Journal of Education for Pure Science- University of Thi-Qar
Vol.12, No2 (December, 2022)
Website: jceps.utq.edu.iq *Email: jceps@eps.utq.edu.iq*

DOI: http://doi.org/10.32792/utq.jceps.12.02.29

Decimal Digits Recognition from Lip Movement Using GoogleNet network

Kwakib Saadun Naif[1] **Prof. Dr. Kadhim Mahdi Hashim[2]**
[1,2] Computer Science Department, College of Education for pure Sciences, University of Thi-Qar , Iraq.

Received 2/6/2022 Accepted 3/7/2022 Published 01/12/2022

Abstract:

Lip reading is a visual way to communicate with people through the movement of the lips, especially the hearing impaired and people who are in noisy environments such as stadiums and airports. Lip reading is not easy to face many difficulties, especially when taking a video of the person, including lighting, rotation, the person's position and different skin colors...etc. As researchers are constantly looking for new techniques for lip-reading.

The main objective of the paper is to design and implement an effective system for identifying decimal digits by movement. Our proposed system consists of two stages, namely, preprocessing, in which the face and mouth area are detected, lips are determined and stored in a temporary folder to used viola jones. The second stage is to take a GoogleNet neural network and insert the flange frame in it, where the features will be extracted in the convolutional layer and then the classification process where the results were convincing and we obtained an accuracy of 87% by using a database consisting of 35 videos and it contained seven males and two females, and the number of the frame was 21,501 lips image.

Key word: viola jones and GoogleNet

1. Introduction

Our research paper aims to read the lips of a group of personal speaking decimal digits in the absence of sound through the neural network GoogleNet. lip reading It is a very important method used by individuals with hearing impairments to recognize or interpret speech by visually watching the movement of the lips [1]. It is a way to predict words and phrases from watching a video only without any audio signal[2]. Lip reading is very difficult to teach because one must learn the language and context of the conversation in order to read correctly [3]. Lip reading has many facets due to the emergence of a wide range of databases covering thousands of vocabulary from sentences, words and digits [4]. This topic has received great attention in recent years due to its many uses in modern applications, such as image processing, pattern recognition, object detection, statistical modeling, artificial intelligence, etc. It is also used in speech recognition [5]. GoogleNet results are more accurate when the lip position is in the correct position and the features extracted are powerful so it is necessary to focus on the lip area. [6]. Our proposed system focuses on detecting the face and then identifying the lips, which are difficult tasks due to differences in sensor presence, background, and light.

This paper is structured as following: Section two overviews the related work. Section three introduces the layout proposed system. Section four describes the results and discussion of conduct tests. Finally, section five the derived conclusions of this paper.

2. Related Work

It is a technique through which speech can be understood, especially mastered by people with hearing difficulties, as it enables them to communicate with others and engage in social activities that would otherwise be difficult. In this section, we will review some of the works that are related to our work. R. Bowden et al. (2009) in [2]. They proposed compared various features for Automatic lip reading, including four types of 2D DCT features, Active Appearance model (AAM) features, Eigen lip features and sieve features in general, AAM features with appearance outperform other types of feature, Which means that appearance is more power than the shape .The proposed obtained accuracy rate 65%. Bang et al. (2014) in [7]. They presented an engineering method, where they used artificial neural networks in the training and classification processes, and to extract the features they used the snake method and they got an accuracy of 60%. I. Anina et al.(2015) in [8]. A method was proposed to a process the problem of non rigid mouth movement analysis using a database OuluVS2 and they got a recognition of 47% . Mr. Befkadu Belete (2019) in [9]. He suggested the audio-visual method for lip reading, where Viola Jones algorithm was used to detect from the mouth, and to extract the features, it used discrete wavelet Transformation (DWT), and the database used was AAVC, and it obtained an accuracy of 72%, and the discrimination was 67.08%.

3. The proposed system

In this proposed system, we are trying to help the hearing-impaired personal by recognizing speech through the movement of the lips and this is done using the convolutional neural networks GoogleNet, where in the beginning a video clip was taken of the person speaking with decimal digits and segmentation it into a sequential frame and then we use the Viola-Jones algorithm that detects the face and the mouth area, then subtracts the mouth area is region of interest (ROI) and stores it in a temporary folder.

And Change the frame size to 224 * 224. Because GooleNet is select to this size then inserts it into the GoogleNet network which extracts the features in the convolutional layer and finally passes these features to classify them Is the pronunciation correct for the decimal numbers or not, and Figure 1 shows that

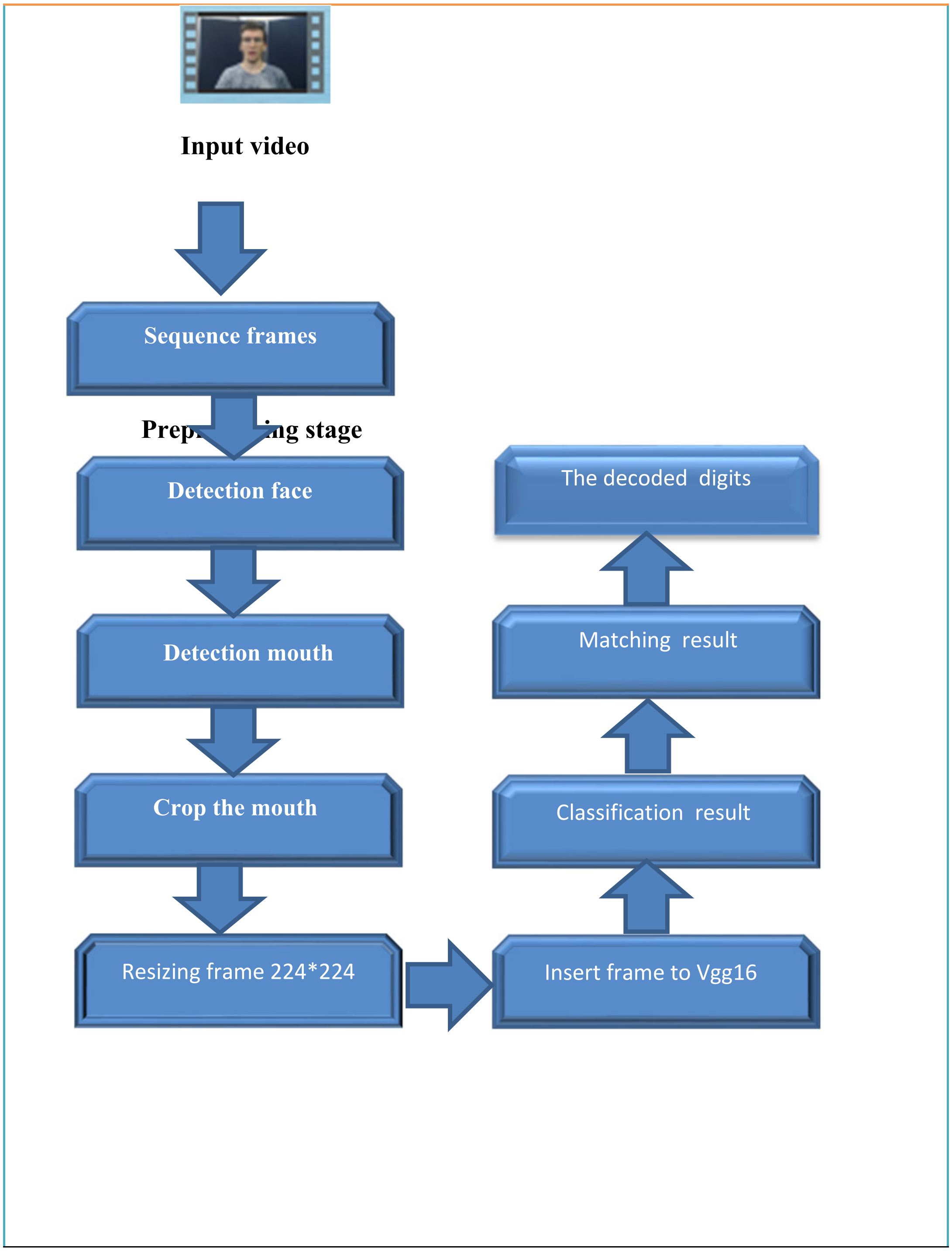

Figure 1: The block diagram of the proposed system automatic lip reading

3.1 The preprocessing stage

Pre-processing begins by cutting the video into a sequential frame then defining a region of interest (RIO) is mouth, which is the region that we closely investigate from a specific region within an image, where it needs engineering operations that modify the spatial coordinates of the image such as cropping, resizing, translate and rotation . Then mouth frames stored in temporary folder**.** Then we change the frame size to 224 * 224 to match the work of the network. The preprocessing stage includes three steps as in the figure 2.

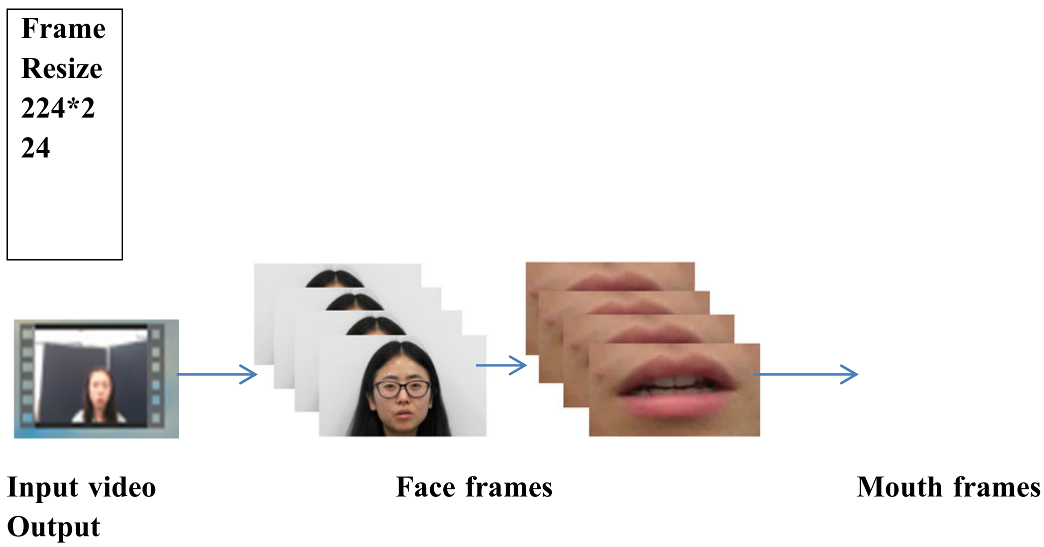

Figure 2: Block diagram the preprocessing steps

3.1.1 Face and mouth detection

At this stage, we use the Viola Jones algorithm (which is a good algorithm for detecting objects that was discovered in 2001 by the two scientists Michael Viola and Jones, and its success was with an accuracy of 95% [10]). The face is detected first, then the mouth area The following figure shows that:

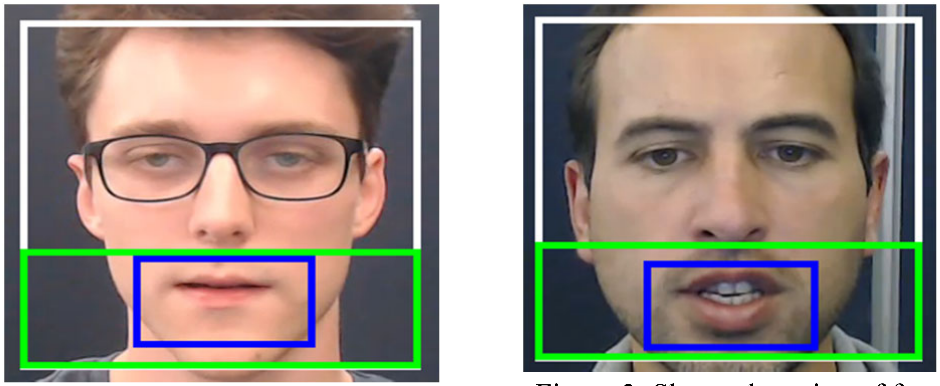

Figure 3: Shows detection of face and mouth by viola jones algorithm

3.1.2 Crop mouth image from the frame

This is the second process in the preprocessing stage Where this process takes place after the detection of the mouth area is now deducted from the image because it is (ROI). Figure 4 is an example of the mouth area crop. The input image to GoogleNet network must be in size 224*224.

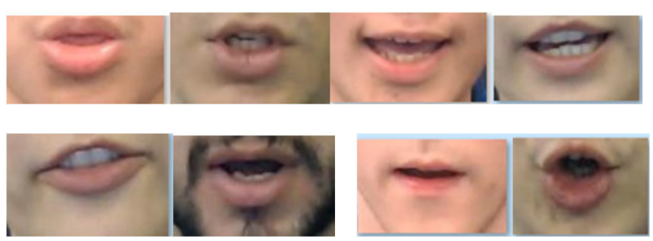

Figure 4: Sample for cropping the mouth

3.2 Convolution neural network stage:

Convolution neural network (CNN). It is a special kind of neural network, a method inspired by the visual cortex of the human brain. It is considered a good model for extracting features in deep learning and achieved remarkable success in image recognition, as it was used by many industry leaders such as Facebook and Google [12]. After completing the pre-processing process, the frames are fed into CNN, and there the features are extracted, reduced, and classified by frames.

3.2.1 Convolution neural network layers

A special type of linear operation used by a neural network [13]. CNN was introduced by Lecun et al.1990 as a solution to a classification task created by Computer Vision [14]. It consists of three types of layers: convolutional layer, pooling layer, and fully connected layers. The function of the first layer is to extract features and the second layer is to reduce the size The function of the third layer is to classify the features extracted at the end [15]. as shown in the figure following [16].

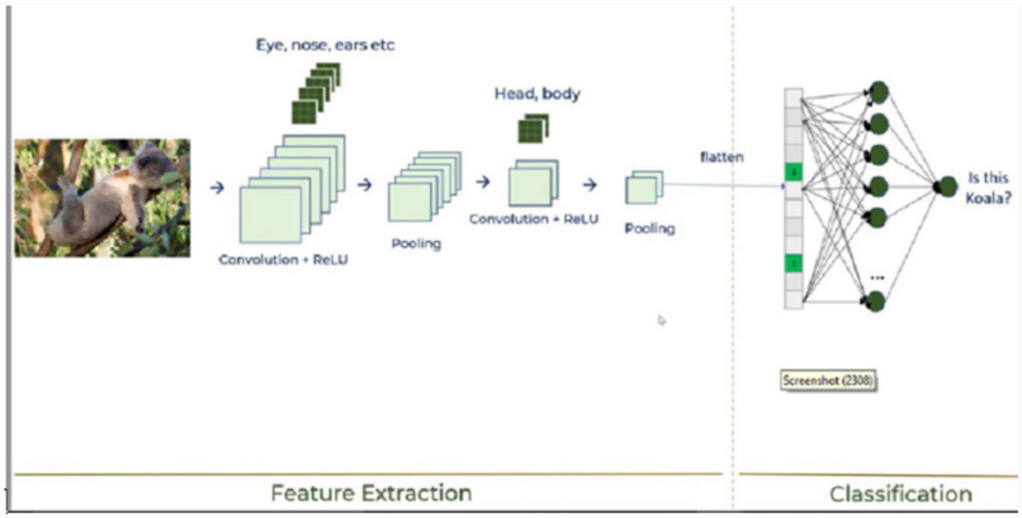

Figure 5: The architecture of convolution neural network CNN [16].

A Convolution layer

It is the most important layer in CNN, which distinguishes it from other layers that contain one or more filters. It is called the convolutional kernel. The filters of the convolution layers are a two-dimensional array, usually 3 * 3 or 5 * 5, and can be 1 * 1. The convolution layer creates maps Features that highlight the features of the original image. This is done by multiplying the kernel by the input matrix [17].

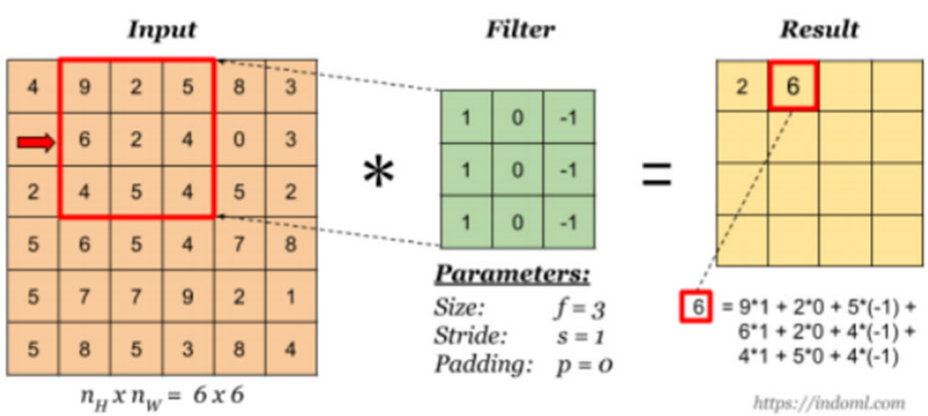

Figure 6: In a convolutional layer, element-wise matrix multiplication, and summation of the results onto feature map [18].

B. Pooling layer

It is the second layer in CNN and it comes after the convolutional layer and it is called the implementation of the reduction process and this is called pooling and it means the process of reducing the size of each dimension of the output maps and this reduces the number of parameters that shorten the training time, this layer preserves the maps Input and Output It also merges pixels adjacent to a specific area of the image [19].

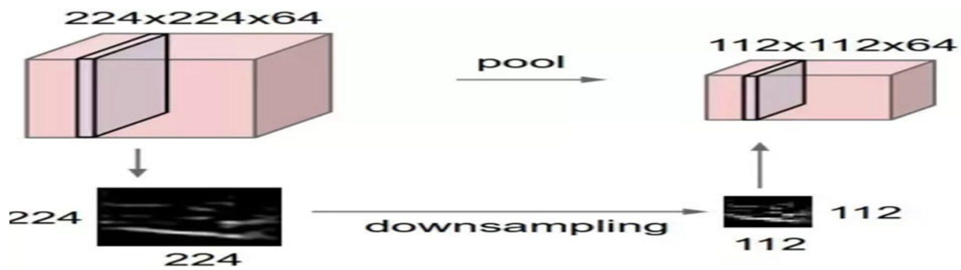

Figure 7: shows pooling operations [20].

C. Fully connected layer

It is the output of the final pooling layer or is the output of the final layer of a CNN [20]. used as a classification layer where the result of each feature class extracted from the convolutional layers in the previous steps is calculated, the FC layer usually takes advantage of the activation function Sotfmax to classify the input appropriately [19].

3.2.3 GoogleNet

It is a convolutional neural network known as (Inception-V1), with a depth of 22 layers and 40 million parameters that provides scattered, multi-domain information It was proposed by GoogleNet company 2014 (ILSVRCI14) and developed by researchers at Google to solve computer vision tasks such as detecting objects and classifying images Google Net managed to achieve the Least top 5 error rate of 6.67% [21]. I also got the best results, and it was very close to the human level. Google Net is a non-sequential network. We can expand the depth as well as the width without placing a large burden on the computer[22]. And the main goal of it is to obtain high accuracy [23].

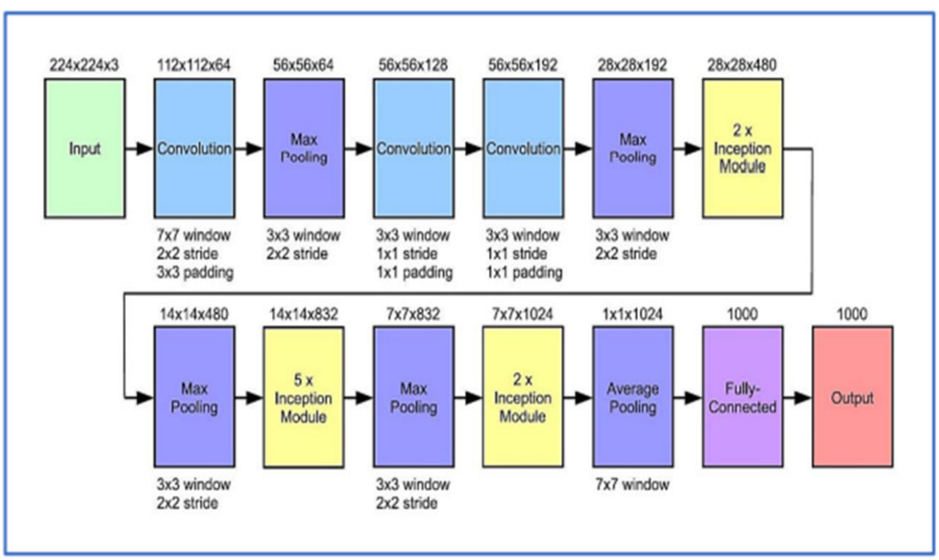

Figure 8: GoogleNet model architecture

3.3 The classification stage

At this stage, the frame of the lips with correct and incorrect reading will be classified by matching the training frames with the test frames. If the matching is done, ___________________this indicates that the reading is correct for the digits in Figure 9 , but if the matching is not done, this indicates an error rate in the work of the network as in figure 10.

Figure9: Example test frame matching training frame Figure10: Example test frame not matching training frame

4 . Dataset Description.

We picked up our database from the site https://ibug-avs.eu/ where personal were in front of the camera Directly when reading the decimal digits, the number of videos was 35 videos for seven personal , 5 of whom were males and 2 females, and the number of the image frame was 21,501 lip image.

4.1 Experimental Details.

We have inserted a number of video lip image frames, which were 21,501 frames, into the convolutional neural network (GoogleNet), where it passed through the layers of the neural network and the training process was 80% (17,200 frames) and then the testing process 20% (4300) frames of the total samples that were randomly selected.

Training and test of Google Net based on MATLAB R2020 language to easily implement the code of CNN. A computer Lenovo that has specifications such as Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz 2.30 GHz for CPU, 8.00 GB Windows 10 Pro of RAM, and 64-bit operating system, x64-based processor.

4.2 performance of GoogleNet

Our proposed method, GoogleNet, achieved an accuracy rate of 87% by using 35 videos for seven personal, five males and two females, and the Google Net network achieved success due to difficulties and from the presence or absence of make-up to females, as well as the presence or absence of the mustache to males, and this was the database from the site https://ibug-avs.eu/.

The following table show the experimental results of the proposed system Table1: Illustrates the main classification criteria for the proposed model.

The measure	ration
Accuracy	87%
Precision	87.9%
Recall	87.4%
Specificity	98.2%

Reference	Number of persons extracted from the database	Number of images in database	Accuracy	Recognition
[7]	10	1200	60%	_
[2]	15	_	65%	_
Our proposed	7	21501	87%	83.5%

Table 2: The recognition rate compared with previous studies

5. conclusion

Our proposed system went through several stages to reach the recognition of the pronunciation of digits by the movement of the lips, the first of which is our use of the Viola Jones algorithm to detect the face, then the mouth area, and then cut out the region of interest (the mouth) and store the lip frame in a temporary folder and then insert the frame into the GoogleNet to extract the features Then classify it. One of the advantages of our proposed system is that it gave us a correct classification, despite the differences in the structure and design of the lips from one person to another.

References :

[1] A. Nagzkshay Chandra Aarkar, "ROI EXTRACTION AND FEATURE EXTRACTION FOR LIP READING OF," vol. 7, no. 1, pp. 484–487, 2020.

[2] R. Bowden, "Comparing Visual Features for Lipreading," no. September 2016.

[3] A. Garg and J. Noyola, "Lip reading using CNN and LSTM," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Jan, p. 3450, 2017.

[4] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Comput. Vis. Image Underst.*, vol. 173, pp. 76–85, 2018, doi: 10.1016/j.cviu.2018.02.001.

[5] A. Mesbah *et al.*, "Lip Reading with Hahn Convolutional Neural Networks moments To cite this version : HAL Id : hal-02109397 Lip Reading with Hahn Convolutional Neural Networks," *Image Vis. Comput.*, vol. 88, pp. 76–83, 2019.

[6] A. H. Kulkarni and D. Kirange, "Artificial Intelligence: A Survey on Lip-Reading Techniques," *2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019*, Jul. 2019, doi: 10.1109/ICCCNT45670.2019.8944628.

[7] A. Bang *et al.*, "Automatic Lip Reading using Image Processing," vol. 2, no. 02, pp. 279–280, 2014.

[8] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," *2015 11th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2015*, no. June 2016, 2015, doi: 10.1109/FG.2015.7163155.

[9] Mr. Befkadu Belete Frew, "Audio-Visual Speech Recognition using LIP Movement for Amharic Language," *Int. J. Eng. Res.*, vol. V8, no. 08, pp. 594–604, 2019, doi: 10.17577/ijertv8is080217.

[10] Y. Lu and H. Li, "Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory," *Appl. Sci.*, vol. 9, no. 8, 2019, doi: 10.3390/app9081599.

[11] K. K. Sudha and P. Sujatha, "A qualitative analysis of googlenet and alexnet for fabric defect detection," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, pp. 86–92, 2019.

[12] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign Language Recognition Using Convolutional Neural Networks BT - Computer Vision - ECCV 2014 Workshops," pp. 572–578, 2015, [Online]. Available: https://core.ac.uk/download/pdf/55693048.pdf.

[13] C. D. Mccaig, "Electric Fields in Vertebrate Repair. Edited by R. B. Borgens, K. R. Robinson, J. W. Vanable and M. E. McGinnis. Pp. 310. (Alan R. Liss, New York, 1989.) $69.50 hardback. ISBN 0 8451 4274," *Exp. Physiol.*, vol. 75, no. 2, pp. 280–281, 1990, doi: 10.1113/expphysiol.1998.sp004170.

[14] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361. pp. 255–258, 1995, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.9297&rep=rep1&type=pdf.

[15] A. Patil and M. Rane, "Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition," *Smart Innov. Syst. Technol.*, vol. 195, pp. 21–30, 2021, doi: 10.1007/978-981-15-7078-0_3.

[16] "Supervised Deep Learning Algorithms _ Types and Applications." .

[17] W. M. Learning and P. Kim, *MATLAB Deep Learning*. .

[18] S. Tammina, "Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images," *Int. J. Sci. Res. Publ.*, vol. 9, no. 10, p. p9420, 2019, doi: 10.29322/ijsrp.9.10.2019.p9420.

[19] M. Z. Alom *et al.*, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," 2018, [Online]. Available: http://arxiv.org/abs/1803.01164.

[20] H. J. Jie and P. Wanda, "Runpool: A dynamic pooling layer for convolution neural network," *Int. J. Comput. Intell. Syst.*, vol. 13, no. 1, pp. 66–76, 2020, doi: 10.2991/ijcis.d.200120.002.

[21] N. A. Muhammad, A. A. Nasir, Z. Ibrahim, and N. Sabri, "Evaluation of CNN , Alexnet and GoogleNet for Fruit Recognition," vol. 12, no. 2, pp. 468–475, 2018, doi: 10.11591/ijeecs.v12.i2.pp468-475.

[22] F. Altenberger and C. Lenz, "A Non-Technical Survey on Deep Convolutional Neural Network Architectures."

[23] C. Szegedy, "Going deeper with convolutions," no. January 2017, 2015, doi: 10.1109/CVPR.2015.7298594.

[24] S. L. Wang, A. W. C. Liew, W. H. Lau, and S. H. Leung, "An automatic lipreading system for spoken digits with limited training data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 12, pp. 1760–1765, 2008, doi: 10.1109/TCSVT.2008.2004924.