Perception of Emotion in Human-Robot Interaction

Author

Muhammad Faisal Zia

Regn Number

318424

Supervisor

Dr. Sara Ali

DEPARTMENT ROBOTICS AND INTELLIGENT MACHINE ENGINEERING

SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

December, 2022

Perception of Emotion in Human-Robot Interaction

Author
Muhammad Faisal Zia
Regn Number
318424

A thesis submitted in partial fulfillment of the requirements for the degree of
MS Robotics and Intelligent Machines Engineering

Thesis Supervisor:
Dr. Sara Ali

Thesis Supervisor's Signature: _____

DEPARTMENT ROBOTICS AND INTELLIGENT MACHINE ENGINEERIN
SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD
December, 2022

Declaration

I certify that this research work titled "*Perception on Emotion in Human-Robot Interaction"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

Muhammad Faisal Zia

2019-NUST-MS-RIME-000318424

Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Signature of Student

Muhammad Faisal Zia

Registration Number

000318424

Signature of Supervisor

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST School of Mechanical & Manufacturing Engineering (SMME). Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST School of Mechanical & Manufacturing Engineering, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the SMME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST School of Mechanical & Manufacturing Engineering, Islamabad.

Acknowledgements

ALHAMDULILLAH, all praises to ALLAH for the gifts and His blessing and guidance in finalizing this thesis. I would like to express special thanks of gratitude to my supervisor Dr. Sara Ali for her commitments and who gave me the golden chance to finish this great task. I would also like to thank Dr. Yasar Ayaz, Dr. Khawaja Fahad Iqbal and Dr. Muhammad Sajid for being on my thesis guidance and evaluation committee and for their support and cooperation. Finally, I would like to thank my parents and all the people who have helped me with my studies.

Dedicated to my adored siblings and exceptional parents and specially to my wife Esha Faisal whose outstanding assistance and collaboration helped me achieve this fantastic feat.

Abstract

Perception of emotion is an intuitive replication of a person's internal state without the need for verbal communication. Visual emotion recognition has been broadly studied and several end-to-end deep neural networks (DNNs)-based and Machine learning-based models have been proposed but they lack the ability to be implemented in low-specification devices like robots, and vehicles. The drawbacks of conventional handcrafted feature-based Facial Emotion Recognition (FER) methods are eliminated by DNNs-based FER approaches. In spite of that, Deep Neural Network based FER techniques suffer from high processing costs and exorbitant memory requirements, their application is constrained in fields like Human-Robot Interaction (HRI) and Human-Computer Interaction (HCI) and relies on hardware requirements. In aforementioned study, we presented a computationally inexpensive and robust FER system for the perception of six basic emotions (i.e., disgust, surprise, fear, anger, happy, and sad) that is capable of running on embedded devices with constrained specifications. In the first step after pre-processing input images, geometric features are extracted from detected facial landmarks, considering the facial spatial position among influential landmarks. The extracted features are given as input to trainthe SVM classifier. Our proposed FER system was trained and evaluated experimentally using two databases, Karolinska Directed Emotional Faces (KDEF) and Extended Cohn-Kanade (CK+) database. Fusion of KDEF and CK+ datasets at the training level were also employed in order to generalize the FER system's response to the variations of ethnicity, race, national and provincial backgrounds. The results show that our proposed FER system is optimized for real-time embedded applications with constrained specifications and yields an accuracy of 96.8%, 86.7% and 86.4% for CK+, KDEF and fusion of CK+ and KDEF databases respectively. As a part of our future research objectives, the developed system will make a robotic agent capable of perceiving emotion and interacting naturally without the need for additional hardware during HRI.

Key Words: *Deep Neural Network (DNN), Facial Emotion Recognition (FER), Human-Robot Interaction (HRI), Human-Computer Interaction (HCI), Karolinska Directed Emotional Faces (KDEF), Extended Cohn-Kanade (CK+)*

Table of Contents

List of Figures

List of Tables

CHAPTER 1:

Introduction

Facial expressions (FEs) are one of the dominating media for emotion recognition. According to [1] person's facial expression determines 55% of the impact of a spoken message. In this way examination of non-verbal sensing, channels are the key part of understanding and synthesis of emotion in robots [2].

Due to automation, the role of industrial and humanoid robots is set to shift from that of an assistant to that of a companion, a caretaker, and an educator [3]. Human-Computer Interaction (HCI) including Human-Robot Interaction (HRI) and Social Robotics (SR) strives on designing, modeling, Implementing and evaluating such systems [4], that collaborate with human operators and meet the societal and emotive needs of their individual users in conjunction with human values [5].

As we known, FER has traditionally been employed in the study of psychology to determine a person's intentions via facial information within a social situation, and detect lies and mental disorders such as schizophrenia, autism, and depression. However, computer-based applications of FER are on the horizon and Recent computer-based systems [6] [7] [8], aren't just able to move, communicate, or perform video analysis like face recognition; there's also growing interest in employing robots to analyze people's emotional states. Several studies have been made using NAO robots due to their design and programming potential. For example, for curriculum's contents are based on students' emotions during teaching [9], Facial Emotion Recognition in Children [10], augmented reality (AR) based games [11], social interaction with the elderly [12], and children with Autism [13].

Automated facial expression identification has emerged as a crucial and challenging field for social interactions. There have been initiatives to create FER systems that incorporate facial expressions [14] [15] [16], but they suffers limitations due to background changes [17], light intensity, face position, intensity of expression [18], variations in age, national and provincial backgrounds, social

attitudes [19]. However, sill there is room for improvement in the FER module. Based on features, the FER system can be classified into, conventional handcrafted feature-based FER and deep-learning based FER. In conventional handcrafted feature-based FER Approach, On the target face, AUs-based features, spatial features, appearance features, or combination of spatial and appearance features are tracked and retrieved.

First FER approach uses models that have been pretrained to recognize AUs, They are then detected in an input image and used, through decoding the observed AUs, to compute the FE (i.e. AU-26, AU-25, AU-22, and AU1, are detected, the system will classify the emotion of the 'surprise' category). However, because AUs are imperceptible micro-muscle movements, it can be challenging to precisely identify them from facial appearance alone. So these FE systems needs multiple classifiers for each AU and decision level fusion is employed for classification, for which exorbitant computing resources are required.

Conventional FER approach appearance features, geometric features, or a fusion of geometric and appearance features are utilized to pre-train the classifier, i.e. Hidden Markov models [24], AdaBoost [20] and Support Vector Machines [20] [21] [22]. Features based on face geometry are more sensitive to variation in head orientation, scale, size and position, While appearance-based features produce higher performance overall.

In deep-learning based FER approaches facial features are identified and classified by deep-neural networks, unlike conventional machine-learning approaches. Deep learning (DL) based methods retrieve optimum features using deep neural networks with appropriate features can be extracted directly from the data. Additionally, improved results for DNN-based FER techniques have been observed. Though It is difficult to gather a hefty amount of training data, the selection of optimal features and parameters in deep learning remains a challenging issue [21] for facial emotion under diverse conditions. However, deep learning-based systems demand more sophisticated and powerful computational resources than conventional approaches to operate with a variety of parameters throughout the training and inference processes [22].

Therefor real-time DNN processing is a challenging task for the majority of embedded systems, including intelligent and robotic systems. The current work is primarily focused on creating a FER system using SVM classifier that address the challenges of real-time FER and operates on low-

spec devices and still achieve an equivalent FER performance. In Section II, we provide a summary of the relevant research on the identification of facial emotions. Section III explains the proposed pre-processing, landmarks detection, feature extraction and the classifier training framework. Detail analyses of the experimental findings are provided in Section IV. Finally, Section V concludes the paper.

CHAPTER 2:

Literature Review

In human-robot interaction, facial emotions are crucial components (HRI) on top of human communication that helps the robot to understand the emotional state of its human counterpart. Humans express their emotions through numerous channels including head gestures, body posture, facial expressions, and speech. Different studies [23] [24], have found that nonverbal components make up two-thirds of human communication and verbal components only account for one-third. According to Mehrabian [1], a person's facial expression accounts for 55% of the effectiveness of a spoken message. Visual signals constitute the majority of the information that humans receive on a daily basis 75%, which is the fundamental reason why visual modality is so common in social robots, according to [25]. This fact drives the majority of social robots to conduct human-like perception via visual signals.

Due to the numerous applications that have recently arisen, notably in the fields of Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI), extensive research [6] [7] [8] effort has recently been conducted in the field of emotion recognition from facial expression. Recent robot-based systems [9] [10] [26] [27], are not only able to talk, move, or perform video analysis like facial recognition but there is growing interest in employing robots for emotional assessment of humans so that the robotic agent interacts in a more natural and socially acceptable way during human-robot interaction [28].

According to [9] using microphones, cameras, and, sensors, the robot can identify the emotional state of the user and acclimate to it by carrying out a set of predetermined actions that fit the current situation. Several studies [10] [12] [29] [30] [31] [32] have been made using NAO [33] robots due to their design and programming potential. For example, for curriculum's contents are based on students' emotions during teaching [9], Facial Emotion Recognition in Children [10], social interaction with the elderly [12], and children with Autism [13].

There is growing interested to develop FER systems for emotion recognition in real time using

computer-based systems e.g. robots [14] [15] [16] [26] [27] [9] [10] [34] [28] [33]. Some of the most primitive attempts were made in [35] [36]. They proposed a facial action coding system (FACS) for facial emotion recognition as the most immediate, natural, and powerful mean for humans to communicate their emotions and intentions. It was revealed that a single emotion is made up of many different Action Units (AUs) that are present in FACS. The most widely used is the facial action coding system (FACS), which was created by Ekman and Friesen [35] and Ekman et al. [37]. FACS depicts all visual and potentially identifiable facial movements (AUs) with reference to 33 action units. A single AU or group of AUs can be modeled to any facial expression and consequently, resolute that an emotion can be mapped to a facial muscle movement.

The visual/facial modality is the utmost extensively used channel, and state-of-the-art approaches usually utilize 2D facial features. Typically used FER systems fall into two broad categories: static images and image sequencing based processing.

In the first category, only the current frame or the peak emotion frame is utilized to recognize facial emotions [38] [39] [40]. However in the second category the neutral face usually the initial frame is used as a reference and variance between the previous frame face and the succeeding face frame act as a feature vector to recognize facial emotions [41] [42] [43], Several FER systems [43] [44] [45] [46] measured the geometrical deformation of fiducial points among the current frame and the subsequent face frame as temporal characteristics and is used to derive appearance features. The main distinction between the two methods of recognizing facial expressions in 2D static images and image sequences is that in the second category new active features are created frame by frame by shifting the landmarks between the apex emotion frame and the current frame.

The frontal face's shape and spatial elements like the nose, lips, and eyebrows are described by geometry-based features in the traditional FER approach, however appearance-based characteristics determine the texture of the face as a result of the expression. Therefore on the bases of the features used conventional FER can be further classified into geometry-features-based FER and appearance-features-based FER. When using the geometric technique, the decision-making process takes into account the geometric spatial relationship between specific key points (also known as fiducial points) on the face, such as distance, angle, and shape, as well as the

location of important facial features like the eyes, brows, mouth, and nose. [47] [48] [45].

Ghimire and Lee [45] employed position and angle data that were derived by the tracking of 52 fiducial points that were modeled as lines and points. features. The distance and angle features calculated between pairs of fiducial points within the frame are subtracted from the corresponding frame's angle and distance. For the classifier, two methods are presented, either employing an SVM on the boosted feature vectors or employing Multi class AdaBoost with active time warping. A. Saeed et al. [48] and A. Poursaberi et al. [49] employ geometric features as the distance between selected facial landmarks from a single frame for FER.

In [50], for the purpose of recognizing six main facial emotions uses the geometrical movement of a small number carefully of chosen child nodes, is defined as the displacement in node coordinates between the peak facial expression intensity frame and the first neutral frame. However, in the static image base method, the spatial feature is extracted to reflect the shape of face components, such as the Euclidean distance between fiducial points [49], whereas, in an image sequence, the geometry feature largely captures the time related features within an image sequence generated by expressions, such as the movement of facial feature points among the neutral frame and the current frame [45].

Appearance-based features describe the variation in the texture of the expressive face [51] [52] [38]. Typically, the global face region [53] or other face areas are used to extract the appearance features, comprising diverse information [54] [54]. In [49] they divide the face area into 29 local domain-specific regions called local regions. Using an iterative search strategy, domain-specific local areas are found, this decreases the feature size and improves the classification and registration accuracy. Happy et al. [53] used principal component analysis to classify a variety of facial emotions utilizing local binary pattern (LBP) the feature vectors are histograms of various block sizes from an entire face region. Contrary to the technique based on global features, various face regions are given varying amounts of weight. For instance, the chin, face, and forehead convey less information than the lips and eyes.

In addition to frame-based systems, image sequence-based methods [55] also employ appearance cues for the identification of face expression [56] [49].

For feature extraction, a lot of well-known handcrafted features, such as Histogram of Orientation Gradient (HOG) [39], Local Binary Pattern (LBP) and its variants [57] [58] [59], Independent Component Analysis (ICA) [60], Linear Discriminant Analysis (LDA) [61] [49], wavelets [50] [62], distance and angle variation between landmarks are used. A large number of classifiers has been deployed for conventional FER on the extracted features, such as Support Vector Machine (SVMs) [39] [45] [48] [58] [50], Hidden Markov Model (HMM) [60] [62] [63], AdaBoost [56], random forest, Dynamic Bayesian Networks (BN) [64] and Gaussian Mixture Model (GMM) [36].

It was suggested by [65] that a FER system should be composed of three steps: data preprocessing, HoG based feature extraction and template matching using normalized cross-correlation for classification. Experimental results on CK+ datasets for this approach showed an accuracy of 83.6% for five FEs.

Real-time FER was introduced by Suk et al. [44] for usage in a mobile application. This procedure begins with neutral frame by removing the neutral features that support vector machine has identified, along with the mouth status. If the face is identified as having a one of the basic emotions, this technique generates new dynamic features while continuing to update features using the distance between the previous feature and the current feature form as neutral frame. Finally, it returns the dynamic characteristics and the predicted resultant expression as determined by SVM classifiers. On the CK+ dataset, this approach generates experimental results with a 10-fold cross-validation that had an accuracy of 86 %.

Utilizing adaptive neuro-fuzzy inference methods, [66] also implements FE recognition. This method uses FEs to identify face deformations in certain areas, such as the lips, eyebrows, and eyes, and to extract attributes like position, length, and width. The adaptive neuro-fuzzy inference methods are then used with the feature vectors defining the movement of facial expression to identify FEs. The average accuracy of this method for Japanese women's facial expressions was around 90% [67].

Contrary to conventional methods, deep neural networks are used in FER methods based on deep

learning to identify and classify facial features. Deep convolutional neural networks are used in DNNs based approaches to directly extract the best features from the data. Even if it is difficult to gather a significant amount of training data, choosing the best features and parameters for deep learning facial expression under different conditions is still a difficult task [68]. Deep learning-based systems, however, need far more sophisticated and powerful processing capacity than traditional approaches to run training and testing [69]. Furthermore, fiducial points are supplied as inputs to the CNNs, highlighting the significance of facial characteristics over face regions that might not have a significant influence on the generation of FEs.

They train a CNN [70] using the CMU MultiPie database as validation to carry out FER. Performance was increased by the proposed method's use of GPU-based parallelism. In [71] they trained a CNN to carry out FER. A 64 by 64 image is sent to CNN as input. After the convolutional pooling procedure, softmax layers is combined with fully connected layer to generate the output class. According to [21], any FER approach comprises three steps: feature extraction, dimensionality reduction, and classification. They claim that the main challenges with FER are dimensionality and feature selection. Additionally, they claimed that analyzing the entire image required a significant amount of memory and processing power. As a replacement, they suggested geometric features, which have been extracted using facial landmark detection and CNN Classifier. They utilize MMI, MUG, CK, and JAFEE databases to test the efficiency of their system. In this particular paper [72] they presented a brand new DNN architecture for expression recognition. This technique uses a single component framework for its network. Therefore, it classifies registered face images as input into either the six fundamental expressions or the neutral expressions. According to testing results, this approach has an accuracy rate of 77.6% for the MMI and 93.2% for the CK+ database.

In [73], they propose DGNN a directed graph neural network FER using a GCNN with facial landmark features. Delaunay method, was used to built the directed graph's edges and the nodes were identified using landmarks. The fundamental characteristics of faces, such as geometrical and temporal information, are used by graph neural networks to capture basic emotional features. Additionally, they used a steady-stable form of a temporal block in the graph framework to avoid the vanishing gradient issue. For AFEW, MMI, and CK+, respectively, they achieve prediction

accuracies of, 32.64%, 69.4%, and 96.02%. Using landmarks and picture information for the fusion network, their network achieves 98.47% and 50.65% performance for CK+ and AFEW respectively.

The system proposed by [41] recognizes facial expressions using the maximum peak frame that was selected. Their strategy was based on measuring the variance between the neutral and expressive faces. They tested the usefulness of their method using the eNTERFACE database and achieved an accuracy of 78.26%. Another FER system was presented by [74] using the active appearance model's 63 facial landmark points. To determine the final 4 landmark points, they calculated the remaining 63 points. The degree of openness was determined using the height-to-width ratio. By taking the product of weights with the sum of ratios, facial expression was derived. They were able to forecast and classify emotions with an accuracy of 88%.

AU-based techniques find pre-defined AU markers, then use the Facial Action Coding System to decode particular expressions (FACS). Recently, the deep learning approach has been used in conjunction with AU-based techniques. Zhao et al. [75] made 8 by 8 patches out of the aligned face images in order to develop the multi-label learning and deep areas to detect AUs and identify facial emotions. By taking into account the correlations between AUs, this method demonstrated a high AU identification performance; nevertheless, the results were influenced by face alignment, and by equally treating all the blocks could lessen the significance of particular regions. In order to investigate the psychological hypothesis that different facial AUs can be used to dissect facial expressions, Liu et al. [76] created AU-inspired deep networks (AUD). An AUDN is composed of three processes: (1) learning the representation of the micro-action-pattern (MAP) through convolutional and max-pooling layers, (2) integrating correlated MAPs through mid-level semantics produced through feature grouping, and (3) developing smaller networks for higher-level representations through multilayer learning. The performance test was done using average accuracy on seven emotion categories, including neutral, and it had shown 75.85% accuracy for CK+ and 93.7% for the MMI database.

In conclusion, since real-time embedded applications may swiftly pick up new information, conventional feature extraction-based algorithms are well suited for such systems and work well

with even less data. Compared to deep learning-based techniques, conventional FER systems use considerably less memory and processing capacity. Due to their high degree of precision and low computational complexity, these techniques are still experimented and use in real-time embedded applications.

CHAPTER 3:

Methodology

We use influential landmarks that represent the six fundamental facial emotions of disgust, surprise, fear, anger, happy, and sad as outlined by Ekman [14] [15] in order to address the limitations in the existing approaches and to reduce the load of 2D real-time emotion recognition. Instead of using the image light intensity, pixels-based and generic geometric features that take more time to process in order to detect emotion, our proposed FER system uses a face geometry-based spatial feature descriptor between important facial landmarks using the angle relations and distance ratio. For FER SVM based Classification technique is used to recognize facial emotions as described in Figure 1.

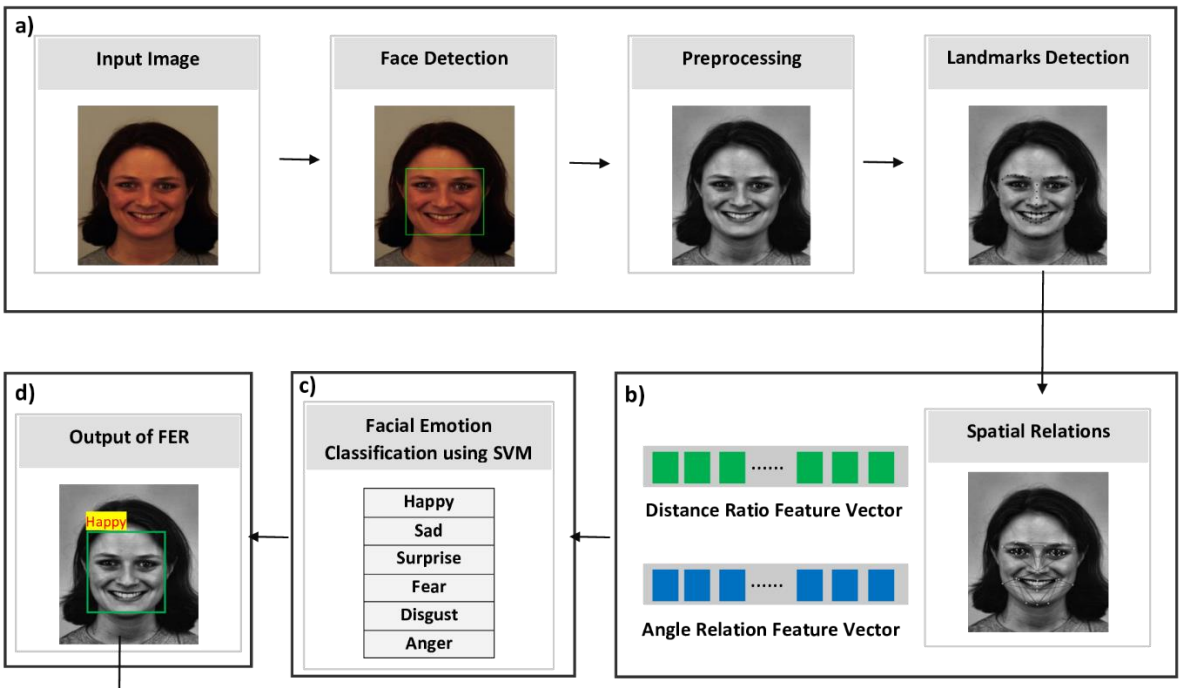


Figure 1. Overview of the proposed FER technique . (a) ROI is detected, CLAHE is applied to the grayscale image and Facial landmarks are detected from the face. (b) spatial facial features angle relations and facial ratios are extracted. (c) Facial Emotion (d) is classified using the SVM classifier.

In addition to cutting down on computing time and system memory, this will also shorten the design of the system. The proposed module has decreased computational complexity as a result (execution time, memory). Fusion of KDEF and CK+ datasets at the training level was also employed in order to generalize the FER system's response to the variations of ethnicity, race, and national and provincial backgrounds. The present work is mainly focused on developing a real-time FER system that is computationally inexpensive and capable of operating in constrained specification devices as compared to systems in earlier works.

3.1 Facial Emotions Datasets

The extended Cohn-Kanade (CK+) dataset [77] and Karolinska Directed Emotional Faces (KDEF) [78] dataset were chosen for training and evaluation of our technique after a thorough and in-depth examination of databases utilized for comparative and comprehensive FER research. Typically, 2D static images or 2D image sequences have been used to study human facial emotions.

3.1.1 Karolinska Directed Emotional Faces (KDEF)

KDEF was created at Stockholm, Sweden's Karolinska Institute's Department of Clinical Neuroscience, Section of Psychology. It consists of 4900 photographs of human faces showing seven various facial emotions (neutral, disgust, surprise, fear, anger, happy, and sad), taken from five different perspectives on 70 people (35 men and 35 women) over the course of two sessions (half right profile, full right profile, straight, full left profile, half left profile). Each face expression image is 562 pixels by 762 pixels in its original size. We use straight-angle facial expression images for our FER system. Figure 2. displays an example of images displaying each of the seven expressions, while Figure 3. lists the frequency of facial expression images taken from the front angle.

3.1.2 Extended Cohn-Kanade (CK+)

The database contains 593 image sequences that depict the various facial expressions made by 123 subjects, starting with a neutral face and ending with an expression that has been FACS-coded.

The majority of the people were female and ranged in age from 18 to 30. Action units and emotional prototypes can both be found in image sequence analysis. Emotion sequences only appear in 327 out of the 593 sequences. It offers guidelines and benchmark outcomes for tracking face features, AUs, and emotion recognition. The original images had 640 x 480 pixel resolutions. Figure 4 displays a sample of images displaying each of the six facial emotions, while Figure 5 lists the frequency of images taken from the front angle.

Figure 2. Sample of KDEF images showing six emotions.

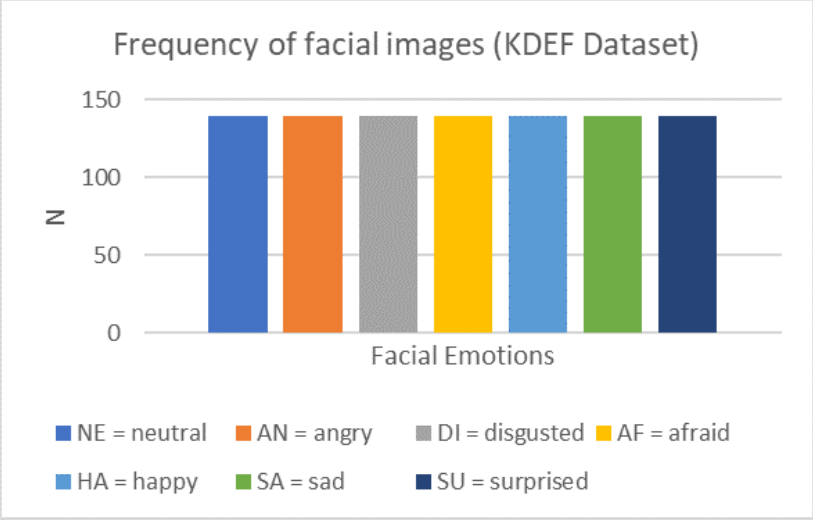


Figure 3. Frequency of Facial images in CK+ dataset

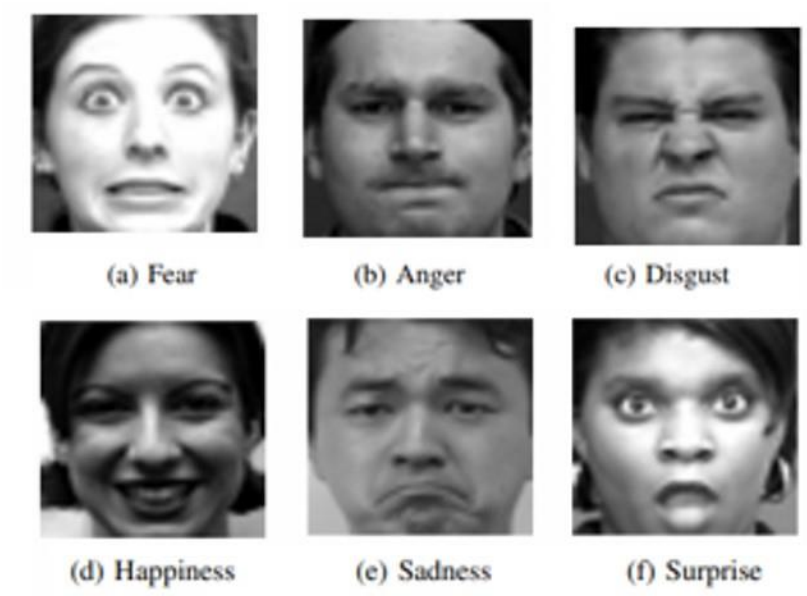


Figure 4. Sample of CK+ images showing six emotions.

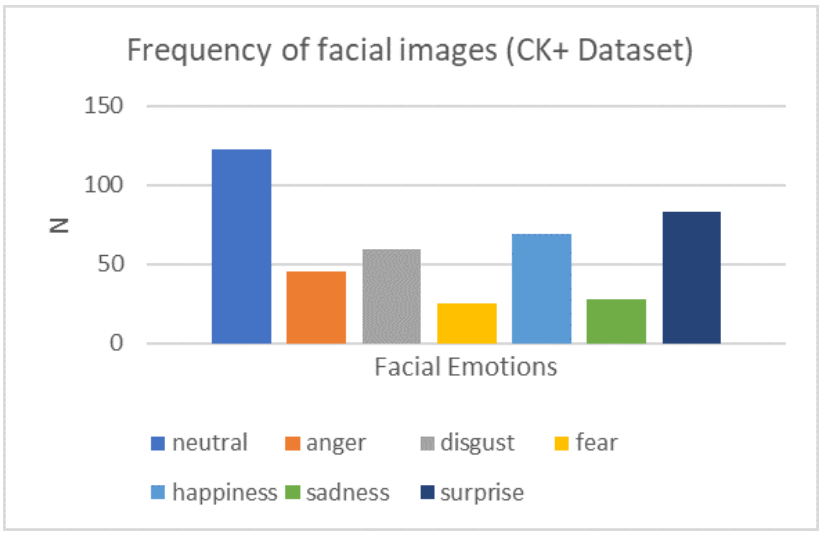


Figure 5. Frequency of Facial images in CK+ dataset

3.2 Image Pre-Processing

Before moving to the processing, the detection of a face in the image frame was one of the key parts of the processing pipeline. we had to detect the face, even though our training data contained only frontal facial expression images. Once the face was detected, it was easier to determine the region of interest(ROI) and extract features from it.

For 2D face detection, several algorithms like Haar-cascades from OpenCV, HOG from Dlib [22], and Multi-Task Cascaded Convolutional Neural Networks (MTCNN) were tried. It was observed that many faces were correctly detected by Haar-cascade and it has the highest frame rate among HOG and MTCN but the biggest weakness was the false-positive detections. Haar cascades lean towards the choice of detectMultiScale parameters that can be tuned to some extent but can't be completely removed. Secondly, the algorithms were tested with varying lighting conditions(very low light and placing a light source behind the person). HOG's output was a little unstable but better than Haar cascade which was able to predict even fewer frames and gave false positives detections as well. MTCNN was unable to detect even a single frame signifying good lighting conditions needed if it is to be used. In conclusion histogram of oriented gradients from Dlib serve us the best for our purpose due to the speed, reliability in varying lighting conditions, and computational efficiency. Figure 6. shows an example of face detection.

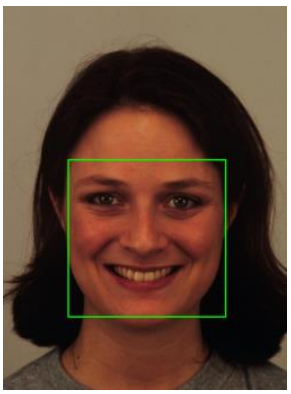


Figure 6. Shows an example of face detection

After face detection, the next step, images are converted to grayscale. The Contrast Limited

Adaptive Histogram Equalization(CLAHE) [79] was applied to ensures that all photos are equalized for similar lighting conditions in the KDEF and CK+ data set using OpenCV [80] built-in createCLAHE() function. Figure 7. shows an example of the grayscale and CLAHE-applied image.

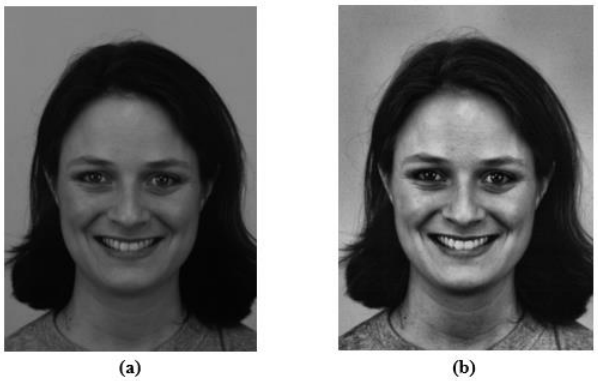


Figure 7. (a) grayscale image (b) CLAHE applied on grayscale image

The benefit of using CLAHE over the normal gradient and generic Histogram Equalization was that they didn't consider the global contrast of the image.

3.3 Facial Landmarks Detection

As soon as the subject face has been detected in the observed scene using Dlib built-in function. It returns a window(x, y, width, height) which is the detected face. Following the face detection, another built-in Dlib function shape_predictor() is utilized to predict the 68 facial landmark points. This function internally utilizes the method described by [81] to achieve better prediction accuracy. The predictor function returns the 68 points at the eyes(left and right), mouth, eyebrows(left and right), nose, and jaw.

The indexes of the 68 coordinates can be visualized on the image below. Figure 8. shows the result of detected landmark points marked with dark dots.

Figure 8. Result of facial landmarks detection using Dlib.

3.4 Selection of Influential Facial Landmarks

An effective feature descriptor should include as many landmarks as feasible to explain the traits that distinguish between different facial expressions. Some landmarks, however, may drastically affect the classifier performance.

Ekman and Friesen developed [35] a comprehensive description facial recognition scheme, which has been regarded as an empirical study for characterizing facial expressions, to notice all potentially perceptible changes that could take place in a face. Figure 9. shows a portion of the FACSAID's facial action unit.

The forty four action units used by FACS to describe facial muscular movement in terms of their position and intensity. Action units combinations or single action unit can be used to model individual expressions. Such a dictionary was introduced for the FACS framework by Friesen and Ekman. Emotion description coding scheme for FACS is given in Table 1. Based on CK+ FACS coding scheme.

Figure 9. Facial Action Coding System scheme (FACS).

Table 1. Facial emotion criteria with reference to actions units.

Emotions	**Dictionary Criteria**
Anger	AU 23 and AU 24 must be present
Disgust	Either AU 9 or AU 10 is present
Fear	Combination of AU 1 + AU 2 + AU 4 must be present.
Happy	AU 12 must present.
Sad	Either combination of AU 1 + AU 4 + AU 15 or AU 11 must be present
Surprise	Either combination of AU 1+AU 2 or AU 5

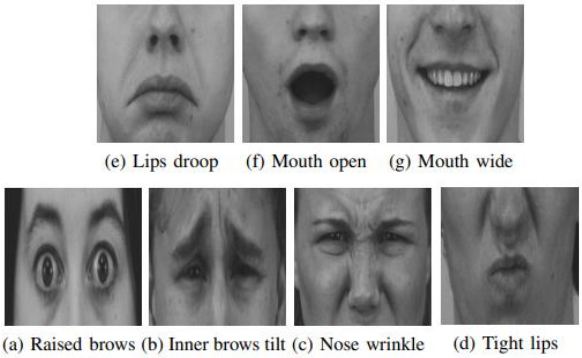


Figure 10. Sample of KDEF images showing marked visible changes for different emotions.

Traditionally, an accurate feature descriptor uses as many landmarks as feasible to distinguish between different facial expressions (position and orientation of 68 facial landmarks relative to each other or relative to mean point) [39]. But, few landmarks decrease the FER system's performance.

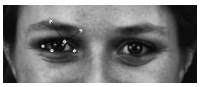
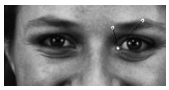
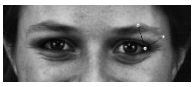
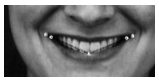
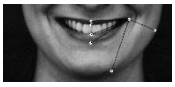
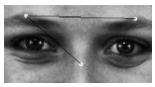
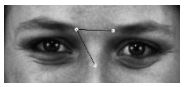
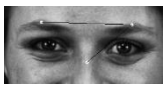
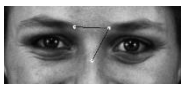
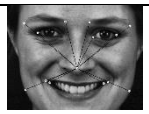
As a result, we describe the spatial relationships between important landmarks around the mouth, nose, eye, eyebrow, and chin regions prin Table 2. for FEs, to create discriminative feature vectors.

3.5 Feature Extraction

Extraction of feature vectors that describes human emotion was a vital part of the FER system. To make it possible real-time facial expression recognition with scant and constrained resources, We used spatial relations-based features which require minimal computational time than image light intensity, pixels-based and generic displacement based features.

As shown in Figure 11. The suggested approach is able to cut processing costs and enhance accuracy by only considering a one half of the facial landmarks rather than all of the landmarks.

Table 2. Manually selected landmarks and spatial relations between fiducial points.

left eye						
Right eye						
Inner lips						
Outer lips						
Left eye brow						
Right eye brow						
Nose						
Chin						

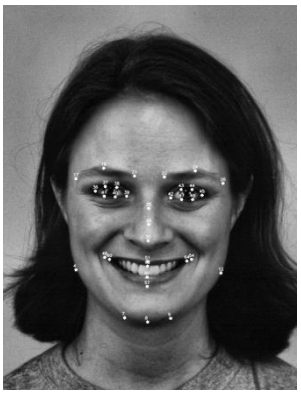


Figure 11. Selected influential landmarks.

Traditionally, an accurate feature descriptor uses as many landmarks as feasible to distinguish

between different facial expressions. In [39] describe the face components and the shape of the relations between face components. Scaling and face orientation can change spatial relations. To address the problem of scaling and face orientation we use angle relations and distance ratios between different placements of fiducial points that are adaptable to face orientation and scaling.

To make it easier to access and manage the fiducial points, we first convert them into an array of x and y coordinates that reflect their locations using NumPy [82].

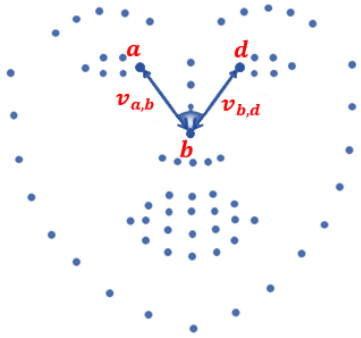


Figure 12. Example of spatial relation between fiducial landmarks {a ,b ,d }.

Secondly, the distance ratio feature of the three landmarks $\{a, b, d\}$ is extracted, as shown in Figure 12. we define two individual vectors for the pairs {a, b} and {b, d} as $v_{a,b}$and $v_{b,d}$. To complement the changes the as a result of face rotation or scaling displacement ratio is calculated using two vector.

$$Dist_{ratio} = v_{a,b}/v_{b,d} \qquad (1)$$

After that The angle relation feature of the three landmarks $\{a, b, d\}$ is then calculated as shown in Figure 12. The angle relation between landmarks is modeled as angle feature.

$$Ang_{rela} = \theta(v_{a,b}/v_{b,d}) \qquad (2)$$

where the vectors $v_{a,b}$and $v_{b,d}$, respectively, point from fiducial landmark point a to b and then from fiducial landmark b to landmark d.

The angle relation and distance ratio can resist changes brought on either rotating or scaling the face. The vectors in Figure 13. are drawn between the markers $a, b,$ and d. The facial marker points $a, d,$ and the central landmark b are connected by a line. As a result, every line drawn is a vector with both magnitude and direction.

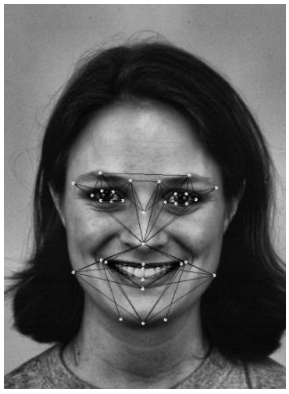


Figure 13. Influential Landmark base spatial feature descriptor.

We extract 60-dimensional angle relations and 60-dimensional distance ratios. So far feature vectors can be generalized as

$$FV_{angl} =< Angle_{rela1},\ Angle_{rela2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ Angle_{rela59},\ Angle_{rela60} > \quad (3)$$

$$FV_{dist} =< Dist_{ratio1},\ Dist_{ratio2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ Dist_{ratio59},\ Angle_{ratio60} > \quad (4)$$

Fusion of $Angle_{rela}$ and $Dist_{ratio}$ feature vector can be generalized as

$$FV =< Dist_{ratio1},\ Angle_{rela1}, Dist_{ratio2},\ Angle_{rela2} \ldots\ldots\ldots Dist_{ratio60},\ Angle_{rela60} > \quad (5)$$

These features are given as input to trainthe SVM classifier.

3.6 Maximum Dissimilarity-based Apex Frame Selection for Real-Time FER

The selection of the apex emotion frame is the most crucial step in real-time FER. The idea behind choosing apex frames is based on hypothesis they should differ from other frames in the sequence the most. First, the distance among selected fiducial points in two consecutive frames compared

to determine the variance between the frames(i.e. $f_i \; and \; f_{i+1} \; or \; f_{i+1} \; and \; f_i$). The algorithm then ranks the frames according to how different they are from other frames in absolute terms, and then chooses the peak frames (or the ones with the K greatest average dissimilarity scores) from that group. To test the effectiveness of our suggested FER approach, we only use the apex expression frames from each sequence.

3.7 Support Vector Machine (SVM)

SVMs are powerful, adaptable supervised learning methods used for regression and classification, due to their effectivity in high dimensional space even They are also memory efficient since they use a smaller sub - set of training data points (known as support vectors) in the decision function when the number of features exceeds the number of samples. SVMs facilitate the use of custom kernels instead of common kernels for decision-making. The only drawback of using SVMs that they calculate the probability estimates using five-fold cross-validation, They can not do it directly. In the case of large data sets high training time is required which might not be suitable. In our setting, since we just have two small data sets (KDEF and CK+), training and classification may be carried out with a high degree of accuracy utilizing only an SVM. SVM and SVC are interchangeably used in literature.

The SVC was employed by using the Scikit-learn library [83] class known as "SVC" is present in the "SVM" module. To provide flexibility, the data items from both datasets were fused, split into training and testing groups at random in an 80:20 ratio. Prior to the training phase, face detection is applied to all input images, and then resized and converted to a grayscale image after that CLAHE was applied on it succeeding which the facial landmarks can be detected and Finally feature vectors were calculated for the fused dataset.

The optimal parameter tuning and selection ($like \; C \; and \; \gamma$) have been performed using the grid search strategy [84] [85]. Where γ optimize the decision boundary and C is the misclassification penalty. If these parameters are not adjusted, they degrade the accuracy of binary and multiclass classifier.

CHAPTER 4:

Experimental Results and Discussion

Using an Intel Core i7 CPU and 8 GB of RAM running Microsoft Windows 10, we carried out all of the experiments to determine effectiveness of the presented FER framework. In addition, SVM training was based on the CPU using two databases:

1- Karolinska Directed Emotional Faces (KDEF) [78] database
2- Extended Cohn-Kanade (CK+) database [77]
3- Fusion of KDEF and CK+ databases

Diverse length of features were applied as SVM input. Eighty percent of the photos from the CK+, KDEF, and fused databases were utilised for training, while the remaining twenty percent were used for testing. Six basic emotions available in each database were considered. Training and testing data were made to be mutually exclusive in order to prevent overfitting. The optimal parameter tuning and selection (like C and γ) have been performed using the grid search strategy [84] [85]. Where γ optimize the decision boundary and C is the misclassification penalty. If these parameters are not tuned, they degrade the accuracy of binary and multiclass classifier.

In experiments on CK+ and KDEF databases, after pre-processing input images, and face detection, influential landmarks base spatial relation features are extracted. The extracted feature vectors are given as input to trainthe SVM classifier.

4.1 Performance Evaluation of proposed FER

We gauged the performance of the presented FER system against six cutting-edge techniques that either employ DNNs or traditional machine learning-based algorithms to confirm its efficacy.

Comparing the classification performance of the presented framework and cutting-edge techniques

using the same dataset is shown in Table 3. (i.e Cohn-Kanade database). The Table clearly shows that the proposed framework produced outcomes on par with cutting-edge techniques.
The presented framework is equivalent compared to any other cutting-edge technique in terms of expression recognition accuracy, as seen in Table 3. The feature extraction technique described in [86] is identical to our technique.

Table 3. Assessment of suggested FER system with state-of-the-art methods for most commonly used CK+ dataset.

Comparison Methods	**No. of Classes**	**(%) Accuracy**
SVM (Gabor features) [87]	7	93.3
SVM (LBP+VLBP) [55]	6	96.26
SVM (Euclidean space) [88]	6	94.5
SVM(PLBP) [54]	6	96.7
NN(three-layer) [89]	6	93.8
SVM (appearance base feature) [90]	6	92.3
SVM (weighted entropy, brightness,local binary pattern) [91]	6	94.9
SVM (Coordinate, Distance base geometrical features) [39]	7	89.0
Inception-ResNet and LSTM [92]	6	92.6
Real-time mobile FER [44]	6	85.5
Hierarchical WRF + Data.Sim [86]	6	92.6
Ours	**6**	**96.8**

With this setup, our system improved recognition accuracy while using a comparatively small feature vector. The proposed FER system attained average recognition rate accuracy of 96.8%, 84.2% and 84.1% for CK+, KDEF and fusion of CK+ and KDEF databases respectively.

As a result of the need for a lightweight algorithm that can run on CPUs rather than expensive GPUs in order to execute in real time, Deep neural network-based methods are not appropriate for

low-specification devices like robots.

4.2 Experiment using Extended Cohn-Kanade (CK+) database

For this experiment we utilized all FACS coded images from CK+ database. The training and testing of classifier was done on the apex front facial frame.

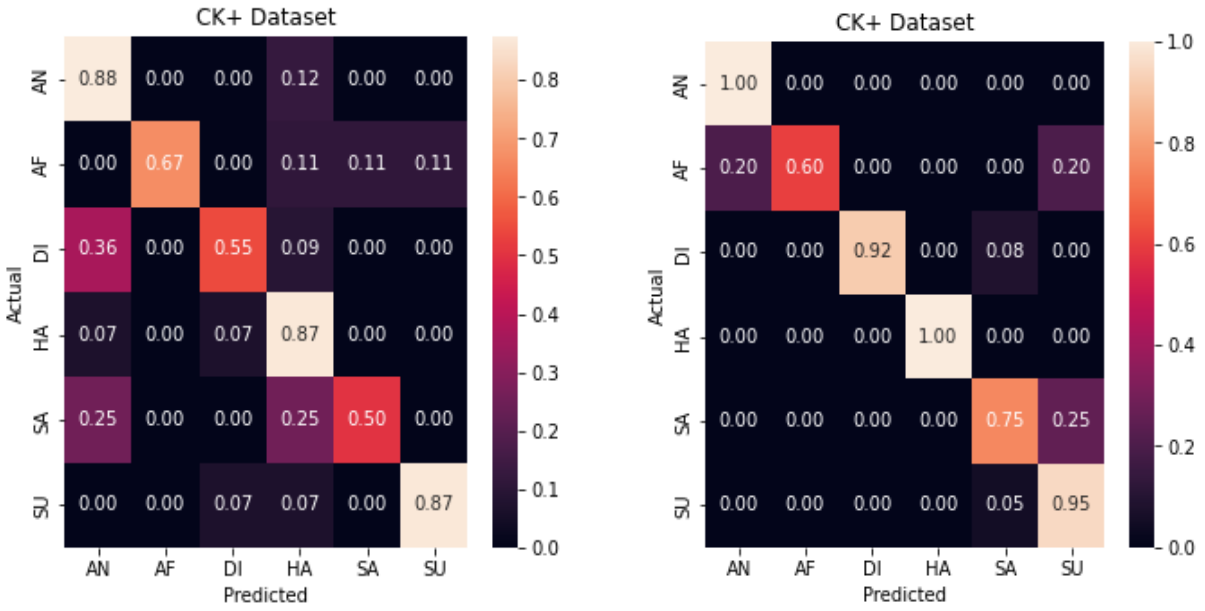
The experiment have been conducted in three parts. In the first part of the experiment the FER was train with generic geometric features such as $< x\ y\ coordinates, mag, ang >$. Where magnitude (mag) is the Euclidean distance and angle (ang) is the direction of line between central point (x_{mean}, y_{mean}) and all 68 facial landmarks of a facial image. In this part of experiment, the performance of the proposed FER was evaluated with the optimal parameters obtained from grid search i.e. "Support vector machine (SVM)" with rbf kernel, $C = 100$ and $\gamma = 0.001$.

In second part of experiment, we excluded x, y coordinates from features vector during classifier training and performance of FER system was evaluated with optimal parameters obtained from grid search using Support vector machine with linear kernel, C=1 and γ=0.001.

In the last part we train the classifier with influential spatial relation based landmarks features $< Dist_{ratio}, Ang_{rela} >$. Performance of system was evaluated again with optimal perimeters obtain from grid search using Support vector machine with linear kernel, C=0.01 and γ=0.001.

We constructed confusion matrices for the CK+ for different number of features, as shown in Figure 14. and Figure 15. respectively, to test the effectiveness of proposed technique whether it distinguishes each of the six facial expressions.

Where magnitude(mag) is the Euclidean distance and angle(ang) is the direction of line between central point (x_mean,y_mean) all 68 facial landmarks.



(a) using features <x, y, ang, mag> **(b)** using features <ang, mag>

Figure 14. Confusion matrix of FER using generic geometric feature for CK+ dataset

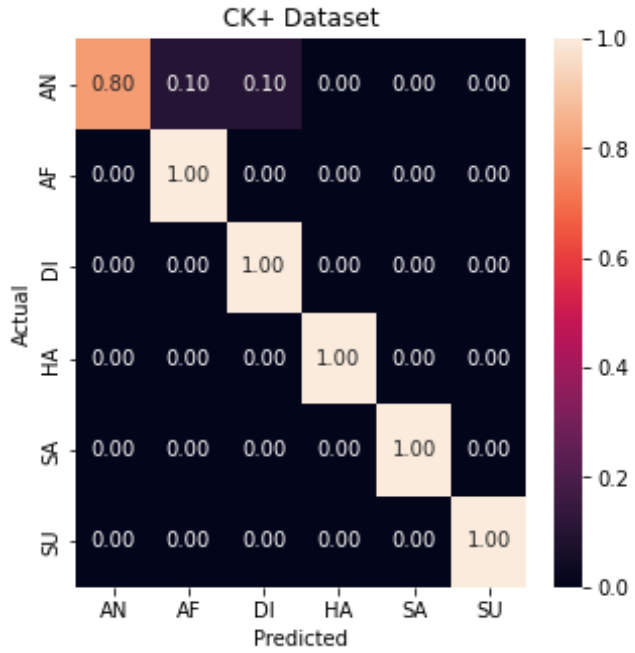


Figure 15. Confusion matrix of FER using influential landmark base spatial relation feature for CK+ dataset

As the results in Table 4. suggest that our proposed FER system is optimized for real-time embedded applications with constrained specifications and yields an accuracy of 96.8%, which is higher than generic geometric base feature i.e. 88.4% and 91.9% for CK+ dataset.

Table 4. Processing time of proposed FER. for CK+ dataset. F.D. (Face Detection), F.E. (Feature Extractor)

Features	F.D. & F.E.	Training Time	Total
$xy\ coordinates, mag, ang$	566.82 ms	2.593 ms	569.41 ms
ang, mag	543.26 ms	1.695 ms	544.96 ms
$Dist_{ratio}, Ang_{rela}$ **(Ours)**	509.56 ms	1.895 ms	511.46 ms

The main objective of this work is to show that proposed influential landmark base spatial relation representation outperforms generic feature based representation with constrained specification so We simply explored with geometric features for both representations and did not study the effectiveness of additional appearance-based characteristics. in literature as they require high processing costs and exorbitant memory requirements.

4.3 Experiment using Karolinska Directed Emotional Faces (KDEF) database

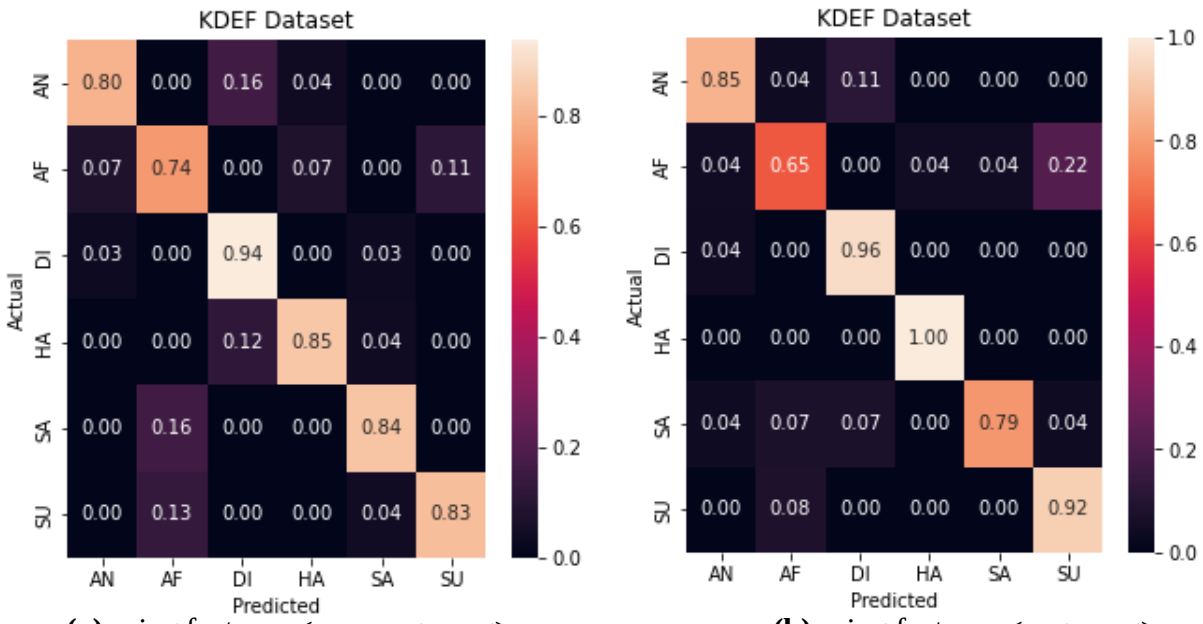
For our FER system, we utilize front angle facial expression images from KDEF database. the experiment on KDEF were conducted in three parts.

In first part of the experiment the FER was train with generic geometric features $< x, y\ coordinates, mag, ang >$. For this part of experiment, the performance of the proposed FER was evaluated with the optimal parameters obtained from grid search using Support vector machine with linear kernel, C=0.1 and γ=0.001.

In the next part of experiment x, y coordinates are exclude from features vector during classifier training and performance of FER framework was evaluated with optimal parameters obtained from grid search using Support vector machine with linear kernel, C=0.1 and γ=0.001.

In the final experiment Influential landmarks base spatial relation features $< Dist_{ratio}, Ang_{rela} >$ are utilized in third part of experiment to train the classifier. Performance of proposed FER framework was evaluated with optimal perimeters obtain from grid search using Support vector machine, C=10, γ=0.001 and with rbf kernel,

Figure 16. and Figure 17. respectively shows the constructed confusion matrices for the KDEF database, to assess the effectiveness of proposed technique whether it distinguishes each of the six FEs.



(a) using features <x, y, ang, mag> **(b)** using features <ang, mag>

Figure 16. Confusion matrix of FER using influential landmark base spatial relation feature for KDEF dataset

Where magnitude(mag) is the Euclidean distance and angle(ang) is the direction of line between central point $(\boldsymbol{x_{mean}}, \boldsymbol{y_{mean}})$ all 68 facial landmarks.

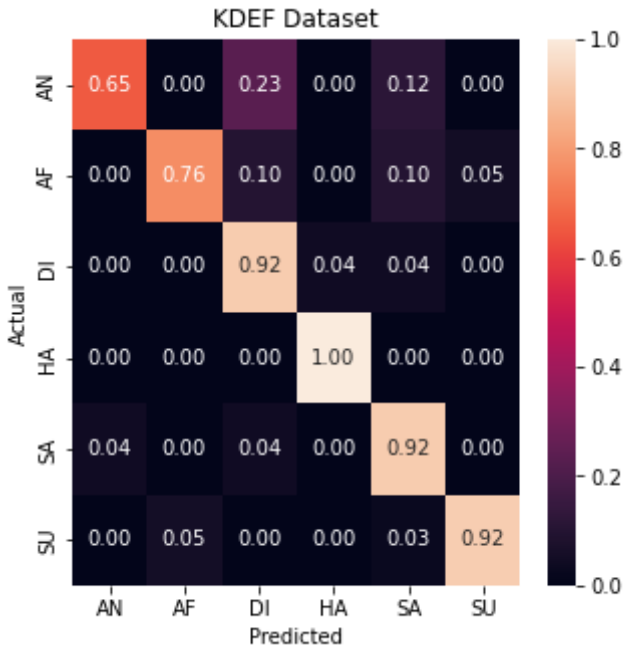


Figure 17. Confusion matrix of FER using influential landmark base spatial relation features for KDEF dataset

As shown in Table 5. the suggested FER system is optimized for real-time embedded applications with constrained specifications and yields an accuracy of 86.7%, which is slightly lower than generic geometric feature (i.e. xy coordinate, ang and mag features) 88.7% and relatively higher than (i.e. ang and mag features) 86.1% for KDEF dataset.

Table 5. Processing time of proposed FER for KDEF dataset. F.D. (Face Detection), F.E. (Feature Extractor)

Features	F.D. & F.E.	Training Time	Total
$xy\ coordinates, mag, ang$	1318.68 ms	8.774 ms	1327.45 ms
ang, mag	1241.07 ms	4.887 ms	1245.96 ms
$Dist_{ratio}, Ang_{rela}$ **(Ours)**	1082.00 ms	4.688 ms	1086.69 ms

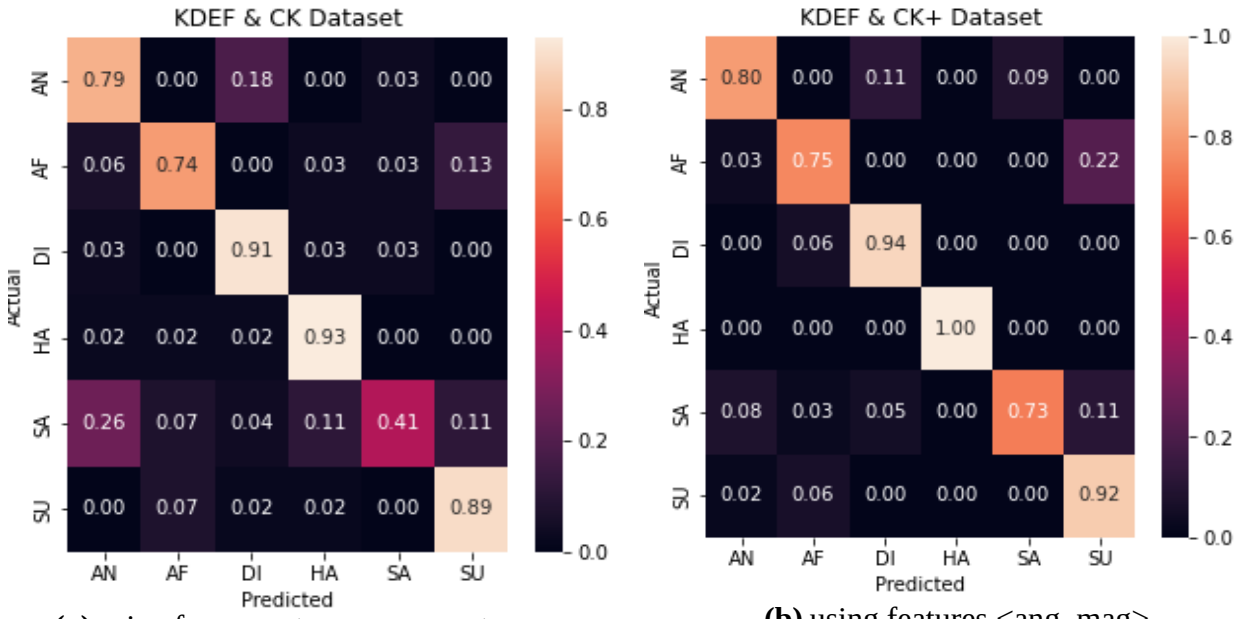
4.4 Experiment on the Fusion of KDEF & CK+ database.

Fusion of KDEF and CK+ datasets at the training level were also employed in order to generalize the FER system's response to the variations of ethnicity, race, national and provincial

backgrounds.

Experiments on fused KDEF and CK+ dataset are performed in three phases. The FER was trained with generic geometric features $< xy\ coordinates, mag, ang >$ in the first phase of the experiment. For this phase of the experiment, the performance of the framework was assessed using Support vector machine with a linear kernel, C=0.1, and γ =0.001 optimal parameters found by grid search.

In the second experiment, xy coordinates are removed from the features vector during classifier training. The performance of the FER framework was then assessed using the best grid search parameters, i.e., for "Support vector machine (SVM)" with a linear kernel, C=0.1, and γ =0.001. To determine whether the suggested technique is effective at differentiating each of the six FEs, the confusion matrices for the fused KDEF & CK+ database are shown in Figures 18. and Figure 19. respectively.



(a) using features <x, y, ang, mag>

(b) using features <ang, mag>

Figure 18. Confusion matrix of FER using generic geometric feature for fused KDEF & CK+ dataset

Where magnitude(mag) is the Euclidean distance and angle(ang) is the direction of line between central point $(\boldsymbol{x_{mean}}, \boldsymbol{y_{mean}})$ all 68 facial landmarks.

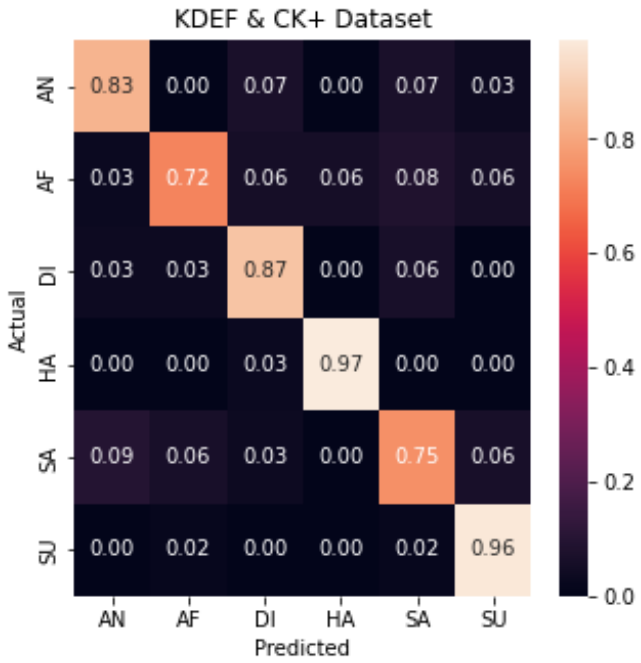


Figure 19. Confusion matrix of FER using influential landmark base spatial relation features for fused KDEF & CK+ dataset

The accuracy of the suggested FER system, which is 86.4%, is slightly lower than that of the generic geometric feature (i.e., xy coordinate, ang, and mag features), which is 88.7%, and relatively higher than that of the (i.e., ang, and mag features) for the KDEF dataset, which is 85.9%, as shown in Table 6. The suggested FER system is optimized for real-time embedded applications with constrained specifications.

Table 6. Processing time of proposed FER for fused KDEF & CK+ dataset. F.D. (Face Detection), F.E. (Feature Extractor)

Features	F.D. & F.E.	Training Time	Total
$xy\ coordinates, mag, ang$	1885.50 ms	14.067 ms	1899.57 ms
ang, mag	1784.33 ms	12.467 ms	1796.80 ms
$Dist_{ratio}, Ang_{rela}$ (**Ours**)	1591.57 ms	8.576 ms	1600.15 ms

The experimental result showed that our FER system provided a performance of 96.8%, 86.7% and 86.4% for CK+, KDEF and fusion of KDEF& CK+ databases respectively, outperforming not only a comparative landmark-based method but also most of the static 2D image and image-

sequence based state-of-the-art DNNs and traditional Machine learning based methods. This result shows that the landmark features can be considered an effective modality to analyze facial emotional information.

As per Section 4.3, we concluded that the proposed model works better on FACS coded image. Thus proposed frameworks prove the validity of influential landmarks base spatial relation features.

Conclusions

In this study, we present a computationally inexpensive, fast FER approach for the perception of six basic emotions that uses influential landmarks based spatial relation features and is capable of running on embedded devices with constrained specifications i.e., robots, smart devices and vehicles. The proposed face representation incorporate movement information makes it more accurate and robust than other appearance and generic geometric feature based face representation. With influential spatial face regions providing the most astute information for facial emotion classification, performance improvement and dimensionality reduction were achieved. In order to confirm the efficacy of the suggested FER technique for six-class facial expressions, several tests with various feature lengths for the CK+ and KDEF dataset were carried out. Fusion of KDEF and CK+ datasets at the training level generalize the FER system's response to the variations of ethnicity, race, national and provincial backgrounds. We compared the proposed influential landmark based spatial relation representation technique with generic geometric based representation. The findings of the experiments demonstrated that the influential landmark-based spatial relation representation outperforms the generic geometric-based feature representation by a significant margin. We produced a comparable and occasionally superior result of FER 96.8% utilizing the suggested technique on the CK+ dataset as compared to the other state-of-the-art techniques in the literature, despite the fact that the paper's main objective was not to compete with other works of literature in terms of accuracy. Which, as far as we can tell, is superior to the results that have been reported in the literature. The experimental results on CK+, KDEF and fusion of CK+ and KDEF databases show that our proposed FER system is optimized for real-time embedded applications with constrained specifications.

Future Work

I believe that, there is a room for the refinement of proposed FER by developing such datasets that help FER system's to generalize the response due to variations of age, ethnicity, race, national and provincial background of subject and looking for more significant facial features within proposed framework. As a part of our future research objectives, the developed system will make a robotic agent capable of perceiving emotion and interacting naturally without the need for additional hardware during HRI.

APPENDIX A

User Agreement for the use of CK and CK+ Dataset

CK and CK+ DATABASE USER AGREEMENT

All requests must be made by a faculty member at a university or college.

CK+ may be used for non-commercial research that is not subject to US export controls. To obtain a copy of the database, please complete the following agreement and return it to Megan Ritter mer160@pitt.edu.

Once the signed agreement is received and approved, you will receive instructions to download the database via Box hosted at the University of Pittsburgh. The database remains the property of Dr. Jeffrey Cohn. Use is subject to the following terms. For questions, please contact Megan Ritter at the address above.

By signing this agreement, you agree:

- To cite the following publications in any paper of yours or your collaborators that makes any use of the database.
    - Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00),* Grenoble, France, 46-53.
    - Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010)*, San Francisco, USA, 94-101.
- To use the images for non-commercial research purposes only.
- Not to provide any portion of the database to other parties.
- In any publications, print, electronic, or other media to use images from only the following subjects and to include notice of copyright (©Jeffrey Cohn):
    - S52, S55, S74, S106, S111, S113, S121, S124, S125, S130, S132

All requests must include the following information.

Faculty member's name: **Dr. Sara Ali**

Faculty member's official title: (e.g., Assistant professor) **Assistant Professor**

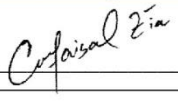
Faculty member's university email address: **sarababer@smme.nust.edu.pk**

Faculty member's signature: 

Name of any additional requestor: **Muhammad Faisal Zia**

Other requestor's official title (e.g., student or postdoc): **Masters's Research Student (of above mentioned faculty)**

Other requestor's university email address: **mzia.rime19smme@student.nust.edu.pk**

Other requestor's signature: 

REFERENCES

[1] A. Mehrabian and J. A. Russell, "An Approach to Environmental Psychology", Cambridge, MA, US: The MIT Press, p. 266," 1974.

[2] F. Alonso-Martin, M. Malfaz, J. Sequeira, J. F. Gorostiza and M. A. Salichs, "A multimodal emotion detection system during human-robot interaction," *Sensors,* vol. 13, p. 15549–15581, 2013.

[3] F. &. Sullivan, *Trend Opportunity Profile: Human-robot Collaboration,* p. 72.

[4] M. I. Ahmad, "Mubin O Orlando J A systematic review of adaptivity in human-robot interaction Multimodal Technol," in *Interact*, 2017.

[5] B. Alenljung and J. Lindblom, "User experience of socially interactive robots: its role and relevance," in *Handbook of Research on Synthesizing Human Emotion in Intelligent Systems and Robotics*, IGI Global, 2015, p. 352–364.

[6] K. Schaaff and T. Schultz, "Towards an EEG-based emotion recognizer for humanoid robots," in *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 2009.

[7] T. Kollar, S. Tellex, D. Roy and N. Roy, "Toward understanding natural language directions," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010.

[8] S. S. Ge, H. A. Samani, Y. H. J. Ong and C. C. Hang, "Active affective facial analysis for human-robot interaction," in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 2008.

[9] I. A. Valagkouti, C. Troussas, A. Krouska, M. Feidakis and C. Sgouropoulou, "Emotion Recognition in Human–Robot Interaction Using the NAO Robot," *Computers,* vol. 11, p. 72, 2022.

[10] A. Lopez-Rincon, "Emotion recognition using facial expressions in children using the NAO Robot," in *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, 2019.

[11] A. Nordahl-Hansen, A. Dechsling, S. Sütterlin, L. Børtveit, D. Zhang, R. A. Øien and P. B. Marschik, "An overview of virtual reality interventions for two neurodevelopmental disorders: intellectual disabilities and autism," in *International Conference on Human-Computer Interaction*, 2020.

[12] D. L. Recio, L. M. Segura, E. M. Segura and A. Waern, "The NAO models for the elderly," in *2013 8th ACM/IEEE international conference on human-robot interaction (HRI)*, 2013.

[13] M. A. Miskam, S. Shamsuddin, M. R. A. Samat, H. Yussof, H. A. Ainudin and A. R. Omar, "Humanoid robot NAO as a teaching tool of emotion recognition for children with autism using the Android app," in *2014 International Symposium on Micro-NanoMechatronics and Human Science (MHS)*, 2014.

[14] D. G. R. Kola and S. K. Samayamantula, "A novel approach for facial expression recognition using local binary pattern with adaptive window," *Multimedia Tools and Applications,* vol. 80, p. 2243–2262, 2021.

[15] A. Agrawal and N. Mittal, "Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy," *The Visual Computer,* vol. 36, pp. 405-412, 2020.

[16] H. Dino, M. B. Abdulrazzaq, S. R. Zeebaree, A. B. Sallow, R. R. Zebari, H. M. Shukur and L. M. Haji, "Facial Expression Recognition based on Hybrid Feature Extraction Techniques with Different Classifiers," *TEST Engineering & Management,* vol. 83, p. 22319–22329, 2020.

[17] A. Hassouneh, A. M. Mutawa and M. Murugappan, "Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods," *Informatics in Medicine Unlocked,* vol. 20, p. 100372, 2020.

[18] A. Savran, B. Sankur and M. T. Bilge, "Regression-based intensity estimation of facial action units," *Image and Vision Computing,* vol. 30, p. 774–784, 2012.

[19] M. V. Mishra, S. B. Ray and N. Srinivasan, "Cross-cultural emotion recognition and evaluation of Radboud faces database with an Indian sample," *PLoS One,* vol. 13, p. e0203959, 2018.

[20] Y. Wang, H. Ai, B. Wu and C. Huang, "Real time facial expression recognition with adaboost," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2004.

[21] N. P. Gopalan, S. Bellamkonda and V. Saran Chaitanya, "Facial Expression Recognition Using Geometric Landmark Points and Convolutional Neural Networks," in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2018.

[22] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, p. 1755–1758, 2009.

[23] K. Kaulard, D. W. Cunningham, H. H. Bülthoff and C. Wallraven, "The MPI facial expression database—a validated database of emotional and conversational facial expressions," *PloS one*, vol. 7, p. e32321, 2012.

[24] A. Mehrabian, "Communication without words," 1968.

[25] "Digital image processing.," 1996.

[26] S. R. Reyes, K. M. Depano, A. M. A. Velasco, J. C. T. Kwong and C. M. Oppus, "Face detection and recognition of the seven emotions via facial expression: Integration of machine learning algorithm into the nao robot," in *2020 5th International Conference on Control and Robotics Engineering (ICCRE)*, 2020.

[27] C. Filippini, D. Perpetuini, D. Cardone and A. Merla, "Improving Human-Robot Interaction by Enhancing NAO Robot Awareness of Human Facial Expression," *Sensors*, vol. 21, p. 6438, 2021.

[28] H. Banaeian and I. Gilanlioglu, "Influence of the NAO robot as a teaching assistant on university students' vocabulary learning and attitudes," *Australasian Journal of Educational Technology*, vol. 37, p. 71–87, 2021.

[29] T. She and F. Ren, "Enhance the Language Ability of Humanoid Robot NAO through Deep Learning to Interact with Autistic Children," *Electronics*, vol. 10, 2021.

[30] L. Ismail, S. Shamsuddin, H. Yussof, H. Hashim, S. Bahari, A. Jaafar and I. Zahari, "Face detection technique of Humanoid Robot NAO for application in robotic assistive

therapy," in *2011 IEEE International Conference on Control System, Computing and Engineering*, 2011.

[31] E. Torta, F. Werner, D. O. Johnson, J. F. Juola, R. H. Cuijpers, M. Bazzani, J. Oberzaucher, J. Lemberger, H. Lewy and J. Bregman, "Evaluation of a small socially-assistive humanoid robot in intelligent homes for the care of the elderly," *Journal of Intelligent & Robotic Systems,* vol. 76, p. 57–71, 2014.

[32] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola and F. Shic, "Improving social skills in children with ASD using a long-term, in-home social robot," *Science Robotics,* vol. 3, p. eaat7544, 2018.

[33] "NAO Robot.Available online:".

[34] D. O. Melinte and L. Vladareanu, "Facial expressions recognition for human-robot interaction using deep convolutional neural networks with rectified adam optimizer," *Sensors*, vol. 20, p. 2393, 2020.

[35] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior,* 1978.

[36] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 21, p. 974–989, 1999.

[37] P. Ekman, W. V. Friesen and J. C. Hager, "Facial Action Coding System (FACS): The Manual & The Investigator's Guide. A Human Face, Salt Lake City, UT: Research Nexus," 2002.

[38] D. Ghimire, S. Jeong, J. Lee and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimedia Tools and Applications,* vol. 76, p. 7803–7821, March 2016.

[39] S. Rohith Raj, D. Pratiba and P. Ramakanth Kumar, "Facial Expression Recognition using Facial Landmarks: A novel approach," 2020.

[40] F. Khan, "Facial expression recognition using facial landmark detection and feature extraction via neural networks," *arXiv preprint arXiv:1812.04510,* 2018.

[41] S. Zhalehpour, Z. Akhtar and C. Eroglu Erdem, "Multimodal emotion recognition based on peak frame selection from video," *Signal, Image and Video Processing*, vol. 10, p. 827–834, 2016.

[42] S. H. Lee, W. J. Baddar and Y. M. Ro, "Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos," *Pattern Recognition*, vol. 54, p. 52–67, 2016.

[43] C. Fabian Benitez-Quiroz, R. Srinivasan and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[44] M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system-a case study," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014.

[45] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, p. 7714–7734, 2013.

[46] "11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015," 2015.

[47] O. Çeliktutan, S. Ulukaya and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP Journal on Image and Video Processing*, vol. 2013, p. 1–27, 2013.

[48] A. Saeed, A. Al-Hamadi, R. Niese and M. Elzobi, "Frame-based facial expression recognition using geometrical features," *Advances in human-computer interaction*, vol. 2014, 2014.

[49] A. Poursaberi, H. Ahmadi, S. N. Yanushkevich and M. L. Gavrilova, "Gauss–Laguerre wavelet textural feature fusion with geometrical information for facial expression identification," *EURASIP Journal on Image and Video Processing*, vol. 2012, pp. 1-13, 2012.

[50] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE transactions on image processing*, vol. 16, p. 172–187, 2006.

[51] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. De la Torre and A. Rossi, "Automated facial expression recognition system," in *43rd annual 2009 international Carnahan conference on security technology*, 2009.

[52] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition*, vol. 36, p. 259–275, 2003.

[53] S. L. Happy, A. George and A. Routray, "A real time facial expression classification system using local binary patterns," in *2012 4th International conference on intelligent human computer interaction (IHCI)*, 2012.

[54] R. A. Khan, A. Meyer, H. Konik and S. Bouakaz, "Framework for reliable, real-time facial expression recognition for low resolution images," *Pattern Recognition Letters*, vol. 34, p. 1159–1168, 2013.

[55] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, p. 915–928, 2007.

[56] D. Ghimire and J. Lee, "Extreme learning machine ensemble using bagging for facial expression recognition," *Journal of Information Processing Systems*, vol. 10, p. 443–458, 2014.

[57] A. C. Cruz, B. Bhanu and N. S. Thakoor, "Vision and attention theory based sampling for continuous facial emotion recognition," *IEEE Transactions on Affective Computing*, vol. 5, p. 418–431, 2014.

[58] C. Shan, S. Gong and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, p. 803–816, 2009.

[59] G. Zhao, X. Huang, M. Taini, S. Z. Li and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and vision computing*, vol. 29, p. 607–619, 2011.

[60] M. H. Siddiqi, S. Lee, Y.-K. Lee, A. M. Khan and P. T. H. Truc, "Hierarchical recognition scheme for human facial expression recognition systems," *Sensors,* vol. 13, p. 16682–16713, 2013.

[61] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park and S. Lee, "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields," *IEEE Transactions on Image Processing,* vol. 24, p. 1386–1398, 2015.

[62] H. M. M. Uddin M, "A depth video-based facial expression recognition system using radon transform, generalized discriminant analysis, and hidden Markov model.," *Multimedia Tools And Applications. 2015 Jun;74(11):3675-9.*

[63] M. B. B. Yeasin and 2. Sharma R., "Recognition of facial expressions and measurement of levels of interest from video.," *IEEE Transactions on Multimedia, 8(3), pp.500-508.*.

[64] Y. Li, S. Wang, Y. Zhao and Q. Ji, "Simultaneous Facial Feature Tracking and Facial Expression Recognition," *IEEE Transactions on Image Processing,* vol. 22, pp. 2559-2573, 2013.

[65] L. Greche and N. Es-Sbai, "Automatic system for facial expression recognition based histogram of oriented gradient and normalized cross correlation," in *2016 International Conference on Information Technology for Organizations Development (IT4OD),* 2016.

[66] I. Perikos, M. Paraskevas and I. Hatzilygeroudis, "Facial expression recognition using adaptive neuro-fuzzy inference systems," in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS),* 2018.

[67] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*, 1998.

[68] O. Sharma, "Deep Challenges Associated with Deep Learning," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon),* pp. 72-75, 2019.

[69] W. Wei, Q. Jia and G. Chen, "Real-time facial expression recognition for affective computing based on Kinect," *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA),* pp. 161-165, 2016.

[70] M. Wang, Z. Wang, S. Zhang, J. Luan and Z. Jiao, "Face Expression Recognition Based on Deep Convolution Network," *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1-9, 2018.

[71] L. Ivanovsky, V. Khryashchev, A. Lebedev and I. Kosterin, "Facial expression recognition algorithm based on deep convolution neural network," in *2017 21st Conference of Open Innovations Association (FRUCT)*, 2017.

[72] A. Mollahosseini, D. Chan and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*, 2016.

[73] Q. T. Ngoc, S. Lee and B. C. Song, "Facial landmark-based emotion recognition via directed graph neural network," *Electronics*, vol. 9, p. 764, 2020.

[74] H. Tang and T. S. Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points," *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1-6, 2008.

[75] K. a. C. W.-S. a. Z. H. Zhao, "Deep region and multi-label learning for facial action unit detection," *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 3391–3399, 2016.

[76] M. Liu, S. Li, S. Shan and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, p. 126–136, 2015.

[77] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, 2010.

[78] D. Lundqvist, A. Flykt and A. Öhman, "Karolinska directed emotional faces," *Cognition and Emotion*, 1998.

[79] G. Yadav, S. Maheshwari and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real-time video system," in *2014 international*

conference on advances in computing, communications and informatics (ICACCI), 2014.

[80] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[81] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees Proc," in *IEEE Conf. on Computer Vision and Pattern Recognition (IEEE) pp*, 1867.

[82] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, p. 357–362, September 2020.

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, p. 2825–2830, 2011.

[84] C.-W. Hsu, C.-C. Chang, C.-J. Lin and others, *A practical guide to support vector classification*, Taipei, Taiwan, 2003.

[85] K. M. Rajesh and M. Naveenkumar, "A robust method for face recognition and face emotion detection system using support vector machines," in *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*, 2016.

[86] M. Jeong and B. C. Ko, "Driver’s facial expression recognition in real-time for safe driving," *Sensors*, vol. 18, p. 4270, 2018.

[87] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[88] I. Kotsia, S. Zafeiriou and I. Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition," *Pattern Recognition,* vol. 41, p. 833–851, 2008.

[89] Y.-l. Tian, "Evaluation of face resolution for expression analysis," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[90] P. Yang, Q. Liu and D. N. Metaxas, "Exploring facial expressions with compositional features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[91] R. A. Khan, A. Meyer, H. Konik and S. Bouakaz, "Saliency-based framework for facial expression recognition," *Frontiers of Computer Science,* vol. 13, p. 183–198, 2019.

[92] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017.

[93] P. Carcagnì, M. Del Coco, M. Leo and C. Distante, "Facial expression recognition and histograms of oriented gradients: a comprehensive study," *SpringerPlus,* vol. 4, p. 1–25, 2015.

[94] P. Ekman, "An argument for basic emotions," *Cognition & emotion,* vol. 6, p. 169–200, 1992.

[95] P. Ekman, "Are there basic emotions?," 1992.

[96] D. Ping Tian and others, "A review on image feature extraction and representation techniques," *International Journal of Multimedia and Ubiquitous Engineering,* vol. 8, p. 385–396, 2013.

[97] A. Poursaberi, H. A. Noubari, M. Gavrilova and S. N. Yanushkevich, "Gauss–Laguerre wavelet textural feature fusion with geometrical information for facial expression identification," *EURASIP Journal on Image and Video Processing,* vol. 2012, p. 1–13, 2012.

[98] M. Z. Uddin, J. J. Lee and T.-S. Kim, "An enhanced independent component-based human facial expression recognition from video," *IEEE Transactions on Consumer Electronics,* vol. 55, p. 2216–2224, 2009.

[99] S. Zhalehpour, Z. Akhtar and C. E. Erdem, "Multimodal emotion recognition based on peak frame selection from video," *SIGNAL IMAGE AND VIDEO PROCESSING*, vol. 10, pp. 827-834, January 2016.