Journal of Information Hiding and Multimedia Signal Processing ©2020 ISSN 2073-4212
Ubiquitous International Volume **11**, Number **3**, September **2020**

Visual Speech Recognition of Lips Images Using Convolutional Neural Network in VGG-M Model

Zhi-Ming Chan, Chee-Yong Lau, Ka-Fei Thang

School of Engineering,
Asia Pacific University of Technology & Innovation,
Kuala Lumpur, Malaysia
laucheeyong@staffemail.apu.edu.my

Received July 2019; revised February 2020

Abstract. *This paper presented a Visual Speech Recognition of lips images using Convolutional Neural Network in VGG-M model. A camera is used to record the video data which to be processed. The video recording is then transmitted for prediction. A variety of models for predicting words from video data without audio data are presented. The dataset used in this project is self-collected which consists of 15 speakers speaking digit 0 to 9 in 10 repetitions. The frames collected for each video are limited to 50 to accommodate different speaking speed. The data is then preprocessed by cropping the speakers' lip in all frames to remove redundant information. Due to hardware limitation, only 2D convolution architecture was explored. Each frame of a video is concatenated into a single image to be fed into the training model. EF3 model is used as a baseline and other types of architectures were also explored. Parameters of the chosen architecture are also tuned to further improving the test accuracy such as kernel size, learning rate, optimizer, etc. In short, this project has achieved a validation accuracy of 87% for seen test and 30% test accuracy for the unseen test.*

Keywords: Lip-reading; Visual Speech Recognition; Convolutional Neural Network

1. **Introduction.** As technology advances, the physiological biometric system is getting harder to be protected. For example, a face recognition system can be deceived by placing the image of a user's face in front of the camera. To solve the problem of cyber hacking, a multimodal solution is proposed by combining several biometric features which is more robust compared to the use of a single biometric feature. Several advanced biometric technologies are emerging on the market such as brainwave biometrics, body odor recognition, handgrip recognition and system that even utilize ear pattern and body salinity. However, they are computationally expensive and requires special equipment. Therefore, this project is aimed to utilize a behavioral feature from the user and is inspired by the work of [1] which is known as lip-reading.

Lip reading is a process of recognizing what a person is saying by analyzing the visual signal of the person?s lip. Visual signal by means is the changing image of the mouth when the mouth is moving during the speech. It remains a very difficult task since the technique requires years of practice with a strong foundation of the speaking language itself. Even professional lip-readers could only see 50% of the speech and guessing is used for the remaining 50%. Different speakers have different speaking speed and accents that made the task even more challenging [1].

There were many researchers tried to implement image processing techniques to solve the problem of lip reading and this new field is known as Visual Speech Recognition

(VSR). However, there are no perfect image processing techniques that can perform well prediction on the uttered words until the rise of Artificial Intelligence (AI). Deep learning is getting popular in recent years due to its excellent performance in solving computer vision problem. Therefore, this project is aimed to tackle the lip-reading problem using deep neural network (DNN).

2. **Related Work.** Salma & Archana [2] had used geometrical features of the lip together with Support Vector Machine (SVM) to predict the phrase uttered by test subjects. 4 key points were placed on the lip once the lip of a test subject is detected. Two features were then calculated based on those 4 points and being used as inputs for SVM classifier. Overall the accuracy of the entire system is about 65.6%.

Terissi et al. [3] proposed a system that utilized wavelet-based features and Random Forests (RF) classification. Discrete Wavelet Transform (DWT) was used to extract image-transformed based features from the lip image. The proposed system achieved 65.68% in AV-CMU, 64.89% in AVLetteres, and 78.28% in AV-UNR.

Sunil [4] also proposed a system that utilized DWT. Artificial Neural Network (ANN) and SVM were chosen in the proposed system. Backpropagation Neural Network (BPNN) was chosen for ANN with sigmoid nodes and 20 hidden layers. Sequential Minimal Optimization (SMO) was chosen for SVM classifier. The proposed system compared the performance of 2D-DWT and 3D-DWT using BPNN and also the performance of BPNN and SVM. The results showed that 3D-DWT achieved 82.5% and 81.25% and 2D-DWT achieved 73.43% and 79.2% in CUAVE and TULIPS respectively. BPNN performed better with an accuracy of 82.5% compared to SVM with 78.56% in CUAVE database.

Kumaravel [5] had researched using History of Oriented Displacements (HOD) for lip-reading purpose. To deal with the problem of different words with different durations, HOD was used to convert the features into fixed-length vectors. Temporal pyramid approach was also used to prevent losing the temporal information of the video. SVM with linear kernel was used as the classifier of the proposed system and the system was evaluated with CUAVE database with 5-fold validation. The average result of evaluating the system with 33 speakers was 81.03%.

Hassanat [6] also proposed another system that utilized hybrid features. There were a total of 8 features extracted for each uttered word. K-nearest neighbor (KNN) method was used as the classifier in this paper to determine the probability of each class according to the testing samples. Two experiments were conducted to determine VSR was a speaker-dependent or speaker-independent problem. The results showed that speaker-dependent testing achieved 76.38% while speaker-independent testing achieved 33%.

Aris et al. [7] proposed a system that was expected to be applicable in real-time. They utilized the method of frame difference to obtain the dynamic features of the lip. For the classifier, Multi-Layer Perceptron (MLP) and SVM were both used to compare their results. The proposed system was also compared to the 2D-DCT to evaluate the performance of the system. The accuracy and area under the curve were used as the measure of performance. The proposed system had achieved 0.9993 in the area under the curve and 96.5% accuracy while the 2D-DCT had achieved 0.9978 in the area under the curve and 94% accuracy.

Neeru [8] utilized geometrical features of the dynamic images to be the input of neural network classifier. Neeru (2016) utilized Learning Vector Quantization (LVQ) neural network which minimized the quantization error. The results showed that the proposed system had achieved 97% accuracy.

Sunil [9] came up with a system that utilized the Localized Active Contour Model (LACM) and Hidden Markov Model (HMM). HMM of 3 states were used to represent the change in lip's height, width, and area. Additional 2 states were added to represent the amount of change in lip movement. The results showed that the recognition rate of 3 states model achieved 66.3% in CUAVE and 64.7% in in-house while 5 states model achieved 78.33% in CUAVE and 76.6% in in-house.

Kuniaki et al. [10] conducted lip-reading research utilizing DNN. A seven-layered Convolutional Neural Network (CNN) was proposed to recognize phonemes from the sequence of lip images. The output of the CNN was then used as input for Gaussian Mixture Model (GMM) HMM. The results stated that features extracted by CNN achieved the highest recognition rate with 21%, 35% and 37% in 8, 16, 32 GMM components respectively. Overall, the CNN approach could achieve a 58% accuracy of recognizing the 40 phonemes.

Another similar type of research had been conducted by Yiting et al. [11]. However, the only difference is the utilization of dynamic images. Two CNN architectures were being created with the one having 1 convolutional layer and pooling layer while the other having 2 convolutional layers and pooling layers. The CNN1 with 71.76% and 58.94% accuracy in dynamic and static images respectively were then being used to compare with features extracted using Discrete Cosine Transform (DCT). CNN1 still outperformed DCT features by 19.91%.

Chung and Zisserman [12] came up with 4 different architectures utilizing 2D and 3D convolution. The models were evaluated with the dataset and Multiple Towers (MT) of 2D convolution had the highest accuracy which is 61.1%. This model was then used to perform prediction on OuluVS1 and OuluVS2 dataset and achieved 91.4% and 93.2% accuracy respectively.

Garg et al. [13] proposed using CNN and Long-Short Term Memory (LSTM). Concatenation of images into one image to be fed into the training process was proposed. The extracted features were then passed to LSTMs for extracting temporal information. Several models were being suggested and the best model achieved 76% of validation accuracy.

Gutierrez and Robert [14] also proposed CNN and LSTM for lip-reading classification. 4 models were proposed which are CNN and LSTM baseline model, deep layered CNN, and LSTM model, ImageNet pre-trained VGG-16 features with LSTM model and a fine-tuned VGG-16 withLSTM model. The best model could achieve validation accuracy of 7%. However, the validation accuracy of unseen test wavered around 10%.

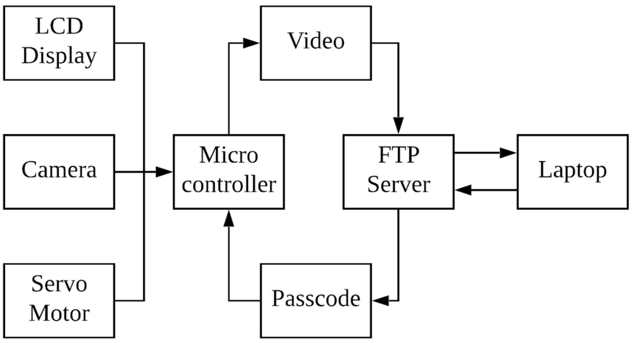

Figure 1. Block diagram of the system

3. **Proposed Methodology.** In this project, the intention was to build a security system employing lip-reading passcode detection. An LCD is used to show the Graphical User Interface (GUI) to allow end-users to record the movement of their lips when uttering the passcode. Real-time input images captured by the camera is hence fed into the

microcontroller. The sequence of these images is stacked together to become a video and be sent for inferencing via file transfer protocol (FTP) server. Initially, the trained deep neural network was planned to be operated within the microcontroller itself. However, the computation speed of Raspberry Pi is slow compared to the traditional laptop and therefore would influence the performance speed of the entire system. Once the trained deep neural network successfully predicted the passcode from the videos, the generated passcode is sent back to the microcontroller for password comparison.

A: *Dataset*

Based on the literature review, most of the researchers utilized CUAVE database and in-house database for training, validating and testing process. CUAVE database is a good option for VSR purpose but requires permission before using it. Therefore, an in-house database was used by recording videos of speaker uttering the words. According to the specification of the recorded videos used by past researchers, the specification of the video recording used in this project is summarized in Table 1.

Table 1. Specifications of Dataset

Aspect	Specifications
Pixels	Minimum 640x480 pixels
Contents	Digits from ‘0’ to ‘9’
Number of Speaker	Minimum 15
FPS	Minimum 30
Frames Per Video	50
Categories	Facial appearance and skin color
Background	Single Color
Number of Repetitions	10 times for each digit

The above database is constructed according to the specifications of MIRACL-VC1 except that depth information is not included and digits were being uttered. The frames per video are limited to 50 to compromise for people with different speaking speed.

B: *Architecture*

The neural network architecture used in this project is shown in Figure 2. It follows the VGG-M model which is the original version of EF-3.

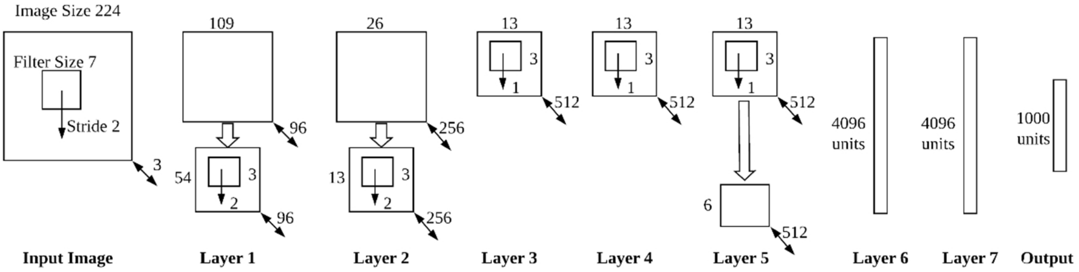

Figure 2. The convolutional network architecture used in this project

The reason that VGG-M is being utilized in this work is due to (a): The size of the lips image is usually fixed. It means that a reasonable depth of deep model should be considered. The depth of VGG-M is sufficient to extract the local and global features of lips image. (b): This system should not require a high computational facility. Based on the scale of training model of VGG-M, it is suitable to be employed in this purpose. And (c):

VGG-M has been proven to have exceptional performance and promising transferability. It has been widely adopted in various vision task and biometric recognition [15].

As illustrated in Figure 2, the VGG-M architectural in this study contains eight (8) layers. The first five (5) are convolution layers and the remaining are fully connected layers. The first and the second layer are followed by response normalization layer and stacked by max-pooling layer. This is repeated in the fifth layer as well. The Rectified Linearity Unit (ReLU) is applied to each layer. The detailed summary is shown in Table 2.

The CNN training was following [16] using Stochastic Gradient Descent (SGD) with momentum 0.9 and initial learning rate of 0.01. The layers are initialized from a Gaussian distribution with a zero mean and variance equal to 0.01. This architecture is similar to [17]. It is categorized by the stride and smaller receptive field in the first convolutional layer. To keep the computational time reasonable, the second convolutional layer was employing larger stride.

Table 2. VGG-M Model

Name	Type	Filter Size\Stride	Output Size
Conv1	Convolution	7x7\2	109x109x96
Relu1	RELU		109x109x96
Norm1	LRN		109x109x96
Pool1	Max-Pooling	3x3\2	54x54x96
Conv2	Convolution	5x5\2	26x26x256
Relu2	RELU		26x26x256
Norm2	LRN		26x26x256
Pool2	Max-Pooling	3x3\2	13x13x256
Conv3	Convolution	3x3\1	13x13x512
Relu3	RELU		13x13x512
Conv4	Convolution	3x3\1	13x13x512
Relu4	RELU		13x13x512
Conv5	Convolution	3x3\1	13x13x512
Relu5	RELU		13x13x512
Pool5	Max-Pooling	3x3\2	6x6x512
Fc6	Fully-Connected		4096x1
Relu6	RELU		4096x1
Fc7	Fully-Connected		4096x1
Relu7	RELU		4096x1
Fc8	Fully-Connected		1000x1

4. Implementation.

A: *Dataset*

In this project, 15 test subjects were recorded according to the specifications. The lips were cropped into videos of 112x112 pixels as shown in Figure 3. The 2D vs 3D convolution test is displayed in a later section. For 2D convolution, the frames of a video were concatenated into an image of 224x224 pixels as illustrated in Figure 4.

5. **Results.** Several tests were done to improve the accuracy of the model and they were 2D vs 3D convolution test, architecture test, environment test, and unseen test. VGG-M model was trained in 2D vs 3D convolution to compare their performance. Due to

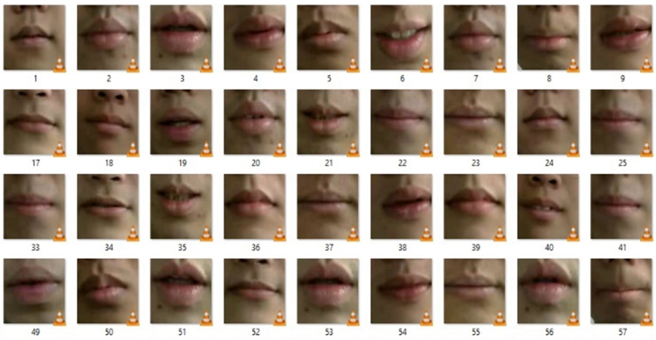

Figure 3. Samples of cropped lip images

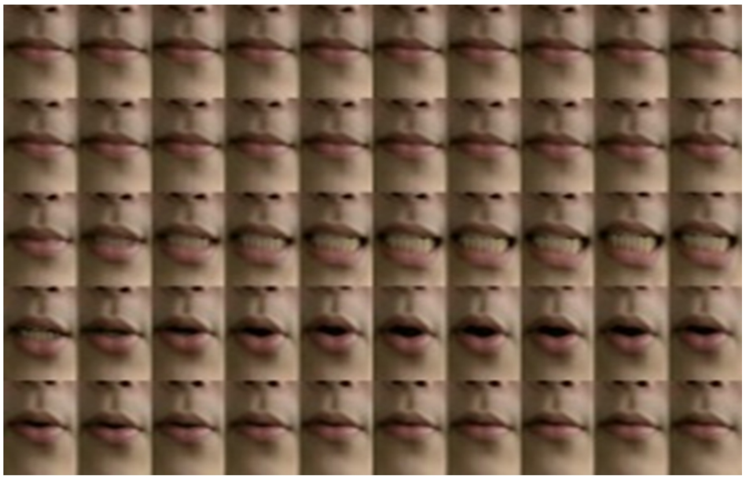

Figure 4. Concatenated Image

hardware limitation, a lighter model of 3D convolution was used. Table 3 shows that 2D convolution outperformed 3D convolution.

Table 3. 2D vs 3D Convolution Test

Types of Convolution	Stochastic Gradient Descent (SGD) (Learning Rate: 0.01, Momentum: 0.9)		Adam (Learning Rate: 0.01)	
	Training	Validation	Training	Validation
2D	91.42%	64.67%	100%	86%
3D	92.25%	64.67%	58.58%	45.33%

Architecture test was done by tuning kernel size, optimizer, nesterov acceleration, learning rate, learning rate decay, and L2 regularizer. Table 4 and 5 shows that model 4 is the best model among all.

Environment test was done by evaluating the models in 10 different environments. Table 6 and Table 7 shows that the trained models are not performing well.

The unseen test was done by evaluating the trained model on data that was not exposed in the training process. The models predict all the samples of the test data from digit 0 to 9. Model 4, 5, 7, 11 and 5 other models (SGD with learning rate of 0.0001 and L2 regularizer (12), SGD with learning rate of 0.00001 (13), SGD with learning rate of 0.00001 and L2 regularizer (14), Adam with learning rate of 0.000001 and L2 regularizer

Table 4. Architecture Test 1

Architecture	1	2	3	4	5	6
Kernel Size	5x5	7x5	7x5	7x5	7x5	7x5
Optimizer	SGD	SGD	SGD	SGD	SGD	SGD
Nesterov Acceleration	False	False	True	True	True	True
Learning Rate	0.01	0.01	0.01	0.001	0.0001	0.001
Learning Rate Decay	False	False	False	False	False	True
L2 Regularizer	False	False	False	False	False	False
Validation Accuracy	67%	68%	71%	87%	70%	51%

Table 5. Architecture Test 2

Architecture	7	8	9	10	11
Kernel Size	7x5	7x5	7x5	7x5	7x5
Optimizer	SGD	Adam	Adam	Adam	Adam
Nesterov Acceleration	True	True	True	True	True
Learning Rate	0.001	0.001	0.0001	0.00001	0.00001
Learning Rate Decay	False	False	False	False	False
L2 Regularizer	True	False	False	False	True
Validation Accuracy	83%	10%	63%	78%	81%

Table 6. Environment Test 1

Model	1	2	3	4	5	6
4	25%	75%	0%	0%	0%	0%
7	0%	25%	25%	0%	0%	25%
11	0%	25%	0%	50%	0%	0%
4(aug)	25%	25%	0%	0%	0%	0%
7(aug)	50%	25%	0%	0%	0%	0%
11(aug)	0%	25%	0%	0%	0%	0%

Table 7. Environment Test 2

Model	7	8	9	10	Percentage
4	50%	75%	0%	25%	25%
7	25%	50%	25%	25%	20%
11	75%	25%	25%	50%	25%
4(aug)	75%	50%	50%	25%	25%
7(aug)	75%	50%	25%	50%	27.5%
11(aug)	75%	50%	50%	25%	22.5%

(15) and SGD with learning rate of 0.00001, L2 regularizer and 200 epochs (16)) were evaluated. Table 8 and 9 verifies the bad performance of the trained model.

The environment test results imply that more data should be collected with a variety of brightness. The samples can be seen in Figure 5.

In the unseen test, the trained model is expected to perform well on people with similar facial features and lighting condition. However, the test subject that's being chosen as test data either moves his head or places his hand on his face before the recording session ends as shown in Figure 6. This type of recording is considered as faulty data.

Table 8. Unseen Test

Model	0	1	2	3	4	5
4	10%	80%	70%	20%	40%	10%
5	0%	30%	50%	0%	0%	10%
7	0%	40%	70%	20%	30%	30%
11	10%	70%	70%	0%	2%	30%
12	30%	80%	60%	10%	10%	10%
13	0%	0%	50%	0%	0%	0%
14	0%	40%	80%	0%	50%	10%
15	0%	0%	30%	0%	0%	10%
16	0%	0%	50%	0%	10%	30%

Table 9. Unseen Test 2

Model	6	7	8	9	Percentage
4	10%	0%	0%	0%	24%
5	0%	10%	0%	30%	13%
7	0%	0%	0%	0%	19%
11	0%	0%	0%	10%	19.2%
12	10%	0%	0%	0%	21%
13	0%	0%	40%	70%	16%
14	0%	0%	50%	70%	30%
15	0%	0%	0%	10%	5%
16	0%	0%	0%	60%	15%

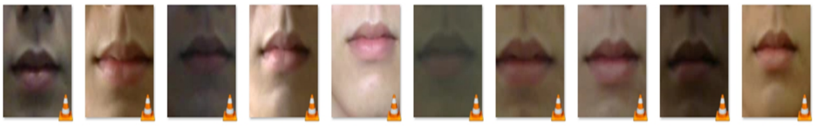

Figure 5. Sample Data with Different Lighting Condition

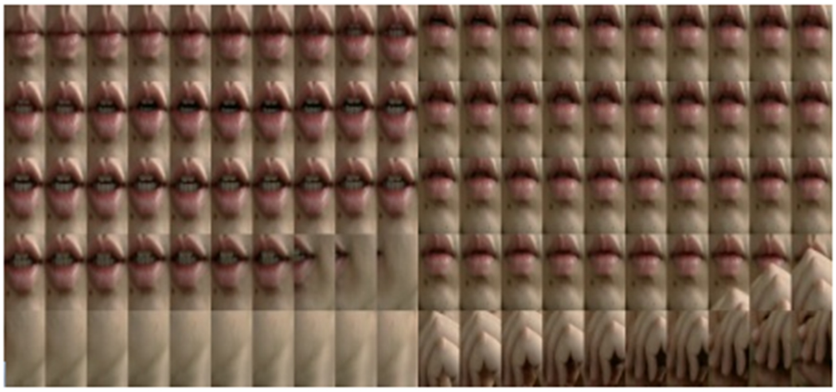

Figure 6. Example of faulty data

The best model in this project was compared with the other two papers that evaluated their models using MIRACL-VC1. The results show that the best model performs better in the seen test but bad in the unseen test. The comparison is summarized in Table 10 and Table 11.

Table 10. Seen Test (Comparison) 1

Architecture	Accuracy(%)		
	Training	Validation	Test
Fine-tuned VGG+LSTM (Gutierrez and Robert, 2017)	100	79	59
Model 4	100	87	90

Table 11. Unseen Test (Comparison) 2

Architecture	Accuracy(%)		
	Training	Validation	Test
Fine-tuned VGG+LSTM (Gutierrez and Robert, 2017)	100	79	59
5x5 stretched version (Garg et al, 2016)	63.1	79	56
Model 4	100	87	90

6. **Conclusion.** In conclusion, the aim and objectives of the project have been achieved by building a physical security system that could perform inferencing on the videos recording movement of the lips. The best-developed model is the 2D convolution architecture with a mixture of 7x7 and 5x5 kernels, SGD optimizer with a learning rate of 0.001 and Nesterov acceleration. 3D convolution is not used in this case due to the hardware limitation. The modified architecture is inspired by the EF3 model on 3D convolution. Adam optimizer speeds up the training process but yields lower accuracy than SGD. Test accuracy could not reach above 80% due to the lack of data. More data is required to achieve a higher accuracy together with deeper network and powerful hardware specifications.

REFERENCES

[1] I. Ipsic, Speech and language technologies. *BoDVBooks on Demand*, 2011.

[2] S. Pathan, A. J. I. J. o. C. S. Ghotkar, and I. Information Technologies, Recognition of spoken English phrases using visual features extraction and classification, *vol. 6, no. 4, pp. 3716-3719*, 2015.

[3] L. D. Terissi, M. Parodi, and J. C. Gomez, Lip reading using wavelet-based features and random forests classification, *in 2014 22nd International Conference on Pattern Recognition, IEEE,*, pp. 791-796, 2014.

[4] S. S. Morade and S. Patnaik, Lip reading by using 3-D discrete wavelet transform with dmey wavelet, *International Journal of Image Processing*, vol. 8, no. 5, pp. 384-386, 2014.

[5] S. S. Kumaravel, Visual speech recognition using histogram of oriented displacements. MS thesis, Clemson University, Clemson, SC. 2015.

[6] A. B. Hassanat, Visual passwords using automatic lip reading, *International Journal of Sciences: Basic and Applied Research*, vol. 13, no. 1, pp. 218-231, 2014.

[7] A. Nasuha, F. Arifin, T. A. Sardjono, H. Takahashi, and M. H. Purnomo, Automatic Lip Reading for Daily Indonesian Words based on Frame Difference and Horizontal-Vertical Image Projection, *Journal of Theoretical Applied Information Technology*, vol. 95, no. 2, 2017.

[8] N. Rathee, A novel approach for lip reading based on neural network, *in 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), IEEE*, pp. 421-426, 2016.

[9] S. S. Morade and S. Patnaik, A novel lip reading algorithm by using localized ACM and HMM: tested for digit recognition, *Optik*, vol. 125, no. 18, pp. 5181-5186, 2014.

[10] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, Lipreading using convolutional neural network, *in Fifteenth Annual Conference of the International Speech Communication Association*, pp.1149-1153, 2014.

[11] Y. Li, Y. Takashima, T. Takiguchi, and Y. Ariki, Lip reading using a dynamic feature of lip images and convolutional neural networks, *in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), IEEE*, pp. 1-6, 2016.

[12] J. S. Chung and A. Zisserman, Lip reading in the wild, *in Asian Conference on Computer Vision, Springer*, pp. 87-103, 2016.

[13] A. Garg, J. Noyola, and S. Bagadia, Lip reading using CNN and LSTM, *Technical report, Stanford University*, CS231n project report2016.

[14] A. Gutierrez and Z.-A. Robert, Lip Reading Word Classification, ed, 2017.

[15] I. Omara, G. Xiao, M. Amrani, Z. Yan, and W. Zuo, Deep features for efficient multi-biometric recognition with face and ear images, *in Ninth International Conference on Digital Image Processing (ICDIP 2017)*, vol. 10420, p. 104200D: International Society for Optics and Photonics, 2017.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, *in Advances in neural information processing systems*, pp. 1097-1105, 2012.

[17] M. D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, *in European conference on computer vision, Springer*, pp. 818-833, 2014.