la Computación (CACIC)*, 2015.

[25] I. Shao and L. Liu, “Learning Discriminative Representations from RGB-D Video Data”, *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI’13),* pp. 1493-1500, May 2013.

[26] Microsoft Research Cambridge-12 Kinect, “Kinect Gesture Data Set”, [Online]. Available: https://www.microsoft.com/en-us/download/details.aspx?id=52283. [Accessed 12 07 2018].

[27] S. Escalera, X. Baro, H. J. Escalante and I. Guyon, “ChaLearn Looking at People: A Review of Events and Resources”, *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, 2017, pp. 1594-1601, 2017.

[28] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, arXiv preprint arXiv:1409.1556, 2014.