RESEARCH ARTICLE OPEN ACCESS

Performance Evaluation Hybrid Visual Speech Recognition Features On Arabic Isolated Words

Nour Sami Ghadban [1], Jafar Alkheir [2], Mariam Saii [3]

PhD student [1], Associate Proffesor [2] & [3]

Department of Computer and Control Engineering

Tishreen University

Syria

ABSTRACT

Visual recognition systems based on the movement of speaker lips are one of the most modern and important recognition systems currently and have received a great deal of attention in the last decade for their potential use in the latest applications that rely on speech recognition systems. It is one of the modern researches used to build systems for understanding or interpreting Speech without hearing it for speakers.

The research suggests using a hybrid system for visual recognition of isolated words spoken in Arabic. This system is a supervised and an offline system. It passes through several stages, beginning with identifying facial features followed by detection and identification of the lips, and then reading the lips to extract a set of visual features that indicate the spoken word. Hidden Markov models were used in the classification phase.

The proposed system was tested on 4155 samples. The results showed that the proposed hybrid system led to rate of up to 73% achieving an increase of up to 12% from the research based on the same database Which can be considered as a starting point in building recognition systems that adopt the integration of audio and visual features together.

Keywords:- Visual speech processing, Visual features, Hybrid approaches, Mouth location/tracking.

I. INTRODUCTION

Lip reading is used to understand or interpret speech without hearing it, a technique especially mastered by people with hearing difficulties. The ability to lip read enables a person with a hearing impairment to communicate with others and to engage in social activities, which otherwise would be difficult. Recent advances in the fields of computer vision, pattern recognition, and signal processing has led to a growing interest in automating this challenging task of lip reading. Indeed, automating the human ability to lip read, a process referred to as visual speech recognition (VSR) (or sometimes speech reading), could open the door for other novel related applications.

VSR has received a great deal of attention in the last decade for its potential use in applications such as human-computer interaction (HCI), audio-visual speech recognition
(AVSR), speaker recognition, talking heads, sign language recognition and video surveillance. Its main aim is to recognise spoken word(s) by using only the visual signal that is produced during speech. Hence, VSR deals with the visual domain of speech and involves image processing, artificial intelligence, object detection, pattern recognition, statistical modelling, etc.

II. FACIAL FEATURE DETECTION

Facial feature detection (FFD) is a necessary front-end to any AVSP system.

In (Ghadban, N. S. et al 2017) article the two basis tasks of eye detection and mouth location/tracking have been defined as being essential in any workable AVSP system. This task can be in many respects simplified by restricting the scope of the problem (i.e. single subject, frontal pose, static background, etc.), but still remains a very difficult problem. Error metrics for evaluating the effectiveness of an FFD system have also been defined.

Although the Viola Jones algorithm achieves a 100% rate, it does not accept rotation in the head except by 10 degrees. Therefore, we add the automatic identification and identification of the skin to overcome this.

III. VSR LITERATURE REVIEW

Most of the work done on VSR came through the development of AVSR systems, as the visual signal completes the audio signal, and therefore enhances the performance of these systems. Little work has been done using the visual only signal. Most of the proposed lip reading solutions consist of two major steps, feature extraction, and Visual speech feature recognition. Existing approaches for feature extraction can be categorised as:

1. Geometric features-based approaches - obtain geometric information from the mouth region such as the mouth shape, height, width, and area.

2. Appearance-based approaches - these methods consider the pixel values of the mouth region, and they apply to both grey and coloured images. Normally some sort of dimensionality reduction of the region of interest (ROI) (the mouth area) is used such as the principal component analysis (PCA), which was used for the Eigenlips approach, where the first n coefficients of all likely lip configurations represented each Eigenlip.

3. Image-transformed-based approaches - these methods extract the visual features by transforming the mouth image to a space of features, using some transform technique, such as the discrete Fourier, discrete wavelet, and discrete cosine transforms (DCT).

These transforms are important for dimensionality reduction and to redundant data
elimination.

4. Hybrid approaches, which exploit features from more than one approach.

3.1 Geometric features-based approaches

A geometric features-based approach includes the first work on VSR done by Petajan in 1984, who designed a lip reading system to aid his speech recognition system. His method was based on using geometric features such as the mouth's height, width, area and perimeter (Petajan, 1984).

Another recent work in this category is the work done by (Werda et al., 2007), where they proposed an Automatic Lip Feature Extraction prototype (ALiFE), including lip localization, lip tracking, visual feature extraction and speech unit recognition. Their experiments yielded 72.73% accuracy of French vowels, uttered by multiple speakers (female and male) under natural conditions.

3.2 Appearance-based approaches

Eigenlips are the compact representation of mouth Region of Interest using PCA. This approach was inspired by the methods of (Turk & Pentland, 1991), and first proposed by (Bregler & Konig, 1994). Another Eigenlips-based system was investigated by (Arsic & Thiran, 2006), who aimed to exploit the complementarity of audio and visual sources.(Belongie & Weber, 1995) introduced a lip reading method using optical flow and a novel gradient-based filtering technique for the features extraction process of the vertical lip motion and the mouth elongation respectively.

In a more recent study, (Hazen et al., 2004) developed a speaker-independent audio-visual speech recognition (AVSR) system using a segment-based modelling strategy. This AVSR system includes information collected from visual measurements of the speaker's lip region using a novel audio-visual integration mechanism, which they call a segment-constrained Hidden Markov Model (HMM). (Gurban & Thiran, 2005) developed a hybrid SVM-HMM system for audio-visual speech recognition, the lips being manually detected. The pixels of down-sampled images of size 20 x 15 are coupled to get the pixel-to-pixel difference between consecutive frames. (Saenko et al., 2005) proposed a feature-based model for pronunciation variation to visual speech recognition; the model uses dynamic Bayesian network DBN to represent the feature stream.

(Sagheer et al., 2006) introduced an appearance-based lip reading system, employing a novel approach for extracting and classifying visual features termed as "Hyper Column Model" (HCM). (Yau et al., 2006) described a voiceless speech recognition system that employs dynamic visual features to represent the facial movements. The system segments the facial movement from the image sequences using motion history image MHI (a spatio-temporal template). The system uses discrete stationary wavelet transform (SWT) and Zernike moments to extract rotation invariant features from MHI.

3.3 Image-transformed-based approaches

(Lucey & Sridharan's, 2008) work was designed to be posing invariant. Their audio-visual automatic speech recognition was designed to recognize speech regardless of the pose of the head, the method starting with face detection and head pose estimation. They used the pose estimation method described by (Viola & Jones, 2001). The pose estimation process determines the visual feature extraction to be applied either on the front face, the left or the right face profile. The visual feature extraction was based on the DCT, which was reduced by the linear discriminative analysis (LDA), and the feature vectors were classified using HMM.

A very recent study which also fits into this category was done by (Jun & Hua, 2009), where they used DCT for feature extraction from the mouth region, in order to extract the most discriminative feature vectors from the DCT coefficients. The dimensionality was reduced by using LDA. In addition, HMM was employed to recognize the words.

3.4 Hybrid approaches

(Neti et al., 2000) proposed an audio-visual speech recognition system, where visual features obtained from DCT and active appearance model (AAM) were projected onto a 41 dimensional feature space using the LDA. Linear interpolation was used to align visual features to audio features.

A comparative Viseme recognition study by (Leszczynski & Skarbek, 2005) compared 3 classification algorithms for visual mouth appearance (Visemes): 1) DFT + LDA, 2) MESH + LDA, 3) MESH + PCA. They used two feature extraction procedures: one was based on normalized triangle mesh (MESH), and the other was based on the Discrete Fourier Transform (DFT), the classifiers designed by PCA and LDA.

Yu (2008) made VSR the process of recognizing individual words based on a manifold representation instead of the traditional visemes representation. This is done by introducing a generic framework (called Visual Speech Units) to recognize words without resorting to Viseme classification.

The previous approaches can be further classified depending on their recognition and /or classification method. Researchers usually use dynamic time warping (DTW), e.g. the work

done by Petajan. Artificial neural networks (ANN), e.g. the work done by Yau et al. and Werda et al.. Dynamic Bayesian Network (DBN), e.g. the work done by Belongie and Weber, and support vector machines (SVM), e.g. the work done by Gurban and Thiran, and Saenko et al.

The most widely used classifier in the VSR literature is the hidden Markov models (HMM).

Methods that use HMM include Bregler and Konig; Neti, et al.; Potamianos et al.; Hazen et al.; Leszczynski and Skarbek; Arsic and Thiran; Sagheer, et al.; Lucey and Sridharan; Yu; and Jun and Hua.

IV. MATERIALS AND METHODS

The algorithm of motion tracking techniques (Ghadban, N. S. et al 2017) used in Hollywood films drew the small green dots on the face, providing a "good" basis for isolating the important area of the mouth and understanding the spoken word:

- Determine the stages of mask implementation to identify the lips to isolate the mouth area.
- Extract features.

There are five steps to build a lip mask**:**

- Crop frame to lip region
- Find green markers
- Remove artefacts
- Reconstruct any missing markers
- Working with markers as objects
- Construct mask polygon

Features Extracted are:

Feature 1 - Width and Height of Lips
Feature 2. Relative Distance of Markers to Centre Point
Feature 3 - Inner Mouth Pixel Count
Feature 4 - DCT representation

V. RESULTS

The system was tested on 4155 sample videos for isolated words spoken in Arabic. These videos are subject to ideal conditions such as lighting.

Several scenarios were used to extract features from the studied videos depending on the methodology used:

- Geometric features-based approaches
- Appearance-based approaches
- Image-transformed-based approaches
- Integrate each approach with each other

- Integration of the three curricula to form a hybrid approache
Figure (1) shows the difference in the recognition ratio according to the features adopted in the identification, where the symbols symbolize: w-H length and width, R: distance changes to the distance of marks from the center (Geometric features-based approache).

DCT: (Image-transformed-based approache)
I : Number of inner mouth pixels (Geometric features-based approache).

The results showed that the proposed algorithm based on the hybrid approach showed that the recognition rate reached 73% and proves that it is complementary to the audio signal and will provide clear results in the development of audiovisual systems based on the integration of audio and visual features.

In comparison with the research conducted in English and adopted the hybrid approach, the highest achieved 94% and justify the reason for the low rate of recognition in the Arabic language than in English because of the privacy of words spoken in the Arabic language there may be some characters that are not spoken in the Arabic language The study did not rely on the integration of works as the proposed research, based on the test of the workbook with different values and the optimal choice that achieves a better knowledge rate. He found that the number of cases of the ideal work is 24 cases.

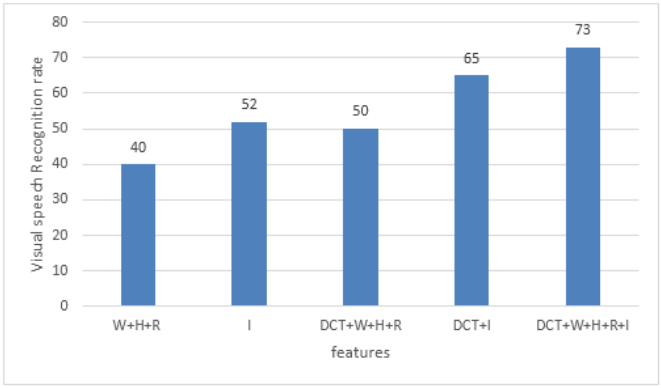

Figure 1: Visual Recognition rate according to the applied Features.

VI. SUMMARY

In this research, we have reached:
• Extracting attributes based on the hybrid approach and determining the optimal number of cases of the workbook to obtain a percentage of knowledge when integrating these features and access to the optimal number of cases of the workbook up to 73% in an ideal environment.
• Studying the effect of different features on the recognition rate.
We can make recommendations for:
• Further testing and evaluation by adding other features and examining their impact on the system.
• Combine visual evidence and audio guides to achieve better speech recognition.
• Test new works or merge several workbooks and study their impact on the system.

ACKNOWLEDGMENT

The research that has led to this work has been supported in part by the Tishreen University Enterprise.
The authors wish to thank Prof Dr.Alaaa Sagheer for his help on the psychological aspects of this work and for the technical support and Submit his database to do our research.

REFERENCES

[1]Ahmad B. A. Hassanat (2011). Visual Speech Recognition, Speech and Language Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-322-4, InTech, Available from:http://www.intechopen.com/books/speechand language-echnologies/visual-speech-recognition

[2]AlaaSagheer, Multimodal Arabic Speech Recognition for Human-Robot Interaction Applications, *Appl. Math. Inf. Sci.* 9, No. 6, 2885-2897 (2015).

[3]Antar,S.; Sagheer,A.; Aly,S.; Tolba.AVAS: Speech datab asefor multimodal recognition applications Hybrid Intelligent Systems (HIS), 2013 13th International Conference on Year:2013 Pages: 123-128.

[4]Cox, S. I. Matthews, and J. A. Bangham, \Combining noise compensation with visual Information in speech recognition," in Auditory-Visual Speech Processing, (Rhodes), 1997.

[5]Dodd, B. and R. Campbell, eds., Hearing by Eye: The Psychology ofLipreading. London, England: Lawrence Erlbaum Associates Ltd., 1987.

[6]Dodd, B. and R. Campbell, eds., Hearing by Eye: The Psychology ofLipreading. London, England: Lawrence Erlbaum Associates Ltd., 1987.

[7]Dodd, B. and R. Campbell, eds., Hearing by Eye: The Psychology ofLipreading. London, England: Lawrence Erlbaum Associates Ltd., 1987.

[8]Elham, S. SalamaReda A. El-Khoribi Mahmoud E. Shoman, Audio-Visual Speech Recognition for People with Speech Disorders, *International Journal of Computer Applications* (0975 – 8887) Volume 96– No.2, June 2014

[9]Fatma Zohra Chelali, Audiovisual speech/speaker recognition, application to Arabic languageMultimedia Computing and Systems (ICMCS), 2011

[10]Fatma Zohra Chelali, Khadidja Sadeddine Amar Djeradi, Visual Speech Analysis, Application to Arabic Phonemes Special Issue of *International Journal of Computer Applications* (0975 – 8887) on Software Engineering, Databases and Expert Systems – SEDEXS, September 2012

[11]Frowein, H. W., G. F. Smoorenburg, L. Pyters, and D. Schinkel, \Improved speech recognition through videotelephony: Experiments with the hard of hearing," *IEEE Journal of Selected Areas in Communications*, vol. 9, pp. 611{616, May 1991.

[12]Ghadban, N. S., Alkheir, J., & Saii, M. (2017). Facial Feature Detection for Arabic Audio Visual Speech Recognition Systems. Pakistan Research Journal of Management Sciences, 1(2).

[13]Ghadban, N. S., Alkheir, J., & Saii, M. (2017). Visual Object Speech Recognitionfor Isolated Arabic Words. International Journal of Information Research and Review Vol. 04, Issue, 08, pp.4459-4463, August, 2017

[14]Hennecke, M. 1996. eds.), pp. 103{114, Springer-Verlag, 1996.

[15]http://uk.mathworks.com/matlabcentral/fileexchange/19665-visualize-output-of-bwlabel/content/vislabels.m

[16]Lavagetto, F. Converting speech into lip movements: A multimedia tele- phone for hard hearing people," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, pp. 90{102, March 1995.

[17]Lippmann, R. P. 1997. "Speech recognition by machines and humans," Speech Commun, vol. 22, pp. 1-15.

[18]Luettin, J., G. Potamianos, and C. Neti, \Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in International Con- ference on Acoustics, Speech and Signal Processing, vol. 1, (Salt Lake City), pp. 169 {172, May 2001.

[19]McGurk, H. and J. MacDonald, \Hearing lips and seeing voices," Nature, pp. 746{748, December 1976.

[20]Sadeddine, K FZ Chelali, R Djeradi, A Djeradi, Visual Speaker Verification System Depending on Arabic Syllables, Speech communication and signal processing laboratory HouariBoumedienne University of sciences and Technologies, USTHB 2013

[21]Seddik, A.F. M El Adawy A Computer-Aided Speech Disorders Correction System for Arabic Language , - ieeexplore.ieee.org ,2013

[22]Tren, D. M. W.Lewis, "Lip Region Detection," 2001.

[23]Yao-Jiunn Chen, Y.-C. L. 2007. "Simple Face-detection Algorithm Based on Minimum Facial Features," in IEEE Industrial Electronics Society (IECON), Taipei, Taiwan.