Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Lip Movement Recognition using Histogram of Oriented Gradient (HOG) and Support Machine Vector (SVM) for Arabic Word

Fahmi Muhammad Rabbani[1], Bima Sena Bayu Dewantara[2], Endra Pitowarno[3]
Email [1]fahmimuhrabbani@gmail.com, [2]bima@pens.ac.id, [3]epit@eepis-its.edu
[1,2,3]Electronic Engineering Polytechnic Institute of Surabaya.

Article Information

Submitted : 12 Dec 2023
Reviewed: 22 Dec 2023
Accepted : 30 Dec 2023

Keywords

Visual speech recognition; Lip-reading; Modern Standard Arabic; HOG Feature Extraction; SVM Classification

Abstract

This research aims to develop a lip gesture recognition system in Arabic words by utilizing Histogram of Oriented Gradient (HOG) feature extraction and Support Vector Machine (SVM) classification. The evaluation was conducted on a dataset of 1749 videos with male and female participation using Modern Standard Arabic. The 10 cross-fold validation method was used to measure the performance of the system. By applying a polynomial kernel, this study achieved an accuracy rate of 95.63%, while the word recognition rate reached 96%. These results confirm the system's ability to recognize lip movements with precision, confirming the effectiveness of the approach used in visual recognition for Arabic.

A. Introduction

Lip reading is a communication technique that allows one to understand speech by simply paying attention to the movement of the speaker's lips without having to listen to their voice [1]. This technique is useful for those with hearing loss or to understand people who have lost their voice due to medical conditions [2]. In addition, this technique is used for various purposes such as improving speech recognition where there is a lot of noise [3], security systems [4], forensic investigations [5], and others. Petar S. Aleksic describes the exploration of changes in visual features extracted from the mouth region to obtain visual feature information that can improve the performance of speech recognition systems and be more resistant to forgery attempts [6].

Feature extraction is a process of retrieving visual content from images for indexing and retrieval. This is an important step in multimedia processing where there are generally 4 main types of features that can be retrieved, namely shape features, color features, statistical features and texture features [7]. Each technique has both advantages and disadvantages, depending on the type of image processing problem to be solved and the type of image being processed. For example, shape features may be more useful in recognizing objects or faces, while texture features are more suitable for distinguishing between types of surfaces. So far, several research studies have compared various feature extraction techniques in lip reading. However, the results show that there is no feature extraction technique that is considered superior.

The development of technology and computing has brought improvements to lip reading methods. Research on lip reading has been conducted in various languages, such as English, Chinese, French [8], Arabic [9]–[13] and even Indonesian [14]. One of the main challenges in lip reading or visual speech recognition is the variation in input, such as differences in skin color, face shape, lighting, and background [15]. Therefore, many systems are limited to reading only a certain number of words or phrases. Nonetheless, many methods have been developed to precisely detect lip contours, such as feature extraction techniques.

B. Related Work

There are research studies that have been conducted on lip reading for Arabic vocabulary. Before the era of deep learning, researchers have discovered lip movements with various machine learning techniques and algorithms. Some used hypercolumn model feature extraction and hidden markov model (HMM) classification techniques [28]. In addition, some used low-level statistical methods and counted ROI pixels for geometric feature extraction in 4 points on the lips and then classified hidden markov model (HMM). They tested their method on 20 words collected from 4 speakers. Their algorithm yielded an accuracy of 81.7% [13].

On vocabulary level identification, Lamiaa A. Elrefai et al. collected 1100 videos of 10 Arabic words spoken by 22 speakers. They manually cropped the mouth region from the video frame then used discrete cosine transform (DCT) feature extraction and support vector machine (SVM) classification model. The study obtained a word recognition rate (WRR) of 70% [9]. In addition, there are those who compare three extraction models with the softmax layer [16].

At sentence-level identification, there are researchers who use deep learning algorithms for Arabic using a dataset of 2400 recorded Arabic digits and 960 recorded Arabic phrases from 24 speakers. The research includes important processes such as face detection, lip localization, feature extraction, classifier training, and word recognition. The accuracy of digit recognition was 94%, phrase recognition was 97%, and phrase and digit recognition was 93% [10].

From the explanation above and the lack of literacy related to lip movement recognition in Arabic pronunciation, we propose to conduct research on visual movement recognition using HOG extraction techniques and SVM as its classification on everyday Arabic vocabulary datasets towards an automatic lip reading system and hope to get maximum accuracy results.

C. Research Method

Lip movement recognition design system using HOG feature and SVM classifier for Arabic words is one of the techniques in recognizing Arabic words. The system uses HOG (Histogram of Oriented Gradients) features to extract patterns and features from lip movements, which are then used as input for SVM (Support Vector Machine) classification model to recognize the spoken word. In the test, we used three SVM kernel comparisons (linear, polynomial, and RBF) to determine the best performance. The system design illustrates the overall process contained in the system which can be shown in Figure 1.

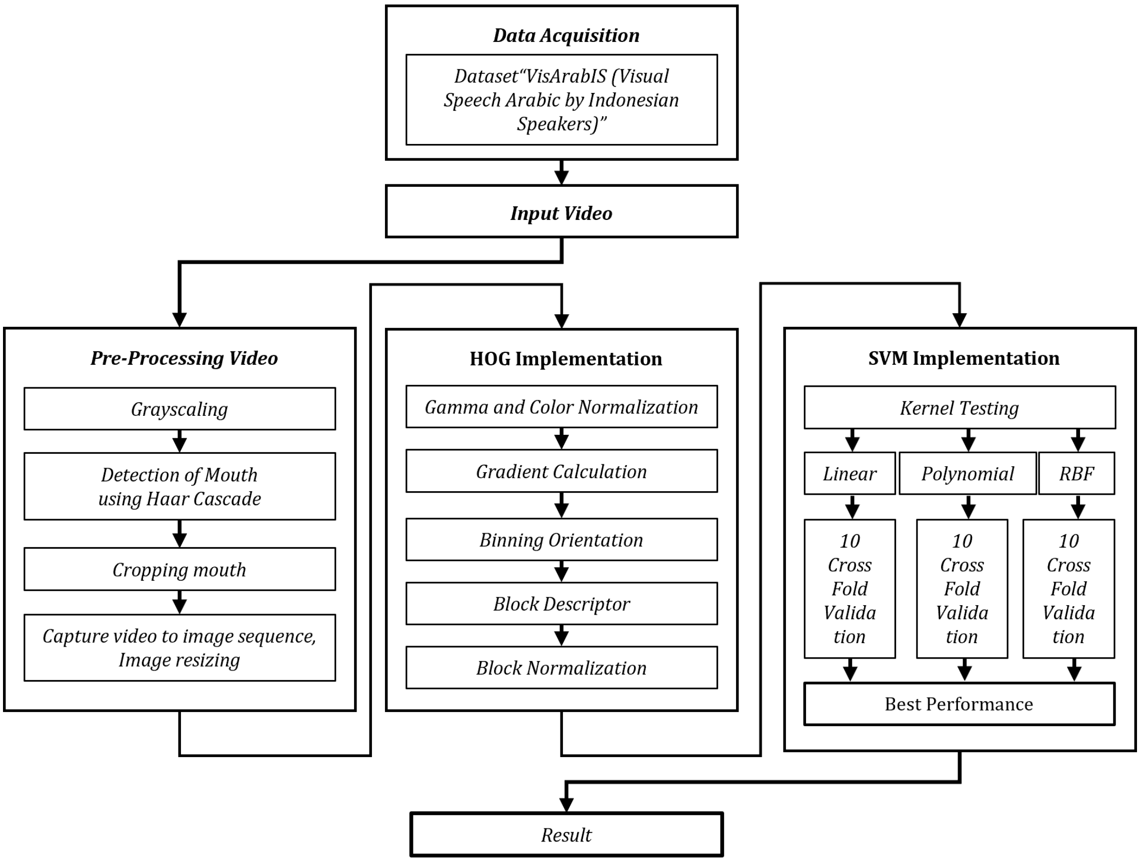

Figure 1. System Design

1. Data Acquisition

In our study we used ten Arabic vocabulary words that were collected independently. Each of the Arabic vocabulary is spoken by fourteen (seven male and seven female) speakers and averaged over ten repetitions, so the total of video datasets is 1749 datasets.

Table 1. Research Dataset Specifications

Specification Category	Specification Type	Specification
Language Specifications	Language	Modern Standard Arabic (MSA)
	Speech Type	Isolated words of communication
	Vocabulary Count	10 words
	Number of Speech Datasets	1749
Participant Specifications	Number of Participants	14
	Gender	7 men & 7 women
	Participant Face Display	Frontal
	Multi-speakers	No
Engineering Specifications	Camera Type	Logitech C270 HD Webcam
	Data obtained	Video (MP4 format)
	Resolution, Frame Rate	1280 x 720 pixel, 30 fps
	Retrieval Distance	±50 cm
	Environment	Lighting using LED lights indoors and simple backgrounds

We named this dataset "VisArabIS" which stands for "Visual Speech Arabic by Indonesian Speakers". Here is the list of vocabularies that we use:

Table 2. Dataset VisArabIS (Visual Speech Arabic by Indonesian Speakers)

No.	Word in Arabic	Word Syllables	Word in English	Number of datasets		Total
				Male	Female	
1	أهلا	ah/lan	Welcome	96	80	176
2	مرحبًا	mar/ha/ban	Hello	96	79	175
3	شكرًا	shuk/ran	Thank you	91	79	170
4	تفضل	taf/fad/dal	Here you go	90	80	170
5	عفوا	af/wan	Sory	91	79	170
6	طيّب	tai/yeb	Ok	95	84	179
7	جيد	Jay/yid	Okay	98	78	176
8	سلام	sa/lam	Congratulations	96	79	175
9	وداعا	wa/da/aan	Goodbye	99	81	180
10	أَنْظُرْ	un/dzur	Look!	100	78	178
		Total		**952**	**797**	**1749**

2. Data Pre-Processing

Pre-Processing is a process carried out to condition the dataset, in this process there are three stages that will be passed, namely, the detection of the mouth using Haar Cascade, grayscaling, cropping the mouth, converting the video into a series of frames, and image resizing.

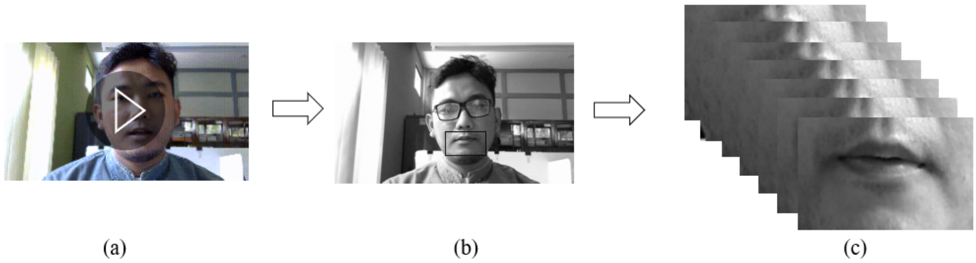

Figure 2. Illustration of the mouth detection to cropping process. (a) Video input, (b) Mouth area detection stage, grayscaling, (c) Frame cropping stage

Mouth detection from video input involves many important steps. The process of identifying the mouth from the video is done in several stages. First, the mouth area is identified in each frame using the Haar cascades method. Once the mouth area is detected, the image is converted into grayscale to make analysis easier. The final stage involves cropping the frame according to the detected mouth area, allowing focus on the relevant part for subsequent analysis. Thus, the process of mouth detection from video becomes more effective.

3. HOG implementation

After the pre-processing stage, the feature extraction process is carried out using the Histogram of Oriented Gradients (HOG) method where the stages are shown in Figure 3.

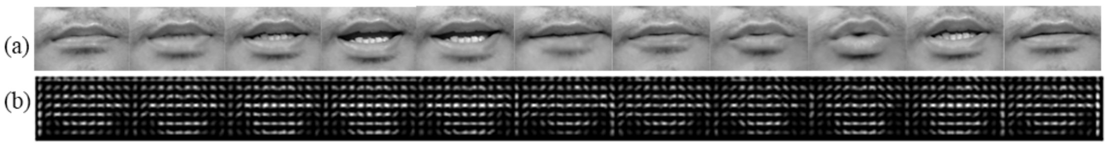

Figure 3. Sample frame of the pronunciation sequence of the word "*afwan*". (a) grayscale frame (b) HOG extracted frame

The process of extracting Histogram of Oriented Gradients (HOG) features from the image dataset uses the OpenCV library. The process continues with the initialization of the dataset path and CSV file, and the processing of the dataset which involves reading, resizing, and calculating the HOG features for each image sequence. The final step saves the feature extraction results into a CSV file with three columns that record the label, frame sequence, and HOG features. This process aims to prepare relevant data for training machine learning models in lip reading identification tasks.

4. SVM implementation

There are two main stages involved in the implementation of classification using the dataset. First, the evaluation should involve testing the RBF, linear, polynomial, and kernel used in the Support Vector Machine (SVM) technique developed by Vladimir Vapnik [17]. Secondly, ten-fold cross validation is used to validate performance. This checks the robustness and accuracy of the model across different subsets of data, so as to obtain the best accuracy.

Table 3. Definition of SVM Kernel

Kernel	Function Definition	
Linear	$K(x, y) = x.y$	(1)
Polynomial	$K(x, y) = (\alpha x.y + c)^d$	(2)
Radial (Gaussian)	$K(x, y) = \exp(-\gamma\|x - y\|^2)$	(3)

D. Result and Discussion

The session evaluates the model in two key areas. First, it evaluates metrics for each class, including accuracy, precision, recall, f1-score, support, and word recognition rate. Then, it compares confusion matrix results from experiments using three kernels: linear, polynomial, and RBF.

1. Model Performance Evaluation

The results of evaluating the model performance using three types of kernels, namely linear, polynomial, and RBF, show significant differences in precision, recall, F1-score, and word recognition rate (WRR) as shown in tables 4 to 6. The linear kernel shows an average result of about 64.87%, while the polynomial kernel is much higher with an average score reaching 95.63%. However, despite being lower, the RBF kernel still showed satisfactory performance with an average score of about 91.25%. This analysis highlights that each type of kernel has its own advantages and disadvantages, and proper selection needs to consider the unique characteristics of the data used.

Table 4. The precision, recall, loss, and testing accuracy for linear kernel

Class	أهلا	مرحبًا	شكرًا	تفضل	عفوا	طيّب	جيد	سلام	وداعا	أنْظُرْ	Average
Precision	68	62	68	58	58	64	66	68	67	73	64,87
Recall	74	62	71	55	62	58	65	67	64	73	64,87
F1-score	71	62	69	56	60	61	66	67	66	73	64,82
WRR	74	62	71	55	62	58	65	67	64	73	65
Overall accuracy 64,88 %											
Loss 35,12 %											

Table 5. The precision, recall, loss, and testing accuracy for polynomial kernel

Class	أهلا	مرحبًا	شكرًا	تفضل	عفوا	طيّب	جيد	سلام	وداعا	أنْظُرْ	Average
Precision	97	95	96	94	95	95	95	96	97	97	95,63
Recall	98	96	96	93	96	94	95	96	96	97	95,62
F1-score	97	95	96	94	96	94	95	96	97	97	95,62
WRR	98	96	96	93	96	94	95	96	96	97	96
Overall accuracy 95,63 %											
Loss 4,37 %											

Table 6. The precision, recall, loss, and testing accuracy for RBF kernel

Class	أهلا	مرحبًا	شكرًا	تفضل	عفوا	طيّب	جيد	سلام	وداعا	أنْظُرْ	Average
Precision	94	90	92	88	91	90	91	91	93	93	91,25
Recall	94	92	91	88	92	89	89	92	92	94	91,25
F1-score	94	91	92	88	91	90	90	92	93	93	91,25
WRR	94	92	91	88	92	89	89	92	92	94	91,25
Overall accuracy 91, 25 %											
Loss 8,75											

2. Kernel Comparison on Confusion Matrix

Figures 4 to 6 show the different patterns in the classification results shown by the three visualizations of the confusion matrix with different types of kernels. With high accuracy and uniform class distribution, the polynomial kernel shows better classification results overall. With a balanced class distribution in the confusion matrix, the RBF kernel is also highly accurate. The linear kernel, on the other hand, showed uneven distribution and inconsistent performance compared to the other kernel types.

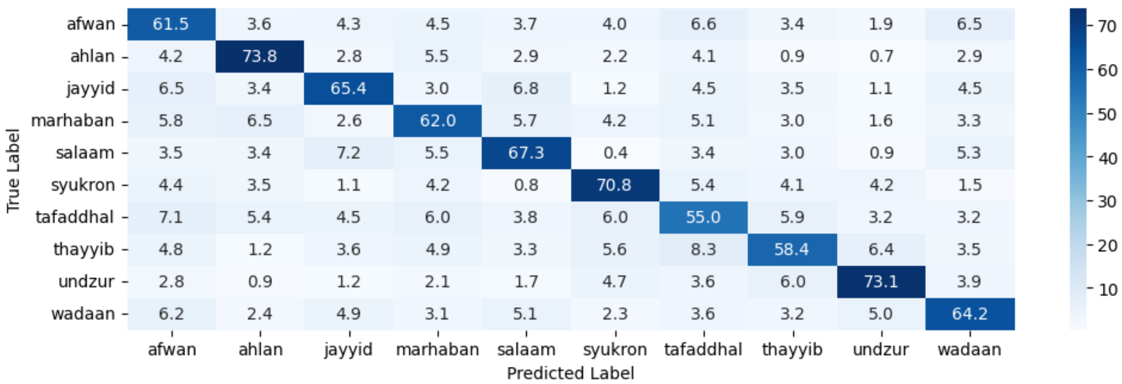

Figure 4. Confusion matrix of the linear kernel experiment

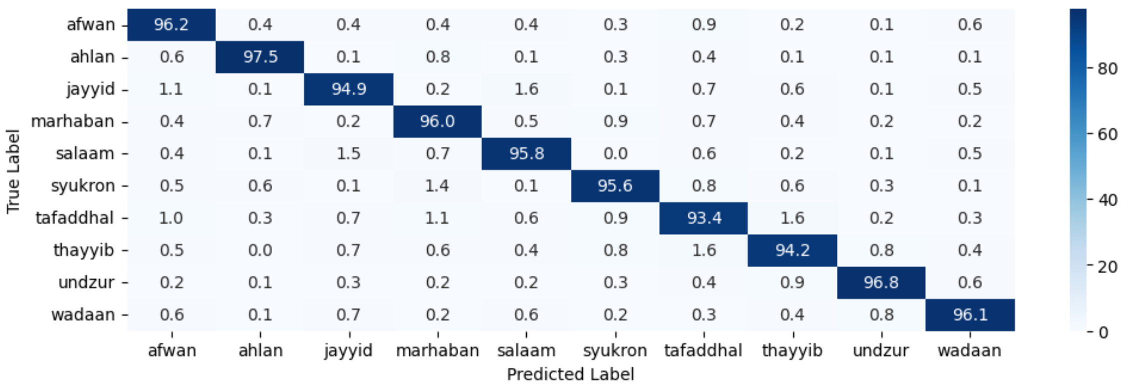

Figure 5. Confusion matrix of the polynomial kernel experiment

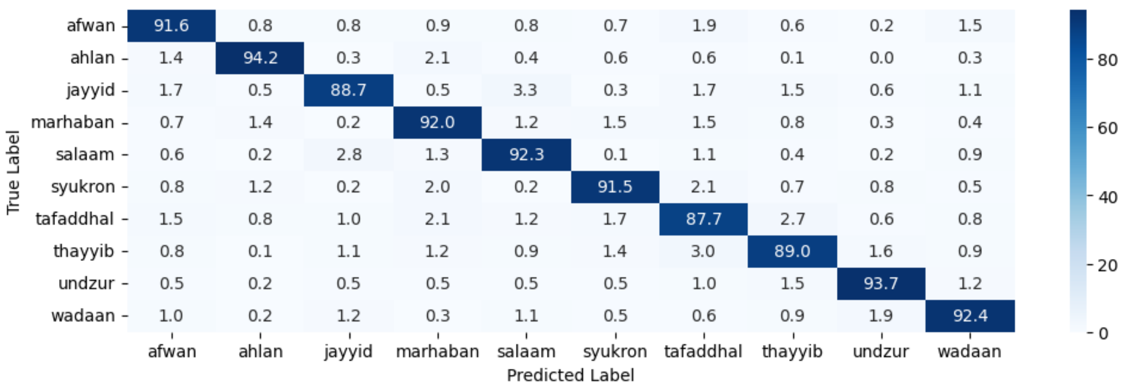

Figure 6. Confusion matrix of the RBF kernel experiment

3. Comparison of Results with other Research

The following are the results of the comparison of lip reading research, especially on Arabic words, as shown in Table 6.

Tabel 6. Comparison with the existing work

Author	Features extraction	Classifier	Level recognition	Accuracy Result
A. Sagheer, T. Naoyuki & R. Taniguchi. [18]	Hypercolumn model (HCM)	HMM, with five states	9 Arabic sentences	Accuracy of 62.9%
D. Pascal [13]	Using statistical methods, extracted pixels of the ROI for geometrical features at 4 lip points (W, H, A, D).	HMM with three states	20 Arabic words	Accuracy of 81.7%
L. Elrefaei, T. Alhassan and S. Omar, [9]	DCT	SVM	10 Arabic words	Word recognition rate of 70%
W. Dweik, S. Altorman and S. Ashour, [16]	Three models: CNN & TD + LSTM & TD + BiLSTM	Softmax layer	10 Arabic words	RGB in CNN (79.2%), grayscale in CNN (76.6%), RGB in TD + LSTM (70.1%), grayscale in TD + LSTM (67.5%), RGB in TD + BiLSTM (74.1%), grayscale in TD + BiLSTM (70.1%), RGB in a voting model (82.8%) accuracies.
N. Alsulami, A. Jamal and L. Elrefaei, [10]	VGG-19 with batch normalization	Softmax layer	10 Arabic digits and 4 Arabic sentences	Digits: 94%, Sentences: 97%, Digits & Sentences: 93% Accuracy.
Fahmi, (Our research)	**HOG**	**SVM**	**10 Arabic words**	**Accuracy of 95,63%, Word Recognition rate of 96%**

E. Conclusion

This study yielded a significant accuracy rate of 95.63% with a word recognition rate of 96% using a polynomial kernel. These findings have many applications in speech recognition and lip movement interpretation, and will help in the development of visual recognition technology for Arabic. This research is an important contribution to the advancement in image processing and speech recognition as the method using Histogram Oriented Gradient (HOG) along with Support Vector Machine (SVM) classification proved effective in recognizing and interpreting lip movements.

F. References

[1] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, "A Survey of Research on Lip-Reading Technology," *IEEE Access*, vol. 8. Institute of Electrical and Electronics Engineers Inc., pp. 204518–204544, 2020. doi: 10.1109/ACCESS.2020.3036865.

[2] M. Bourguignon, M. Baart, E. C. Kapnoula, and N. Molinaro, "Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech," *J. Neurosci.*, vol. 40, no. 5, pp. 1053–1065, Jan. 2020, doi: 10.1523/JNEUROSCI.1101-19.2019.

[3] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1325, 2003, doi: 10.1109/JPROC.2003.817150.

[4] M. B. Pranoto, K. N. Ramadhani, and A. Arifianto, "Face Detection System Menggunakan Metode Histogram of Oriented Gradients (HOG) dan Support Vector Machine (SVM) Face Dtection System using Histogram of Oriented Gradients (HOG) Method amd Support Vector Machine(SVM)."

[5] S. Deepika and P. Philip, "Lip Movement Detection on Online Exam Along with Machine Learning," JETIR, 2021. [Online]. Available: www.jetir.org

[6] M. Ezz, A. M. Mostafa, and A. A. Nasr, "A Silent Password Recognition Framework Based on Lip Analysis," *IEEE Access*, vol. 8, pp. 55354–55371, 2020, doi: 10.1109/ACCESS.2020.2982359.

[7] W. K. Mutlag, S. K. Ali, Z. M. Aydam, and B. H. Taher, "Feature Extraction Methods: A Review," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Aug. 2020. doi: 10.1088/1742-6596/1591/1/012028.

[8] Y. Lu, J. Yan, and K. Gu, "Review on Automatic Lip Reading Techniques," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 7, Jul. 2018, doi: 10.1142/S0218001418560074.

[9] L. A. Elrefaei, T. Q. Alhassan, and S. S. Omar, "An Arabic Visual Dataset for Visual Speech Recognition," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 400–409. doi: 10.1016/j.procs.2019.12.122.

[10] N. H. Alsulami, A. T. Jamal, and L. A. Elrefaei, "Deep learning-based approach for arabic visual speech recognition," *Comput. Mater. Contin.*, vol. 71, no. 1, pp. 85–108, 2022, doi: 10.32604/cmc.2022.019450.

[11] A. H. Reda, A. A. Nasr, M. M. Ezz, and H. M. Harb, "AN ARABIC FIGURES RECOGNITION MODEL BASED ON AUTOMATIC LEARNING OF LIP MOVEMENT," 2017.

[12] A. Al-Ghanim, N. Al-Oboud, R. Al-Haidary, S. Al-Zeer, S. Altammami, and H. A. Mahmoud, "I See What You Say (ISWYS): Arabic lip reading system," in *Proceedings of the 2013 International Conference on Current Trends in Information Technology, CTIT 2013*, IEEE Computer Society, 2013, pp. 11–17. doi: 10.1109/CTIT.2013.6749470.

[13] P. Damien, "Visual speech recognition of Modern Classic Arabic language," *SHUSER 2011 - 2011 Int. Symp. Humanit. Sci. Eng. Res.*, pp. 50–55, 2011, doi: 10.1109/SHUSER.2011.6008499.

[14] F. N. Effendy, "Pengenalan Pola Gerak Bibir Dalam Pengucapan Fonem Vokal

Bahasa Indonesia."
[15] G. Potamianos and C. Neti, "Audio-Visual Speech Recognition in Challenging Environments."
[16] W. Dweik, S. Altorman, and S. Ashour, "Read my lips: Artificial intelligence word-level arabic lipreading system," *Egypt. Informatics J.*, vol. 23, no. 4, pp. 1–12, 2022, doi: 10.1016/j.eij.2022.06.001.
[17] S. Dabbaghchian, M. P. Ghaemmaghami, and A. Aghagolzadeh, "Feature extraction using discrete cosine transform and discrimination power analysis with a face recognition technology," *Pattern Recognit.*, vol. 43, no. 4, pp. 1431–1440, 2010, doi: 10.1016/j.patcog.2009.11.001.
[18] A. El Sagheer, N. Tsuruta, and R. I. Taniguchi, "Arabic lip-reading system: A combination of hypercolumn neural network model with hidden Markov model," *Proc. Eighth IASTED Int. Conf. Artif. Intell. Soft Comput.*, no. January, pp. 311–316, 2004.