Humanized Computing and Communication
Vol. 1, No. 1 (2020) 23-42

DOI: 10.35708/HCC1868-148363

AI at the Edge for Sign Language Learning Support

Pietro Battistoni[1], Marianna Di Gregorio[1], Marco Romano[1], Monica Sebillo[1], and Giuliana Vitiello[1]

University of Salerno, Salerno, Italy
{pbattistoni, madigregorio, msebillo, gvitiello, marromano}@unisa.it

Received (11/12/2019)
Revised (02/15/2020
Accepted (03/16/2020)

Abstract. In the field of multimodal communication, sign language is and continues to be, one of the most understudied areas. Thanks to the recent advances in the field of deep learning, there are far-reaching implications and applications that neural networks can have for sign language mastering. This paper describes a method for ASL alphabet recognition using Convolutional Neural Networks (CNN), which allows to monitor user's learning progress. American Sign Language (ASL) alphabet recognition by computer vision is a challenging task due to the complexity in ASL signs, high interclass similarities, large intraclass variations, and constant occlusions. We produced a robust model that classifies letters correctly in a majority of cases. The experimental results encouraged us to investigate the adoption of AI techniques to support learning of a sign language, as a natural language with its own syntax and lexicon. The challenge was to deliver a mobile sign language training solution that users may adopt during their everyday life. To satisfy the indispensable additional computational resources to the locally connected end- user devices, we propose the adoption of a Fog-Computing Architecture.

Keywords: Fog-Computing; Neural Network; Sequence Learning; Sign Language.

1 Introduction

The United Nations Convention on the Rights of Persons with Disabilities recognizes and promotes the use of sign languages, establishing that sign languages are equal in status to spoken languages [30]. It recommends states parties to facilitate the learning of sign language and promote the linguistic identity of the deaf community. One of the major challenges is to raise awareness of the importance of providing sign language learning support. Most deaf children are

born to hearing parents and are not exposed to sign language until school age, missing a vital window of time for language acquisition, which corresponds to the first three years of life. As a result, damage due to acoustic sensory deprivation may occur and the associated effects will not only affect the learning process of the word but will also negatively affect the global perceptual mechanism and consequently the behavior of the subject [12]. Early access to sign language and services in sign language, including quality education available in sign language, is recognized to be vital to the growth and development of the deaf individual and critical to the achievement of personal goals. While special education promotes the integration of children with and without disabilities as the least restrictive environment (LRE), Deaf cultural perspective holds that a language-rich environment is best achieved through sign language. Therefore, an LRE for deaf children involves access to information through sign language and interaction with peers. The goal of our research has been to provide IT support to the cumbersome process of sign language learning, relying on the use of deep learning techniques. In the first part of the paper, we present a recognition system that uses Convolutional Neural Networks (CNN), which was developed at the HCI- UsE laboratory of the University of Salerno to provide sign language learners with an advantageous interactive experience. Demonstrating through the first experimental phase that our system elicits more positive feelings than the traditional video-based learning technique. Several studies confirm that there has been increasing recognition of the influence of emotion in human computer interactions and the use of the software by users [10] [11] [23]. The main objectives of subsequent experiments we conducted on the system were the study of user's ability to efficiently and effectively reproduce the letters submitted to him/her. The results obtained from the tests carried out were positive about user learning. Most participants developed, during the various experimental sessions, a higher level of familiarity with the sign language, leading to an increase in the accuracy detected by the neural network. The system has proven to be a valid teaching tool for children or adults with no experience with this language, who may start learning the sign language alphabet. A natural step beyond has been to conceive a similar system that would deal with the complexities of the whole language of signs, as described in the second part of the paper. Sign languages are natural languages with their own grammar and lexicon. This raised some new challenges when conceiving the new system, not only because of the complexities of these signs, the high interclass similarities, the large interclass variation, and constant finger occlusion, but also for the computational resources required and the low-latency requirements to achieve answers in near real-time. In a previous research we experienced the high acceptance gained by mobile applications within the deaf community as valid communication means [9]. Therefore, we also aimed at delivering a deep learning solution on mobile smart devices to facilitate sign language training activities during everyday life. This gave rise to the idea of a *Fog-Computing* Architecture, which could provide additional computational resources to the locally connected end- user devices. The paper is structured as follows. Some related work is presented in Section II. Section

III presents the system we realized to recognize the American Sign Language (ASL) alphabet, Section IV describes a comparative studies between the traditional video-based self-learning and the one based on our interactive system. and Section V describes the experiment performed and summarizes the results. Section V proposes the *fog-computing* architecture underlying the complete ASL learning support. Section VI concludes the paper.

2 Related Work

In general, the ASL alphabet recognition task is formulated as two subtasks: feature extraction and subsequent classification. Researchers have been using different methods to extract discriminative features and create powerful classifiers. In [27], Pugeault and Bowden apply Gabor filters to extract features from both color and depth images at 4 different scales. Then a multiclass random forest classifier is used to recognize the 24 static ASL alphabet signs. They report a 49% recognition rate in the leave-one-out experiment. Half of the signs could not be recognized, showing that Gabor filters cannot capture enough discriminative information for differentiating different signs. Also, Wang et al. used color and depth images for recognition [31]. They proposed a Superpixel Earth Mover's Distance (SP- EMD) metric, and they reported a 75.8% recognition rate on the benchmark dataset. In [19] using a Support Vector Machine (SVM) classifier, Maqueda et al. gained 83.7% leave-one-out accuracy on the benchmark dataset. In [22] features were extracted from only depth images on randomly positioned line segments and a random forest was used for classification, with 81.1% accuracy. Some studies attempted to exploit the 3D information embedded in the depth images (3D approach) [15] [32] [34] [33] [28]. Such 3D approaches are promising to achieve better performance than image representations due to the extra dimension. However, the 3D point cloud obtained from the depth image is sparse at the regions with large gradients and absent at the occluded areas, which affects the overall performance. Due to the articulated structure of hands, some studies implemented a hand part segmentation step before the gesture recognition (bottom-up approach). In [13], Keskin et al. extracted depth comparison features from depth images following the method proposed in [29] and fed them into a per-pixel random forest classifier. They reported their leave-one-out recognition rate as 84.3% on the benchmark dataset. This classifier was trained using synthetic depth images which have the parts' ground truth of a hand. To generate more realistic training data, a colored latex glove was employed by Dong et al., resulting in a 70% recognition rate on the benchmark dataset [6]. One of the major drawbacks for those bottom-up approaches is that the sign recognition performance is highly dependent upon the result of the hand part segmentation, and it is challenging to improve the performance of the hand part segmentation because of the high complexities and constant occlusions. Recently, deep learning methods, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated their extraordinary performance in various classification and recognition tasks.

2.1 From Pose to Gesture

The Sign language (SL), seen as a crucial language for hearing and speech impaired people, can be considered as the most grammatically structured gestural communication [24]. A Gesture can be seen as a continuous sequence of poses. In [20] Marcon et al. use Hidden Markov Models (HMMs), trained on different gestures, to identify a set of key postures and to classify their sequences over a set of possible actions. Cui et al. in [7] proposed a CNN with temporal convolution and pooling for spatiotemporal representation learning from a video, and an RNN with a bidirectional Long Short Term Memory (LSTM) for the mapping of feature sequences to sequences of annotations. Madhuri et al. in [18], present a real-time vision-based system for recognizing finger spelling continuous Sign Language (SL) using a single camera to track user's unadorned hands. The goal is to help hearing or speech impaired people to communicate with people who do not know SL. Although facial expressions add relevant information to the emotional aspect of the sign, in [18], they are not considered since their analysis complicates the problem. They only focus on translating a one-handed sign of representation of alphabets (A-Z) and numbers (0-9). Some researchers have proposed techniques to detect predefined signs from a continuous video stream (sign spotting), while others have handled the classification of isolated gestures into the correct category. Continuous SL recognition, instead, deals with transcribing videos of SL sentences into ordered sequences of annotations, possibly in real-time. Cui et al. in a recent work [8], focus on continuous SL recognition on videos, where learning the spatiotemporal representations as well as their temporal matching for the labels is crucial. They claim their framework based on recurrent convolutional neural networks shows a superior capability of learning temporal dependencies compared to HMMs. Most related researches are using Markov models (HMM) or LSTM networks. Although both those networks take into account the spatiotemporal aspect of Gesture, they face different issues. The HMM solution has a problem with a more prolonged gesture, while the LSTM requires powerful computational resources, and it is slow on a small device like a smartphone. Panzner e Cimiano in [25] compare a purely generative model based on Hidden Markov Models to a discriminatively trained recurrent LSTM network in terms of their properties and their suitability to learn and represent models of actions. They highlight the limitation of temporal context with HMM and the need for an extraordinary computational resource with LSTM networks.

3 Learning ASL Alphabet relying on CNN Support

The proposed system has as its main objective to be a bridge of communication between deaf people and today's society. The system offers a chance to those who are inexperienced in this language, a simple and interactive learning method. Our classification of letters is carried out using a convolutional neural network (CNN or ConvNet). CNNs are machine learning algorithms that have been incredibly successful in managing a variety of activities related to video and image processing. Our network, after being trained, allows the recognition of gestures, through the use of images made by the user. During the execution

of the application, the user will have to take photos containing one of the 24 static letters of the American alphabet. After this, the classification of the latter will be carried out, i.e. the neural network will define the letter corresponding to the gesture made by the user. Finally, the number of correctly executed letters and the number of incorrect ones will be shown to the user, with the consequent level of learning developed during the use of the system.

3.1 Methods Used

1. *Transfer Learning*: is a machine learning technique where models are trained on (usually) larger data sets and refactored to fit more specific or niche data. This is done by recycling a portion of the weights from the pre-trained model and reinitializing or otherwise altering weights at shallower layers. The most basic example of this would be a fully trained network whose final classification layer weights have been reinitialized to be able to classify some new set of data. The primary benefits of such a technique are its less demanding time and data requirements. However, the challenge in transfer learning stems from the differences between the original data used to train and the new data being classified. Larger differences in those data sets often require re-initializing or increasing learning rates for deeper layers in the net.

2. *AlexNet:* We employed Matlab in order to develop, test, and run our CNNs. Specifically, we used AlexNet. AlexNet is a Convolutional network that has had a great impact in the field of machine learning, designed by Alex Krizhevsky. The network acquires an image as input and generates a label for the object in the image along with the probabilities for each category of objects. AlexNet consists of 8 layers: the first 5 are Convolutional layers, and 3 layers are fully connected and can classify images into 1000 categories of objects, such as keyboard, mouse, pencil and many animals.

3.2 The System

In order to obtain images of the user signing in real-time, we created a desk application that is able to access a native camera. Image capture rate was a massive problem we struggled with. Our desk application sends images to our net one by one. Each time, the net classifies the image and presents probabilities for each letter. Our system, as we can see in figure 1, presents on the left side the GIF of the letter to be reproduced by the user, and the consequent reproduction of the same. On the right, instead, we see the image acquired by the camera. When the user feels confident about the sign being made, he/she can take a photo. The letters are randomly generated (from 1 to 24) by the system. Once the photo is taken, it will be shown on the left side of the page. Moreover, it will be possible to change the photo just taken. Moreover, if the user is not able to reproduce the requested letter, he/she may decide to change it.

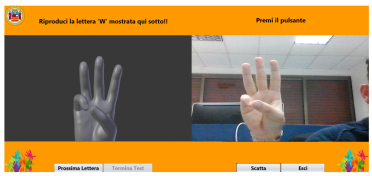

Fig. 1. Core Page of the system.

Furthermore, our system presents another page containing the results of the tests performed by the user. In figure 2 we can see a table divided into three columns: in the first one there are the letters that the user made during the test; in the second column the relative results are present for each letter, and that is how the neural network has classified the images sent to it; in the last column we can see a graphical response of the single test carried out. Marking the wrong result with a red sticker and a positive response with a green one. Finally, at the bottom of the page, we find the level of learning detected during the test. This result comes from the classification of all images made by the user during the experiment. The level of learning coming from the network is not based on an average of the correct or wrong letters, but the neural network produces this value based on the skill and precision that the user has used in reproducing the letters. A high percentage of learning assumes, therefore, that the user had better precision than the other users.

Fig. 2. Result page of the system.

3.3 Dataset and Features

ASL Alphabet was the most incisive DataSet in the training of our network. It contains about 87,000 images of various kinds and depths. The dataset is divided into two directories. The second dataset we used within the system, Sign Language and Static gesture recognition, was less predominant for training. It contains a very limited number of data and, consequently, it is inefficient for training an effective network. The total number of data present is 1,687. With the use of those datasets, we have created a new dataset including all the data, about 73,488 images, and we have decomposed it to develop two happy directories: the first including 80% of the images, to perform training, the second, including the remaining 20%, for tests. The network training was carried out on 58,944 images of various kinds. The images have been reduced in size to be consistent with the specifications of the network used. Each image was resized to 227x227 pixels.

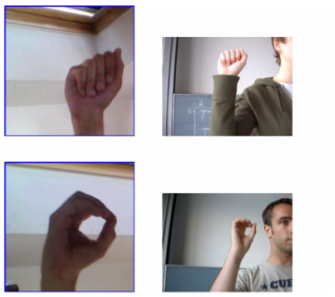

Fig. 3. Dataset example. Left: ASL Alphabet. Righ: Sign Language and Static gesture recognition.

3.4 Train and Test

In order to compare our results with state of the art, we considered two metrics, *accuracy* and *confusion matrix*.

1. *Accuracy and loss*: Accuracy in the validation set is the most popular criterion in the literature, defined as the percentage of correctly classified examples. Splitting the dataset into two for training and validation, we achieved a high validation accuracy, namely 99.96% on the alphabet gestures. The running time of the network train was 491 minutes and 24 seconds, performing 13,800 iterations. The network training has led, therefore, to the development of an efficient network, with very high accuracy in classifying the images, so that we can build a reliable sign recognition system.

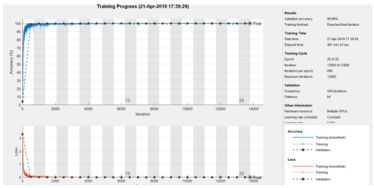

Fig. 4. Train Convolutional Neural Network.

2. *Confusion Matrix*: Additionally, we used a confusion matrix, which is a specific table layout that allows visualization of the performance of the classification model by class. This allowed us to evaluate which letters are the most misclassified and draw insights for future improvement.

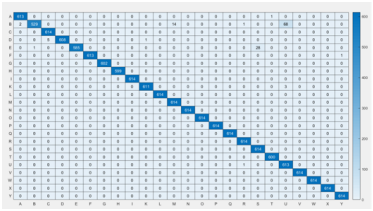

Fig. 5. Test Convolution Neural Network – Confusion Matrix.

4 Comparative Study based on Emotions

The goal of this experiment is to study the different effects on the emotions between the traditional video-based self-learning and the one based on an interactive CNN system. Indeed, when users study videos cannot interact with other students or teachers and this, in general, can be less motivating respect to other forms of learning. The literature suggests that maximizing user's positive emotions and minimizing their negative emotions leads to greater learning and productivity, and less frustration for the user. Then, the objectives of this experiment are both measuring the difference in terms of learning efficacy and in terms of learning emotions and feelings. For this, we designed a between-subjects experiment to use as a pilot for further investigation. The experiment involved 10 participants who were asked to do one personal session of sign language self-learning in a controlled environment. The participants were divided into two groups and were asked to practice with sign language and then test their knowledge through an assessment; one group using the CNN application and one using traditional learning videos. During the experiment, we collected the assessment results of each participants and evidence of their emotions and feelings using a questionnaire and a facial expression analysis system.

4.1 Independent and dependent variables

The independent variable in the experiment is the application used to learn the sign language which the subjects use before performing a language assessment. More precisely, the prototype based on CNN and a traditional learning video. The dependent variables are the learning efficacy measured through the assessment and the users' emotions and feelings measured during the learning task and the assessment. Regarding the learning video, it shows the animated sign besides the corresponding letter and shows the same animations used in the CNN software. In this way, the two different learning approaches offer exactly the same content.

4.2 Between-subjects Design

In our experiment, we decided to adopt a between-subjects design with the aim to limit the effects of the bias. We picked 10 test subjects selecting them taking into account the following aspects: they are technologically proactive, with at least the basic technical skills, and they have no sign language previous knowledge.

The participants are divided into two groups: Gr1 and Gr2. Gr1 will use the CNN application and Gr2 the video. Additionally, to limit bias due to users' variability, participants were randomly assigned to each group.

4.3 Environment

The experiment took place in a controlled environment, properly the HCI-UsE Lab of the University of Salerno, with the aim to have all the factors held constant and controlled except for the independent variable.

4.4 Assessment metrics

Formal usability tests in a lab setting are an excellent method to evaluate whether users can complete tasks; however, the techniques employed have limited effectiveness for measuring the user's emotional experience and desirability of the use of the product. One standard method used to evaluate these intangible aspects is a questionnaire with Likert scales. To measure the users' emotions and feelings we selected the questionnaire presented in [26] that is designed to measure emotions in learning, and we tailored it to fit our experiment necessities. Users were not given ratings on predetermined scales but rather created their own scales with an opportunity to explain their answers. The questionnaire is presented in table 1. Questions were answered using a 5-point Likert scale with 1 for strongly disagree, 5 for strongly agree and 3 for a neutral response. According to the literature, a 5-point Likert scale increases the response rate and quality and it also reduces respondents' "frustration level" [5] [2]. The questionnaire is formed of three parts. The first part must be filled out before starting the learning session and it is about uses' emotions and feelings towards their capabilities and study. The second part is related to the learning session and the last part is about the users' feelings after an assessment. Then table 1 is divided into three sections: "Before the learning session", "During the learning session", "After the evaluation session".

Questionnaires are a useful means to go through aspects indicated by researchers, but sometimes there are aspects that can elude. For this reason, it was decided to complete the study of the emotions adding a facial expressions analysis. Analysing facial expressions allows rapid, quantifiable insights into expressed facial emotions. Facial expression analysis has been carried out manually for decades, now this can be carried out in real-time, helping you understand the facial emotions elicited by stimuli as fast as they are generated, through software. As measurement metrics for the evaluation we used summary scores of engagement and valence are provided, as they provide an overview of emotion.

Table 1. Comparison of ellipse fitting models.

Before the learning session		
1b	Anxiety	Thinking about this activity makes me feel uneasy (b)
2b	Hopelessness	I feel hopeless when I think about studying
3b	Hopelessness	I feel hopeless
4b	Hope	I am confident when I go to start the experiment
5b	Hope	I have an optimistic view toward studying
6b	Hope	I have great hope that my abilities will be sufficient to learn sign language
During the learning session		
1d	Shame	I feel ashamed that I can't absorb the simplest of details
2d	Shame	I get embarrassed
3d	Boredom	I get bored
4d	Boredom	The material bores me to death
5d	Pride	I'm proud of my capacity
6d	Anger	Studying a gesture makes me irritated
7d	Anxiety	I get tense and nervous while learning a gesture
8d	Enjoyment	For me the test is a challenge that is enjoyable
9d	Enjoyment	I enjoyed learning
10d	Enjoyment	I enjoyed acquiring new knowledge
11d	Hopelessness	I have lost all hope that I have the ability to do well on the test
After the assessment		
1a	Pride	I am proud of myself
2a	Pride	I'm proud of how well I mastered the test
3a	Anger	I am angry
4a	Anger	I am fairly annoyed
5a	Relief	I feel very relieved
6a	Shame	I feel ashamed

4.5 Tasks execution

Task.1 - Depending on the group belonging, the participants were asked to execute a learning session using the CNN software or the videos on the letters of the American language sign. Task.2 - Then, they were asked to complete an interactive learning assessment in which they watched an animation and answer to multiple choice questionnaire. Before and after the first task and after the assessment they were required to fill out the questionnaire about their emotions during the learning. Each session lasted an average of 30 minutes.

4.6 Result

In this section we report and compare the results of the experiment obtained through questionnaires and through facial analysis. In this first part, we will see the results of the questionnaire for the two groups. Each of the following tables reports the mean of answers for the two groups for each questionnaire item, the difference between the means and each standard deviation (S.D.) for the relative mean. We can state that in general the results between the means

are similar and all the standard deviations are low. This means that it is not possible to appreciate particular differences between the two groups' reactions. The only noteworthy items are d5 and a2 that are both related to the feeling to be proud of their own capacities demonstrated during the learning and the assessment. Precisely, d5 has a difference of 0.8 in favour of the control group and a2 0.6 again in favour of the control group. Even though there is no such a big difference between these means, researchers noticed that during the experiment Gr1 expended some more efforts watching the videos and interacting with the CNN software than Gr2 who had only to watch videos and this can affect the users' perception of the one's abilities. Further investigation is needed to explore this sentiment.

Table 2. Emotional questionnaire results related to the users' emotions before starting the learning

Before the learning session						
	b1	b2	b3	b4	b5	b6
Gr1 mean	4.8	4.4	4.6	4	4	3.5
Gr2 mean	5	3.8	4.4	4	3.8	3.2
difference	-0.2	0.6	0.2	0	0.2	0.3
Gr1 S.D.	0.89	0.89	0.45	0.71	0.84	1.64
Gr2 S.D.	0	1.64	0.89	1	1.30	0.84

Table 3. Emotional questionnaire related to the users' emotions during the learning session

During the learning session											
	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11
Gr1 mean	4.4	4.7	4.6	4.8	3	4.9	4.9	4.1	4.3	4.5	4.6
Gr2 mean	4.8	5	5	5	3.8	4.8	4.8	4	4.4	4.6	4.6
difference	-0.4	-0.3	-0.4	-0.2	-0.8	0.1	0.1	0.1	-0.1	-0.1	0
Gr1 S.D.	1.00	0.55	1.30	0.55	0.45	0.00	0.00	0.84	0.84	0.89	0.55
Gr2 S.D.	0.45	0.00	0.00	0.00	0.84	0.45	0.45	1	0.89	0.55	0.89

Although there are no differences through the test, the study carried out through facial analysis gives us some more insights. The results obtained through the facial analysis in terms of valence and engagement are reported below. The valence is a measure of the positive or negative nature of the recorded person's experience: positive, negative, neutral. The engagement is a measure of facial muscle activation that illustrates the subject's expressiveness. Each of the fol-

Table 4. Emotional questionnaire results after the assessment

After the assessment						
	a1	a2	a3	a4	a4	a6
Gr1 mean	4.2	4	5	4.9	3.3	4.9
Gr2 mean	4.4	4.6	5	5	3.6	5
difference	-0.2	-0.6	0	-0.1	-0.3	-0.1
Gr1 S.D.	1.00	1.14	0.00	0.45	1.22	0.45
Gr2 S.D.	0.55	0.55	0.00	0.00	1.67	0.00

lowing pie charts shows the average of the time percentages of the positive, negative, and neutral valence for Gr1 and Gr2. We can say that by using our system, the positive value of the participants' experience is more significant than that obtained by a group that had to watch the videos.

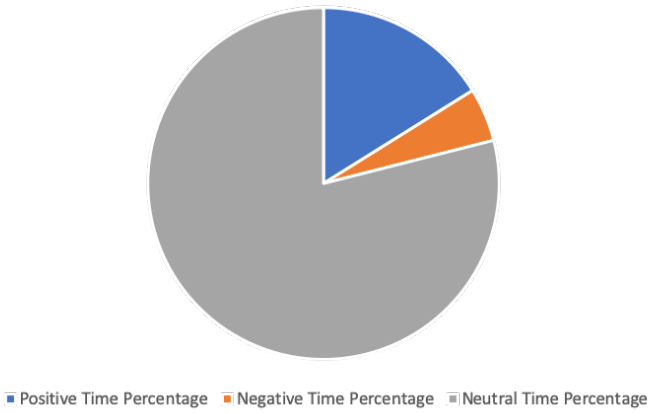

Fig. 6. Average valence using the our system.

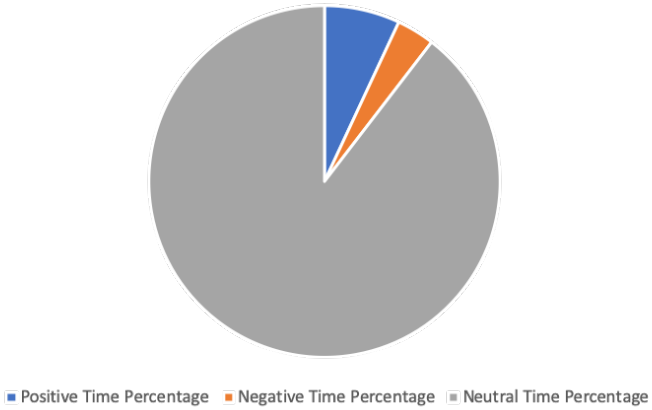

Fig. 7. Average valence using the traditional system.

Moreover, analyzing the engagement for the two groups, it is possible to see in 8 that even in this case, the one obtained for the group that used our system was higher. This means that our system can be considered a more than valid alternative for learning sign language, as it is much more stimulating for the user.

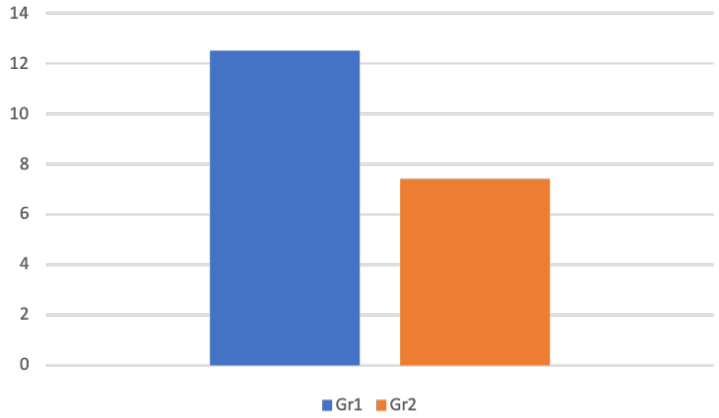

Fig. 8. Average percentage of engagement for each group.

This led us to the development of the next section, in which we analyze the real effectiveness and efficiency of the learning system through a longitudinal study.

5 Usability Evaluation

In this section, we present a longitudinal study we performed to measure both the usability of the software and its effects on the learning efficacy over the time.

5.1 Experiment

The main objective of our experiment was to carry out a study for the learning of sign language. Each participant used the application performing the gesture of the 24 letters of the American alphabet. The 12 participants of the experiment were divided into two groups: 5 participants were tested on the application once a day, for three days. This made it possible to detect improvements in user learning. To accomplish that, two different metrics were chosen:

- *Accuracy:* percentage of gestures performed correctly;
- *User satisfaction:* the subjective impression of the user (SUS).

We focused on accuracy rates as a basic metric as we can immediately realize how many errors the user has committed. In addition to empirical data, it is also important to collect subjective information. As the opinions of users are very important to define any future developments that can improve our application. The experiment started with an experimenter explaining the task and the system to the participants. It was explained to them that the purpose of the experiment was to test and evaluate this new intelligent sign language learning system, not their abilities. The participants were asked to replicate the gesture reproduced by the GIF, trying to simulate it as precisely as possible. The participants were positioned less than a meter from the computer on which the application was launched, comfortably seated in an armchair to create as much as possible a natural environment to be used. At the end of the experiment, the participants were asked to complete a SUS (System Usability Scale) questionnaire [1]. It consisted of 10 statements to which the participant assigned a score on a scale from 1 (strongly discouraged) to 5 (strongly agreed). The final score of the SUS varies from 0 to 100. A higher score indicates greater usability by the participant. Furthermore, after the experiment, comments and suggestions were also collected from the participants. The experiments were evaluated according to the Within-subject design, i.e., all the subjects participating in the experiment were tested in each condition (on each letter of the alphabet). The order of the letters to be executed was counterbalanced among the participants, to avoid problems of order effects.

5.2 Result

Each experimental session was allocated a 10 minutes time slot. All participants completed the experiment. To assess the accuracy of the method used for each participant, we recorded the level of learning achieved. As shown in figure 9 the participants obtained an accuracy ranging between 17% and 65% (M = 33.9%, DS = 16.3). The figure refers to the percentage obtained after only one use of the system by the 12 participants.

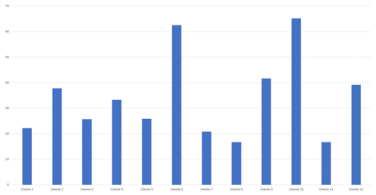

Fig. 9. User learning achievements.

As for the results gained over several days, performed by 5 of the participants, figure 10 shows how the accuracy increases through the different sessions. This result was very useful as the basic hypothesis of the experiment was verified, namely that the system can be used as a support for learning sign language.

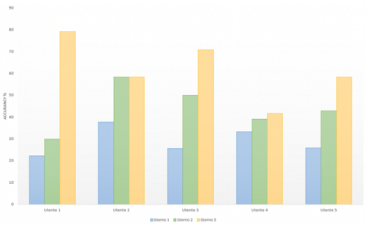

Fig. 10. Learning progress through the three experimental sessions.

The average SUS score was 73.7 (SD = 29) for our system. All participants said that they find our application very stimulating for learning sign language. The use of the application by the participants was considered a choice that is always preferable to self-learning through video or other applications, above all because it is possible to have immediate feedback on one's level of learning. Moreover, since our application provides feedback for every gesture made, the user can also see his/her mistakes and become aware of the gesture to be improved.

6 Artificial Intelligence at the Edge

Sign languages are made of sequences of signs. Although, to use regular neural network for the machine learning of sequential data, as videos (a sequence of frames), could be possible feeding the CNN with entire sequence, the constraint of a fixed size of input could be an unacceptable limit. What we need, instead, is to feed an arbitrary length of images sequence, one element per time step and a neural network which has some kind of memory to remember events happened many step times in the past. This behavior is best performed by RNN with LSTM models. While these kinds of network are successfully used for Sequence Learning applications, they require extraordinary computational resource. To offer additional computational resources to the locally connected end-user devices, this paper proposes a *Fog-Computing* Architecture.

6.1 The Fog-Computing paradigm

Fog-Computing is an emerging distributed computing model aimed at bringing computation close to its data sources, which can reduce the latency and cost of delivering data to a remote cloud. This feature and related advantages have been profitably exploited in different application scenarios, especially for latency sensitive and mission intensive services [3] [16]. The definition and architecture of Fog-Computing model [17] are briefly presented in this paper, while it proposes a new specific hardware platform.

Fog-Node is the core component of *Fog- Computing* architecture. They are between the edge of the network, and Cloud resources, distributed on layers, offering connectivity and computing resources to the smart end-devices (figure 11). The paradigm is that, from Edge to the Cloud, there are layers which elaborate the data, forwarding the results to the upper layer, and eventually, collaborate with other *Fog-Nodes* to distribute the processes "in press" [4] and storage [21]. Usually, the lower layer offers connectivity to the end-user device and is part of the local network. Furthermore, the lower layer offers computation resources and limited storage to locally connected devices, with low latency and without the need of the Internet connection.

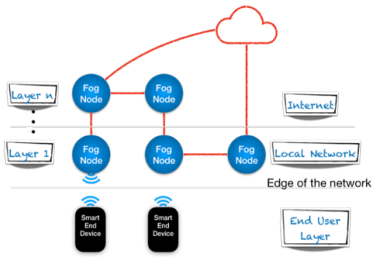

Fig. 11. An exemple of Generic *Fog-Computing* Architecture.

6.2 The Proposed Framework

In the Deep Neural Network (DNN) models, the large number of identical neurons, makes it natural to consider high parallelism in the computation. Actually, in 2012, the team led by Alex Krizhevsky, the creator of AlexNet, used for the first time a Graphics Processing Unit (GPU) -accelerated DNNs, winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by a large margin [14]. The parallelism naturally fits with GPUs architecture, which provides significant speed-up over traditional CPU. Currently, an increasing number of promising devices are available on the market, which claim to bring modern AI to the makers and to the embedded developers. We selected the new low-cost NVIDIA Jetson Nano module, which features a 128-core GPU and allows to use libraries like Python, C++, CUDA X and OpenGL. A wide variety of deep neural network models that enable tasks like image recognition and object detection can be accelerated with support from NVIDIA CUDA® Deep Neural Network

library (cuDNN) and TensorRT tm. The cuDNN library has support for RNN, which are widely used for sequence learning in many fields. TensorRT is a platform for high-performance deep learning inference. It contains a deep learning inference optimizer and a runtime that delivers low latency and high-throughput for deep learning inference applications. Applications based on TensorRT perform up to 40x faster than CPU-only platforms, during inference. The idea is to optimize trained neural network models and calibrate them for lower precision with high accuracy on the Cloud data center, and finally deploy the optimized Inference Engine to the Fog-Node, at the edge of the network (figure 12), deploying there the NVIDIA Jetson Nano module. Smart user devices, which could not have computational resources to deploy a fast enough inference engine, may demand the elaboration of image sequences to the locally networked *Fog-Node*. The optimized Inference Engine of *Fog-Node* can then answer with low latency, and parallelism in the computation, to the end-user device requests. Single *Fog-Node* can serve more than one end- user device, and more *Fog-Nodes* can be added whenever it needs to scale up for more computational resources.

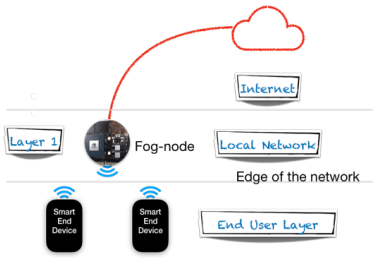

Fig. 12. Simplified schema of the proposed framework.

7 Conclusions

This paper proposed a simple and interactive learning method and application, which relies on the potentials of deep learning tools to offer a chance to those who are inexperienced in Sign Languages for hearing and speech impaired people. The method has been initially adopted to develop a system which supports ASL beginners during alphabet learning tasks and incite them to exercise. The accuracy of the proposed system has been calculated; the user's learning curve has been evaluated and compared against traditional video training. The outcomes obtained from the tests have produced positive results, indicates that the solution may not only improve the user's learning curve but also encourage training, due to its efficient methodology. The last is particularly crucial for self-learning courses, where the absence of a tutor or teacher who encourages the student to exercise has to be rendered by an attractive and desirables instrument of learning. Furthermore, in order to apply a similar approach and support to learning of a complete sign language, we have adopted a new architecture that may deal with the increased complexities of sign languages, as natural languages

with their grammar and lexicon. So, to rely on more powerful computational resources while delivering the final support to learners on their mobile smart devices, our idea is to bring AI at the Edge, proposing the adoption of a modern *Fog-Computing* architecture with low-cost small AI Computer hardware. The development of the new system is an ongoing activity at our lab, and the next step of our study will be devoted to benchmarking the user experience of learning with the proposed solution.

Acknowledgment

This work was supported in part by E3APP project, POR CAMPANIA FESR 2014/2020

References

1. System usability scale, `https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html`
2. Babakus, E., Mangold, W.G.: Adapting the servqual scale to hospital services: an empirical investigation. Health services research **26**(6), 767 (1992)
3. Battistoni, P., Sebillo, M., Vitiello, G.: Experimenting with a fog-computing architecture for indoor navigation. In: 2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC). pp. 161–165 (June 2019). https://doi.org/10.1109/FMEC.2019.8795307
4. Battistoni, P., Sebillo, M., Vitiello, G.: Computation offloading with mqtt protocol on a fog-mist computing framework. In: International Conference on Internet and Distributed Computing Systems. pp. 140–147. Springer (2019)
5. Buttle, F.: Relationship marketing
6. Cao Dong, Leu, M.C., Yin, Z.: American sign language alphabet recognition using microsoft kinect. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 44–52 (June 2015). https://doi.org/10.1109/CVPRW.2015.7301347
7. Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1610–1618 (July 2017). https://doi.org/10.1109/CVPR.2017.175
8. Cui, R., Liu, H., Zhang, C.: A deep neural framework for continuous sign language recognition by iterative training. IEEE Transactions on Multimedia **21**(7), 1880–1891 (July 2019). https://doi.org/10.1109/TMM.2018.2889563
9. Di Gregorio, M., Sebillo, M., Vitiello, G., Pizza, A., Vitale, F.: Prosign everywhere-addressing communication empowerment goals for deaf people. In: Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good. pp. 207–212 (2019)
10. Forlizzi, J.: Emotion and the design of new technology. In: 2003, C. (ed.) booktitle (2003)
11. Jakob, N., Jonathan, L.: Measuring usability: Preference vs. performance. Commun. ACM **37**(4), 66–75 (Apr 1994). https://doi.org/10.1145/175276.175282, `http://doi.acm.org/10.1145/175276.175282`
12. J.B. De Quirós; Schrager, O.L.: Neuropsychological fundamentals in learning disabilities. Academic Therapy Publications. Academic Therapy Publications (1979,)

13. Keskin, C., Kıraç, F., Kara, Y.E., Akarun, L.: Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In: European Conference on Computer Vision. pp. 852–863. Springer (2012)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. pp. 1097–1105. NIPS'12, Curran Associates Inc., USA (2012). https://doi.org/10.1145/3065386, http://dl.acm.org/citation.cfm?id=2999134.2999257
15. Kuznetsova, A., Leal-Taixé, L., Rosenhahn, B.: Real-time sign language recognition using a consumer depth camera. In: 2013 IEEE International Conference on Computer Vision Workshops. pp. 83–90 (Dec 2013). https://doi.org/10.1109/ICCVW.2013.18
16. Liu, Y., Fieldsend, J.E., Min, G.: A framework of fog computing: Architecture, challenges, and optimization. IEEE Access **5**, 25445–25454 (2017). https://doi.org/10.1109/ACCESS.2017.2766923
17. M. Iorga, N.G., Feldman, L., Barton, R., Martin, M., Mahmoudi, C.: Fog computing conceptual model, special publication (nist sp)
18. Madhuri, Y., Anitha., G., Anburajan., M.: Vision-based sign language translation device. In: 2013 International Conference on Information Communication and Embedded Systems (ICICES). pp. 565–568 (Feb 2013). https://doi.org/10.1109/ICICES.2013.6508395
19. Maqueda, A.I., Del Blanco, C.R., Laureguizar, F., García, N.: Human-computer interaction based on visual recognition using volumegrams of local binary patterns. In: 2015 IEEE International Conference on Consumer Electronics (ICCE). pp. 583–584 (Jan 2015). https://doi.org/10.1109/ICCE.2015.7066536
20. Marcon, M., Paracchini, M.B.M., Tubaro: A framework for interpreting, modeling and recognizing human body gestures through 3d eigenpostures **10**(5), 1205–1226 (May 2019)
21. Moysiadis, V., Sarigiannidis, P., Moscholios, I.: Towards distributed data management in fog computing **2018**, 14. https://doi.org/10.1155/2018/7597686
22. Nai, W., Yue, Rempel, D., Wang, Y.: Fast hand posture classification using depth features extracted from random line segments. Pattern Recognition **65**, 1 – 10 (2017). https://doi.org/https://doi.org/10.1016/j.patcog.2016.11.022, http://www.sciencedirect.com/science/article/pii/S0031320316303806
23. Norman, D.: Emotional Design: Why We Love (Or Hate) Everyday Things (2004)
24. Ong, W., S.C., Ranganath, S.: Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Trans. Pattern Anal. Mach. Intell. **27**(6) (Jun 2005). https://doi.org/10.1109/TPAMI.2005.112
25. Panzner, M., Cimiano, P.: Comparing hidden markov models and long short term memory neural networks for learning action representations. In: International Workshop on Machine Learning, Optimization, and Big Data. pp. 94–105. Springer (2016)
26. Pekrun, R., Goetz, T., Frenzel, A.C., Barchfeld, P., Perry, R.P.: Measuring emotions in students' learning and performance: The achievement emotions questionnaire (aeq). Contemporary Educational Psychology **36**(1), 36 – 48 (2011). https://doi.org/https://doi.org/10.1016/j.cedpsych.2010.10.002, http://www.sciencedirect.com/science/article/pii/S0361476X10000536, students' Emotions and Academic Engagement
27. Pugeault, Nicolas, Bowden, Richard: Spelling it out: Real-time asl fingerspelling recognition. pp. 1114–1119 (11 2011). https://doi.org/10.1109/ICCVW.2011.6130290

28. Rioux-Maldague, L., Giguère, P.: Sign language fingerspelling classification from depth and color images using a deep belief network. In: 2014 Canadian Conference on Computer and Robot Vision. pp. 92–97 (May 2014). https://doi.org/10.1109/CRV.2014.20
29. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR 2011. pp. 1297–1304 (June 2011). https://doi.org/10.1109/CVPR.2011.5995316
30. v.v.: Convention on the rights of persons with disabilities (crpd). accessed in august 2019, `www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html.`
31. Wang, C., Liu, Z., Chan, S.: Superpixel-based hand gesture recognition with kinect depth camera. IEEE Transactions on Multimedia **17**(1), 29–39 (Jan 2015). https://doi.org/10.1109/TMM.2014.2374357
32. Wohlkinger, W., Vincze, M.: Ensemble of shape functions for 3d object classification. In: 2011 IEEE International Conference on Robotics and Biomimetics. pp. 2987–2992 (Dec 2011). https://doi.org/10.1109/ROBIO.2011.6181760
33. Zhang, C., Yang, X., Tian, Y.: Histogram of 3d facets: A characteristic descriptor for hand gesture recognition. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–8 (April 2013). https://doi.org/10.1109/FG.2013.6553754
34. Zhang, C., Tian, Y.: Histogram of 3d facets: A depth descriptor for human action and hand gesture recognition. Computer Vision and Image Understanding **139**, 29 – 39 (2015). https://doi.org/https://doi.org/10.1016/j.cviu.2015.05.010, `http://www.sciencedirect.com/science/article/pii/S1077314215001216`