Multidimensional Systems and Signal Processing (2021) 32:189–215
https://doi.org/10.1007/s11045-020-00733-0

Gravitational search-based optimal deep neural network for occluded face recognition system in videos

C. P. Shirley[1] · N. R. Ram Mohan[2] · B. Chitra[3]

Received: 6 July 2019 / Revised: 5 June 2020 / Accepted: 8 June 2020 / Published online: 17 June 2020

Abstract

Video surveillance is an effective method to improve public safety and privacy. Video surveillance technology has entered a stage where increasing video cameras are inexpensive but requiring open staff to evaluate the videos is expensive. The extensive researches conducted using machine learning for automatic face recognition doesn't provide accurate results as of human evaluation. To enhance the biometric features of the security applications, automatic face recognition is used. Surveillance records incorporate various challenges for face recognition and face detection. For instance, facial recognition systems can be affected by the size of a face image, occlusion, posture, lighting conditions, and establishment, while recognition accuracy may be affected as a result of low objectives, occlusion, posture, light, and dimness. To conquer these obstacles, an effective face detection and recognition framework is proposed with optimal feature extraction methods. At first, the keyframes with face pictures are extracted using a strategy known as keyframe extraction using wavelet information. After the extraction of keyframes, the multi-angle movement feature, SURF feature, holo-entropy, and appearance features are used for feature extraction. Finally, the recognition can be done using optimal deep neural network based on the gravitational search algorithm. Therefore, the proposed method's performance is evaluated using various benchmark video dataset. The efficiency of the proposed approach is evaluated by comparing it using sensitivity, specificity, accuracy, keyframe extraction time, etc.

Keywords Face recognition · Deep neural network · Gravitational search algorithm (GSA) · Video surveillance · Keyframe extraction · Discrete wavelet

✉ C. P. Shirley
cpshirley.id@gmail.com; shirleydavidlivingston@gmail.com

✉ N. R. Ram Mohan
nrrammohan5050@gmail.com; nrrammohan@yahoo.co.in

Extended author information available on the last page of the article

1 Introduction

Video surveillance technologies are gradually being implemented in defense applications. Therefore, possible potential risks are recognized at an early stage; yet the existence of video cameras acts as a deterrent to potential criminals. Recently, the human face identification from the video has turned into an interesting exploration point because of video surveillance and other security issues. Effective face discovery from the video has turned into a tremendous need as it can give different personality measures in the field of barrier and other security-related territories (Rejeesh 2019).

The computer vision problem and object identification are focused on the development and recognition of the human face by the variations of images, occlusions, external lighting conditions, and appearance of the human face (Ngo et al. 2008). Research in the area of face detection mainly focused on the goals of high accuracy of recognition and constant implementation under shifting conditions. The reduction of components of the information space tends to be continuously constrained by the sophistication of calculations based on information space elements (Tsagkatakis and Savakis 2009). Past biometric validation and ID applications, depends on human faces collected from a network of surveillance cameras. In our proposed technique, we have built up an effective strategy for face recognition to file a specific face from various video shots. Currently, a lot of research was performed based on an automatic face recognition concept and it is still an ongoing process, due to its various difficulties involving varying lighting conditions and pose variations. Recent research works have shifted their area from two dimensions to the three-dimension representation of a human face. In Smeets et al. (2011), 3D face recognition using a symmetric surface feature was presented, resulting in the improvement of face recognition rate and mean average precision for face recognition purposes. Moreover, the different direction of limited or no overlap into bilateral symmetry, the performance improvement was not said to be provided. Based on this issue, detector ensembles (Pagano et al. 2012) were used in video surveillance by applying dynamic niching particle swarm optimization.

Many methods for effective facial identification have been developed for better recognition. However, not many works have provided answers to the self-assuring fixation of a face image. In Liao et al. (2013), an arrangement free face representation technique dependent on multi-keypoint descriptors notwithstanding Gabor ternary example was utilized, bringing about the face recognition for both hearty and fractional faces without requiring arrangement. A near report by different annotators for video arrangement was exhibited in Srivastava et al. (2013). A crossover Euclidean-and-Riemannian measurement learning was connected in Huang et al. (2013) for extensive scale video-based face recognition. An overview of face discovery techniques with robust computer vision algorithms was planned in Zafeiriou et al. (2015).

The manner in which static and moving articles can be defined is through order techniques (Mishra and Saroha 2016) and shape, posing, and thoughts. The aim of determining the right recognition rate was to provide the model-put together methodology (Sarode and Anuse 2014) using the strategic plan for the movement-based element extraction procedure. Another Eigen probabilistic elastic part (PEP) was designed by Li et al. (2019) that created a moderate high dimensional pose-invariant portrayal that guaranteed recognition precision as well as guaranteed versatile arrangement as for many subjects. A relative investigation of human activity recognition was given in Shen et al. (2015). The keyframe yields the best quality of the face images from the video.

Conventional security or video surveillance frameworks need somebody to screen a centralized screen continuously (Shieh and Huang 2009). Be that as it may, video surveillance condition is overwhelming and pictures caught by video surveillance cameras more often are of low quality, notwithstanding uncontrolled and poses illumination conditions that make it hard to recognize individuals' faces from video surveillance (Hu et al. 2015). Various such recognition techniques have been created in the later past, which can be utilized for productive video surveillances (Ragashe et al. 2015). However, various impediments exist in multiple techniques such as most popularized surveillance cameras containing a casing rate is very small, and small in targets, but high in chaos (Yew and Suandi 2011), the identification time depends on the range of the database, which can contain a huge number of faces in security applications at occupied spots such as aircraft terminals. Since the process of face recognition requires complex computation, dedicated equipment is vital (Sudha et al. 2011).

Face recognition is a functioning area of research for a long time with a major trial. Although numerous papers have been published over the past few decades on image-based face recognition, look into video-based face recognition is not yet significantly established. The amount of camcorders that are required for video recognition in the recognition framework is rising irreducibly than any other moment. In this application, we have used broad elements to identify the individuality of the face image by eliminating the SURF highlights, appearance highlights, and holo-entropy. The appearance highlight is removed using the dynamic appearance. With the dynamic appearance, the aspect highlight is removed. The removed highlights are then used for perceiving the face utilizing DNN, where we use the gravitational search optimization algorithm for effective feature selection. In recent years, evolutionary methods achieved better performance in computer vision applications (Sundararaj 2016, 2019a, b; Vinu et al. Vinu et al. 2018).

1.1 Objectives

- We have also structured a keyframe extraction method using Wavelet information (KEWI) to ensure that the keyframes are extracted at a minimum execution to increase the accuracy of keyframe extraction.
- To identify the frames carrying the discriminatory motion features, "multi-angle movement feature selection (MMFS)" based on the development of key-frame selection is used.
- To remove the unwanted noise, a temporal average filtering-based pre-processor is applied.
- Finally, face recognition can be done using optimal deep neural network (ODNN) combined with gravitational search algorithm (GSA).

The structure of the paper is as follows. In Sect. 2, the detailed related works are presented. The proposed method is presented in Sect. 3. In Sect. 4, the results and discussions are discussed. Section 5 describes the practical implications. Section 6 concludes the paper.

2 Review of related works

Recently the detection of a human face from the video has turned into a fascinating region of exploration. Video observation has been expanded to ascend crest as a security issue in the various areas. Yoganand and Aruldoss (2015) implemented a detailed strategy for human face detection from video sequences, generated with step guides such as division, highlight extraction, and group the modified neural network. Consequently, the classification results showed a technique that had been progressively effective in grouping the faces of the film.

Camera organized video-based face recognition, as suggested by Du and Chellappa (2016). They planned to perform present uniqueness by using the repetition in the information on the multiview video. In the presence of scattering lighting, they arranged an item for lively face recognition. Along these lines offering variations, differentiating from traditional methodologies in which they gage the face's stance unambiguously. By using the repetition in the details on the multiview video, they intended to achieve present singularity. They designed an element for lively face recognition in the presence of dispersive illumination. Along these lines offering variations, separating themselves unambiguously from traditional methodologies in which they check the location of the face.

Atan et al. (2013) introduced a precise learning system dependent on multi-client multi-furnished crooks in the direction of adjusting each gadget to the transmission, including extraction and inquiry parameter. Also, the longer-term disappointment in the expected acceptance rate per face recognition effort adjacent to the result, which might presume earlier device output results in any possible setting. Arceda et al. (2016) established the method to recognize faces in security application scenes. In addition, the author used marvelous goals calculation for non-versatile insertion and a face identifier for Kanade–Lucas–Tomasi (KLT) to improve the video quality. The handling time is low and super goals are paralleled and face identifier calculations with CUDA.

Wang et al. (2018) inevitably suggested a different way in which video anomalies could be detected and controlled. The determination of the content lobe area is more precise as a further closer view restriction based on Robust PCA conspires. The ULGP-OF descriptor flawlessly consolidates the great *2D* surface descriptor LGP and optical stream. This is introduced to portray the movement insights of neighborhood district surface in the regions situated by the frontal area restriction plot. Omaima and Al-Allaf (2014) acquainted a fake neural system that was used for seeing the face response of a person. A variety of facial responses found that are utilized in the field of the picture breaking down and data grouping. The predominant procedure did not yield the apparent review of the face response acknowledgment method by ANN. Based on the accessible methodologies, few novel methodologies are used combine with its calculations to face detection accentuation using ANN strategy. To see the face responses of humans, we are using the neural system. Thus, the method distinguishes face responses by the acquisition of the huge highlights of faces. After the isolation of required highlights from the face, the isolated bits are a group to obtain the full view, which is connected with the first view. Back engendering calculation was used to deliver optimal answers for recognizing face response. Pandey (2014) saw the utilization of unique calculations for acknowledgment of face responses.

2.1 Problem definition and objectives

The success of high-quality images under controlled circumstances in face recognition is the major challenging issue. Nevertheless, the analogous level of performance obtaining in video-based face recognition is hard. Video frame analytic systems are receiving increasing interest, not only for automatic detection of abnormal events but also for enforcing human face recognition. As most video frame analytics programs are unable to dismiss a worker who builds up gradually and refines face templates over time. In fact, the first stage of programmed face recognition (image acquisition) is an essential element of face recognition. Nonetheless, face detection isn't clear as loads of picture appearance varieties, for instance, picture introduction, impediment, present variety (front, non-front), outward appearance, and enlightening condition (Sadeghipour and Sahragard 2016; Ramalingam and Chandra Mouli 2016; Ganguly et al. 2015). The basic issue in the previous face recognition methods is delineated as beneath,

- The main considerations affecting the framework for face recognition are Posture, illumination, personal characteristics, expression, and occlusion.
- Face recognition in the market and research networks has gained considerable importance but at the same time remains highly appreciated in real-time applications.
- The partial occlusions observed when capturing a face in a video recognition system makes the computational task very complex.
- Most of the technologies developed recently are too computationally intensive according to facial recognition to meet the demands for real-time devices.

The above-mentioned issues of traditional works which inspire us to investigate face recognition. Thus, the real goal of our examination is to distinguish the face with different circumstances, for example, posture, light, and so forth. The following article should be eligible for significant video surveillance. These conventional techniques pave the way for a novel facial recognition that can identify both pedestrian and terrorist attacks which creates an opportunity for structuring for powerful applications. It may very well be linked to the existing face recognition estimates in a collaborative effort to such a degree that facial biometric systems are increasingly invariant to specific forms of varieties owing to shifts in posture, walking, appearances, and enlightenment, etc.

3 Proposed approach

The primary intention of this research is to develop an efficient face recognition technology in favor of a video surveillance system. As a result, the proposed approach delivers an accurate outcome under different conditions such as occlusion, illumination, and pose. The proposed model consists of two phases and it is illustrated in Fig. 1. However, the brief explanations of the proposed methodologies are provided in the following section.

3.1 Keyframe extraction using wavelet information (KEWI)

KEWI calculation is separated into four stages. Per sub-band in the primary stage is evaluated using subtracting point of interest component estimations of present and next (for example, sequential) face locale frame. Next, the standard deviation and mean are registered from the distinction estimations of each face area extraction sub-band. In stage three, limit an opportunity for every sub-band ignoring mean and standard deviation. At last, with the limit, the distinction estimation of each band is analyzed. On the off chance that two distinct estimations of any two sub-groups have related limits, the key face district frame is the consideration of the last frame.

3.1.1 KEWI scheme frequency components

KEWI conspire distinctive video scenes cut for pre-preparing. Every keyframe characterizes a connected face that consists of exceptionally vital information of the face. In the

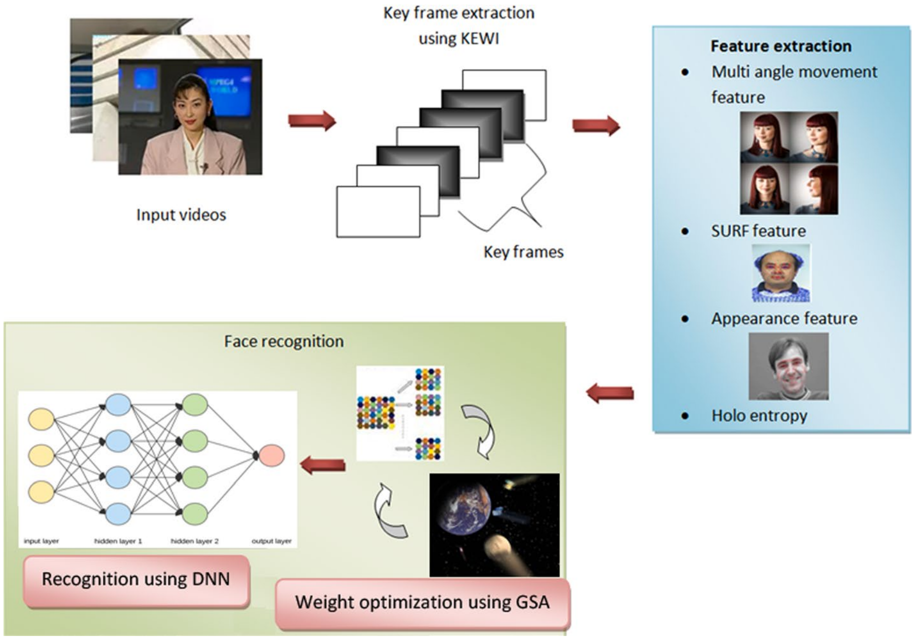


Fig. 1 Architectural diagram of the proposed methodology

underlying stage, the preprocessed ground-truth dataset is obtained from video successions. In the KEWI plot, a benefit is recognized between two progressive frames for the extraction of keyframes as a scene and visible contrasts. Thus, KEWI conspire utilizes the pair of the final frame as a key for the keyframe extraction. At that point, the keyframe is separated with the help of discrete wavelet change coefficients. The frame-filling of discrete wavelet coefficients are used. If frame-filling is modified, then the coefficient details are never totally equivalent. The keyframes used in image round down, highlight selection, and other readings are applied in the next major face area object extraction. Therefore a composite of the lower KEWI calculation and a low extraction time is required.

In the video indexing and retrieval process, keyframe and object detection is the essential step. Comprehensively, the utilized KEWI creates the distinction measurements by assessing the element information obtained from the packet stream. The technique of key-frame extraction is executed using contrast measurements by discrete wavelet information. Keyframe extraction dependent on forwarding movement examination and DWT coefficients of lingering blunder is gotten. Each frame scans for the ideal coordination in the relation of reference frames, at this point the prescient error of movement remuneration with DWT coding is reduced. In the meantime, a couple of movement vectors are exchanged. Keyframes are removed dependent on the peculiarity of video streams utilized for preparing. On the off chance that a face cut happens, frame principal is picked as a keyframe. The video frames are coded with forwarding movement returns in the video stream. At the point, when a transformation happens at a frame, extraordinary modifies occur in the frame comparing to the past reference frames. Therefore, the reference frames' impact is neglected using an encoder. In the course of working out the proportion without compensating movement, a method is introduced and it used to detach the frame and pick the keyframe.

Here two progressive frames are perused and changed with DWT. For key extraction, the face area is accomplished into four sub-groups such as *LL, HL, LH,* and *HH*. Inside the four sub-groups, just three sub-groups of *HL, LH,* and *HH* are utilized to separate keyframe since the low recurrence band is *LL* and is never utilized for KEWI handling. Figure 2 demonstrates the key face region extraction with algorithmic strategy.

3.2 Feature extraction

After keyframe extraction, feature extraction is an important step. The features are the main aspect needed in the recognition process. The feature extraction contains simplifying some resources required to depict a huge set of data precisely. In each region of an individual image in a video shot, four features are computed using our proposed model. Feature Selection that combines the optical flow with biologically inspired face features to extract the most discriminatory information from the face is presented. The features are namely (1) multi-angle movement feature selection, (2) SURF feature, (3) appearance features, and (4) holo entropy. Therefore, the four features are the extensive feature set that we use to extract. The three features can aid in effectively recognizing the face from multiple angles present in any video. The brief explanation of each feature is given in the below sections,

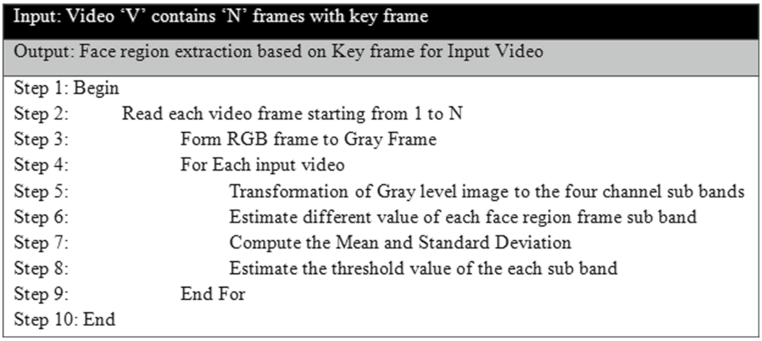
3.2.1 Multi-angle movement feature selection

In this section, the multi-angle movement feature selection (MMFS) combines the optical flow with biologically inspired face features to extract the most discriminatory information from the face. With the pre-processed raw video sequences using the temporal Wiener average denoising algorithm, a de-noised frame is obtained, which is then fed as input to extract multi-angle movements for human face recognition.

3.2.1.1 Optical flow extraction for human face recognition To obtain stable measures, the MMFS initially identifies the difference between adjacent frames and is expressed as below.

$$\left(R_{i+1}, R_i\right) \rightarrow R_{i+(1/2)} \tag{1}$$

With the obtained difference between adjacent frames, the TA-MMFS identifies the optical flow between adjacent frames and is expressed as given below.



```
Input: Video 'V' contains 'N' frames with key frame
Output: Face region extraction based on Key frame for Input Video
Step 1: Begin
Step 2:     Read each video frame starting from 1 to N
Step 3:         Form RGB frame to Gray Frame
Step 4:         For Each input video
Step 5:             Transformation of Gray level image to the four channel sub bands
Step 6:             Estimate different value of each face region frame sub band
Step 7:             Compute the Mean and Standard Deviation
Step 8:             Estimate the threshold value of the each sub band
Step 9:         End For
Step 10: End
```

Fig. 2 Key face region extraction algorithm

$$\left(R_{i+(1/2)}, R_{i-(1/2)}\right) \rightarrow F \tag{2}$$

From the above equation, the optical flow of human faces is extracted between adjacent frames that form as an input to the biologically inspired features. With these optical flow features, the low-level features (i.e., intensity, size, shape, and structure) are mapped with the high-level features (i.e., expression, scar, moles, and skin) and the mapping is explained in the next section.

3.2.1.2 Biologically inspired features of the face Motivated by the work of biologically inspired features for scene classification (Pagano et al. 2012), the MMFS extracts the features for human recognition as they encode intensity information with multi-angle movement (i.e., intensity, size, shape, score). The feature extraction process focuses on the multi-angle movement that detects video frames moving the most discriminative information. The MMFS is motivated by mechanisms in the face that consists of two different layers S_1 and C_1. The complex patterns of images mimic C_1 in the face forming an intensity feature.

Initially, S_1 is first calculated by using Gabor convolution kernel with multiple intensities and shapes in the direction of target images $R(i, j)$ and is expressed as given below.

$$GF(i,j) = R(i,j)\left[\exp\left(-\frac{(i^2 + \gamma^2 j^2}{2\sigma^2}\right) * \cos\left(\frac{2\pi}{\lambda} i\right)\right] \tag{3}$$

From (3), with the two values $i = i\cos\theta - j\sin\theta$ and $i = i\cos\theta + j\sin\theta$ the Gabor convolution, Kernel factor is evolved for the de-noised target image $R(i, j)$. The biologically inspired features of the face with two layers S_1 and C_1 are shown in Fig. 3.

In the MMFS method, the Gabor filter is defined at four different intensities from 5×5 to 11×11 an incremental size of two pixels. Also, four different shapes 0–180 at 45° increment have been accepted. Similarly, $4 \times 4 = 16$ and S_1 feature maps are computed. As the increment size of the pixel is 2, precise information resulting in the key extraction efficiency is evaluated.

Next, to obtain the C_1 MMFS method measures the adjacent frames with an identical intensity at various window structures from 2×2 to 8×8 with size increment of two pixels, and upon identification of the maximum intensity pixel box that has been utilized to denote the equivalent pixel inside feature map as C_1. This provides precise and robust features improving the key extraction efficiency.

Finally, the mapping is performed because the features extracted from the keyframes comprises of low-level features such as intensity, size, shape, score. The MMFS maps these low-level features with its corresponding high-level biological features such as expression, scar, moles, and skin to create a template by assigning a score value to each restored image (i.e., de-noised image). The score value is proportional to the similarity between the extracted features (i.e., low-level features) and the normal feature (i.e., expression, scar, moles, and skin) respectively. Higher, the score value (i.e., based on the similarity function), higher the rate of extraction is said to be, and the mapping function (based on similarity function) is expressed as given below.

$$sim_f = \sum_{i=1}^{4} (f_i)(\mu_i)(\sigma_i) \tag{4}$$

The similarity score value of each extracted feature is performed with the normal face, where f_i corresponds to the face independent random variable, μ_i corresponds to the mean

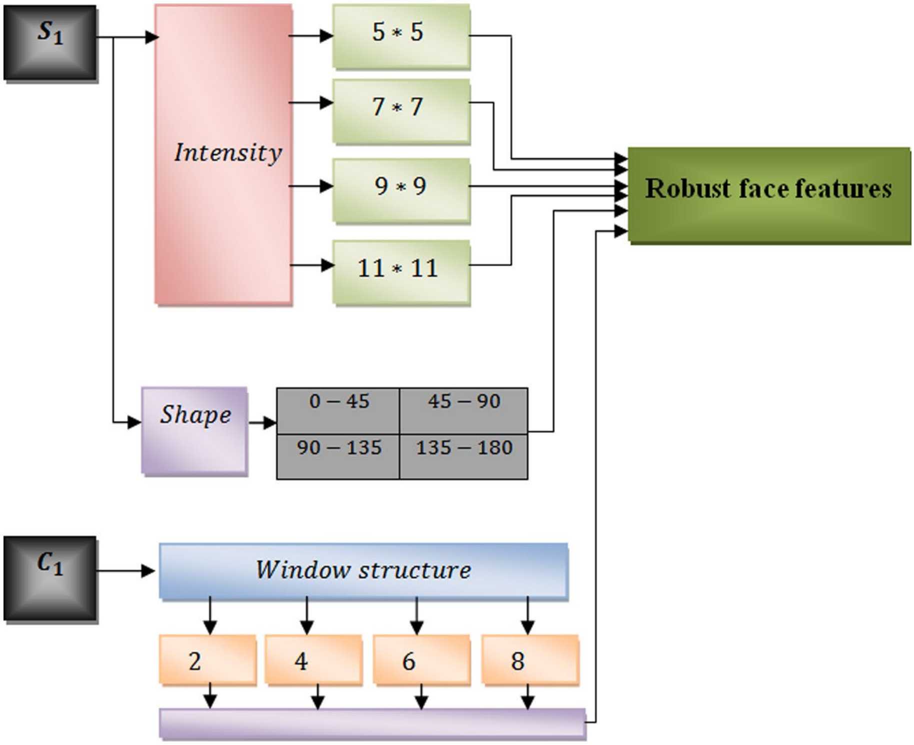


Fig. 3 Biologically inspired features of the face with two different layers S_1 and C_1

random value, and σ_i corresponds to the standard deviation random value. In this way, by applying the MMFS algorithm can extract the discriminatory features from every video sequence frame from the faces. Figure 4 shows the algorithmic step for Multi-angle Movement Feature Selection.

Each de-noised image from Fig. 4, the multi-angle movement feature selection initially performs an optical flow model to extract features from the images sequence. Next, depending upon the biologically inspired features a face extraction model to extract the most discriminative information using the two layers S_1 and C_1 is presented. The feature extraction with the aid of S_1 and C_1 in turn extracts robust features, improving the feature extraction efficiency.

3.2.2 SURF feature

For both images during training and testing, the proposed model uses speed-up robust feature extraction methods (SURF) to extract the features for facial recognition. One of the important rotation and scale-invariant feature extraction method is known as SURF. When compared to the scale-invariant feature transform, the SURF is the faster feature extraction method. The major focus of the SURF method is image descriptors, in-plane, and scale rotation invariant detectors. The pixel intensities in the integral image are calculated using Eq. (5).

```
Input: De-noised frame $f'(i,j)$
Output: Improved feature extraction accuracy and time
Begin
    For each De-noised frame $f'(i,j)$
        Measure the difference between adjacent frames using (5)
        Measure optical flow between adjacent frames using (6)
    End for
    For each De-noised frame $f'(i,j)$ with multiple intensities and shapes
        For two different layers $S_1$ and $C_1$
            Measure Gabor convolution kernel using (7)
            Measure intensity and shape for $S_1$
            Measure window structure for $C_1$
        End for
    End for
End
```

Fig. 4 Multi-angle movement feature selection algorithm

$$h_{\xi}(p,q)=\sum_{a=0}^{a\leq p}\sum_{b=0}^{b\leq p}h(a,b) \tag{5}$$

During SURF feature extraction, the intensity points are determined by the Hessian matrix and it represented in Eq. (6).

$$H_m(g,\phi)=\begin{bmatrix} I_{g,g}(g,\phi) & I_{g,h}(g,\phi) \\ I_{g,h}(g,\phi) & I_{h,h}(g,\phi) \end{bmatrix} \tag{6}$$

where the convolution of the Gaussian second-order derivative function is noted as $I_{g,g}(g,\phi)$. Based on image locations is to detect the maximum determinant of the Hessian matrix. Therefore, the maximum intensity point is provided by the determinant of the Hessian matrix also the proposed system is implemented using extracted features of maximum intensity points.

3.2.3 Appearance features

The next feature is the appearance feature for image recognition. From the face image, the appearance model features are determined using the Active appearance model (AAM), and it is a statistical template matching technique. The training set is more helpful to represents a texture and shape variability.

Here, the AAM model is generated by combining shape variation with appearance variation. The point sets are aligned into a common coordinate frame, which is represented in terms of the vector. This is expressed as in Eq. (7) below,

$$p=\bar{p}+V_s z_s \tag{7}$$

From Eq. (7), $\bar{p}$ is the mean shape. The shape variation set of orthogonal modes and the set of shape parameters are represented as V_s and z_s.

The grey level appearance of the statistical model is obtained; we combine every query image position to join the mean shape. Next, the sampling of the grey level is performed. The final model is obtained using the expression below,

$$g_L = \bar{g}_L + V_g z_g \tag{8}$$

where the grey level vector of mean normalized is $\bar{g}_L$. Also, V_g and z_g are the grey level variation set of orthogonal modes and the parameters set of grey level model.

By using the Eqs. (7) and (8), the shape and appearance of some images are obtained in vector form. Finally based on the above expression we generate the concatenated vector which is given below,

$$z = \begin{pmatrix} D_s z_s \\ z_g \end{pmatrix} = \begin{pmatrix} D_s V_s^T (p - \bar{p}) \\ V_g^T (g_L - \bar{g}_L) \end{pmatrix} \tag{9}$$

In Eq. (9), each shape parameter with the diagonal matrix of weights is denoted as D_s and apply PCA to these vectors we get,

$$z = Ea \tag{10}$$

Therefore, E and z is the eigenvector, a is the appearance vector parameter that manages grey levels and shapes. The appearance feature is estimated using the above expressions and these values can be further applied to the next stage for the recognition of the face from the video.

3.2.4 Holo entropy

The texture of the input image characterization by the statistical measure of randomness is known as holo entropy and it is based on image histogram value. The total correlation of the random vector Y is known as holo entropy. All attributes with some of the entropies are expressed and the removal of outlier candidates is used to minimize the holo entropy. The equation below shows the expression to calculate holo entropy.

$$EN_{hl}(Z) = \sum_{i=1}^{n} E(z_i) \tag{11}$$

Therefore, the holo entropy for the image Z is represented as $EN_{hl}(Z)$. The image pixel entropy is denoted as $E(z_i)$. When the independent component of Z has a single component, $EN_{hl}(Z) = E(z_i)$ i.e., entropy and holo entropy coincides with each other.

3.3 Classification using optimal deep neural network (ODNN)

After the feature extraction, the chosen features have been fed to the input for the classification phase. The classification stage has two phases such as (1) training and (2) testing. In this, the training process used 80% of images, and the remaining 20% of images are utilized in the testing process. We use the optimal deep neural network (DNN) for the classification stage and the weight values of DNN are optimally selected using gravitational search algorithm (GSA). The DNN is a model of an artificial neural network, which consists of multiple layers of hidden units and outputs. Additionally, the parameter learning consist of fine-tuning and pre-training (using generative deep belief network or DBN) stage. From the training data set, feature selection is the major goal of this research; likewise, the input features are correctly classified using optimal weight determination. In this paper, fine-tuning stage weight values are optimized using GSA.

3.3.1 Deep belief networks

The deep architecture with a feed-forward neural network is an important part of DBN and it has a number of the hidden layers. The classification approach of DBN is shown in Fig. 5, which has a visible unit of the input layer, L the number of hidden layers, and the output layer. The weights $W^{(j)}$ between the DBN parameters are biases $b^{(j)}$ of layer j and among the unit layers $j-1$.

3.3.1.1 Pre-training stage How to initialize this parameter is one of the common issues for training deep neural network architectures. In low generalization, the poor local minima of the fault function determined by arbitrary initialization affect optimization algorithms. Restricted Boltzmann machines (RBM) of the training sequence problems is utilized by Hinton (2010). The two-layer repeated neural network is called RBM and the binary stochastic inputs used are weighted connections. The RBM with the initial layer takes the inputs (visible units v) and the net layer concealed units h. After RBM training, the hidden units are taken as feature detectors and the input vector compact is represented. The RBM structure is represented in Fig. 5 also the energy function of RBM is denoted in Eq. (12):

$$E(v,h) = -h^T W v - b^T v - c^T h \tag{12}$$

Therefore, visible and hidden layer bias vectors are b, c and weight is W with the conditional distributions functions are denoted as follows:

$$P(v_i = 1 \mid h) = \sigma\left(b_i + \sum\nolimits_j W_{ji} h_j\right) \tag{13}$$

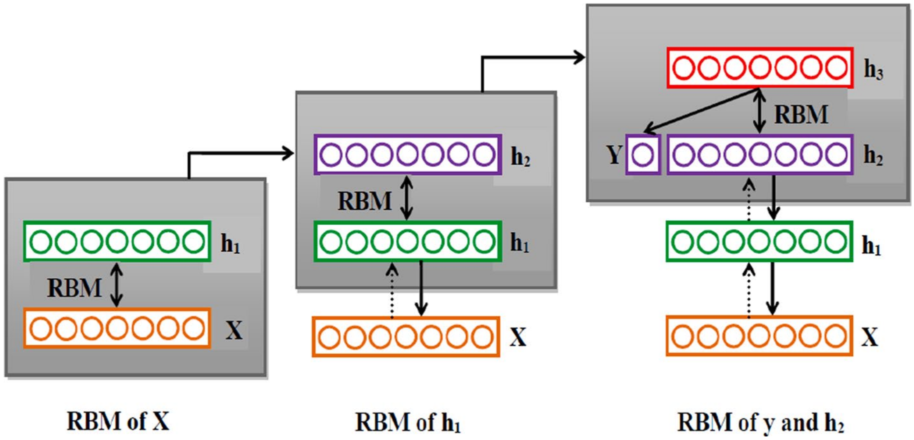


Fig. 5 Deep belief network with three hidden layers *h1, h2, h3*, one input layer *x*, and one output layer *y*

$$P(h_i = 1 \mid v) = \sigma\Big(c_i + \sum\nolimits_j W_{ji}v_j\Big) \tag{14}$$

where logistic function become $\sigma(a) = 1/(1 + e^{-a})$ tends to the range (0, 1).

The RBM training and applied procedures are explained in this section. Primarily, the RBM training is unsupervised and ignores the class label at specific training examples. In Eq. (28), the conditional distribution is followed by the hidden units' outputs generating a binary vector of the next generous distribution. RBM can conversely direct the circulated vector and it impacts the similar input data (confabulation). Finally, the confabulation of RBM is to propagating the condition of the hidden units. Thus, the process is continued and the parameters obtained are listed as follows:

$$\Delta W_{ji} = \eta\Big(\big\langle v_i h_j\big\rangle_{data} - \big\langle v_i h_j\big\rangle_{reconstruction}\Big) \tag{15}$$

$$\Delta b_i = \eta\big(\langle v_i\rangle_{data} - \langle v_i\rangle_{reconstruction}\big) \tag{16}$$

$$\Delta c_j = \eta\Big(\big\langle h_j\big\rangle_{data} - \big\langle h_j\big\rangle_{reconstruction}\Big) \tag{17}$$

3.3.1.2 DNN pre-training procedures

- Initially, the training vector with visible units v is initialized.
- Next, Eqs. (28) and (29) is used to update the hidden units in parallel.
- In the same way, Eq. (12) is used to update the visible units in parallel.
- Based on the observed reconstruction of the same equation used in step 2, the hidden units in parallel are again updated.
- Weights are updated using $\Delta w_{ij}\alpha\langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{reconstruction}$.

Once the RBM is trained, the multilayer model is created and one more RBM can be *stacked* at the top. In every instance, the stacked RBM with units present in the already-trained RBM layers is allocated and the visible input layer with an initialized vector is found with the current biases and weights. The new RBM input with already-trained layers and the procedures are continued till the desired stopping criterion is met. In this work, we use the input layer as 253, hidden layer three, and only one output layer. The fine-tuning stage is initialized via obtained deep network weights.

3.3.2 Fine-tuning stages

In this fine-tuning stage, we adjust weight value and minimize the error using the gravitational search algorithm (GSA). Here, at first, we randomly initialize the weight values W_{ij}. The solution representation is an important process for solving the problem in the entire optimization algorithm. Then, we calculate the fitness for each solution. In this paper, classification accuracy is considered as the fitness function. The fitness is measured based on Eq. (21). Thereafter, fitness estimation and GSA solutions are updated. The updation steps are given in the next section. The output layer is presented at the top of the deep neural network that classifies the images. The training dataset D^T is used to optimize the weight in the training stage. Initially, few features are only given to the DNN thereafter the weight is given. Ultimately, the testing of the dataset D^T is to categorize the images based on the optimal weight (w).

Chiefly, the reduced characteristics are offered to the DNN, while the weight is arbitrarily altered. During optimal weights selection, the GSA algorithm is used by the proposed model and the algorithm steps are explained in beneath.

3.3.3 Gravitational search algorithm (GSA)

The law of gravity and motion are the basis of the gravitational search algorithm (GSA). Hence, the algorithm is clustered in a population-based technique involving different masses. Based on the gravitational force, the masses are distributing data to direct the search to the best location in the search space. Newton's law of gravity and Newton's law of motion are to build the GSA and it is shown in Eqs. (18) and (19).

$$F_{ij}^{d}\ (t) = G(t)\ \frac{AM(t) \times PM(t)}{D_{ij}^{2}} \tag{18}$$

$$a_i(t) = \frac{F_{ij}^{d}(t)}{M_i(t)} \tag{19}$$

where AM_i is the active gravitational mass that is linked with the agent i and PM_j is the passive gravitational mass is connected j. The t is the Gravitational constant at a time and denoted by $G(t)$ as well as the Euclidian distance among two agents i and j is denoted as D_{ij}. Similarly, the acceleration and inertial mass are denoted as $a_i(t)$, and $M_i(t)$.

Based on our work, the object is the consideration of agents and their masses to evaluate their routine. Force of gravity is to attract each object and trigger the entire object toward heavier masses in order to find the global solution. The weightier mass ensures an excellent performance-based approach. Based on GSA, every mass (agent) includes four requirements: (1) the passive gravitational mass, (2) position, (3) inertial mass, and (4) active gravitational mass. Hence, the mass location is representing a solution to the problem; also gravitational and inertial masses used to determine fitness function. Thus, every mass suggests a solution, and the procedure is focused on appropriately acclimatizing the inertial and gravitational masses. The heaviest mass obtained over a period of time brilliantly brings in an optimum solution for the search space.

Step 1 Initialization of agents.

The N number of agents positions are randomly initialized with the population weight and ith agent positions are denoted in Eq. (20).

$$w_i = \left(w_i^1\ ,\ w_i^2, \ldots, w_i^d, \ldots, w_i^n \right) \quad for\ = (i = 1, 2, \ldots, N) \tag{20}$$

In Eq. (20), the dth dimension with ith agent weight position is represented as w_i^d.

Step 2 Best fitness computation.

Based on the issue of minimization, the number of iteration with best and worst fitness for each agent is represented using Eq. (21)

$$Fit\,(t) = \min error \tag{21}$$

$$best\,(t) = \min_{j \in (1,\ldots,N)} Fit_j(t) \tag{22}$$

$$worst(t) = \max_{j \in (1,\ldots,N)} Fit_j\,(t) \tag{23}$$

Step 3 Computation of gravitational constant (G).

The time period until the beginning is scaled down and the gravitational constant G is initialized. Where, the function of the initial value is specified as G_0 and time (t) is given as shown below:

$$G(t) = G\big(\,G_0\,,\,t\big) \tag{24}$$

In fitness evaluation, gravitational and inertial masses are effectively analyzed. So, the weightier mass is indicated using an incredibly efficient agent. Otherwise, superior magnetism was incredibly slow when achieving the excellent agents. According to the gravitational and inertial mass equality, the masses' values are estimated as the fitness map.

Step 4 Calculation of agent mass.

The iteration t to consider all agents with its inertial and Gravitational masses $AM_i = PM_j = M_i$, where $i = 1, 2, \ldots, N$

$$m_i\,(t) = \frac{Fit_i(t)\;-\;worst\,(t)}{best\,(t)\;-\;worst\,(t)} \tag{25}$$

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)} \tag{26}$$

Therefore, the agent i fitness value at the time t is denoted by $Fit_i(t)$.

Step 5 Calculation of accelerations agent.

At the time t, the acceleration of the agent i is expressed as:

$$a_i(t) = \frac{F^d_{ij}(t)}{M_i(t)} \tag{27}$$

Step 6 Agents velocity and positions.

Moreover, current velocity can deem the fraction of resulting velocity and acceleration. Therefore, the velocity and position are evaluated using the equations that are given below.

$$V_i^d(t+1) = rand_i \times V_i^d(t) + A_i(t) \tag{28}$$

$$R_i^d(t+1) = R_i^d(t) + v_i^d(t+1) \tag{29}$$

where $rand_i$ is the uniform random variable in the interval [0, 1].

Step 7 Repeat steps 2–6.

Repeat steps 2–6 until the iterations are reached their maximum limit. The global fitness with the best fitness value at the last iteration is processed and the position of the consistent agent at stated dimensions has been calculated by the specific problem of a global solution. Hence by using the above method, objects in the input video sequence are tracked. The presentation of the projected technique is estimated and the outcomes are elucidated beneath.

The above-mentioned classification technique competently achieves superlative classification accuracy. In this work, we apply the feature vector to the neural network, which is taken from video shots. Based on the available feedback image and feature values, the classifier classifies the image. As a result, the classifier produces relevant and irrelevant images.

4 Results and discussion

The proposed method execution is done in MATLAB (version 2015a). The recordings of information are selected from an ordinary database, and recordings are prepared using our proposed method. Once the division is finished, highlight extraction is conducted where various highlights are extracted for each frame as well as the next extraction of important highlights. The face is identified and classified by using an improved neural network. Thus, the test outcomes are shown in the following section.

4.1 Dataset description

Here, the recordings are taken as of Derf's accumulation (video dataset) and HMDB51 dataset. Each video arrangement is uncompressed by the design of YUV4MPEG utilized by the JPEG devices scheme, except if generally showed. Hence, the organization acknowledged via Theora encoder devices. For example, few encoding parameters lack a frame rate from the rough information that has been speculated or deduced and could be off-base. A portion of the video selected from Akiyo video, vehicle telephone, foreman, and so forth and explored different avenues regarding some continuous video too. Hence, the sample video image dataset is formulated in Fig. 6. The pose, illumination, and occlusion images from the sample database are shown in Fig. 7.

4.2 Performance metrics

The proposed approach performance is checked via different kinds of evaluation metrics such as sensitivity (*Sen*), accuracy (*A*), specificity (*Spec*), precision (*P*), and recall (*R*). Each performance metrics formulas are explained in the following equation.

$$A = \frac{PT + NT}{Total\, no.\, of\, samples} \times 100 \tag{30}$$

$$Sen = \frac{PT}{PT + NF} \times 100 \tag{31}$$

$$Spec = \frac{NT}{NT + NF} \times 100 \tag{32}$$

$$P = \frac{PT}{PT + PF} \tag{33}$$

$$R = \frac{PT}{PT + NF} \tag{34}$$

where *PT* and *NT* are the true positive and true negative values of the ground-truth keyframe. Also, the *PF* and *NF* are represented the false positive and false negative values of the ground-truth keyframe respectively.

4.3 Performance analysis

4.3.1 Keyframe and feature extraction performance

The proposed performance of this research is analyzed using execution time, feature extraction time, keyframe extraction accuracy (Shirley et al. 2016), sensitivity, specificity, and accuracy. The video stream of the representative frame is keyframe, and it provides the most accurate video content with a compact summary. The determination of

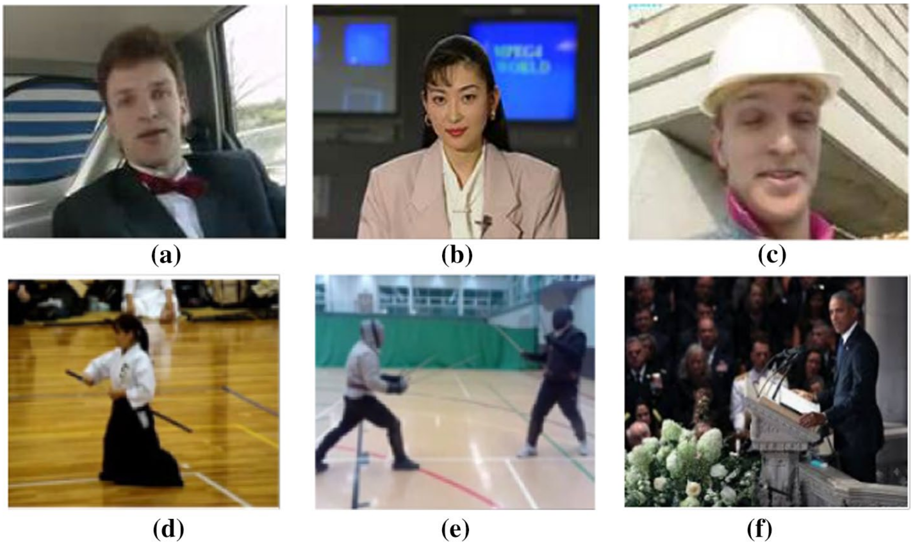

(a) (b) (c)
(d) (e) (f)

Fig. 6 Sample videos from the database **a**–**f**

Fig. 7 Sample videos from the database. **a**–**c** Pose, **d**–**f** illumination and **g**–**i** occlusion

accuracy and superiority of the proposed framework is more important and we conduct the experiment using different video frames. The keyframe extraction performance is assessed by the performance metrics namely precision and recall. The number of keyframe detection with precision and recall of proposed work is formulated in Table 1.

In the initial keyframe extraction phase, KEWI's performance is evaluated. The Bayesian keyframe extraction using Wavelet information (KEWI) (Shirley et al. 2016) methods is compared with state-of-art methods such as three dimensional face modeling (TDFM) (Medioni et al. 2019), and automatic object reconstruction and extraction (AOE-R) (Lu and Li 2008). The execution time for separating keyframes is also expanded with an increment measure of video, but this is not observed directly due to the relative proximity of disturbance in the video. The keyframe extraction of KEWI is compared with existing TDFM and AOE-R techniques. As a result of the proposed algorithm delivers an optimal output because it has taken only a few execution times while comparing to other methods. The memory size is chosen from 113.6 MB to 936.2 MB for this analysis. In addition, the KEWI plan links a proposed condition of work share without movement compensation to effectively use the frames reference impact. The KEWI plan is used to extract the rate of keyframe from human face recognition with two previous methodologies such as AOE-R and TDFM as explained in Table 2. Different numbers of frame/second are used in the range from 10 frames to 70 frames and are analyzed using MATLAB.

Table 1 Proposed approach with the keyframe detection result

Name of the video	Number of detected keyframes	Precision	Recall
Carphone video	15	0.879	0.8
Akiyo video	31	0.945	0.87
Multiface student video 1	20	0.93	0.86
Multiface student video 2	16	0.97	0.79
Multiface student video 3	34	0.84	0.77
Soccer	29	0.897	0.73
Container	14	0.945	0.82
Foreman video	21	0.90	0.71

At Fig. 8, we describe the keyframe extraction accuracy rate with different sizes of recordings taken as information available from a range of 113.6–936.2 MB with the final objective of the analysis, by sequences and designed in parallel form. The keyframe extraction accuracy suggested by KEWI conspires higher, contrasting with existing TDFM (Medioni et al. 2019) and AOE-R strategies (Lu and Li 2008). Other than it can correspondingly expand the extent of video, the rate of keyframe extraction accuracy is additionally expanded by every one of the strategies. Be that as it may, relatively, the accuracy obtained is higher using the KEWI scheme.

Figure 9 as shown above, calculates the accuracy of the keyframe extraction rate and it is higher with KEWI conspire usage. The keyframe extraction rate accuracy has been checked for various number of frames with fluctuating video sizes. The pixel scheme utilization, related pixel estimations of two diverse video frames are assessed and the distinction esteem is derived by applying edge esteem. The accuracy of the KEWI keyframe extraction is 5.96% compared to TDFM. In addition, the rate of keyframe extraction precision is 11.56% better than AOE-R by making an unbiased adjustment by the option of a fitting reward limit in the KEWI plot (Lu and Li 2008).

Figure 8 shows the effects of highlighting the extraction time using contrasting TA-MMFS techniques, and two cutting-edge techniques, namely CBDFT (Yoder et al. 2010) and FS-PV (Passalis et al. 2011), for visual comparisons based on the relevant information. As shown in the Fig. 8, the extraction time of the component is increased bit by bit as the video is extended. The CBDFT and FS-PV strategies vary from the TA-MMFS strategy by consolidating the optical stream extraction which distinguishes between the video

Table 2 Comparative analysis of execution time (ms) for keyframe extraction

Size of the video	Proposed KEWI	TDFM	AOE-R
113.6	3.7	5.37	6.37
323.7	5.23	7.54	9.23
349.5	8.45	10.72	11.45
454.5	9.23	13.17	14.83
635.2	10.67	15.62	16.79
905.3	12.78	17.6	18.78
936.2	13.67	18.67	20.8

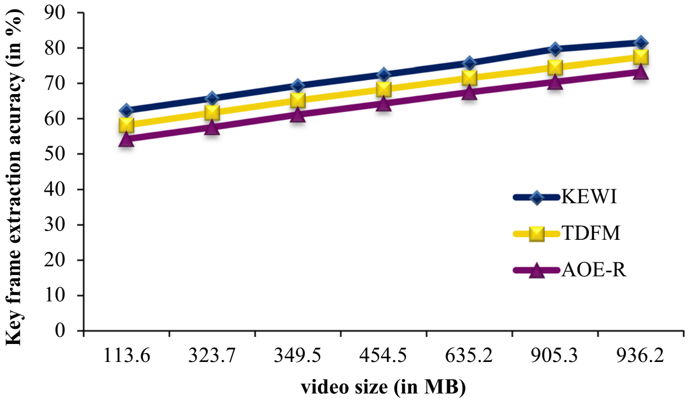


Fig. 8 Comparison of keyframe extraction accuracy

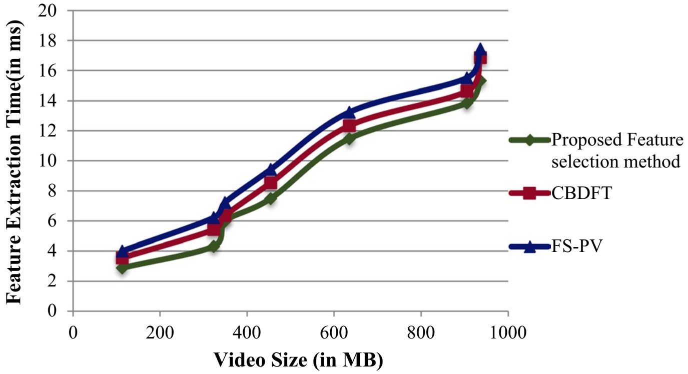


Fig. 9 Performance variance of feature extraction time

frames and conveys the most discriminatory information. In this way, it decreases the element extraction time of the TA-MMFS strategy by 14% when contrasted with CBDFT. Furthermore, the optical stream between nearby frames is estimated, which further diminishes the component extraction time of the TA-MMFS technique by 25% when contrasted with FS-PV. The various feature extracted values using the proposed work is described in Table 3.

Table 3 shows the recognized output from each video input. The video input is fed to our system and based on that the feature values are extracted from each image. In this proposed work, we extract AAM features, holo entropy, Multi-angle features, and Surf features. When compared to all other features, the combined feature only delivers a better result of classification. So, we use the combined features for classification and its performance is shown in Fig. 10.

4.3.2 Classification performance

The face is compared and classified using the DNN for facial recognition. The performance of proposed DNN-GSA for face recognition results is estimated in terms of pose,

Table 3 Proposed feature extraction values

S. no.	Input videos	SURF	AAM	Holo entropy	Multi-angle features
1	Carphone video	4.621	114.2853	0.00215	1.67
2	Akiyo video	5.38	102.0416	0.026952	5.678
3	Multiface student video 1	5. 47	94.93071	0.00751	6.45
4	Multiface student video 2	3.81	105.5182	0.00331	2.34
5	Multiface student video 3	0.52	132.24	0.00310	0.34
6	Soccer	1.611	121.8984	0.00412	1.45
7	Container	1.16	95.53801	0.022277	9.45
8	Foreman video	5.01	100.543	0.00198	3.5

illumination, and occlusion and are tabulated in Table 4. All the datasets with different resolution images are chosen as well as the number of poses also mentioned. Recognition without occlusion and with 40% occlusion results is formulated in Table 4.

For the evaluation of the proposed method, we use accuracy, sensitivity, and specificity as evaluation metrics. Thus, the received metrics are compared with state-of-art methods such as neural network and fuzzy and its effectiveness is tabulated in Table 5. The results obtained for the Akiyo video, Carphone video, and real-time video are also tabulated in Table 6. Here, the proposed method outperforms the existing methods because of the extensive feature sets and unique points detection using the AAM method.

The performance analysis of the proposed DNN-GSA algorithm with state-of-art methods in terms of the Akiyo video, Car phone, and multi-face video is depicted in Table 6. The performance of the proposed DNN-GSA algorithm is compared with other existing methods such as artificial neural networks (ANN) (Omaima and AL-Allaf 2014) and

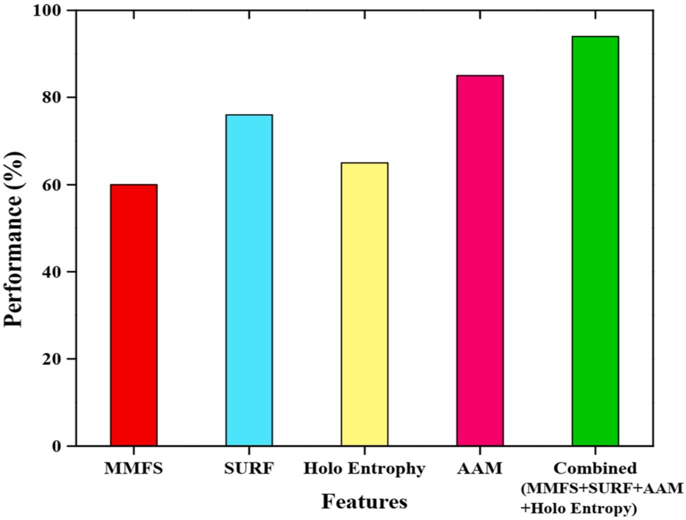


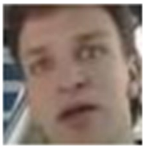
Fig. 10 Performance analysis of feature usage

Table 4 Proposed DNN-GSA face recognition performance

Database images	Frame size (resolution)	Pose	Illumination	Occlusion	
				Recognition without occlusion result (%)	Recognition with occlusion 40% result
Real-time video 1	134×352	1	4	90.67	88.78
Real-time video 2	124×152	10	43	88.45	86.56
Real-time video 3	234×252	23	8	93.67	91.67
Soccer	134×256	10	9	94.34	93.12
Akiyo video	125×312	12	2	92.45	89.56
Foreman	125×125	35	12	92.34	90.56
Carphone	134×152	17	45	91.45	89.23
Container	154×158	23	35	98.23	92.13

k-nearest neighbour (k-NN) (Saravanan 2016). From this, we observed that the Akiyo video obtained more accurate results when compared to other videos mainly because of less motion and a single face. In the multi-face video, more than single face images with various pose and occlusion factor occurs which slightly affects the accuracy. The proposed DNN-GSA is compared with some other existing methods such as (Li et al. 2015),

Table 5 Recognized output from each video input

S. no.	Input videos	Recognized output	S. no.	Input videos	Recognized output
1	Real-time video 1		5	Akiyo video	
2	Real-time video 2		6	Foreman	
3	Container		7	Carphone	
4	Soccer		8	Real-time video 3	

DNN-HML (Salman et al. 2019), DNN-DIBR (Xie et al. 2016) to ensure efficiency. Figure 11 shows the comparison of recognition accuracy between our proposed methods with existing DNN approaches. Finally, the proposed DNN-GSA delivers optimal and higher accuracy of face recognition output than other methods such as DNN, DNN + DIBR, and DNN + HML.

5 Practical implications

Security is the biggest issue in information technology for every individual, organization, and even nation and worldwide. Face recognition is the most important security aspect of video surveillance systems. The identification and authentication methods that are using the Face recognition process have been used in various areas like entrance control system in buildings, access control for computers in general or ATMs, withdrawing money from a bank account or post office, and in a criminal investigation. Our proposed face recognition system will handle various factors. So it will be used for video surveillance and in the defense department for crime investigation.

6 Conclusion

In this paper, we designed a face recognition approach from the video. The key features are extracted using Keyframe extraction using the Wavelet information (KEWI) method. The intention of giving a proposed methodology that guarantees the least execution time as of separating keyframes and build the keyframe extraction precision in support of different video frames and videos. The process of extracting keyframes is scheduled to identify the final pair of the keyframe to retrieve the keyframe correctly, with negligible time to identify the human face.

In the second stage, video surveillance by extensive feature set with occlusion and pose invariant face recognition system is proposed. For various video inputs, we apply the proposed strategies for face recognition, which include pose, illumination, and occlusion. The features extracted are mainly AAM features, holo entropy, multi-angle features, and SURF features. For recognition of face from the video, we have used a productive framework where the ideal DNN is combined with the GSA for achieving an improved recognition rate.

Table 6 Evaluation metrics for the proposed DNN-GSA and existing method

Methods	Akiyo video			Carphone video			Real-time video		
	Sensitivsity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
Proposed DNN-GSA	97.58	100	98.72	95.91	99.4	98.41	96.03	97.3	97.35
ANN	96.11	90	95.16	93.67	92	93.76	92.76	90	94.23
k-NN	94.47	88	93.83	95.04	90	93.67	93.87	88	92.01

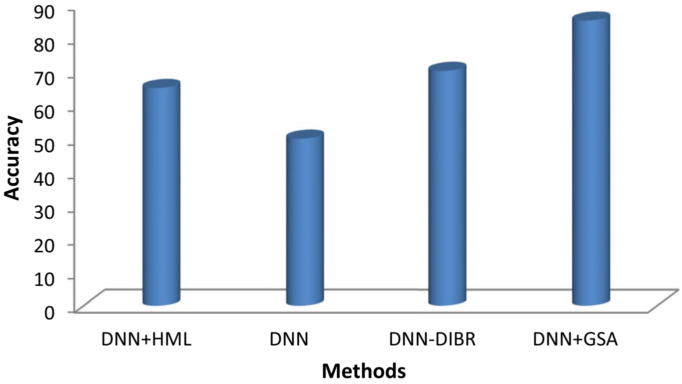


Fig. 11 Recognition accuracy with state-of-art methods

References

Arceda, M., Fernández Fabián, V. E., Laguna Laura, K. M., Rivera Tito, J. J., & Gutiérrez Cáceres, J. C. (2016). Fast face detection in violent video scenes. *Electronic Notes in Theoretical Computer Science, 329,* 5–26.

Atan, O., Andreopoulos, Y., Tekin, C., & van der Schaar, M. (2013). Bandit framework for systematic learning in wireless video-based face recognition. *IEEE Journal of Selected Topics in Signal Processing*. https://doi.org/10.1109/JSTSP.2014.2330799.

Du, M., & Chellappa, R. (2016). Face association for videos using conditional random fields and max-margin markov networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Vol. 38, no. 9, pp. 1762–1773).

Ganguly, S., Bhattacharjee, D., & Nasipuri, M. (2015). Wavelet and decision fusion-based 3D face recognition from range image. *International Journal of Applied Pattern Recognition, 2*(4), 306–324.

Hinton, G. (2010). *A practical guide to training restricted Boltzmann machines*. University of Toronto, UTML TR 2010-003.

Hu, X., Liao, Q., & Peng, S. (2015). Video surveillance face recognition by more virtual training samples based on 3D modeling. In *Proceedings of 11th international conference on natural computation (ICNC)*, 2015.

Huang, Z., Wang, R., Shan, S., & Chen, X. (2013). Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning. *Elsevier, Computer Vision and Image Understanding, 117*(10), 1384–1399.

Li, H., Hua, G., Shen, X., Lin, Z., & Brandt, J. (2015). Eigen-PEP for video face recognition. In *Computer vision—ACCV 2014, 9005* (pp. 17–33). Springer.

Li, S., Neupane, A., Paul, S., Song, C., Krishnamurthy, S., Roy-Chowdhury, A. K., & Swami, A. (2019). Stealthy adversarial perturbations against real-time video classification systems. In *NDSS,* 2019.

Liao, S., Jain, A. K., & Li, S. Z. (2013). Partial face recognition: alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(5), 1–14.

Lu, Y., & Li, Z.-N. (2008). Automatic object extraction and reconstruction in active video. *Pattern Recognition, 41*(3), 1159–1172.

Medioni, G., Choi, J., Kuo, C.-H., & Fidaleo, D. (2019). Identifying noncooperative subjects at a distance using face images and inferred three-dimensional face models. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems And Humans, 39*(1), 12–24.

Mishra, P. K., & Saroha, G. P. (2016). A study on classification for static and moving object in video surveillance system. *International Journal of Image Graphics and Signal Processing*. https://doi.org/10.5815/ijigsp.2016.05.07.

Ngo, T. D., Le, D.-D., Satoh, S., & Duong, D. A. (2008). Robust face track finding in video using tracked points. In *Proceedings of IEEE international conference on signal image technology and internet based systems*, 2008.

Omaima, N., & AL-Allaf, A. (2014). Review of face detection systems based artificial neural networks algorithms. *International Journal of Multimedia and Its Applications, 6,* 1–16.

Pagano, C., Granger, E., Sabourin, R., & Gorodnichy, D. O. (2012). Detector ensembles for face recognition in video surveillance. In *The 2012 international joint conference on neural networks* (IJCNN) (pp. 1–8).

Pandey, S. (2014). Review: Face detection and recognition techniques. *International Journal of Computer Science and Information Technologies, 5,* 4111–4117.

Passalis, G., Perakis, P., Theoharis, T., & Kakadiaris, I. A. (2011). Using facial symmetry to handle pose variations in real-world 3D face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(10), 1938–1951.

Ragashe, M. U., Goswami, M. M., & Raghuwanshi, M. M. (2015). Approach towards real time face recognition in streaming video under partial occlusion. In *Proceedings of IEEE sponsored 9th international conference on intelligent systems and control (ISCO)*, 2015.

Rejeesh, M. R. (2019). Interest point based face recognition using adaptive neuro fuzzy inference system. *Multimedia Tools and Applications, 78*(16), 22691–22710.

Ramalingam, S. P., & Chandra Mouli, P. V. S. S. R. (2016). Two-level dimensionality reduced local directional pattern for face recognition. *International Journal of Biometrics, 8*(1), 52–64.

Sadeghipour, E., & Sahragard, N. (2016). Face recognition based on improved SIFT algorithm. *International Journal of Advanced Computer Science and Applications, 7*(1), 548–551.

Salman, A., Siddiqui, S. A., Shafait, F., Mian, A., Shortis, M. R., Khurshid, K., et al. (2019). Automatic fish detection in underwater videos by a deep neural network-based hybrid motion learning system. *ICES Journal of Marine Science*. https://doi.org/10.1093/icesjms/fsz025.

Saravanan, D. (2016). Video substance extraction using image future population based techniques. *ARPN Journal of Engineering and Applied Science, 11*(11), 7041–7045.

Sarode, J. P., & Anuse, A. D. (2014). A framework for face classification under pose variations. In *International conference on advances in computing*, *communications and informatics (ICACCI)* (pp. 1886–1891).

Shen, H., Zhang, J., & Zhang, H. (2015). Human action recognition by random features and hand-crafted features: A comparative study. In *Computer vision—ECCV 2014 workshops, 8926* (pp. 14–28). Springer.

Shieh, W.-Y., & Huang, J.-C. (2009). Speedup the multi-camera video-surveillance system for elder falling detection. In *Proceedings of international conferences on embedded software and systems*, 2009.

Shirley, C. P., Lenin Fred, A., & Ram Mohan, N. R. (2016). Video key frame extraction through wavelet information scheme. *ARPN Journal of Engineering and Applied Sciences*, *11*(7).

Smeets, D., Keustermans, J., Hermans, J., Claes, P., Vandermeulen, D., & Suetens, P. (2011). Symmetric surface-feature based 3D face recognition for partial data, biometrics (IJCB). In *2011 International joint conference on IEEE biometrics compendium, IEEE RFIC virtual journal, IEEE RFID virtual journal* (pp. 1–6).

Srivastava, G., Yoder, J. A., Park, J., & Kak, A. C. (2013). Using objective ground-truth labels created by multiple annotators for improved video classification: A comparative study. *Computer Vision and Image Understanding, 117*(10), 1384–1399.

Sudha, N., Mohan, A. R., & Meher, P. K. (2011). A self-configurable systolic architecture for face recognition system based on principal component neural network. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2011.2133210.

Sundararaj, V. (2016). An efficient threshold prediction scheme for wavelet based ECG signal noise reduction using variable step size firefly algorithm. *International Journal of Intelligent Engineering and Systems, 9*(3), 117–126.

Sundararaj, V. (2019a). Optimised denoising scheme via opposition-based self-adaptive learning PSO algorithm for wavelet-based ECG signal noise reduction. *International Journal of Biomedical Engineering and Technology, 31*(4), 325.

Sundararaj, V. (2019b). Optimal task assignment in mobile cloud computing by queue based ant–bee algorithm. *Wireless Personal Communications, 104*(1), 173–197.

Tsagkatakis, G., & Savakis, A. (2009). Random projections for face detection under resource constraints. In *Proceedings of 16th IEEE international conference on image processing (ICIP).*

Video dataset from https://media.xiph.org/video/derf/.

Vinu, S., Selvi, M., & Kumar, R. S. (2018). An optimal cluster formation based energy efficient dynamic scheduling hybrid MAC protocol for heavy traffic load in wireless sensor networks. *Computers and Security, 77,* 277–288.

Wang, S., Zhu, E., Yin, J., & Porikli, F. (2018). Video anomaly detection and localization by local motion based joint video representation and OCELM. *Neurocomputing, 277*(14), 161–175.

Xie, J., Girshick, R., & Farhadi, A. (2016). Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European conference on computer vision* (pp. 842–857).

Yew, C. T., & Suandi, S. A. (2011). A study on face recognition in video surveillance system using multi-class support vector machines. In *Proceedings of IEEE region 10 conference, TENCON*, 2011.

Yoder, J., Medeiros, H., Park, J., & Kak, A. C. (2010). Cluster-based distributed face tracking in camera networks. *IEEE Transactions on Image Processing, 19*(10), 2551–2563.

Yoganand, A., & Aruldoss, C. K. (2015). A region growing and modified neural network classifier based face detection technique from video. *International Journal of Applied Engineering Research, 10*(12), 30231–30248.

Zafeiriou, S., Zhang, C., & Zhang, Z. (2015). A survey on face detection in the wild: Past. *Present and Future, Elsevier, Computer Vision and Image Understanding, 138,* 1–24.

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

Affiliations

C. P. Shirley[1] · N. R. Ram Mohan[2] · B. Chitra[3]

B. Chitra
chitrarammohan09@gmail.com

[1] Department of Computer Science and Engineering (CSE), C. S. I. Institute of Technology, Thovalai, Tamilnadu, India

[2] Department of Computer Science and Engineering, Annai Vailankanni College of Enginneering, Nagercoil, India

[3] Department of Electronics and Communication Engineering, Arunachala College of Engineering for Woman, Anna University, Chennai, India