EasyChair Preprint

№ 10780

Word Construction Through Sign Concatenation Based on Deep Learning

Cherrate Meryem, Sabri My.Abdelouahed, Yahyaouy Ali and Aarab Abdellah

August 27, 2023

Word construction through sign concatenation based on Deep Learning

CHERRATE Meryem

Department of Computer science,

Faculty of Sciences Dhar-Mahraz, USMBA,

Fez, Morocco

meryem.cherrate@usmba.ac.ma

My Abdelouahed SABRI

Department of Computer science,

Faculty of Sciences Dhar-Mahraz, USMBA

Fez,Morocco

abdelouahed.sabri@usmba.ac.ma

YAHYAOUY Ali

Department of Computer science,

Faculty of Sciences Dhar-Mahraz, USMBA,

Fez, Morocco

ali.yahyaouy@usmba.ac.ma

AARAB Abdellah

Department of physique,

Faculty of Sciences Dhar-Mahraz, USMBA

Fez,Morocco

aarab_abdellah@yahoo.fr

Abstract

Many deaf-mute people suffer from hearing loss, which gives us problems communicating between them and other normal mute people, especially for those who don't understand sign language, which is the language used by deaf-mute people to express their emotions and opinions, as well as to transfer an idea to others. So we're going to try and find a solution to ensure communication between these two categories of people, in the following form: a tool capable of transcribing American Sign Language (ASL) into natural language by concatenating these signs to construct a word, with the application of various deep learning techniques.

Keywords: sign language deaf-mute hearing loss natural language ASL transcription concatenation communication deep learning

1. Introduction

One of the major problems faced by a person who can't speak is that they can't express their emotions clearly in this world, and it's communication with normal people that poses the problem of exchanging information. These people represent 5% of the world's population (according to World Health Organization statistics for 2020) and use sign language as a means of communicating with others. Numerous researchers [2], [3], [4], [5], [6], [7], [8], [9], have put considerable effort into creating a tool to help these people communicate and exchange information, using artificial intelligence techniques as well as conventional solutions, but the problem remains, That's why in this article we're going to focus on American Sign Language (ASL) to find a solution that will transcribe the signs into natural language by applying Deep Learning algorithms, and to achieve this task we're going to follow the following plan:

- The pre-processing step" Resize the images in the training database ".
- Extract features (Color, Shape, Texture).
- Create model.
- Performance analysis.
- Classification and prediction stage.

In the following, we will detail each phase in the processing of a sign recognition problem: - The pre-processing stage: is a process applied to images of insufficient quality or containing artifacts that could negatively influence the following stage.

- Feature extraction: is a crucial step in obtaining a good classification rate. Among the features we extract are the color, shape and texture of the sign.
- Model creation.
- Performance analysis.
- The classification & prediction stage: this is the last stage of such a system, which classifies the signs and their appropriate alphabet.

The following section presents a general overview of sign language and, as a special case, American Sign Language, followed by the 3rd section, which presents the state-of-the-art in approaches to the same problem. The 4th section presents the proposed approach in detail. In the 5th section, we present the results of the application of our approach and the dataset used in our project. A synthesis of our application will be presented in the final section.

2. The Sign language

2.1 Statistics on deaf mutes

There are a significant number of deaf-mutes in the world, with 1.5 billion people suffering from some degree of hearing impairment. Of these, 430 million require rehabilitation services. France is a case in point, with over four million people affected to a greater or lesser degree. This represents 6% to 8% of the French population, divided as follows: 2 million are mildly hard of hearing, 1.8 million are moderately deaf and 200,000 are profoundly deaf.

Even in Morocco, we have a significant number of these people in our society, which presents 23,000 deaf children of school age (according to the national survey of 2004). About 800 children are in school according to the figures of the Ministry of National Education and more than 1'200 are in school through associations. Therefore, there is a difference of 21'000 children who are on the waiting list. Therefore, according to these statistics, we can observe that the number is important and that we should not neglect this problem, even if these people communicate with other people with their own language, which is the sign language. So we can see that the number of these people is significant, and they use sign language as a means of communicating with others. In what follows, we'll introduce sign language.

2.2 General information on sign language

2.2.1 Definition

Sign language can be defined as a gestural language produced by the movements of the hands, face and body as a whole. This language is the language used by deaf-mute people to communicate with others.

2.2.2. American Sign Language

American Sign Language (ASL) is a language used by deaf and dumb people in the United States and especially by North American people, which is a combination of hand signs, facial expressions and body postures. This langage is the main language used by these people as a means of communication to express their emotions, expressions to other normal people.

In addition to North America, ASL dialects and ASL-based Creoles are used in many countries around the world, including much of West Africa and parts of Southeast Asia.

This figure below represents the American signs:

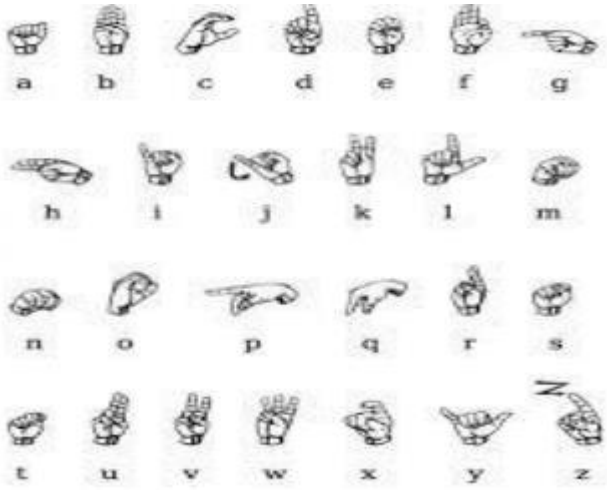

Figure 1: the American sign alphabet

3. State of the art

The problem of sign recognition is posed everywhere, as we must not overlook the large number of deaf-mute people worldwide, who represent over 450M of the world's deaf-mute population according to statistics from the World Health Organization [1], and it is estimated that by 2050 almost 2.5 billion people will be affected by a hearing impairment of varying degrees of severity, and at least 700 million will require rehabilitation services. This is why researchers from all over the world are working hard to find a solution to this problem and ensure communication between deaf-mutes and normal people. These solutions are hardware and software solutions proposed by researchers for sign recognition. Some of them propose solutions using hardware such as connected gloves, sensors to recognize gestures and hand and finger movements. These embedded solutions do not facilitate exchanges between deaf-mutes and normal people, nor do they recognize signs made by full-body gestures. As well as software solutions using image processing and artificial intelligence.

Recognition from gloves

A team of American researchers have [2] developed a glove capable of translating the 26 letters of the sign alphabet. This device is in the form of a sports glove made of leather material on which nine movement sensors have been placed at the level of the joints, this glove is composed of a set of electrodes that allow to translate the gestures made by the hand and the wrist in order to designate the 26 letters of the American alphabet, unfortunately this solution has limitations due to lack of funding because it is expensive and not very effective, as well as this system can only spell words. Another research done by Boon Giin Lee et al. in [3] proposed a portable intelligent hardware to do sign recognition, this system proposed by these researchers uses 10 sensors, five of them are flexible sensors, two pressure sensors and the last one for inertial movement with three axes, these researchers find with this system an accuracy rate of 65.7%.In parallel a set of researchers (Shukor et al.) [4] have proposed a glove to ensure the recognition of signs of discomfort, they based their system on ten sensors including the type of clicks, an accelerometer, a microcontroller and a Bluetooth module, this system has experienced problems during the test phase cat it was not tested properly in depth to approve its effectiveness but on 4 uses, they were able to obtain an accuracy of recognition of signs equal to 89%.

Sign recognition based on image processing and artificial intelligence

The solutions that we have already mentioned before, based on gloves and sensors, have limitations in their use, as well as not being able to manage all the gestures made by deaf-mute people. That's why some researchers propose to use solutions based on artificial intelligence and image processing that approve their efficiency in the recognition of the signs of deaf-mute people

In image processing and artificial intelligence based approaches to sign language recognition, image processing is used to capture and identify gestures while artificial intelligence is used for sign recognition and transcription into written language. Hand detection is the first step in such a system. And it is essential to detect it correctly in order to be able to effectively recognize the sign made. This step is considered a challenge.

The authors [5] have proposed an approach to detect and localize the hand landmarks using color images, this approach consists of using a mask to identify the skin and the use of a distance to detect the hand landmarks.

Ravikiran et al. in [6] proposed an approach for finger detection. Fingers are detected based on boundary tracing for tip detection. For sign recognition, we use machine learning methods for their classification. Thus, the objective is to design a classification model using supervised classification algorithms able to recognize the sign of a gesture image. In this step, a database of labeled images is used to train a classification model.

In the paper [7], the authors used two deep learning architectures, AlexNet and VGG16 for feature extraction. Classification is performed using the SVM algorithm. The authors obtained a very good classification rate, but using only the training data. Deriche et al. in [8] proposed to use a classification based on a combination of a Bayesian approach and a mixture of Gaussian models (GMM) with the use of linear discriminant analysis (LDA) for dimensionality reduction. The simulation results showed that the proposed approach performs moderately well with a classification rate equal to 92%.

Moroccan researchers in [9] proposed a new approach based on artificial intelligence and more precisely machine learning to help deaf mutes communicate with others and they followed the following steps:

- Image feature extraction (color, shape, texture)
- Features Engineering
- Classification

After these steps, they captured an image and as a result they had the alphabet suitable for the captured signs and they found that the system having a fully recurrent architecture offers the best performance with an accuracy of 98.33% for the recognition of static signs. Despite these hardware and software solutions, the problem still exists, and as researchers, we will try to propose our approach in the following sections.

4. Proposed approach

In this part, we will detail the proposed approach to transcribe American Sign Language through the concatenation of words and then the construction of a set of words.

The proposed approach consists in using different Deep Learning techniques based on an artificial neural network inspired by the human brain. This network is composed of tens or even hundreds of "layers" of neurons, each one receiving and interpreting the information of the previous layer. The system will learn, for example, to recognize letters before tackling the words in a text, or determine whether there is a face in a photo before discovering which person it is.

More exactly we will use CNN which have been very much used for image analysis. The two main characteristics of convolutional networks are that they use filters (kernel) and implement pooling. The filters analyze the images zone by zone. Each filter specializes in order to recognize patterns. For example, one filter may specialize in edge detection, while another will recognize certain shapes. Convolution increases the depth of the image matrix, as each filter adds a layer to it. Pooling reduces the size of an image by keeping only the most important pixels. This has the effect of distorting the image by losing the precise positioning of the pixels. This effect is in fact beneficial, since it limits the risks of overlearning. For example, a face detection system will be well advised to learn that a face consists of two eyes, a nose and a mouth, but it is preferable that it does not learn by heart the pixel

spacing between these different elements of the face, since their position can vary from one person to another. There are other techniques than pooling, in particular capsule networks, a new technique that will not be covered in this tutorial on convolutional networks.

The most common form of a convolutional neural network architecture stacks a few Conv-ReLU layers, follows them with Pool layers, and repeats this pattern until the input is reduced to a space of sufficiently small size. At some point, it is common to place fully connected (FC) layers. The last fully connected layer is connected to the output. Here are some common convolutional neural network architectures that follow this model:

- INPUT -> FC implements a linear classifier
- INPUT -> CONV -> RELU -> FC
- INPUT -> [CONV -> RELU -> POOL] * 2 -> FC -> RELU -> FC

Here, there is a single CONV layer between each POOL layer

- INPUT -> [CONV -> RELU -> CONV -> RELU -> POOL] * 3 -> [FC -> RELU] * 2 -> FC
  Here there are two CONV layers stacked before each POOL layer.

Stacking CONV layers with small pooling filters (rather one large filter) allows for more powerful processing with fewer parameters. However, with the disadvantage of requiring more computational power (to hold all the intermediate results of the CONV layer).

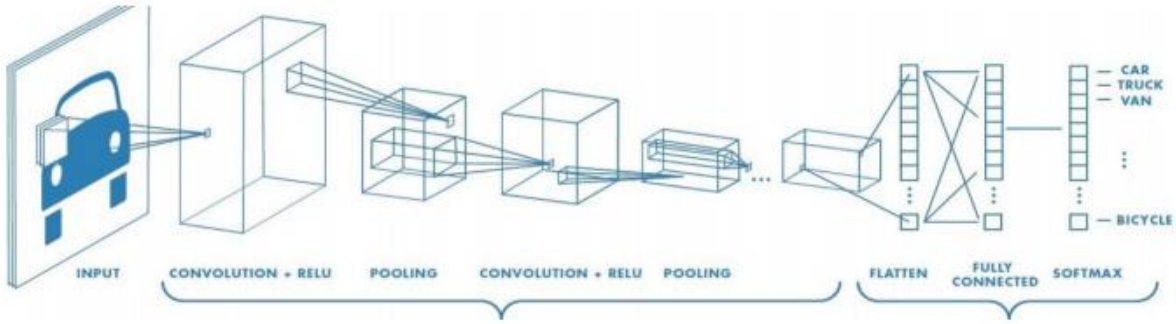

Figure 2: CNN Architecture

And below is the diagram of our application

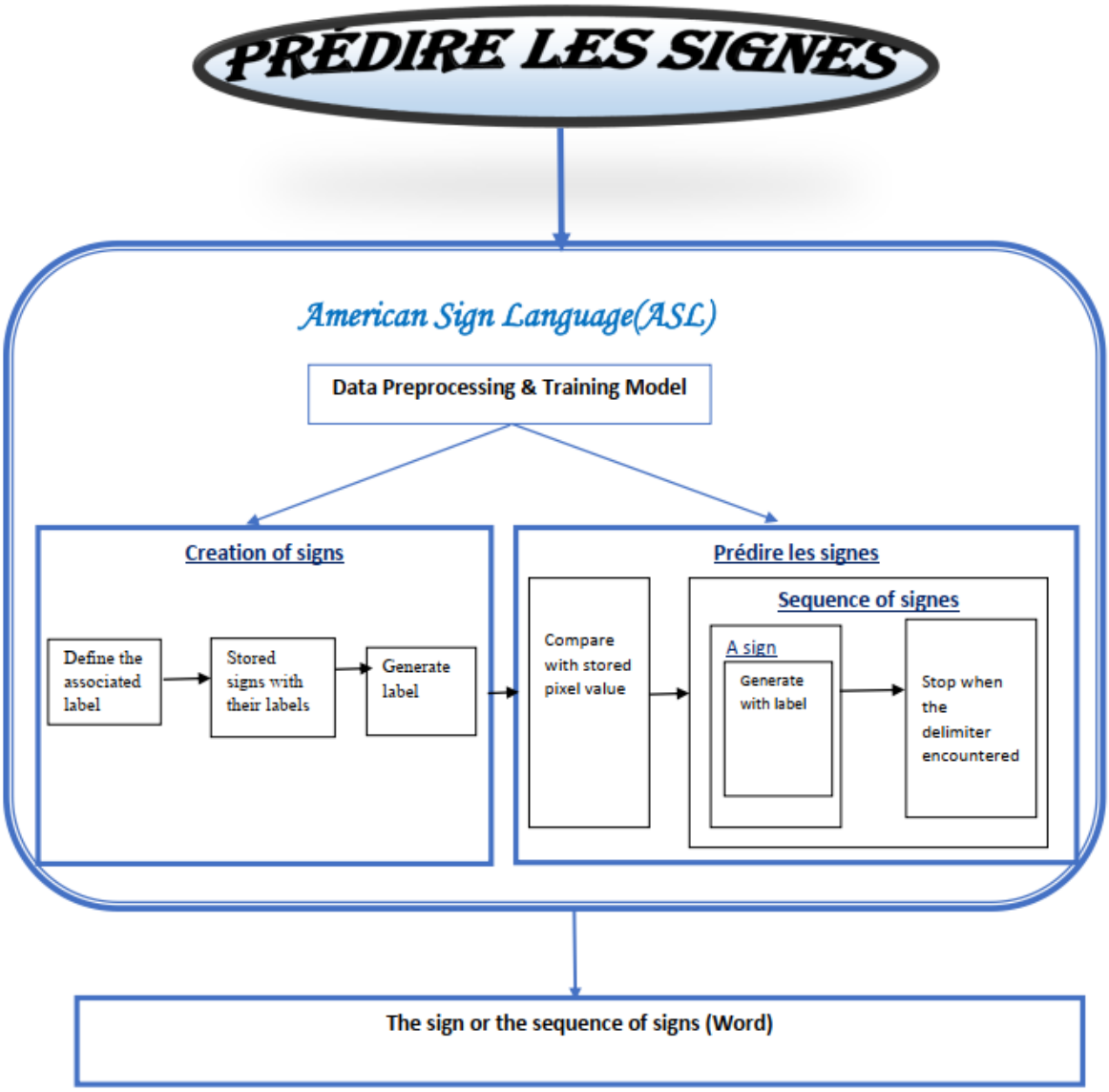

Figure 3: Prediction scheme of a sign and a set of signs

As we can clearly see, after applying the different Deep Learning mechanisms using CNN to have good classification results. The details of the mechanism of this process used are given in the previous subsections.

We summarize the main steps as the following descriptions:

- Data pre processing & training model: In this module, based on the object detected in front of the camera, its binary images are filled. This means that the object will be filled with solid white and the background will be filled with solid black. Depending on the regions of the pixel, their numerical value in a range of 0 or 1 is given to the next process for the modules.
- Predict Sign: a gesture scanner will be available in front of the user end where the user will have to make a hand gesture. Based on pre-processed output from the module, a user shall be able to see the associated label assigned for each hand gesture, based on the predefined American Sign Language (ASL) standard inside the output window screen.
- Create Sign: A user will give a desired hand gesture as input to the system with the text box available at the bottom of the screen where the user needs to type in what they wish to associate

with that gesture. This customize the gesture will then be stored for future use and will be detected in the time to come.

- Predict Sentence: A user will be able to select a delimiter and until that delimiter is met each character of digitized gesture will be appended with the previous results forming a stream of meaningful words and phrases.

Here is the diagram or we can say also it is the activity diagram of this application which is schematized in another way than the architecture above:

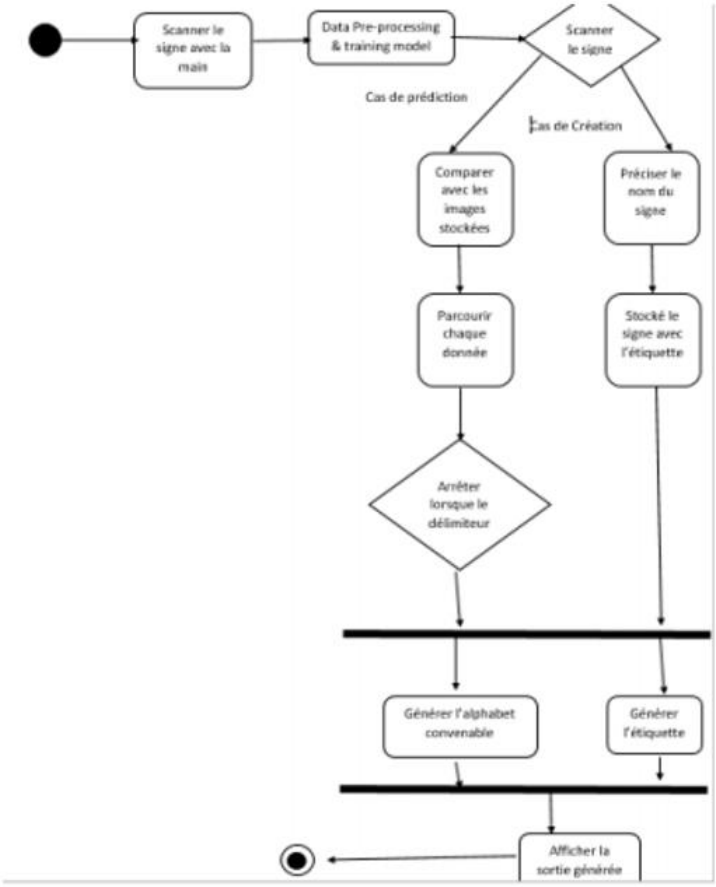

Figure 4: American Sign language activity diagram

5.Simulations results

4.1 Dataset utilisé

Afin d'assurer la communication entre les sourds-muets et les personnes normales qui utilisent deux langues différents pour communiquer on a mené à créer un outil capable de faire la transcription et la concaténation des signes pour construire un mot compréhensible par l'être humain .Donc on a besoin d'une base de données des images des signes américaines afin de réaliser cette tache.

Cette dataset est une base de données composé par un ensemble des images de 87 028 fichiers et 30 dossiers de taille 1,03Go séparé entre les images d'apprentissage et de test. L'ensemble de données est une collection d'images d'alphabet de la langue des signes américaine, séparé en 29 dossiers qui représentent les différentes classes. L'ensemble de données de formation contient 87 000 images de 200 x 200 pixels. Il y a 29 classes, dont 26 pour les lettres A-Z et 3 classes pour SPACE, DELETE et NOTHING. Ces 3 sont très utiles dans les applications en temps réel et la classification.

Cette figure montre les dossiers qui composent training dataset :

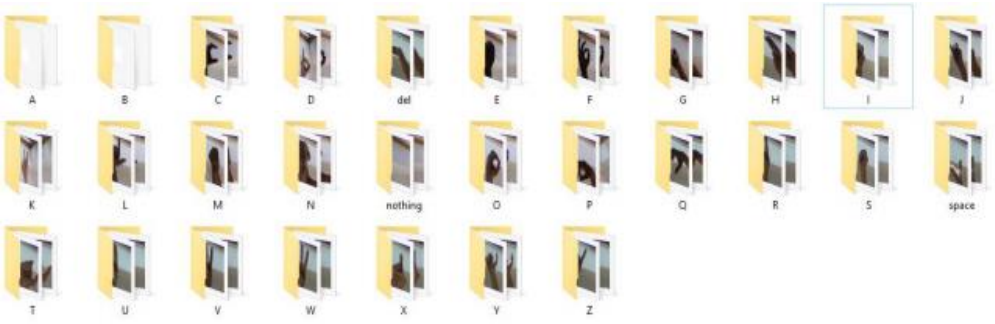

Figure 5:les images d'apprentissage

Pour les images de test on travailler avec ce jeu de données de test qui est de taille 323 Ko, composé de 28 Fichiers(images) d'extension .jpeg (.JPG) La figure ci-dessous montre les images de test :

Figure 6:les images de test

4.2 Résultats de simulation

En Deep Learning, on utilise CNN (Convolutional Neural Network) pour créer notre modèle en poursuivant les étapes suivantes :

- Création d'un réseau de neurones vide
- Ajout de la première couche de convolution, suivie d'une couche ReLU
- Ajout de la première couche de pooling
- Ajout de la deuxième couche de convolution, suivie d'une couche ReLU
- Ajout de la troisième couche de convolution, suivie d'une couche ReLU
- Flattening
- Full Connection
- Compiler CNN
- Fittting the CNN to the image
- Sérialisation du modèle(enregistrer)

On peut schématiser les étapes ci-dessus comme suivant :

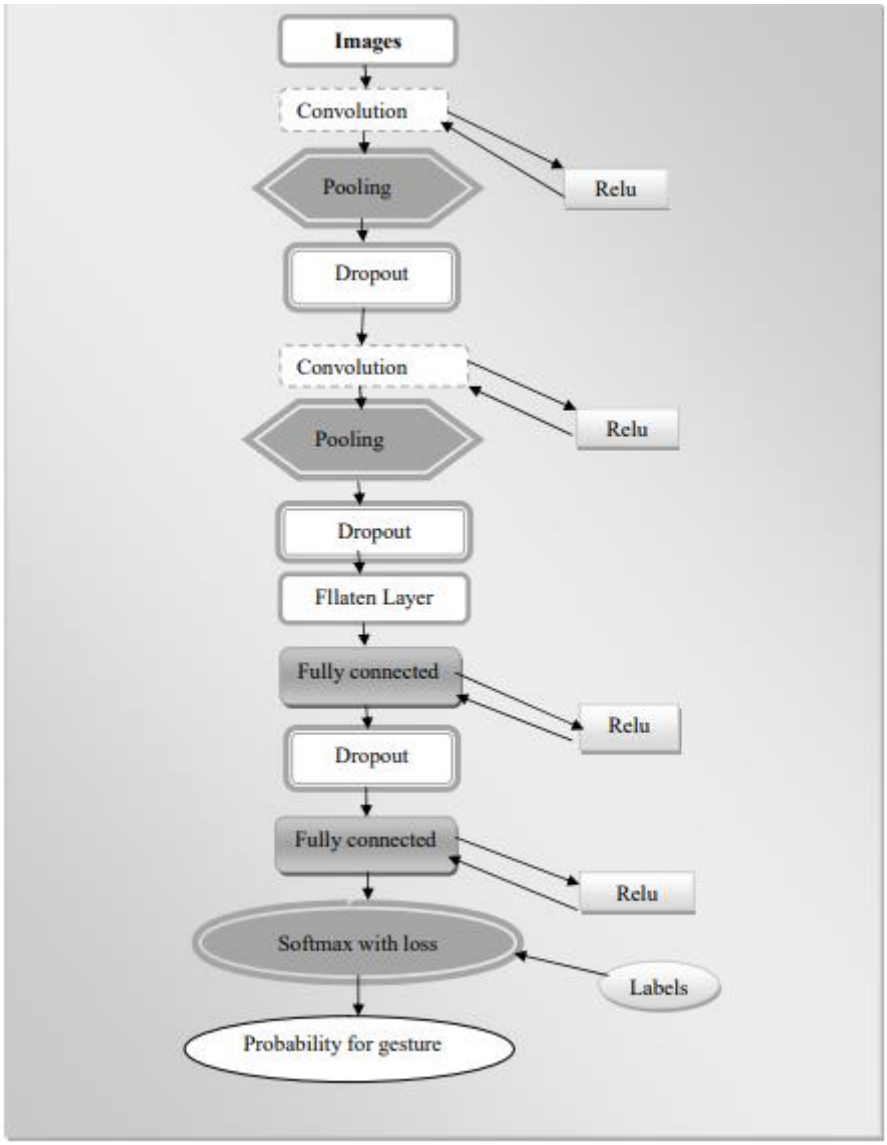

Figure 7:Architecture CNN utilisé pour la prédiction d'un ensemble des signes

Après la préparation des modèles en utilisant CNN , ça vient l'étape de prédiction d'un signe aussi un ensemble des signes .

Et pour cela on va baser sur La méthodes des SIFT [(scale-invariant feature transform= transformation de caractéristiques visuelles invariantes à l'échelle), qui est une méthode développée par David Lowe en 2004, permettant de transformer une image en ensemble de vecteurs de caractéristiques qui sont invariantes par transformations géométriques usuelles (homothétie, rotation) et de manière moins fiables aux transformations affines et à l'illumination.

Le point fort de la méthode de Lowe est qu'elle est capable de mettre en correspondance des points distants avec des variations de caméra importantes.

L'algorithme des SIFT vient combler en grande partie les limites des méthodes d'extraction de points remarquables déjà développées avant lui par Harris, et plus tard par Mohr et Schmid. En effet, il a contribué à l'amélioration des techniques d'extraction d'information dans une image en apportant un algorithme robuste et satisfaisant les propriétés que requièrent les procédés de vision artificielle notamment le recalage d'images," technique consistant à trouver une transformation géométrique permettant de passer d'une image (dite source) à une autre image (dite cible) "] pour assurer la réalisation de cette fonctionnalité.

Cette figure ci-dessous montre le processus de reconnaissance d'un signe afin de construire un mot et pourquoi pas une phrase.

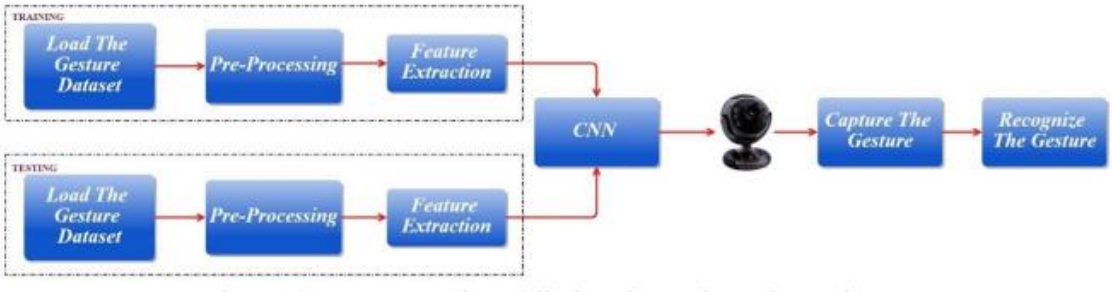

Figure 8:Processus de prédiction d'un signe & un ensemble des signes

Conclusion

A partir de cet article, nous avons essayé de faire oublier certains des problèmes majeurs auxquels sont confrontées les personnes handicapées en termes de conversation. Nous avons découvert la cause profonde de pourquoi ils ne peuvent pas s'exprimer plus librement. Le résultat que nous avons obtenu était de l'autre côté du public ne sont pas en mesure d'interpréter ce que ces personnes essaient de dire ou quel est le message ils veulent transmettre. Dans ce projet nous avons travaillé sur la reconnaissance des signes américains en utilisant les techniques de Deep Learning et plus précisément CNN.

La tâche principale c'est d'améliorer la précision de la classification avec utilisation du mécanisme de de Deep Learning, les résultats obtenus prouvent l'efficacité de notre système en donnant une précision moyenne de 98% sur le jeu de données des signes.

Dans travail futur, nous avons l'intention de proposer une dataset des mots afin de donner une précision exacte pour la classifications des mots et de construire une phrase cette technique sur des autres signes (MSL, LSF, …) afin de détecter, classer et prédire automatiquement les signes et par suite faire la transcription de ces signes en langue naturelle et pourquoi pas de construire notre propre dataset en utilisant des morceaux de vidéo et faire la transcription des mots et des phrases en langue naturelle.

Références

[1] W. H. O. (WHO, 1 april 2021. [En ligne]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2]B. G. L. a. S. M. Lee, ""Smart Wearable Hand Device for Sign Language Interpretation System With Sensors Fusion,"," *IEEE Sensors Journal,* vol. 18, pp. 1224-1232, Feb.1, 2018.

[3]M. F. M. M. H. J. F. A. I. M. F. A. a. M. B. B. A. Z. Shukor, ""A new data glove approach for Malaysian sign language detection"," *Procedia Comput. Sci.,* vol. 76, pp. 60-67, Decembre 2015.

[4]T. K. M. &. G. A. Grzejszczak, "Hand landmarks detection and localization in color images.," *Multimed Tools,* pp. 75-16363–16387 , 2016.

[5]J. M. K. M. S. D. R. S. S. &. P. N. V. Ravikiran, "Finger detection for sign language recognition," in *In Proceedings of the international MultiConference of Engineers and Computer Scientists*, March 2009.

[6]A. K. R. &. J. R. Barbhuiya, "CNN based feature extraction and classification for sign language.," in *Multimed Tools*, 2021.

[7]S. O. A. a. M. M. M. Deriche, " "An Intelligent Arabic Sign Language Recognition System Using a Pair of LMCs With GMM Based Classification,"," in *IEEE Sensors Journal*, Sept.15, 2019.

[8]M. &. M. (. Al-Rousan, "Automatic recognition of Arabic sign Language finger spelling," *International Journal of Computers and Their Applications,* pp. 80-88.

[9]C. M. Y. A. A. A. Sabri abdelouahed, "Moroccan sign language recognition based on machine learning," in *ICSV22*, 2022.