Sign Language Gesture Recognition Using HMM

Zuzanna Parcheta[1(✉)] and Carlos-D. Martínez-Hinarejos[2]

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

Abstract. Gesture recognition is very useful in everyday life for tasks related to computer-human interaction. Gesture recognition systems are usually tested with a very large, complete, standardised and intuitive database of gesture: sign language. Unfortunately, such data is typically very large and contains very similar data which makes difficult to create a low cost system that can differentiate a large enough number of signs. This makes difficult to create a useful tool for allowing deaf people to communicate with the hearing people. The present work presents a sign recognition system for the Spanish sign language. The experiments conducted include separated gesture recognition and sequences of gestures. This work extends previous work by augmenting the size of data set (91 signs), higher than most of the state of the art systems (around 20 gestures). Apart from that, the proposed recognition system performs recognition of dynamic gestures, in contrast to most studies that use static gestures, which are easier to recognise. Finally, the work studies the recognition of sequences of gestures corresponding to grammatically correct phrases in Spanish sign language. For both tasks, Hidden Markov Models are used as recognition models. Results presented for classification of separated gestures are compared with other usual classification techniques, showing a better recognition performance. The current data set, which has been captured using the *Leap Motion* sensor, will be publicly available to the research community in gesture recognition.

Keywords: Gesture classification · Sign language translator · HTK · Leap Motion · HMM · K-NN · DTW

1 Introduction

In the world there are 70 million deaf people who use sign language as their first language [1]. In case of Spain, there are over 100,000 people native users of Spanish sign language (LSE). To communicate with non-users, interpreter service is required. However, this service is available only in special situations, such as interactions with police, help with administrative matters, etc. At the moment there is no tool that facilitates communication between deaf and hearing

L.A. Alexandre et al. (Eds.): IbPRIA 2017, LNCS 10255, pp. 419–426, 2017.
DOI: 10.1007/978-3-319-58838-4_46

people in everyday life. From a long time ago, researchers have been struggling to create a tool that can automatically translate sign language into oral language. Therefore, the main motivation of this work is to explore tools and models that, in the future, can become an aid in the daily life of the deaf people.

More formally, this work deals mainly with gesture recognition of Spanish sign language by means of Hidden Markov Models (HMM) and their comparison with other techniques such as K-Nearest Neighbour (K-NN) and Dynamic Time Warping (DTW). The gesture acquisition tool is from the *Leap Motion* sensor[1]. Paper is organised as follows: Sect. 2 presents previous related work, Sect. 3 describes the followed methodology, Sect. 4 provides the experimental framework and the results, and Sect. 5 summarises conclusions and future work lines.

2 Related Work

Current state of the art in the area of gesture recognition covers different approximations. There are many studies on sign gesture recognition but most of them are related gestures corresponding to alphabetical signs. In such case, the set of data is very reduced (about 25 gestures) and signs are typically static. Hence, the gestures can be captured with a digital image camera. For example, in [2], they recognised 23 gestures from the Colombian sign language alphabet in which they obtained a 98.15% of accuracy. In [3] similar techniques were used, but in addition to a digital camera to capture data, they also used the *Kinect* sensor. From 30 gestures they could recognise 20 signs using a digital camera and 25 using the *Kinect* sensor. The work presented in [4] also used *Kinect* combined with Dynamic Time Warping (DTW) to obtain a 96.7% of accuracy using a data set consisting of 8 different gestures with 28 samples per gesture class.

Other studies only used three dimensional data captured with different sensors. For example, in [5], experiments of classification of 18 dynamics signs were conducted by using *Leap Motion* for data capture. The gesture data was separated into 2 parts: *(1)* the static part, which was the initial shape of the hand before conducting a gesture, which was recognised using K-NN; and *(2)* the trajectory followed by the hand during the gesture, which was recognised by DTW. The accuracy in case *(1)* was 95.8% and in case *(2)* was 86.1%. However, the gesture was only considered to be correctly recognised if both parts match, giving a final accuracy of 44.4%. In [6], phrases from American sign language formed by 40 different signs were recognised. This data set comprised 395 of 5 words each one. 99 phrases were dedicated to evaluation purposes and the rest to training. The training and the recognition was done using Hidden Markov Models. The gestures were captured using a digital camera. The accuracy of this experiment was a 95%. Surprisingly, very few works are found in the literature (apart from this one) that employ HMM for sign language recognition.

The current work provides a study about recognition of phrases of sign language with a larger size (when compared to previous work) data set. Also,

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

the novelty of this work is about providing a study about the gestures recognition using HMM and a sensor which provides 3D data.

3 Methodology

3.1 The *Leap Motion* Sensor

Leap Motion is an optimised sensor in order to obtain three-dimensional information from hands gestures. The gesture device emits infrared light with which illuminates its effective range. Objects in the effective range reflect infrared light, and depending on the amount of light incident on the two-camera lenses incorporated in the sensor, it is possible to determine how far a given object is from the sensor. Of all the gestural interfaces available in the market, this one was chosen because it is optimised to obtain three-dimensional information from the hands. In addition, it is very small and discreet, it has a suitable effective range for gestures of sign language, it has a very complete application programming interface (API), it is very precise, and it has a good quality/price ratio. Apart from all these advantages it has certain limitations. For instance, it has many problems with gestures where hands overlap or touch one another. Therefore, the gestures used in this work are approximations to the real gestures of sign language. Also, sunlight interferes with the sensor, producing noise, because its operating principles are based on the reflection of infrared light.

3.2 Hidden Markov Models

For years, HMMs have had special success in speech recognition [7]. Nowadays they constitute a popular technique applied for modelling time sequence data, like gesture recognition [8], recognition of handwritten text [9], etc. The HMM represents probability distributions over sequences of hidden states and observations. Two types of HMMs exist according to the kind of symbols observed: discrete HMMs, where the symbols correspond to discrete magnitudes, and continuous, if the symbols correspond to continuous magnitudes [10]. In this work, HMMs are used for gesture recognition. In this case, the emitted symbols are feature vectors in a real space (i.e., continuous HMM were used) obtained from the gestures corresponding to signs from LSE.

3.3 Hidden Markov Model Toolkit

The Hidden Markov Model Toolkit (*HTK*) [11] is a set of tools created to manipulate HMMs. *HTK* was primarily created to manipulate data for speech recognition, but it can also be used to solve other problems by using stochastic analysis techniques, such as image recognition, determination of valid sequences of human DNA, or gesture recognition. The architecture of *HTK* for HMM management is very flexible and allows to adjust multiple parameters depending on the type and the complexity of the problem to solve. For all these reasons, this tool was used for the experiments commented in this article.

4 Experimental Framework

4.1 3D Data Set

For the experiments, an own data set was created with the idea of making it public[2]. The gestures that compose our database approach the LSE gestures due to the limitations of the sensor. Sampling has been carried out with the software described in [5], where only one-hand signs were performed. For this work, gestures were made with one and both hands. This program segments different gestures when the movement amount threshold[3] is exceeded. The program starts capturing the data, and ends the capture when the variable that describes the amount of movement produced goes below the threshold. The sampling rate is fixed at 30 frames per second. The sensor provides the data in numerical format. Finally, a gesture is a matrix where the number of rows is the number of captured images and the number of columns is the number of observed variables.

The variables involved were the directions on the X, Y and Z axes of each hand and each finger, where each of them describe the direction in one of the axes (i.e., the hand/finger points) and the inclination of the hand according to the X, Y and Z axes (which expresses the hand angle with respect to the horizontal plane). In total, 21 variables are available for each hand, that is, 42 variables in total. In order to avoid training problems with *HTK*, when single-hand gestures appeared, the data of the other hand were replicated from the data of the moving hand. In the case of very short samples, which might cause training problems, the number of frames was incremented by using interpolation.

The data set for the classification task of separated gestures consisted of 91 isolated words with 40 samples per word (a total of 3640 samples). Data were acquired from 4 different people, because each person performs the gestures in a different way. The words that have been chosen come from an on-line course [12] and a LSE dictionary [13]. The words were chosen in order to give the possibility of forming different sentences that have a correct grammatical structure and semantics in sign language. Attending to that, the chosen words pertained to the followings groups: colours, numbers, personal pronouns, possessive adjectives, adverbs, greetings, courtesy phrases, verbs, names, confirmation, quantities, prepositions, and interrogative pronouns.

For the task of consecutive gesture recognition, 274 sentences were captured. The vocabulary is the same than that appears in first task. These data have the same format than the isolated gestures. In both tasks the evaluation was conducted with cross validation using 4 partitions.

4.2 Gesture Classification

This section provides results on the recognition of the 91 separated gestures. Data were generated using the aforementioned *Leap Motion* sensor. Training

[2] The data is available in https://github.com/Sasanita/spanish-sign-language-db.git.
[3] The value of the movement is the weighted sum of several variables, such as the speed of each hand between two consecutive images [5].

and evaluation were performed using the *HTK* toolkit. For this experiment, the data were divided were into 4 partitions to use cross validation technique. For training different HMM topologies were tried by employing initially models of 4, 5, 6, 7 and 8 states (including artificial initial and final states). In speech recognition the transcription file used to train the model contains the phonemes transcribed from the audio sample, but in the case of this work, the transcription file contains whole words as basic units. A standard training approximation with common initialisation for all words and Forward-Backward estimation was employed. Effects of the number of Forward-Backward iterations were explored. Finally, the results shown in Fig. 1 were obtained. Models with 7 and 8 states obtained the same results. The best accuracy was obtained with the model of 7/8 states and 7 training iterations. This model obtained 84.6% word accuracy.

To improve this result, increment of the number of Gaussians within the emission probability mixture model was conducted. The model with the highest accuracy was used in the increment. In Fig. 2 the comparison of models of 8 states with 1 Gaussian and 2 Gaussians is shown. The accuracy increased to 87.4%. A mixture model with 4 Gaussians was tried but the number of samples by Gaussian was insufficient to train the model and caused training errors. Apart from accuracy, training and test speed were measured. The time required to obtain the best topology was 8:32 min. The second training with 2 Gaussian Mixtures lasted 1:45 min. In total, the time consumed by the complete experiment was about 10:17 min. In addition, the experiment using the methodology described in [5] with the data from the current work was conducted. The purpose was to achieve an objective comparison of results of both works. First, the classification using K-NN and DTW algorithms was done. For K-NN we took the first frame (42 values) of each sample, which represents the initial shape of hands. The next step was calculating the Euclidean distance between each vectors of test samples and all vectors of training samples. The smallest distance determines the class which belongs to. In order to recognise the trajectory with DTW, we took the complete data matrix of each sample. We calculated the sum of the distances between each column (vector of each variable).

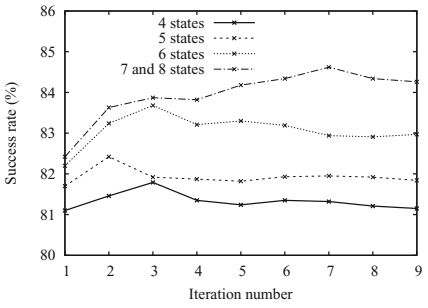


Fig. 1. Topologies comparison.

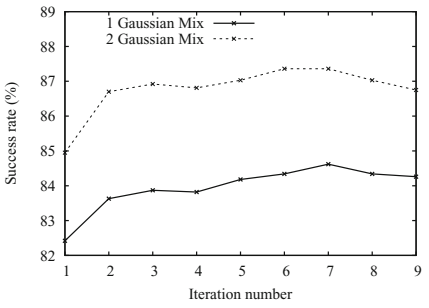


Fig. 2. Model of 8 states trained with 1 and 2 Gaussian Mixtures.

Table 1. Results using the methodology of [5] and comparission with actual metodology. Times in seconds.

	HMM	KNN + DTW
Accuracy	87.4%	88.4%
Elapsed time	519	9383

Table 2. Confusion matrix of most confused signs (40 samples for each sign). Original words in Spanish.

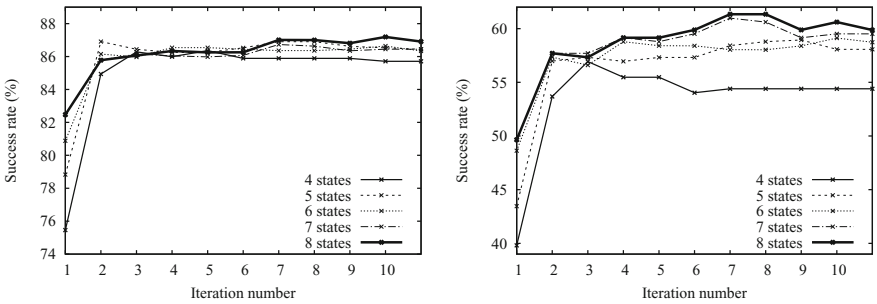
Sign	Confused sign and corresponding frequency				
She	You(female)-8	We(female)-7			
Not	One-10	You(female)-6	He-2	Eleven-2	Forget-1
How are you	Well-13	Grey-2			
So-so	Please-7	Good night-4	How are you-3	Born-2	
Red	Spain-6	Eyes-4	Me-4	Shirt-1	

When the recognition of the static part (initial configuration of hands) using K-NN and the trajectory using DTW, which both belong to the correct class, the final recognition result for each sample is considered as correctly recognised. Apart from accuracy, elapsed time was also measured. Table 1 shows classification accuracy obtained by using HMM and by using K-NN + DTW. The accuracy of the previous work is about 1% better, but the time required is much higher. Our method is 18 times faster, since DTW has to perform the sum of differences of each variable and of classification it is necessary to calculate the difference between each pair of samples. Table 2 provides the confusion matrix of most confused signs. Different groups of confused signs have similar initial or final shape of hands, similar trajectory or some part of gesture are exactly the same. For example, to indicate male or female gender, only an additional movement is required, which makes it prone to confusion.

4.3 Consecutive Gesture Recognition

This task performed recognition of consecutive gestures. In Sect. 1 it has been mentioned that context may help in determining the meaning of a sign, which is the focus of this task. The experiment consists in using the consecutive signs contained within the data set in order to build logical sentences. These sentences were automatically generated using a grammar file created in collaboration with the *Fesord*[4] association. Only 68 words were used in this grammar file, since data acquisition was very costly. Listing 1.1 shows part of this grammar. In order for training to be statistically significant, at least 3 examples of use of each word

[4] http://www.fesord.org.



(a) Word level. (b) Sentence level.

Fig. 3. Topologies comparison for consecutive gesture recognition.

in each partition are needed. That is the reason why the number of used words was reduced in this task. In total, 274 sentences were implemented and cross validation with 4 partitions was used.

Listing 1.1. Part of the grammar file used to generate logical LSE sentences.

```
$colour = black | yellow | blue | red | green | brown;
$number = one | two | three | ... | eight | nine | ten;
$person = father | mother | child | brother;
$personalpronoun = I | you | we | she | he | them;
$family = $personalpronoun ( brother | child ) $number (
    to_have | not_to_have );
$like = ( $personalpronoun | $personalpronoun $person ) (
     $activity | $colour )  ( to_like | not_to_like);
( $like | $understand | ... | $give | $know | $family )
```

Figure 3 shows the word accuracy and Fig. 3b the sentence accuracy for different topologies and training iterations. The best result was obtained with the model with 8 states after 7 training iterations, which led to 61.3% sentence accuracy and 87.0% word accuracy. In the case of consecutive gesture recognition it was impossible to increase the number of Gaussians due to lack of data.

4.4 Discussion

The results of the isolated gestures recognition provided 87.4% gesture accuracy. In the consecutive gesture recognition part, accuracies were 87.0% at the word level and 61.3% at the sentence level. We consider these results satisfactory as a first step towards creating a useful tool for deaf people. Regarding previous work, several improvements have been introduced: more complex models (such as HMMs) have been used, the speed of training and recognition was improved, and the use of a parametric classifier leads to lower storage space requirements in the device where the gesture recognition system will be used.

5 Conclusions and Future Work

In this work we presented a first approximation to gesture classification and recognition for LSE. Results are promising and allow us to think of a future implementation that may aid deaf people in their daily communication.

In future work, we would like to expand the data set and investigate the use of more Gaussians, as well as deep learning, in the consecutive gesture recognition task. Different techniques of surrogate data, described in [14], will be tested to avoid the expensive manual generation of data.

As mentioned earlier, this work is only a small step for the creation of a tool that could enhance the quality of people's life who use LSE. It is expected that with the continuation of this work, a system could be achieved that could effectively break down communicative barriers between many people.

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.

References

1. World federation of the deaf. wfdeaf.org/human-rights/crpd/sign-language/
2. Guerrero-Balaguera, J.D., Pérez-Holguín, W.J.: FPGA-based translation system from colombian sign language to text. DYNA **82**, 172–181 (2015)
3. Priego Pérez, F.P.: Reconocimiento de imágenes del lenguaje de señal Mexicano. TFG, Instituto Politécnico Nacional (2012)
4. Celebi, S., Aydin, A.S., Temiz, T.T., Arici, T.: Gesture recognition using skeleton data with weighted dynamic time warping. In: VISAPP, pp. 620–625 (2013)
5. Parcheta, Z.: Estudio para la selección de descriptores de gestos a partir de la biblioteca "LeapMotion". TFG, EPSG, UPV (2015)
6. Starner, T.E.: Visual recognition of American sign language using hidden Markov models. MIT, DTIC Document (1995)
7. Huang, X., Ariki, Y., Jack, M.: Hidden Markov Models for Speech Recognition. Columbia University Press, New York (1990)
8. Liu, K., Chen, C., Jafari, R., Kehtarnavaz, N.: Multi-HMM classification for hand gesture recognition using two differing modality sensors. In: Circuits and Systems Conference (DCAS), pp. 1–4. IEEE, Dallas (2014)
9. Patil, P., Ansari, S.: Online handwritten devnagari word recognition using HMM based technique. Int. J. Comput. Appl. **95**(17), 17–21 (2014)
10. Salcedo Campos, F.J.: Modelos Ocultos de Markov. Del reconocimiento de voz a la música. LuLu (2007)
11. Young, S., et al.: The HTK Book. Entropic Cambridge Research Laboratory, Cambridge (2006)
12. Curso online de LSE. http://xurl.es/zjuqm
13. Diccionario online de LSE. https://www.spreadthesign.com/es/
14. Schreiber, T., Schmitz, A.: Surrogate time series. Physica D **142**(3–4), 346–382 (2000)