INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Spanish Sign Language Recognition with Different Topology Hidden Markov Models

Carlos-D. Martínez-Hinarejos[1], *Zuzanna Parcheta*[2]

[1] Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camino de Vera, s/n, 46022, Spain
[2] Sciling S.L., Carrer del Riu 321, Pinedo, 46012, Spain
cmartine@dsic.upv.es, zparcheta@sciling.com

Abstract

Natural language recognition techniques can be applied not only to speech signals, but to other signals that represent natural language units (e.g., words and sentences). This is the case of sign language recognition, which is usually employed by deaf people to communicate. The use of recognition techniques may allow this language users to communicate more independently with non-signal users. Several works have been done for different variants of sign languages, but in most cases their vocabulary is quite limited and they only recognise gestures corresponding to isolated words. In this work, we propose gesture recognisers which make use of typical Continuous Density Hidden Markov Model. They solve not only the isolated word problem, but also the recognition of basic sentences using the Spanish Sign Language with a higher vocabulary than in other approximations. Different topologies and Gaussian mixtures are studied. Results show that our proposal provides promising results that are the first step to obtain a general automatic recognition of Spanish Sign Language.

1. Introduction

Natural language technologies have been used in the last decades to provide solutions for many language related problems, such as speech recognition [1], machine translation [2], dialogue systems [3], or sentiment analysis [4], among others. All these tools have as final aim to provide easier communication between people or to obtain valuable information from language resources.

In the case of people with functional diversity, the use of language technologies could make their life easier when they communicate with other people. This is the case of deaf people, where the communication must be done by using visual means such as written language and sign language, which is the most common and natural used.

Sign language uses a combination of hand shapes with different orientations and movements, although they may include arms and body movements and facial expressions. Words are encoded in different combinations of their corresponding hand features, and sentences are formed by the concatenation of these features. This is quite similar to what happens in speech: single words are concatenated to form sentences.

As spoken languages have different implementations, sign languages present differences according to geographical distribution. Sign languages are not comparable to spoken grammar because they only express concepts and ideas (e.g., "He has a yellow shirt" would be in sign language "He yellow shirt").

In any case, sign language communication requires participants to know the sign code. In case one participant does not know the code, human interpreters can provide the service. Of course, this is a limitation because it is not always possible to access to this service. The possibility of obtaining an automatic method that could recognise the sequence of hands features and obtain its corresponding words (although not in the usual spoken grammar) would be a great value for sign language users.

In this paper, we propose the use of Hidden Markov Models (HMM) to recognise Spanish Sign Language (LSE, from *Lengua de Signos Española*), at both word and sentence level. Section 2 presents related work on the matter, Section 3 details the data acquisition and features, Section 4 features the experimental framework and results, and Section 5 offers the conclusions and the possible future work lines.

2. Related work

The idea of having automatic recognition of sign language is not new. There have been many approaches about this task and an overview of them is given in this section.

Sign language recognition can be restricted to sign languages where each sign represents a static single character. This is the case of [5], where video capture and Multilayer Perceptron (MLP) neural networks are used to recognise 23 static signs from the Colombian Sign Language, with an error of 1.85%. Another example is [6], where a Gaussian Mixture Model (GMM) allows to recognise 5 static signs obtained from image captures with an error rate of 5.8%

When dynamic features are taken into account, recognition can be restricted to single signs or to whole words. The former approach is followed in [7], that employs Microsoft *Kinect* features to identify 8 different signs using weighted Dynamic Time Warping (DTW) with an error rate of 3.3%. The latter approach is considered in [8], where a combination of k-Nearest Neighbour classifiers and DTW allows to recognise 18 signs extracted by using the *Leap Motion*[1] sensor with an error rate of 55.6%.

Several works have employed Hidden Markov Models (HMM) to exploit the temporal nature of dynamic features in recognition. HMM have been applied in both single word recognition and sentence recognition. For example, [9] employs HMM to recognise a set of 262 gestures obtained from coloured video images with recognition errors from 52.4% to 5.7% in different conditions. [10] presents the recognition of 40 words from American Sign Language, including the recognition of sentences with some fixed length formed by those words; features were extracted from video images obtained by two different recording systems; recognition errors vary from 25.5% to 8.1% for single words, and from 3.2% to 2.2% for fixed length sentences, in different conditions.

[1] https://www.leapmotion.com/product/desktop

http://dx.doi.org/10.21437/Interspeech.2017-275

A more sophisticated approach was taken in [11], where HMM output distributions were modelled by using Self Organising Features Maps (SOFM). Data were extracted by using tracking gloves and it is collected from a vocabulary composed of 5113 signs. In this case, the isolated word task has an average recognition error of 9.5%, and the sentence recognition task has errors between 27.6% and 8.7% in different conditions.

With respect to employing HMM for LSE, a first approach was taken in [12], that employed a dataset to recognise 91 single words and sentences combining a subset of 68 words of those 91. Acquisition was performed by using the *Leap Motion* sensor (which makes acquisition cheaper and simpler comparing to other works). Reported results are a 12.6% of error in single word recognition and 13.0% of Word Error Rate (WER) in sentence recognition. In this work, we employ the same dataset than in [12], but we explore how different HMM topologies influence the system performance.

3. Data acquisition and features

Data acquisition for sign language was done by using the *Leap Motion* sensor. *Leap Motion* is a sensor optimised to extract information on the hands position. Among their advantadges, *Leap Motion* is optimised for extracting hand three-dimensional information with high precision, has a small size, a complete API for its programming, and a good quality/price relation. Among their limitations, it has problems for overlapping hands gestures and it is sensitive to sunlight. The limitation on overlapping gestures led us to define a slightly modified set of gestures for some words of LSE.

Data acquisition was done with the software described in [8], modified to allow the acquisition with both hands and not only one hand. For each hand and finger 3 features are generated, and for each hand 3 more extra features are acquired, giving 21 features for each hand, 42 feature for each sample.

Some issues had to be solved in order to prevent training problems. The first one happens with single-hand movements; in that case, the features for the other hand remain null, and this could cause training errors because of the null variability of that part of data; to avoid that, in those cases the features for one hand are copied for the other. The second issue appears with very short gestures, that cannot be used to estimate HMM with more states than samples; in that case, the number of samples was artificially incremented by interpolation between samples.

The resulting dataset is publicly available[2]. The dataset is composed of a set of isolated word gestures and another set of sentences. Isolated word subcorpus is formed by samples corresponding to 91 words, with 40 examples performed by 4 different people for each word (3640 acquisitions). The words were taken from an on-line LSE course[3] and allow to form some meaningful sentences.Continuous (sentence) subcorpus was acquired by a single person. The number of acquisitions is of 274 sentences, using a subset of the vocabulary of the isolated corpus (68 words).

4. Experiments and results

For training and decoding, HTK [13] was used. A cross validation approach using 4 partitions was used in the different experiments. Confidence intervals of 95% for the error rates were calculated by using bootstrapping [14] with 10,000 repetitions.

[2] https://github.com/Sasanita/spanish-sign-language-db

[3] http://xurl.es/zjuqm

Table 1: *Cross validation test classification error results in isolated words for fixed topology HMM. Topology number corresponds to number of states for each HMM. Best result in boldface, with a 95% confidence interval of 0.9.*

Gaussian number	Topology						
	1	2	3	4	5	6	7
1	16.5	14.6	13.6	13.0	12.1	12.1	11.9
2	15.0	13.8	13.2	12.2	11.6	11.6	11.4
4	14.7	13.5	12.6	12.2	11.6	11.6	**11.3**
8	14.5	13.3	12.5	12.1	11.4	11.6	**11.3**

Table 2: *Cross validation test classification error results in isolated words for variable topology HMM. Topology number corresponds to k factor. Best result in boldface, with a 95% confidence interval of 0.9.*

Gaussian number	Topology				
	1	1.5	2	2.5	3
1	11.7	11.5	14.9	15.4	16.4
2	11.0	11.5	14.2	14.5	15.1
4	10.7	11.4	13.9	14.4	14.7
8	**10.6**	11.3	13.6	14.5	14.6

4.1. Isolated word classification

The 91 different words were modelled with 91 HMM. Two options for the HMM topology were selected: uniform number of states for each word (fixed topology), and variable number of states according words lengths and transitions (variable topology). Our hypothesis is that using variable topology must give better results since the length of the samples is quite different depending on the words that are employed; this is in contrast with the usual use of HMM in speech recognition, where all phones are usually modelled with the same topology.

For the fixed topology, number of states selected were from 1 to 7 (longer HMM provided training errors because lack of data). For the variable topology, the following process was adopted: features' values for each word were normalised between 0 and 1 (scaling between the minimum and maximum value for each feature in the acquired samples), Euclidean distance from sample i to sample $i+1$ (d_i) was calculated, and average distance value for all n samples in the word ($\bar{d}$) was computed. After that, the number of states was equal to $\sum_{i=1}^{n-1} \delta(d_i, \bar{d}, k)$, where:

$$\delta(d_i, \bar{d}, k) = \begin{cases} 1 & d_i \geq k\bar{d} \\ 0 & \text{otherwise} \end{cases}$$

with $k \in \mathbb{R}^+$. In our case with used $k \in \{1, 1.5, 2, 2.5, 3\}$. Lower k values were disregarded because of training problems (many samples were not long enough to contribute to training). Figure 1 shows HMM number of states distribution for each k.

In both types of topologies, cross validation with 3 partitions for training and 1 partition for test was performed. All HMM followed a left-to-right without skips transition topology. Initial prototypes were initialised to average mean and covariance for the data on the training partitions (by using HCOMPV

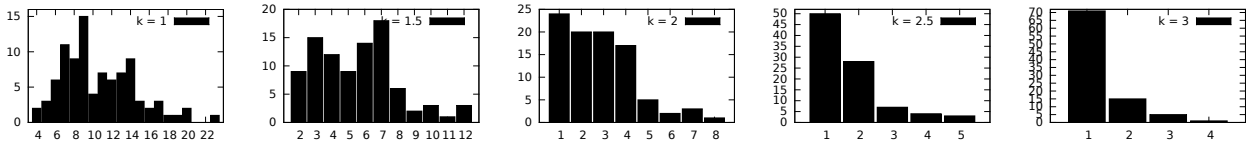

Figure 1: *Number of states distribution for variable topology for* $k \in \{1, 1.5, 2, 2.5, 3\}$.

Table 3: *Most confused gestures in isolated gesture classification.*

Num of confusions	Reference	Recognition
9	five	hello
9	thank you	nineteen
10	yellow	how are you
10	yes	make signs

(a) "hello" sign. (b) "five" sign.

Figure 2: *Pair of confused signs.*

HTK tool). After initialisation, several training iterations (by using HEREST HTK tool) were performed (up to 9 iterations). The best average iteration (according to the cross validation test) set of models was employed to duplicate gaussian numbers (by using HHED HTK tool), and a new set of reestimation iterations was performed. The process continued until 8 gaussian per state models were obtained. Recognition tests were performed on the different sets of HMM (by using HVITE HTK tool) with a parallel word language model.

Test classification error results (corresponding to best training iteration for each number of gaussians) are presented in Table 1 for fixed topologies and in Table 2 for variable topologies.

As was expected, recognition errors with variable topologies are slightly better than those obtained using fixed topologies. Results are coherent since for lower k number, higher number of average states, which coincides with the results tendency for fixed topology. However, differences between best results in the two options are not significant.

Table 3 presents the most confused signs using best topology. These signs present some similarities in shape of hands or trajectory during gesture performance. Figure 2[4] shows the representation for signs "hello" and "five". The two signs have the same shape of hand, although the "hello" sign implies a swinging movement not present in "five".

[4]Images taken from online sign language dictionary https://www.spreadthesign.com/es/ .

Table 4: *Sentence recognition WER using isolated words for fixed topology HMM. Topology number corresponds to number of states for each HMM.*

Gaussian number	Topology						
	1	2	3	4	5	6	7
1	67.6	58.3	57.5	57.2	57.5	56.9	55.2
2	63.5	56.8	57.7	56.7	58.7	56.3	55.9
4	63.0	57.2	57.9	56.6	58.6	56.0	55.9
8	62.8	56.7	57.2	56.0	58.1	55.6	55.8

Table 5: *Sentence recognition WER using isolated words for variable topology HMM. Topology number corresponds to* k *factor.*

Gaussian number	Topology				
	1	1.5	2	2.5	3
1	52.3	55.8	62.5	65.4	65.4
2	52.2	55.1	61.1	64.8	64.0
4	51.7	55.4	61.4	64.5	63.9
8	51.1	54.6	60.8	63.2	64.4

4.2. Sentence recognition

The amount of data for continuous subcorpus (274 sentences) is not enough to use it alone for training tasks. Thus, all results for sentence recognition employ as initial step the models trained with the isolated words. A first approach is to employ isolated words models to directly recognise sentences. In that case, all isolated corpus is employed to train the HMM (including several training iterations and gaussian duplication). For all sentence decoding experiments, the language model was a manual grammar and recognition error was computed with WER. Table 4 presents the WER for the different models of fixed topology, and Table 5 for those of variable topology.

In this approach, results are really bad, which was expected since isolated word models do not model transitions between each word. In any case, the effect of the variable topology in this case is quite beneficial with respect to the use of fixed topology. Anyway, in both cases recognition errors are so high that would prevent to employ the system in a real environment.

Another approach is to use continuous sentence data to reestimate the HMM. With this approach, transitions between words would be properly modelled. A cross validation approach with 4 partitions was used in order to obtain more reliable average results. Thus, for each best isolated set of models corresponding to each topology and number of gaussians, iterative training reestimations with the sentence data was performed.

Table 6: *Sentence recognition WER using fixed topology HMM and reestimating isolated word HMM with sentence data. Topology number corresponds to number of states for each HMM. Best result in boldface. 95% confidence interval of 2.4.*

Gaussian number	Topology						
	1	2	3	4	5	6	7
1	28.3	22.3	19.6	17.8	17.7	**14.8**	16.2
2	27.0	22.6	21.2	22.0	19.1	19.5	20.2
4	31.2	25.2	24.8	24.8	21.7	23.5	22.7
8	31.9	26.9	28.3	26.6	24.9	26.3	24.5

Table 7: *Sentence recognition WER using variable topology HMM and reestimating isolated word HMM with sentence data. Topology number corresponds to k factor. Best result in boldface. 95% confidence interval of 2.5.*

Gaussian number	Topology				
	1	1.5	2	2.5	3
1	**16.7**	15.6	21.8	27.7	26.4
2	20.7	19.9	23.7	26.7	26.0
4	22.0	23.9	24.5	28.1	30.2
8	22.8	26.3	25.4	31.7	33.3

Table 8: *Sentence recognition WER using fixed topology HMM and training with isolated word and sentence data. Topology number corresponds to number of states for each HMM. Best result in boldface. 95% confidence interval of 2.3.*

Gaussian number	Topology						
	1	2	3	4	5	6	7
1	36.2	24.7	21.4	17.3	16.7	14.4	15.6
2	30.5	21.5	18.4	16.5	16.0	14.7	**12.9**
4	24.3	20.2	18.3	15.7	13.8	**12.9**	13.4
8	25.0	20.4	18.2	15.2	14.6	**12.9**	13.2

Table 9: *Sentence recognition WER using variable topology HMM and training isolated word and sentence data. Topology number corresponds to k factor. Best result in boldface. 95% confidence interval of 2.3.*

Gaussian number	Topology				
	1	1.5	2	2.5	3
1	13.3	16.7	25.1	36.8	35.1
2	12.3	16.1	22.1	27.3	29.3
4	**11.8**	14.2	19.9	23.1	26.8
8	12.1	14.5	18.6	22.8	26.3

Best iteration results with this approach and fixed topology are shown in Table 6, and in Table 7 for variable topology.

As can be seen, error rates decrease dramatically, making feasible the potential use as real system. Apart from that, in this approach fixed topologies behave better than variable topologies, although differences are not significant. Thus, retraining with continuous data seems suitable to obtain good results.

Anyway, the continuous data is not fully used since it does not participate in gaussian increment. Final approach will train from scratch the HMM using the whole set of isolated data and the corresponding cross validation partitions of sentence data. This process includes initialisation, single gaussian estimation and retraining, and gaussian duplication and retraining. Decoding results (best training iteration) for fixed topologies can be seen in Table 8 and for variable topologies in Table 9.

With this last option, variable topologies present a better recognition result than fixed topologies, although results are not significantly better. However, in general the tendency is that topologies fitted to the length of signs work better than using the same topology for all of them.

As happened with isolated words, most confused signs in sentence recognition have similar hand shapes or similar trajectories during gesture. In fact, most of confused signs are the same, as can be seen in the confusion matrix for sentence recognition presented in Table 10.

5. Conclusions and future work

In this work we have presented a sign language recognition system for Spanish Sign Language (LSE). The acquisition system is the *Leap Motion* sensor, a cheap and versatile hardware that makes data acquisition feasible for a wide range of users. The covered vocabulary size is higher than in many other works that employ Hidden Markov Model recognition for this task, and both isolated word and sentence recognition is performed with promising error rates (about 10% for isolated words and about 12% WER for sentences). The proposal of using variable topologies obtains better results than fixed lenght topologies, although statistical significant improvements are not obtained.

Future work will be directed to the use of separated recognition for each hand and posterior integration of the recognised posibilities using output combination techniques like ROVER [15] or Lattice Rescoring [16]. The use of more sophisticated models based on Deep Neural Networks [17] is another option, although for that case we think that a more massive acquisition would be needed. This extended acquisition, including more words and a less restricted language model for sentences, will be explored in another task.

6. Acknowledgements

Work partially supported by MINECO under grant DI-15-08169, by Sciling under its R+D programme, by MINECO/FEDER under project CoMUN-HaT (TIN2015-70924-C2-1-R), and by Generalitat Valenciana (GVA) under reference PROMETEOII/2014/030.

Table 10: *Most confused gestures in sentence recognition.*

Num of confusions	Reference	Recognition
9	thank you	nineteen
9	be born	good morning
9	how are you	good
10	yes	make signs
11	yellow	how are you

7. References

[1] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA, USA: MIT Press, 1997.

[2] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[3] O. Lemon, A. Gruenstein, and S. Peters, "Collaborative activities and multi-tasking in dialogue systems: Towards natural dialogue with robots," *TAL. Traitement automatique des langues*, vol. 43, no. 2, pp. 131–154, 2002.

[4] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 347–354.

[5] J. D. Guerrero-Balaguera and W. J. Pérez-Holguín, "FPGA-based translation system from colombian sign language to text," *DYNA*, vol. 82, pp. 172 – 181, 2015.

[6] F.-H. Chou and Y.-C. Su, "An encoding and identification approach for the static sign language recognition," in *Advanced Intelligent Mechatronics (AIM), 2012 IEEE/ASME International Conference on*. IEEE, 2012, pp. 885–889.

[7] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici, "Gesture Recognition Using Skeleton Data with Weighted Dynamic Time Warping," in *VISAPP*, 2013, pp. 620–625.

[8] Z. Parcheta, "Estudio para la seleccin de descriptores de gestos a partir de la biblioteca "LeapMotion"," 2015, tFG, EPSG, UPV.

[9] K. Grobel and M. Assan, "Isolated sign language recognition using hidden markov models," in *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, vol. 1. IEEE, 1997, pp. 162–167.

[10] T. Starner, A. Pentland, and J. Weaver, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, Dec. 1998. [Online]. Available: http://dx.doi.org/10.1109/34.735811

[11] W. Gao, G. Fang, D. Zhao, and Y. Chen, "A chinese sign language recognition system based on sofm/srn/hmm," *Pattern Recogn.*, vol. 37, no. 12, pp. 2389–2402, Dec. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2004.04.008

[12] Z. Parcheta and C.-D. Martínez-Hinarejos, "Sign language gesture recognition using hmm," in *Proceedings of IbPRIA 2017: 8th Iberian Conference on Pattern Recognition and Image Analysis*, 2017, p. To appear.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK book*. Cambridge university engineering department, 2006.

[14] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performanc e evaluation," in *Proc. of ICASSP*, vol. 1, 2004, pp. 409–412.

[15] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proceedings of Automatic Speech Recognition and Understanding (ASRU 1997)*, 1997, pp. 347–354.

[16] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: lattice-based word error minimization." in *Eurospeech*, 1999, pp. 495–498.

[17] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.