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.12.02.30

Automatic Lip reading for decimal digits using ResNet50 Model

Kwakib Saadun Naif[1] **Prof. Dr. Kadhim Mahdi Hashim**[2]

[1,] Computer Science Department, College of Education for pure Sciences, University of Thi-Qar , Iraq.

[2]Imam Ja,afar AI-Sadiq University.

Received 3/5/2022 Accepted 2/6/2022 Published 01/12/2022

Abstract:

Lip reading is a method to understand speech through the movement of the lips, as audio speech is not inclusive of all Categories of society, especially the hearing impaired or people in noisy environments. Lip reading is the best and alternative solution to this problem. Our proposed system solves this problem by taking a video of the person speaking with digits. Then the pre-processing process is carried out by Viola Jones algorithm, by cutting the video into a sequential frame, then detecting the face, then the mouth, deducting the mouth region of interest(ROI), and inserting the mouth frame into the convolutional neural network (ResNet50), where the results are classified and the test frames is matched with the training frames if it is done Matching, the network is working correctly and the correct digit is spoken. But if the test frame is not matched with the training framework, then there is an error rate in the network's work and there is an error rate in the network. For that, we used a standard database to pronounce the digits from 0 to 9, and we took seven speaking people, 5 males and 2 females, and we got an accuracy of 86%.

Keyword: CNN, ResNet50 and viola jones

2. Introduction

Lip reading is a method of interpreting speech by watching only the movements of the lips without examining the auditory signal [1]. Has been documented since the sixteenth century and is often used for people with hearing impairments and is also used as an aid to understand speech fluently [2]. That reading lips is of great importance as You gave more than one concept Including it is known as the decoding process from the movement of the speaker's mouth [3]. And also knows that it is the prediction of words and phrases from the video only without any audio signal [4]. The movement of the lips can be considered as a biometric because it is very difficult to imitate or shape the movement of the lips [5]. That reading the lips is not an easy thing because it faces many challenges due to differences in inputs such as skin colors White and brown, facial features, languages, speaking ability and intensity, as these difficulties can be reduced by using more visual input data collected. Automatic lip reading mainly includes face detection, lip positioning, feature extraction, classifier training, and decimal recognition through lip movement [6].

The lip-reading process is applied to the levels of sentences, digits, alphabets and words where the success rate of lip-reading is directly related to the correct segmentation of the image frame for an region of interest [8].

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.

3. Related works

Automatic Lip reading is used to interpret or understand speech without the hearing it, a technique mastered by human with hearing difficulties. In this Paragraph, we review previous work that is related to our work. Zhao et al. (2009) in [9]. They used method local spatiotemporal descriptors an automatic lip reading, they represented the isolated words by extraction spatiotemporal Local Binary Patterns (LBP) from mouth area, this method obtained an accuracy of 58.85%. R. Bowden et al. (2009) in [2]. They proposed compared various features for Automatic lip reading, including four types of Active Appearance model (AAM) features, 2D DCT features, sieve features and Eigen lip features.in general, AAM features with appearance outperform other types of feature, Which means that appearance is more useful than the shape .The proposed achieved accuracy rate 65%. Ngiam et al. (2011) in [10]. They used a deep learning method to recognize speech using Video and audio, where they used two data set Cuave to digits and AVletters where their accuracy was 68.7%and 64.4%. A.Rekik et al.(2016) in [11]. They proposed an adaptive system for automatic lip reading. They used three different classifications to obtain three different results Support Vector Machine (SVM), Hidden Markov Model(HMM) and K-Nearest Neighbors(KNN) It was noted that SVM is the best, with an overall accuracy of 71.15%, HMM with an accuracy of 65.35% and no KNN with an accuracy of 50.58%. Petridis and Pantic (2016) in [12]. They present a method based on Deep bottleneck feature extraction directly, and they trained a model using Long Short Term Memory(LSTM) this method achieved 58.1%. J.Chung (2017) in [13]. They presents a method called spell, attend, listen, and watch, that goal to convert mouth motion videos to characters. using both LSTM and CNN to recognize the spoken words, They obtained on an accuracy of 76.2%. Chung and Zisserman (2017) in [14]. They shown that convolution neural network architectures can be used to classify temporal sequences, they achieved an accuracy of 65.4%.

4. The Proposed System

In our proposed system, we used a video clip of the digit decimal speaking people by installing the camera in front of the person speaking the digits, where the camera captures a video clip of the person and then segmentation the video into a sequential frame and then enters this frame into the Viola Jones algorithm to detect the face first and then detect the mouth area of interest and then Subtracting the mouth into a frame and inserting the mouth frames into the convolutional neural network (ResNet50), where in the first layer the features are extracted and then passed these features to the fully connected layer to classify the correct or incorrect mouth frame.

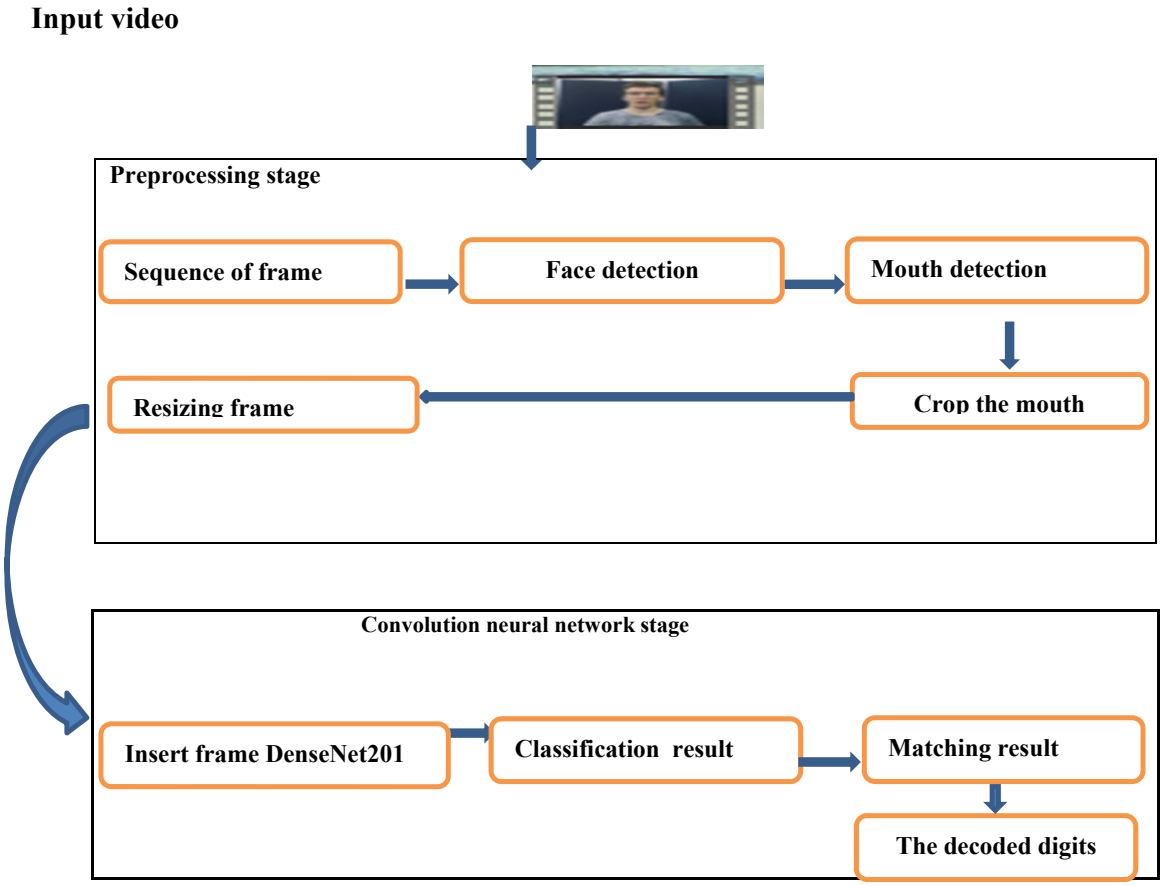

Figure 1: The block diagram of the proposed system automatic lip reading

4.1 The preprocessing stage

It is the stage that facilitates the work of the following stages and includes operations related to extracting an region of interest (ROI) and performing arithmetic and quantizing operations on the image as ROI includes cropping, resizing, rotation and translate [15].

4.1.1 Face and mouth detection

After capturing a video of the person speaking the digits, it is stored in a temporary folder and the face and mouth region are detected, and this is done by Viola Jones algorithm [16]. As shown in Figure (2) Detecting the face and mouth region

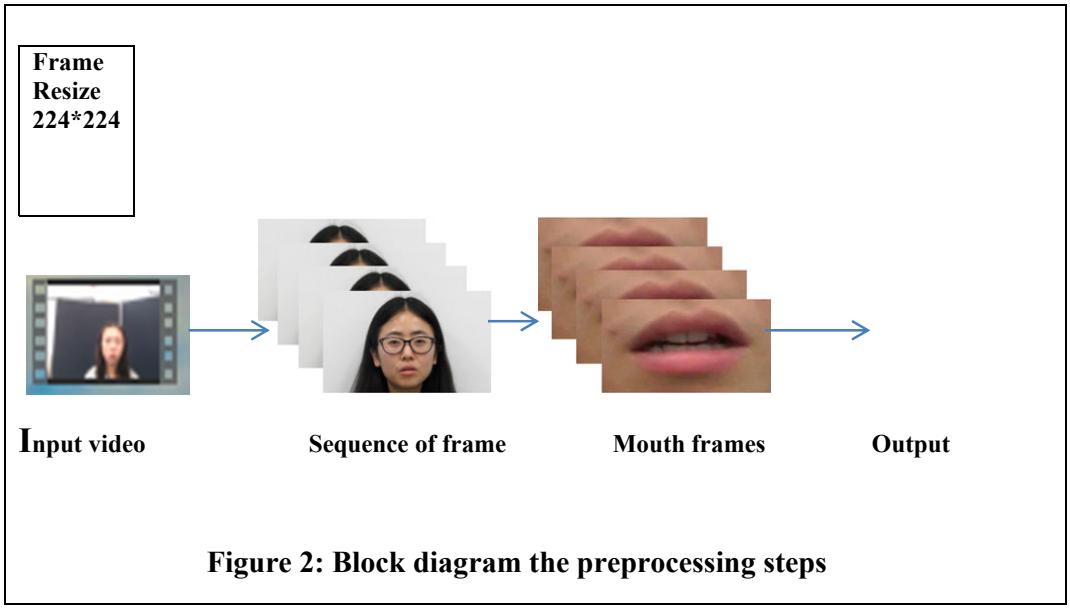

Figure 2: Block diagram the preprocessing steps

Figure 3: Sample for face and mouth region detection by viola jones algorithm.

4.1.2 Crop mouth image from the frame

It is the process of cutting out areas of the image that are of interest called a sub-image and that benefit us in the analysis [17]. At this stage the mouth is identified and

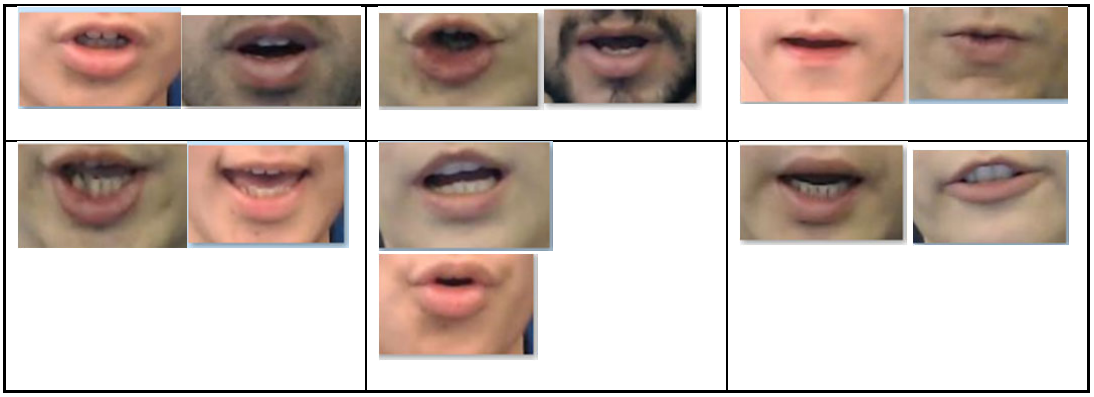

Figure 4: Examples for cropping the mouth

4.1.3 Resize mouth image to 224×224.

In ResNet50 network, the size of the input image must be 224 * 224, and this means that we need to modify the size of the input image of the CNN to reduce or enlarge it to reach the required size and the mouth image is the cutout area.

4.2 Convolution neural network stage

In this stage, we use the Convolutional Neural network, after completing the pre-processing of the specified image, where the movements of the lips must be matched with the spoken number and to know whether the pronunciation is correct or not. This is done using the proposed method ResNet50 to classify the state of the lips.

4.2.1 convolution neural network layers

It is a specific type of artificial neural network that uses sensory perception [18]. Also known as (conNet) [19]. It contains sequential layers, and each layer has its own task, as the CNN consists of three layers [20]:

A. Convolution layer

It is the most important component of the CCN architecture and it is a set of automatic kernels also called filters that are executed with the input image to create an output feature map [19]. Also called feature maps [21]. And it is a generalized linear model of the basic local image [22]. Where the convolution layer Creates feature maps that highlight the unique features of the original image. Inserts the image through convolution filters into the feature map.

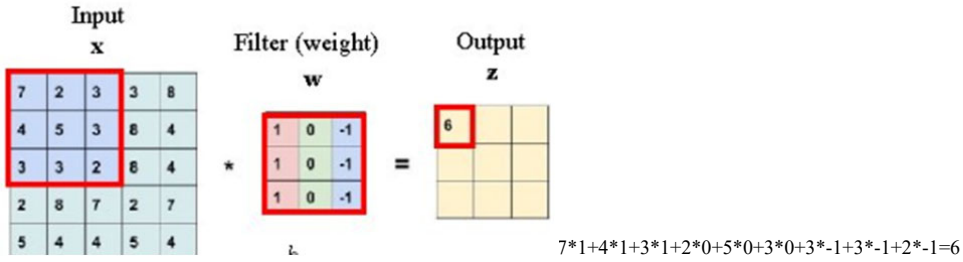

Figure 5: Illustrating an example that shows convolution operation [23].

B- Pooling layer

It is the process that comes immediately after each convolution process and is used to subsample the feature maps of the largest size and reduce them to the feature maps of smaller size, because the feature maps preserve the most common information during the downsampling process in each step of the pooling. This pooling process is done by defining the collected area, and there are many types of pooling techniques, including: max pooling, tree pooling, min pooling, average pooling, gated pooling, etc. Max Pooling is the most common and mostly used pooling technique [24].

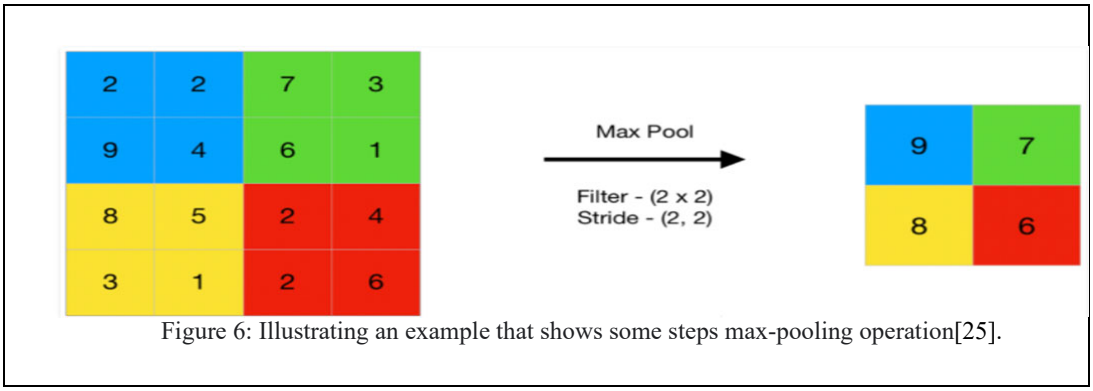

Figure 6: Illustrating an example that shows some steps max-pooling operation[25].

C - Fully connected layer

It is the last layer of the convolutional neural network and it comes directly after the pooling layer, which is the layer that feeds the neural networks forward and its inputs are the outputs of the final pooling layer or the convolutional layer that has been regularization and then inserted into this layer. The goal of this layer is to make the convolutional network more able to classify images [22].

4.2.3 ResNet50

Acronym Residual network is a specific type An important convolutional neural network that was introduced in 2015 by kaiming He and join sun and others consists of fifty layers that can receive more than one million trained images from a database and can also classify the trained images into 1000 object classes [18] .

The goal is to solve a complex problem by stacking some extra features in the network and this leads to an increase in performance and accuracy because the increase in the number of layers leads to gradual learning of more complex feature [26].

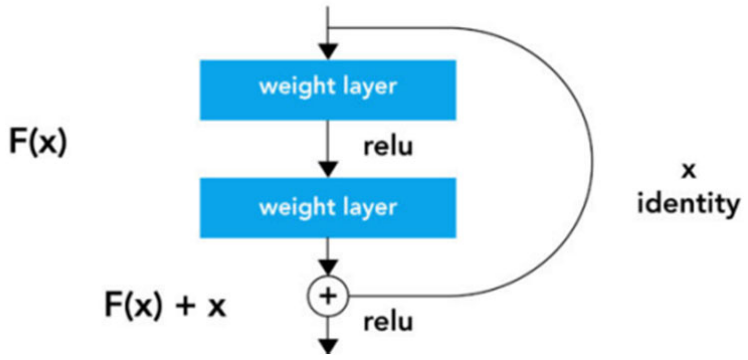

Figure 7: A residual learning building block [26].

4.3 The classification stage

To classify the picture frame to identify the decimal digits by lip movement, we use the convolutional neural network ResNet50 by inserting a frame into the network, where the network calculates the features and classified them in the classification process if the test frame is correctly classified and matched with the training frames as in Figure (8) This indicates that the network is working correctly. But if the matching is not done, this indicates that the network work was done incorrectly, as in Figure (9), and there is an error rate.

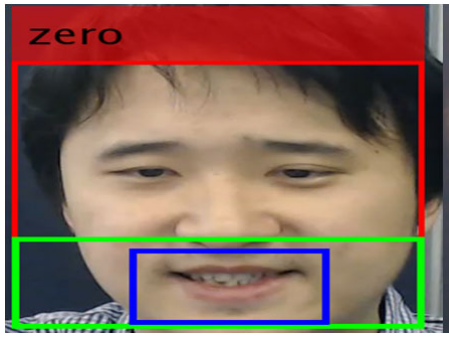

Figure 8: Test frame match training frame

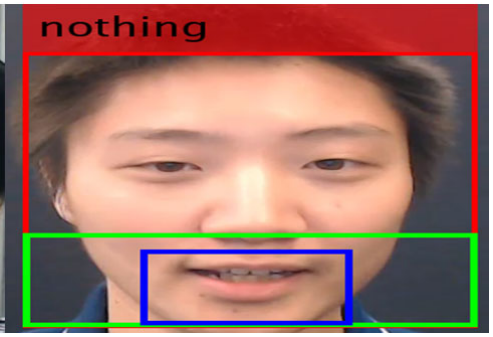

Figure 9: Test frame not match training frame

5. Dataset Description

Our database consists of 35 videos of seven people of different nationalities and languages, 5 males and 2 females, and each person pronounces the digits from 0 to 9 five times at random. And we find 21501 frames and this was done by installing a camera in front of the digits speaking. The site from which we captured the videos is https://ibug-avs.eu/

5.1 Experimental Details

We inserted the lip frame into the convolutional neural network ResNet50, which numbered 21501 frames. It will pass through the training phase, which was 80% (17200) images from the training images samples randomly. Then the testing phase starts after the training process which was 20% (4300) images were selected from The sum of randomly selected samples. and achieved an accuracy of 86%.

Training and test of DenseNet201 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.

The following table show the experimental results of the proposed system

Table1: Illustrates the main classification criteria for the proposed model

Table1: Illustrates the main classification criteria for the proposed model

Accuracy	**86%**
Precision	**86.6%**
Recall	**87.7%**
Specificity	**98%**
F-measure	**87%**

Table 2: The recognition rate compared with previous studies

Reference	Number of persons extracted from the database	Number of images in database	Accuracy	Recognition
[10]	36	_	68.7%	_
[9]	1000	_	58.85%	_
Our proposed	7	21501	86%	83%

6- Conclusion

We used the proposed method ResNet50 to determine the lip reading of decimal digits from 0 to 9 for people who speak different languages and of different nationalities as well. We used the Viola Jones algorithm to detect the face, then detect the mouth area and deduct the mouth area and store it in a temporary folder. Where we encountered some difficulties in the Viola Jones algorithm, as it does not determine the mouth area correctly and determines the eye instead, because there is some similarity between them. Our proposed method gave us good results with high accuracy, and its accuracy reached 86%.

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. 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.
[4] 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.
[5] 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.
[6] 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.
[7] B. O. Li, "Deep Learning-Based Automated Lip-Reading :," vol. 9, 2021, doi: 10.1109/ACCESS.2021.3107946.
[8] T. As for training and classification, this is done with the help of artificial neural networks and A. Basturk, "Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models," vol. 7, no. 2, pp. 195–201, 2019, doi: 10.17694/bajece.479891.
[9] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading with local spatiotemporal descriptors,"

IEEE Trans. Multimed., vol. 11, no. 7, pp. 1254–1265, 2009, doi: 10.1109/TMM.2009.2030637.
[10] J. Ngiam and A. Y. Ng, "Multimodal Deep Learning," 2011.
[11] A. Rekik, A. R. A. Ben-Hamadou, and W. Mahdi, "An adaptive approach for lip-reading using image and depth data," *Multimed. Tools Appl.*, vol. 75, no. 14, pp. 8609–8636, 2016, doi: 10.1007/s11042-015-2774-3.
[12] S. Petridis and M. Pantic, "DEEP COMPLEMENTARY BOTTLENECK FEATURES FOR VISUAL SPEECH RECOGNITION Stavros Petridis Imperial College London Dept . of Computing , London , UK Maja Pantic Imperial College London / Univ . of Twente Dept . of Computing , UK / EEMCS , Netherlands," *Icassp 2016*, pp. 2304–2308, 2016.
[13] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3444–3450, 2017, doi: 10.1109/CVPR.2017.367.
[14] J. S. Chung and A. Zisserman, "Lip reading in the wild," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10112 LNCS, pp. 87–103, 2017, doi: 10.1007/978-3-319-54184-6_6.
[15] I. Horace H-S, *Digital image processing and computer vision*, vol. 8, no. 3. 1990.
[16] H. VAIBHAV, "Face Identification using Haar cascade classifier," *Medium*, pp. 1–5, 2020, [Online]. Available: https://medium.com/geeky-bawa/face-identification-using-haar-cascade-classifier-af3468a44814.
[17] J. Shemiakina, A. Zhukovsky, I. Konovalenko, and D. Nikolaev, "Automatic cropping of images under projective transformation," no. March, p. 117, 2019, doi: 10.1117/12.2523483.
[18] T. Bezdan and N. Bačanin Džakula, "Convolutional Neural Network Layers and Architectures," no. July, pp. 445–451, 2019, doi: 10.15308/sinteza-2019-445-451.
[19] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, *Fundamental concepts of convolutional neural network*, vol. 172, no. January. 2019.
[20] J. Wu, "Introduction to Convolutional Neural Networks," *Introd. to Convolutional Neural Networks*, pp. 1–31, 2017, [Online]. Available: https://web.archive.org/web/20180928011532/https://cs.nju.edu.cn/wujx/teaching/15_CNN.pdf.
[21] W. M. Learning and P. Kim, *MATLAB Deep Learning*. .
[22] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," *Neurocomputing*, vol. 323, pp. 37–51, 2019, doi: 10.1016/j.neucom.2018.09.038.
[23] "Convolutional Layer について（画像からの特徴量抽出）." [Online]. Available: https://qiita.com/icoxfog417/items/5fd55fad152231d706c2.
[24] "Enhanced Reader." .
[25] GeeksforGeeks, "CNN | Introduction to Pooling Layer," *Https://Www.Geeksforgeeks.Org/Cnn-Introduction-To-Pooling-Layer/*. p. duction to Pooling Layer, 2022, [Online]. Available: https://www.geeksforgeeks.org/cnn-introduction-to-pooling-layer/.
[26] C. Vision, H. Resnet, R. Block, H. Resnet, and U. Resnet, "What is Resnet or Residual Network | How Resnet Helps? Introduction to Resnet or Residual Network," pp. 1–8, 2020, [Online]. Available: https://www.mygreatlearning.com/blog/resnet/.