IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Sign Language Gesture Classification Using Neural Networks

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

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

Abstract

Recent studies have demonstrated the power of neural networks for different fields of artificial intelligence. In most fields, such as machine translation or speech recognition, neural networks outperform previously used methods (Hidden Markov Models with Gaussian Mixtures, Statistical Machine Translation, etc.). In this paper, the efficiency of the LeNet convolutional neural network for isolated word sign language recognition is demonstrated. As a preprocessing step, we apply several techniques to obtain the same dimension for the input that contains gesture information. The performance of these preprocessing techniques on a Spanish Sign Language dataset is evaluated. These approaches outperform previously obtained results based on Hidden Markov Models.

Index Terms: gesture recognition, human-computer interaction, gesture classification, sign language, Leap Motion, Convolutional Neural Networks

1. Introduction

Nowadays, technology has advanced to an unbelievable point in helping people with almost any kind of disability. People who are deaf or hard of hearing often experience limitations in their everyday life. However, assistive technology helps deaf people in their daily life. There are two main types of assistive devices for deafness: 1) Devices to enhance listening, e.g. frequency modulated (FM) systems [1], infrared (IR) systems [2], cochlear implants [3], and hearing aids [4] and 2) Devices to convey information visually, e.g. alerting devices [5], captioning [6], and real-time transcription systems [7]. Apart from that, most deaf people use sign language as their preferred way to communicate. However, there is no tool to help them interact with non-users of sign language. Fortunately, contemporary technology allows for developing systems of automatic translation from sign language into oral language.

To solve this problem it is necessary to create a tool which allows deaf people to communicate with non-users of sign language in a comfortable and fluid way. The researchers' commitment to the deaf community is to improve the quality of the automatic systems that are used for this task by exploring new technologies and recognition patterns.

In this work, we explore Convolutional Neural Networks (CNN) to perform gesture classification for isolated words coming from a sign language database generated using the Leap Motion[1] sensor. Previous work [8, 9] demonstrated that it is possible to classify this dynamic gesture database using Hidden Markov Models with a 10.6% error rate. In our current work, this score is significantly improved to an error rate of 8.6%.

The article is organised as follows: Section 2 presents the related work on sign language recognition, Section 3 details the experimental setup (database, models, partitions, etc.); Section 4 details the experiment parameters and their corresponding results; finally, Section 5 offers conclusions and future work lines.

2. Related work

The first step to developing data based gesture recognition systems is data acquisition. For data acquisition there are devices available which use accelerometers [10], electromyography (Myo's armband physical principle) [11], or infrared light (Kinect's and Leap Motion physical principle) [12, 13]. Also, to get gesture data it is possible to use cameras warn on an arm [14] or data gloves [15]. Publications about gestures recognition can be divided into two main groups depending on whether signs are static or dynamic.

There are quite a few studies about static sign recognition. In [16], video capture and Multilayer Perceptron (MLP) neural networks are used to recognise 23 static signs from the Colombian Sign Language. In [17], Gaussian Mixture Models are applied to recognise 5 static signs obtained from image captures.

Some authors also consider dynamic gestures recognition. In [18], the Microsoft *Kinect* is used to obtain features to identify 8 different signs using weighted Dynamic Time Warping (DTW). The latter approach is mentioned in [8], where a combination of k-Nearest Neighbour classifiers and DTW allows the recognition of 91 signs extracted by using the Leap Motion sensor. This work was continued in [9] where different typologies of Continuous Density Hidden Markov Models were applied. [19] describes a system to recognise very simple dynamic gestures which uses Normalised Dynamic Time Wrapping. In [20], authors introduce a new method to obtain gesture recognition, called Principal Motion Components (PMC). However, they do experiments with gestures not used in sign language.

Neural networks have gained a lot of interest in recent years. Consequently, deep architectures for gesture recognition have appeared. In [21], authors explore convolutional and bidirectional neural network architectures. In experiments, they use the Montalbano[2] dataset, which contains 20 Italian sign gesture categories. In [22], authors modified the Neural Machine Translation (NMT) system using the standard sequence to sequence framework to translate sign language from videos to text. Also in [23], the deep architecture is applied to tasks such as gesture recognition, activity recognition and continuous sign language recognition. The authors employ bidirectional recurrent neural networks, in the form of Long Short Term Memory (LSTM) units.

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

[2]http://chalearnlap.cvc.uab.es/dataset/13/description/

10.21437/IberSPEECH.2018-27

In this work, Convolutional Networks are used to classify gestures from the Spanish sign language. Some improvements in recognition accuracy are obtained with respect to results from our previous studies [8, 9].

3. Experimental setup

3.1. Data preparation

The experimental data is the isolated gestures subset from the "Spanish sign language db"[3], which is described in [8]. All gestures in this database are dynamic. Data was acquired using the Leap Motion sensor. Gestures are described with sequences of 21 variables per hand. This gives a sequence of 42 dimensional vector features per gesture. The isolated gesture subset is formed of samples corresponding to 92 words; for each word, 40 examples performed by 4 different people were acquired, giving a total number of 3,680 acquisitions. The dataset was divided into 4 balanced partitions with the purpose of conducting cross-validation experiments in the same fashion as those described in [9].

Data corresponding to each gesture is saved in a separated text file. Each row in the file contained information (42 feature values) about the gesture in a determined frame. Gestures have different numbers of frames; therefore, a preprocess step is required because we need fixed length data to train a neural network. Three different methods were used to fix the number of rows of each gesture to an equal length of matrix max_length:

- Trim data to fixed length (keeping the last max_length vectors) or pad with 0 (at the beginning) to get the same number of rows.
- Our implementation of the trace segmentation technique [25]
- Linear interpolation of data using the `interp1d` method from the `scikit-learn` library.

3.2. Model

In our experiments we use the LeNet network [26]. LeNet is a convolutional network that has several applications, such as handwriting recognition or speech recognition. The architecture of this network is characterised to ensure invariance to some degree of shift, scale, and distortion.

In our case, the input plane receives the gesture data, and each unit in a layer receives inputs from a set of units located in a small neighbourhood in the previous layer. In Figure 1 the scheme of the LeNet architecture is shown. The first layer is a 2 dimensional convolution layer which learns convolution filters, where each filter is 20×20 units. After that, the ReLU activation function is applied and followed by a max-pooling with kernel size of 2. Once again, a convolution layer is applied, but this time with 20 convolution filters. The ReLU activation function is applied, previously applying max-pooling. A dropout layer with $p = 0.5$ is added and finally, a fully-connected layer is used. The number of input features depends on the max_length parameter, and it is calculated by following Equation (1).

$$\frac{max\_length - 12}{4} \cdot 140 \tag{1}$$

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

Table 1: *Results with trim/zero padding preprocessing. Best results marked with "*".*

Patience \ Max_length	40	50	60	70	80	90	97	100
5	12.0	13.1	12.0	12.5	11.6	12.7	12.6	11.4
10	11.0	11.0	10.1	11.6	10.6	11.8	11.4	11.2
15	10.8	11.2	9.8	11.3	9.8	11.4	11.1	10.3
20	10.5	10.5	10.5	10.9	9.4	11.1	10.8	9.9
25	10.4	10.0	9.6	10.9	9.5	10.5	11.1	10.2
30	11.1	10.6	9.4	10.6	*9.3	11.9	10.2	10.3
35	10.9	10.7	10.0	10.5	9.4	11.0	10.3	10.2
40	10.7	10.4	9.9	10.5	*9.3	10.8	10.3	9.9
45	11.2	10.1	10.4	10.5	9.7	11.0	10.3	10.1
50	11.2	10.1	10.3	10.5	9.6	10.5	10.3	10.2

The number of output features is fixed to 1,000. Finally, there is an output layer with 92 neurons corresponding to the number of classes.

4. Experiments and results

We conducted experiments using our LeNet network. In each experiment we varied the value of patience $p = \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ and $max\_length = \{40, 50, 60, 70, 80, 90, 97, 100\}$. Patience (also called early stopping) is a technique for controlling overfitting in neural networks, by stopping training before the weights have converged. We stop the training when the performance has stopped improving in a determined number of epochs. The original code and LeNet configuration is dedicated to speech recognition, where the maximum length was fixed to 97 by default. That is the reason why this value is used in the experiments. The experiments compare against the baseline provided by [9], which used Hidden Markov Models (HMM) to obtain a classification error rate of 10.6 ± 0.9. Confidence intervals were calculated by using the bootstrapping method with 10,000 repetitions [27] and they are all around the same value for all the experiments (± 0.9). All the results shown in the following are classification error rates obtained through cross-validation using the same four partitions stated in [9].

For each experiment we used the three different techniques of data preprocessing described above. Table 1 shows gesture classification results using the trim/zero padding preprocessing technique. We added a colour scale to make it easier to see the performance of classification depending on patience and max_length values. Dark colours are for bigger error rates and light colours are assigned to lower error rates. We can appreciate that the best score is obtained with a patience of 30 or 40 epochs and max_length of 80 rows (9.3% error rate). In general, the higher the patience, the lower the error, but only until a certain value. With respect to max_length, the behaviour does not present a clear pattern.

In Table 2 we show most frequent confused gestures in gesture classification using trim/zero padding as preprocessing step. The confusions comes from global confusion matrix generated during cross-validation experiments.

In Table 3, trace-segmentation preprocessing results are shown. In this case, the best score is obtained with a patience of 45 epochs and max_length of 100 rows (9.2% of error rate). This result is slightly better than the best score obtained with the trim/zero padding technique. However, according to confidence

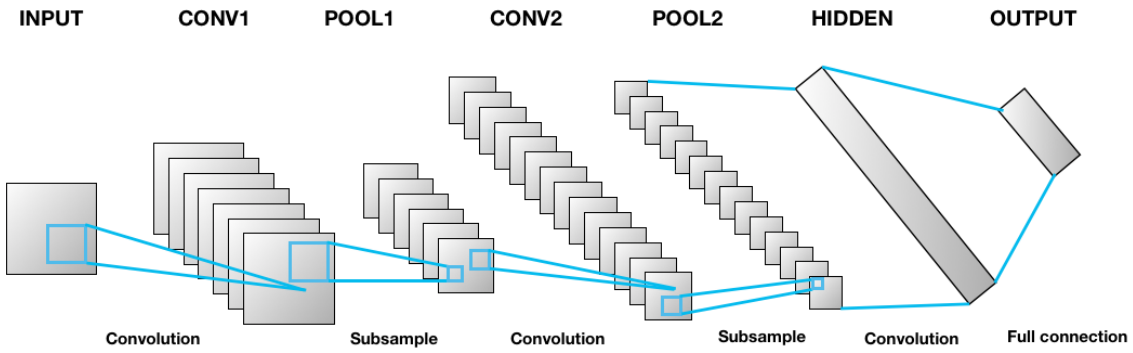

Figure 1: *Schematic diagram of LeNet architecture [24].*

Table 2: *Most frequent confused gestures in gesture classification using trim/zero padding preprocessing.*

Num of confusions	Reference	Recognition
9	sign	a lot
9	thank you	nineteen
8	thirteen	fourteen
8	sign	regular
7	you (male)	we (male)
7	one	eleven
7	red	eyes

Table 3: *Results with trace-segmentation preprocessing. Best result marked with "*".*

	Max_length							
Patience	40	50	60	70	80	90	97	100
5	12.5	12.6	13.2	11.9	11.5	11.1	12.4	11.4
10	11.4	10.8	11.5	10.5	12.6	11.7	10.7	11.1
15	11.2	11.3	11.3	11.3	10.6	10.3	10.9	10.0
20	11.1	9.8	10.7	10.5	10.8	10.2	10.0	11.0
25	10.2	9.7	10.3	10.1	10.7	10.1	9.7	9.7
30	10.7	9.7	9.9	10.4	11.1	10.2	9.7	9.8
35	10.7	9.3	9.9	10.5	10.7	10.3	9.3	9.6
40	10.1	9.7	10.0	10.1	10.0	9.5	9.6	9.7
45	10.0	9.9	10.0	10.7	10.1	10.0	9.4	*9.2
50	10.2	9.6	10.0	10.1	10.1	10.2	9.6	9.8

intervals, neither trim/padding nor trace-segmentation preprocessing provide statistically significant improvements with respect to the best result obtained with HMM. With this technique, the behaviour for patience is similar, but it seems that the higher the max_length, the lower the error.

In Table 4 we show most frequent confused gestures in gesture classification using trace-segmentation as preprocessing step. Some of confused gestures match with confusions from Table 2.

The last experiments were conducted using the interpolation technique as a preprocessing step. This time the best result was obtained with a patience of 25 epochs and 97 for max_length. These are definitely the best results, since the lowest error rate is 8.6%, which is significantly better than the HMM baseline (which did not happen with the other preprocessing techniques). Therefore, the previous result from [9] is improved by about 2% in an absolute manner. This result is considered to be very satisfactory.

Table 4: *Most frequent confused gestures in gesture classification using trace-segmentation preprocessing.*

Num of confusions	Reference	Recognition
10	one	no
10	sign	a lot
10	red	eyes
10	gray	colour
9	no	one
9	thank you	study
8	be born	good morning
8	eighteen	nineteen
7	eyes	red
7	five	hello

In Table 6 we show most frequent confused gestures in gesture classification using interpolation as preprocessing step.

We can appreciate that some of confusions match in each of confusion table e.g. "sign" confused with "a lot" sign. In most of confused signs there are some similarity in shape of hands or trajectory during gesture e.g. sign "hello" and "five" have the same shape of hand only that the sign "hello" make swinging movement. We show image of both signs in Figure 2[4]. Exactly the same thing happens with signs "one" and "no" where the only difference between them is swinging movement.

5. Conclusion and future work

In this work, we developed a classification system for the Spanish sign language. We used the LeNet convolutional network. As a preprocessing step we employed three different methods. The data set used in this work is the "Spanish sign language db" which contains data from 92 different gestures captured with the Leap Motion sensor. The results obtained are very satisfactory, since we were able to improve our baseline by lowering

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

Table 5: *Results with interpolation preprocessing. Best result marked with "*".*

Patience \ Max_length	40	50	60	70	80	90	97	100
5	12.1	11.4	12.3	11.6	12.0	11.3	10.7	12.7
10	11.1	11.4	11.2	10.6	10.9	10.1	9.9	10.6
15	10.3	9.5	11.9	10.0	10.3	10.1	10.1	9.2
20	10.5	9.3	10.3	9.4	11.1	10.0	9.5	9.0
25	9.6	9.8	9.9	9.5	9.8	9.9	*8.6	9.6
30	10.0	10.1	10.3	9.9	9.6	9.7	8.9	9.1
35	9.7	10.0	9.8	9.6	9.8	9.2	8.7	9.0
40	10.0	9.7	10.0	9.6	9.6	9.2	8.9	9.0
45	10.0	9.7	9.9	9.4	9.8	9.5	9.2	9.6
50	10.0	9.7	9.9	9.5	9.6	9.5	9.1	9.0

Table 6: *Most frequent confused gestures in gesture classification using interpolation preprocessing.*

Num of confusions	Reference	Recognition
10	sign	a lot
10	eighteen	nineteen
9	red	eyes
9	he	brother
8	good morning	be born
7	one	no
7	do not know	no
7	hello	five

the error rate by 2% to a level of 8.6%. We also, present tables containing most frequent confused gestures using different methods of preprocessing.

As we obtained such a good result, we would like to continue working in this area. Now we want to work with sign language sentence recognition to be able to create a system that recognises a signed sentence composed of several signs. For that, we will use Recurrent Neural Networks (RNN) commonly used in machine translation and speech recognition. Also, we will conduct experiments on isolated gestures classification to compare the performance of LeNet and RNN.

6. Acknowledgements

Work partially supported by MINECO under grant DI-15-08169, by Sciling under its R+D programme, and by MINECO/FEDER under project CoMUN-HaT (TIN2015-70924-C2-1-R). The authors would like to thank NVIDIA for their donation of Titan Xp GPU that allowed to conduct this research.

7. References

[1] American Speech-Language-Hearing Association, "Guidelines for Fitting and Monitoring FM Systems," https://bit.ly/2udMuOs, 2002.

[2] M. Kaine-Krolak, , and M. E. Novak, "An Introduction to Infrared Technology: Applications in the Home, Classroom, Workplace, and Beyond ..." https://bit.ly/2udMuOs, 1995.

[3] National Institute on Deafness and Other Communication Disorders, "Cochlear Implants," https://bit.ly/27R6aWd, 2016.

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

Figure 2: *Pair of confused signs.*

[4] National Institute on Deafness and Other Communication Disorders, "Hearing Aids," https://bit.ly/1UwxUYN, 2013.

[5] S. Jones, "Alerting Devices," https://bit.ly/2Eri6Y8, 2018.

[6] National Association of the Deaf, "Captioning for Access," https://bit.ly/2NCpLVi.

[7] The Canadian Hearing Society, "Speech to text transcription (CART Communication Access Realtime Translation)," https://bit.ly/2N2QhFV.

[8] Z. Parcheta and C.-D. Martínez-Hinarejos, "Sign language gesture recognition using HMM," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2017, pp. 419–426.

[9] C.-D. Martınez-Hinarejos and Z. Parcheta, "Spanish Sign Language Recognition with Different Topology Hidden Markov Models," *Proc. Interspeech 2017*, pp. 3349–3353, 2017.

[10] V. E. Kosmidou, P. C. Petrantonakis, and L. J. Hadjileontiadis, "Enhanced sign language recognition using weighted intrinsic-mode entropy and signer's level of deafness," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 6, pp. 1531–1543, 2011.

[11] "MYO armband," https://www.myo.com/.

[12] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[13] L. E. Potter, J. Araullo, and L. Carter, "The leap motion controller: a view on sign language," in *Proceedings of the 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration*. ACM, 2013, pp. 175–178.

[14] D. Kim, O. Hilliges, S. Izadi, A. D. Butler, J. Chen, I. Oikonomidis, and P. Olivier, "Digits: freehand 3d interactions anywhere using a wrist-worn gloveless sensor," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 2012, pp. 167–176.

[15] T. Kuroda, Y. Tabata, A. Goto, H. Ikuta, M. Murakami *et al.*, "Consumer price data-glove for sign language recognition," in *Proc. of 5th Intl Conf. Disability, Virtual Reality Assoc. Tech., Oxford, UK*, 2004, pp. 253–258.

[16] J. D. Guerrero-Balaguera and W. J. Pérez-Holguín, "Fpga-based translation system from colombian sign language to text," *Dyna*, vol. 82, no. 189, pp. 172–181, 2015.

[17] 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.

[18] 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.

[19] Y. Zou, J. Xiao, J. Han, K. Wu, Y. Li, and L. M. Ni, "Grfid: A device-free rfid-based gesture recognition system," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 381–393, 2017.

[20] H. J. Escalante, I. Guyon, V. Athitsos, P. Jangyodsuk, and J. Wan, "Principal motion components for one-shot gesture recognition," *Pattern Analysis and Applications*, vol. 20, no. 1, pp. 167–182, 2017.

[21] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 430–439, 2018.

[22] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," *CVPR 2018 Proceedings*, 2018.

[23] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] H. Dwyer, "Deep Learning With Dataiku Data Science Studio," https://bit.ly/2mfx8Wd.

[25] E. F. Cabral and G. D. Tattersall, "Trace-segmentation of isolated utterances for speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 1995, pp. 365–368 vol.1.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

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