ARTICLE

JOURNAL`S UNIVERSITY OF BABYLON FOR ENGINEERING SCIENCES (JUBES)

مـــجلـة جـــامعة بـــابـل للعلـــوم الهندسية

Vol. 32, No. 4. \ 2024 ISSN: 2616 - 9916

Grammatical Facial Expression Recognition Based on Machine Learning

Hayder A. Ismael[1], Sana A. Nasser[2], Rula Sami Aleesa[3], Zahraa H. Al-Obaide[4], Ivan A. Hashim[5]

[1]*Department of Biomedical Engineering, Al-Khwarizmi College of Engineering, University of Baghdad, Baghdad, Iraq*

hayder.a@kecbu.uobaghdad.edu.iq

[2]*Department of Mechatronics Engineering, Al-Khwarizmi College of Engineering, University of Baghdad, Baghdad, Iraq*

Sana.a@kecbu.uobaghdad.edu.iq

[3]*Faculty of Materials Engineering, University of Babylon, Babylon, Iraq*

mat.rula.sami@uobabylon.edu.iq

[4]*Department of Systems Engineering, College of Information Engineering, Al-Nahrain University, Baghdad, Iraq*

zahraa.h.n@nahrainuniv.edu.iq

[5]*Department of Electrical Engineering, University of Technology, Baghdad, Iraq*

30095@uotechnology.edu.iq

Received:	23/6/2024	Accepted:	31/7/2024	Published:	14/8/2024

Abstract

Facial expression recognition is an evolving field of research with various applications in system development. Recognizing facial expressions is particularly crucial in discourse construction. This study focuses on the design of a grammatical facial expressions (GFEs) recognition system that depended on extracted features from the estimation of head pose and detection of Action Units (AUs) in order to recognize grammatical expressions. The Facial Action Coding System (FACS) is utilized to depict AUs, which effectively capture and classify the intricate movements of facial muscles. Among the AUs, AUs 1 and 4 serve as potential indicators for recognizing grammatical expressions. The process of estimating head pose produces characteristics, including Euler angles (namely, pitch, roll, and yaw), as well as 3D coordinates, which signify the relative arrangements of facial landmarks in correspondence to the camera. The present study employs a dataset comprising video recordings obtained from a sample of 53 individuals whose ages range from 18 to 44 years. Two distinct classifications,

namely Multilayer Perceptron (MLP) and K-Nearest Neighbor (KNN), are employed in the final stage of the proposed system. The experimental outcomes show that the KNN classifier attains better efficacy in contrast to the MLP classifier.

Keywords: – Grammatical Facial Expressions (GFEs); Facial Action Coding System (FACS); Multilayer Perceptron (MLP); K-Nearest Neighbor (KNN).

1. Introduction

In recent times, there has been a significant surge in the interest surrounding grammatical facial expression (GFE) recognition systems, particularly in light of the marked advancements made in computer technology, sensor capabilities, and camera technology. The GFEs recognition system has several possible uses, including aiding in daily decision-making [1] and facilitating communication among the hearing-impaired [2].

Expressions exhibited on the face are a fundamental component of body language. The term "body language" refers to the many forms of nonverbal human communication. There exists no human civilization on Earth that does not utilize this language as a mode of communication, however, its exact connotations may differ across diverse cultures. The majority (93% of all communication) is nonverbal, including tone of voice (38%), body language (55%), and only 7% spoken words [3].

The facial expressions exhibited by humans are a highly potent manifestation of nonverbal communication. Facial expressions are caused by facial muscles and fascia. The muscles move, which causes the skin to move. This causes lines and folds to show up on the face, which move facial features like the mouth and eyebrows. Facial expression research goes back to 1872, when Darwin said that a person's emotions and thoughts show what's going on inside their head [4]. The analysis of facial expressions has gained substantial prominence in recent years, predominantly owing to advancements in closely associated academic fields, namely face detection, face tracking, and face recognition, as well as the availability of low-cost computing power.

In order to extract characteristics from tracked faces, researchers have developed methods based on deformable models like Active Shape Models (ASM) [5]. In order to extract facial characteristics, the Facial Action Coding System (FACS) used Action Units (AUs) to characterize a set of facial muscle movements [6]. The Euler angles (roll, pitch, and yaw) and 3D coordinates of the face landmark locations in relation to the camera are the characteristics of head pose estimation [7]. These characteristics are obtained from a facial landmark monitoring system [8].

The primary goal of this research is to create a system that can identify nine different types of grammatical phrases by using characteristics taken from the detection of AUs and head posture estimation.

2. Mechanism Work of The GFEs Recognition System

The recognition system of GFEs comprises three principal components, namely video capture and pre-processing, feature extraction, and decision-maker, as evident in Fig. 1. Capturing and preprocessing video is the initial stage for the GFEs recognition system. The video capture process involves filming the participant's face while they say a series of grammatical words. At this stage, the acquired video is edited, faces are detected, and important landmarks on the face pictures are located and identified so that characteristics may be extracted. The second step in recognizing GFEs involves extracting features from tracking a collection of landmarks projected on the facial pictures, which greatly improves the recognition rate. The decision-maker is the last part of the GFEs recognition system. It uses the extracted features to decide the best output choice. This section will describe in depth each step of the proposed system.

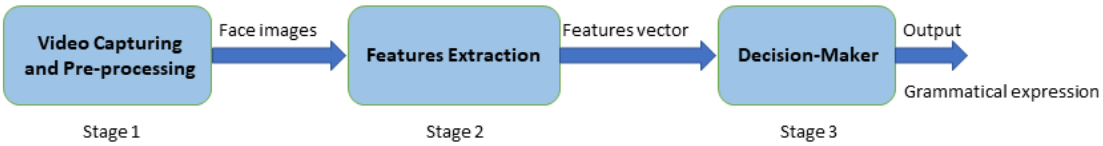


Fig. 1: An overview of the GFEs recognition system block diagram.

2.1 Video Capturing and Pre-processing

A Video of the participants' faces expressing a series of grammatical statements is recorded using a digital camera. To capture changes in facial behavior and head attitude, a digital camera should be positioned at a consistent distance from the participant's face. Video editing is required to separate the grammatical words expressed by each participant.

In order to extract features, it is required to first identify and then track faces. The Viola-Jones (VJ) algorithm is a highly favored technique for real-time face detection, as well as being extensively applied in GFE recognition systems. The VJ algorithm's main properties are its versatility and accuracy in identifying several faces inside a single picture, regardless of their skin tone or whether or not they wear glasses [9]. Facial tracking includes landmark detection, tracking, and head posture estimation. Facial landmark detection is used to specify a certain number of points of interest in a face picture. Facial landmark tracking refers to the process of following a series of points in a video by either treating each frame as separate or by making use of the video's temporal data. The tracking of facial landmarks is sometimes referred to as "non-rigid tracking" since the face is a very flexible object. Face alignment and face registration are two more names for it. Head posture estimate is based on tracking a collection of interest points

Vol. 32, No. 4. \ 2024 ISSN: 2616 - 9916

in a video. There are multiple methods for facial tracking. Constrained Local Neural Fields (CLNF) is a strong method for detecting facial landmarks in unrestricted environments [10]. In the CLNF model, the global motion (rigid) parameters are represented by (s, w, t), while the local non-rigid parameter is represented by (q). These four distinct categories of parameters are employed to control the instance of the face in an image. Using weak perspective projection, the following equation is used to place a single point of the 3D Point Distribution Model (PDM) in an image [11].

$$X_i = s \, . R_{2D} \, . (\bar{X}_i + \Phi_i q) + t \qquad (1)$$

Where $t = [t_x, t_y]^T$ is the translation term, $\bar{X}_i = [\bar{X}_i, \bar{Y}_i, \bar{Z}_i]^T$ is the average value of the i^{th} feature, R_{2D} is 2×3 rotation matrix, s is a scaling factor that controls the relative distance of the face from the camera, q is a vector of parameters that affect the non-rigid form, and it has m dimensions, and Φ_i is a $3 \times m$ principal component matrix.

2.2 Features Extraction

The GFEs recognition system relies on feature extraction, which in turn relies on a number of various methods and techniques.

FACS is the most well-known method that Paul Ekman developed to classify how people move their faces [12]. FACS employed AUs to characterize emotional facial muscle movements. Each AU is linked to one or more muscular movements [13]. FACS uses 18 AUs and several types of eye and head angles to explain face actions. Each AU has a number that represents one or more face muscle moves.

The CLNF approach is able to estimate the head pose based on three parameters of the shape model (rotation, scaling, and translation). Head posture estimation is sometimes called rigid tracking because the head is often thought of as a rigid object. The Generalized Adaptive View-based Appearance Model (GAVAM) was used by CLNF to track the rigid head posture under different lighting situations [14]. GAVAM is a keyframe-dependent differential tracker. To predict the next frame's position based on previously captured images, it employs a 3D scene flow [15]. Keyframes are assembled and refined using the Kalman filter to apply to different points in time in the video. This leads to accurate tracking and a reduction in deviation. Several different techniques exist for the representation of rotation. Euler angles (roll, pitch, and yaw) are employed for this function. As demonstrated in Fig. 2, rotations about the three axes align with the triumvirate of angles comprising roll, pitch, and yaw.

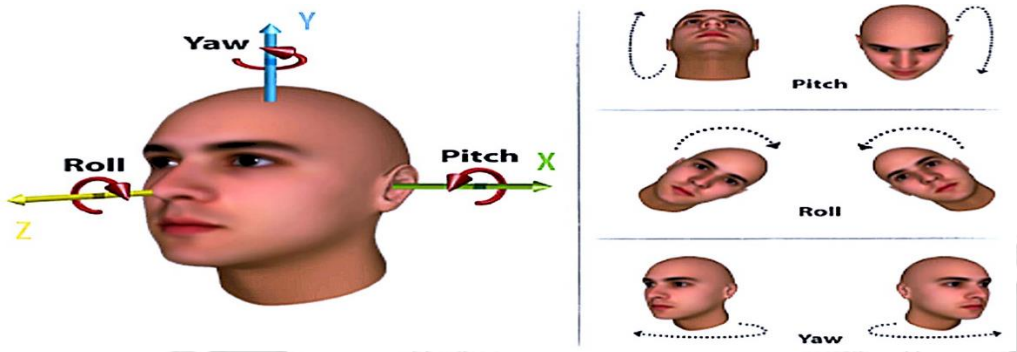


Fig. 2: Pitch, roll, and yaw angles of the head [16].

2.3 Decision-Maker

The final step in the system for recognizing GFEs is the decision-maker. In order to distinguish nine different types of grammatical expressions, face-tracking characteristics are extracted and then introduced into the classification stage. Two types of classifiers are proposed in this study for classifying grammatical expressions: K-Nearest Neighbor (KNN), and Multilayer Perceptron (MLP).

The KNN algorithm stands out as a typical instance of a machine learning technique that is both lucid and exact. The GFEs recognition system recognizes nine types of grammatical expressions using the KNN algorithm. The selection of the distance metric employed in identifying the closest training sample to each new data point substantially affects the algorithm's accuracy potential. Fig. 3 depicts the KNN classification based on the number of the K closest neighbors. There are numerous methodologies available for computing the distance between test samples and all training samples. These techniques include the Euclidean and Manhattan methodologies. The present study employs the Euclidean distance to compute distance, as demonstrated in the equation below [17]:

$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad (2)$$

The symbol $p_i = (p_1, p_2, \dots, p_n)$ denotes the test sample whose values are unknown. On the other hand, $q_i = (q_1, q_2, \dots, q_n)$ represents the training sample extracted from the dataset. The variable n indicates the size of the features.

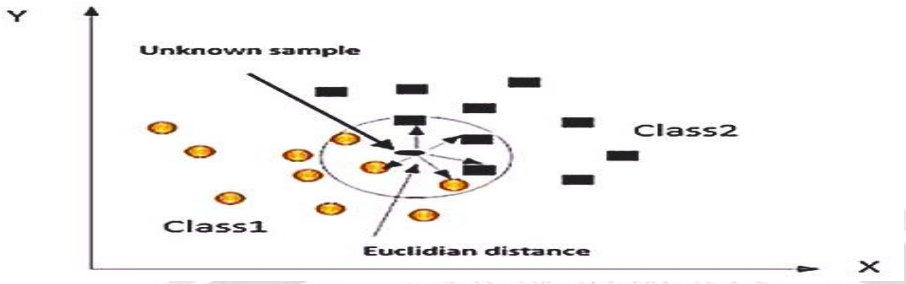


Fig. 3: A KNN algorithm example with two classes [18].

The realm of artificial neural networks often relies on the Multilayer Perceptron (MLP) as a prominent classification technique. A creative approach to this method is essential for optimal results. It comprises an input layer that receives data, hidden layers that facilitate data processing, and an output layer that presents the categorization outcomes.

There are a set number of nodes (processing) in each layer. The input layer is distinct from the other processing layers because it performs no operations. The input layer assumes the role of receiving data, while the ensuing layers undertake the responsibility of executing computations at every node until the output value is derived from each of the output nodes. Layer-by-layer biases are used to transmit the input signal through the MLP neural network, with the neurons of the first hidden layer coupled to the lower layers through weights. In the quest for relevance, the output is subjected to a rigorous comparison against a target pattern that bears a direct correlation to the input. The MLP neural network's weight is adjusted to decrease the error between the output layer and the desired pattern. The equation below describes how the output layer is calculated [19]:

$$Vj = \sum_{i=1}^{p} W_{ji} X_i + \theta_j \qquad (3)$$
$$y_j = f_j(Vj)$$

Where Vj is linear combination of inputs ($X_1, X_2,..., X_p$), The weight of the link between neuron j and input X_i is denoted by W_{ji}, θ_j represent the bias, $f_j(.)$ is the activation function of the jth neuron, and y_j is the output.

The GFEs system uses an MLP neural network split into distinct recall and training stages. The MLP network employs a backpropagation (BP) algorithm during its training stage. In order to optimize the neural network's weights, the input characteristics are fed forward. During the training process, the BP makes adjustments to the weights by sending mistakes at the output layer back to the prior layer. The MLP uses the goal as well as the input values in the training data to determine how much the weights should be changed each time in order to minimize the gap between the goal and the output values. Errors in the network may be reduced with enough

training at the specified epoch. An MLP will converge to a target degree of accuracy at each epoch [20]. During the recall stage, the MLP network reacts to the input features that display characteristics that are comparable to the learned characteristics during the training stage [21].

3.Collected Dataset

A dataset is a valuable collection of data that serves as a platform to confidently test the suggested system. Since uses a few datasets and is not available to the public in previous grammatical facial expression research, it is necessary to acquire a new dataset in order to assess the efficacy of the suggested GFE recognition system. The study included a cohort of individuals aged 18 to 44, consisting of 31 males and 22 females who were participants.

The data is gathered in a realistic setting without the use of specific lights or sophisticated equipment. The participant is observed to be positioned in front of a solitary camera, as depicted in Fig. 4.

Fig. 4: The recording session. Recording video for the participant's face during the performance of grammatical sentences.

4. The Proposed GFEs Recognition System

The proposed system for recognizing GFEs is designed in three phases. In the proposed system's initial phase, video is recorded and preprocessed in order to recognize faces and identify landmarks on them so that characteristics may be extracted. The second phase involves the extraction of features, which are obtained from the identification of AUs and head posture estimation. The decision-maker is the last step, and it is used to categorize sentences into nine different types of grammatical expressions using the features extracted from the previous phase. Two classification methods, KNN and MLP, are proposed in this study to identify nine types of grammatical statements. Fig. 5 depicts the steps involved in the suggested GFE recognition system.

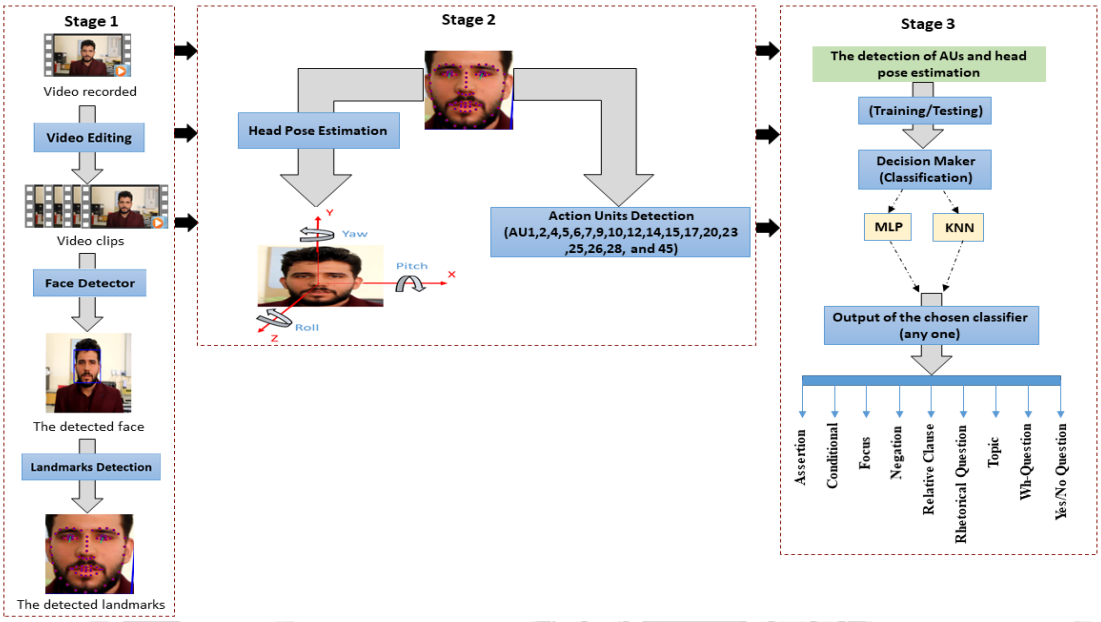


Fig. 5: The configuration stages of the suggested GFE recognition system.

4.1 The Processing of the Proposed System

The Canon 4000D is a digital camera used for video recording. The recorded video has a MOV file extension. One video has been recorded for each participant, and then the video has been edited into several sub-videos in order to acquire the necessary parts when the participant expresses grammatical sentences. The editing was done with Windows Movie Maker.

After the videos are edited, 548 clips are acquired, split as follows: 65 for assertion expressions, 63 for conditional expressions, 46 for focus expressions, 64 for negation expressions, 61 for relative clause expressions, 65 for rhetorical question expressions, 68 for topic expressions, 56 for wh-question expressions, and 60 for yes/no question expressions. The range of the duration of video clips spans from 0.6 to 1.8 seconds, thereby implying that the quantity of frames encompassed within each video clip lies between 15 to 45 frames.

Face detection is the next phase in the video recording and pre-processing phase, and it's utilized to examine whether or not the face area is available in the provided video. The VJ algorithm is frequently utilized for the objective of detecting facial features in a series of images. This algorithm employed the grayscale values of the image to extract feature pixels through the utilization of the feature block technique. This technique is based on a Haar-like feature block set that incorporates an AdaBoost classifier. The Haar-like feature block set encompasses three distinct types of Haar-like features. The extracted features encompass the essential information required to characterize the facial region. In this algorithm, a total of 31 cascades of AdaBoost

layers were employed, with a specified threshold value of 3. Fig. 6 depicts the outcomes obtained from the application of the Viola-Jones algorithm to the participant's face.

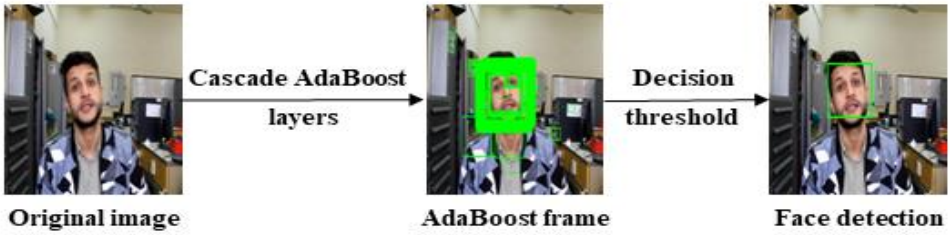


Fig. 6: Face detection for the participant using the VJ algorithm.

As a consequence of the procedure of detecting the face, a rectangular framework is depicted encircling the countenance, thereby establishing the preliminary positioning of key features and structural attributes. The process of fitting the CLNF model can be initiated by utilizing the initial shape parameters. The fitting model is designed to identify the most optimal shape parameters. Fig. 7 depicts the flow map for a CLNF model-based landmark point identification and tracking system. Tracking landmark points in picture sequences assists in obtaining features.

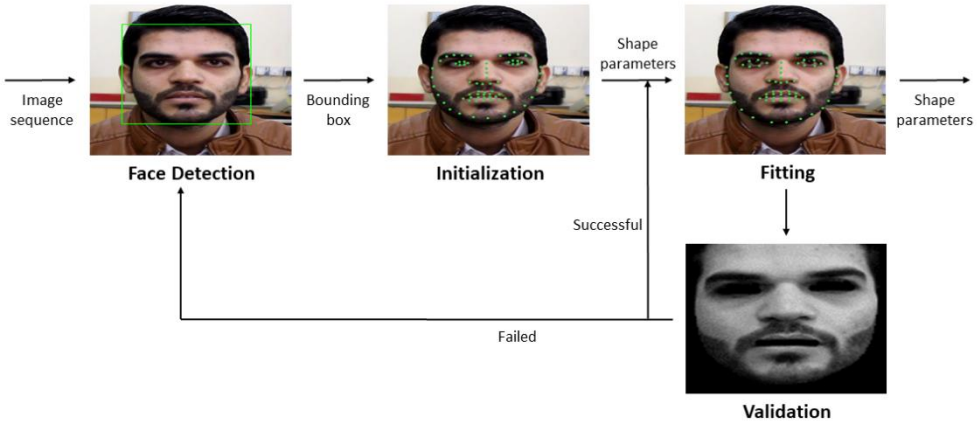


Fig. 7: The flow map for a CLNF model.

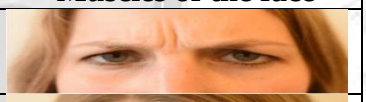
4.2 Result of the Features Extraction

The extracted features in this study encompass Action Units (AUs), Euler angles, and the 3D coordinates' positions. The GFEs recognition system suggested in this study identifies a total of 18 AUs (1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45). It should be noted that while all these AUs are considered, some of them do not significantly impact the grammatical expression recognition process. Head orientation estimation employs Euler angles (pitch, yaw, and roll) which are described in degrees to represent the orientation of the head. In

this particular study, the 3D coordinates are utilized to indicate the precise locations of facial landmarks when viewed in relation to the camera. The translation term signifies the face's location relative to the camera in two-dimensional (x, y), and the scaling term reflects the participant's facial proximity to the camera $(s = \frac{1}{z})$.

In this study, it has been observed that only two specific Action Units (AUs 1 and 4) have a notable impact when participants express grammatical sentences. Table 1 provides a summary of these two Action Units (AUs 1 and 4) as promising indicators for the recognition of grammatical expressions.

Table 1: AUs with their region effect and description.

Action Unit	Characterization	Muscles of the face
AU4	Brow Lowerer	
AU1	Inner Brow Raiser	

To effectively classify the nine categories of grammatical expressions, it is essential to identify the distinctive features associated with a head pose. The main feature that has been extracted to classify the assertion expression in this study is the pitch. Negative pitch values are indications of upward head movements, whereas positive pitch values indicate to signify downward head movements. To quantify the rate of change between raised and lowered head positions, the differences between pitch values in each video clip are calculated and then subjected to a signum function. The signum function assigns $+1$ for positive differences, -1 for negative differences, and 0 for zero differences. These differences are then compared to ± 1 and 0. The resulting values include 0, positive values, and negative values. Zero values indicate either upward or downward head movement based on the corresponding positive or negative values. The existence of negative values implies the manifestation of head transitions in an ascending trajectory, whereas the appearance of positive values indicates head transitions in a descending trajectory. Finally, the locations of non-zero values (positive and negative) are determined to calculate the length of the 0 values. A length greater than 20 is considered indicative of an assertion expression, as depicted in Fig. 8.

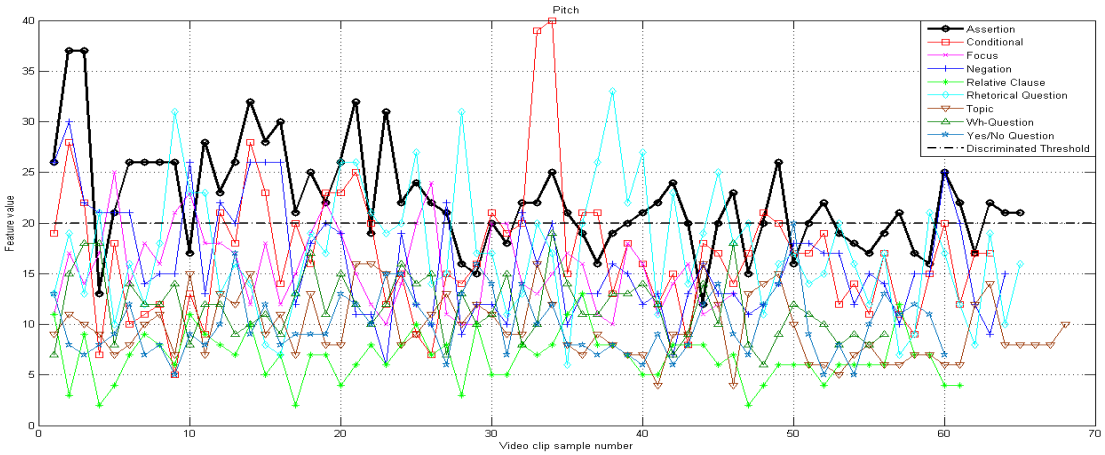


Fig. 8: A specific feature value that identified for all video clips for identifying the assertion expression.

The distinction between the relative clause and assertion expressions relies on the positioning of the face in relation to the camera, specifically the y-coordinate location. To quantify the rate of change between raised and lowered head positions, the differences between y-coordinate values in each video clip are calculated. Subsequently, the utilization of standard deviation is implemented to ascertain the mean deviation of the difference outcomes from the arithmetic mean. A standard deviation result greater than 3 is identified as an assertion expression, while a result less than 1 is designated as a relative clause expression, as illustrated in Fig. 9. The mathematical equation below represents the calculation of the standard deviation:

$$S = \sqrt{\frac{1}{r-1}\sum_{i=1}^{r}(Y_i - \bar{Y})^2} \quad (4)$$

As shown in the equation above, Y_i represents an individual data value, $\bar{Y}$ represents the mean, and r denotes the total number of data points.

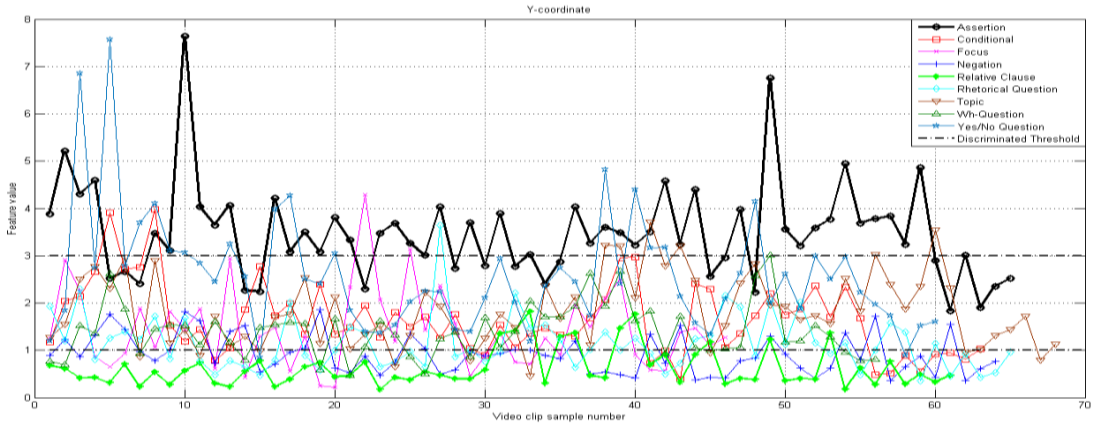


Fig. 9: A specific feature value that identified for all video clips for identifying the assertion and relative clause expressions.

Regarding the negation expression, the extracted feature for analysis is yaw. Head gestures towards the right are depicted by positive yaw values, while head gestures towards the left are represented by negative yaw values. To determine the mean's average deviation from the yaw values, one calculates the standard deviation for the yaw values of each video clip. A standard deviation result greater than 4 is considered indicative of a negation expression, as depicted in Fig. 10.

As depicted in Fig. 10, a standard deviation result lower than 1.5 is designated as indicative of a relative clause expression.

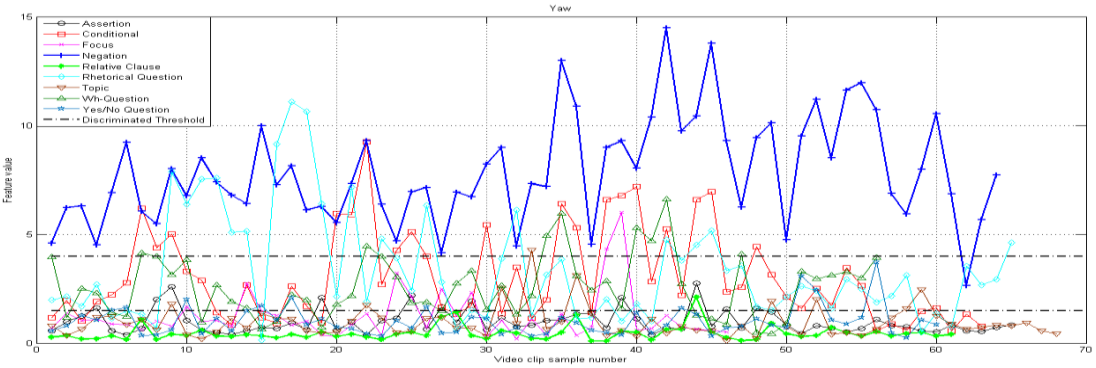


Fig. 10: A specific feature value that identified for all video clips for identifying the relative clause and negation expressions.

The recognition of negation expression can be accomplished by considering the positioning of the face in relation to the camera, particularly with respect to the x-coordinate location. To assess the rate of change associated with continuous right-to-left head movement, the differences between the x-coordinate values in each video clip are calculated. Subsequently,

the variance is applied to measure the prevalence of the difference results from the mean. A variance result exceeding 16 is designated as indicative of a negation expression, as illustrated in Fig. 11. The subsequent equation illustrates the arithmetic operation used for determining the variance:

$$S^2 = \frac{1}{r-1}\sum_{i=1}^{r}(Y_i - \bar{Y})^2 \quad (5)$$

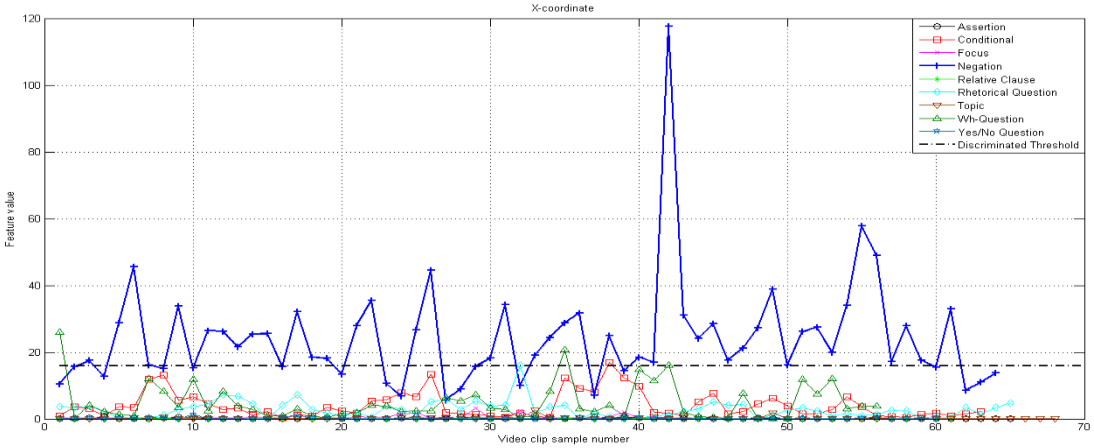


Fig. 11: A specific feature value that identified for all video clips for identifying the negation expression.

By examining the z-coordinate location, the focus expression can be distinguished based on the proximity of the face to the camera. To differentiate the focus expression from other grammatical expressions, the computational process involves determining the discrepancy between the highest and lowest values of the z-coordinate and subsequently dividing it by the minimum z-coordinate value. Consequently, a result exceeding 0.1 is identified as indicative of a focus expression, as depicted in Fig. 12.

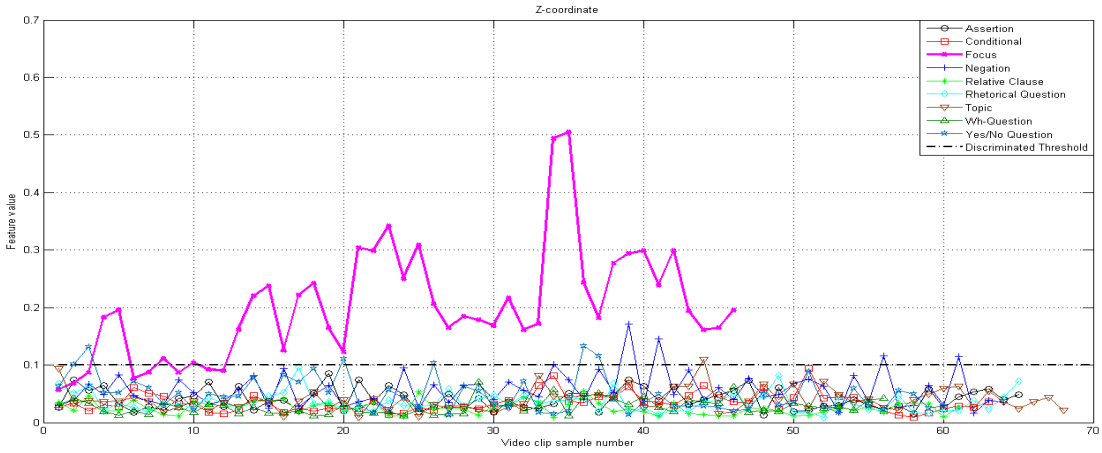


Fig. 12: A specific feature value that identified for all video clips for identifying the focus expression.

In order to distinguish between rhetorical question and conditional expressions, the feature extracted for this purpose is the roll. Leftward head tilts are denoted by affirmative roll values, conversely, rightward head tilts are symbolized by negative roll values. To assess the rate of change in head tilt for the rhetorical question and conditional expressions, the differences between roll values in each video clip are calculated, and the resulting differences are then accumulated. A result exceeding 5 is designated as a rhetorical question expression, while a result lower than -5.5 is specified as a conditional expression, as depicted in Fig. 13.

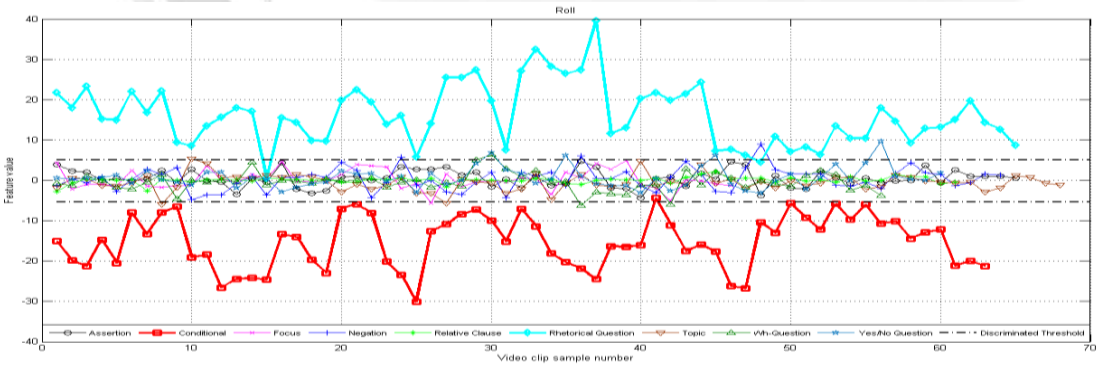


Fig. 13: A specific feature value that identified for all video clips for identifying the conditional and rhetorical question expressions.

To differentiate between yes/no question and topic expressions, the key feature extracted for this task is the pitch. The statistical analysis involves determining the differences between pitch values to gauge the rate of change in head lowering for the yes/no question expression and head raising for the topic expression. The mean of these differences is then calculated. A result

exceeding 0.5 is designated as a yes/no question expression, while a result below -0.5 is specified as a topic expression, as depicted Fig. 14.

Fig. 14 indicates that a result below -0.5 is designated as a wh-question expression. The participation of AUs is pivotal in distinguishing between the wh-question and topic expressions. The subsequent equation illustrates the arithmetic operation used for determining the mean:

$$\bar{Y} = \frac{1}{r}\sum_{i=1}^{r} Y_i \qquad (6)$$

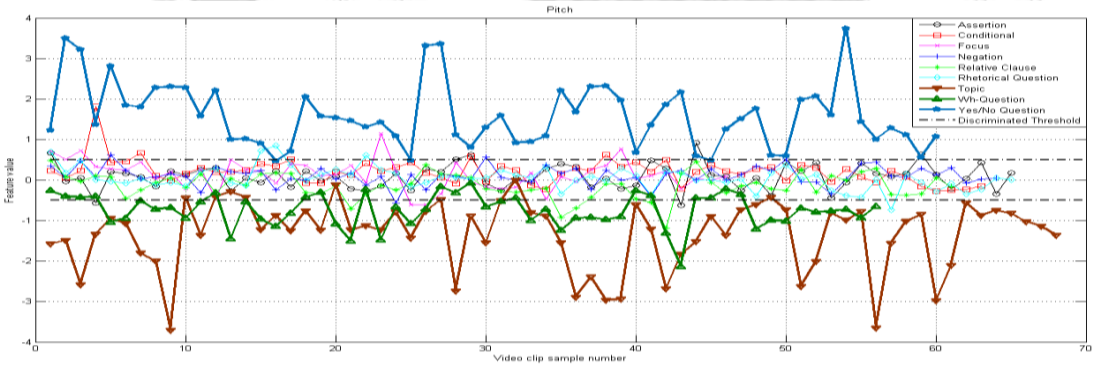


Fig. 14: A specific feature value that identified for all video clips for identifying the wh-question expressions, yes/no question, and topic.

4.3 Classification

The classification phase suggested the utilization of two kinds of classifiers, specifically KNN and MLP, to recognize and classify nine distinct grammatical expressions. The dataset that was collected comprises a total of 548 video clips. The distribution of these clips is as follows: assertion expressions account for 65 clips, while conditional expressions are represented by 63 clips. Additionally, 46 clips were collected for focus expressions, 64 for negation expressions, 61 for relative clause expressions, 65 for rhetorical question expressions, 68 clips for topic expressions, 56 for wh-question expressions, and finally, 60 clips for yes/no question expressions. The dataset is split into two sets: training and testing. The training set comprises of 272 video clips, whereas the testing set encompasses 276 video clips. The present study has distributed a set of randomly selected video clips for training purposes. This distribution encompasses 32 clips for assertion expressions, 31 for conditional expressions, 23 for focus expressions, 32 for negation expressions, 30 for relative clause expressions, 32 for rhetorical question expressions, 34 for topic expressions, 28 for wh-question expressions, and 30 for yes/no question expressions. The remaining video clips were utilized during the testing stage.

5. Performance Results

The performance outcomes of the recommended classifiers (KNN and MLP) are presented in Table 2. The table provided in this context represents the accuracy of recognizing nine categories of grammatical expressions using the gathered database, which includes participation from both male and female individuals.

Table 2: The overall rate of accuracy achieved by the suggested classifiers.

	The accuracy achieved by the classifier (%)	
Grammatical expression	KNN	MLP
Assertion expression	84.85	75.76
Conditional expression	96.88	100
Focus expression	95.65	95.65
Negation expression	100	93.75
Relative Clause expression	67.74	64.52
Rhetorical Question expression	96.97	96.97
Topic expression	97.06	100
Wh-Question expression	92.86	89.29
Yes/No Question expression	90	90
Overall accuracy rate %	**91.3**	**89.49**

As indicated in Table 2, the negation expression obtained the highest accuracy in its recognition rate when employing the KNN classifier. In contrast, the conditional and topic expressions obtained the highest accuracy in their recognition rate when using the MLP classifier. The recognition rate of the relative clause expression obtained the lowest accuracy when employing KNN and MLP classifiers. Regarding the remaining classes classified by KNN and MLP, their accuracy ranges from 75.76% to 97.06%. The proposed GFEs recognition system achieves a final accuracy rate result of 91.3% when utilizing the KNN classifier, while achieved 89.49% when using the MLP classifier.

6. Results Comparison With Prior Studies

The performance of the proposed GFEs recognition system is evaluated using the newly described feature extraction approach. Tracking a preset set of markers on the face is used to extract features. These collected properties are fed into multiple classifiers, which categorize nine different types of grammatical expressions. When contrasted with the MLP classifier, the KNN classifier yielded the most favorable outcomes. The KNN and MLP classifiers' discoveries are evaluated against pertinent research to gauge the effectiveness of the proposed system in a remarkably innovative manner. Table 3 compares the proposed system to prior studies in terms of recognition techniques, grammatical expression categories, participant numbers, number of video clips or sentences, and accuracy. This table depicts the numerous recognition strategies used in prior studies to recognize different kinds of grammatical statements. The recommended system's performance is dependent on both participant numbers and accuracy, which are the two key obstacles it faces.

Table 3: Evaluating the effectiveness of the suggested system in comparison to prior studies.

Ref. No.	Recognition Techniques	Grammatical Expression Categories	Participant Count	No. of Video clips / No. of Sentences	Accuracy %
[22]	HMSVM	5	3 participants	330 sentences	Recall = 80.37 & Precision = 81.08
[23]	Combination of HMM and NN	4	4 participants	84 video clips	84
[24]	HMM	3	3 participants	147 video clips	84.1
[25]	Combination of HMM and SVM	6	7 participants	297 video clips	87.71
[26]	MLP	9	2 participants	45 sentences	89.4
[27]	2-layer conditional random fields	6	3 participants	129 video clips	Recall = 85.54 & Precision = 93.76
[28]	HMSVM	5	3 participants	45 video clips	91
[29]	SVM	2	8 participants	136 video clips	Over 95
Suggested system	KNN	9	53 participants	548 video clips	91.3

NN – Neural Network; HMSVM – Hidden Markov Support Vector Machine

HMM – Hidden Markov Model; SVM – Support Vector Machine

As shown in Table 3, the suggested system exhibits superior performance compared to related studies concerning the collected dataset and achieved accuracy.

7. Conclusions

The main goals of the suggested GFEs recognition system are to identify Action Units (AUs) and estimate head pose, aiming to utilize them as dependable indicators for recognizing nine categories of grammatical expressions. The recognition of grammatical expressions is significantly influenced by two specific Action Units (AUs1 and 4), which prove to be highly effective. The removal of certain features (AUs 2, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45) does not impact the recognition rate. To differentiate between grammatical expressions,

specific features related to head pose estimation are extracted. These features are obtained through the tracking of 68 facial landmark points. The suggested system operates independently and is capable of functioning in unconstrained environments.

References

[1] R. S. Muhammad and M. I. Younis, "The limitation of pre-processing techniques to enhance the face recognition system based on LBP," Iraqi J. Sci., pp. 355–363, 2017.

[2] C. Neidle *et al.*, "A method for recognition of grammatically significant head movements and facial expressions, developed through use of a linguistically annotated video corpus," in *Proc. of 21st ESSLLI Workshop on Formal Approaches to Sign Languages*, Citeseer, 2009.

[3] N. Berthouze, T. Fushimi, M. Hasegawa, A. Kleinsmith, H. Takenaka, and L. Berthouze, "Learning to recognize affective body postures," in *The 3rd International Workshop on Scientific Use of Submarine Cables and Related Technologies, 2003.*, IEEE, 2003, pp. 193–198.

[4] C. Darwin, "1965. The expression of the emotions in man and animals," *London, UK John Marry*, 1872.

[5] T. F. Cootes and C. J. Taylor, "Active shape models—'smart snakes,'" in *BMVC92: Proceedings of the British Machine Vision Conference, organised by the British Machine Vision Association 22–24 September 1992 Leeds*, Springer, 1992, pp. 266–275.

[6] T. Zeyad, "Human Face Recognition Using GABOR Filter And Different Self Organizing Maps Neural Networks," Al-Khwarizmi Eng. J., vol. 1, no. 1, pp. 38–45, 2005.

[7] S. J. A. Al-Atroshi and A. M. Ali, "Facial Expression Recognition Based on Deep Learning: An Overview," Iraqi J. Sci., pp. 1401–1425, 2023.

[8] W. A. Mahmoud, A. I. Abbas, and N. A. S. Alwan, "Face Identification Using Back-Propagation Adaptive Multiwavenet," J. Eng., vol. 18, no. 3, 2012.

[9] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," 2010.

[10] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 354–361.

[11] T. Baltrušaitis, "Automatic facial expression analysis," University of Cambridge, 2014.

[12] P. Ekman and E. L. Rosenberg, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), New York, NY, US: Oxford University, 2005.

[13] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, 2001.

[14] L.-P. Morency, J. Whitehill, and J. Movellan, "Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation," in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, IEEE, 2008, pp. 1–8.

[15] S. Vedula, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 475–480, 2005.

[16] E. N. Arcoverde Neto *et al.*, "Enhanced real-time head pose estimation system for mobile device," *Integr. Comput. Aided. Eng.*, vol. 21, no. 3, pp. 281–293, 2014.

[17] I. Babaoğlu, M. S. Kıran, E. Ülker, and M. Gündüz, "Diagnosis of coronary artery disease using artificial

bee colony and k-nearest neighbor algorithms," *Int. J. Comput. Commun. Eng.*, vol. 2, no. 1, pp. 56–59, 2013.

[18]P. Ray, "Satellite image processing using KNN rules," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 5, no. 9, pp. 125–127, 2015.

[19]H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, "A multilayer perceptron-based medical decision support system for heart disease diagnosis," *Expert Syst. Appl.*, vol. 30, no. 2, pp. 272–281, 2006.

[20]M. Azarbad, S. Hakimi, and A. Ebrahimzadeh, "Automatic recognition of digital communication signal," *Int. J. energy, Inf. Commun.*, vol. 3, no. 4, pp. 21–33, 2012.

[21]B. Kröse and P. Van der Smagt, "An introduction to neural networks . Amsterdam, The Netherlands: University of Amsterdam." 1996.

[22]D. N. Metaxas, B. Liu, F. Yang, P. Yang, N. Michael, and C. Neidle, "Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking.," in *LREC*, 2012, pp. 2414–2420.

[23]T. D. Nguyen and S. Ranganath, "Tracking facial features under occlusions and recognizing facial expressions in sign language," in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, IEEE, 2008, pp. 1–7.

[24]N. Michael, C. Neidle, and D. Metaxas, "Computer-based recognition of facial expressions in ASL: from face tracking to linguistic interpretation," in *sign-lang@ LREC 2010*, European Language Resources Association (ELRA), 2010, pp. 164–167.

[25]T. D. Nguyen and S. Ranganath, "Facial expressions in American sign language: Tracking and recognition," *Pattern Recognit.*, vol. 45, no. 5, pp. 1877–1891, 2012.

[26]F. A. Freitas, S. M. Peres, C. A. M. Lima, and F. V Barbosa, "Grammatical facial expression recognition in sign language discourse: a study at the syntax level," *Inf. Syst. Front.*, vol. 19, pp. 1243–1259, 2017.

[27]T. D. Nguyen and S. Ranganath, "Recognizing continuous grammatical marker facial gestures in sign language video," in *Asian Conference on Computer Vision*, Springer, 2010, pp. 665–676.

[28]N. Michael, P. Yang, Q. Liu, D. N. Metaxas, C. Neidle, and C. Center, "A Framework for the Recognition of Nonmanual Markers in Segmented Sequences of American Sign Language.," in *BMVC*, 2011, pp. 1–12.

[29]N. Michael, D. Metaxas, and C. Neidle, "Spatial and temporal pyramids for grammatical expression recognition of American Sign Language," in *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, 2009, pp. 75–82.

ARTICLE
JOURNAL`S UNIVERSITY OF BABYLON FOR ENGINEERING SCIENCES (JUBES)
مـــجلـة جـــامعة بـــابـل للعلـــوم الهندسية
Vol. 32, No. 4. \ 2024 ISSN: 2616 - 9916

التعرف على تعبيرات الوجه النحوية بناءً على التعلم الآلي

حيدر عبدالقادر إسماعيل

قسم هندسة الطب الحياتي، كلية الهندسة الخوارزمي، جامعة بغداد، بغداد، العراق

hayder.a@kecbu.uobaghdad.edu.iq

سنا احمد ناصر

قسم هندسة الميكاترونيكس، كلية الهندسة الخوارزمي، جامعة بغداد، العراق

Sana.a@kecbu.uobaghdad.edu.iq

رلا سامي خضير العيسى

قسم هندسة المعادن، كلية هندسة المواد، جامعة بابل، بابل، العراق

mat.rula.sami@uobabylon.edu.iq

زهراء حمزه نزال

قسم هندسة المنظومات، كلية هندسة المعلومات، جامعة النهرين

zahraa.h.n@nahrainuniv.edu.iq

ايفان عبدالزهرة هاشم

قسم الهندسة الكهربائية، الجامعة التكنولوجية، بغداد، العراق

30095@uotechnology.edu.iq

الخلاصة:-

يعد التعرف على تعبيرات الوجه مجالًا متطورًا للبحث وله تطبيقات مختلفة في تطوير النظام. يعد التعرف على تعبيرات الوجه أمرًا بالغ الأهمية بشكل خاص في بناء الخطاب. تركز هذه الدراسة على تصميم نظام التعرف على تعبيرات الوجه النحوية (GFEs) الذي يعتمد على الميزات المستخرجة من تقدير وضعية الرأس والكشف عن وحدات السلوك (AUs) من أجل التعرف على التعبيرات النحوية. يتم استخدام نظام ترميز حركة الوجه (FACS) لتصوير وحدات السلوك، والتي تلتقط وتصنف بشكل فعال الحركات المعقدة لعضلات الوجه. من بين وحدات السلوك، تعمل وحدات السلوك 1 و4 كمؤشرات محتملة للتعرف على التعبيرات النحوية. تنتج عملية تقدير وضعية الرأس خصائص، بما في ذلك زوايا أويلر (pitch, roll, and yaw)، بالإضافة إلى الإحداثيات ثلاثية الأبعاد، التي تشير إلى الترتيبات النسبية لمعالم الوجه بالنسبة إلى الكاميرا. تستخدم الدراسة الحالية مجموعة بيانات تشتمل على تسجيلات فيديو تم الحصول عليها من عينة مكونة من 53 فردًا تتراوح أعمارهم بين 18 و44 عامًا. يتم استخدام تصنيفين متميزين، وهما شبكة عصبية متعددة الطبقات (MLP) والمصنف الذي يعتمد بالاختيار على اقرب جار (KNN)، في المرحلة النهائية من النظام المقترح. تظهر النتائج التجريبية أن مصنف KNN يحقق فعالية أفضل على النقيض من مصنف MLP.

الكلمات الدالة:- تعبيرات الوجه النحوية (GFEs) ، نظام ترميز حركة الوجه (FACS) ،شبكة عصبية متعددة الطبقات (MLP)، المصنف الذي يعتمد بالاختيار على اقرب جار (KNN).