Analyzing multiple attributes in real-time videos: Integrating age, emotion, and gender estimation databases using deep learning and YOLOv3

Faisal Ghazi Abdiwi **م.د. فيصل غازي**
Department of Computer Science & information systems
Al Mansour University College
faisal.ghazi@muc.edu.iq

تاريخ تقديم البحث: 2024/03/27
تاريخ قبول البحث: 2024/04/07

Abstract

This paper presents a deep learning approach to estimate age, emotion, and gender based on real-time video without relying on facial landmarks or geometric calculations. By integrating age prediction databases with emotions and gender, the proposed model aims to gain a comprehensive understanding, which may benefit various scientific fields. Convolutional neural networks (CNNs) pre-trained on Caffe's ImageNet were used for image classification, with the YOLOv3 algorithm for fast, real-time face detection. Caffe, a platform that supports deep learning algorithms, has proven efficient large-scale computing. The proposed framework uses Deep Convolutional Neural Network (DCNN) features to automatically understand facial expressions and achieve state-of-the-art recognition on public datasets. Combining these features improves classification scores with minimal processing time, especially when using graphics processing units (GPUs)

Keywords : deep learning, YOLOv3, Image expression.

الملخص

تقدم هذه الورقة أسلوب التعلم العميق لتقدير العمر والعاطفة والجنس بناءً على الفيديو في الوقت الفعلي دون الاعتماد على معالم الوجه أو الحسابات الهندسية. ومن خلال دمج قواعد بيانات التنبؤ بالعمر مع العواطف والجنس، يهدف النموذج المقترح إلى الحصول على فهم شامل، مما قد يفيد مختلف المجالات العلمية. تم استخدام الشبكات العصبية التلافيفية (CNNs) التي تم تدريبها مسبقًا على ImageNet من Caffe لتصنيف الصور، مع خوارزمية YOLOv3 للكشف السريع عن الوجه في الوقت الفعلي. أثبتتCaffe ، وهي منصة تدعم خوارزميات التعلم العميق، كفاءة الحوسبة واسعة النطاق. يستخدم الإطار المقترح ميزات الشبكة العصبية التلافيفية العميقة (DCNN) لفهم تعبيرات الوجه تلقائيًا وتحقيق التعرف على أحدث التقنيات في مجموعات البيانات العامة. يؤدي الجمع بين هذه الميزات إلى تحسين درجات التصنيف مع الحد الأدنى من وقت المعالجة، خاصة عند استخدام وحدات معالجة الرسومات.(GPUs)

الكلمات الرئيسية :التعلم العميق، YOLOv3، التعبير بالصور

I. INTRODUCTION

The study of emotional facial expressions, which reflect internal emotional states, meanings, and social communications, has been a continuing interest of scientists since Darwin's work in 1872[1]. It is important to distinguish, from a computer vision perspective, between facial expression recognition and emotion recognition. While the former classifies facial movements based on visual information, the latter involves various factors and manifests itself across multiple channels, including voice, posture, gestures, gaze direction, and facial expressions[2,3,4]. Understanding the extent to which facial expressions correspond to emotions requires prior knowledge of the categories of human emotions[5]. Emotions, as complex and nuanced phenomena, are influenced by cultural and social conditioning and evolutionary factors, leading to physiological and behavioral changes[6]. Facial expressions, a powerful way to convey emotions and signal intentions in face-to-face interactions, play a vital role in communication. A proposed thesis aims to explore facial expression recognition using basic emotion categories, drawing on the Facial Action Coding System (FACS) and Ekman's six universal emotions[7,8]. This study seeks to evaluate existing frame recognition and address the challenge of reliable automatic recognition of facial expressions. Interchangeability between the terms “facial expression” and “facial emotion” is adopted throughout the thesis, considering each category of facial expression as a direct representation of the equivalent human emotional state[9,10].

II. RELATED WORK

In this section, previous work related to the study is presented, highlighting important techniques and identifying strengths and limitations in the studies, and this contributes to determining the basic background of this study.

Caruana [11] was the first to investigate multi-task An early technique proposed by [12] focused on human face recognition, including facial landmark location and face estimation. This approach, which was later popularized by [13], used a tree model containing a common set of components, each representing a specific salient point. Recent developments have seen the multi-task learning model being combined with deep convolutional neural networks (CNNs) for various face-related tasks. Levy et al. [14] proposed a CNN for simultaneous age and gender estimation, while HyperFace [15] improved feature extraction by integrating the middle layers of CNNs for face recognition, pose, historical location, and gender estimation. Ehrlich et al. [16] introduced time-limited multitasking using a Boltzmann computer to study facial characteristics, and Zhang et al. [17] improved landmark localization through simultaneous head pose prediction and facial feature inference. However, these approaches face limitations in performing large-scale multi-task learning, a gap that is addressed in this thesis.

Person face recognition activities have been the subject of extensive research. DP2MFD [18], Faceness [19], Hyperface [15], Faster-RCNN [20], and other deep CNN-based face recognition methods have greatly outperformed standard approaches like TSM [12] and NDPFace [22].

Owing to a lack of training evidence, only a few approaches for face orientation functions, deep CNNs have been used. [17][23][24][15]. Existing landmark localization approaches have mostly concentrated on near-frontal faces [25][26][27], where all of the important key points are clear. Face alignment has been studied PIFA [28], 3DDFA [24], HyperFace [15], and CCL [29] are examples of recent approaches.

In the field of facial feature inference, the focus has expanded to include tasks such as gender and smile identification based on unconstrained images. A major contribution in this field is the CelebA dataset presented by Liu et al. [12],includes nearly 200,000 frontal images annotated with 40 attributes,,including gender and facial expressions. This dataset has played a pivotal role in accelerating research in the field of gender and smile identification from unconstrained images, as recent studies have shown [22] [5] [23] [24]. The Faces of the Universe [25] challenge dataset specialized analysis on these projects with faces of different scales, lighting, and poses [26][27][28].

The process of determining a person's true or apparent age based on their face picture is known as age estimation. Using deep CNNs [29][30], a few approaches have already exceeded human error for the obvious age prediction task [31].

In the proposed system, face verification, which is not in line with our primary focus, serves the purpose of determining whether two faces belong to the same person. Recent developments in this area, such as DeepFace [32], Facenet [33], and VGG-Face [34], leverage deep convolutional neural network (CNN) models trained on millions of annotated data. Notably, these methods significantly enhanced the verification accuracy, especially on datasets such as LFW [37]. However, for unconstrained faces with vast differences in perspective and lighting, it remains a difficult challenge (IJB-A [33] dataset). This problem has been solved by using the multitask learning method to regularize the CNN parameters with just half-a-million samples [35][36][38] for training.

III. PROPOSED METHODOLOGY

The main purpose of this paper is to analyze multiple attributes in real-time videos: integrating age, emotion, and gender estimation databases using deep learning and YOLOv3. When running the process for a live video stream, a frame is captured and a duration is stored in a buffer. A frame splitter start to extract frames from the duration of video stream stored in the buffer. The frames are preprocessed and fed to the detection/recognition algorithm obtaining a result as a structure of data returned to the calling procedure. This will be explained in details showing the usage of algorithm in the process and how the results were obtained.

The proposed systems goes through a number of steps as manifolds to produce the final desired output, for this reason it is convenient to list the manifolds as steps explaining along the way the peculiarities of each step.

A. *Data collection*

In this step collecting labeled data for age, gender and emotion where some problem such as the fact human applied real age estimation is prone to large number of errors (sometimes greater than the computer vision estimations) and one cannot rely on human. There for it is somewhat important to either rely on pre-collected data or building the dataset altogether from scratch.

1. Age dataset

For the proposed age estimation, the Group dataset [15] and Adience [12] results were studied and compared. Images are assigned age groupings based on these datasets and others. A dataset of that has 302,818 images was used as a combination from different sources. Each image has a label of age between 10 and 90 that corresponds to it. The dataset also includes age labels for all images, but unlike the Adience and Group datasets. The results were compared to, which is built off of IMDB (the Internet Movie Database) and IMDB-Wiki data. Also, available APIs like Microsoft [23] and Face++ [24] were tested against the system: (MAE) which means Mean absolute error, is widely used in literature. This data can clearly show that the proposed approach significantly outperformed other methods.

2. Gender dataset

The gender recognition model is based on the Adience benchmark. There are 17,492 faces in the Adience benchmark that have been labelled with their gender. For the tested results a 91% accuracy was achieved based on 20% of the images found in the used data set as validation data during the training process

The proposed system is compared to other leading methods in regards to gender model. Using the Adience facial feature index combined with other gender data sets, there are more than 463,085 different images representing genders of people totally. The faces are scored into five sections since we used the same model on all sides, without the need to apply further fine-tuning to it. Also a look at two commercial APIs: [24] and [25], the proposed recognition of gender is in good agreement with research and existing products.

3. Emotion dataset

Caffe runs on a graphics processing unit (GPU) to perform feature extraction, using a convolutional neural network architecture designed for object recognition ImageNet 15. The ImageNet model includes eight learning layers, five convolutional layers and three fully connected layers, with only the five base layers used extract features. These layers include operations such as convolution, modified linear units (RELU), local response normalization (LRN), and pooling, as shown in Figure 1. Repeating these operations creates different layers[51].

B. Data pre processing

In this research, the majority of datasets used contain images of faces that undergo processing and alignment to simulate real-world applications. Ideal face images should be of a fixed size, centered on the screen, and display minimal background interference. Using an off-the-shelf face detector is the preferred method to accurately locate and record the location and size (scale) of the face in each image. Introducing the cropping feature before subjecting the image to age calculation through face detection significantly improves performance. While many methods typically rely on precise facial landmarks and image forgery, preliminary experiments have revealed that landmarks may lead to misalignment errors. By focusing on faces in diverse environments, the proposed convolutional neural network (CNN) shows the ability to tolerate small alignment errors. To ensure the best results from the proposed system, pre-processing steps include image resizing, where the captured video images are resized to a smaller meaningful size, and region of interest (ROI) convolution [52]. Algorithm 1 shows image resize.

Algorithm 1 : Image resize[52]
Step 1 Obtain image from Real-time stream Step 2: find Outimage(x,y)=$\sum_{i=1}^{2}\sum_{i=1}^{2} \boldsymbol{aijx^{i-1}y^{i-1}}$ Step 3: Move to next x,y in image Step 4: go to step 2

The face detection image is resized to 448 x 448 using a high-speed algorithm, and has an effective resizing rate of about 4.8 ms per 128 x 128 image - a remarkably fast nonlinear resizing algorithm implemented through the Matlab imresize function. For neural networks (NNs) for gender, age and emotion detection, images are resized to 224 x 224 using the same algorithm as above, taking advantage of the smaller size contributing to a higher resizing rate[53].

C. Face detection

It represents the initial and pivotal stage in the detection procedure. In the visual system, YOLO uses a single neural network to directly predict bounding boxes and class probabilities from input images. This approach involves dividing the input image into grid cells, then explicitly predicting the positions and labels of each cell. While the detection accuracy may be lower compared to two-stage detectors, the speed is much greater, making it well-suited for real-time live video streams that feature multiple faces within the same scene[54].

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 1 YOLOV3 internal shortcuts[54].

To extract features, YOLOv3 uses a new network. Since the current network is a combination of the YOLOv2 (Darknet-19) network and the residual network, it contains many shortened links. It is known as Darknet-53 because it contains 53 convolutional layers as seen in Figure 1

D. Age Estimation

Using a deep convolutional neural network (CNN) to evaluate an individual's age from 2D images is an important aspect of this study. Deep CNNs, known for their versatility in information processing, face the difficult problem of overfitting when faced with small datasets, a common complexity in machine learning. This problem is especially apparent in deep neural networks, due to their vast layers, numerous nodes, and millions of parameters.

It is worth noting that databases specifically designed for age classification and estimation are relatively restricted in size compared to those created for image classification and face recognition. In order to mitigate the effect of overfitting, the proposed deep CNN for age estimation adopts a model trained on a large-scale database, to meet the challenge presented by limited dataset sizes[55].

E. machine learning

Supervised methods rely on historical information and class labels, while unsupervised methods excel at revealing underlying structures within the data. Artificial neural networks, inspired by neurobehavior, have gained great importance in data mining to model large and complex problems efficiently.

The process involves multiplying the input by the weight, adding the bias, and passing through the transfer function. The output is determined by the chosen transfer function, with adjustable parameters guided by the learning rule. Realizing that a single neuron may not be enough to solve practical problems, the concept of layers was introduced, forming a complete neural network[56]. As shown in Figure2.

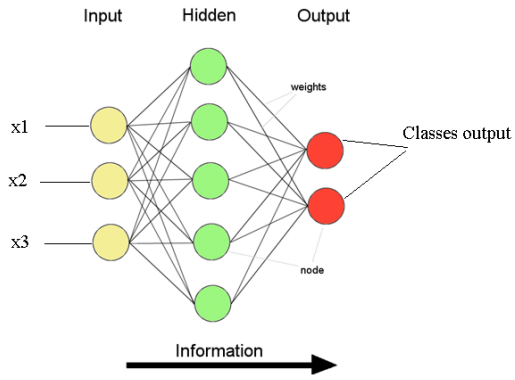


Figure 2. Diagram of a three-layer neural network[56].

Learning is pivotal to a neural network's ability to adapt, as synapses are fine-tuned during the training process to have diverse responses to environmental stimuli. The training process adjusts the weights and biases, enabling the network to generate outputs from the input data.

Artificial neural networks have proven their versatility in solving various problems, including classification, identification, diagnosis, improvement or prediction, especially in data-driven scenarios, which require on-the-run learning and fault tolerance. In such cases, artificial neural networks adapt dynamically by constantly adjusting the weights of their interconnections[57].

A professional convolutional neural network (CNN) was used as the initial facial feature extractor in an attempt to train another convolutional neural network specifically to predict age from facial images. This approach has proven to be highly effective for training purposes. The proposed CNN architecture is based on deep face recognition CNN technology which is famous for extracting distinct and consistent facial features, thus reducing exposure to overfitting. In the context of this study, remarkable results have been achieved, especially on the LFW [36] and YFT [37] datasets. The VGG-Face model consists of eleven layers, eight of which are convolutional layers and three are fully connected. Each convolutional layer is preceded by a normalization layer, and each convolutional block ends with a max-pooling layer, as shown in Figure 3 and detailed in Table 1 [47,48].

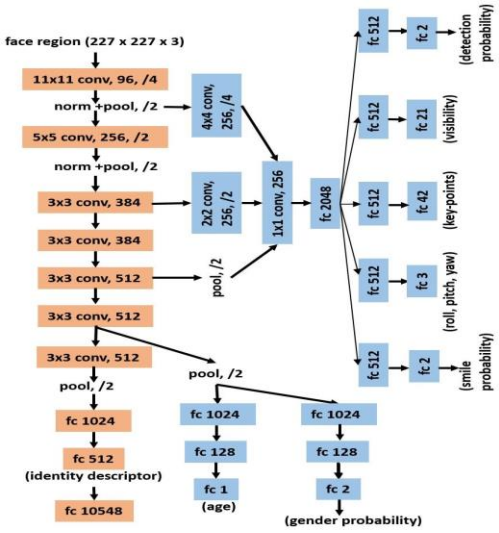


Figure 3: The suggested method's CNN architecture[48].

F. prediction

The comparison image is resized to 256 x 256 pixels. Then, the qualified CCN feeds these three extracted images to the model. For each pixel, this function calculates a softmax probability output vector. By combining the output score vectors of the three images, the final probability vector of the class scores of the original test image is produced. Low resolution and occlusions are reduced using this approach.

G. Expression Extraction

The feature extraction process is performed by Caffe on a graphics processing unit (GPU), taking advantage of a convolutional neural network architecture originally designed for object detection ImageNet 15. The specific network has eight learning layers, including five convolutional layers and three fully connected layers designed for object detection. Objects in ImageNet. For the purpose of feature extraction in this study, only the five basic layers of this architecture are used. These layers include a combination of convolutional operations and modified linear unit (ReLU) activation, as shown in the algorithm 2:

Algorithm 2 Algorithm for recognizing facial expressions using CNN
for all image (i), depicting facial expressions ∈ face dataset do
convert the image (i) to gray-scale
detect frontal face in (i) and crop only the face (c)
extract POOL5 (256 × 6 × 6) features for cropped face (c) using CNN
copy predefined facial expression label for each image(i) as per the input dataset
use the resulting POOL5 vector of dimension 9216(256 × 6 × 6) for tenfold and leave-one-out cross validation
with SVM classifier to recognize facial expression

A. Proposed system

Caffe carefully monitors each aspect with a layer-directed acyclic graph, ensuring accurate forward and backward passes. In the world of machine learning systems, Caffe models are designed to run seamlessly from start to finish. A standard network starts with a data layer responsible for reading data from disk and culminates in a loss layer that calculates the target for specific tasks such as classification or recovery.

The system seamlessly switches between CPU and GPU with a single switch, with layers backed by CPU and GPU routines that produce identical results, backed by rigorous testing. This transition from CPU to GPU remains smooth and independent of the underlying model architecture. To ensure accurate training and quantifiable results, 20% of the data processed and trained within the system is allocated for validation purposes across each dataset.

The 2D network proposed in this paper contains seven convolutional layers followed by a connected network (also called "fully connected"). The training network used for face recognition and as an origin point for sharing the parameters of the six convolutional layers with other related network processes. A parametric rectifier has the activation The architecture includes a convolutional neural network (CNN) with a linear, rather than parametric, functional parameter set, which enhances the extraction of discriminative facial features for comprehensive face analysis. This analysis includes face detection, key point identification, pose estimation, smile detection, age detection, gender identification, and facial feature recognition. Combining the third and fifth convolutional layers with the base layer facilitates training on subject-independent tasks.

Using a pooling layer and two subsequent convolutional layers for the underlying layers, a function map size of 6 × 6 is achieved. Furthermore, a dimensionality reduction layer is applied to reduce the function maps to 256. This is followed by a fully connected layer with dimensions of 2048, which serves as a global representation of non-objective activities The underlying tasks are then separated into 512 fully connected layers, culminating in the output layers.

For training, the input images (shown in Table 2) are resized to 256 x 256 pixels and then randomly cropped into patches of 224 x 224 pixels. Stochastic gradient descent optimization is used with mini-batches of 256 segments, a momentum value of 0.9, and a weight decay of 10-3. Regularization is achieved with a dropout rate of 0.6 during training.

Table 2 Training images datasets

Table 1 Architecture and connection of VGG-Face model

Layer#	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Name	n/a	Conv1	Relu1	Norm1	Pool1	Conv2	Relu2	Norm2	Pool2	Conv3	Relu3	Conv4	Relu4	Conv5	Relu3_2
Support	n/a	3	1	3	1	2	3	1	3	1	2	3	1	3	1
Filt dim	n/a	3	n/a	64	n/a	n/a	64	n/a	128	n/a	n/a	128	n/a	256	n/a
Num filts	n/a	64	n/a	64	n/a	n/a	128	n/a	128	n/a	n/a	256	n/a	256	n/a
Stride	n/a	1	1	1	1	2	1	1	1	1	2	1	1	1	1
pad	n/a	1	0	1	0	0	1	0	1	0	0	1	0	1	0
Layer#	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Name	Conv3_3	Relu3_3	Mpool3	Conv4_1	Relu4_1	Onv4_2	Relu4_2	Conv4_3	Relu4_3	Mpool4	Conv5_1	Relu5_1	Conv5_2	Relu5_2	Conv5_3
Support	3	1	2	3	1	3	1	3	1	2	3	1	3	1	3
Filt dim	256	n/a	n/a	256	n/a	3	1	3	1	2	3	1	3	1	3
Num filts	256	n /a	n/a	512	n/a	512	n/a	512	n/a	n/a	512	n/a	512	n/a	512
Stride	1	1	2	1	1	512	n/a	512	n/a	n/a	512	n/a	512	n/a	512
pad	1	0	0	1	0	1	1	1	2	2	1	1	1	1	1
Layer#	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
Name	relu5_3	pool5	fc6	relu6	drop6	fc7	relu7	drop7	fc8	relu8	drop8	fc9	prob		
Support	1	2	7	1	1	1	1	1	1	1	1	1	1		
Filt dim	n/a	n/a	512	n/a	n/a	4096	n/a	n/a	5000	n/a	n/a	5000	n/a		
Num filts	n/a	n/a	4096	n/a	n/a	5000	n/a	n/a	5000	n/a	n/a	8	n/a		
Stride	1	2	1	1	1	1	1	1	1	1	1	1	1		
pad	0	0	0	0	0	0	0	0	0	0	0	0	0		

If the consistency of the validation set does not improve, the training rate will be reduced tenfold(as mentioned in Algorithm 2). Biases are set to zero, and weights between newly introduced fully connected classes are initialized using a Gaussian distribution with a negative mean and standard deviation of 2-10. The RGB input image is fed to the network input layer, and the output of each hidden layer serves as input for the subsequent hidden layer before reaching the network output layer.

In Figure 4, the stochastic gradient descent approach optimizes the parameters of the fully connected layer for age estimation, while leaving the parameters of the convolutional layer unchanged. This strategy involves optimizing the parameters of the fully connected layer for age prediction while preserving the parameters of the convolutional layer, which is pre-trained for face recognition tasks..

During the testing phase, a two-stage approach is used, as shown in Figure 4. Initially, from a test image, selective search [49] generates region proposals, which are subsequently processed by the proposed end-to-end neural network to extract detection scores. . .

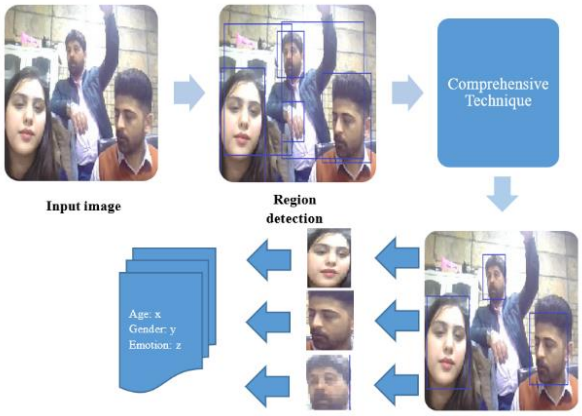


Figure 4: The proposed method's end-to-end pipeline during testing.

The testing pipeline includes head pose detection, orientation, and vision. Iterative region proposals and feature-based non-maximum suppression (NMS) techniques are applied to filter out non-face regions and enhance orientation and pose detection [50]. The reliability scores collected in the second stage are then used to match and align the observed face to a valid display using a similarity transformation.

Next, details of emotion, gender, age and identity are obtained by passing the matching faces, as well as their mirror versions, through the network again. The identity descriptor is based on a 512-dimensional function extracted from the penultimate layer fully connected to the identity network.

IV. Results and discussion

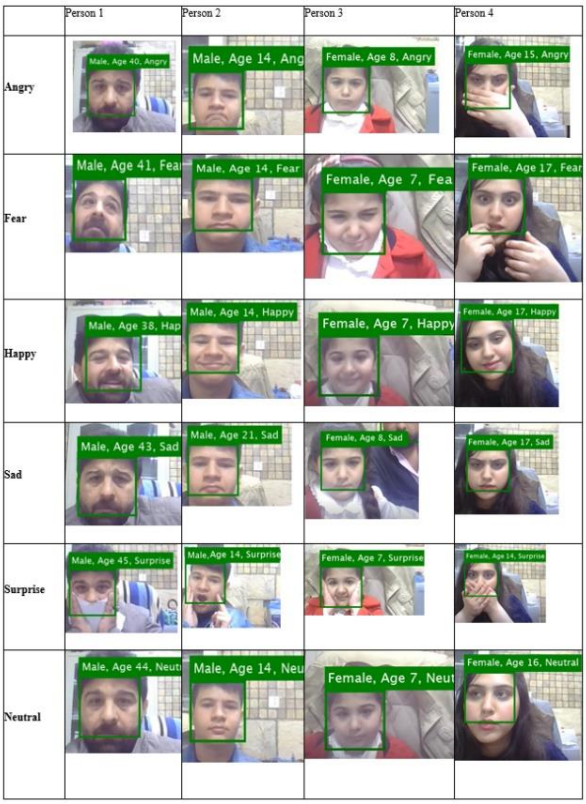
Table 3 shows the classification performance using state-of-the-art methods on the CAFFE database. The top 1 and top 5 errors were reported by 96.10% and 100%, respectively.

Table 3: Performance Comparison on Caffe Dataset

Study	Method	Recognition (%)
The proposed method	CNN model	Top1:96.10 Top 5:100
[23], 2016	Local curvelet transform	94.65
[17], 2011	Patch-based-Gabor	91.00
[18], 2010	Salient feature vectors	85.92
[19], 2008	WMMC	65.77
[20], 2007	KCCA	67.00
[21], 2008	DCT	79.30
[22], 2010	FEETS + PRNN	83.84

However, Tables 4 and 5 display gender, age, and emotion detection results for eight individuals (four males and four females), including six distinct emotions (anger, fear, happiness, sadness, surprise, neutral).

Table 4 :A Set of 4 people performing a number of expressions



While some results achieve accuracy levels of 94% to 96%, some cases may be misidentified due to factors such as face coverings, beards and mustaches. Features related to the mouth and cheeks can also affect detection accuracy. Additionally, issues inherent in the accuracy of the face detection algorithm contribute to variations in the overall success rate of the detection process.

Table 5: - a Set of 4 people performing a number of expressions

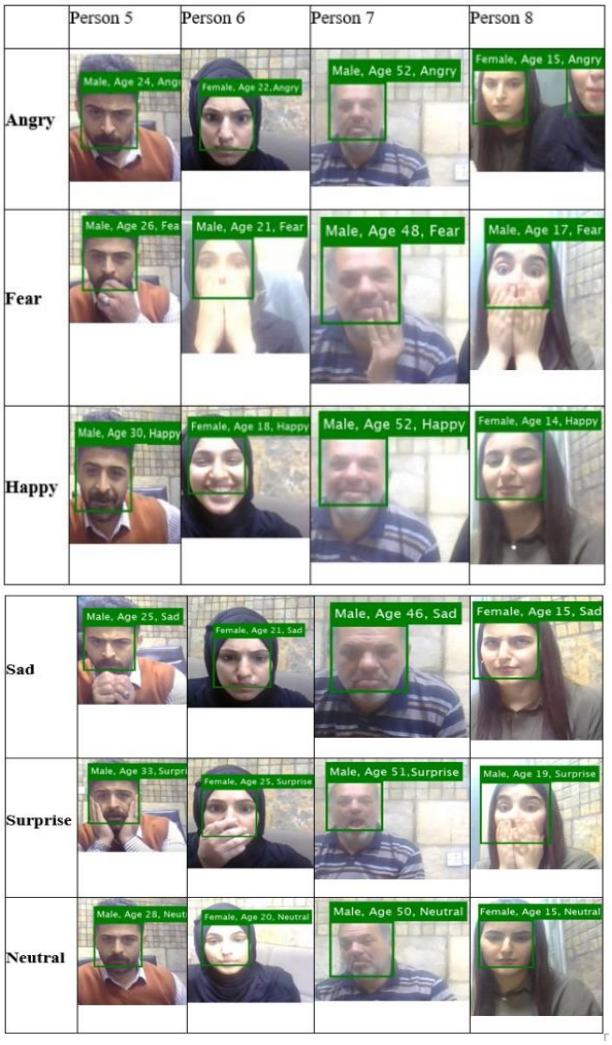


Table 6 : obtained results of 8 people performing multiple emotions

	Face detection	Emotion recognition	Gender estimation	Age estimation
Person 1	98%	98%	98%	±3
Person 2	97%	97%	97%	±2
Person 3	95%	93%	95%	±5
Person 4	96%	96%	96%	±2
Person 5	97%	97%	97%	±3
Person 6	97%	93%	94%	±5
Person 7	97%	97%	97%	±2
Person 8	96%	96%	96%	±6

V. CONCLUSION

This study evaluates the performance of a deep learning facial expression recognition system across eight real-time tests. Both the Caffe model and YOLO face detection techniques were used. Caffe is a deep learning framework typically used for image classification and convolutional neural networks (CNNs), while YOLO is known for its real-time object detection capabilities. The proposed model achieved high accuracy in face detection, reaching (95% to 98%), while accuracy in emotion recognition reached (93% to 98%), and gender estimation (94% to 98%). and age estimate (±2 to ±6 years). The achieved accuracy highlights the system's efficiency in capturing facial features and expressions in real-world applications. The proposed model can be applied in many fields that require human-computer interaction and demographic analysis. These comprehensive metrics confirm the reliability and robustness of the applied deep learning model.

References

[1] C. Darwin, The expression of emotion in man and animals, 1872.

[2] B. a. L. J. Fasel, "Automatic facial expression analysis: a survey. Pattern," vol. 36, no. 1, p. 259–275, 2003.

[3] 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.

[4] 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.

[5] M. a. R. L. J. Pantic, "Facial action recognition for facial expression analysis from static face images," IEEE Transactions on Systems, Man, and Cybernetics, vol. 34, no. 4, p. 1449–1461, 2004.

[6] C. Darwin, "The expression of the emotions in man and animals," University of Chicago press, 1965, p. 526.

[7] P. Eckman, in Emotions revealed, New York, St. Martin's Griffin, 2003.

[8] P. a. F. W. V. Ekman, "Unmasking the face : A guide to recognizing emotions from facial clues," Ishk, 2003.

[9] P. a. F. W. V. Ekman, Manual for the facial action coding system, Consulting Psychologists Press, 1978.

[10] Anderson, K. and McOwan, P. W. (2006). A real-time automated system for the recognition of human facial expressions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 36(1) :96–105.

[11] R. Caruana. Multitask learning. In Learning to learn, pages 95–133.

Springer, 1998.

[12] 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.

[13] X. Zhu and D. Ramanan. FaceDPL: Detection, pose estimation, and landmark localization in the wild. preprint 2015.

[14] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In IEEE Conf. on Computer Vision and Pattern,2015, doi: 10.1109/CVPRW.2015.7301352

[15] 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.

[16] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer. Facial attributes classification using multi-task representation learning. (2016)In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

[17] 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.

[18] R. Ranjan, V. M. Patel, and R. Chellappa. A deep pyramid deformable part model for face detection. In International Conference on Biometrics Theory, Applications and Systems, 2015

[19] S. Yang, P. Luo, C. C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In IEEE International Conference on Computer Vision, 2015.

[20] H. Jiang and E. Learned-Miller. Face detection with the faster r-cnn. arXiv preprint arXiv:1606.03473, 2016

[21] S. Liao, A. Jain, and S. Li. A fast and accurate unconstrained face detector. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In International Conference on Computer Vision, Dec. 2015.

[23] A. Kumar, R. Ranjan, V. M. Patel, and R. Chellappa. Face alignment by local deep descriptor regression. CoRR, abs/1601.07950, 2016.

[24] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. arXiv preprint arXiv:1511.07212, 2015.

[25] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. International Journal of Computer Vision, 107(2):177–190, 2014.

[26] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1685–1692, 2014.

[27] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1867–1874, June 2014.

[28] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In The

IEEE International Conference on Computer Vision (ICCV), December 2015.

[29] S. Zhu, C. Li, C. C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning.

[30] X. Zhu and D. Ramanan. FaceDPL: Detection, pose estimation, and landmark localization in the wild. preprint 2015.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In International Conference on Computer Vision, Dec.2015.

[32] 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.

[33] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1637–1644, 2014.

[34] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer. Facial attributes classification using multi-task representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 47–55, 2016.

[35] 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

[36] 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.

[37] 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.

[38] K. Zhang, L. Tan, Z. Li, and Y. Qiao. Gender and smile classification using deep convolutional neural networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2016.

[39] R. Rothe, R. Timofte, and L. V. Gool. Dex: Deep expectation of apparent age from a single image. In IEEE International Conference on Computer Vision Workshops (ICCVW), December 2015.

[40] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa.

A cascaded convolutional neural network for age estimation of unconstrained faces.

[41] 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. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 1–9, 2015.

[42] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 815–823, 2015.

[43] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1701-1708, 2014.

[44] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition.

In British Machine Vision Conference, volume 1, page 6, 2015.

[45] 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 Recognitionm(CVPR), pages 1931–1939. IEEE, 2015.

[46] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. CoRR, abs/1411.7923, 2014.

[47] Salas, R., 2004. Redes neuronales artificiales. Universidad de Valparaıso. Departamento de Computación, vol.1

[48] Durán Suárez, J., 2017. Redes neuronales convolucionales en R: Reconocimiento de caracteres escritos a mano. S.l.: Universidad de Sevilla.

[49] Arshid, S., Hussain, A., Munir, A., Nawaz, A., and Aziz, S. (2017). Multi-stage binary patterns for facial expression recognition in real world. Cluster Computing, pages 1–9.

[50] Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M., and Solomon, P. E. (2009). The painful face–pain expression recognition using active appearance models. Image and vision computing, 27(12) :1788–1796.

[51] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland and S. Wermter, "The OMG-Emotion Behavior Dataset," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 2018, pp. 1-7, doi: 10.1109/IJCNN.2018.8489099.

[52] Witten, I.H. Y Frank, E., 2005. Data Mining: Practical machine learning tools and techniques. S.l.: Morgan Kaufmann.

[53] Friedman, J., Hastie, T. Y Tibshirani, R., 2001. The elements of statistical learning. S.l.: Springer series in statistics New York.

[54] 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.

[55] Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., & Li, S. Z. (2020). Towards Fast, Accurate and Stable 3D Dense Face Alignment. In Computer Vision – ECCV 2020 (pp. 152–168). Lecture Notes in Computer Science (LNIP), volume 12364. Springer.

[56] Celiktutan, O., Sariyanidi, E., & Gunes, H. (2015). Let me tell you about your personality!: Real-time personality prediction from nonverbal behavioural cues. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).

[57] Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns, 2(8), 100336. https://doi.org/10.1016/j.patter.2021.100336.