

Original Article

PERFORMANCE ENHANCEMENT FOR MOUTH DETECTION TIME BASED ON MODIFIED VIOLA-JONES ALGORITHM

Mohammed E. Safi ^{1,*}, Eyad I. Abbas ¹, and Alaa H. Ali ²

¹College of Electrical Engineering, University of Technology- Iraq, Baghdad, Iraq.

²Department of Medical Instrumentation Techniques Engineering, Al-Mustaqbal University, 51001, Babylon, Iraq.

*Corresponding author, E-mail: mohammed.e.safi@uotechnology.edu.iq (Tel: +964-7901848481)

ABSTRACT

Received
August 6, 2026

Accepted
October 9, 2026

Published
October 20, 2026

Mouth detection is an important issue in visual speech recognition. The de facto algorithm used to achieve this is the Viola-Jones algorithm. Nevertheless, time enhancement for the auto-detection of this algorithm is required. Therefore, the goal of this research is to propose a modified approach based on the Viola-Jones algorithm to enhance the detection time. The main contribution of the proposed algorithm is limiting the auto-valuing for the threshold of the Viola-Jones algorithm and enhancing the time by configuring the maximum step for the adaptive threshold value. The UOTletters database was used in the experiments of this research with 30 speakers pronouncing English letters with overall frames of 28164. The experiment results have shown that the average detection time for the proposed method was 0.585 sec, compared with the baseline method's 2.097 sec. As a result, the proposed approach enhanced the mouth detection time with respect to the baseline algorithm (358.46%). Consequently, the proposed algorithm is a good solution for real-time visual speech recognition regarding the detection time enhancement.

Keywords: Mouth detection; Viola-Jones algorithm; adaptive threshold value.

1. INTRODUCTION

Machine learning is witnessing a rise in many applications in communications, enabling smarter resource allocation, enhanced signal processing, and more adaptive algorithms (Radhi *et al.*, 2022). Speech recognition nowadays has become one of the most common features in many applications and devices (Mohammed E. Safi & Abbas, 2016). Nevertheless, the challenge of enhancing accuracy in audio speech recognition is required (Mohammed Ehsan Safi & Abbas, 2022). As a result, visual speech recognition was adopted to assist audio speech recognition. Visual speech recognition, basically in common algorithms, consists of three parts: Mouth detection, lips isolation, and finally classification (Mohammed Ehsan Safi & Abbas, 2023). Mouth detection is a machine-learning process that is used to detect the human mouth and separate it from its surroundings (Kaur & Sharma, 2017). Many approaches are used in this field, as shown in research (Balasubramanian *et al.*, 2009; Nguyen *et al.*, 2006; Ryu & Oh, 2001), but the Viola-Jones algorithm became a de facto algorithm (Wisam H Ali *et al.*, 2020; Ibrahim & Mulvaney, 2015; Viola & Jones, 2001). The Viola-Jones algorithm (VJA) is a machine-learning algorithm for detecting and analyzing Human face parts from images with high speed and accuracy (Maghraby *et al.*, 2014). VJA in color images is converted to grayscale. Therefore, in Haar-like features, it's important to specify the threshold for it to configure edges. The thresholds in some pieces of research are static thresholds, such as (Wisam H Ali *et al.*, 2020; Setu & Rahman, 2016). In research (Joshi *et al.*, 2021), the optimal threshold for mouth detection was studied. The experiment results of this study have shown that the maximum

accuracy at the threshold value of 150, by accuracy of 73.07% and 92.30%, respectively. The database was selected at 5ft and 8ft distance respectively. In practical and real-time videos, not all frames have the same threshold due to differences in environments like illumination and skin color. As a result, auto-configuring to the optimal threshold was proposed by (Indrabayu *et al.*, 2019). They suggested a modification for VJA by using a zero value as the initial for the threshold parameter. The auto-valuing threshold 938 face frames for 20 videos yields 5.665 sec. in mouth detection feature time with an accuracy of 99.49%. In this paper, a modification Viola-Jones algorithm using an adaptive threshold for optimized algorithms (Indrabayu *et al.*, 2019) and (Joshi *et al.*, 2021), respectively, was proposed. The drawbacks of the previous works based on a fixed threshold value are problems related to moving objects and varying light and distance. The drawback of the baseline auto-valuing threshold is that it is time-consuming because the criteria determination is based on practical observation only and is not supported by analytical calculations.

The main contribution of the proposed algorithm is limiting the auto-valuing of the threshold of the Viola-Jones algorithm and reducing the time by configuring the maximum step for the adaptive threshold value. The selection of the search step was improved by analyzing the group of threshold values using polynomial curve fitting of the linear equation. Moreover, the direction of the adaptive threshold value step (increment or decrement) was determined based on experimental results and analytical determination. The experiment results were performed on the UOTletters database (Wisam H Ali *et al.*, 2020), which consists of 30 speakers pronouncing English letters with overall frames of 28164.

Access this article online



DOI: <https://doi.org/10.25271/sjuoz.2026.14.4.1993>

Printed ISSN 2663-628X;
Electronic ISSN 2663-6298

Science Journal of University of Zakho
Vol. 14, No. 04, pp. 687-692, October-2026

This is an open access article under a CC BY-NC-SA 4.0 license
(<https://creativecommons.org/licenses/by-nc-sa/4.0/>)

Table 1: Related works' drawbacks.

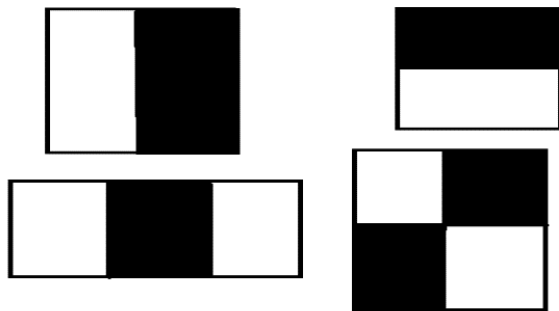
Reference	Algorithm	Threshold	Drawback
(Wisam H Ali <i>et al.</i> , 2020; Ibrahim & Mulvaney, 2015; Viola & Jones, 2001)	Standard Viola & Jones algorithm	Static, based on distance and light contrast	Misses detection with varying light and distance
(Joshi <i>et al.</i> , 2021)	Standard Viola & Jones algorithm with a fixed threshold based on distance	The threshold value is 150 at 5 ft and 8 ft distances, respectively	Miss detection with varying light and distance more than 8 ft distances and less 5 ft.
(Indrabayu <i>et al.</i> , 2019)	Modified Viola & Jones algorithm	Auto-valuing for Threshold starting from 1 to unlimited value	Cost of time and cause of error if no object detection.
(Joshi <i>et al.</i> , 2021)	Modified Viola & Jones algorithm	Auto-valuing for Threshold starting from 1 to unlimited value maximum values of 150 at 3 ft to 8 ft of distance	Cost of time.
The proposed algorithm	The proposed Modified Viola & Jones algorithm	Auto Valuing for Threshold starting from 150 to 1 at 3 ft to 8 ft of distance	Starting value for distance more than 8 ft. (Future works)

1. Viola-Jones Algorithm

VJA has many features, like dealing with different skin colors, the ability to detect even with eyeglasses, and the ability to recognize many faces in one image (Zhang & Zhang, 2010). The algorithm is based on rectangular-shaped features called Haar-Like features (Papageorgiou *et al.*, 1998). This feature distinguishes the boundaries of the faces from their surroundings in the frame. This technique is effective even for different face images with different ranges. The algorithm with features is faster than the algorithm that works on a pixel operation. As a result, employing characteristics of pictures is used more than pixel-based ones (Wissam Hassan Ali, 2016). Figure 1 illustrates most shapes of rectangle features, which consist of a white pixel area and a black pixel area (Indrabayu *et al.*, 2019). The concept of using Haar-like features is to follow the boundaries of objects by subtracting the white area from the black area:

$$V = \sum P_{(whiteregion)} - \sum P_{(blackregion)} \quad (1)$$

Where: P is the value of the pixel, and V is the feature value.

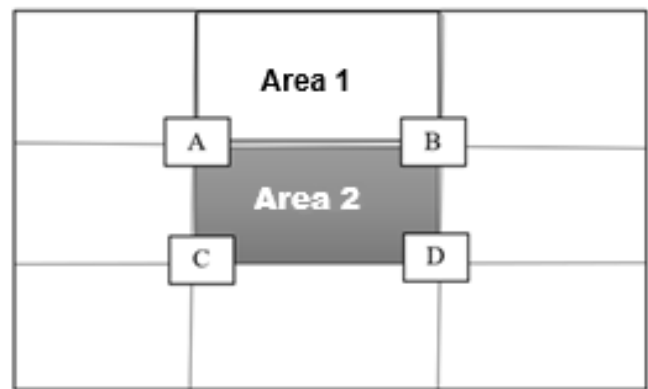
**Figure 1:** Different shapes of Haar-like features.

To minimize the complexity of the mathematical operations, the Viola-Jones algorithm converts the input image into grayscale and into a new mathematical form called the integral image to minimize the calculation time. The integral image is determined by the sum of the upper rectangle for the current pixel (Kherchaoui & Houacine, 2010):

$$In_{(i,j)} = \sum_{\tilde{i}=1, \tilde{j}=1}^{i-1, j-1} Image(\tilde{i}, \tilde{j}) \quad (SEEquation \setminus * ARABIC2)$$

Where: In is the integral image, Image is the input image, and (i,j) is the spatial pixel at the x and y coordinates, respectively.

The integral image is determined by taking the spatial value of the bottom-left pixel of the feature window and subtracting the remaining unwanted region of the integral image to obtain the contrast of the Haar-like features, which leads to decreased detection time. For example, as shown in Figure 2, two areas of Haar's feature window represent the cases of calculation for areas in the integral image (Berbar *et al.*, 2006).

**Figure 2:** Integral image example.

The two cases are the region with an edge surrounding it and the region with an unwanted area boundary. The edge surrounding the Area illustrated in Area 1 of Figure 2 can be calculated as given below:

$$v_1 = B - A \quad (SEEquation \setminus * ARABIC3)$$

The Area with unwanted Areas surrounding it, illustrated in Area 2 of Figure 2, can be calculated as given below:

$$v_2 = D - (B + C) + A \quad (SEEquation \setminus * ARABIC4)$$

It is obvious that point A is added because it is subtracted twice from point D (point A is compact with points B and C in the undesired areas). In the same example in Figure 2, it could estimate the Haar-like feature value by subtracting v_1 from v_2 using the same principle of Equation 1.

VJA has another concept to reduce the massive number of features; eg., a 24*24 window gives 160,000 features. Viola and Jones proposed two techniques to overcome this problem: using a cascade filter and AdaBoost training, as shown in Figure 3. Some of the obtained features are not useful for face detection. Therefore, a cascade filter was utilized to crumble the detection operation of faces into sub-stages like eyes, nose bridge, and mouth region. In addition, to build a strong classifier, the subsequent weak classifier corrected the error in the previous classifier by AdaBoost training (Setu & Rahman, 2016).

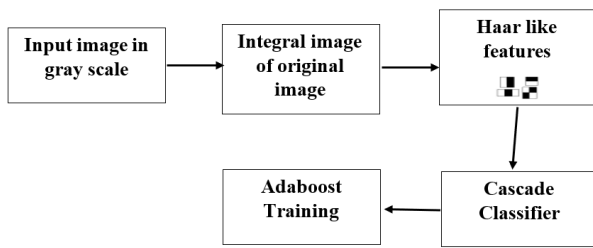


Figure 3: Viola and Jones algorithm.

The final results of detection for the images that have multi-candidate detection are surrounded by merging the bounding boxes around the targets. The configuration step of the threshold governs the grouping of the collocated detection boxes. The selection of the optimal merged bounding box threshold would correctly overcome error type 1 (false positive) and time delay or error type 2 (miss detection). The effect of the minimum value of the threshold which led to a false positive illustrated in Figure 4. Nevertheless, error type 2 could occur due to an increase in the threshold (Joshi *et al.*, 2021). Consequently, that parameter is a tunable value in the function (vision.CascadeObjectDetector) in MATLAB (MathWorks, n.d.).



Figure 4: (a). Detection when Threshold=1 (b). Detection when Threshold = 10 (c). Detection when Threshold =80 (d). Detection when Threshold =90.

1. Proposed Mouth Detection Algorithm

The proposed algorithm is inspired by the adaptive value of the algorithm proposed in (Indrabayu *et al.*, 2019). Yet, the selection of the auto-valuing of the optimal threshold is the opposite way. Moreover, the proposed algorithm adopts a technique to maximize the search step.

Figure 5 shows the proposed algorithm procedure; the starting point for the threshold is 150 instead of one initial value in the study (Indrabayu *et al.*, 2019), due to our observation in Section 4 that most values of the optimal threshold are in the upper limit. Moreover, the selection of the upper limit of the optimal threshold (150) is based on observations in the study (Joshi *et al.*, 2021) and our practical experiments for the maximum value that the threshold values reached. In addition, another contribution of the proposed algorithm is to select the search step, which is set to 5. Consequently, the adaptive threshold value (TH) is decreased by (5) until reaching the ROI or reaching the last value (1):

$$TH_t = TH_{t-1} - 5(\text{SEEquation} \setminus * \text{ARABIC5})$$

Equation 5 is determined based on analytic and experimental results in Section 4.

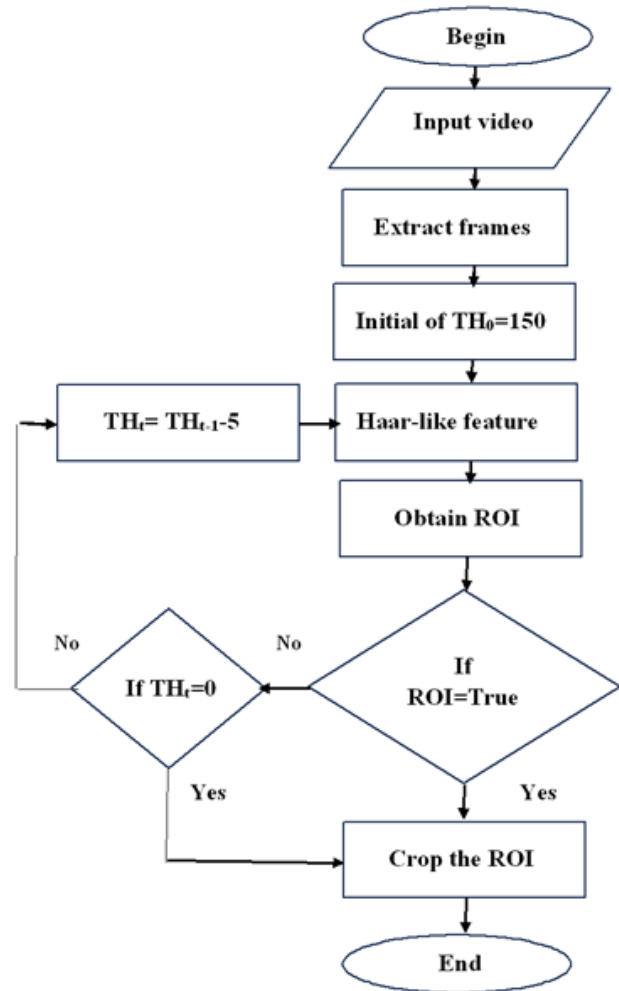


Figure 5: The proposed adaptive algorithm of the optimal threshold for mouth detection.

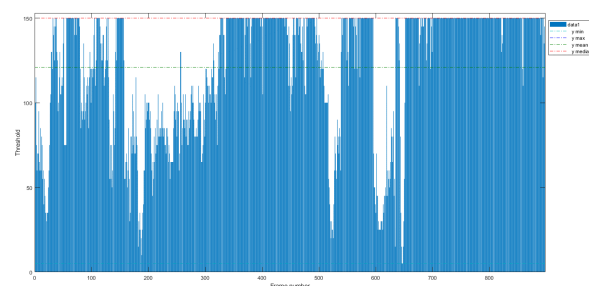
1. Experimental Results

Figure 6 shows sample images of faces and the detected mouth using the proposed algorithm for the UOTletters database. In this section, many experiments were done to ensure that the analytical results satisfy the practical results. Moreover, make an overall evaluation of the proposed algorithm with respect to the baseline algorithm for the thirty speakers pronouncing English letters, with overall frames 28164. Equation 5 is determined from the curve fitting for the threshold based on the detection results, in addition to the distribution of the detection results. As we can see in Figure 7, the sample values of the threshold for the letter A for thirty speakers are most of the values located in the upper area for the proposed range (1 to 150). Consequently, we proposed starting at 150 and then decreasing the threshold value to adopt the optimal value instead of the increment approach in the baseline algorithm. Moreover, the selection of the search step was improved by analyzing the group of threshold values using polynomial curve fitting of the linear equation:

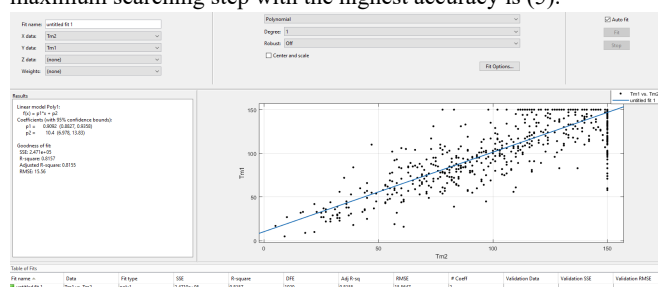
$$TH_t = aTH_{t-1} + b(\text{SEEquation} \setminus * \text{ARABIC6})$$

Where THt is the predicted value of the threshold, and THt-1 is the previous value of it.

Practically, the two coefficients, a and b, are limited to integer values due to the threshold criteria, in addition to the detection ratio.



For the second criterion (detection ratio), a practical experiment for accuracy versus values of the searching step was applied to examine the b coefficient values (1 to 10). As shown in Figure 9, the maximum searching step with the highest accuracy is (5).



A line graph showing the detection ratio (Y-axis, 90 to 100) versus the delta threshold (X-axis, 2 to 10). The ratio is constant at 100% for thresholds 2 through 5. At threshold 5, a vertical dashed red line marks the 'delta threshold'. For thresholds 6 through 10, the ratio decreases: approximately 97.5% at 6, 96.2% at 7, 95.8% at 8, 94.5% at 9, and 92.8% at 10.

delta threshold	detection ratio
2	100
3	100
4	100
5	100
6	97.5
7	96.2
8	95.8
9	94.5
10	92.8

Figure 9: Accuracy versus (b coefficient).

The final experiment evaluation based on accuracy and time delay with respect to the baseline algorithm (Indrabayu *et al.*, 2019) was illustrated in Table 1. The experiment results show that the baseline algorithm has an accuracy of 99.567% with an average detection time (TD) of 2.097 s. Nevertheless, the proposed algorithm has the same accuracy of 99.567% with an average detection time (TD) of 0.585 sec.). Consequently, the proposed algorithm has sped up the calculation process by 3.5846 times that of the baseline algorithm.

DOI:<https://doi.org/10.25271/sjuoz.2026.14.4.1993>

Table 2: Evaluation of the proposed and baseline methods.

No. of speakers	No. of frames	Detected	Accuracy (%)	Average thresholds	average T _D (Sec.) (Indrabayu et al., 2019)	The proposed method's average T _D (Sec.)
1	1019	1019	100	144	4.126	0.1748
2	985	985	100	137	0.068	0.9864
3	1057	1057	100	40	3.97	0.206
4	1161	1160	99.9	130	3.94	0.212
5	962	953	99.1	45	0.15	1.03
6	786	786	100	142	4.33	0.134
7	992	992	100	135	2.77	0.446
8	785	785	100	66	0.665	0.867
9	803	803	100	120	2.87	0.426
10	896	896	100	129	3.34	0.332
11	898	898	100	140	3.86	0.228
12	819	819	100	144	3.835	0.233
13	841	841	100	126	1.91	0.618
14	794	794	100	91	1.55	0.69
15	788	788	100	77	1.025	0.795
16	771	771	100	90	1.03	0.794
17	889	889	100	91	0.18	1.001
18	945	945	100	135	3.515	0.297
19	992	992	100	113	2.5515	0.489
20	957	957	100	105	1.575	0.685
21	867	867	100	129	2.13	0.574
22	908	908	100	75	0.29	0.942
23	932	932	100	134	0.56	0.888
24	965	965	100	107	2.155	0.569
25	973	973	100	138	2.205	0.559
26	1018	1018	100	145	3.025	0.395
27	1010	1010	100	121	0.58	0.539
28	1223	1150	94	85	0.58	0.884
29	1045	1045	100	145	0.12	1.024
30	870	870	100	144	4.015	0.197
Overall	28164	28008	99.567	114	2.097	0.585

CONCLUSIONS

The mouth detection time is an important issue in visual speech recognition. In this paper, a modification of VJA using an adaptive threshold was proposed. The experimental results have shown that our algorithm enhanced the mouth detection time with the baseline algorithm reduced to more than 3.5 times approximately. Consequently, the proposed algorithm gives a good solution for real-time visual speech recognition regarding the detection time enhancement. The drawback of this paper is that the experimental results are limited to a maximum distance of 8 ft from the camera; therefore, the maximum limit (150) may change at greater distances depending on the application and real-time implementation. Moreover, the limitation of the present study is an important direction for future validation using more diverse datasets.

ACKNOWLEDGEMENTS:

“The authors sincerely thank the original investigators, Wisam H Ali, R. Saeed, and H. Al-Muifraje, for their foundational work and for making the dataset.”

Ethical Statement:

“The original data collection was conducted by the primary investigators (Wisam H Ali et al., 2020) with the official approval of their institutional university authorities at that time. Informed consent was explicitly obtained from all human volunteers by the original research team. For the present study, a secondary analysis was performed using this existing dataset, which was provided in a fully anonymized and de-identified format to protect participant confidentiality.”

Author Contributions

Study concept and design: M. E. Safi, E. I. Abbas, A. H. Ali; Analysis and interpretation of data: M. E. Safi, E. I. Abbas; Drafting of the manuscript: M. E. Safi, E. I. Abbas, A. H. Ali; Critical revision of the manuscript for important intellectual content: M. E. Safi, E. I. Abbas; Statistical analysis: M. E. Safi, E. I. Abbas, A. H. Ali; Administrative, technical, and material support: M. E. Safi, E. I. Abbas, A. H. Ali; Study supervision: M. E. Safi, E. I. Abbas, A. H. Ali.

Conflict of Interest:

“The authors declare that they have no conflicts of interest related to the publication or funding of this manuscript.”

Funding:

“The authors declare that no financial support, grants, or funding were received from any public, commercial, or non-profit organizations for the research, authorship, or publication of this article.”

2. REFERENCES

- Ali, W., H. R. Saeed, T., & H. Al-Muifraje, M. (2020). FPGA Implementation of Visual Speech Recognition System based on NVGRAM-WNN. *2020 International Conference on Computer Science and Software Engineering (CSASE)*, 132–137. DOI:<https://doi.org/10.1109/CSASE48920.2020.9142095>
- Ali, W. H. (2016). Pattern Recognition System for Human Utterance (Issue June). University of Technology-Iraq.
- Balasubramanian, M., Palanivel, S., & Ramalingam, V. (2009). Real time face and mouth recognition using radial basis function neural

- networks. *Expert Systems with Applications*, 36(3, Part 2), 6879–6888. DOI:<https://doi.org/10.1016/j.eswa.2008.08.001>
- Berbar, M. A., Kelash, H. M., & Kandeel, A. A. (2006). Faces and Facial Features Detection in Color Images. *Geometric Modeling and Imaging–New Trends (GMAI'06)*, 209–214. DOI:<https://doi.org/10.1109/GMAI.2006.18>
- Ibrahim, M. Z., & Mulvaney, D. J. (2015). Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping. *Journal of Visual Communication and Image Representation*, 30, 219–233. DOI:<https://doi.org/10.1016/j.jvcir.2015.04.013>
- Indrabayu, Hasanuddin, & Nurtanio, I. (2019). An Approach in Auto Valuing for Optimal Threshold of Viola Jones. *Symposium of Emerging Nuclear Technology and Engineering Novelty*. DOI:<https://doi.org/10.1088/1742-6596/1198/9/092003>
- Joshi, A., Chavan, V., & Kaveri, P. (2021). Semantic Gap Reduction From Mouth Feature Threshold Value Using Viola Jones Algorithm. *ICCRDA 2020*. DOI:<https://doi.org/10.1088/1757-899X/1022/1/012065>
- Kaur, J., & Sharma, A. (2017). Performance Analysis of Face Detection by using Viola-Jones algorithm. *International Journal of Computational Intelligence Research*, 13(5), 707–717.
- Kherchaoui, S., & Houacine, A. (2010). Face detection based on a model of the skin color with constraints and template matching. *2010 International Conference on Machine and Web Intelligence*, 469–472. DOI:<https://doi.org/10.1109/ICMWI.2010.5648043>
- Maghraby, A. El, Abdalla, M., & El, M. Y. (2014). Detect and Analyze Face Parts Information using Viola-Jones and Geometric Approaches. *International Journal of Computer Applications*, 101(3), 23–28. DOI:<https://doi.org/10.5120/17667-8494>
- MathWorks. (n.d.). vision.CascadeObjectDetector (Detect objects using the Viola-Jones algorithm). Retrieved December 27, 2022, from <https://www.mathworks.com/help/vision/ref/vision.CascadeObjectDetector-system-object.html>
- Nguyen, D., Halupka, D., Member, S., Aarabi, P., Sheikholeslami, A., & Member, S. (2006). Real-Time Face Detection and Lip Feature Extraction Using Field-Programmable Gate Arrays. *IEEE Transactions on Systems, Man, and Cybernetics*, 36(4), 902–912. DOI:<https://doi.org/10.1109/TSMCB.2005.862728>
- Papageorgiou, C. P., Oren, M., & Poggio, T. (1998). A general framework for object detection. *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 555–562. DOI:<https://doi.org/10.1109/ICCV.1998.710772>
- Radhi, A. A., Akkar, H. A. R., & Abdullah, H. N. (2022). Skewness and access kurtosis as denoised mixed features-based K-Medoids for cooperative spectrum sensing. *Physical Communication*, 54, 101831. DOI:<https://doi.org/10.1016/j.phycom.2022.101831>
- Ryu, Y.-S., & Oh, S.-Y. (2001). Automatic extraction of eye and mouth fields from a face image using eigenfeatures and multilayer perceptrons. *Pattern Recognition*, 34(12), 2459–2466. DOI:[https://doi.org/10.1016/S0031-3203\(00\)00173-4](https://doi.org/10.1016/S0031-3203(00)00173-4)
- Safi, Mohammed E., & Abbas, E. I. (2016). Microcontroller-Controlled security door based on speech recognition. In *Al-Sadiq International Conference on Multidisciplinary in IT and Communication Techniques Science and Applications, AIC-MITCSA 2016* (pp. 48–53). DOI:<https://doi.org/10.1109/AIC-MITCSA.2016.7759909>
- Safi, Mohammed Ehsan, & Abbas, E. I. (2022). Isolated word recognition based on PNCC with different classifiers in a noisy environment. *Applied Acoustics*, 195, 108848. DOI:<https://doi.org/10.1016/j.apacoust.2022.108848>
- Safi, Mohammed Ehsan, & Abbas, E. I. (2023). Lip Segmentation for Visual Speech Recognition Based on the Convolution Process. *2023 International Conference on Engineering Applied and Nano Sciences (ICEANS)*, 102–106. DOI:<https://doi.org/10.1109/ICEANS58413.2023.10630463>
- Setu, T. A., & Rahman, M. (2016). Human Face Detection and Segmentation of Facial Feature Region. *Global Journal of Computer Science and Technology: G Interdisciplinary*, 16(1).
- Viola, P., & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. DOI:<https://doi.org/10.1109/CVPR.2001.990517>
- Zhang, C., & Zhang, Z. (2010). A Survey of Recent Advances in Face Detection. *Microsoft Research*, 1(MSR-TR-2010-66). <https://www.microsoft.com/en-us/research/publication/a-survey-of-recent-advances-in-face-detection/>