An automatic lip reading for short sentences using deep learning nets

Maha A. Rajab [a,1,*], Kadhim M. Hashim [b,2]

[a] Department of Biology Science, College of Education for Pure Sciences /Ibn AL-Haitham, University of Baghdad, Baghdad, Iraq
[b] Department of Computer Technolongy Engineering, College of Information Technolongy, Imam Ja'afar AI-Sadiq University, Baghdad, Iraq
[1] maha.a.r@ihcoedu.uobaghdad.edu.iq; [2] Kadhem@sadiq.edu.iq
* corresponding author

ARTICLE INFO

Article history
Received September 21, 2022
Revised January 21, 2023
Accepted January 28, 2023
Available online March 31, 2023

Keywords
Lip-reading
CNN
AlexNet
VGG-16 net
Short sentences

ABSTRACT

One study whose importance has significantly grown in recent years is lip-reading, particularly with the widespread of using deep learning techniques. Lip reading is essential for speech recognition in noisy environments or for those with hearing impairments. It refers to recognizing spoken sentences using visual information acquired from lip movements. Also, the lip area, especially for males, suffers from several problems, such as the mouth area containing the mustache and beard, which may cover the lip area. This paper proposes an automatic lip-reading system to recognize and classify short English sentences spoken by speakers using deep learning networks. The input video extracts frames and each frame is passed to the Viola-Jones to detect the face area. Then 68 landmarks of the facial area are determined, and the landmarks from 48 to 68 represent the lip area extracted based on building a binary mask. Then, the contrast is enhanced to improve the quality of the lip image by applying contrast adjustment. Finally, sentences are classified using two deep learning models, the first is AlexNet, and the second is VGG-16 Net. The database consists of 39 participants (32 males and 7 females). Each participant repeats the short sentences five times. The outcomes demonstrate the accuracy rate of AlexNet is 90.00%, whereas the accuracy rate for VGG-16 Net is 82.34%. We concluded that AlexNet performs better for classifying short sentences than VGG-16 Net.

1. Introduction

There is increasing attention on visual speech recognition, often called lip-reading. It is a natural supplement to audio-based speech recognition and enables silent dictation in offices and public places, making it easier to dictate in noisy environments [1]. One of the key elements of virtual reality (VR) and human-computer interaction is an automatic lip-reading system. It is required for visual processing as well as human-language interaction [2]. Automatic lip-reading can be performed when the audio is present as an aid or if it is not. In the case of the absence of audio, it is often known as visual speech recognition [3]. The hearing-impaired can also find it to be of great help as a hearing aid. But similar to speech recognition, lip-reading algorithms face many difficulties because of variations in the inputs, such as features of the face, colors of skin, talking speeds, and intensity levels [4]. The effectiveness of lip reading is also closely related to the feature selection process and the widely utilized classification approaches. The most fundamental examples of traditional classification techniques are hidden Markov models, support vector machines (SVM), and the k-nearest neighbor approach. Deep learning techniques have been applied to classification problems frequently, and researchers have begun applying them to lip reading classification problems [5].

Typical conventional lip reading systems include two phases, the first is feature extraction, which depends on the pixel value extracted from the mouth area. Then, the extracted features are retrieved and put into the (SVM) in the second phase [6], [7]. In computer vision, deep learning has now achieved important advancements, including (representation of images, object detection, recognition of human behavior, and recognition of video). Therefore, the end-to-end deep learning architecture is the natural direction of scientific research as opposed to the previous manual feature extraction classification methodology for automatic lip-reading technology. Researchers use convolutional neural networks (CNN) to focus on interest regions in the last few years, and they have also been quite successful at classifying images and detecting targets [8], [9].

Studies on automatic lip reading have recently gained more attention, and significant progress in lip recognition is a complete reflection of image, speech, and natural language processing technology. This section outlines many techniques for an automatic lip-reading system. Nikita et al. [10] describes a technique for a lip-reading system that combines a (CNN) with attention-based Long Short-Term Memory (LSTM). The pre-trained model CNN extracts features from preprocessed video frames, which LSTM then processes to learn distinctive features. The SoftMax layer of the architecture provides the outcome of lip reading. The current work compares the experiments performed with two pre-trained models, VGG19 and ResNet50. Using ResNet50 and ensemble learning, the system has an accuracy of 85%. Ümit and Furkan [11] presents a model for lip-reading to extract features from the frames utilizing CNN-based models and classifies them using Bidirectional (Bi-LSTM). Experiment results show that the ResNet-18 and Bi-LSTM pair produce the best outcomes with accuracy values of 84.5% and 88.55%, respectively. Shashidhar and Sudarshan [12] presents an approach employed to identify only words from the lip movement utilizing video in the lack of audio, and this primarily aids in the extraction of words from a video without audio. For data classification and recognition, the proposed method makes use of the VGG16 pre-trained CNN architecture. The validation accuracy obtained is 76%. Zhi-Ming et al. [13] presents a VGG-M model for Visual Speech Recognition of Lip images utilizing CNN. A camera is employed to capture the video data that will be handled. The video recording is then sent to be predicted. There are several models presented for estimating words from video data without voice data. To eliminate unnecessary information, the data is then normalized by cutting the speakers' lips in all frames. The validation accuracy achieved is 87%. Pooventhiran et al. [14] presented a system consisting of two stages; the first stage aims to extract features from the lip region, and the second one uses a CNN model trained to classify the sentences. The accuracy achieves 76.89%.

The process of extracting the lip area to recognize the short sentences spoken by the speaker faces many problems, such as, the input video not only contains the face area but a complex background. It must be deleted and the face area determined, and the database contains males and females, so the mouth area in males contains the mustache and beard that cover the lip area, which must be removed and the lip area extracted. To find solutions to these problems, this paper proposes an automatic lip reading system to recognize short sentences using deep learning techniques.

The remaining portions of the article are arranged as follows: Section 2 offers the materials, existing deep learning techniques, and proposed lip-reading system layout. Section 3 contains the results and discussion. Section 4 of this article presents its conclusions.

2. Method

2.1. Database

The database includes normal, whispered, and silent speech. It required the participants to read ten brief phrases. The phrases are: “Excuse me”, “Goodbye”, “Hello”, “How are you”, “Nice to meet you”, “See you”, “I am sorry”, “Thank you”, “Have a good time”, “You are welcome”. Each participant repeats ten phrases five times in three various manners: normal, whispered, and silent. This section included 39 participants, including 32 men and 7 women. The database used in this work was recorded by Petridis et al. [15]. We only used a normal speech in this work.

2.2. Face Detection

The Viola-Jones algorithm is a method for object detection in images, specifically for faces. It consists of four main steps:

1) Haar feature selection: Haar-like features are used to represent the object to be detected. These features are calculated based on the difference in intensity between adjacent rectangular regions of the image.
2) AdaBoost training: A machine learning technique called AdaBoost is used to train a classifier using the selected Haar features. AdaBoost is used to select a subset of the most useful features for detecting the object.
3) Cascading classifiers: The classifier is then applied in a cascaded manner, where multiple stages of classifiers are applied, each one becoming more selective as the cascade progresses.
4) Non-maximum suppression: Finally, non-maximum suppression is applied to remove multiple detections of the same object. This step is used to eliminate duplicate detections of the same object and to improve the overall accuracy of the algorithm [16], [17].

2.3. Convolutional Neural Network (CNN) Models

Define (CNNs) are a specific kind of deep learning technique that attained advanced efficiency in computer vision tasks, including object identifying and recognizing, image retrieval, image classification, and segmentation. Thus, instead of extracting features from images directly using some feature extraction methods like SIFT, and HoG, CNN techniques detect and classify features from image data and generate grades based on their outcome [18]. The main building blocks of a CNN are [19].

1) Convolutional layers: Perform the convolution operation on the input image, which involves applying a set of filters to the image to extract features.
2) Pooling layers: used to reduce the spatial dimensions of the convolutional layer's output, also known as down-sampling. This is done by applying a pooling operation to the output, such as max pooling or average pooling.
3) Fully connected layers: These layers are used to make the final prediction.
4) Activation function: Is performed to the output of each layer to introduce nonlinearity into the neural network. The most popular activation functions are ReLU, sigmoid, and tanh.

The most popular pre-trained CNN are AlexNet, VGG-16, VGG-19, and GoogleNet [5]. AlexNet and VGG-16 Net will be used in this paper for the experiments as described below.

2.3.1. AlexNet

The AlexNet is a deep convolutional neural network that is utilized to categorize images into one of the thousands of classes. Many issues are resolved using AlexNet, such as indoor sense classification, which is widely utilized in artificial neural networks. It is an effective way of understanding an image's features with larger variability vision in the computer field of pattern recognition. The AlexNet consists of 5 convolutional layers, 3 sub-sampling layers, and 3 fully connected layers [20], [21] . Fig. 1 illustrates the AlexNet model's architectural layout [22].

2.3.2. VGG – 16 Net

The VGG-16 is a deep (CNN) architecture. The "16" in the name refers to the network having 16 layers of convolutional and fully connected layers. The VGG-16 architecture is known for using small convolutional filters (3x3) and deep architectures with a stride size of 1, while all the pooling layers are 2×2 with a stride size of 2 and the same padding. The size of the entered image to the VGG-16 is 224×224 by default. Before the fully connected layers, there is a 7×7 feature map with 512 channels. As the resulting feature map, this subset of features is expanded it into vector with 25,088 (7×7×512) channels. Fig. 2 illustrates the VGG-16 Net model's architectural layout [23], [24].

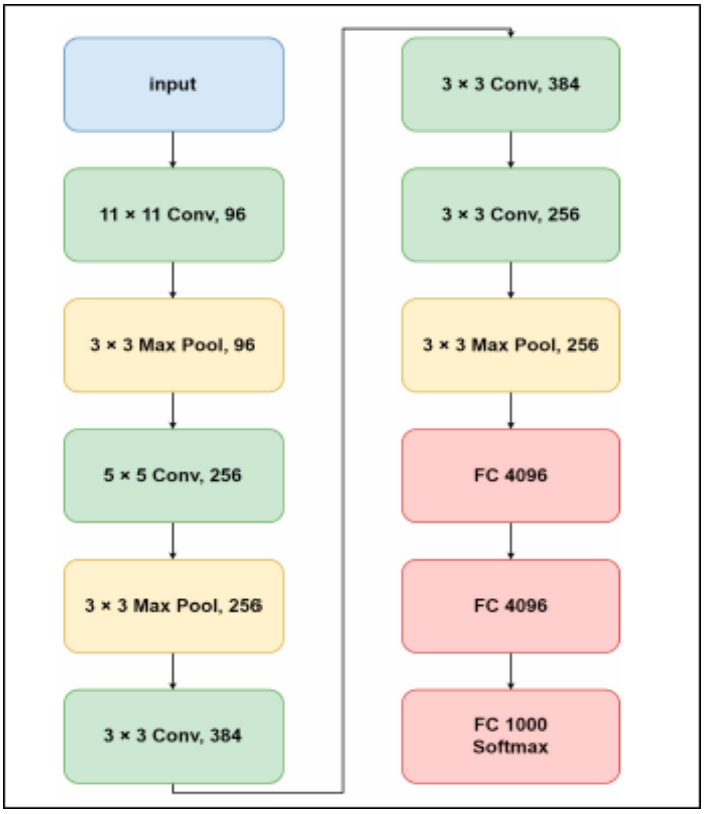

Fig. 1. AlexNet architecture [22]

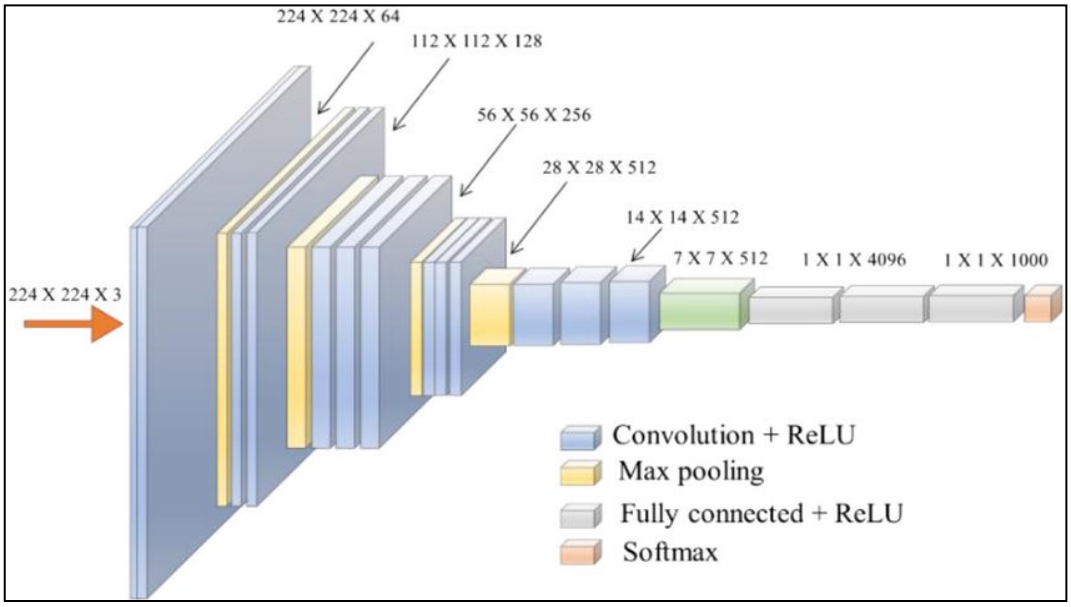

Fig. 2. VGG-16 Net model's architectural layout [23]

2.4. Proposed Lip Reading System

The proposed automatic lip reading system layout and main stages are thoroughly addressed in this section. Fig. 3 demonstrates the three main stages of the proposed system, including the lips region extraction stage, the lips enhancement stage, and the classification stage.

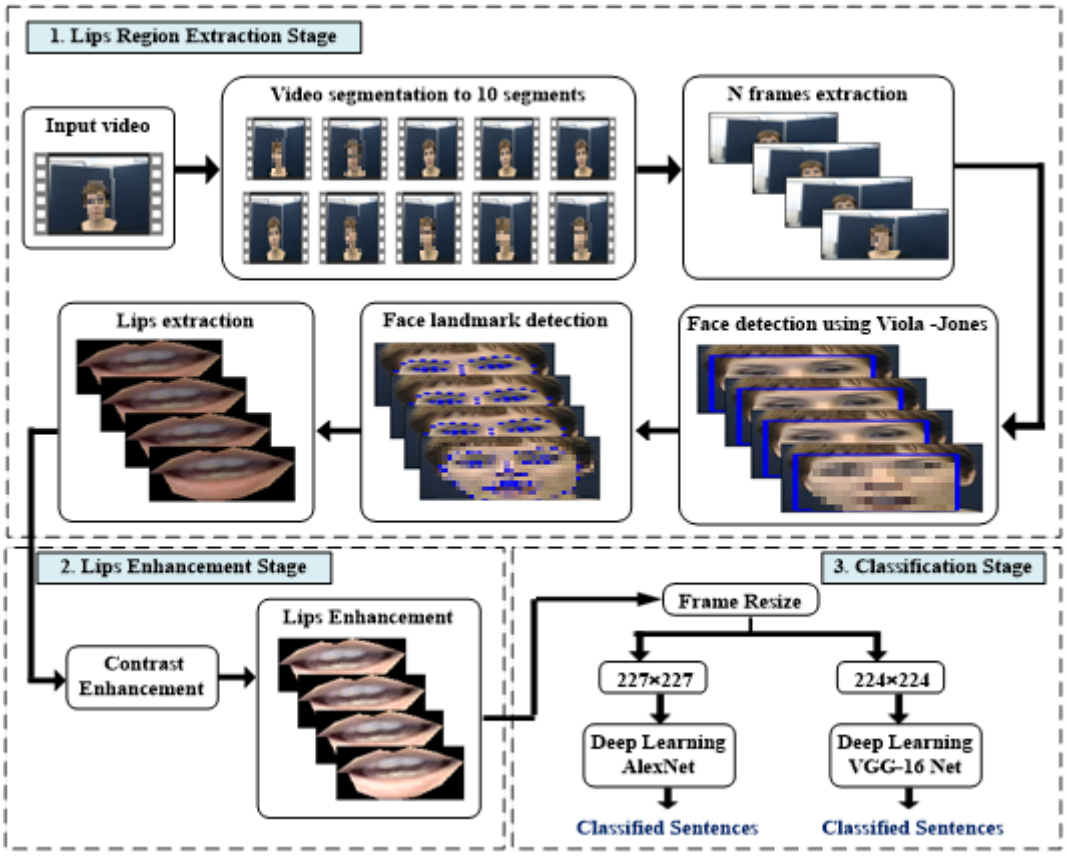

Fig. 3. The proposed an automatic lip reading system layout.

2.4.1. Lips Region Extraction Stage

This stage represents the first stage of the proposed system, which aims to extract the mouth area only from the input video, which consists of six main steps described as follows.

1) Input Video

The video is recorded by asking the speaker to say ten short sentences which are recorded by the camera. Then, the video is entered into our proposed system.

2) Video Segmentation

The input video is segmented into ten videos where, each video represents a short sentence said by the speaker, as illustrated in Fig. 4.

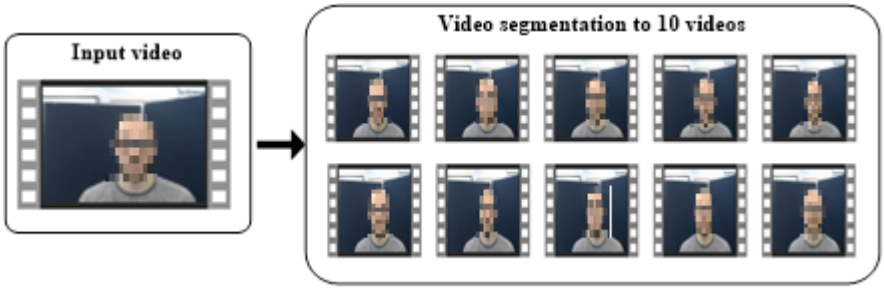

Fig. 4. Video segment to 10 videos

3) Frames Extraction

This step aims to extract frames from the ten segmented videos for each individual, which are kept in a temporary folder, as illustrated in Fig. 5.

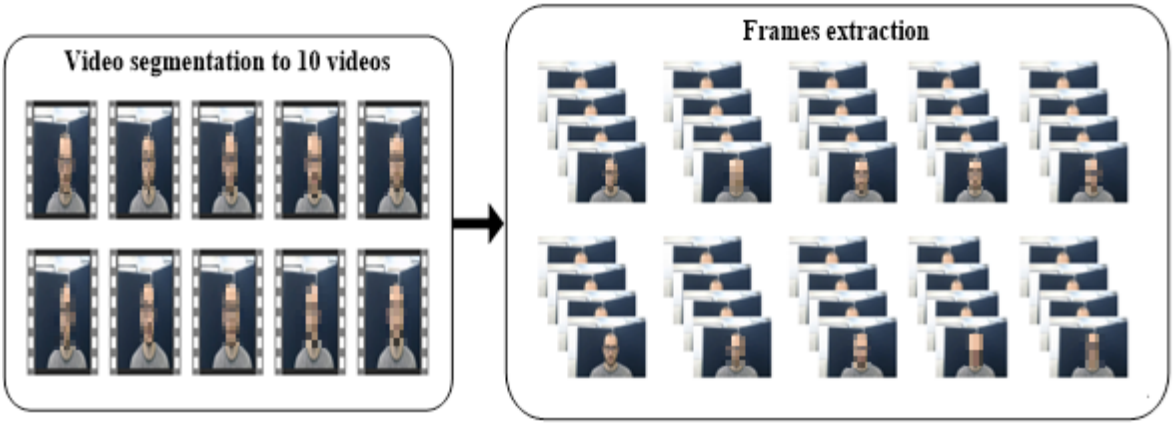

Fig. 5. Frames extraction

4) Face Detection Using Viola-Jones

The frames which are extracted from the original video do not contain only the image of the person's face, but also contain a complex background and some unimportant objects and noise, thus, we only need to determine the interest region from the frame, which represents the person's face and ignore the rest of the components of the frame. To accomplish this task, we use the Viola-Jones algorithm to detect and extract the face image where the face area is surrounded by a blue rectangle, as shown in Fig. 6.

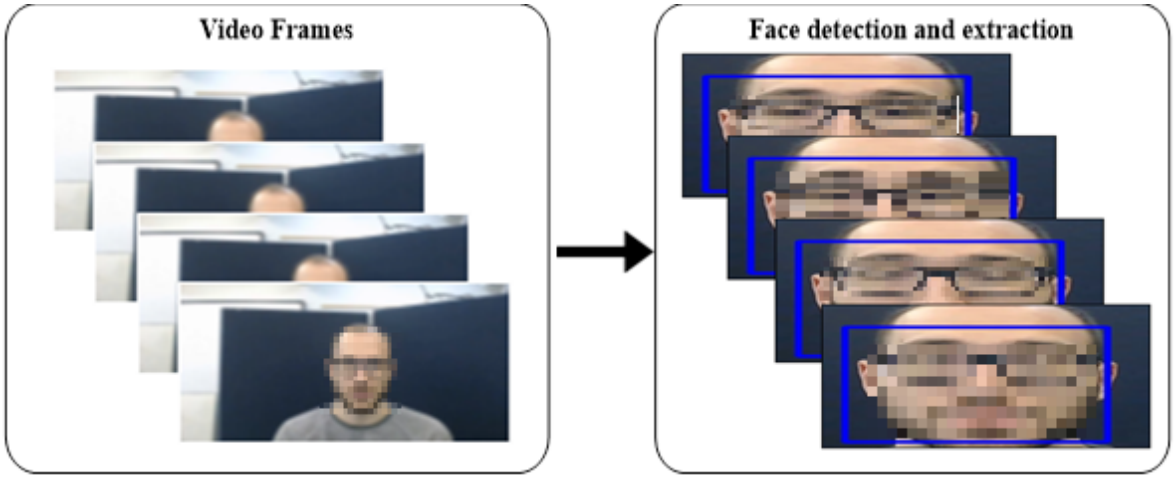

Fig. 6. Face detection and extraction

5) Face Landmark Detection

The process of detecting the landmarks of the face is done by identifying 68 points on the face, each group of points representing a part of the facial features including the eyes, nose, mouth, lip, etc., the landmarks detection is done by using dlip detector. The essential part of our work is the lip extracted by cropping points from 48 to 68 of the facial landmarks. Determining the lip area is very necessary, especially for male people, because the mouth area contains the mustache and beard, which may cover the lip area, and these problems affect the classification stage. Fig.7 illustrates the example of a man whose mouth area contains a mustache and beard; it also shows how the points of the lip area are determined for them. Fig. 8 shows an example of determining the facial landmarks of a woman selected from the database.

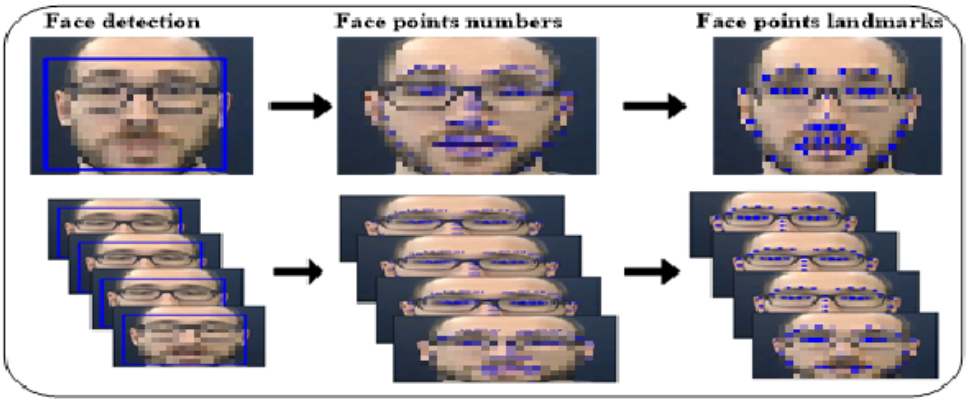

Fig. 7. Face landmarks determination for a man

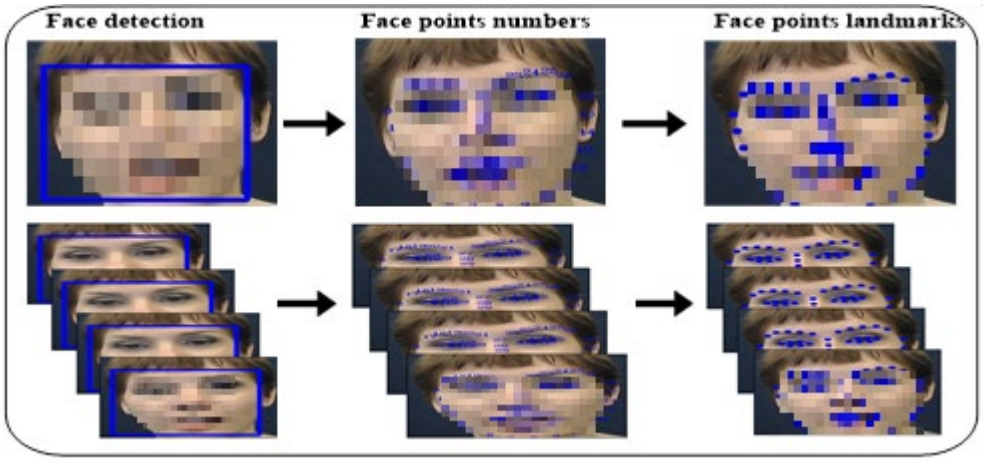

Fig. 8. Face landmarks determination for a woman

6) Lips Extraction

The process of lip extraction is done by building a binary mask, and this mask is a zero matrix whose size is the same as the size of the face image. Since the points from 48 to 68 of the facial landmarks represent the landmarks of the mouth area, we fill this area with white color in the mask image depending on the index of these landmarks, hence the mask image becomes black, except for the lips area is white color. After that, the mask image is multiplied by the image face to extract the lip area, and this extraction process is shown in Fig. 9.

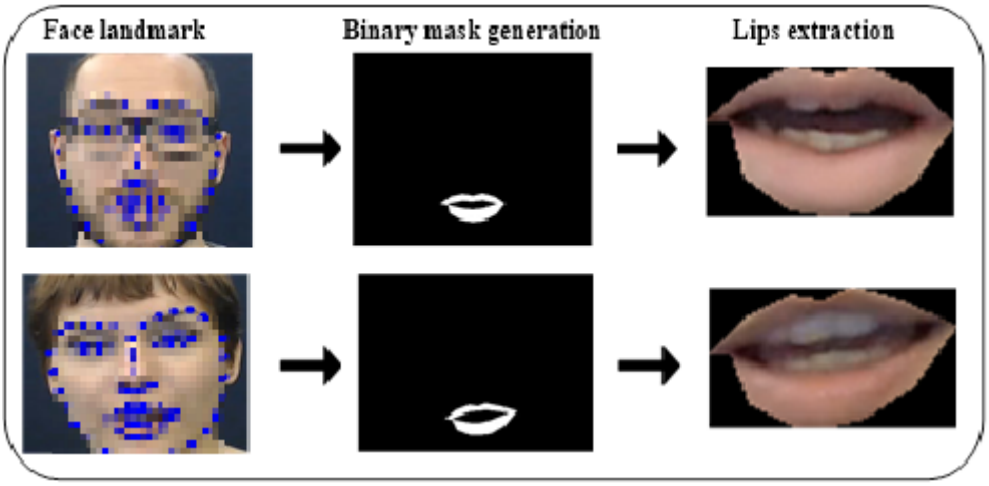

Fig. 9. Lips extraction

2.4.2. Lips Enhancement Stage

This stage aims to enhance the quality and brightness of the lips' image. It consists of the following steps, 1) Change the color space of an image from RGB to Lab; 2) Enhance the contrast of the image, contrast adjustment is only applied on the luminosity layer L using the image adjust function; and 3) The image is changed back into RGB format. Modifying luminosity affects pixel intensity while maintaining the original colors. Fig. 10 explains the results of applying contrast adjustment to enhance the contrast of the lips image.

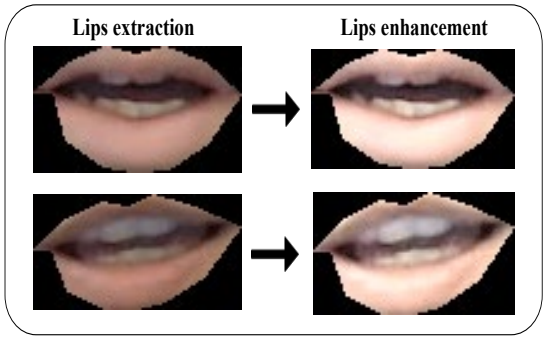

Fig. 10. Lips enhancement

2.4.3. Classification Stage

In this stage, we describe the various deep learning nets used to classify the sentences correctly. Since CNN are perfect for extracting the appropriate features from images, they form the foundation of the proposed system. In this stage, two nets are used to classify the short sentences are AlexNet and VGG-16 Net as described as follows:

- **AlexNet.** A sequence of N frames will be processed into 227 × 227 pixels and is used as input to the AlexNET, and is processed by 5 convolutional layers, each one of them is followed by a max-pooling layer, and 3 fully connected layers represent the outcomes of the last layer of a CNN.
- **VGG – 16 Net.** A sequence of N frames will be processed into 224 × 224 pixels and is used as input to the VGG-16 Net, and is processed by 13 convolutional layers, each one of them is followed by a max-pooling layer, and 3 fully connected layers represent the outcomes of the last layer of a CNN.

2.5. System Evaluation Criteria

Accuracy, Precision, recall, and Specificity are four metrics used to evaluate the proposed system's effectiveness and accuracy. Accuracy is the proportion of correct estimates as in (1) [25], [26], while Precision calculates the proportion of all truly positive detections, as in (2) [27], [28]. Recall calculates the proportion of detected ground truth annotations, as in (3) [29], [30], and Specificity measures the percentage of negative values which are correctly identified, as in (4) [31], [32].

$$Accuracy = (TP + TN)/\,(TP + TN + FP + FN) \quad (1)$$

$$Precission = (TP)/\,(TP + FP) \quad (2)$$

$$Recall = (TP)/\,(TP + FN) \quad (3)$$

$$Specifity = (TN)/\,(TN + FP) \quad (4)$$

3. Results and Discussion

The efficiency of the suggested automatic lip-reading system is evaluated using the database described in 2.1. The database consists of 39 speakers. Each speaker repeats the ten sentences five times, bringing the total number of videos to 195. After the videos have passed over all the phases of our suggested model until we get the lip area frame, thus the total number of frames is 130,000 frames, 80% will be

chosen for training, while the remaining 20% of the tested. The proposed system performance is tested by training two networks, the first is AlexNet and the second is VGG-16 Net. Therefore, the classification process begins by passing the video over all the phases of our suggested model until we get the frames. These frames are entered into the networks for training and measuring accuracy.

Table 1 displays the results from testing the suggested system using the above metrics, where the results of accuracy, Precision, recall, and Specificity for AlexNet are 90.00%, 89.7%, 89.8%, and 88.7%, respectively, while the results of accuracy, Precision, recall, and Specificity for VGG-16 Net are 82.34%, 84.4%, 81.9%, and 83.9%, respectively. According to the test results, the AlexNet gave better results than the VGG-16 Net network, and therefore the Alex network is considered better in classifying short sentences.

Table 1. The results of an automatic lip reading system

Model	Results			
	Accuracy%	*Precision%*	*Recall%*	*Specifity%*
AlexNet	**90.00%**	89.7%	89.8%	88.7%
VGG-16 Net	82.34%	84.4%	81.9%	83.9%

Fig. 11 shows the comparison between AlexNet and VGG-16 Net established on the accuracy, Precision, recall, and Specificity. It also shows the effectiveness of AlexNet in classifying short sentences. Fig. 12 and Fig. 13 offer the outcomes of Precision, recall, and Specificity for each short sentence pronounced by the speaker for AlexNet and VGG-16 Net, respectively. One main limitation of a proposed lip-reading system is that it is not robust to lighting and head poses variations. Additionally, the system may not perform well on individuals with unique facial features or speech patterns. Another limitation is that a deep learning system is trained on a huge database, so it may not generalize well to new, unseen data.

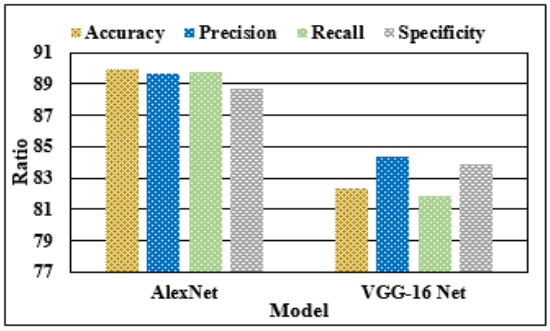

Fig. 11. Comparison between AlexNet and VGG-16 Net

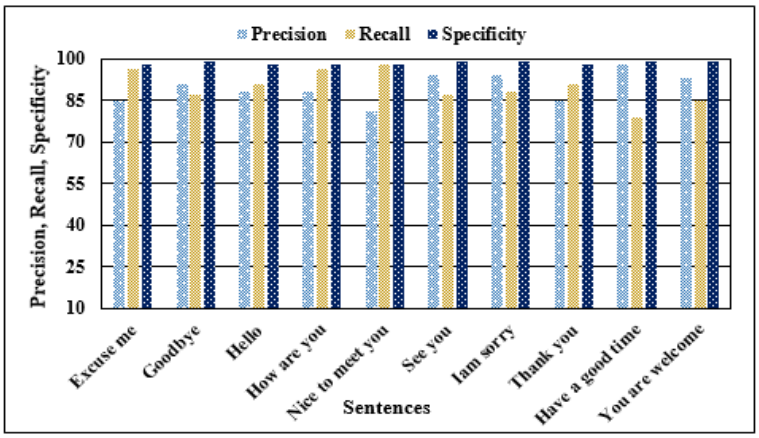

Fig. 12. The outcomes for each sentence when using AlexNet

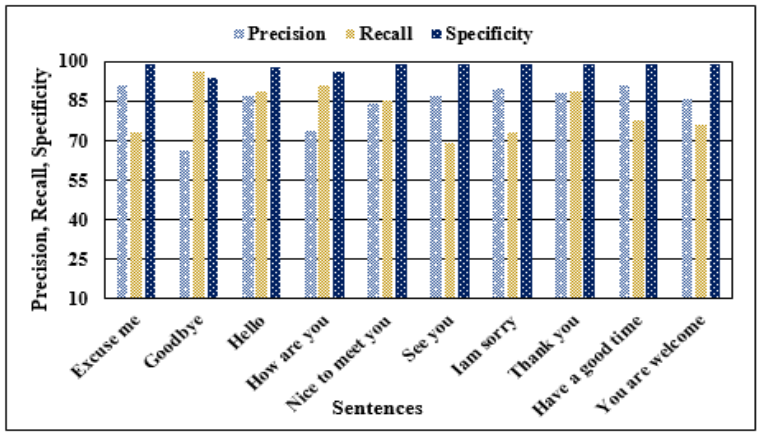

Fig. 13. The outcomes for each sentence when using VGG-16 Net

Table 2 compares our suggested strategy with other previously published studies and demonstrates that it yields superior outcomes from other previous tests. The results in Table 2 showed the suggested approach provides a greater accuracy, Precision, recall, and Specificity than other earlier studies. Thus, the effectiveness of our suggested system has been proven.

Table 2. Compared results with previous experiments

Reference	Comparison			
	Accuracy%	*Precision%*	*Recall%*	*Specificity%*
[10]	85%	-	-	-
[11]	88.55%	-	-	-
[12]	76%	-	-	-
[13]	87%	-	-	-
[14]	76.89%	-	-	-
Our proposed with AlexNet	**90.00%**	89.7%	89.8%	88.7%
Our proposed with VGG-16 Net	82.34%	84.4%	81.9%	83.9%

4. Conclusion

In this section, a new lip-reading system is suggested that utilizes deep learning to expect short sentences with a variety of vocabulary in natural speaker recordings. Also, we focus on extracting the lip area based on determining the face area from the video using Viola Jones, then, we determine 68 points in the face, which represent the parts of the face, and the points from 48 to 68 represent the lip area, which is extracted based on the binary mask. The image of the lip area is also enhanced to enhance quality. To classify the short sentences, we used AlexNet and VGG-16 Net. The outcomes demonstrate the accuracy rate for AlexNet is 90.00%, whereas the accuracy rate for VGG-16 Net is 82.34%. We concluded that AlexNet performs better for classifying short sentences than VGG-16 Net. Future work can be carried out from the study by developing deep learning models specifically designed for lip reading, taking into account the unique characteristics of lip movements during speech and integrating lip reading with other modalities, such as audio and visual information, to improve performance.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] N. Akhter *et al.*, "Diverse Pose Lip-Reading Framework," *Appl. Sci. 2022, Vol. 12, Page 9532*, vol. 12, no. 19, p. 9532, Sep. 2022, doi: 10.3390/APP12199532.

[2] Y. Lu and H. Li, "Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory," *Appl. Sci. 2019, Vol. 9, Page 1599*, vol. 9, no. 8, p. 1599, Apr. 2019, doi: 10.3390/APP9081599.

[3] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip Reading Sentences Using Deep Learning with only Visual Cues," *IEEE Access*, vol. 8, pp. 215516–215530, 2020, doi: 10.1109/ACCESS.2020.3040906.

[4] K. Srilakshmi and R. Karthik, "A Novel Method for Lip Movement Detection using Deep Neural Network," *J. Sci. Ind. Res.*, vol. 81, no. 06, pp. 643–650, Jun. 2022, doi: 10.56042/JSIR.V81I06.53898.

[5] T. OZCAN and A. BASTURK, "Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models," *Balk. J. Electr. Comput. Eng.*, vol. 7, no. 2, pp. 195–201, Apr. 2019, doi: 10.17694/BAJECE.479891.

[6] H. Huang *et al.*, "A Novel Machine Lip Reading Model," *Procedia Comput. Sci.*, vol. 199, pp. 1432–1437, Jan. 2022, doi: 10.1016/J.PROCS.2022.01.181.

[7] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, "A survey of research on lipreading technology," *IEEE Access*, vol. 8, pp. 204518–204544, 2020, doi: 10.1109/ACCESS.2020.3036865.

[8] A. Pyataeva and A. Dzyuba, "Artificial neural network technology for lips reading," *E3S Web Conf.*, vol. 333, p. 01009, 2021, doi: 10.1051/E3SCONF/202133301009.

[9] S. Jeon, A. Elsharkawy, and M. S. Kim, "Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition," *Sensors 2022, Vol. 22, Page 72*, vol. 22, no. 1, p. 72, Dec. 2021, doi: 10.3390/S22010072.

[10] N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali, and K. Warkari, "Vision based Lip Reading System using Deep Learning," *2021 Int. Conf. Comput. Commun. Green Eng. CCGE 2021*, 2021, doi: 10.1109/CCGE50943.2021.9776430.

[11] Ü. Atila and F. Sabaz, "Turkish lip-reading using Bi-LSTM and deep learning models," *Eng. Sci. Technol. an Int. J.*, vol. 35, p. 101206, Nov. 2022, doi: 10.1016/J.JESTCH.2022.101206.

[12] R. Shashidhar and S. Patilkulkarni, "Visual speech recognition for small scale dataset using VGG16 convolution neural network," *Multimed. Tools Appl.*, vol. 80, no. 19, pp. 28941–28952, Aug. 2021, doi: 10.1007/S11042-021-11119-0/METRICS.

[13] Z. M. Chan, C. Y. Lau, and K. F. Thang, "Visual speech recognition of lips images using convolutional neural network in vgg-m model," *J. Inf. Hiding Multimed. Signal Process.*, vol. 11, no. 3, pp. 116–125, 2020.[Online]. Available: http://bit.kuas.edu.tw/~jihmsp/2020/vol1/2_jihmsp-1522_vol3.pdf

[14] G. Pooventhiran, A. Sandeep, K. Manthiravalli, D. Harish, and R. D. Karthika, "Speaker-Independent Speech Recognition using Visual Features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 616–620, 2020, doi: 10.14569/IJACSA.2020.0111175.

[15] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-Only Recognition of Normal, Whispered and Silent Speech," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 6219–6223, Sep. 2018, doi: 10.1109/ICASSP.2018.8461596.

[16] T. H. Obaida, N. F. Hassan, and A. S. Jamil, "Comparative of Viola-Jones and YOLO v3 for Face Detection in Real time," vol. 22, no. 2, pp. 63–72, 2022. [Online]. Available: https://iraqjournals.com/article_175825_0.html

[17] Dherya Bengani and Prof. Vasudha Bah, "Face Detection Using Viola Jones Algorithm," *Int. J. Mod. Trends Sci. Technol.*, vol. 6, no. 11, pp. 131–134, 2020, doi: 10.46501/ijmtst061124.

[18] I. T. Ahmed, C. S. Der, N. Jamil, and M. A. Mohamed, "Improve of contrast-distorted image quality assessment based on convolutional neural networks," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 6, pp. 5604–5614, Dec. 2019, doi: 10.11591/IJECE.V9I6.PP5604-5614.

[19] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image Video Process.*, vol. 12, no. 2, pp. 355–362, Feb. 2018, doi: 10.1007/s11760-017-1166-8.

[20] M. M. Krishna, M. Neelima, M. Harshali, and M. V. G. Rao, "Image classification using Deep learning," *Int. J. Eng. Technol.*, vol. 7, no. 2.7, pp. 614–617, Mar. 2018, doi: 10.14419/IJET.V7I2.7.10892.

[21] P. Haripriya, "Deep learning pre-trained architecture of alex net and googlenet for DICOM image classification," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, pp. 3107–3113, 2019.[Online].Available: http://www.ijstr.org/final-print/nov2019/Deep-Learning-Pre-trained-Architecture-Of-Alex-Net-And-Googlenet-For-Dicom-Image-Classification.pdf

[22] W. Ketwongsa, S. Boonlue, and U. Kokaew, "A New Deep Learning Model for the Classification of Poisonous and Edible Mushrooms Based on Improved AlexNet Convolutional Neural Network," *Appl. Sci. 2022, Vol. 12, Page 3409*, vol. 12, no. 7, p. 3409, Mar. 2022, doi: 10.3390/APP12073409.

[23] F. D. Adhinata, N. A. F. Tanjung, W. Widayat, G. R. Pasfica, and F. R. Satura, "Comparative Study of VGG16 and MobileNetV2 for Masked Face Recognition," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 7, no. 2, pp. 230–237, Jul. 2021, doi: 10.26555/JITEKI.V7I2.20758.

[24] Q. Guan *et al.*, "Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study," *J. Cancer*, vol. 10, no. 20, pp. 4876–4882, 2019, doi: 10.7150/JCA.28769.

[25] M. A. Rajab and K. M. Hashim, "Dorsal hand veins features extraction and recognition by correlation coefficient," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 20, no. 4, pp. 867–874, Aug. 2022, doi: 10.12928/TELKOMNIKA.V20I4.22068.

[26] M. Arhami, A. Desiani, S. Yahdin, A. I. Putri, R. Primartha, and H. Husaini, "Contrast enhancement for improved blood vessels retinal segmentation using top-hat transformation and otsu thresholding," *Int. J. Adv. Intell. Informatics*, vol. 8, no. 2, pp. 210–223, Jul. 2022, doi: 10.26555/ijain.v8i2.779.

[27] S. Ghosh *et al.*, "Evaluation and Optimization of Biomedical Image-Based Deep Convolutional Neural Network Model for COVID-19 Status Classification," *Appl. Sci.*, vol. 12, no. 21, p. 10787, Oct. 2022, doi: 10.3390/APP122110787/S1.

[28] A. A. Azmer, N. Hassan, S. H. Khaleefah, S. A. Mostafa, and A. A. Ramli, "Comparative analysis of classification techniques for leaves and land cover texture," *Int. J. Adv. Intell. Informatics*, vol. 7, no. 3, pp. 357–367, Nov. 2021, doi: 10.26555/ijain.v7i3.706.

[29] M. J. J. Ghrabat, G. Ma, I. Y. Maolood, S. S. Alresheedi, and Z. A. Abduljabbar, "An effective image retrieval based on optimized genetic algorithm utilized a novel SVM-based convolutional neural network classifier," *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, pp. 1–29, Dec. 2019, doi: 10.1186/S13673-019-0191-8/FIGURES/20.

[30] A. E. Maxwell, T. A. Warner, and L. A. Guillén, "Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 1: Literature Review," *Remote Sens. 2021, Vol. 13, Page 2450*, vol. 13, no. 13, p. 2450, Jun. 2021, doi: 10.3390/RS13132450.

[31] R. Ruslan, S. Khairunniza-Bejo, M. Jahari, and M. F. Ibrahim, "Weedy Rice Classification Using Image Processing and a Machine Learning Approach," *Agric. 2022, Vol. 12, Page 645*, vol. 12, no. 5, p. 645, Apr. 2022, doi: 10.3390/AGRICULTURE12050645.

[32] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Res. Notes*, vol. 15, no. 1, pp. 1–8, Dec. 2022, doi: 10.1186/S13104-022-06096-Y/FIGURES/2.