DOI: 10.21917/ijsc.2024.0496

DEEP Q-NETWORK (DQN) PARTIALOCCLUSION SEGMENTATION AND BACKTRACKING SEARCH OPTIMIZATION ALGORITHM (BSOA) WITH OPTICAL FLOW RECONSTRUCTION FOR FACIAL EXPRESSION EMOTION RECOGNITION

S.S. Sudha[1] and S.S. Suganya[2]

[1]Department of Applied Mathematics and Computational Sciences, PSG College of Technology, India
[2]Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science, India

Abstract

Video facial expression recognition (FER) has garnered a lot of attention recently and is helpful for several applications. Although many algorithms demonstrate impressive performance in a controlled environment without occlusion, identification in the presence of partial facial occlusion remains a challenging issue. Solutions based on reconstructing the obscured area of the face have been suggested as a way to deal with occlusions. These options mostly rely on the face's shape or texture. Nonetheless, the resemblance in facial expressions among individuals appears to be a valuable advantage for the reconstruction. For semantic segmentation based on occlusions, Reinforcement Learning (RL) is introduced as the initial stage. From a pool of unlabeled data, an agent learns a policy to choose a subset of tiny informative image patches to be tagged instead of full images. In the second stage, a trained Backtracking Search Algorithm (BSA) is used to rebuild optical flows that have been distorted by the occlusion. On obtaining optical flows estimated from occluded facial frames, AEs restore optical flows of occluded regions. These recovered optical flows become inputs to anticipate classes f expressions. Optical flux reconstructions then classify stages. This study evaluates classification model's performances for face expression identification based on Very Deep Convolution Networks (VGGNet). Furthermore, it produces more accurate confusion matrices and proposes approaches for the KMU-FED and CK+ databases, respectively. The results are evaluated using metrics including recall, f-measure, accuracy, and precision.

Keywords:

Facial occlusions, facial expressions, optical flow, Reinforcement Learning (RL), Deep Q Network (DQN), Backtracking Search Optimization Algorithm (BSOA), Very Deep Convolution Networks (VGGNet), and Driver Facial Expression Emotion Recognition (DFEER)

1. INTRODUCTION

Humans communicate their emotions through facial expressions. They can affect how we conduct our lives by changing our focus, perception, and memory, which makes it easier for us to understand the intentions of other people [1]. Since facial expressions convey information, they are a suitable means of expressing human emotions. People can also gain insight into the inner thoughts of others by studying their facial expressions [2]. Automated image-based FER has developed from computer vision technology as potential tools for standardizations, scales, and facilitate research on facial emotions [3].

The results of research on automated FERs are now being used in virtual reality [5], augmented reality games [6], consumer marketing and advertising, academics [7], and HCIs (human-computer interfaces) [4,]. FERs are essential parts of sophisticated driving assistance systems because, when combined with intelligent automotive technology, they can identify driver fatigue and promote safe driving [8].The advancement of artificial intelligence and other fields depends on FER research [9]. Though facial recognition technology has many uses and a bright future, there are a few technical issues that need to be resolved before the real product arriving.

In practice, a number of unfavorable external factors alter the facial image, such as unsuccessful face frame extractions or poor recognition; hence, for facial expression recognition (FER) in a video image to be a successful system assistance, it must have a high recognition rate [10]. There will be an increase in the adoption of FER applications due to their superior speed and precision. The PMVO is mentioned for optical occlusions in this current paper. Compute optical fluxes between the frames of an occluded face sequence as the first step. The second stage involves training a Parallel Adaptive Multi-Verse Optimizer (PAMVO) to reconstruct optical flows that have been deformed due to occlusion. After each fixed iteration, the initial solutions are divided using the parallel approach, which permits information sharing across groups, into pairs of occluded and non-occluded optical flows at random..

Solutions based on reconstructing the obscured area of the face have been suggested as a way to deal with occlusions. These options mostly rely on the face's shape or texture. Nonetheless, the resemblance in facial expressions among individuals appears to be a valuable advantage for the reconstruction. To solve this issue suggested a novel RL for semantic segmentation based on segmenting occlusions using the FER. From a pool of unlabeled data, an agent learns a policy to choose a subset of tiny informative image patches to be tagged instead of full images. To enhance the efficiency of activity segments, task model information on the partial sequence of activities is utilized in conjunction with an online search procedure. In the second stage, a trained Backtracking Search optimization algorithm (BSOA) is used to rebuild optical flows that have been distorted by occlusion. On obtaining optical flows estimated from occluded facial frames, AEs restore optical flows of occluded regions. The repaired optical flow is then used as an input by the recognition to forecast the expression class. Although the segmentation is improved by the simulation findings, there are still problems with accuracy and improper occlusion creation like low repeatability and poor generalizability across image domains.

2. LITERATURE REVIEW

Fobi et al [11] suggested a novel and broadly applicable two-stage methodology for improved pixel-by-pixel image segmentation when missing and misaligned annotations are present. First displayed is the Alignment Correction Network, which will be used to rectify open-source labels that have been wrongly registered. Next, we show how, in the presence of missing annotations, the segmentation model Pointer Segmentation Network forecasts infrastructure footprints using updated labels. We apply Alignment Correction Networks on OpenStreetMaps labels for correcting building footprints, demonstrating usability of even less qualitative data sources. We also show that the Pointer Segmentation Network accurately predicts California's cropland boundaries from medium resolution data. The suggested technique is durable across a range of applications using different quantities of training data, enabling a mechanism to extract useful information from noisy, imperfect data.

Nakisa et al [12] suggested a temporal multimodal fusion technique that captures non-linear emotional links inside and between blood volume pulse (BVP) and electroencephalography (EEG) data, applying deep learning (DL) to increase classification precisions of emotions. The evaluation of the suggested models uses Early and late fusions. Initially, distinct deep networks for modalities are trained on merged EEG and BVP data using ConvNet long short-term memory (LSTM) models. Consequently, concurrent examinations evaluate intricately linked representations of emotions across several modalities. Temporal multimodal DL model's performances are evaluated based on early and late fusion procedures and contrasted with other techniques on smart wearable sensor dataset where experimental findings demonstrate that temporal multimodal DL models classified human emotions in their respective defined quadrants.

Liu et al [13] proposed that Attention is calculated using partial class activation attention (PCAA), which employs class-level representations at the local and global levels. PCAA gathers local class centers and calculates local relationships for pixels-classes after obtaining the partial CAM. The usage of local-specific representations ensures consistent outcomes in various local contexts. The characteristics are aggregated and global representations from each local class center collected to ensure global consistency. According to experimental findings, Partial CAM performs better than the other two approaches on pixel relationships in benchmarks, including Cityscapes, Pascal Contexts, and ADE20K.

Lu et al [14] presented the multiple spatio-temporal feature fusion (MSFF) framework, which combines two mutually complimentary sources—the face image and audio—to more correctly capture spatial and temporal emotional information. Facial image and audio models are components of the framework. Three alternative architectures of spatial-temporal neural networks are utilized in the facial image model to extract discriminative features about various emotions from images of people's facial expressions. The first step involves using pre-trained convolutional neural networks (CNN) like ResNet-50 and VGG-Face to extract high-level spatial information from video images. The speech spectrogram images that are acquired by preprocessing audio are also modeled in a VGG-BLSTM framework for the audio model in order to more effectively define the emotional fluctuation. Lastly, to improve emotion identification performance, a fusion strategy based on the score matrices of many spatiotemporal networks obtained from the previous framework is suggested. Our proposed MSFF has an overall accuracy of 60.64%, outperforming the winning team's performance and a considerable improvement over the baseline, according to extensive experiments.

Siu et al [15] proposed unique convexity shape before segmentation frameworks that split zones either fully or partially convex based on users' choices. Convexity constraint with segmentation models based on registrations is fundamental ideas where they are applied on conformal features of image meshes. This work provided iterative techniques to trace borders of target items by gradually deforming templates for solving segmentation models. Projections are employed for upholding convexity criteria. Target objects are then caught by (fully or partially) convex zones and subsequently fully convex shapes need to know convexity as previous knowledge, but partially convex forms must know positions of partial convexities. This work used synthetic as well as real image graphs where their outcomes demonstrated the usefulness of the framework.

Chung et al [16] proposed a variety of techniques and techniques are used to reduce bad emotions and enhance driving experiences, which reduce stress related to driving. Systems with notifications, driver assistance and environmental comforting are the three categories into which these technologies are separated. Notification alert systems improve driving experience by raising the driver's awareness of their physiological state and lowering the chance of an accident. In bad weather, driving assistance systems direct and support the driver. Driver stress brought on by environmental changes can be reduced with the use of the environmental calming approach.

Chang et al [17] designed a DL system that combines V-A estimation, AU detection, and face attribute identification. Since both AUs and V-A space may be used to identify different emotion kinds, the basic idea is to utilize AUs to estimate V-A intensity. Additionally, a CNN is used to train the AU detector to recognize face attributes. Trials were conducted using the outcomes of the three previously described activities to verify the functionality of our suggested network architecture.

Theagarajan et al [18] suggested a deep driver Automated System for Measuring Arousal and Valence in Automobile Driver Videos to show that deep learning networks are capable of extracting more robust and powerful face features from massive volumes of data than the most advanced human-crafted features. The network surpasses state-of-the-art methods since it was trained just on raw face images. This work's approach employs CNN for face extraction and recognition and an LSTM to represent changes in CNN characteristics over time. The AFEW-VA dataset and videos from the 2014–16 Motor Trend Magazine Best Driver Car of the Year were used to assess methodologies where the methodology performed better than the other seven methods.

Bhatti et al [19] suggested a feedforward learning paradigm and the use of FER by a teacher in the classroom. to accomplish effective high-level feature extraction, faces are detected from collected lecture recordings, and relevant frames are selected after removing any unnecessary frames. Deep features are then retrieved and fed into a classifier that employs several convolutional neural networks with parameter adjustments. In the classroom, regularized extreme learning machine (RELM) classifiers categorized unique expressions of instructors, promoting efficient learning and algorithm generalization. The experiments are carried out in classroom settings utilizing three benchmark face datasets: the Cohn-Kanade, Japanese Female Face Expression (JAFFE), and FER 2013 (FER2013) datasets, as well as generated instructor FER dataset. Furthermore, the suggested method is compared to convolutional neural networks, traditional classifiers, and cutting-edge approaches. The trial results show a significant improvement in parameters like as recall, accuracy, and F1-score.

Zhu et al [20] suggested MVML framework a new block-row regularizer which selected informative views (i.e., eliminate the uninformative views) for high-level feature selections using the F-norm regularizer. Lower level feature selections were then carried out on informative views using the 12,1-norm regularizer. The block-row regularizer is designed to carry out hierarchical feature selections. Over-fitting can be avoided by using a block-row regularizer; duplicate views can be removed and the data's natural group structures preserved by using an F-norm regularizer; and noisy features may be removed by using a 12,1-norm regularizer. Finally, compared against three cutting-edge techniques and two baseline algorithms, the suggested method outperforms the others in terms of classification performance on real image datasets.

Guo et al [21] proposed an algorithm called Image segmentation. Partial differential equations are used in the investigation of image segmentation, a technique that has become more and more efficient with the advancement of computer technology. Along with a study and assessment of SegNet-v2 segmentations in medical images, thorough explanations of curve representations are offered using in plane differential geometry. The test findings demonstrate that greater development of the partial differential equation image segmentations are required for enhancing accuracies, particularly in the domain of medical image segmentation.

Al Machot et al [22] suggested a model hyper-parameter optimization technique and a CNN architecture that may advance the area of human emotion detection they guaranteed findings for detections of subject dependent or independent human emotions that were robust. CNN models were trained and parameters of suggested CNN architectures were optimized using grid search strategies. The overall performances of the proposal were verified and highlighted using the MAHNOB and DEAP datasets. Their findings demonstrated remarkable increases in robustness for multiple assessments and increased the accuracy of subject dependent and subject independent classifications (four classes/labels) for both MAHNOB and DEAP datasets. The study unequivocally shows that a robust categorization of human mood is achievable using only non-intrusive EDA sensors, even in the absence of additional or different physiological inputs.

Cruz et al [23] created a novel FER technique. Volumetric expansions of the well-known POEM features, TPOEM (Temporal Patterns of Oriented Edge Magnitudes) features look at temporal derivatives and neighboring frames. We provide a unique binary code coding strategy to address the increase in code length resulting from TPOEM. TPOEM features are computed inside non-overlapped patches of images, and the per-patch scores are averaged to get the final classification. The outcomes showed shorter execution times while exhibiting accuracy on par with or better than cutting-edge methods.

3. PROPOSED METHODOLOGY

This work suggests direct rebuilding of obstructed optical flows and use of generative algorithms for recreating occluded facial expressions where after receiving as inputs optical flows computed between two obscured images, RL produces unobscured flows. To be clear, although facial image reconstruction using generative algorithms has been documented in the literature, our method is the first to handle facial movements to carry out FER in the presence of occlusions [24].

3.1 DATASET COLLECTION

In terms of acquiring datasets, publicly accessible datasets are taken into account for identifying facial expressions, and the literature's current face occlusion techniques are utilized to establish a highly suitable testing strategy. Firstly, publicly accessible datasets pertaining to the recognition of expressions while occlusions are present are aggregated. The suggested technique can be trained on larger datasets thanks to the advantage of having an ensemble of numerous datasets. Second, different areas of the face are covered with occlusions. Normalized optical flows are finally computed.

- ***CK+ database:*** Expanded CK+ is a commonly used database for FERs [25]. 327 image sequences from 118 distinct patients are included in this collection, in addition to FE labels that depend on DFEER. These graphic sequences have the most emotional ending and a neutral beginning. Each subsequent image displays the emotion labels, the FACS code, and the facial landmarks. Seven distinct feelings are categorized by the emotion labels: fear, happiness, sorrow, surprise, contempt, disgust, and rage. The experiment compares this strategy to several approaches based on the six primary expression categories using six emotions (sadness excluded). The images have 640 × 480 and 640 x 490 pixel resolutions, with grayscale values precisiond to 8 bits.
- ***KMU-FED database:*** The standard dataset KMU-FED database for FERs is used which proves that the recommended approach is efficient when driving in the actual world. The dataset was created by using an NIR camera to record regular dataset series while driving in the actual world. KMUFED dataset has driver FEs obtained by NIR cameras mounted on dashboards or steering wheels. They encompass 55 image sequences with varying intensity (front, left, right, and rear light) and possible semi-occlusions of hairs or sunglasses on 12 persons. While analyzing the suggested method cross-validation approach were used on the dataset. Because there are no published findings from past research investigations utilizing the dataset accessed from the web [26].the suggested approach's accuracy values are investigated and evaluated using images resolutions of 1600 × 1200 pixels.

3.2 DATA PREPARATION

The data preparation process relied on publicly accessible FER datasets and published face occlusion techniques to suggest a more appropriate testing regimen. Merging publically accessible datasets pertaining to facial expression identification had occlusions. Our suggested method can be trained on bigger datasets because to the benefit of integrating several datasets. Second, cover various facial areas using occlusions. Lastly, we compute optical flows that are normalized.

3.3 OCCLUSIONS GENERATION

To address the majority of occlusions already examined in the literature [27], occlusions around mouths and eyes are duplicated as seen in Fig.1. Mouths and eyes are crucial components of FER.

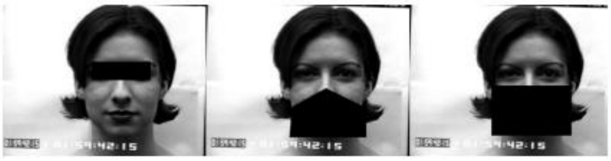

Fig.1. An instance of produced occlusions, utilized in our assessment, applied to a image from the CK+ dataset.

3.4 OPTICAL FLOW CALCULATION

When calculating the optical flow, we aim to preserve as much information as possible with normalized inputs that deep networks require for recognitions and reconstructions. The original images (i.e., keeping their original resolution) are used, and optical flows are immediately scaled. First, process the images at their original resolutions. Cropping faces based on positions of eyes and inter-pupilar distances are second stages. Thirdly, optical flows from clipped faces are computed using DQN. The optical flows from clipped faces are estimated in the third stage using Farneback approach, which is highly helpful for identifying facial expressions and can also be used to compute optical fluxes [28]. The section on assessment discusses how to determine the best parameter size. In real flows, new x and y values are computed using sliding window components. Eq.(1) is utilized to determine the new value for every coordinate (*i*,*j*).

$$resize\left(OF\left(i,j\right)\right)=\begin{pmatrix}\mu_{OF}\left[\left(\lfloor dx*i\rfloor,....,\left(\lfloor dx*i+1\rfloor\lfloor *dx\rfloor-1\right)\right)\right],\\ \mu_{OF}\left[\left(\lfloor dy*i\rfloor,....,\left(\lfloor dy*i+1\rfloor\lfloor *dy\rfloor-1\right)\right)\right]\end{pmatrix} \tag{1}$$

where values for means of optical flows in windows and *dx* and *dy* are coefficients between actual and final sizes (*dx* = *origSizes x*/*finalSizes x* and *dy* = *origSizes y*/*finalSizes y*) where values for *origSizes x* and *finalSizes x*, reflect image's actual sizes and μ stands for final widths.

3.5 REINFORCEMENT LEARNING FOR PARTIAL OCCLUSION SEGMENTATION

RL systems and deep Q-learning are detailed in this section.

3.5.1 Reinforcement Learning:

RL is a well-known machine learning (ML) technique for creating appropriate rules to handle sequential choices by maximizing cumulative reward signals [29]. (S,A,P,R) represent states namely actions, state transition probabilities, and reward functions of RL models. Through contact with its surroundings, the agent learns how to operate in a certain condition to maximize rewards in the future. It is crucial to remember that the agent's short- and long-term profits must be balanced while making judgments [30]. The RL agent may choose the optimal courses of actions for states with highest cumulative rewards based on cumulative learning experiences.

3.5.2 Deep Q-learning:

Q-learning are binary discrete functions depicted by Equation(2):

$$Q(S,A) = f(s,a) \tag{2}$$

The standard Q-learning method works well in low dimensions, but not so well in high dimensional data state spaces. Q-tables are unable to cover states in high-dimensional state spaces due to massive volumes of data, which also significantly increase processing burdens. Therefore, in order to get over the Q-learning algorithm's restriction for high-dimensional problems, the function fitting approach is applied, as shown in Figure 2.

Since neural networks are good at approximating functions, they have been included into the Q-learning processes. On RL tasks, DQN—hybrid neural networks and Q-learning systems have demonstrated competitiveness in their performances. They describe relationships between state action values and updates targets using dual network structures referred to as target networks and Q-networks.

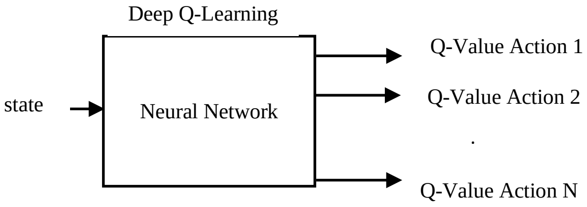

Fig.2. Basic of Deep Q Network

Q-network outputs are solutions to state-value function Q(S, A), and target network outputs act as Q-network labels. Notably, both networks have the same architecture and the parameter update is asynchronous. Cycles modify Q-network's parameters in iterations. When Q-networks modify target networks' settings, target networks do not get modified.

There will always be a correlation between the samples, it should be emphasized. In order to decrease sample correlation and increase sample efficiency, the DQN uses a replay buffer. Stated differently, the samples produced by the agent during its interactions with the environment are stored in a replay buffer. Replay buffer samples are selected in tiny batches for the DQN training procedure.

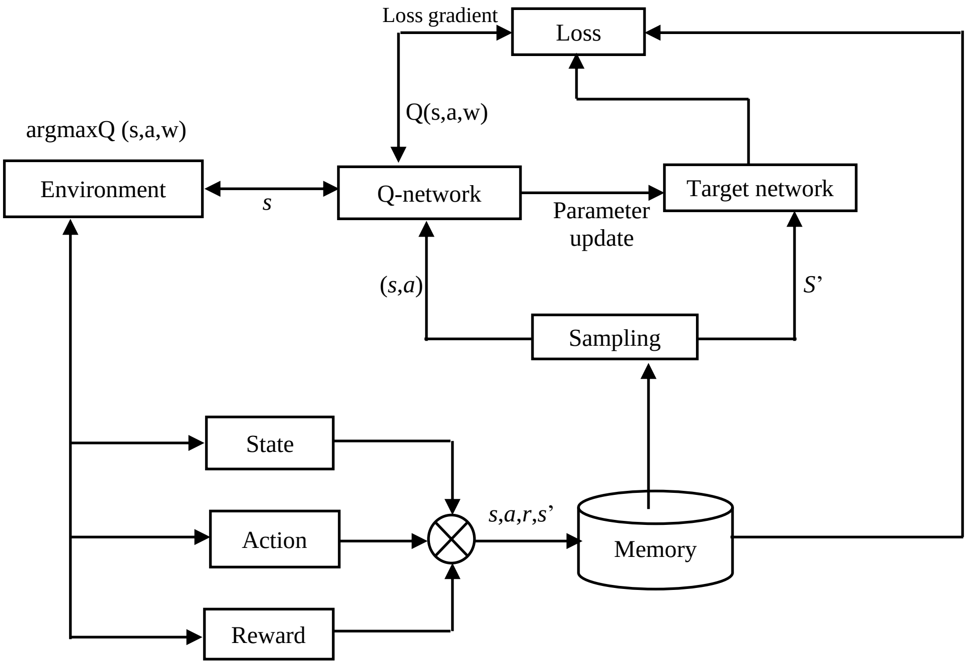

Fig.3. The schematic diagram of DQN

Subsequently stochastic gradient descents update parameters of both target and Q networks. This strategy significantly reduces sample correlations and partially addresses local optimum problems. It should be mentioned that the performance of the DQN method in method 1 is significantly influenced by the network design.

Algorithm 1: Deep- Q Network in Pseudocode

```
The dataset was accessed from the web [26].
Replay memories $D$
Define starting  action value functions $Q$
using randomized weights
Observe initial states $s$
**Repeat**
  Select actions $a$
    With probabilities $\varepsilon$ select random actions
    Otherwise select a=argmax$_{s'}$ $Q(s,a')$
Execute actions a
Observe rewards $r$ and new states $s'$
Store experiences <$s,a,r,s'$> in $D$
Sample random transitions <$ss,aa,rr,ss'$> from replay memories
Calculate targets for transitions of mini batches
If $ss'$ are terminal states then $tt=rr$
Else $tt=rr+y\max_{a'} Q(ss',aa')$
Train $Q$ networks using $(tt-Q(ss,aa))^2$ as losses
 $s=s'$
Until **terminated**
```

The schematic diagram of DQN is represented in Fig.3. The loss function of DQN is shown as follows Eq.(3):

$$Loss(\theta,Q,y)=\frac{1}{2}\left[y(s,a)-Q(s,a,\theta)\right]^2 \tag{3}$$

where $y(s,a)$ represents Q-network labels determined by maximizing values of state-value function and Eq.(4):

$$y(s,a)=r+\max_{a'}Q\left(s',n',\bar{\theta}\right) \tag{4}$$

where $\bar{\theta}$ denotes target network parameters, and is constant fixed during computations of $y(s,a)$.

3.6 BSOA

BSA techniques [32] are the latest population-based evolutionary algorithms. Its basis is an iterative process aimed at minimizing the objective function. The five evolutionary mechanisms that make up BSA are crossover, mutation, initiation, selection-I, and selection-II. The Fig.5 displays the primary BSA flowchart.

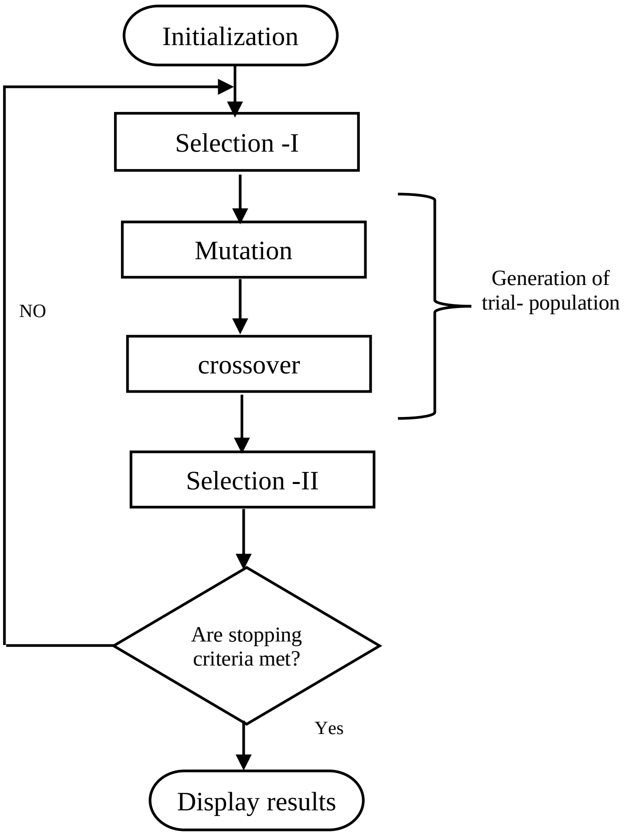

Fig.5. General Flowchart of BSOA

3.6.1 Initialization:

BSA initially scatters populations in solution spaces using uniform random distribution functions as shown in Eq.(5)

$$P_{i,j}\sim(low_j,\ up_j),\quad i=1,2,3,\ldots.,N,\ j=1,2,3,\ldots..,D, \tag{5}$$

where N represents population sizes, D represents population dimensions, U implies uniform distribution functions and P_i implies positions of i^{th} population members in solution spaces low_j, and up_j represent solution space's lower and upper bounds;

3.6.2 Selection-I:

BSOA creates the historical population that is used to establish search directions in Selection-I stages. Historical populations are initialized using Eq.(6) as follows:

$$\text{old}P_{i,j}\sim Ulow_j,up_j) \tag{6}$$

BSOA redesigns oldP_o in iterations using Equation (7)

$$\text{if } a<b \text{ then old } P{:=}P|a,b\sim U(0,1) \tag{7}$$

where a and b are uniform real numbers between [0,1] for selecting *oldP* from previous generations. Eq.(8) shuffles population's members:

$$oldP{:=}permuting(oldP) \tag{8}$$

where the *permuting*() function is a random shuffling function.

3.6.3 Mutation:

BSOA generates mutant members with Eq.(9):

$$Mutant=P+F.(old\text{-}P), \tag{9}$$

where F implies real numbers to control step size amplifications in searches. By taking *oldP* values into account, BSA uses past data to ascertain the population members' search direction.

3.6.4 Crossover:

The BSA's crossover procedure produces the trial-population T's ultimate form. Algorithm 2 provides the BSOA crossover mechanism in pseudocode form. There are two stages in the crossing step. Mixrate is used in the first strategy (Algorithm 2, lines 2-4). In each trial, just one randomly selected individual may mutate under the second approach (Algorithm 2, line 6).

Algorithm 2 Crossover strategy of BSOA

Input: Mutants, mix rates, N and D.

Output: T: Trial-Populations.

```
1:  map$_{(1:N,1:D)}$ =1
2:  If $a<b|a,b\sim U(0,1)$ then
3:  For i from 1 to N do
4:  map$_{i,u(1,\lceil \text{mixrate.end.}D\rceil)}$ = $0|u = permuting((1,2,3,\ldots.,D))$
5:  End
6:  Else
7:  For i from 1 to N do, $map_{i,randi(D)} = 0$,end
8:  End
9:  T: =Mutant
10: For $i$ from 1to $N$ do
11: For $j$ from 1to $D$ do
12: If $map_{i,j}=1$ then $T_{i,j} \coloneqq P_{i,j}$
13: End
14: End
```

Following the crossover procedure, values could go beyond search space limits. As required, the process described in Algorithm 3 is triggered to limit individuals' ranges of motions as the limitation mechanism offers random regenerations in search spaces which cross defined limits,.

Algorithm 3: Boundary control Mechanism of BSOA

Input: T, search space limits

Output: T

```
For $i$ from 1 to $N$ do
For $j$ from 1 to $D$ do
    If $T_{i,j}<low_j$ or $T_{i,j}>up_j$ then
    $T_{i,j}$ = rnd.$up_j$-$low_j$+$low_j$
End
End
End
```

3.6.5 Selection -II:

Greedy selection strategies that employ T_i*s*, with higher fitness than corresponding P_i*s* values, update P_i*s* values in selection-II steps of BSA. When (P_{best}) values are greater than global minimum values, the best individuals of $P(P_{best})$ replace global minimum values.

- **Fitness function**: Many losses functions help find the best reconstructions of optical flows while maintaining FER data.
- **Mean Squared Error (MSE) losses:** The well-known MSE loss function uses AE to lower noise. To calculate the MSE loss, each pixel in the network output is compared to a pixel in the baseline image that has the same coordinates. The

well-known MSE loss function uses an auto-encoder to cut down on noise. To calculate MSE losses, pixels in network outputs are compared to baseline image pixels that have the same coordinates.

- **Wing losses:** The wing loss was first used to complete the landmark localization process [33]. The losses are specifically focused on medium and minor errors enabling incredibly accurate predictions in contrast to employment of MSE losses.
- **Endpoint losses:** When assessing optical flows, the endpoint loss is the final standard error measure to be used [34]. Endpoint errors are widely used metrics to assess deviations between estimated and baseline optical flows.

3.7 DRIVER FACIAL EXPRESSION EMOTION RECOGNITION (DFEER)

Further non-linear combination results are made possible by the feature fusion layer of the FER approach, which is capable of performing non-linear fusion of the feature information produced by the VGG16 network.

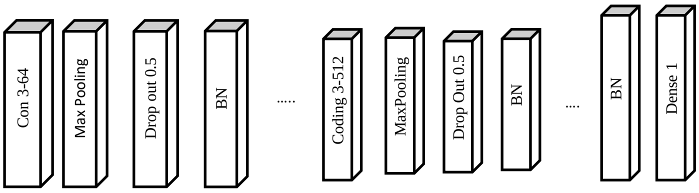

Fig.6. Improved VGG16 Model

Fig.6 displays a more thorough model diagram of the VGG16 network that was developed for this study. Batch normalization and dropout are the two new layers added to the VGG16 network model. Three completely linked layers, two sets of standardized layers, and global pooling layers comprise feature fusion layers [35]. Convolution kernels of network models are 64, 128, 256, and 512 [36], while kernels of feature fusion layers are 1024, 512, and 1.

Node hiding rate parameters of Dropout layers are typically set to a value between 0.3 and 0.7 when the process is finished. The expression for the completely linked layer is given by Eq.(10).After this study conducted several tests, it was eventually determined that the ideal concealment rate was 0.5 [37]. Essential feature fusion layer settings are modified using the BP (backpropagation method) after every model training iteration is depicted below:

$$Y_{out}=f(WX_{in}+b) \tag{10}$$

where W denotes training weights, $f(x)$ represents activation functions ReLU, and Y denotes outputs obtained from fully connected layers and X denotes inputs sent to fully connected layers in Eq.(11) [38]:

$$f(x)=\begin{cases}0 & x<0\\ x & x>0\end{cases} \tag{11}$$

ReLU function's demonstration of unilateral inhibition meets loss function's conditions in this experiment. In training of neural networks, gradient explosions or disappearances commonly happen, which complicates the process. Training will become harder and convergence slower as the network gets deeper. Therefore, during the DL process, BN layers maintain uniform distributions of inputs across all neural network layers.

After a few layers of feature extraction in the DL network, the Dropout layer randomly eliminates a portion of the neurons, preventing overfitting [39]. This prevents overfitting. A well-liked ML approach called L2 regularisation seeks to limit overfitting and regulate model complexity. Because of the fundamental regularisation technique [40], the genuine goal function may be penalized for the model's increased complexity by inserting a penalty term.

4. EXPERIMENTAL EVALUATIONS

DFEER performance in image sequences is investigated for detailed and comparative analysis using several datasets. The two FE-related databases with the greatest number of images frequently utilized in DFEER-relevant analysis are CK+ and KMU-FED. This effort's goal is to ascertain the driver's FEs, in contrast to earlier investigations. Using a near-infrared (NIR) camera, the KMU-FED image database is created in a driving environment, capturing the driver's FE in a realistic driving scenario. Every experiment is done on an Intel Core i7 CPU with 8 GB of RAM and Microsoft Windows 10. There are 100 iterations of this assessment. The Equation from (12–16) represents the evaluation metrics.

4.1 EVALUATION METRICS

- ***Precision:*** Precision values are positive class prediction counts, which are truly from positive classes and formulated as,

$$Precision=TP/(FP+TP) \tag{12}$$

- ***Recall:*** Recall values provide measures of positive class prediction counts obtained from all positive samples and expressed as,

$$Recall=TP/(FN+TP) \tag{13}$$

- ***F1-Score:*** F1-Scores balance both issues of precision and recall values in one number and expressed as,

$$F1\text{-}Score=2\times(Precision\times Recall)/(Precision+Recall) \tag{14}$$

- ***Accuracy:*** Accuracy values refer to ratios of examples, which were rightly classified i.e., ratios between sums of true positive and true negative counts, and counts of samples present in datasets and formulated as,

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \tag{15}$$

- ***Error:*** Errors are ratios of wrongly classified samples and computed as,

$$Error=1\text{-}Accuracy \tag{16}$$

3.1. *Reconstruction Methods Comparison*

The results of the accuracy and error of expression identification on reconstructed optical flows for different occlusions are displayed in Table.1 and Table.2, which contrast the suggested method with other existing methods like AE and PMVO of the CK+ dataset. Those tables can be used to find it; the suggested method is quite flexible in this situation. The endpoint loss is fixed in the experiments that follow, and only findings with this loss are shared. The suggested method decreases error for MSE, Wing, and Endpoint loss functions for Eyes occlusion by

3.58%, 5.44%, and 6.28%, respectively, when compared to all occlusions.

Table.1. Reconstructed optical flow accuracies vs. differential losses for (CK+ Dataset)

Loss Functions	Eyes occ. (%)			occ. (%) of Mouths			occ. (%) of Lower parts		
	AE	PMVO	DQN-BSOA	AE	PMVO	DQN-BSOA	AE	PMVO	DQN-BSOA
MSE	87.63	93.62	96.42	76.00	82.36	86.45	69.20	75.36	79.21
Wing	86.52	92.33	94.56	80.50	85.21	87.51	70.82	79.21	82.41
EndPoint	88.63	91.82	93.72	80.63	88.12	90.10	71.26	82.83	87.64

Table.2. Reconstructed optical flow errors vs. differential losses for (CK+ Dataset)

Loss Functions	Eyes occ. (%)			occ. (%) of Mouths			occ. (%) of Lower parts		
	AE	PMVO	DQN-BSOA	AE	PMVO	DQN-BSOA	Flownet	PMVO	DQN-BSOA
MSE	12.37	6.38	3.58	24.00	17.64	13.55	30.80	24.64	20.79
Wing	13.48	7.67	5.44	19.50	14.79	12.49	29.18	20.79	17.59
EndPoint	11.37	8.18	6.28	19.37	11.88	9.90	28.74	17.17	12.36

Table.3. Reconstructed Optical Flow Accuracies vs. Differential Losses for (KMU-FED Dataset)

Loss Function	Eyes occ. (%)			occ. (%) of Mouths			occ. (%) of Lower parts		
	AE	PMVO	DQN-BSOA	AE	PMVO	DQN-BSOA	AE	PMVO	DQN-BSOA
MSE	86.48	92.28	94.35	74.92	80.94	82.84	67.45	74.34	77.21
Wing	85.45	91.18	93.21	79.35	83.81	87.71	69.64	78.00	81.70
EndPoint	87.42	90.64	92.15	79.31	86.87	89.71	70.05	81.91	83.41

Table.4. Reconstructed Optical Flow Accuracies vs. Differential Losses for Errors (KMU-FED Dataset)

Loss Function	Eyes occ. (%)			occ. (%) of Mouths			occ. (%) of Lower parts		
	AE	PMVO	DQN-BSOA	AE	PMVO	DQN-BSOA	AE	PMVO	DQN-BSOA
MSE	13.52	7.72	5.65	25.08	19.06	17.16	32.55	25.66	22.79
Wing	14.55	8.82	6.79	20.65	16.19	12.29	30.36	22.00	18.30
EndPoint	12.58	9.36	7.85	20.69	13.13	10.29	29.29	18.09	16.59

The Table.1 display the accuracy results of comparing three distinct occlusions using three different reconstruction techniques and loss functions (CK+ dataset). The results show that the suggested DQN-BSOA rebuilt technique improves end point loss function accuracy by 93.72%, 90.10%, and 87.64% for lower part, mouth, and eye occlusions.

The suggested DQN-BSOA rebuilt procedures show the accuracy gains of 96.42%, 94.56%, and 93.72% for MSE, Wing, and end point loss function with eye occlusion, as shown in Fig.7(a). When occlusion of the eyes occurs, MSE loses function Fig.7(c) demonstrates that the recommended DQN-BSOA rebuilt approach has a greater accuracy of 96.42%, while other strategies like AE and PMVO have lesser accuracy of 86.45% and 79.21% (See Table 1).

The evaluation of expression recognition accuracy and error on recovered optical flows for different occlusions is presented in Table.3 and Table.4, along with a comparison with the findings of other methods that are currently accessible, including PMVO and AE of the KMU-FED dataset. These tables show how flexible the suggested approach is in this particular situation. Only the results with this loss are presented in the tests that follow since the endpoint loss is repaired. The suggested approach yields an error reduction of 5.65%, 6.79%, and 7.85% for MSE, Wing, and Endpoint loss function for Eyes occlusion when compared to all occlusion.

Table.3 display obtained accuracies of comparisons from three different occlusions using three different reconstruction techniques and loss functions (KMU-FED dataset). The results show that the end point loss function with lower part occlusions is more accurate when using the suggested DQN-BSOA rebuilt technique by 83.41%, 89.71%, and 94.35%, respectively. The Table.3 shows that the suggested DQN-BSOA rebuilt technique improves accuracy by 86.48%, 74.92%, and 67.45%, respectively, for MSE, Wing, and end point loss function with eye occlusion.

The accuracy of the MSE loss function with eyes occlusion is improved to 93.21% by the suggested DQN-BSOA rebuilt approach, as shown in The Table.3. In contrast, strategies like AE and PMOV provide lesser accuracy, at 87.71% and 81.7%, respectively (Table.3).

4.2 RECOGNITION METHODS COMPARISON

The analysis compares the results with methods like Weighted Random Forest (WRF), CNN, Inception-ResNet+LSTM, FERDERnet, and VGGNet using precision, recall, F1-score, accuracy, and error.

Table.5. Comparative Results of CK+Dataset

Methods	Precision	Recall	F1-Score	Accuracy	Error
CNN	85.19	87.14	86.16	89.46	10.54
Inception-ResNet+LSTM	87.28	88.26	87.77	91.72	8.28
WRF	89.82	91.45	90.63	92.18	7.82
FERDERnet	92.71	92.41	92.56	93.47	6.53
VGGNet	93.57	93.32	94.60	94.78	5.22

The Table.5 evaluates the performance of classifiers using metrics like precision, recall, F1-score, and accuracy, including WRF, CNN, Inception-ResNet+LSTM, FERDER net, and VGGNet (CK+ dataset). Based on the results, the recommended VGGNet classifier has the best accuracy (94.78%), whereas the accuracy of the other classifiers is 93.47%, 91.72%, 92.18%, 89.46%, and 94.78% (see Table 5).

Table.6. Comparative Results for KMU-FED Dataset

Methods	Precision	Recall	F1-Score	Accuracy	Error
CNN	83.25	85.41	84.33	88.15	11.85
Inception-ResNet+LSTM	84.62	87.16	85.89	92.45	7.55
WRF	88.18	90.40	89.29	93.21	6.79
FERDERnet	90.21	91.82	91.01	94.49	5.51
VGGNet	92.54	94.42	93.48	95.92	4.08

The Table.6 shows the outcomes of comparing the performance of several classifiers, including CNN, WRF, FERDER net, VGGNet, Inception-ResNet+LSTM, and Inception-ResNet. These measures include accuracy (KMU-FED dataset), precision, recall, and F1-score. Table 6 shows that the recommended VGGNet classifier produces the highest accuracy of 95.92%, whereas the other classifiers only yield 85.15%, 87.45%, and 90.21%, respectively.

5. CONCLUSION AND FUTURE WORK

A novel method is suggested for the purpose of semi-facial occlusion FER where optical flows for the identification stage are rebuilt first. This work’s Driver Facial Expression Emotion Recognition (DFEER) technique involves rebuilding directly computed optical flows that have been influenced by semi-occlusion.

The suggested method uses the reconstructed optical fluxes, which were initially computed directly before being harmed by a partial blockage, in the recognition stage. To do that, a backtracking search optimization algorithm with optical flow reconstruction and a DQN for partial occlusion segmentation were developed for on-road driver facial expression emotion identification. DQN is presented for semantic segmentation based on segmenting occlusions in the first stage. With pixel-level annotations, it was specifically created for the fully supervised semantic segmentation problem and incorporated partial class activation attention.

The second stage involves simulating the optical flows broken by occlusion using a trained BSA. After estimating the optical flow from two frames of an obstructed face, the reconstruction uses an AE to restore the optical flow of the occluded parts. The recognition system then receives the recovered optical flow and utilizes it as an input to forecast the expression class.

Using a feature fusion layer created by the VGGNet method, results may be obtained by nonlinearly combining feature data. Since the proposed network seeks to quickly and accurately identify the emotions on drivers’ faces, recognition speed is essential. It is evident from a comparison of the results and benchmark findings that the suggested technique yields improved accuracy. The suggested VGGNet classifier was evaluated for precisions, recalls, F1-scores, and accuracies using the CK+ and KMU-FED datasets, along with other methods. Future plans are examining occlusions induced by variances in head position. In this case, in addition to the loss of mobility caused by the blockage, the head also produces loud movement.

REFERENCES

[1] F. An and Z. Liu, “Facial Expression Recognition Algorithm based on Parameter Adaptive Initialization of CNN and LSTM”, *The Visual Computer*, Vol. 36, No. 3, pp. 483-498, 2020.

[2] J. Serrano-Puche, “Affect and the Expression of Emotions on the Internet: An Overview of Current Research”, *Second International Handbook of Internet Research*, pp. 529-547, 2020.

[3] O. Arriaga, M. Valdenegro-Toro and P. Ploger, “Real-Time Convolutional Neural Networks for Emotion and Gender Classification”, *Proceedings of International Conference on Machine Learning*, pp. 1-7, 2017.

[4] T. Küntzler, T.T.A. Höfling and G.W. Alpers, “Automatic Facial Expression Recognition in Standardized and Non-Standardized Emotional Expressions”, *Frontiers in Psychology*, Vol. 12, pp. 1-6, 2021.

[5] H. Yang, U. Ciftci and L. Yin, 2018, “Facial Expression Recognition by De-Expression Residue Learning”, *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 2168-2177, 2018.

[6] S. Hickson, N. Dufour, A. Sud, V. Kwatra and I. Essa, “Eyemotion: Classifying Facial Expressions in VR using Eye-Tracking Cameras”, *Proceedings of International Conference on Winter Applications of Computer Vision*, pp. 1626-1635, 2019.

[7] C.H. Chen, I.J. Lee and L.Y. Lin, "Augmented Reality-based Self-Facial Modeling to Promote the Emotional Expression and Social Skills of Adolescents with Autism Spectrum Disorders", *Research in Developmental Disabilities*, Vol. 36, pp. 396-403, 2015.

[8] H. Yang, U. Ciftci and L. Yin, "Facial Expression Recognition by De-Expression Residue Learning", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pp. 2168-2177, 2018.

[9] J.H. Kim, B.G. Kim, P.P. Roy and D.M. Jeong, "Efficient Facial Expression Recognition Algorithm based on Hierarchical Deep Neural Network Structure", *IEEE Access*, Vol. 7, pp. 41273-41285, 2019.

[10] F. Kong, "Facial Expression Recognition Method based on Deep Convolutional Neural Network Combined with Improved LBP Features", *Personal and Ubiquitous Computing*, Vol. 23, No. 3-4, pp. 531-539, 2019.

[11] S. Fobi, T. Conlon, J. Taneja and V. Modi, "Learning to Segment from Misaligned and Partial Labels", *Proceedings of the ACM SIGCAS Conference on Computing and Sustainable Societies*, pp. 286-290, 2020.

[12] B. Nakisa, M.N. Rastgoo, A. Rakotonirainy, F. Maire and V. Chandran, "Automatic Emotion Recognition using Temporal Multimodal Deep Learning", *IEEE Access*, Vol. 8, pp.225463-225474, 2020.

[13] S.A. Liu, H. Xie, H. Xu, Y. Zhang and Q. Tian, "Partial Class Activation Attention for Semantic Segmentation", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pp. 16836-16845, 2022.

[14] C. Lu, W. Zheng, C. Li, C. Tang, S. Liu, S. Yan and Y. Zong, "Multiple Spatio-Temporal Feature Learning for Video-based Emotion Recognition in the Wild", *Proceedings of International Conference on Multimodal Interaction*, pp. 646-652.2018.

[15] C.Y. Siu, H.L. Chan and R.L. Ming Lui, "Image Segmentation with Partial Convexity Shape Prior using Discrete Conformality Structures", *SIAM Journal on Imaging Sciences*, Vol. 13, No. 4, pp.2105-2139, 2020.

[16] W.Y. Chung and B.G. Lee, "Methods to Detect and Reduce Driver Stress: A Review", *International Journal of Automotive Technology*, Vol. 20, pp. 1051-1063, 2019.

[17] W.Y. Chang and J.H. Chien, "FATAUVA-Net: An Integrated Deep Learning Framework for Facial Attribute Recognition, Action Unit Detection, and Valence-Arousal Estimation", *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 17-25, 2017.

[18] R. Theagarajan and A. Cruz, "Deepdriver: Automated System for Measuring Valence and Arousal in Car Driver Videos", *Proceedings of IEEE International Conference on Pattern Recognition*, pp. 2546-2551, 2018.

[19] R. Guo and H. Kang, "Image Segmentation Algorithm based on Partial Differential Equation", *Journal of Intelligent and Fuzzy Systems*, Vol. 38, No. 4, pp. 3903-3909, 2020.

[20] Y.K. Bhatti and S.A. Velastin, "Facial Expression Recognition of Instructor using Deep Features and Extreme Learning Machine", *Computational Intelligence and Neuroscience*, Vol. 2021, pp.1-17, 2021.

[21] X. Zhu, X. Li and S. Zhang, "Block-Row Sparse Multiview Multilabel Learning for Image Classification", IEEE Transactions on Cybernetics, Vol. 46, No. 2, pp. 450-461, 2015.

[22] F. Al Machot and K. Kyamakya, "A Deep-Learning Model for Subject-Independent Human Emotion Recognition using Electrodermal Activity Sensors", Sensors, Vol. 19, No. 7, pp. 1659-1667, 2019.

[23] E.A.S. Cruz and C.H.E. Franco, "Facial Expression Recognition using Temporal POEM Features", *Pattern Recognition Letters*, Vol. 114, pp. 13-21, 2018.

[24] Y. Cui, S. Wang and L. Jiao, "Remote Sensing Object Tracking with Deep Reinforcement Learning under Occlusion", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp. 1-13, 2021.

[25] L.F. Barrett, S. Marsella, A.M. Martinez and S.D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements", *Psychological Science in the Public Interest*, Vol. 20, No. 1, pp. 1-68, 2019.

[26] L. Zhang, B. Verma and V. Chandran, "Facial Expression Analysis under Partial Occlusion: A Survey", ACM Computing Surveys, Vol. 51, No. 2, pp. 1-49, 2018.

[27] B. Allaert, C. Djeraba and M. Bennamoun, "Optical Flow Techniques for Facial Expression Analysis-A Practical Evaluation Study", *Proceedings of International Conference on Machine and Deep Learning*, pp. 1-12, 2019.

[28] N. Zeng and X. Liu, "Deep-Reinforcement-Learning-based Images Segmentation for Quantitative Analysis of Gold Immunochromatographic Strip", *Neurocomputing*, Vol. 425, pp. 173-180, 2021.

[29] E. Mocanu, M.E. Webber and J.G. Slootweg, "On-Line Building Energy Optimization using Deep Reinforcement Learning", *IEEE Transactions on Smart Grid*, Vol.10, No. 4, pp. 3698-3708, 2018.

[30] A. Gosavi, "*Simulation-Based Optimization*", Springer, 2015.

[31] V. Mnih and S. Petersen, "Human-Level Control through Deep Reinforcement Learning", *Nature*, Vol. 518, pp. 529-533, 2015.

[32] M.G. Asogbon, A.A. Miller, G. Li and K.K. Wong, "GBRAMP: A Generalized Backtracking Regularized Adaptive Matching Pursuit Algorithm for Signal Reconstruction", *Computers and Electrical Engineering*, Vol. 92, pp. 107189-107194, 2021.

[33] M. Cicero, A. Bilbily, B. Gray and J. Barfett, "Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs". *Investigative Radiology*, Vol. 52, No. 5, pp. 281-287, 2017.

[34] S. Jun and N. Kim, "Development of a Computer-Aided Differential Diagnosis System to Distinguish between Usual Interstitial Pneumonia and Non-Specific Interstitial Pneumonia using Texture-and Shape-based Hierarchical Classifiers on HRCT Images", *Journal of Digital Imaging*, Vol. 31, No. 2, pp. 235-244, 2018.

[35] F.L. Da Silva and A.H.R. Costa, "A Survey on Transfer Learning for Multiagent Reinforcement Learning Systems", *Journal of Artificial Intelligence Research*, Vol. 64, pp. 645-703, 2019.

[36] C. Szegedy, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions", *Proceedings of IEEE*

International Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

[37] H. Wang, “Garbage Recognition and Classification System based on Convolutional Neural Network VGG16”, *Proceedings of IEEE International Conference on Advanced Electronic Materials, Computers and Software Engineering*, pp. 252-255, 2020.

[38] M. Jeong and B.C. Ko, “Driver’s Facial Expression Recognition in Real-Time for Safe Driving”, *Sensors*, Vol. 18, No. 12, pp. 1-17, 2018.

[39] D. Poux, C. Djeraba and M. Bennamoun, “Dynamic Facial Expression Recognition under Partial Occlusion with Optical Flow Reconstruction”, *IEEE Transactions on Image Processing*, Vol. 31, pp. 446-457, 2021.

[40] B. Hasani and M.H. Mahoor, “Facial Expression Recognition using Enhanced Deep 3D Convolutional Neural Networks”, *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 30-40, 2017.