Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 5, October 2021, pp. 2557~2565
ISSN: 2302-9285, DOI: 10.11591/eei.v10i5.2927

Constructed model for micro-content recognition in lip reading based deep learning

Nada Hussain Ali[1], Matheel E. Abdulmunim[2], Akbas Ezaldeen Ali[3]
[1]Imam Ja'afar Al-Sadiq University, Baghdad, Iraq
[2,3]Computer Science Department, University of Technology, Baghdad, Iraq

Article Info

Article history:

Received Feb 28, 2021
Revised Jun 14, 2021
Accepted Jul 8, 2021

Keywords:

CNN
Deep learning
Lip reading
Micro-contents

ABSTRACT

Communication between human beings has several ways, one of the most known and used is speech, both visual and acoustic perceptions sensory are involved, because of that, the speech is considered as a multi-sensory process. Micro contents are a small pieces of information that can be used to boost the learning process. Deep learning is an approach that dives into deep texture layers to learn fine grained details. The convolution neural network (CNN) is a deep learning technique that can be employed as a complementary model with micro learning to hold micro contents to achieve special process. In This paper a proposed model for lip reading system is presented with proposed video dataset. The proposed model receives micro contents (the English alphabet) in video as input and recognize them, the role of CNN deep learning is clearly appeared to perform two tasks, the first one is feature extraction and the second one is the recognition process. The implementation results show an efficient accuracy recognition rate for various video dataset that contains variety lip reader for many persons with age range from 11 to 63 years old, the proposed model gives high recognition rate reach to 98%.

Corresponding Author:

Nada Hussain Ali
Department of information technology
Imam Ja'afer Al-Sadiq University
Baghdad, Iraq
Email: cs.19.47@grad.uotechnology.edu.iq, nada.hussien@sadiq.edu.iq

1. INTRODUCTION

In machine learning vision, visual speech recognition (VSR), also known as automatic lip-reading, is the process of recognizing the words through processing and observing the visual lip movement of a speaker's talking without any audio input. Although visual information itself cannot be considered as enough resource to provide normal speech as intelligibility, it may succeed with several cases especially when the words to be recognized are limited [1]. Visual lip-reading plays an important role in the interaction between human and computer in noisy environments where audio speech may be difficult to recognize. It can also be very useful for the hearing-impaired as a hearing aid tool [2]. Despite the fact that audio signals are in much more informative than video signals, it has been noticed that most people use lip-reading gestures to understand speech [3]. Lip reading is difficult task for both machines and humans due to the considerably high similarity of lip shape and movements corresponding to uttering letters (e.g., letters b and p, or d and t). In addition to the lip movement the, lip size, wrinkles around the mouth, orientation, brightness and the environment around the speaker also affect the quality of the detected words. Sarhan, *et al.* [4] micro learning presents the opportunity to absorb and retain the information provided and the activities that are more digestible and manageable easily. The way micro-learning identifies small portions of learning content which

consists of fine-grained and loosely-coupled that are interconnected and shortened learning activities which defines the concentrate on the individual learning needs [5]. Deep networks, which are considered robust and precise learning techniques, are able to learn from data in the same way that babies are able to learn from the world around them, starting with fresh eye sight and gradually acquiring more skills needed to navigate environments around them. Many difficult problems can be solved using the same learning networks; their solutions can be generalized and need much less work than writing a different program for each problem. The deep learning revolution has two convoluted themes: how artificial intelligence (AI) evolved and how human intelligence is evolving. The difference between the two types of intelligence is the time needed for evolving, human intelligence took many years to evolve, but AI is evolving faster on a trajectory measured in decades. The conversion from AI based on logic, symbols and rules to deep learning approach based on learning algorithms and big data is not easy [6]. Deep learning techniques will be the efficient solution that empowers classification techniques spatially on images [7]. The remaining sections of this paper are as; section 2 related work description is provided, section 3 the deep learning and convolution neural network (CNN) technique is presented, section 4 micro-learning basic concept is presented, section 5 the proposed model frame work is provided and the experimental results are discussed and section 6 conclusion and future work are discussed.

In the literature, several works are presented for the most relevant that are relates to the proposed model in this paper as; Drakidou [8], proposed that using microlearning in e-learning courses enhance the long life learning and continuous learning. The author implanted several example courses that are carefully designed, supervised and implemented by well-trained instructors-facilitators. The author proved that microlearning can be used as an e-learning technique that will improve learning outcomes. Mohammed, *et al.* [9] proposed that an important requirement for successful learning is experiencing learning activities on a regular basis and keeping it memorable for long time. Microlearning can be delivered in small chunks which make memorable and easy to understand the authors test microlearning technique on primary school student and they found that student which learned using micro learning gained better learning than student that were subjected to traditional learning. Rettger [10] presented the idea of employing microlearning using mobile devices for academic studies and how the delivery of instruction-distributed presentation will affect the learning outcome and the author proved that students receiving small units of instruction and information over a series of days would perform much better than students receiving the instruction and information in a massed unit. Friesen [11], suggested that the traditional learning is forcing constrains on the learner. Micro learning is giving the ability for personalized learning and freeing the learner from those constrains. The author thinks that these features of micro learning are important and valuable. Lu and Li [12] proposed a lip reading system using deep learning to recognize numbers from 1-9 in videos, they used CNN to capture features and RNN to extract the sequence relationship between the video frames, the CNN and RNN are used as encoder and decoder respectively in decoding process an attention mechanism is used to learn attention weights, therefore the model take the whole video as attention area, the model gave accuracy 88.2% on the tested dataset. Mesbah, *et al*. [13] proposed a visual based lip reading system from videos by presenting a novel convolution neural network called Hahn by changing the first layer of CNN and using Hahn moment as first layer, the proposed HCNN helped in reducing the dimnstionality of the videos or images and gave good results with 90% accuracy on different datasets. Chung and Zisserman [14] proposed model for profile lip reading instead of frontal view lip reading. They used a ResNet to classify the faces into 5 groups (frontal-left profile-left three quarter-right three quarter-right profile), and they used a SyncNet for achieving the purpose of the proposal by synchronous the audio with the video lip motion, active speaker detection and sequence to sequence feature generation model. The model reached good results compared to other methods frontal face 91%, 30 face angle 90.8, 45 face angle 90%, 60 face angle 90% and profile face 88.9%. Cruz, *et al.* [15] proposed a lip reading model to recognize the English letters in filipino speakers, the dataset were gathered from 30 speakers, 15 male and 15 female, the videos were pre-recorded for the speakers, the model depends on lip movement only and using point distribution model (PDM) and kanade lucas tomasi (KLT) tracking algorithms template to extracted features from 16 key frames, a J48 decision tree algorithm is used for classification, the model achieved 45.26% average accuracy. Ibrahim and Mulvaney [16] proposed a system for lip reading that can recognize the English digit from 0-9, the model contains four steps, the first step is to extract the face from video then the mouth area using Viola jones object recognizer. In the second step, two regions are detected from the mouth area which are lip and non-lip regions. The third step is to extract lip geometry using a proposed approach depends on borders and convex hull computation to generate a shape based features. The final step, a novel approach, is used to classify the geometric features. This model achieved word recognition accuracy about 71%.

2. THEOREMS AND ALGORITHMS

In this section the used thermos and algorithms in the proposed work are explained

2.1. Convolutional neural networks

Deep learning in recent years has proven to be accurate on some tasks that surpass that of a human. Actually, the recent results gained from deep learning algorithms that transcend human ability and performance in image recognition tasks that can't likely considered by computer vision experts in the last decade. Many architectures of deep learning that presents such phenomenal performance are not a results of a random connections of computational units. The outstanding performance shown by deep neural networks reflect the fact that biological neural networks obtained much of their strength and power also from depth. Furthermore, it is not fully understood how biological networks are connected. In the cases that the biological network structure is understood at some grade, great achievements have been reached by modeling artificial neural networks based on those networks [17]. The main goal in applying deep learning to computer vision (CV) is to remove the exhausting, and limiting, feature selection process. Deep neural networks are very efficient for this process because it works in layers and each layer of a neural network is responsible for building up features and learning to represent the receives input [18]. The architecture of deep-learning is a like stack of modules that is considered as multilayer, all of these models or most of them are undergo to learning, all or (many) of them process non-linear input-output mappings. In this stack each module diverts its input to boost both the invariance and selectivity of the representation of the model. With several layers that are non-linear, say a depth of 5 to 20, the system will be able to implement extremely complex functions of its inputs that are sensitive to details-the system can distinguish a dog from a muffin-and incurious to variations that are irrelevant such as the pose, background, surrounding objects and lighting [19]. CNNs are a powerful combination of math, biology and computer science, these neural networks have been one of the most effective innovations in the field of artificial intelligence and computer vision [20]. CNN enables learnings and obtaining large quantities of information from raw data abstraction level [21]. CNN consist of serval component, these components are convolution layers, pooling layers, fully connected layers activation function dropout layers. The first layers which are the convolution layers contain number of filters these filters are responsible of feature extraction process and they learn as the fully connected layers do [22]. these filters provide a chance to recognize and detect features not caring of their positions in the image for that reason these layers are called convolutions. In these layers (convolutional) the filters are initialized, then they go through training procedure shape filters, which are suitable for the feature extraction task. For more benefits of this process, more layers can be added for more in details features by employing different filters in each layer [23]. Smaller objects are extracted from the input image these objects are deep features from the original image, this process gets iterated in every convolution layer. The convolution process that leads to feature extraction can be considered as compression of important information extracted from the input image. After feature compression and deeper information representation in the convolution layer another layer is needed called max pooling layer, this layer may precede or follow the convolution layer. The max-pooling layer use several hyperprameters that that are often organized as 2 by 2 grid, the image is divided into several areas the same size as the pool size (hyperpramerters grid) and chooses from each pool (four pixels) the maximal value. These pixels Compose new image, while preserving the order of the pixels in the original image. This process will produce an image that is half in size from the original image while keeping the channel number. An alternative of the maximal value can be choosing like minimum or average in a way that better serve the process. The idea that lies behind the max pooling layer is that the important pixels that hold information about features are rarely adjacent in an image so picking the maximum value from a surrounding of four pixels will catch the pixel that is highly informative. This layer gives the best results when it's implemented on feature map rather than the original image [24]. After several convolution and pooling layers, the architecture end with number of fully connected layers. The feature maps extracted from the convolution layers and pooling layers are transferred into a vectors, at this point to avoid overfitting a dropout layers can be added these layer are virtual layers that drop some of the connections in the fully connected layers. The finale fully connected layer in the architecture contains the same amount of output neurons as the number of classes to be recognized [25].

2.2. Micro content

Micro-content and micro-learning together determine how to submit a quantum of information and knowledge, structured in many short sections, fine-grained, interconnected and well-defined. The piece of information whose size is determined by a single topic, content that covers a single concept or idea and can be accessed via a single URL, being suitable for using in handheld devices, web-browsers, emails all that are refers to micro-content. Thus, micro-content is the part that merges into micro-learning [5]. In micro learning knowledge are acquired using instructional design techniques, abilities and skills which happen on a daily basis. The way that micro learning works is by taking information naturally by learner's brain, so that the body and brain does not get stressed. One of the essential features of micro learning that works saliently is that it allows the learner to find what he or she is looking for exactly. It enables the learner's brain to explore and satisfy its own patterns and its own curiosity [26]. Micro-learning proved its flexibility and adaptability

to deliver micro-content using easy to access techniques like email, mobile and network social society. Using micro- content make it easy to update and it can considered as standalone learning units though can be used as supporting units in other learning techniques. The researcher found that using micro-learning can improve the e-learning and can be very helpful for the people who are seeking continuous learning [8].

3. RESEARCH METHOD

The proposed model is divided into several stages as illustrated in the flowchart of this model, in the subsections below a full description of the model is presented.

3.1. The proposed dataset

The dataset was built by the authors, using more than 2700 pre-recorded videos of 11 persons (male and female from different ages). The videos were one to two seconds long consisting of the pronunciation of the English alphabet. The dataset contains 20 letters only, due to the difficulty to differentiate between similar pronounced letters, this similarity originates from the mouth geometry during letter utterance, but not from the acoustic information, these letters like (A, U), (F, V), (P, B), (Q, W), (K, C), (S, X). The recording process was held in several artificial lighting condition, the distance between the camera and the persons were 30 centimeters and the height was horizontal to the face, each video has the top part (from shoulders) of the person pronouncing the letters.

3.2. Preprocessing

The preprocessing plays an important role in any system, in the proposed model the preprocessing is implemented in two stages, dataset preprocessing and constructed model preprocessing.

a. Dataset preprocessing: The videos in the dataset is passed into several steps in order to prepare it to be used in the model, these steps are as:

- Convert the video into frames, in this step the videos are converted into frames (29 frame per second), the frames are saved for next steps.
- Face detection step, in this step, Haar Cascade face detection technique is used to detect the face in the frame and crop the face area only.
- Mouth detection step, in this step, the output from the previous step is fed as input to this step, the mouth area is cropped using spatial coordinate detection technique.
- Key frame selection step, in this step, a key frame (or frames) is selected based on visual features, this frame (or frames) represents the utterance letter and distinguish it from other letters.

After these steps a prepared dataset is formulated and constructed which consist of utterance letters key frames of the mouth area only, Figure1 shows the dataset through several steps.

(a) (b) (c)

Figure 1. Dataset preprocessing steps; (a) the frame extracted from the video without preprocessing, (b) the frame after detecting the and cropping the face, (c) after cropping the mouth are only

b. Model preprocessing: After the dataset has been preprocessed and prepared as a formulated and constructed form for the recognition process, the model preprocessing stage is achieved as the data will be ready for the recognition process. The following steps illustrates the model preprocessing stage:

- Extracting the labels from the dataset, each letter frames are stored in a file with a name as the letter name (A for letter A, so the others), these names are compared with the labels given to consider them as a target.
- Reshape, in this step, the frames are reshaped into square 224*224 images.

– Dataset partitioning, the dataset is partitioned into two categories, training set 75% and testing set 25%.

3.3 Data augmentation and normalization

Data augmentation technique is used to expand the dataset because when using deep learning, the data must be large enough in order to avoid overfitting problem, this problem happen when the neural networks can't generalize to the testing set because the neural network learned the features of the training set to well it can't generalize. Employing data augmentation on the dataset is as follows:

– Rotating the images within 30 degree.
– Zooming the images with 0.15 percentage.
– Shafting the images in the width 0.2 degree.
– Shafting the images in the height 0.2 degree.
– Shearing the images in rang equals to 0.15.
– Horizontal flipping.

After employing data augmentation, each frame has several copies that are rotated, zoomed, shafted, sheared or flipped. Now the data is large enough to proceed with deep learning, the next step is to normalize the data before feeding it to CNN. The mean subtraction technique is used to normalize the data, in this technique the mean RGB value for the training data set is computed and then subtracted from every pixel.

3.4 Micro content recognition using convolution neural network

In this work a convolution neural network is used for recognizing the letters as 20 class for 20 letters. The visual geometry group (VGG)19 pre-trained CNN is used with image-net weights, the VGG consist of several layers, 16 convolution layers and 3 fully connected layers and 5 max polling layers, the fully connected layers of the VGG19 CNN were altered in this work and replaced with other layers. The purpose of using the convolution layers (the operation of convolution is declared in (1) of the VGG is to make use of the pre-trained weights and not starting with a completely random weights, the network and the weights are loaded and used for feature extraction process only, the process were as follows: First: the network is loaded with the weights of image net dataset, which is a dataset that have over a million images and can classify more than 1000 object classes. Second: the network is trained with the proposed dataset in order to extract feature map using the convolution layers and the loaded weights, the layers of the VGG are as:

1	Conv3x3(64)	6	MaxPool(2,2)	11	MaxPool(2,2)	16	MaxPool(2,2)	21	Maxpool (2,2)
2	Conv3x3(64	7	Conv3x3(256)	12	Conv3x3(512)	17	Conv 3x3(512)		
3	MaxPool(2,2)	8	Conv3x3(256)	13	Conv3x3(512)	18	Conv 3x3(512)		
4	Conv3x3(128)	9	Conv3x3(256)	14	Conv3x3(512)	19	Conv 3x3(512)		
5	Conv3x3(128)	10	Conv3x3(256)	15	Conv3x3(512)	20	Conv 3x3(512)		

Where 3x3 means a 3 by 3 mask with stride 1 that will be convolved over the image while the number between brackets (64), (128), (265), (512) are the number of parameters in each layer and the numbers (2,2) are the mask of maxpool layer with stride2.

$$\text{Convolution}=\left|\frac{\sum_{i=1}^{q}\{\sum_{i=1}^{q} f(ij)d(ij)\}}{F}\right| \quad (1)$$

where: f(ij)=the coefficient of a convolution kernel at position (ij) in the kernel
d(ij)=the data value of the pixel that correspond to f(ij)
q=the dimension of the kernel if the kernal 3*3 then q=
F=either the sum of the coefficients of the kernel or 1 if the sum of the coefficient is zero
Convolution=the output pixel value

Maxpool=Maximum value of {4 values from the 2x2 maxpolling layer kernel} (2)

The layering of VGG is illustrated in Figure 2. After the extraction of the feature maps by using the VGG, the next step is to build a head model for classification process, the feature maps are fed to several layers as:

a. max pooling layer with pool size (3,3)
b. flatten layer
c. fully connected layer with 512 nodes
d. dropout layer with 0.5 percent

e. fully connected layer with 20 output nodes (number of classes) using soft max activation function.

The final step in the training process is to compile the model using stochastic gradient descent (SGD) optimizer with learning rate=0.0001 and momentum term=0.9 and decay=0.0001. The Gradient descent optimizer is a method to minimize an objective function J(θ) given parameter values by a model's parameters θ ∈ R d, it works by updating the parameters used in the model in the opposite direction of the gradient of the objective function ∇θJ(θ) to the parameters. The learning rate η determines the size of the steps we take to reach a (local) minimum. The SGD optimizer updates the parameters in each training epoch for training $x^{(i)}$ and label $y^{(i)}$ [27].

$$\theta=\theta - \eta \nabla\theta J(\theta; x(i); y(i)) \quad (3)$$

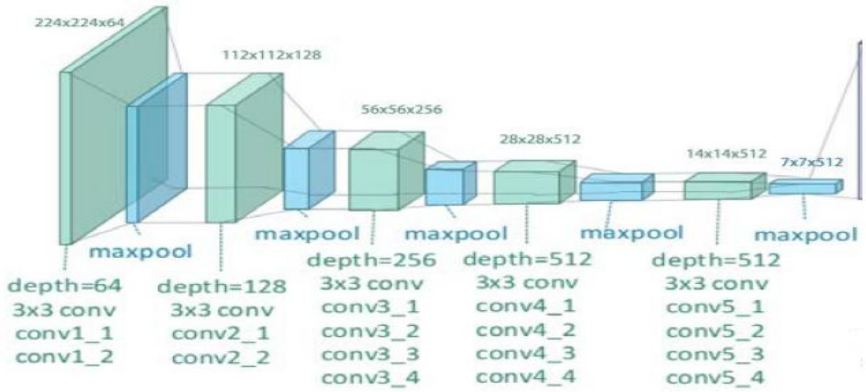

Figure 2. VGG architecture

The algorithm micro content recognition, illustrate the steps of the proposed model and Figure 3 shows the flow chart of the proposed model.

Algorithm Micro Content Recognition
Input: video
Output: Letter Label
Process **Step1:** convert video to frames **Step2:** face cropping using HAAR Cascade face recognition technique **Step3:** mouth cropping using spatial coordinate detection **Step4:** key frames selection
Step5: extracting labels from dataset **Step6:** reshape the frames into 224*224 images **Step7:** partitioning dataset into training and testing **Step8:** data augmentation **Step9:** data normalization **Step10:** using VGG model and image net weights for feature extraction **Step11:** building head base model for classification **Step11.1:** max pooling layer with pool size (3,3), **Step11.2:** flatten layer **Step11.3:** fully connected layer with 512 nodes **Step11.4:** dropout layer with 0.5 percent **Step11.5:** fully connected layer with 20 output nodes and soft max activation function **Step12:** compiling the training phase using SGD optimizer **Step13:** testing phase using precision, recall and F-score metrics

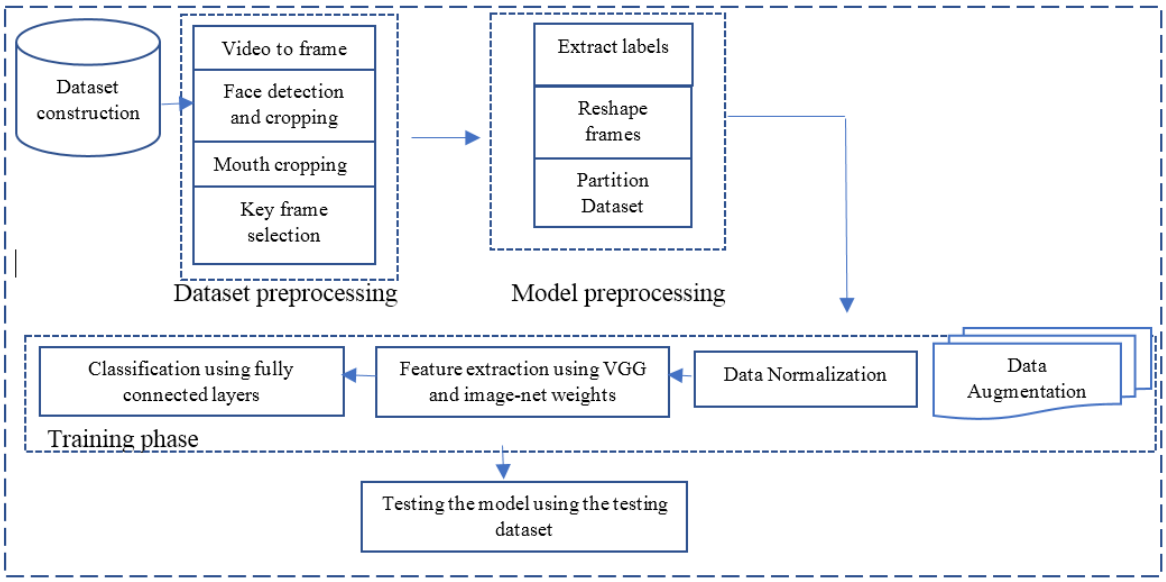

Figure 3. model flow chart

4. RESULTS AND DISCUSSION

The testing stage is implemented on 25% of the dataset, the model achieved a remarkable result on the testing set. Table 1 shows the results of the dataset. The results show that the training was successful and the model can recognize 20 letters with accuracy of 95% on the training dataset and 98% on the testing dataset, the training set had more near miss classification in regards to testing set near miss classification which led to slight difference in the computed accuracy.

Table 1. Measurements criterion results

Letters	precision	recall	f1-score	support
A	0.99	0.99	0.99	276
B	0.98	0.97	0.97	127
C	0.99	0.98	0.99	177
D	0.97	0.97	0.97	119
E	0.96	0.88	0.92	170
F	1.00	0.98	0.99	447
G	0.96	0.99	0.97	233
H	0.95	0.96	0.95	134
I	0.98	0.99	0.98	372
J	1.00	1.00	1.00	201
L	0.94	0.97	0.95	163
M	0.98	1.00	0.99	628
N	1.00	0.99	0.99	142
O	0.99	1.00	0.99	549
R	0.93	0.97	0.95	143
S	0.99	0.99	0.99	320
T	0.99	0.94	0.96	87
W	0.99	0.99	0.99	320
Y	0.99	1.00	0.99	292
Z	0.97	0.92	0.94	73
Total accuracy			0.98	5078

From the above table we can notice that several letters have results of 99-100 these letters had distinguished features that can more easily recognize them from other letter, whereas the letters with less than 99% accuracy they were more difficult to recognize due to the big similarity with other letters. This challenge of similar letters like the letter E which is very similar to letter A but the model recognize the frames that have the same features as A more than as E Although it was hard to distinguish them but the model achieved an excellent results, whereas the letter J had an accuracy of 100% because there were no other letter that have the same features as the letter J.

5. CONCLUSION

The proposed model for English alphabet lip reading succeed in achieving the aim of the model with high efficiency by using deep learning technique with a proposed dataset which was constructed by the author containing more than 2700 videos for 20 letters recorded for11 persons (male and female from different ages). From the experiment results, it is clear that the proposed model achieved an excellent recognition results for 20 letters English alphabet using deep learning, points below represent the proposed model conclusions: the use of an appropriate CNN model in regard of the number of layers avoid trapping in over fitting problem, when removing the letters that is very similar to other letters it enhanced the average accuracy, the preprocessing stage play an important role in achieving high accuracy recognition rate, this is clear by extracting the region of interest from the video frames which contains relevant effective features and ignoring unnecessary features that have negative impact on the recognition results. For the future work, a trial will be conducted to recognize whole words depending on the proposed model according to lip words reading, this is required labeling each resulted letter from the presented proposed model.

REFERENCES

[1] Z. Zhou, X. Hong, G. Zhao and M. Pietikäinen, "A Compact Representation of Visual Speech Data Using Latent Variables," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 1-1, Jan. 2014, doi: 10.1109/TPAMI.2013.173.

[2] A. G. Amit, J. N. Jnoyola and S. B. Sameepb, "Lip reading using CNN and LSTM," Technical report, Stanford University, CS231 n project report, 2016.

[3] A. Fernandez-Lopez, and F. M. Sukno," Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing*, vol. 78, pp: 53-72, 2018, doi: https://doi.org/10.1016/j.imavis.2018.07.002.

[4] A. M. Sarhan, N. M. Elshennawy, and D. M. Ibrahim, "HLR-Net: A Hybrid Lip-Reading Model Based on Deep Convolutional Neural Networks," *Computers, Materials & Continua*, vol. 68, no. 2, pp: 1531-1549, 2021, doi:10.32604/cmc.2021.016509.

[5] L. Giurgiu, "Microlearning an Evolving Elearning Trend," *Scientific Bulletin*, vol. 22, no. 1, 2017, doi: 10.1515/bsaft-2017-0003.

[6] F. Zantalis, G.s Koulouras, S. Karabetsos, and D. Kandris," A Review of Machine Learning and IoT in Smart Transportation," *Future Internet* , vol. 11, no. 4, 2019, doi: doi.org/10.3390/fi11040094.

[7] W. M. Salih, I. Nadher, and A. Tariq, "Deep Learning for Face Expressions Detection: Enhanced Recurrent Neural Network with Long Short Term Memory," *In book: Applied Computing to Support Industry: Innovation and Technology*, pp: 237-247, 2020, doi: 10.1007/978-3-030-38752-5_19.

[8] C. Drakidou, "Micro-learning as an Alternative in Lifelong eLearning," *Thesis for: Master's Advisor: Pr. Panagiotis Panagiotidis*, 2018.

[9] G. S. Mohammed, K. Wakil, and S. S. Nawroly," The Effectiveness of Microlearning to Improve Students' Learning Ability," *International Journal of Educational Research Review*, vol. 3, no. 3, pp: 32-38, 2018. doi: 10.24331/ijere.415824

[10] E. Rettger, "Microlearning with Mobile Devices: Effects of Distributed Presentation Learning and the Testing Effect on Mobile Devices," Ph.D. Dissertation, Arizona State University, USA, 2017.

[11] N. Friesen, "The Microlearning Agenda in the Age of Educational Media," Thompson Rivers University, Canada 2007.

[12] Y. Lu, and H. Li, "Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory," *Applied Sciences*, vol. 9, no. 8, p: 1599, 2019, doi:10.3390/app9081599.

[13] A. Mesbah, H. Hammouchi, A. Berrahou, H. Berbia, H. Qjidaa, and M. Daoudi, "Lip Reading with Hahn Convolutional Neural Networks moments," *Image and Vision Computing,Elsevier 88*, pp: 76-83, 2019, doi: 10.1016/j.imavis.2019.04.010.

[14] J. S. Chung, and A. Zisserman, "Lip Reading in Profile," *British Machine Vision Conference*, September 2017, doi: 10.5244/C.31.155.

[15] H. M. Cruz, J. K. T. Puente, C. Santos, L. A. Vea, and R. Vairavan," Lip Reading Analysis of English Letters as Pronounced by Filipino Speakers Using Image Analysis," *1st International Conference on Green and Sustainable Computing (ICoGeS) Journal of Physics*, vol. 1019, no. 1, p: 012041, 2017, doi :10.1088/1742-6596/1019/1/012041.

[16] M. Z. Ibrahim, and D. J. Mulvaney, "Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping," *Journal of Visual Communication and Image Representation*, vol. 30, pp 219-233, 2015, doi: https://doi.org/10.1016/j.jvcir.2015.04.013.

[17] C. C. Aggarwal, "*Neural Networks and Deep Learning*," *Springer*, vol. 10, p: 978, 2018.

[18] N. Buduma, and N. Lacascio," Fundamentals of Deep Learning Designing Next-Generation Machine Intelligence Algorithms," O'Reilly Media, Inc., pp: 92-122, 2017.

[19] Y. L. Cun, Y. Bengio, and G. Hinton "Deep learning Review," *Macmillan Publishers Limited*, vol. 521, pp: 436-444, 2015, doi:10.1038/nature14539.

[20] Y. Zhenga, C. Yangb, and A. Merkulov, "Breast Cancer Screening Using Convolutional Neural Network and Follow-up Digital Mammography," *Conference: Computational Imaging III*, 2018, doi: 10.1117/12.2304564.

[21] W. M. Salih, I. Nadher, and A. Tariq, "Modification of Deep Learning Technique for Face Expressions and Body Postures Recognitions," *International Journal of Advanced Science and Technology*, vol. 29, no. 3s, pp. 313-320, 2020.
[22] T. Ozcan, and A. Basturk, "Lip Reading Using Convolutional Neural Networks with and Without Pre-Trained Models," *Balkan Journal of Electrical & Computer Engineering*, vol. 7, no. 2, April 2019, doi: 10.17694/Bajece.479891.
[23] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.
[24] S. Skansi, "*Introduction to Deep Learning from Logical Calculus to Artificial Intelligence*," Springer, 2018.
[25] T. Bezdan, and N. B. Džakula, "Convolutional Neural Network Layers and Architectures," *International Scientific Conference On Information Technology and Data Related Research*, 2019, doi: 10.15308/Sinteza-2019-445-451.
[26] O. Jomah, A. K. Masoud, X. P. Kishore, and S. Aurelia, "Micro Learning: A Modernized Education System," *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, vol. 7, no. 1, pp: 103-110, 2016.
[27] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv:1609.04747v2 [cs. LG]*, 2017.

BIOGRAPHIES OF AUTHORS

Nada Hussain Ali: PhD student at University of technology, Iraq. She got her B.Sc and M.Sc Degree in computer science, from university of technology, Iraq. Her research interests include Artificial Intelligence, Image Processing, Machen Learning, Pattern Recognition

Matheel E. Abdulmunim: Professor qualified to Direct Research at University of Technology, Iraq. She got her B.Sc in 1995 from university of technology, Iraq, and her M.Sc degree in 2000 from university of technology, Iraq, and her Ph.D in 2004 university of technology, Iraq.

Akbas Ezaldeen Ali: Assist Professor qualified to Direct Research at University of Technology, Iraq. MSc. and Ph.D. in Computer Science from the University of Technology-Iraq/department of computer science in 1996 and 2016 respectively. The area of interest is image and video processing.