Research Article

Age, Gender, and Emotion Recognition based Deep learning models

Mithaq Nazar Jassim[1], Asst. Prof. Dr. Ali Hussainmary[2]

[1]Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers and Informatics, Baghdad, Iraq
[2]Mechatronics Engineering Department, Al-Khwarizmi College of Engineering, University of Baghdad, Iraq.
[2]Alimary76@kecbu.uobaghdad.edu.iq

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: A deep learning approach was proposed in this study for estimating age, emotion expression, and gender from a real-time video source without using facial landmarks or other geometric calculations of café features. For image classification, a convolutional neural networks (CNNs) pre-trained and used on ImageNet from Caffe (Convolutional Architecture for Quick Feature Embedding), a modifiable platform for state-of-the-art deep learning algorithms and a set of reference models. The (you only look once v3) YOLOv3 algorithm was employed for such purposes having a desirable abilities to serve the required purpose. Deep Convolutional Neural Network (DCNN) features are used to propose a framework for automatically understanding facial expressions. The suggested model focuses on understanding an individual's facial expressions from a single image.

1. Introduction

Over the past decade, the computer vision research community showed great interest in the analysis and automatic recognition of facial expressions. Most of these facial expression analysis systems attempt to classify expressions into a few broad emotional categories, such as happy, sad, angry, surprised, fear and disgust.

Caruana [1] was the first to investigate multi-task learning in depth. Since then, multi-task learning has been used in a variety of approaches to solve a variety of computer vision problems. [2] Suggested an early technique for studying human-face recognition, locality of facial landmarks, and estimation of face all together, which has been later generalized to [3]. Levi et al. [4] suggested using a CNN to estimate age and gender at the same time. By merging the middle layers of CNNs for better extraction of features, HyperFace [5] educated a multi-function learning network for face recognition, pose, landmark locality, and gender estimation. Ehrlich et al. [6] suggested a time limited multitasking to study facial characteristics, a Boltzmann computer is used, and on the other hand Zhang et al. [7] increased landmarks localization by simultaneously Head-pose prediction and facial feature inference are two techniques that can be used to train it. While these approaches can perform multitask learning on a limited number of tasks, they cannot perform multitask learning on a large number of tasks.

As part of facial trait inference, the gender and smile identification tasks from unconstrained photographs been taken into consideration. Liu et al. [12: 8] recently published the CelebA dataset, which contains approximately 200000 near-frontal photographs with 40 attributes such as gender and facial expressions, accelerating study in this area [9][5][10][11]. The role of Face Verification (which is not an interest in the proposed system) is to determine if two faces belong to the same individual. By using millions of annotated data to train deep CNN models, DeepFace [12], Facenet [13], and VGG-Face [14] are examples of recent approaches that have been dramatically improved the accuracy of verification on the LFW [15] dataset.

2. Convolutional Neural Networks

A convolutional neural network according to [16] is a type of multilayer network that consists of several alternating convolutional and pooling (subsampling) layers, and at the end it has a series of full-connected layers as a multilayer perceptron network, as shown in Figure 1.The input of a convolutional layer network is usually an image $m \ast m \ast x \ast r$, where m is height and width of the image together and r is the number of channels. Convolutional layers have 1 kernels (or filters) whose dimensions are $n \ast n \ast q$, where n and q are specified by the designer (where generally *q is equal to r*). Each filter has a size of $(m - n + 1) \ast (m - n + 1) \ast q$ which generated by convolution map of features or characteristics, where *p is* the number of filters to use.

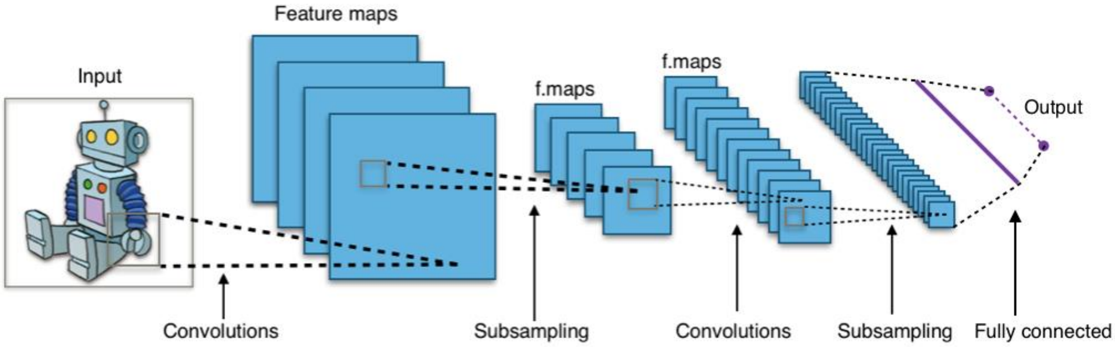


Figure 1Schematic of a convolutional neural network.

Then each map is further sampled in the pooling layer with the mean pooling or maximum pooling operation on contiguous regions of size $p * p$ where p can take values from 2 to 5 depending on the size of images (small to large respectively). Before or after subsampling, a sigmoidal activation function is applied with addition to a bias for each feature map.FasterRCNN is a network that does object detection. As its name explains, it is faster than its RCNN and FastRCNN descendants. This network has use cases in self-driving cars, manufacturing, safety, and is even used on Pinterest.The algorithm passes the image through a CNN to obtain a feature map. It runs the activation map through a separate network (RPN), which produces regions of interest. For each RPN region, several layers fully connected to the outputs are used together with the link box coordinate [16].The Fast R-CNN architecture essentially reduces the computational load, with respect to CNN, and for this reason decreases the detection time presented by the R-CNN layer

3. You only look once

YOLO consists of a single conventional network that simultaneously predicts multiple bounding boxes and class probabilities for those boxes. It is trained with full set of pictures. It is extremely fast compared to other traditional methods, getting approximately 45 fps, in its first version (processed on a Titan X GPU) and in a faster version more than 150 fps. This means that it would be possible to process the video transmission in real time with less than 25 milliseconds of latency [9] with optimal computational performance. Yolo also achieves twice the accuracy of the average of other real-time systems. All training and testing code is open source, and there are a variety of pre-trained models that can be downloaded.

4. Data pre processing

The majority of datasets used in this work show images of faces which are processed and aligned in real-world applications. A well-positioned face image should be the same size or approximately the same, and positioned in the middle of the display with little or no background. The off-the-shelf face detector is the preferred method to calculate and record the face location and size (scale) in each image.The advantage of cropping prior to submitting the image to recognition using face detection in order to calculate age gives a massive boost to performance. Some preprocessing have to be done on the images to ensure better results from the proposed system, these preprocessing include image resize where the video captured image is resized to a smaller size having only meaningful data 63including ROI (region of interest). In Face Detection image resized to 448x448 based on the following Matlab code.

```
.Algorithm 1 : Image resize

Step 1 Obtain image from Real-time stream
Step 2: find Outimage(x,y)=$\sum_{i=1}^{2}\sum_{i=1}^{2} aijx^{i-1}y^{i-1}$
Step 3: Move to next x,y in image
Step 4: go to step 2
```

This algorithm provides a high speed resize rate time which is about 4.8ms per 128x128 images and considered one of the fastest nonlinear resize algorithm which is used by Matlab imresize function.For the gender, age and emotion detection NNs image are resized to 224x224 using same previously explained algorithm. Since the size is lower which provided a higher resize rate, finally for emotion detection image resized to 64x64 also using the imresize algorithm to obtain the image of interest to be passed to the next step. Furthermore, image grayscale is also used to reduce the number of color data in the image to further speedup performance. Using im2gray Matlab built-in function provided a convenient tool for such purpose working within provided time constrains.

5. Network Configuration and Architecture

A professional CNN as a facial feature extractor was used as a starting point for the purpose of training a CNN to predict age from face images may be a helpful way to practice. The CNN network that is proposed in such architecture is built on a deep face recognition CNN that can remove distinct and consistent facial features. It would also be less subjected to over recognition.Using a small number of CNN models, face recognition has been successfully trained. The VGG-face model proposed by [17] was used in this study, and it generated outstanding outcome on the LFW [18] and for YFT [19] datasets. VGG-Face has eleven layers, eight of which are convolutional and three of which are fully related. Every convolutional layer is preceded by a rectification layer, and each convolutional block ends with a max pool layer,The completely linked layers are a kind of convolutional layer in which the filter and input data sizes are all the same. For the first two completely connected layers, the number of input features is 4096.

An N-way unit predictor is represented by the model's final completely connected sheet, where N is the database's number of marks (classes). In this article, we architect and retrain the VGG-Face model for age estimation, keeping the convolutional layers but replacing the fully connected layers with four new fully connected layers. Dropout and rectification layers match the first three totally connected layers. The first fully connected layer has a size of 4096 bytes, while the second and third fully connected layers have a size of 5000 bytes. The output layer's output size corresponds to the number of age marks, which is eight.

6. Expression Extraction

Caffe on Graphics Processing Unit (GPU) was used to extract features. To detect the ImageNet 15 object convolution neural network architecture was employed for the extraction of facial features. In the ImageNet, there were eight learned layers with five convolutions and three completelinked layers to detect objects. This study usesonly themain five layers for the extractions. The layers combine convolution and Rectified Linear

Algorithm 2 Algorithm for recognizing facial expressions using CNN

```
for all image (i), facial expressions ∈ face dataset do alter the image (i) to gray-scale
for the detection of the frontal face in (i) and crop only the face (c)extract POOL5 (256 × 6 × 6) features. This is for the cropped face (c) by CNN
copy which is predefined facial expression in any image(i) in an input dataset.
It uses POOL5 vector for the dimension 9216(256 × 6 × 6) in tenfold and leave-one-out cross validationwith SVM classifier for the recognition of facial expressions.
```

6.1 Validation

For each dataset processed and trained in the system a 20% quantity of this data are used as validation data to ensure accurate training and quantified outcome.

The Architecture of CNN shown in Fig. 3.5 a pre-trained face recognition network is used from Sankayan et al. (2013) [20] to provide initial training for the algorithm. The CNN's discriminative face feature filters provide a better base to begin a generic face analysis, these activities are divided into two categories:

1) Face detection, main point localization and visualization, pose estimation, and smile detection are all part of this category.

2) Specific operations: age and gender detection and face feature recognition are all examples of specific operations. The third, and fifth convolutional layers are fused with the first, for training the subject-independent tasks, similar to HyperFace [21], since they focus more on local knowledge accessible from the network's lower layers. To achieve a compatible function map size of 6 x 6, we apply to these basic layers, there are a pooling layer and two convolution layers, respectively. To reduce the number of function maps to 256, a dimensionality reduction layer is applied. It is preceded by a 2048-dimensional completely connected layer that serves as a generalized representation for activities that are not dependant on the subject. The basic tasks are then branched into 512-layer fully connected layers, which are then followed by output layers.

6.2 Training

Input images are then arbitrarily trimmed into 224x224 patches of pixels after being rescaled to 256 x 256 pixels. With 256-piece mini-batches and a momentum value of 0.9, the proposed network is optimized using the stochastic gradient descent technique. In addition, weight decay is also fixed for a value of 10-3. The network parameters are regularized using a 0.6 dropout rate during the training process. When the validation set consistency result does not improve, the training rate is lowered by a factor of ten. The biases are set to zero. Also, the weights in the newly suppliedcompletely linked layers are initialized by Gaussian distribution with a negative mean. It uses a 10 to 2 standard deviation. Within the network's input layer,the RGB input image is inserted. Also, the output of each hidden layer is input to the next layer before the network's output layer (last layer) is determined.

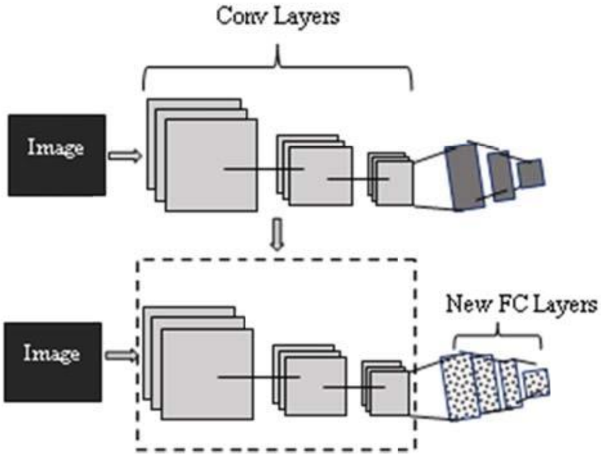


Figure 2 Convolutional model

As seen in Fig. 2, the stochastic gradient decent approach optimizes and seeks the layer parameters mitigating the softmax-log-loss prediction to estimate age. Today, there is no update for the parameters of the convolutional layers which remained frozen. To put it another way, we optimize the parameters of the completely linked layers to find the age of people while depositing the results found in the convolutional layers alone, which were optimized and trained for the face recognition task by [17]. The training CNN model is divided into five sub-networks, each with its own set of parameters. Face recognition, since they both use the same dataset, key-point localization and visibility, as well as pose estimation, are all trained in the same sub-network (AFLW [11]). Separate sub-networks are used to train for the gender identification, expression recognition and estimation of age. These sub-networks are intergraded within into one comprehensive CNN during testing. Caffe [19] is used to train all activities concurrently from beginning to end. Here are the failure roles and training datasets for each.

1) Localization, detecting face, and estimating pose: using AFLW [11] dataset these tasks were learned in a similar way to HyperFace [21]. We choose 1000 images at random from the dataset for processing, and the rest of images are used for merely preparation. To generate area suggestions for faces included an image, we use the Selective Search [22] algorithm. Good examples are regions with a join relationship to overlap of more than 0.5 probability of similarity in the bounding box, while negative examples are regions with a join relationship of less than 0.35 to train the face detecting based on asoftmax loss feature.

2) Gender Recognition: Similar to face identification, gender detection acts as binary classification problem. The datasets that were used to train gender as Table 3.2. The facial key-points in the dataset or HyperFace [21] are used to fit the training images first. As seen in equation, a cross entropy loss LG is used for equation (3.1)

"$LG = -(1 - g) \cdot \log(1 - pg) - g \cdot \log(pg)$…………………..(3.1)"

Where g is equal to 0 or 1 for men and women respectively, the input face is regarded as a women then approximate likelihood found in pg.

3) Expression Detection: For face expression recognition, the happy, sad, angry, surprised, and disgust features are studied for making the network resist expression variations. In the preparation stage, the CelebA [23] dataset was used. Before sending the photos across the network, they are synchronized in the same was as the gender recognition task. The loss function LS is defined as follows: (5)

$LS = -(1 - s) \cdot \log(1 - ps) - s \cdot \log(ps)$ ……………..(3.1)

Where s is 1 if the face is laughing and 0 if it is not. The estimated probability is given as the input face is smiling by ps.

4) Age Estimation: This problem is conceived as a regression issuebecause network have need to be familiar with the prediction of an individual's age from a face image. The databases IMDB+WIKI [24] are used for planning, Adience [19], and MORPH [25]. Where the standard deviation of age is given, Ranjan et al. [26] demonstrate that Gaussian loss performs better than Euclidean loss for estimating apparent age. When the expected age is far from the actual age (Fig. 4), the gradient of Gaussian failure is close to zero, slowing the training phase. As a result, as seen in, we use a linear combination of these two loss functions weighted by (6)

$$L_A = (1-\lambda)\frac{1}{2}(y-a)^2 + \lambda\left(1 - exp(-\frac{(y-a)^2}{2\sigma^2})\right), \quad (6)$$

Where the value of LA represents the loss of age amount, *y* is the expected age, in addition *a* is the ground-truth age, and is the annotated age value's standard deviation is represented by σ. ƛ is set to 0 at the start of the training and incremented by 1 over time. In our implementation, we start with a value of 0 and then change it to 1 after 20k iterations. If the training range does not have a value, it is set to 3.

6.3 Testing

During analysis, we employ a two-stage approach, as seen in Figure 3.6. From a test picture, Selective Search [22] generates area proposals, which are then processed by the proposed comprehensive neural network for detecting scores. The pose of the head detection, orientation and visibility for which an iterative Region Proposals and Landmarks-based NMS for scaning out non-faces was used. This could also be helpful for the improvement of the orientation and detectingpose [21].The collected fiducial points are then used in the second stage to match and observed face to a valid sceneby the resemblance transform. The emotion, gender, age, identification details are obtained by passing the matched faces, as well as their flipped copies, through the network once more.

7. Simulation results

Table 1 consists of a performances of classification with the methods ofstate of the art in the world for CAFFE database. For CAFFE databases, most errors obtained as 96.10 and 100 ordered from top1 through top 5, respectively. However in Table 2 and 3 the captured values of gender, age and emotion detection for 8 people (4 males and 4 Females) showing 6 different types of emotions namely (Angry, Fear, Happy, Sad, Surprise, Neutral). Some of the results are 94% to 96% accurate however some results will be recognized wrongfully due to some reasons such as covering the face and having a beard and a mustache, some of the mouth and cheeks features affect the result of the detection process, in addition there are some inherited problems from the face detection algorithm accuracy that affect the overall success rate of the detection process. In table 4.2 a set of 4 people performing a number of emotion expressions and the system detects the intended ones in table III another set of also 4 people performing the same thing.

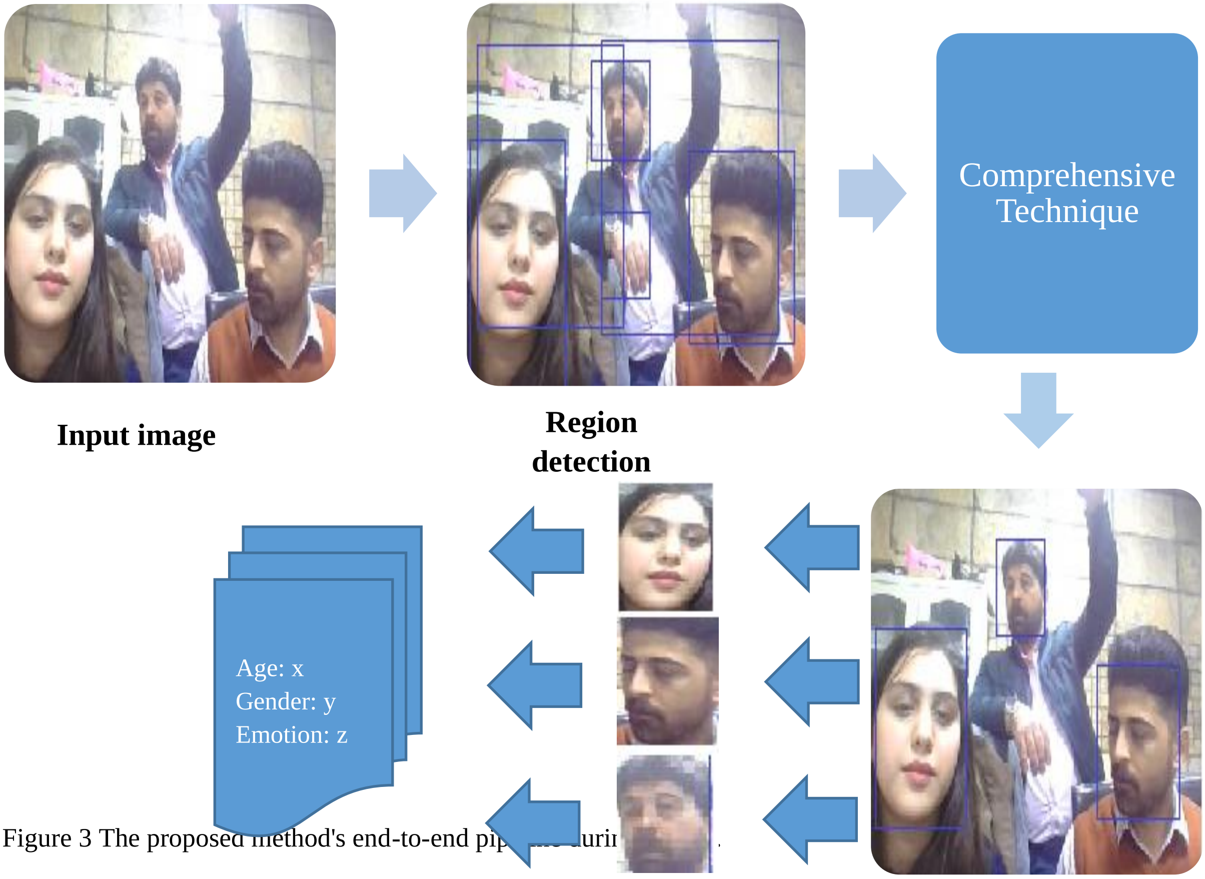


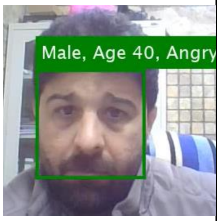
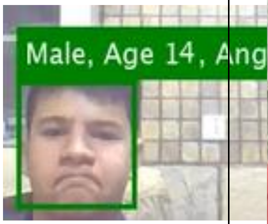
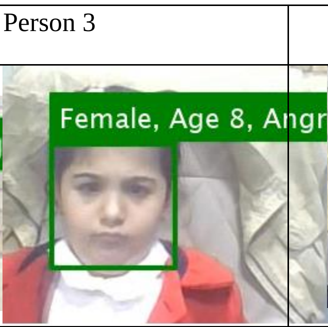
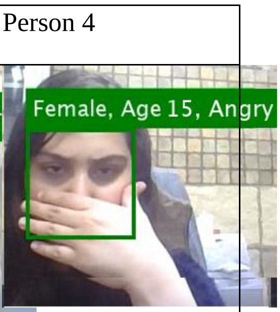
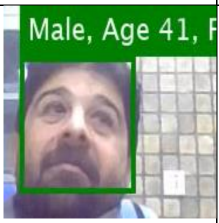
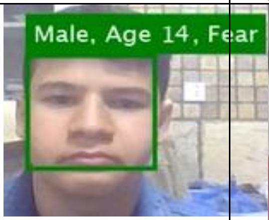
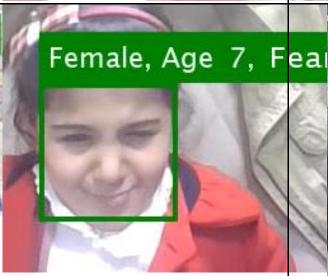
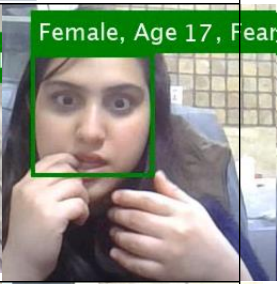
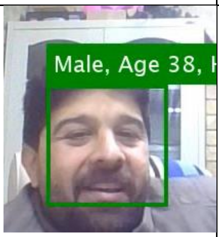
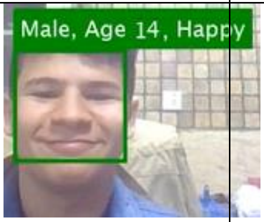
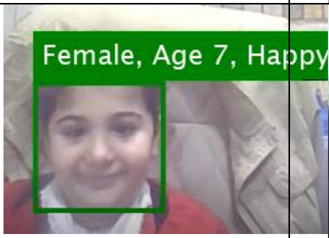
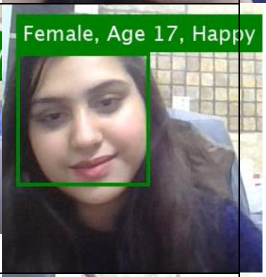
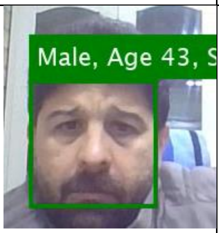
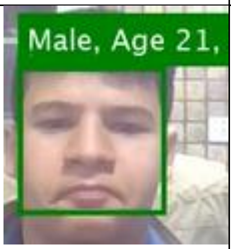
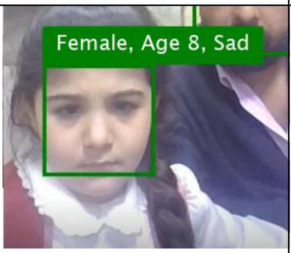
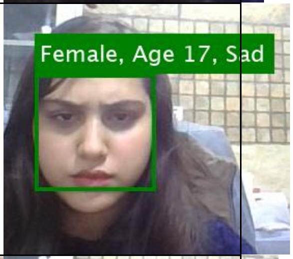
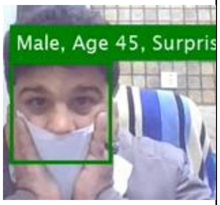
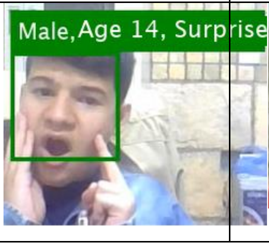
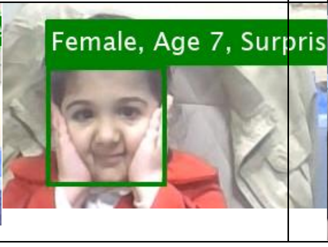
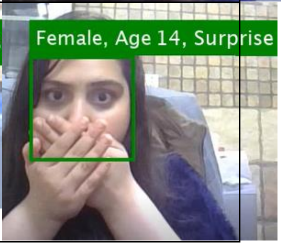
Figure 3 The propose method's end-to-end pip urir .

Table 1 Performance Comparison on Jaffe Dataset

Reference	Recognition rate

Proposed	98%
[23]	94.6%
[17]	91%
[18]	85.1%
[19]	65%
[20]	67%
[21]	79%
[22]	83%

Table 2 ASet of 4 people performing a number of expressions

	Person 1	Person 2	Person 3	Person 4
Angry				
Fear				
Happy				
Sad				
Surprise				

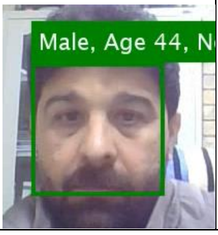
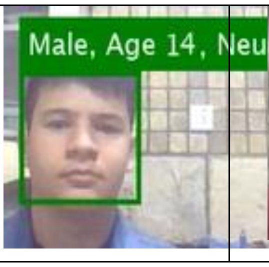
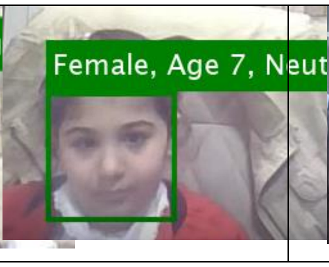
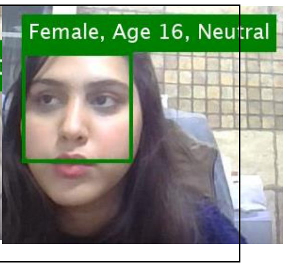
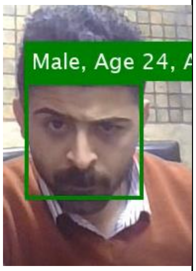
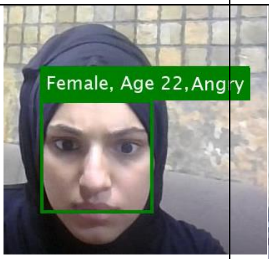
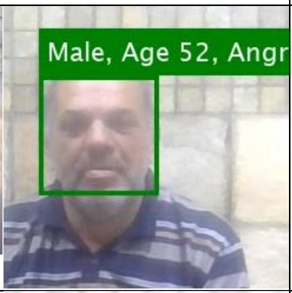
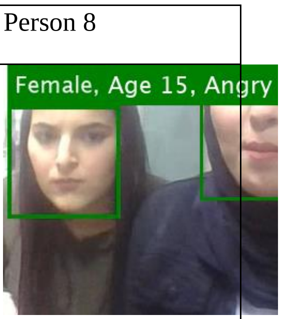
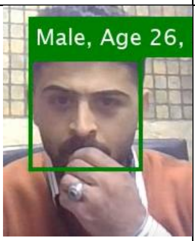
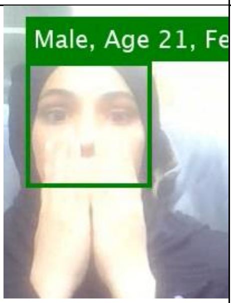
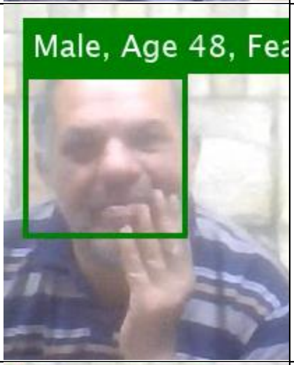
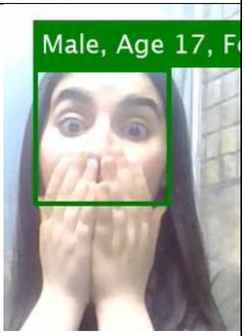
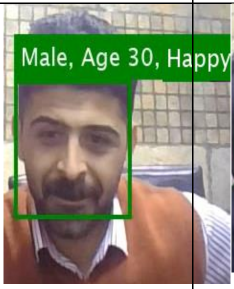
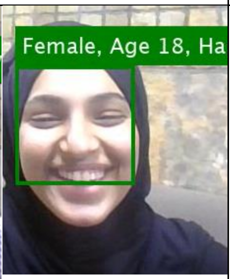
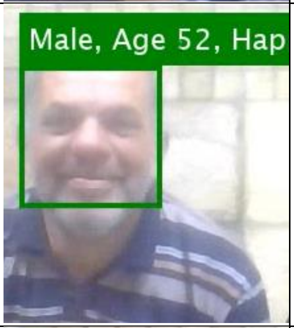
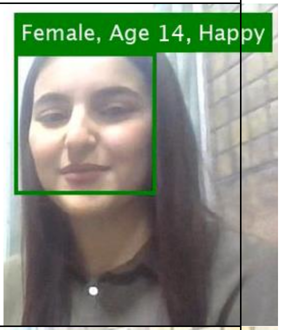
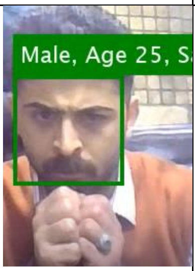
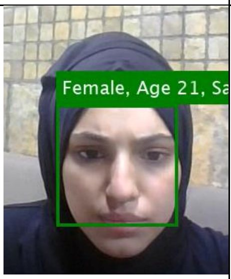
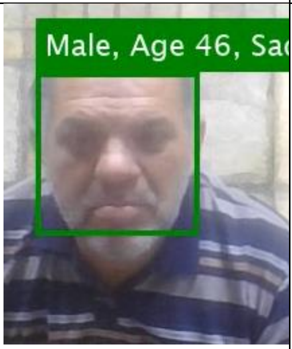
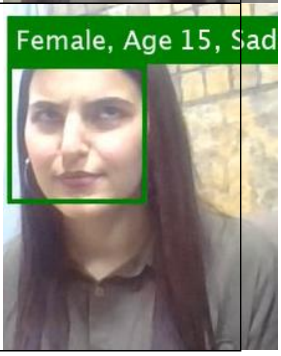
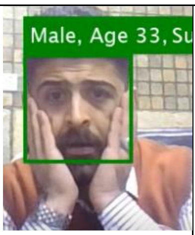
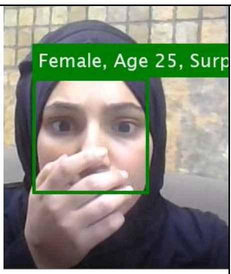
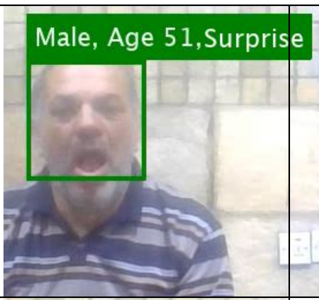
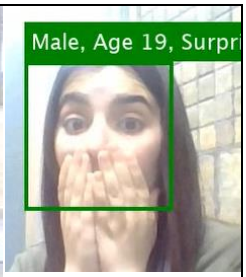
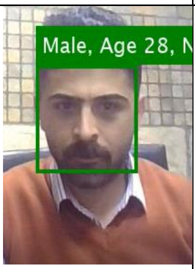
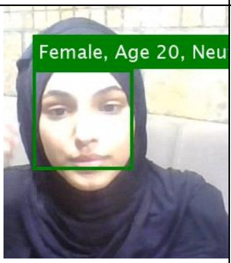
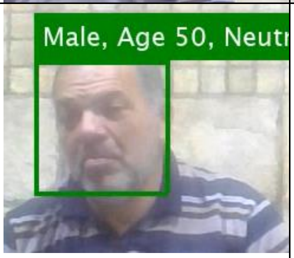
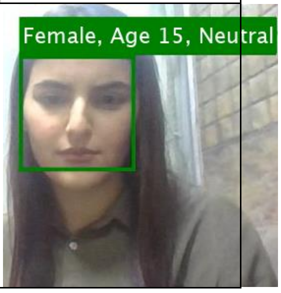
Neutral				

Table 3 Set of 4 people performing a number of expressions

	Person 5	Person 6	Person 7	Person 8
Angry				
Fear				
Happy				
Sad				

Surprise				
Neutral				

8. Conclusions

To implement a Deep learning facial expression recognition system we need to initialize and configure a work environment: To carry out this, it has been necessary to install a large number of libraries for the Matlab and Python programming languages. The use of such libraries as blocks to assist in data sifting where the integration of these tools into the system one of the biggest challenges of the research due to the difference in their implementation and platforms. It is beneficial to use facial expression recognition system to represent an important social ability given its relation to real-world social behavior and other characteristics and emotional abilities. It is important to use a suitable face detection algorithm that can not only be 100% accurate rather can detects multiple faces in a scene and doing this as fast as possible, due to the fact that the proposed systems require real-time processing rather than offline processing o stored data.

References

1. R. Caruana. Multitask learning. In Learning to learn, pages 95–133.Springer, 1998.
2. X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2879–2886, June 2012.
3. X. Zhu and D. Ramanan. FaceDPL: Detection, pose estimation, and landmark localization in the wild. preprint 2015.
4. G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In IEEE Conf. on Computer Vision and Pattern
5. R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multitask learning framework for face detection, landmark localization, pose estimation, and gender recognition. CoRR, abs/1603.01249, 2016.
6. M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer. Facial attributes classification using multi-task representation learning. In Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition
7. Z. Zhang, P. Luo, C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In European Conference on Computer Vision, pages 94–108, 2014.
8. Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In International Conference on Computer Vision, Dec. 2015.
9. J. Wang, Y. Cheng, and R. S. Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. arXiv preprint arXiv:1604.06433, 2016.
10. Mary, Ali Hussien, Zubaidah Bilal Kadhim, and Zainab Saad Sharqi. "Face Recognition and Emotion Recognition from Facial Expression Using Deep Learning Neural Network." IOP Conference Series: Materials Science and Engineering. Vol. 928. No. 3. IOP Publishing, 2020.
11. Mary, AliHussien. "Face Recognition Based Wavelet-PCA Features and Skin Color Model." Journal of Engineering and Development (2011).

12. F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition, pages 815–823, 2015.
13. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition, pages 1701-1708, 2014.
14. V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In IEEE Conference on Computer Visionand Pattern Recognition, pages 1867–1874, June 2014.
15. Durán Suárez, J., 2017. Redes neuronales convolucionales en R: Reconocimiento de caracteres escritos a mano. S.l.: Universidad de Sevilla.
16. Ren, S., He, K., Girshick, R. Y Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. S.l.: s.n., pp. 91-99.
17. S. Escalera, M. Torres, B. Martinez, X. Bar´o, H. J. Escalante, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In Proceedings of IEEE conference
18. C. Li, Q. Kang, G. Ge, Q. Song, H. Lu, and J. Cheng. Deepbe: Learning deep binary encoding for multi-label classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 39–46, 2016.
19. M. Uric´ar, C. FEE, R. Timofte, E. CVL, R. Rothe, J. Matas, and L. Van Gool. Structured output svm prediction of apparent age, gender and smile from deep features.
20. I.-O. a. T. G. A. Stathopoulou, "An improved neural-networkbased face detection and facial expression classification system," Man and Cybernetics, 2004 IEEE International Conference on systems, vol. 1, p. 666–671, 2004.
21. B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen,P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pages 1931–1939. IEEE, 2015.
22. P. Eckman, in Emotions revealed, New York, St. Martin's Griffin, 2003.
23. S. Escalera, J. Fabian, P. Pardo, X. Bar´o, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. Inroceedings of the IEEE International Conference on Computer Vision Workshops, pages 1–9, 2015.
24. G. B. M. S. H. J. C. E. P. a. S. T. J. Donato, "Classifying facial actions," IEEE Transactions on pattern analysis and machine intelligence, vol. 1, no. 10, p. 974–989, 1999.
25. B. a. L. J. Fasel, "Automatic facial expression analysis: a survey. Pattern," vol. 36, no. 1, p. 259–275, 2003.
26. D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. CoRR, abs/1411.7923, 2014.