Arab Sign language Recognition with Convolutional Neural Networks

Conference Paper · July 2019

DOI: 10.1109/ICCSRE.2019.8807586

4 authors, including:

Salma Hayani
University Ibn Zohr - Agadir
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Mohamed Benaddy
University Ibn Zohr - Agadir
22 PUBLICATIONS **35** CITATIONS
SEE PROFILE

O.E. Meslouhi
University Ibn Zohr - Agadir
15 PUBLICATIONS **54** CITATIONS
SEE PROFILE

Arab Sign language Recognition with Convolutional Neural Networks

Salma Hayani
LabSIE Laboratory
Polydisciplinary Faculty of Ouarzazate,
Ibn Zohr University
Ouarzazate, Morocco
salma.hayani@edu.uiz.ac.ma

Mohamed Benaddy
LabSIE Laboratory
Polydisciplinary Faculty of Ouarzazate,
Ibn Zohr University
Ouarzazate, Morocco
m.benaddy@uiz.ac.ma

Othmane El Meslouhi
LabSIE Laboratory
Polydisciplinary Faculty of Ouarzazate,
Ibn Zohr University
Ouarzazate, Morocco
o.elmeslouhi @uiz.ac.ma

Mustapha Kardouchi
IT departement
Faculty of science, Moncton University
Moncton, Canada
mustapha.kardouchi@umoncton.ca

***Abstract*— The implementation of an automatic recognition system for Arab sign language (ArSL) has a major social and humanitarian impact. With the growth of the deaf-dump community, such a system will help in integrating those people and enjoy a normal life. Like other languages, Arab sign language has many details and diverse characteristics that need a powerful tool to treat it. In this work, we propose a new system based on the convolutional neural networks, fed with a real dataset, this system will recognize automatically numbers and letters of Arab sign language. To validate our system, we have done a comparative study that shows the effectiveness and robustness of our proposed method compared to traditional approaches based on k-nearest neighbors (KNN) and support vector machines (SVM).**

***Keywords*— Arab sign language, convolutional neural networks, automatic recognition system.**

I. Introduction

Sign language (SL) is a natural, visual, and non-verbal language. However, both sign and spoken languages have the same functions. SLs are used by deaf, hard-hearing and dumb people. This allows them to communicate with others through signs and gestures.

In fact, SLs are not a novel innovation; they exist similarly as the verbally expressed dialects. The sign languages are developed naturally in each region, like other verbal dialects as independent communication systems. Their emergence cannot be attributed to anyone. Hence, with the growing of deaf community setting up automatic systems becomes a necessity.

The main components that compose sign languages are manual and non-manual signs [1]. The manual signs are: hand position, orientation, shape, and trajectory. The non-manual represents body motion and facial expressions. However, most researchers focus on manual signs because it contains the main information [2][3], which non-manual signs allow for signers to clarify and emphasize the meaning of manual signs. Our work will focus on the manual part.

The two main methodologies for Sign Language Recognition (SLR) are vision-based and sensor-based techniques. Vision-based SLR uses cameras just to catch motions (signs). It has the upside of ease of use since they don't require the signer to wear any gadgets such as information gloves or movement trackers. The computational cost is high for this method. In addition, the method's biggest challenges are the varieties out of sight and the luminosity change conditions. The sensor-based method uses wearable gadgets to capture the signs. Although it probably won't be as helpful as the vision-based, it accompanies immense improvement in acknowledgment exactness. Gloves and movement trackers are the most famous wearable gadgets for SLR.

In the present paper, we aim to build an Arab sign language recognition system that automatically recognizes 28 letters and numbers from 0 to 10 using a CNN model feed with RGB images.

We organize the rest of this paper on five sections: section two introduces the related works achieved in this field. Section three summarizes some CNN models. Section four exposes the proposed model. Section five discusses the experimental results and section six presents a general conclusion and future work.

II. Related work

In the following section, we present a survey of some previously used approaches in Arab sign language recognition.

In [4], authors have introduced an ArSL system based on Adaptive Neuro-Fuzzy Inference System (ANFIS), this system uses data collected using color gloves. With the same database and feature extraction techniques, Assaleh et al. [5] set up a recognition system with a polynomial classifier and they obtained improved results compared with those got by ANFIS approach.

In the same context, Al-Jarrah and Halawani [6] implemented an automatic system that translates 30 letters of Arab sign language alphabets based on neuro-fuzzy systems. In contrast with previously cited works, the system doesn't use any devices to collect data, and experiments have achieved an accuracy of 93.55%.

978-1-7281-0826-1/19/$31.00 ©2019 IEEE

Shanableh et al. [7], use various spatiotemporal feature extraction methods to recognize isolated ArSL gestures in the offline and online modes. The proposed system achieves the classification with the simplest method K-nearest neighbor (KNN) and it proved its performance with an accuracy of 97%.

An automatic Arab sign language (ArSL) recognition system based on Hidden Markov models (HMMs) has been established by AL-Rousan et la. in [8]. They use this model in different modes online and offline, signer-dependent and signer-independent. In the experiments, the signer-dependent case, the obtained accuracy for the offline mode is 96.74% and 93.8% for the online mode. In the signer-independent case, the system gives a precision of 94.2% and 90.6% in the offline and the online modes, respectively.

In [9], Maraqa et al. propose the use of a system based on feed-forward neural networks and recurrent neural networks with its different architectures; partially and fully recurrent networks. The obtained results show that the recommended system with the fully recurrent architecture has an exhibition with an accuracy of 95% for static signs recognition.

Alzohairi et al. in [10], introduce a recognition system of alphabets. This system extracts the HOG descriptor and transfer features to One Versus All soft-margin (SVM). The proposed system has reached an accuracy of 63.5% for Arab Alphabet signs.

III. Convolutional neural networks

Convolutional neural network (CNN, or ConvNet) is a class of deep neural networks used in the examination of visual images. The viability of convolutional nets in image recognition is one of the primary reasons researchers have woken up to the adequacy of deep learning. They are controlling significant advances in computer vision (CV), which has obvious applications for self-driving vehicles, robotics, drones, security, medical technology, and treatments for the visually impaired.

Convolutional neural networks use an architecture that is particularly well suited to image classification. Using this architecture makes neural networks quick to learn. This helps us to create powerful deep multi-layer networks for classifying images. CNN still uses the Backpropagation and its derivatives training methods to learn from data. Modern implementations take advantage of specialized GPUs to further improve performance.

.

There are several architectures based on convolutional networks. In this paragraph, we go over some of the most powerful Convolutional Neural Networks which laid the foundation of today's Computer Vision achievements, achieved using Deep Learning.

- **LeNet-5** (1998) is a 7 layers Convolutional Neural Network [11], deployed in many banking systems to recognize handwritten numbers on cheques scanned in 32 x 32pixels grayscale format. The ability to process higher resolution images requires larger and more convolutional layers.

- **AlexNet (2012)** [12] is very similar to LeNet but is much deeper and has around 60 million parameters. This architecture contains 5 convolutional layers and 3 fully connected layers. These 8 layers, combined with two concepts MaxPooling and ReLU activation, give an edge to the architecture..

- **VGGNet (2014):** Because of the simplicity of its uniform architecture [13], it appeals to a new-comer as a simpler form of a deep convolutional neural network. VGGNet has two simple rules:
    1. Each Convolutional layer has configuration windows size = 3×3, stride = 1×1, padding = same. The only thing that differs is the number of filters.
    2. Each Max Pooling layer has configuration windows size = 2×2 and stride = 2×2. Thus, we half the size of the image at every Pooling layer.

- **GoogleNet (2014)**: This network has a big similarity with LeNet, but it implements a new element called the creation module [14]. It uses batch normalization, image distortions and RMSprop optimizition algorithm. This model uses many tiny convolutions to reduce the number of parameters.

- **ResNet (2015)**: The Residual Neural Network (ResNet) [15] introduces a new architecture with "skip connections" and features heavy batch normalization. Such skip connections called gated units or gated recurrent units and have a solid closeness to recent successful units applied in RNNs. This approach is capable to train a NN with 152 layers while still having lower complexity than VGGNet..

IV. Proposed model

This section describes the proposed model. Figure 1 gives the used CNN architecture which is inspired from LeNet-5. It is composed of seven adjacent layers.

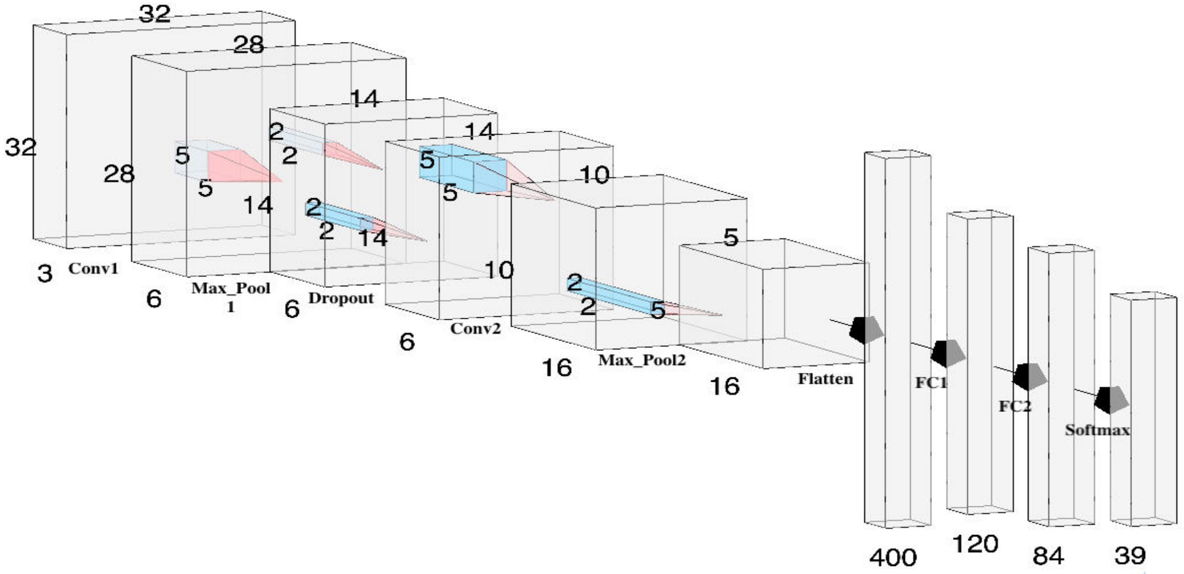

Figure 1: Proposed system architecture

The first four layers extract deep features from images and the three last ones classify them.

The layer “Conv1” is a convolution layer with 6 feature maps of 28 × 28 pixels. Every neuron does a convolution of the kernel size 5× 5 × 3 and adds the shared bias. We use a leaky rectified linear unit (Leaky-ReLU) as an activation function, as shown in Eq.1, which allows a small positive gradient when the neuron is not active.

$$f(x) = \begin{cases} x & if\ x > 0 \\ \alpha x & otherwise \end{cases} \qquad \textit{Eq. 1}$$

Layer “Max_Pool1” is responsible of grouping the results of the Conv1 layer with a kernel of 2 × 2 for which we add a dropout layer with a probability of 0.75 for regularization.

In layer “Conv3” we use 5×5 kernel with 16 feature maps of size 10 ×10 then we apply a max-polling with 2 × 2 to get finally 16 feature maps of 5×5 then we flatten the maps to get 400 neurons.

The three remaining layers represent the classification stage. The first one (FC1) contains 120 Leaky-Relu neurons that are fully connected to the flatten layer. In the second layer (FC2) 84 Leaky-Relu neurons are fully connected to the previous layer. The last layer (softmax) is composed of 39 neurons to give the class of the input image. In this layer, we use the softmax activation function. We train the system using AdamOptimizer with a learning rate of 0.03.

V. Experement and results

A. *Datasets:*

To evaluate our system, we use a dataset of images containing 2030 images of numbers (from 0 to 10) and 5839 images of 28 letters of Arab sign language, i.e. 7869 RGB color images with 256 × 256 pixels (**Figure 2**). These images are collected by a group of Polydisciplinary Faculty of Ouarzazate students, Ibn Zohr University, Agadir. These images are taken by a professional Canon® camera from different signers and different luminosity intensities. We notice that we resize all images to 32×32 pixels before feeding our recognition system.

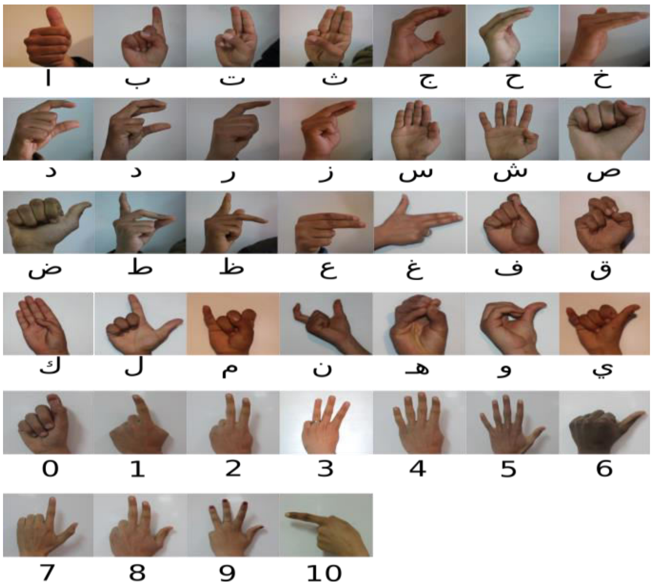

Figure 2: Samples of ArSL numbers and letters from our Database

B. *Results and discussions*

To improve the proposed system, we have done various experiments by varying the number of training and test sets. First, we split the database into training and testing sets. Second, we change the percentage of images in the training/validation phases between [50%, 80%] and [20%,50%] to measure the performance of our proposed model. Finaly, we save the obtained accuracy in every using these sets.

Table 1 shows the accuracy of the different training set sizes. As we can observe, the accuracy achieves the best value of 90.02% when we train our system with 80% of images from the database.

Table 1: test accuracy according to percentage of training sets.

Training images	Recognition rate
50%	83.27%
60%	85.64%
70%	88.47%
80%	90.02%

In Figures 3 and 4 we can visualize the plot of the accuracy and error rate, change respectively, depending on epochs with 80% of training data. As we can see, the accuracy continues to increase and the error rate decrease during the training and test phases and we observe no sign of the over fitting.

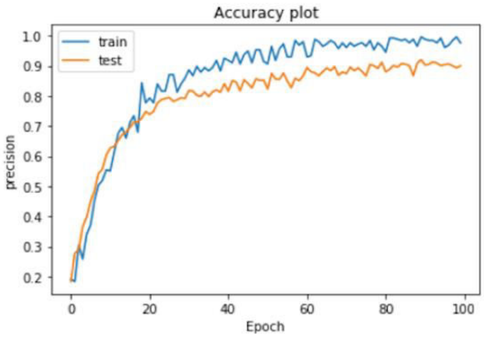

Figure 3: Accuracy plot using 80% of training data

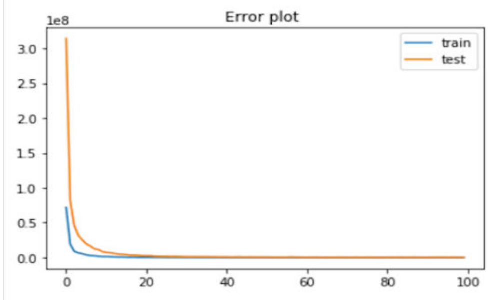

Figure 4: Error plot using 80% of training data

To show the performance of the proposed system, we compare its obtained results with KNN (k-nearest neighbor) with Euclidean distance and SVM (support vector machines) with different kernels classifiers commonly used in this field.

The Table 2 shows the performance obtained by our proposed method and the other classifiers using 80% of the data for learning. As we can remark, the proposed system achieves the best accuracy despite the diversity of the used classifiers.

Table 2*:* comparison of the proposed system with different classifiers

Classifier	Recognition rate
KNN[16]	66%
SVM with lineair kernel[17]	84%
SVM with polynomial kernel[18]	85%
SVM with RBF kernel[19]	88%
Prposed system	90.02%

VI. Conclusion

In summary, the present paper introduces an offline recognition system of Arab sign numbers and letters based on deep convolutional neural networks. The suggested method is a vision-based approach, and the system learns directly from the original data. The evaluation of our system shows its efficiency in recognizing both numbers and letters; it gives interesting performances compared with other existing systems based on the KNN and SMV.

As a future work, we expect to extend our system for video-based recognition to recognize sentences in real time.

VII. References

[1] A. Subhash Chand, A. S. Jalal and R. Kumar Tripathi, «A survey on manual and non-manual sign language recognition for isolated and continuous sign,» *Applied Pattern Recognition,* 2016.

[2] H. Cooper, B. Holt and R. Bowden, «Sign language recognition,» *Springer,* 2011.

[3] S. Ong and S. Ranganath, chez *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[4] M. Al-Rousan and M. Hussain, «Automatic recognition of Arabic sign language finger spelling,» *International Journal of Computers and Their Applications,* 2001.

[5] K. Assaleh and M. Al-Rousan, «Recognition of Arabic sign language alphabet using polynomial classifiers,» *EURASIP Journal on Applied Signal Processing,* 2005.

[6] O. Al-Jarrah and A. Halawani, «Recognition of gestures in Arabic sign language using neuro-fuzzy systems,» *Atificial Intelligence,* 2001.

[7] T. Shanableh, K. Assaleh and M. Al-Rousan, «Spatio-temporal feature-extraction techniques for isolated gesture recognition in Arabic sign language,» *IEEE Transactions on Systems, Man, and Cybernetics Part B (Cybernetics),* 2007.

[8] M. AL-Rousan, K. Assaleh and A. Tala'a, «Video-based signer-independent Arabic sign language recognition using hidden Markov models,» *ELSEVIER,* 2009.

[9] M. Maraqa, F. Al-Zboun , M. Dhyabat and. R. Abu Zitar, «Recognition of Arabic Sign Language (ArSL) Using Recurrent Neural Networks,» *Intelligent Learning Systems and Applications,* 2012.

[10] R. Alzohairi, R.Alghonaim, W.Alshehri, S.Aloqeely, M.Alzaidan and O.Bchir,«Image based Arabic Sign Language Recognition System,» *International Journal of Advanced Computer Science and Applications,* 2018.

[11] Y. LeCun, L. Bottoux, Y. Bengio and P. Haffner, «Gradient-Based Learning Applied to Document Recognition,» *IEEE,* November 1998.

[12] A. Krizhevsky, I. Sutskever and G. Hintton, «Imagenet classification with deep convolutional neural networks,» chez *Proceedings of the 25th international conference on neural information processing systems*, 2012.

[13] K. Simonyan and A. Zisserman, «Very deep convolutional networks for large-scale image recognition,» *arXiv preprint,* 2014.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinov, C. Hill and A. Arbor, «Going Deeper with Convolutions,» *IEEE Xplore,* 2015.

[15] K. He, X. Zhang, S. Ren and J. Sun, «Deep Residual Learning for Image Recognition,» 2015.

[16] D. Coomans and D. Massart, «Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearestneighbour classification by using alternative voting rules ,» *AnalyticaChimicaActa 136, 15-27,* 1982.

[17] C. Williams and M. Seeger, «Using the Nyström method to speed up kernel machines ,» 2001.

[18] Y-W. Chang, C.-J. Hsieh, K-W. Chang, M. Ringgaard and C-J. Lin, «Training and testing low-degree polynomial data mappings via linear SVM ,» *Journal of Machine Learning Research,* april 2010.

[19] J-P. Vert, K. Tsuda and B. Schölkopf, «A primer on kernel methods ,» *Kernelmethods in computationalbiology, 47, 35-70,* 2004.