Social Network Analysis and Mining (2024) 14:65
https://doi.org/10.1007/s13278-024-01231-y

ORIGINAL ARTICLE

Deep facial expression detection using Viola-Jones algorithm, CNN-MLP and CNN-SVM

Hadhami Aouani[1,2] · **Yassine Ben Ayed**[2]

Received: 12 August 2023 / Revised: 8 December 2023 / Accepted: 23 February 2024 / Published online: 21 March 2024

Abstract

Computer vision researchers are now studying the process of recognizing emotions from facial expressions. Our system is based on his three-step method in this article, which includes face detection, feature extraction, and classification. Capture a photo/video to get facial recognition information and find the face area in this image. Face extraction uses the Viola-Jones algorithm to find reflective areas (eyes, mouth, nose, and temples) in specific faces. In order to extract the faces, we have built a database of frontal face images. We offer two systems. The first facial emotion detection system is based on classification using raw facial images, and the second extracts the oriented gradient histogram (HOG) from facial images. For the classification phase, we use three classifiers: support vector machines (SVM), Convolutional Neural Network (CNN) and hybrid CNN-SVM. To increase the performance of our facial emotion recognition system, we propose to merge the two CNN outputs of the two systems to create deep features that are merged as inputs of two classifiers (MLP and SVM). The experiments are performed the Ryerson Multimedia Laboratory (RML) dataset. The objective is to compare the performances of these methods and to identify the most suitable approach. Our experimental results showed good accuracy compared to previous studies.

Keywords Face detection · Viola-Jones · Emotion · HOG · SVM · CNN · CNN-SVM · Fusion deep features

1 Introduction

The face, which is the most noticeable portion of the human body, is used to identify a person as well as to indicate their age and gender. As many facial recognition studies show, facial emotion recognition is increasing and becoming more familiar. Scientific studies have shown that facial emotion recognition can be used for a variety of applications. Facial expression recognition technology is widely used by businesses that provide a wide range of consumer goods to evaluate client feedback and identify future consumers. In the realm of machine learning, these facial expressions are becoming increasingly important.

Facial expressions of emotion are utilized in education to assess pupils' understanding. Teachers are allowed to instruct their pupils in a method that suits them best. Different platforms, both digital and conventional ones, are used to teach the courses. The major objective is for the pupils to comprehend each lesson at the conclusion of it. Teachers can gauge whether or not their pupils understand their lessons by identifying facial expressions. The instructor can alter her delivery of the lesson based on the feedback identified using facial expression patterns. Expression recognition systems, especially emotional ones, consist of three main steps ((Tian et al. 2002)):

- Face detection: This step can be completed by estimating head pause.or by face detection
- Extraction of facial information related to expression: this information is related either to the appearance of the expression or to the geometry of the deformations. In the context of an image sequence, information related to expression dynamics is also useful.
- Expression recognition: This step is performed by classification.

✉ Hadhami Aouani
hadhami.aouani@enis.tn; hadhamiaouani22@gamil.com

Yassine Ben Ayed
Yassine.benayed@isims.usf.tn

[1] National Engineering School of Sfax (ENIS), University of Sfax, BP 1173, 3038 Sfax, Tunisia

[2] MIRACL:Laboratory,Higher Institute of Computer Science and Multimedia (ISIMS), University of Sfax, Sfax, Tunisia

We concentrate on face detection from a video library as part of our study, then on descriptor extraction, and ultimately on emotion recognition.

In this article, we initially suggested applying Viola and Jones to video frame face detection before considering these faces as inputs to our classification system. We take the Histogram of Oriented Gradients (HOG) characteristic from the detected faces in order to compare them. Support vector machines (SVM), Convolutional Neural Networks (CNN), and a hybrid CNN-SVM were the three classifiers we used, and we propose combining the CNN output from systems using raw images and HOG descriptors to build a fusion deep feature to create two systems, one with an MLP classifier and the other with an SVM classifier. The RML database is used to evaluate our research.

The rest of this article is organized as follows: Sect. 2 presents the latest research on facial emotion recognition. Section 3 describes the method of the proposed system. In Sects. 4 and 5, presents the experimental setup with results. And finally Sect. 6, we conclude our work.

2 Related works

Facial emotion recognition (FER) is an important part of human-computer interaction, enabling computers to understand facial expressions based on human thinking. The three main modules that comprise the facial expression identification process, depending on how it is handled, are face detection, feature extraction, and classification. In the realm of facial recognition methods, face detection has undergone significant development, as evidenced by established research (Insaf et al. 2020; Zhang et al. 2022). The retrieval of exceptional features from original facial images and the precise classification of these features play a crucial role in determining recognition outcomes. Notably, Gao and Ma (2020) utilized facial expression characteristics derived from facial images to predict emotional states through changes in facial expressions. Facial expression feature extraction can be categorized into geometry and appearance-based approaches. Early endeavors focused on facial geometry, encompassing aspects such as position, distance, and angles (Russell 2017; Tian et al. 2011; Valstar et al. 2017), but these methods require accurate detection of face components and are challenging to apply in real-time scenarios. In contrast, appearance-based methods have addressed the limitations of geometric feature-based approaches, leveraging techniques such as changes in facial structure, intensity, histograms, and pixel values. Principal Component Analysis(PCA) (Franco and Treves 2001), Independent Component Analysis(ICA) (Uddin et al. 2009), Gabor Wavelet (Hegde 2017) and Local Binary Pattern (LBP) (Khan et al. 2018) have been employed to extract face-specific feature descriptors. Notably, Noroozi et al. (2017) proposed a multimodal emotion recognition system integrating visual geometric features with acoustic features using random forest classifiers and CNN, SVM, achieving recognition rates of 31.67% and 36.10% on the RML and SAVEE databases using the SVM classifier. Furthermore, (García et al. 2017) introduced a framework employing a hidden Markov model (HMM) with active appearance features for facial expression recognition, achieving an accuracy of 89.04% on the RML dataset. In a separate study, Noushin et al. (2021) utilized the SVM classifier with geometric features from the RML database, achieving recognition rates of 36.06%, 39.76%,and 37.51% for the full face region, eye and eyebrow region, and nose and mouth region, respectively.

In this study, we propose the first emotion recognition system based on faces detected by Viola-Jones as geometric features, and the second system based on HOG features extracted from detected faces. The three classifiers are used: Support Vector Machines (SVM), Convolutional Neural Network (CNN) and the hybrid CNN-SVM. In order to increase the performance, we propose to combine the output of CNN of system using raw image with the output of CNN using HOG descriptor to build a fusion deep features to create two system one with MLP classifier and the second with SVM classifier. The RML database was used in our work.

3 Proposed method

The general architecture diagram of the proposed FER system is shown in Fig. 1.The main steps are, first, preprocessing that contains framing steps and face detection steps with resizing and normalization of face. Second, feature extraction and classification in the third phase.

In proposed approach, two systems (S1, S’1) are showing in Fig. 2, each system decomposed by three systems (S1, S2, S3; S’1, S’2, S’3). And a new approach that fusion deep features of the two systems to build new FER (FS1, FS2).

3.1 Data and preprocessing

In this work, the RML (Wang and Guan 2008)emotion database used which contains 720 sample audiovisual emotional expressions that were collected at the Ryerson Multimedia Lab. Six basic human emotions are expressed: Anger, Disgust, Fear, Happy, Sad, and Surprise. The database was collected from subjects speaking six different languages such as English, Mandarin, Urdu, Punjabi, Persian and Italian.

In our work, we extracted a new image dataset made by framing each video. This image dataset used as input to the algorithm of face detection by the Viola- Jones to extract the region frontal face. The experiments are performed in the Python environment.

Fig. 1 General architecture of our system of emotion recognition

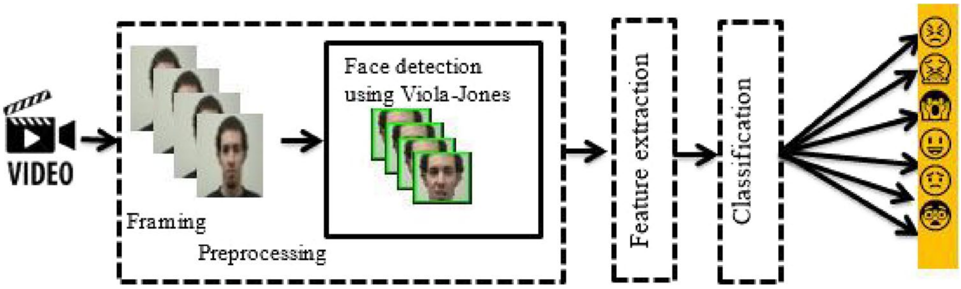
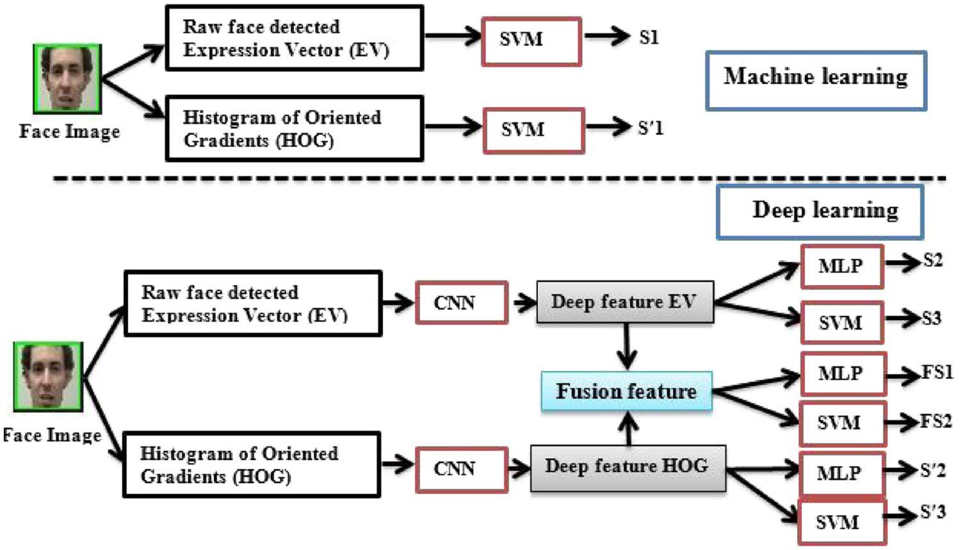


Fig. 2 Architecture of ours systems of emotion recognition


Face Image
Raw face detected Expression Vector (EV)
Histogram of Oriented Gradients (HOG)
SVM
SVM
S1
S'1
Machine learning
Deep learning
Face Image
Raw face detected Expression Vector (EV)
Histogram of Oriented Gradients (HOG)
CNN
CNN
Deep feature EV
Fusion feature
Deep feature HOG
MLP
SVM
MLP
SVM
MLP
SVM
S2
S3
FS1
FS2
S'2
S'3

3.2 Face detection

Face detection is the ability of a computer to recognize a face. Haar's feature-based waterfall classifier is simple and robust, making it a very popular facial recognition model(Aljaloud and Ullah 2020). The De Haar Cascade was used only for mouth and eye detection (Yang et al. 2018). The Viola Jones algorithm is a popular facial recognition model proposed by Paul Viola and Michael Jones in 2001 (Viola and Jones 2004). It uses rectangular bars to detect a human face in an image (Dandil and Ozdemir 2019).

The input image is converted to a grayscale image. Face detection is performed on a grayscale image using the Viola-Jones algorithm. The steps of the Viola-Jones algorithm are shown in Fig. 3.

The Viola Jones object detector is based on a binary classifier which returns a positive result when the search box contains the desired object, otherwise it returns a negative result.

The classifier can be used multiple times as the window moves over the tested image. The binary classifier used in the algorithm is realized using several hierarchy layers that form an ensemble classifier (Lo and Chow 2012). Such classifier works by classifying images based on the value of simple features. It is observed that this works much faster than a system that bases the classification on a pixel-based system (Viola and Jones 2004).

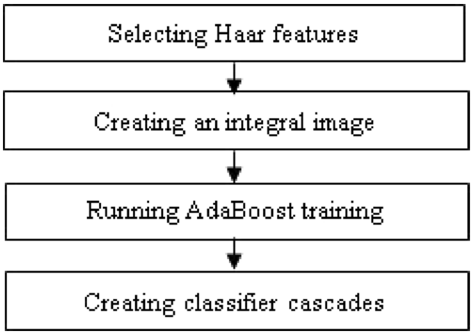

Selecting Haar features
Creating an integral image
Running AdaBoost training
Creating classifier cascades

Fig. 3 The steps of the Viola-Jones algorithm (Sehra et al. 2019)

The Viola Jones algorithm exercises control over three features dictated by Viola et al. in Viola and Jones (2004) namely, the functionality of two rectangles, the functionality of three rectangles and the functionality of four rectangles.

The framework proposed by the group has the following steps:

- The Haar feature selection: is calculated using Haar basis functions which are based on the three features listed above and usually include the summation of the pixels of the adjacent rectangular areas involved, and then calculates the difference between these sums. A representation of the Haar characteristics with respect to the corresponding detection window is shown in Fig. 4.
- The integral image is then created and is used to evaluate rectangular features in a constant time. Since the number of features can vary greatly,
- The Adaboost or Adaptive Boosting algorithm is used to select the best features and to train the classifiers using them. This is responsible for creating a "strong" classifier which is considered as a linear weighted combination of simple "weak" classifiers.
- Finally, in cascade, each step consisting of “strong” classifiers is grouped into several steps. Each step is responsible for determining whether a sub-window consists of a face or not, as shown in Fig. 5.

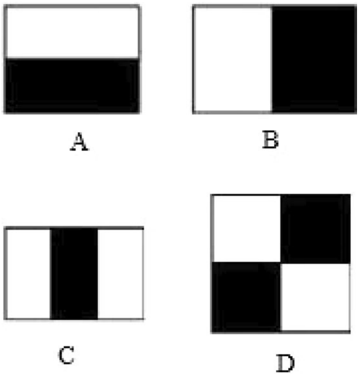


Fig. 4 Representation of the rectangular features displayed in relation to the detection window (Sehra et al. 2019)

4 Machine learning for facial emotion recognition

The Support Vector Machine (SVM) is a well-known machine learning technique that is frequently employed in applications involving emotion identification. It was first presented in 1964, and during the 1990 s, it has swiftly advanced thanks to a variety of enhanced and expanded algorithms. Although multi-class detection is a possibility, it is often employed for binary classification. Multiclass SVMs have been demonstrated to be successful in detecting the many classes of data supplied to them and are actually employed in many sectors (Dellaert et al. 1996). It enables data classification by locating an appropriate hyperplane that can separate the data with the greatest threshold, i.e., the best separation of training data projected into feature space (Ioffe and Szegedy 2015) by the function Kernel K, Linear, Polynomial, RBF, New Learning Network values are allocated and analyzed based on sets. Therefore, the use of this classification method is mainly to select good kernel functions and adjust the parameters to achieve the maximum identification speed.

We want to use SVM with these three kernel functions, so:

- Linear:

$$K(x_i, x_j) = x_i^{\Gamma} x_j \tag{1}$$

- Polynomial:

$$K(x_i, x_j) = (\gamma x_i^{\Gamma} x_j + r)^d ; \gamma > 0 \tag{2}$$

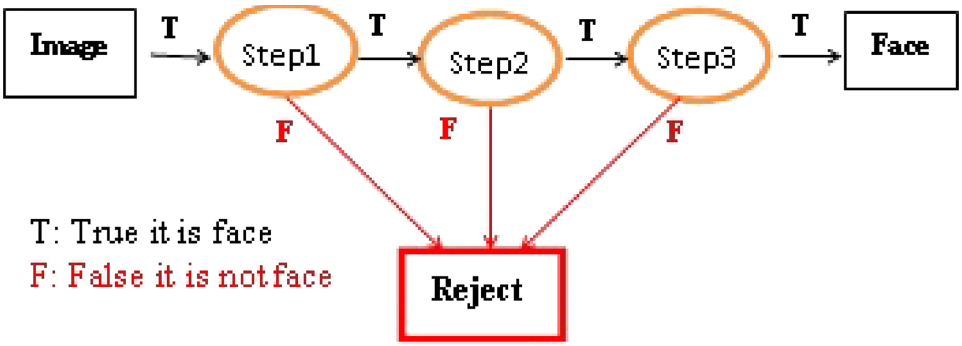
- RBF:

$$K(x_i, x_j) = \exp(-\gamma || x_i x_j ||^2) ; \gamma > 0 \tag{3}$$

With: d: degree of the polynomial, r: weighting parameter (used to control the weights) γ: kernel flexibility control parameter.

Then, the adjustment of the different parameters of the SVM classifier is done empirically, each time one changes the type of SVM kernel to determine the values γ, r, d and c the user

Fig. 5 Representation of classifier work flow in the Viola-Jones algorithm



chooses these values, in order to find the most suitable kernel parameters for our search.

We present our proposed model that is the use of raw frontal facial image from RML dataset firstly and secondly we present the other system uses feature HOG extracted from raw facial image.

4.1 Expression vector

We use the raw frontal face of dimension (64*64) as input for the SVM classifier to emotion detection. This raw frontal face called Expression Vector (EV). Figure 6 shows the architecture of SVM-based Expression Vector.

4.2 Histograms of oriented gradients

Histograms of Oriented Gradients (HOG) have applications in object recognition and pattern recognition because they can extract important information even from images captured in scrambled environments (Dalal and Triggs 2005). This was originally introduced by Dalal and Triggs (2005) to look things up. Counts the number of times gradient trends occur in local image corrections.

In this article, each 64x64 image is split into overlapping 8x8 blocks. This will generate 196 of his 8x8 blocks. Each 8x8 block is then represented by a 9-dimensional integrated template histogram to describe each image block. These extracted features are then concatenated into one block to form a 1764-dimensional feature vector for the final facial appearance (Fig. 7).

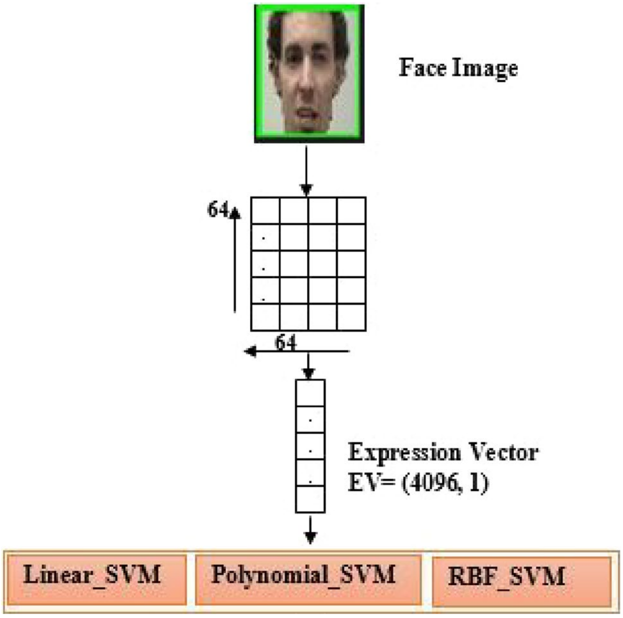


Fig. 6 SVM system of facial emotion recognition based in EV (S1)

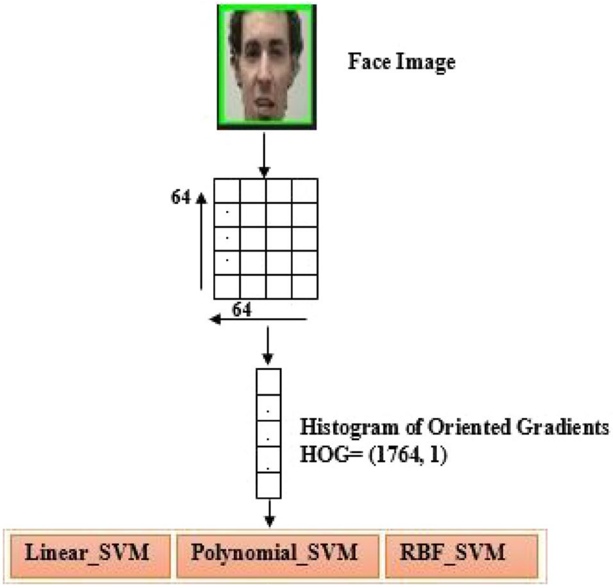


Fig. 7 SVM system of facial emotion recognition based in HOG (S'1)

The use of SVM in the two systems show the performance, after series of experiences and variation the parameters of each kernel SVM,in the second system using HOG in the two kernel RBF and polynomial with accuracy equal to 99.46%. The Table1 present the best accuracy of the system S1 that uses the raw frontal face compared by the best accuracy returned by each kernel SVM of the second system S' using the HOG feature.

Tables2 and 3 illustrate the metrics performance (Precision, Recall and F1-score) of each emotion for the two systems using SVM with the best kernel RBF.

5 Deep learning for facial emotion recognition

Deep learning plays an important role in emotion recognition. Using deep learning techniques, it is possible to create computer models that can analyze and interpret facial expressions, voice tone or even physiological data to identify human

Table 1 The accuracy rate on the test corpus obtained with the SVM for our two systems (S1, S'1)

SVM	Kernel Linear	Kernel Polynomial	Kernel RBF
S1: SVM-based EV facial emotion recognition	97.30	98.23	98.82
S'1: SVM-based HOG facial emotion recognition	96.96	99.46	99.46

Table 2 The Precision, Recall and F1-score on the test corpus obtained by our S1 system with the best kernel RBF

Systems	S1: SVM-based EV facial emotion recognition		
Emotion	Precision	Recall	F1-score
Angry	98.91	97.85	98.38
Disgust	99.51	100	99.75
Fear	98.46	97.85	98.15
Happy	99.41	99.41	99.41
Sad	97.70	99.41	98.54
Surprise	98.80	97.92	98.35

Table 3 The Precision, Recall and F1-score on the test corpus obtained by our S'1 system with the best kernel RBF

Systems	S'1: SVM-based HOG facial emotion recognition		
Emotion	Precision	Recall	F1-score
Angry	99.63	98.57	99.09
Disgust	100	99.75	99.87
Fear	99.07	98.77	98.91
Happy	99.70	100	99.84
Sad	99.55	99.70	99.12
Surprise	99.70	99.70	99.70

emotions. For example, a deep Neural Network can learn to recognize specific visual patterns that are associated with particular emotions, through successive layers of processing. To improve performance of our system, we propose the use of CNN-MLP and CNN-SVM. MultiLayer Perceptron classifier (MLP) is a class of feed forward Artificial Neural Network (ANN).It composed by input layer that is passed to one or several hidden layer and finally the output layer.

Convolutional Neural Network (CNN) is an extension of deep Neural Network that is an ANN with many hidden layer is the furthermost representative models of deep learning. It is the most widely used in recently systems. The architecture of our proposed CNN is: the raw image or the HOG descriptor is passed to a two blocks each one contains of two Convolutional layers Conv2D with Relu activation each one followed by layer batch normalization, then alternated with one of pooling layer variants and dropout. To build a feature that called Deep features. Then, Classification model it can be a one or more Fully Connected (FC) layers in which the last one outputs the class label. Our proposed CNN, it followed in the first system by the MLP classifier and SVM classifier for the second system.

We have 4 systems that employ the use of deep learning the common thing is the use of CNN to select the deep parameters for the two systems S and S' then we apply the two classification methods MLP and SVM as it is shown in the following (Fig. 8).

For the CNN model, we selected two blocks; the first block consists of two Convolutional layers (Conv2 D) with filter size equal to (3.3) and filter number 32 and 64 respectively for each Convolutional layer and followed each other by a batch normalization layer (BN) (Ioffe and Szegedy 2015). We applied a max pooling layer (Maxpooling2D) with the grouping size (4.4) at each Convolutional block to extract the most descriptive characteristics followed by dropout rate of 0.25. The second block has two Convolutional layers followed each one by batch normalization (BN) with filter size (3,3) and for the filter number is equal to 64 and 128 respectively for both layers and similarly the four layers followed by a maximum grouping layer (MaxPooling 2D) of size (4.4) followed by dropout layer rate of 0.25. Characteristic maps are called "deep feature" were then given to the classification model.

For the MLP contains of an input layer that is the deep feature of CNN model and a hidden layer with 128 cells for modeling. Batch normalization (BN) was used to improve the problem of explosion and gradient disappearance with the dropout layer with a rate of 0.25. Finally, the activation function Softmax is used in the output layer to find the classification results.

5.1 Deep learning using EV

The table below shows the system's highest accuracy rate when the Expression Vector and two deep learning models are combined. After a series of experiments in which the learning rate and optimizer function were varied, we came to the conclusion that for both systems, the best recognition rate was found to be equal to 99.76% for CNN-MLP using EV features and that the best accuracy was founded by kernel RBF equal to 99.80% for EV features for CNN-SVM when learning rate = 0.001 and RMSprop optimizer function (Table 4).

These Tables (5,6) below show the performance of the using of deep learning with raw image.

5.2 Deep learning using HOG

For Facial emotion recognition system using HOG parameter which are extracted from raw frontal face, we apply CNN model to select deep HOG feature to refine system performance then these deep hog feature are input in the two classifiers MLP (S'2) and SVM (S'3) to recognize emotion and we make the comparison between the two systems based on the performance metrics which are: precision, recall and f1-score. The following tables illustrate all of this (Tables 7, 8 and 9).

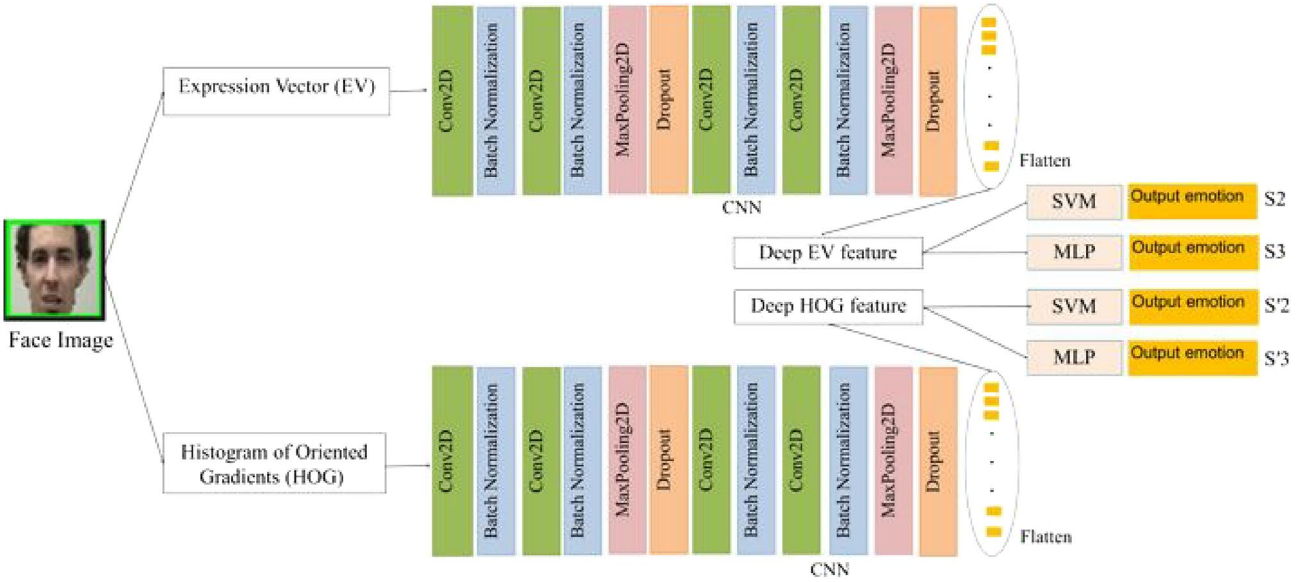


Fig. 8 Deep learning for the both systems using CNN-MLP and CNN-SVM

Table 4 The best accuracy for the system with EV using CNN-MLP and CNN-SVM classifiers

Systems	Accuracy
S2: CNN-MLP based EV feature	99.76
S3: CNN-SVM based EV feature	99.80

Table 5 The Precision, Recall and F1-score on the test corpus obtained with the CNN-MLP facial emotion recognition using EV feature

Systems	S2: CNN-MLP based EV feature		
Emotion	Precision	Recall	F1-score
Angry	100	100	100
Disgust	100	100	100
Fear	100	99.32	99.66
Happy	100	99.45	99.72
Sad	99.38	99.69	99.53
Surprise	99.11	100	99.55

5.3 Our fusion system

From the previous tables, we shows that the system using raw facial (EV) feature have the best accuracy for the two classifiers CNN-MLP and CNN-SVM. To increase the performance of the system we propose to fusion the deep features coming from CNN using EV with the deep features coming from CNN using HOG features.

Table 6 The Precision, Recall and F1-score on the test corpus obtained with the CNN-SVM facial emotion recognition using EV feature

Systems	S3: CNN-SVM based EV feature		
Emotion	Precision	Recall	F1-score
Angry	100	100	100
Disgust	100	100	100
Fear	100	100	100
Happy	100	99.45	99.72
Sad	99.07	100	99.53
Surprise	99.70	99.40	99.55

Table 7 The best accuracy for the system with HOG using CNN-MLP and CNN-SVM classifiers

Systems	Accuracy
S'2: CNN-MLP based HOG feature	97.76
S'3: CNN-SVM based HOG feature	99.31

So, Fig. 9 presents our system and Table 10 represents the best accuracy obtained by our proposed system using MLP (FS1) and SVM (FS2) using fusion deep features.

The combined deep features EV called "DEV" and HOG deep features "DHOG" defined by this equation:

$$FS = DVE + DHOG \tag{4}$$

where DVE: denote Deep Feature EV, DHOG: denote Deep feature HOG.

Table 8 The Precision, Recall and F1-score on the test corpus obtained with the CNN-MLP facial emotion recognition using HOG feature

Systems	S'2: CNN-MLP based HOG feature		
Emotion	Precision	Recall	F1-score
Angry	98.68	95.83	97.24
Disgust	98.12	98.58	98.35
Fear	95.42	98.65	97.01
Happy	98.35	98.62	98.49
Sad	98.40	95.64	97.00
Surprise	97.35	98.80	98.07

Table 9 The Precision, Recall and F1-score on the test corpus obtained with the CNN-SVM facial emotion recognition using HOG feature

Systems	S'3: CNN-SVM based HOG feature		
Emotion	Precision	Recall	F1-score
Angry	100	98.72	99.35
Disgust	99.29	99.76	99.53
Fear	99.32	98.65	98.98
Happy	100	98.90	99.45
Sad	98.46	99,69	99.07
Surprise	98.82	100	99.40

Table 11 shows the performance metrics of our proposed models used the combined deep feature. The use of deep learning methods (CNN-MLP and CNN-SVM) show the effectiveness in the system using raw facial image (EV) with an accuracy rate equal to 99.76% and 99.80% respectively for CNN-MLP and CNN-SVM. But in the system using HOG descriptor the SVM is better than CNN-MLP and CNN-SVM accuracy is 99.46%.

Our proposed FER is compared with a several recent published studies using RML dataset with different methods and features showing in Table 12.

6 Conclusion

In this research, we offer two systems: one that uses the HOG descriptor and a raw frontal face retrieved by Viola and Jones from the RML database after being tested using various classification techniques. The Support Vector Machines (SVM) is the first way of classification, and it produces the best results only when combined with the HOG features, which have a 99.46% accuracy rate in the second system. The two additional techniques are, first, the Convolutional Neural Network with Multilayer Perceptron (CNN-MLP), which, for the systems FER employing EV and the second using HOG features, respectively, obtains recognition rates of 99.76% and 97.76%. The second method is the Convolutional Neural Network with Support Vector Machines (CNN-SVM), which has accuracy rates of 99.80% and 99.31% for the two systems, respectively.

In order to improve the results, we have proposed the combination of the output of CNN from the two systems and make them as a single vector called fusion deep features as input to MLP classifier and to the SVM classifier. Fusion Deep features which shows its efficiency with two models MLP and SVM with accuracy rate equal to 99.85% and 99.90% respectively.

This work achieves better accuracy using the two systems and exactly our proposed deep features with MLP and SVM

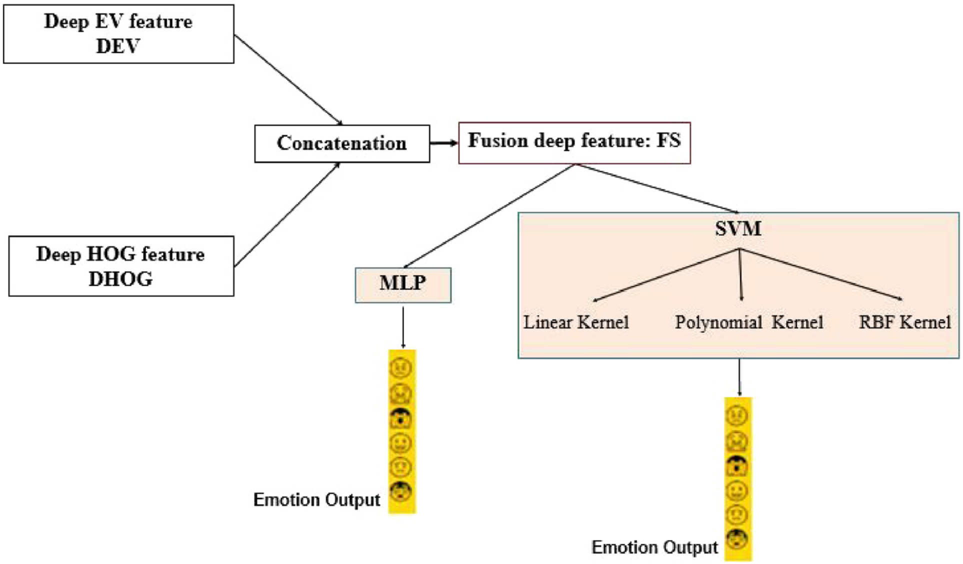


Fig. 9 Deep Facial emotion recognition using fusion features

in comparison with recent previous studies that used the same dataset.

In the future, we can think of using other types of features and apply our system on other broader bases and use methods for the reduction of the dimension of the features like the autoencoder, finally we can also consider to perform emotion recognition using an audiovisual base and in this case to benefit from descriptors of speech and others of the image. This allows us to improve the recognition rate of each emotion.

Table 10 The accuracy rate obtained using different classifiers for proposed fusion deep features (FS) for facial emotion recognition

Systems	Accuracy
FS1: Deep FER using fusion features and MLP classifier	99.85
FS2: Deep FER using fusion features and SVM classifier	99.90

Author Contributions All authors have read and agreed to the published version of the manuscript

Funding This research received no external funding.

Table 11 The Precision, Recall and F1-score on our proposed fusion deep features (FS) for FER using MLP and SVM

Systems	FS1: Deep FER using fusion features			FS2: Deep FER using fusion features		
	and MLP classifier			and SVM classifier		
Emotion	Precision	Recall	F1-score	Precision	Recall	F1-score
Angry	100	99.68	99.84	100	100	100
Disgust	100	100	100	100	100	100
Fear	100	100	100	100	100	100
Happy	100	99.72	99.86	100	99.72	99.86
Sad	99.38	100	99.69	99.38	100	99.69
Surprise	99.70	99.70	99.70	100	99.70	99.85

Table 12 Comparative table between our proposed system and other systems

State of the art	Methods and features	Accuracy (%)
Avots et al. (2019)	Frame-based emotion recognition obtained	60,20
	Using CNN using RML database	
Xianzhang (2020)	SVM using RML database, CNNs for extracting	
	Useful information from each original image frame	
	(Displacement, scale and deformation invariance	65,12
	+ features)HOG features	
Noushin et al. (2021)	SVM using RML database, Geometric characteristics	
	Whole Face Region	36,06
	Eyes and eyebrows region	39,76
	Nose and mouth region	37,51
Proposed system	SVM-based EV dimension 4096	98,82
	CNN-MLP -based EV dimension (64,64,3)	99.76
	CNN-SVM based EV (64,64,3)	99.80
	SVM-base 1764 dimension HOG feature of face image	99.46
	CNN-MLP base HOG feature of face image	97.76
	CNN-SVM base HOG feature of face image	99.31
	MLP-based Fusion deep features of CNN	
	Of the system using EV	
	Output with the output of CNN of system	99.85
	SVM-based Fusion deep features of CNN	
	Of the system using EV	
	Output with the output of CNN of system	99.90

Availability of data and materials Not applicable.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval not applicable.

References

Tian Y-L, Kanade T, Cohn JF (2002) Facial Expression Analysis. Springer, New York, pp 247–275

Insaf A, Ouahabi A, Benzaoui A, Taleb-ahmed A (2020) Past present and future of face recognition: a review. Electronics 9:1188. https://doi.org/10.3390/electronics9081188

Zhang L, Linjun S, Lina Y, Xiaoli D, Jinchao C, Weiwei C, Chen W, Xin N (2022) Arface: attention-aware and regularization for face recognition with reinforcement learning. IEEE Trans Biometrics, Behav, Identity Sci. https://doi.org/10.1109/tbiom.2021.3104014

Gao H, Ma B (2020) A robust improved network for facial expression recognition. Front Signal Process. https://doi.org/10.22606/fsp.2020.44001

Russell JA (2017) Toward a broader perspective on facial expressions. The Sci Facial Exp, 93–105

Tian Y, Kanade T, Cohn J (2011) Facial Expression Recognition, pp. 487–519. https://doi.org/10.1007/978-0-85729-932-1_19

Valstar M, Zafeiriou S, Pantic M (2017) Facial Actions as Social Signals, pp. 123–154. https://doi.org/10.1017/9781316676202.011

Franco L, Treves A (2001) A neural network facial expression recognition system using unsupervised local processing

Uddin MZ, Lee JJ, Kim T-H (2009) An enhanced independent component-based human facial expression recognition from video. Consumer Electr, IEEE Trans 55:2216–2224. https://doi.org/10.1109/TCE.2009.5373791

Hegde G (2017) Subspace based expression recognition using combinational gabor based feature fusion. Int J Image, Gr Signal Process 9:50–60. https://doi.org/10.5815/ijigsp.2017.01.07

Khan S, Hussain A, Usman M (2018) Reliable facial expression recognition for multi-scale images using weber local binary image based cosine transform features. Multimedia Tools Appl. https://doi.org/10.1007/s11042-016-4324-z

Noroozi F, Marjanovic M, Njeguš A, Escalera S, Anbarjafari G (2017) Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing PP, 60–70 https://doi.org/10.1109/TAFFC.2017.2713783

García H, Álvarez M, Orozco A (2017) Dynamic facial landmarking selection for emotion recognition using gaussian processes. J Multimodal User Interf. https://doi.org/10.1007/s12193-017-0256-9

Noushin H, Bashirov E, Demirel H (2021) Video-based person-dependent and person-independent facial emotion recognition. Signal, Image Video Process 15(5):1049–1056

Wang Y, Guan L (2008) Recognizing human emotional state from audiovisual signals*. Multimed, IEEE Trans 10:936–946. https://doi.org/10.1109/TMM.2008.927665

Aljaloud AS, Ullah AAH (2020) Facial emotion recognition using neighborhood. Int J Adv Comput Sci Appl 11:299–306

Yang D, Alsadoon A, Prasad PC, Singh AK, Elchouemi A (2018) An emotion recognition model based on facial recognition in virtual learning environment. Procedia Comput Sci 125:2–10. https://doi.org/10.1016/j.procs.2017.12.003

Viola P, Jones M (2004) Robust real-time face detection. Int J Comput Vision 57:137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

Dandil E, Ozdemir R (2019) Real time facial emotion classification using deep learning. Int J Data Sci Appl 2:13–17

Sehra K, Rajpal A, Mishra A, Chetty G (2019) Hog based facial recognition approach using viola jones algorithm and extreme learning machine. Computational Science and Its Applications - ICCSA 2019. Springer, Cham, pp 423–435

Lo C, Chow P (2012) A high-performance architecture for training viola-jones object detectors, pp. 174–181. https://doi.org/10.1109/FPT.2012.6412131

Dellaert F, Polzin T, Waibel A (1996) Recognizing emotion in speech. International Conference on Spoken Language Processing, ICSLP, Proceedings **3**

Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 448–456. PMLR, Lille, France

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection 1:886–893. https://doi.org/10.1109/CVPR.2005.177

Avots E, Sapinski T, Bachmann M, Kaminska D (2019) Audio-visual emotion recognition in wild. Mach Vis Appl. https://doi.org/10.1007/s00138-018-0960-9

Xianzhang P (2020) Fusing hog and convolutional neural network spatial-temporal features for video-based facial expression recognition. IET Image Process. https://doi.org/10.1049/iet-ipr.2019.0293

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.