Abyss. J. Sci. Technol. Vol. 6, No. 2, 2021, 42-50
ISSN 2616 – 4728 (Online); 2616 – 471X (Print)

Abyssinia Journal of Science and Technology

Amharic Language Visual Speech Recognition using Hybrid Features

Zelalem Tamrie

Department of Computer Science, Kombolcha Institute of Technology,
Wollo University, Kombolcha, Ethiopia

ABSTRACT

Lip motion reading is a process of knowing the words spoken from a video with or without an audio signal by observing the motion of the lips of the speaker. In the previous studies its accuracy is limited because of not applying appropriate image enhancement methods and the algorithms used for feature extraction and feature vector generation. In the present study, we propose automatic visual speech recognition machine learning and computervision techniques for Amharic language lip motion reading. The objective of the study to improve the existing Amharic lip motion reading and the performance of speech recognition systems operating in noisy environments. The collected the video of Amharic speech by recording directly using mobile devices. In this study 14 Amharic words that are frequently talked by patients or health professional in the hospital were recorded. The total number of patients used for the study were 1260 (945 for training and 315 for testing our proposed model. To extract the features, we used Convolutional Neural Networks (CNN), Histogram of Oriented Gradients (HOG) and their combination methods were employed so as to extract the features. We feed these features to random forest independently and with combination to recognize the spoken word. Each of these features were tested by using precision, recall and fl-score classifiers for measuring the performance of our model and to compare the accuracy of our model with previous related works.Our model system records 66.03%, 75.24% and 76.51% accuracy on HOG, CNN and combined features (random forest), respectively.

Keywords: Convolution Neural Network, Histogram of Oriented Gradients, Lip motion reading, Random Forest.

INTRODUCTION

Lip motion reading is very crucial in order to help hearing-impaired people and even the normal people to have good communication when the audio is corrupted (Lu & Li, 2019). Many lip motion reading systems have been developed by traditional and end to end deep learning methods. Most traditional methods extract features based on appearance-based or image transforms and shape-based manner and feed this feature to Hidden Markov Model (HMM) while an end to end deep learning methods use deep learning algorithms for both feature extraction and classification. Traditional methods (Frew, 2019); (Wu & 'Ruan, 2014); (He, 2010) end to end deep learning methods (Lu & Li, 2019); (Faisal & Manzoor, 2018). In traditional methods, the feature extraction method is either appearance-based or shape-based. The problem encountered in appearance-based feature extraction is its performance is limited when the environment is changed and also when the face orientation is changed. In the shape-based feature extraction method, all pixels in the image are not considered when features are selected.

*Corresponding author: zelalem.t8@gmail.com

Machine learning methods have a great role in different social activities by solving different practical problems, especially in artificial intelligence, natural language processing, computer vision, etc. Nowadays human-machine interaction technology becomes multimedia and multimode technology from computer-centered to people-centered because of the high growth of artificial intelligence and the increasing popularity of the smart device. Using natural language to interact with a computer is inevitable or the most suitable and efficient one method between all ways of human-computer interaction methods (Lu & Liu, 2018).

In the past decades, this lip motion reading process using a machine learning method has two steps: feature extraction and classification. In order to extract features from the speaker most recent researchers use traditional machine learning algorithms (i,e. discrete cosine transform, discrete wavelet transform, principal component analysis, the HOG, etc.). After they extract features from the video they also use a traditional classifier to

classify. However, these traditional methods are outperformed by deep learning methods (Lu & Li, 2019).

An Amharic audio-visual speech recognition system was proposed by using traditional methods for both feature extraction and recognition recently. In this work the author firstly, record a video of phone (vowels) and isolated word for recognition secondly, change videos to image frames and calculate the region of interest(ROI) thirdly, feature extraction was takes placed by using Discrete Wavelet Transform (DWT) and finally the recognition was performed by using HMM. However, in this paper the main problem is Their performance is limited due to HMM limitation and in feature extraction they use DWT, they did not use any preprocessing like remove redundant frame enhancement, feature vector of this study was generated from a single frame. However, single frames do not represent the video. Because of this reason, they get an overall accuracy of 60.42% and 61%, for speaker-independent and speaker-dependent respectively (Frew, 2019).

However, the previous Amharic lip motion reading studies have not enough performance because of the methods used for detection of mouth, image enhancement, and extract features, feature vector generation. In addition, the previous studies process irrelevant data such as framing before face detection and do not remove redundant frames generated from identical videos. Therefore, it needs an improvement by applying appropriate methods for mouth detection, image enhancement, and extract features and by generating suitable feature vector. Irrelevant processing must be removed also, to decrease the processing computational resources ((Frew, 2019).

Recently, the growth of human-computer interaction and virtual reality is increasing exponentially. In nature, human communication is bimodal: acoustic and visual. Acoustic information may be lost due to nature like impairment or manmade noise in highly crowded traffic in a street, in a factory the sound generated from the motor, in the market, in the train station, in silence movies, etc (Thein & San, 2018). Researchers in the past decades on lip motion reading using the video sequence of the speaker's mouth has fascinated significant attention. Lip motion reading is a one way of speech recognition in which the spoken words were predicted by observing the movements of the lip.Several methods for lip motion reading wereproposed in the literature. Traditional and deep learning approaches are general categories. The design of this system depends on the choice of visual features, the classification approach, and the speech database used. In lip motion reading the spoken word may be differentiated by the angle of the sequence frames. This angle is detected by traditional feature extraction. And also texture features are extracted by the traditional extraction method. Deep feature extraction in many research areas provides a promising result.

Combining traditional and deep feature extraction methods enables us to use the above advantages. In general, Lip motion reading is an inevitable solution for the above problems. Therefore, in this study we implemented hybrid feature extraction method and this enables us to use the advantage of the two feature extraction methods. Therefore, the objective of the study was to improve the existing Amharic lip motion reading and the performance of speech recognition systems operating in noisy environments.

MATERIALS AND METHODS

The proposed system has three stages: preprocessing, feature extraction, and classification. In the preprocessing stage, we first detect the face. Since detecting face before applying any preprocessing tasks saves our computing time. Because our relevant information is located around the mouth and mouth are located on the face. For example, if we convert the video into frames before face detection, non-face frames may be converted and this non-face frame does not include our relevant information. Therefore, the profit here is nothing rather than defalcating our time. After detecting the speaker's face, we extract ROI/mouth because our sensitive/useful information for visual speech recognition is located on the mouth. Finally, we convert videos to image frame sequences that mean our ROI/lips counter. In feature extraction, we use both hand-crafted and deep features extraction techniques. From hand-crafted feature extraction machine algorithms, we use the HOG and from deep learning machine learning algorithms, we use a CNN. Hence, combined features extracted by different algorithms helps to use the advantage of both algorithms we combine features extracted by the HOG and convolution neural network. After extracting features in both methods, we classify using random forest independently and by concatenating both features. Classification encompasses three main phases: training, validation, and testing phase (Fig.1).

Preprocessing:

Before applying training or recognition on the image sequences in lip motion reading, visual data must be preprocessed to remove data irrelevant to speech and to enhance certain characteristics that are used to improve lip motion reading system performance. The preprocessing task in this study includes detecting the speaker's face, calculating the ROI (i.e. lip counter), image enhancement was

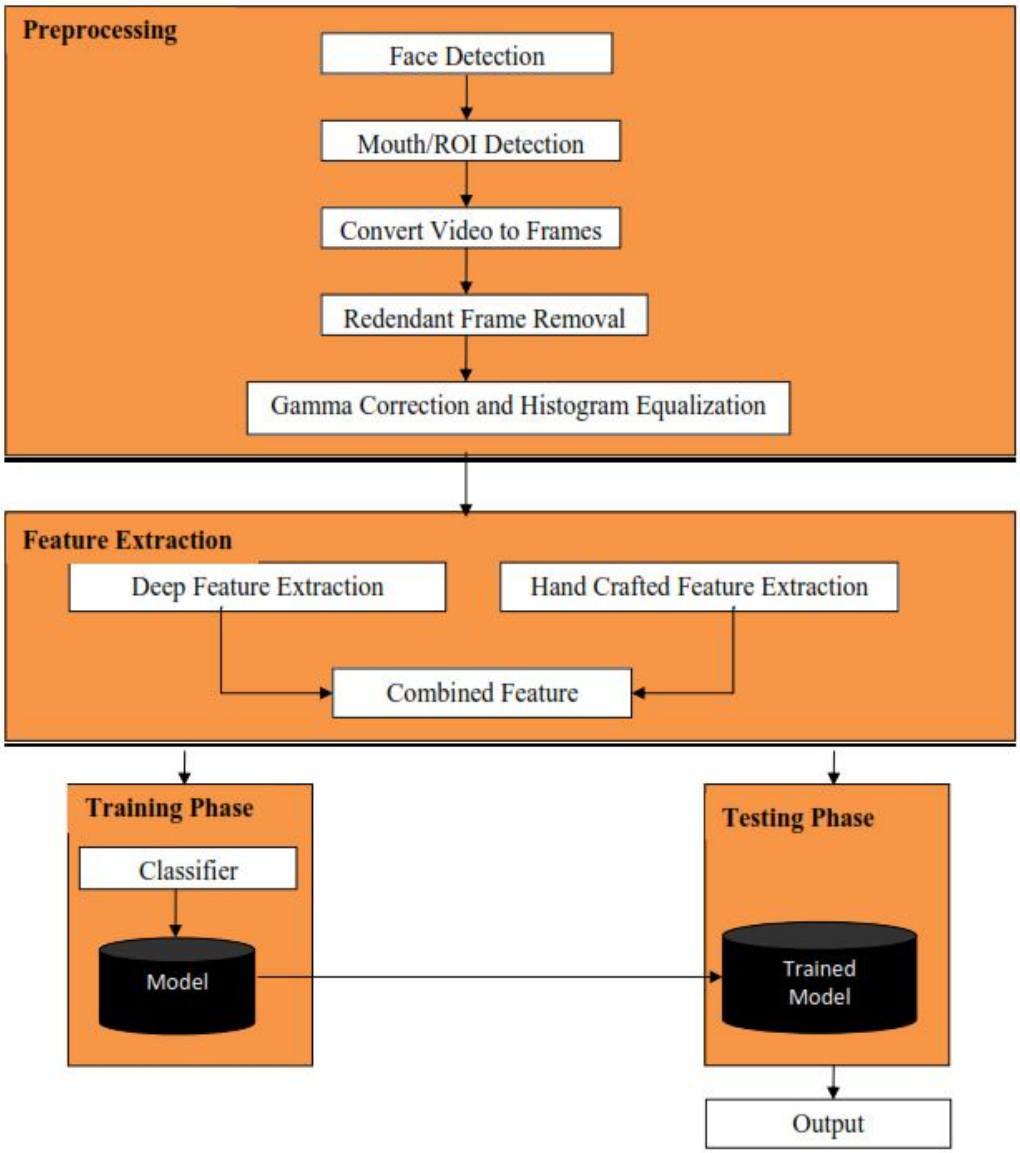

Fig. 1: Proposed system architecture

performed. In the following sub-sections, we discuss in detail the preprocessing tasks we have used throughout this study.

1) Face and mouth Detection: Before converting videos of the speaker when uttered the selected words were converted to equivalent frame sequence face detection must be performed. Because, the video may contain non-face frames, so converting this non-face frame is not necessary for lip motion reading since it does not contain the lip that is the most important region and it contains useful information for the lip-reading system. In addition to this, converting non-face frames to an image increases our computational time and requires high memory. Therefore, after recording the videos uttered by the speaker face detection was performed on this video. Several methods are available in the literature for face detection. Viola-Jones object recognizer was implemented in this study because it is simple rectangular Haar features and is applied to each frame in a wide range of translations and at many different scales. To select specific Haar features, the AdaBoost technique was used to train a weak classifier. Single strong object classifiers can be formed by cascading such weak classifiers. In addition to this viola-jones algorithm is robust, real-time, features are computed very quickly, feature selection is efficient, and also it is a generic detection scheme that can be used to detect other objects. Following face detection, mouth detection was performed.

2) Remove Duplicated Frames: The speakers' lip motion may become identical by different factors such as the nature of the visem, the speaking style of the speaker. This identical motion produces duplicate frames since frames were created from videos and the movement is identical. Processing these duplicated frames increases our computational resources. To reduce this problem, we remove duplicate frames. Two operations were

performed to remove duplicated images: (i) we compare the pixel values of adjacent frames. (ii) Remove the duplicated frame if they are equally based on the comparison performed in the above step. The following algorithm shows clearly the steps to remove duplicated frames.

3) Image Enhancement: The image quality may be affected by different factors. Among these factors, the air condition, camera nature, the cameraman is some of the common factors. In order to get useful information, the quality of the frames must be increased. To increase the image quality image enhancement techniques were applied. Enhancement is a process of increasing the image effectiveness for the computer process by increasing its clarity or quality. There are many methods to enhance the image in the literature. Two major categories of image enhancement techniques exist: global and local. In the global method of enhancement consider the global pixel of the image and then adjust based on global information. However, a different region of the image may not require a different level of enhancement, and applying global enhancement removes the detail of this region. In the local enhancement method, the neighboring pixel was considered and then adjustments are applied based on the local information. Even though the enhancement by the local method is advisable, it requires high computational time. Gamma correction and contrast limited adaptive histogram equalization was adopted one after the other in this study. Because both enhancement methods have a minimum signal to noise ratio did not remove detail of the image.

Feature extraction:

For the lip motion reading feature extraction process can be either traditional (hand-crafted) or end to end deep teaching (Fernandez-Lopez &Sukno, 2018). The work (Frew, 2019) Uses DWT to extract features from ROI identifies 8 features from ROI like height, width, area, etc. The author in (Lu & Li, 2019) uses CNN for feature extraction. DCT and LBP are also used by for feature extraction of visual speech recognition. The performance of the lip motion reading system is better when applying deep learning architectures compared to traditional systems (Fernandez-Lopez &Sukno, 2018). HOG is also used for feature extraction in the Hindi language (Upadhyaya, Farooq, &Abidi, 2018). In this study, CNN and handcrafted based feature extraction is adopted. Therefore, for hand-crafted feature extraction, we use the HOG. Because HOG computes edge gradients and this enables it to capture the shape of the image. Features are extracted from each frame of videos. After extracting features from each frame average feature is computed from each frame feature. Both handcrafted and deep features are fed to classifiers independently and after combining the features.

i. Traditional Feature Extraction: In traditional feature extraction method features are appearance-based or shape-based. HOG is implemented in this study for hand-crafted features. Because it is widely used in the literature and the features in this technique are calculated from magnitude and orientation. This feature extraction is grouped under appearance-based feature extraction in the traditional feature extraction method. In this method, two values are calculated. The first one is the magnitude and the second one is orientation. To calculate the magnitude first gradients must be calculated. To calculate the gradient first select the central pixel from the region of the image. After the central pixel is selected gradient of X and Y is calculated as: (Singh, 2020).

$$GX(m, n) = I(m, n + 1) - I(m, n - 1) \quad (1)$$

$$GY(x, y) = I(x, y - 1) - I(x, y + 1) \quad (2)$$

Where GX and GY is the X and Y gradients, $I(m, n + 1)$ and $I(m, n - 1)$ is the pixel value from the right and left to the selected central pixel value respectively and $I(x, y - 1)$ And $I(x, y + 1)$ the pixel value from the top and bottom to the selected central pixel value respectively.

Following gradients calculation, the magnitude for each pixel was calculated by applying Pythagoras theorem as follows:

$$\text{Total Magnitude} = \sqrt{GX2 + GY2} \quad (3)$$

After calculating the magnitude of each pixel, the next step is calculating the orientation by using its magnitude. The orientation or direction of each pixel was calculated as follows:

$$\tan \alpha = GY/GX \quad (4)$$

After calculating the magnitude and the orientation the final step is generating the histogram from these two values. The histogram is a graphical representation of continuous data in correspondence with its frequency. Therefore, to create the histogram we put the angle or orientation on the x-axis and the frequency on the y-axis.

ii. Deep Feature Extraction: In this study, we use CNN as a feature extraction method for deep feature extraction. Because CNN has great power in image processing in the literature. Images with size MxN are the inputs in CNN based feature extraction. This image passed through many layers of the so-called hidden layers and by applying operations like convolution, pooling on the image

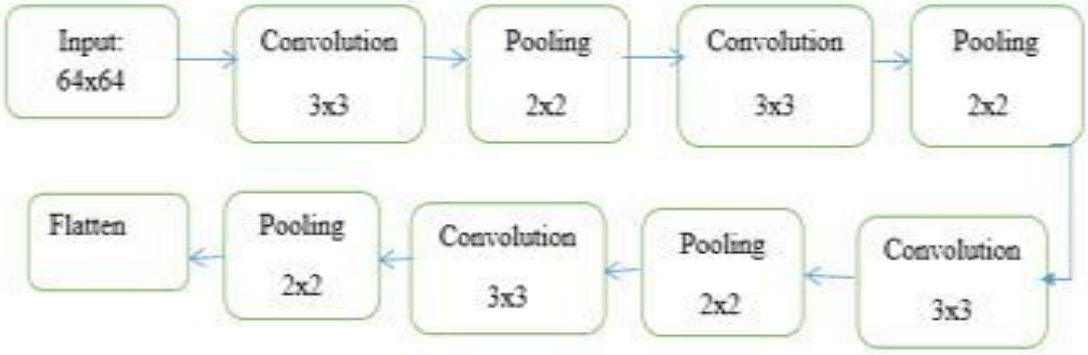

Fig.2: CNN Architecture

we get useful information for the recognition stage. After through experiment, we select relu activation function because our model performance is better when compared with other activation functions. We use maximum pooling because it reduces the dimension of the feature vector and it is also suitable for an image that has noise. The following illustrates demonstrates the architecture of our network (Fig. 2).

As we see in the above figure 2 the first stage is providing the input image by specifying its height and width. The height and width of the image are set to be 64 x 64. Bicubic Interpolation image resampling techniques are selected for image resizing. Because it effective in all applications of image processing (S & Patil, 2018). This size is selected randomly and it is within the range that CNN performs better in the literature which is 64 up to 360. The next stage in the network is convolution and pooling layers. The network contains four convolution and four pooling layers. The kernel size of the convolution and pooling layer is set to be three and two respectively. Finally, the feature obtained from the last pooling layer was fed to the flatten layer and this layer changes the feature from multi-dimensional to one-dimensional vector.

iii. Average Feature: As we have discussed in the above sections a video is a sequence of frames. A single video in our study is the utterance of a single word. The number of frames is different from video to video. This is because of the length of the word and the speed of the speaker. The number of frames defined in this study is based on the Frame per Second (FPS) of the camera and our camera is 30 FPS. When we predict the words from a video a single frame has not complete information. The video to frame relationship mathematically is:

$V = \{Fj, Fj + 1, Fj + 2 \dots , Fm\}\ 0 < j < m$ (5)

Where Fi is the frame created at j time and V is video n is the number of frames in a video. Therefore features must be extracted from all frames of a video and then the average features are selected. The feature of a single frame is represented as follows:

$fFi = \{fi, fi + 1, fi + 2 \dots \dots , fn\} 1 < i < n$ (6)

Where n isthe length of the feature vector extracted by using CNN and HOG, fi is the single feature of a frame (Fig.3).

INPUT: frame in videos
Step 1: declare list to store features, store average feature
Step 2: for frame in video
    2a: read frame
    2b: extract features by HOG and CNN
    2c: append features on the declared list
Step 3: for each column in the list
    3a: summing up column values to single value
    3b: divide sum in step 3a by the number of columns
    3c: append value of 3b to list declared to store average feature
OUTPUT: average feature

Fig. 3: Algorithm for average feature extraction

Therefore, in order to get useful information in a video, we must include the features extracted from each frame sequence. The average feature is calculated as shown in the following equation.

$$favg(i) = \sum_{i=1}^{n} \left(\frac{(fi)}{n} \right) 0 < i < k$$

(3.14)

Where $favg(i)$ is the average feature at location i, fi is the feature of each frame at location i, n is the number of frames and k is the length of a feature vector. The algorithm shows the algorithm in order to calculate the average feature (Fig. 3).

Classification:

For recognition purpose, we used random forest classifier. In machine learning, the challenging task is selecting the optimal hyper parameters of each machine learning algorithm. We use grid search algorithm for selecting the optimal parameter in this study on random forest.

RESULTS

Dataset:

Videos recorded from different speakers were the datasets for this study. Because we did not find the data collected by previous researchers, we recorded14 isolated words. Each word was uttered by 9 subjects (four males and five females) whose age is between 20 and 30 times. Each word was talked 10 times by a separate subject. Therefore, a total of 1, 260 videos were recorded. Samsung Galaxy A30s mobile phone with a triple camera (i.e 25MP+8MP+5MP) and 1920x1080 pixel full HD was used as a video recorder. The video format was MP4 and its frame was 30FPS. In order to minimize noise, we recorded video for this we used the YNUTFNG fixer that was used to hold on a stable condition and we also used YNUTFNG wireless camera controller to control the camera. The selected words used for this study are showed in (Table 1).

In this study, we used precision, recall, and f1-score for measuring the performance of our model. In addition, we have also calculated the micro-average, macro-average, and weighted average for all the aforementioned performance metrics. Our features were extracted by using HOG, CNN and CNN and HOG combined feature extraction techniques). Each feature was tested by using the above classifier. Finally, we compared our model with previous done related works. The data set was portioned in to training and testing test: 75% of our data is the training set and the remaining 25% was our testing set.

Table 1: Selected words used for the study

Index	Amaharic word (አማርኛ)	Word (English Version)
1	ራስምታት [rasemetate]	Headache
2	ትኩሳት [tikusat]	Fever
3	ማዞር [mazor]	Dizziness
4	አሞኛል [āmonyal]	I'm sick
5	ማላብ [malab]	Sweats
6	ጭንቀት [ch'ink'et]	Depressed
7	ድካም [dkam]	Fatigue, Tired and Weak
8	ውሃ [wha]	Water
9	ምግብ [mgb]	Food
10	ሰገራ [segera]	Stool
11	ሽንት [shnt]	Urine
12	ሐኪም [hạ̄ kīm]	Doctor
13	አዎ[aāwo]	Yes
14	ተሸሎኛል [teshlonyal]	I am better

Result of random forest on HOG feature extraction method:

The result of random forest on HOG feature: In this experiment, we use features extracted by HOG that have 144 dimensions and 945 training samples used as input for the RF classifier. The remaining 315 samples are used for testing purposes. The classification result of this experiment is shown (Fig.4). Based on this experiment the least recognition rate was the word 'ማላብ'.

Result of random forest on CNN feature extraction method:

Similarly, to the above experiment the features extracted by CNN used the same samples for training and testing in the experiment above using RF classifier. The result of this experiment done on random forest extraction using CNN extraction technique is shown in (Fig. 5). As shown in figure 5 the word 'ድካም' has the least recognition rate (55%) , this was because of its 18% sample is classified as ማላብ, 9%, 5%,5%,5%,5% of its sample are classified as ምግብ, አዎ, ጭንቀት, ራስምታት, ትኩሳት, respectively. The word ማዞር has the highest recognition rate.

Result of random forest on combined feature extraction method:

Here, we also tested the features created by combining CNN and HOG using RF classifier with the samples used in the above experiment, and the result of this experiment shown(Fig.6). In this experiment, the 'ድካም' has the least recognition rate (55%). And, words such as ጭንቀት, ምግብ, አሞኛል, ማላብ, አዎ,ተሸሎኛል, had 9%, 9%, 9%, 9%, 5%,5%, recognition rates, respectively. The word ማዞር has the highest recognition rate.

DISCUSSION

Features of this study were extracted by using HOG and CNN and CNN and HOG combined feature extraction methods. Therefore, each feature was tested by the classifier. Finally, we compare our model with previous related work. The data set of this experiment wasdivided in training and testing test. From the total samples (1260), 75% (945) of the samples for used for the training set and the remaining 25% (315)was our testing set purpose. In this study, we have used precision, recall, and f1-score for measuring the performance of our model.

Based on the Random Forest on HOG feature extraction experiment shown in figure 4 the least recognition rate was the word 'ማላብ' this because the visual information was similar to many classes,

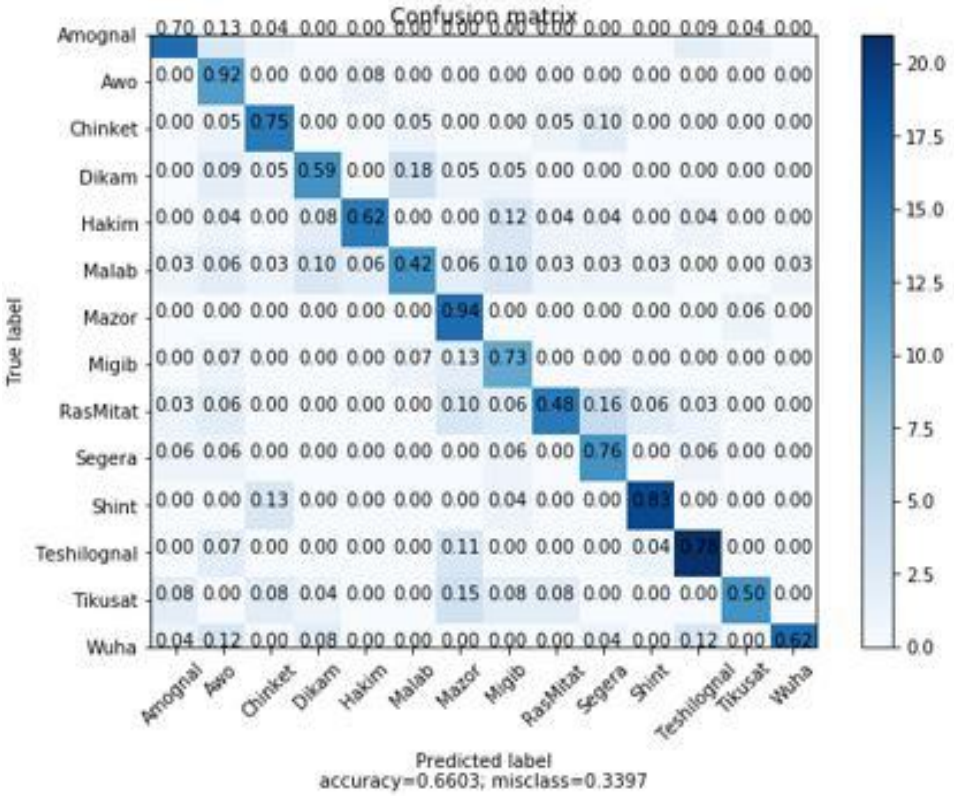

Fig. 4: Test result of HOG feature using RF

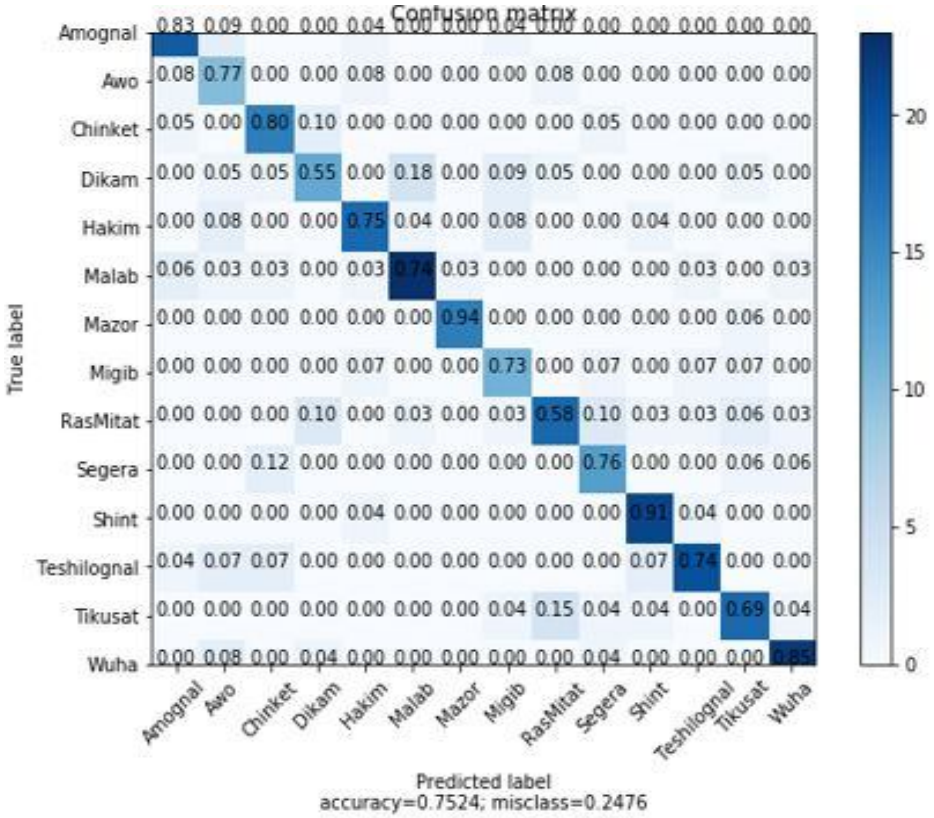

Fig. 5: Test result of CNN feature using RF

for example , it was similar with the word 'ምግብ'. This means 10% of 'ምግብ' was classified as 'ማላብ', which was also the same for the word 'ድካም'(Fig.4).

The result of random forest on CNN feature extraction the word 'ድካም' has the least recognition rate (55%), this was because of its 18% sample was classified as ማላብ, 9%, 5%,5%,5%,5% of its sample were classified as ምግብ, አዎ, ጭንቀት, ራስምታት, ትኩሳት respectively. The word ማዞር has the highest recognition rate (Fig. 5).

The result of this experiment based on random forest on HOG and CNN combined feature extraction method), the word 'ድካም' has the least recognition rate (55%). And, words ጭንቀት, ምግብ, አሞኛል, ማላብ, አዎ,ተሸሎኛል, had 9%, 9%, 9%, 9%, 5%,5%, recognition rates, respectively. The word ማዞር has the highest recognition rate (Fig.6)

Moreover, in this study model records of features extracted using HOG, features extracted using CNN and features extracted on HOG and CNN combined feature extraction methods which were measured in RF 66.03%, 75.24% and 76.51% accuracies on HOG, respectively. The accuracies of each feature extraction methods were higher as compared to the finding of (Frew, 2019), who reported an overall accuracy of 60.42% and 61%, for speaker-independent and speaker-dependent respectively. This might be due to the limitations of feature extraction methods HMM and DWT used for his study; and he didn't use any preprocessing to remove redundant frame enhancement, the feature vector used in his study was generated from a single frame.

The result of the present study was also higher than 62% and 70% accuracies which were recorded in speaker-independent and speaker dependent recognition, respectively which was reported by (Barnard & Pietik¨ainen, 2009); and 62 %, 56 % accuracies for word recognition using Long Short-Term Memory LSTM and Deep Neural Network (DNN) respectively (Faisal & Manzoor, 2018). Another study (Befkadu, 2019) reported 60.42% and 61% accuracies for speaker dependent and speaker independent. The reasons for this variation might be also the tradition feature extraction method and the number of 30 Amharic words used in his experiment.

However, the finding of the present study was

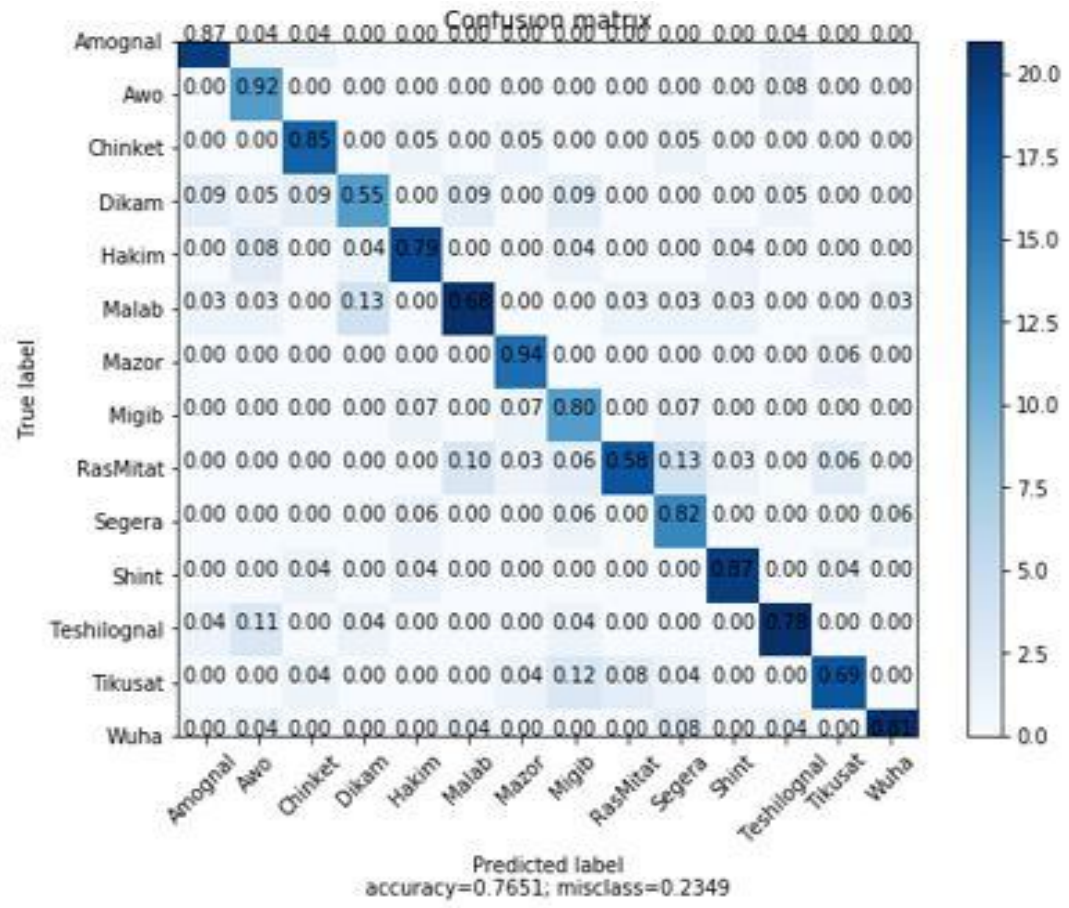

Fig. 6: Test result of combined feature using RF

lower than the finding of (Lu & Li, 2019). The reason for this difference might be the method they used. In their study they used CNN in a combination with attention-based long short-term memory which was implemented to recognize digits using lip reading. Moreover, their data was the English numbers from zero to nine and this has a big difference in phonological structure with the Amharic language. The over accuracy registered in their study was 88.2%.

In conclusion, visual lip movement was used as principal source of speech information in automatic speech recognition systems. The purpose of this study was to develop an automatic visual speech recognition for Amharic language using the lip movement detection, region of interest (ROI), visual features extraction, visual speech recognition. The architecture of the system that we adopted in our study is the decision fusion architecture. As a result of this architecture, we used three classifiers (precision, recall and fl-score). Speakers' mouth was the primary source of information for visual speech recognition. The video, features were extracted by using HOG and CNN and CNN and HOG in combination). Each feature was tested by the classifiers (precision, recall, and f1-score for measuring the performance of our model). Fourteen Amharic words that were frequently talked by patients were recorded and used for this study. The total number of patients used for the study were 1212 (960 for training and 252 for testing our proposed model. Our model records 66.03%, 75.24% and 76.51% accuracies on HOG, CNN and combined features (random forest), respectively. These accuracies are slightly higher than from most of previous related works. The performance of each featureswas evaluated by RF by different performance measurement metrics. Among the three feature extraction methods the combined feature was outperformed when we compared with HOG and CNN feature extraction methods. Our systems(models) recorded 66.03%, 75.24% and 76.51% on features extracted using HOG, features extracted using CNN and combined features measured in RF classifier, respectively. Therefore, the proposed model improves the existing model around 10% accuracy when compared with the previous study even if the dataset was different.

REFERENCES

Faisal, M., & Manzoor, S. (2018). Deep learning for lip reading using audio-visual information for urdu language. *arXiv preprint arXiv:1802.05521*.

Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, *78*, 53-72.

Frew, B. B. (2019). Audio-Visual SpeechRecognition using LIP Movement for AmharicLanguage. Inter*national Journal of Engineering Research & Technology*, *8(08)*, 594-604.

He, L. (2010). Study on Lipreading Recognition Based on Computer Vision. *Institute of Electrical and Electronics Engineers*, 1-4.

Lu, Y., & Li, H. (2019). Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. *Applied Sciences*, *9*(8), 1599.

Lu, Y., & Liu, Q. (2018). Lip segmentation using automatic selected initial contours based on localized active contour model. *EURASIP Journal on Image and Video Processing*, *2018*(1), 1-12.

Jain, A., &Rathna, G. N. (2017). Visual speech recognition for isolated digits using discrete cosine transform and local binary pattern features. *Institute of Electrical and Electronics Engineers,* 368- 372.

Manjunath, S., & Patil, M. M. (2018). Interpolation Techniques in Image Resampling. *International Journal of Engineering Technology*, 7, 567- 570.

Thein, T., & San, K. M. (2018). Lip movements recognition towards an automatic lip-reading system for Myanmar consonants. *Institute of Electrical and Electronics Engineers,*791-796.

Upadhyaya, P., Farooq, O., & Abidi, M. R. (2018). Block Energy Based Visual Features Using Histogram of Oriented Gradient for Bimodal Hindi Speech Recognition. *Procedia Computer Science*, *132*, 1385-1393.

Wu, D., &Ruan, Q. (2014). Lip reading based on cascade feature extraction and HMM. Institute *of Electrical and Electronics Engineers*, 1306-1310.

Zhao, G., Barnard, M., & Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, *11*(7), 1254-1265.