Zayed University
ZU Scholars

All Works

1-1-2021

A Deep Learning Approach for Real-Time Analysis of Attendees' Engagement in Public Events

Sujith Samuel Mathew
Zayed University, sujith.mathew@zu.ac.ae

Manar AlKhatib
British University in Dubai

May El Barachi
University of Wollongong in Dubai

Follow this and additional works at: https://zuscholars.zu.ac.ae/works

Part of the Computer Sciences Commons

Recommended Citation

Mathew, Sujith Samuel; AlKhatib, Manar; and Barachi, May El, "A Deep Learning Approach for Real-Time Analysis of Attendees' Engagement in Public Events" (2021). *All Works*. 4218.
https://zuscholars.zu.ac.ae/works/4218

This Article is brought to you for free and open access by ZU Scholars. It has been accepted for inclusion in All Works by an authorized administrator of ZU Scholars. For more information, please contact scholars@zu.ac.ae.

A Deep Learning Approach for Real-Time Analysis of Attendees' Engagement in Public Events

Sujith Samuel Mathew, Manar AlKhatib, and May El Barachi

***Abstract*—Smart city analytics requires the harnessing and analysis of emotions and sentiments conveyed by images and video footage. In recent years, facial sentiment analysis attracted significant attention for different application areas, including marketing, gaming, political analytics, healthcare, and human computer interaction. Aiming at contributing to this area, we propose a deep learning model enabling the accurate emotion analysis of crowded scenes containing complete and partially occluded faces, with different angles, various distances from the camera, and varying resolutions. Our model consists of a sophisticated convolutional neural network (CNN) that is combined with pooling, densifying, flattening, and Softmax layers to achieve accurate sentiment and emotion analysis of facial images. The proposed model was successfully tested using 3,750 images containing 22,563 faces, collected from a large consumer electronics trade show. The model was able to correctly classify the test images which contained faces with different angles, distances, occlusion areas, facial orientation and resolutions. It achieved an average accuracy of 90.6% when distinguishing between seven emotions (Happiness, smiling, laughter, neutral, sadness, anger, and surprise) in complete faces, and 86.16% accuracy in partially occluded faces. Such model can be leveraged for the automatic analysis of attendees' engagement level in events. Furthermore, it can open the door for many useful applications in smart cities, such as measuring employees' satisfaction and citizens' happiness.**

***Index Terms*—smart cities, sentiment analysis, facial recognition, convolutional neural networks, deep learning.**

I. Introduction

THE principal goals of smart cities are to improve citizens' quality of life, foster economic growth while ensuring that development remains sustainable, and efficiently deliver the necessary public services. Intelligence gathering and tracking of residents, events and assets constitute core functions in any smart city. One category of data of particular interest to businesses and governments is citizens' sentiments about current events and developments. Indeed, emotions and sentiments are central to all human activities and are considered to be key influencers of human behavior. Thus, the ability of analyzing emotions and sentiments saw great interest not only in the computer science field, but also in political science, marketing, finance, and health sciences [1][2].

This interest was accentuated with the rapid growth of social networks (such as Facebook and Twitter) that offer rich and vast amounts of data from which sentiments and emotions can be inferred. Enabling the gaining of insights about the public's views and feelings about products or services is important for companies seeking to enhance their reputation and offerings, and maintain a competitive edge in the market. Smart city governments are also interested in measuring citizens' satisfaction and engagement with government services.

In [3], the authors proposed a system for classifying customer shopping behavior based on surveillance camera feed analysis. An SVM model classified the orientation of the customer's head and body into one of eight directions, to determine whether the customer is satisfied with the product or turning away from it. A similar research about the use of surveillance cameras to analyze customer satisfaction was proposed in [4].

In [5], the authors developed a still-to-video facial recognition (FR) multiple classifier system (MCS) to recognize and detect individual faces through surveillance camera feeds. These systems were designed to accurately detect the presence of the individuals of interest across a distributed network of video cameras based on their corresponding facial models. Face Recognition systems capture faces appearing in videos, and then match these faces against facial models generated based on high-quality target images. Similarly, the authors in [6] employed a model to detect facial depression behavior. The proposed model was built using a deep neural network architecture based on features extracted from full face and eyes regions that are relevant for analyzing depression. This model has interesting applications in the areas of medical diagnosis and affective state monitoring.

In terms of sentiment analysis techniques, facial sentiment analysis of images has been typically conducted along three different dimensions: 1) low-level image features; 2) mid-level or semantic-level features attributes; and 3) most recently deep neural network-based architectures.

The low-level features approach relies on basic hand-crafted images. The authors in [7] proposed a semi-supervised approach using a partially-labeled factor graph model (PFG), for detecting and predicting image affects. The model leveraged color features and social correlation among images by

Manuscript received March 23, 2021; revised April 26, 2021. Date of publication May 4, 2021. Date of current version May 4, 2021.

The paper was presented in part at the 5th International Conference on Smart and Sustainable Technologies (SplitTech20) 2020.

This research was supported in part by Zayed University RIF grant number R19099, UAE, in 2020.

Sujith Samuel Mathew is with the College of Technological Innovation, Zayed University, UAE (e-mail: Sujith.Mathew@zu.ac.ae). Manar Alkhatib is with the College of Information Technology, British University in Dubai, UAE, (e-mail: Manar.Alkhatib@buid.ac.ae). Corresponding author: May El Barachi is with the Faculty of Engineering and Information Sciences, University of Wollongong in Dubai, UAE, (e-mail: maielbarachi@uowdubai.ac.ae).

Digital Object Identifier (DOI): 10.24138/jcomss-2021-0072

1845-6421/06/2021-0072 © 2021 CCIS

classifying the images into 16 categories. The experiments were made on 20,000 random Flickr images, and yielded a precision of 49% with a recall of 24%. The authors in [8] designed a sentiment classification system based on the evaluation of local image statistics, while in [9], advanced features based on psychology and art theories (e.g. faces, colors, skin, texture, composition) were employed for sentiment prediction. Mid-level models focus on conducting image sentiment analysis using semantic-level features that are related to objects or scenes and are inferred from low-level image features. Typically, mid-level attributes of an image make the prediction more interpretable than using the low-level features. In [10], the authors proposed an image sentiment prediction framework by leveraging eigenface-based facial expression detection model. The framework yielded a Facial Emotion Detection accuracy of 73.86%. Similar research was conducted in [11], in which a model was proposed to develop an image sentiment ontology and sentiments' bank based on to a 1, 200 adjective noun pairs (ANP) representing different emotions.

With the rapid development of convolutional neural networks (CNNs), researchers have attempted to propose deep learning architectures for analyzing image sentiments. In [12], the authors proposed a multimodality sentiment analysis technique by combining texts and images together. The authors in [13] proposed a novel Sentiment Network with visual Attention (SentiNet-A) architecture which explored the visual attention that enhanced image sentiment analysis. The work in [14] used a hybrid approach that integrated two classifiers (CNN and SVM) in order to speed up the detection process. This hybrid model focused on eye detection only without sentiment analysis capabilities, and relied on the eye variance filter (EVF) to detect and eliminate non-eye images.

Another work on sentiment analysis [15] proposed a unified CNN-RNN model to predict emotions. In this model, a bidirectional recurrent neural network (RNN) model is integrated with the learned features from the multiple layers in the CNN model. Finally, in [16], the authors proposed a new loss function along with a deep learning approach based on a two CNN channels architecture to detect and verify partially occluded faces.

Conducting accurate facial sentiment and emotional analysis for individuals appearing in unconstrained real-life videos and images remains a challenging task. Indeed, natural settings imply large variations of facial appearance caused by the diversity of human face angles, orientations, occlusions, scales, poses, expressions, blur, and ambient lighting. Conducting such analysis in real time, for a large diversity of noisy images that may be of different resolutions, taken at different distances and angles is an open research area. Despite the merits of the proposed approaches, there is still a significant gap in the capabilities of such solutions, when compared to human vision.

In this work, we address some of those challenges by proposing a deep learning approach for the real-time analysis of noisy images extracted from large events' surveillance cameras' feeds.

The contributions of this work are as follows:

1. The proposed solution represents an advanced framework enabling accurate facial sentiment and emotional analysis for individuals appearing in unconstrained real-life videos and images.
2. Unlike existing deep learning and supervised learning approaches, which analyze complete faces only, considering the image as a whole and converting it into one representation, our approach aims at identifying the most relevant regions within the image and using this information to conduct sentiment and emotions analysis from complete and partial (incomplete) faces. It should be mentioned that some recent solutions addressed the detection of partial faces, such as the work in [17] that detects faces with and without masks, and the approach proposed in [16] for identity detection in occluded and non-occluded faces. Yet those approaches focus on facial detection only, without addressing the sentiment and emotion analysis aspect.
3. Finally, unlike existing solutions, our approach has practical applications in natural settings encompassing a large diversity of noisy images that represent participants in large scale events. Such approach can have interesting applications for the analysis of events' participants engagement and satisfaction.

The rest of the paper is organized as follows: Section II details our facial sentiment and emotional analysis approach. Section III presents the results obtained. We conclude the paper in section IV.

II. A Deep Learning Approach for Facial Emotion Detection in Noisy Images

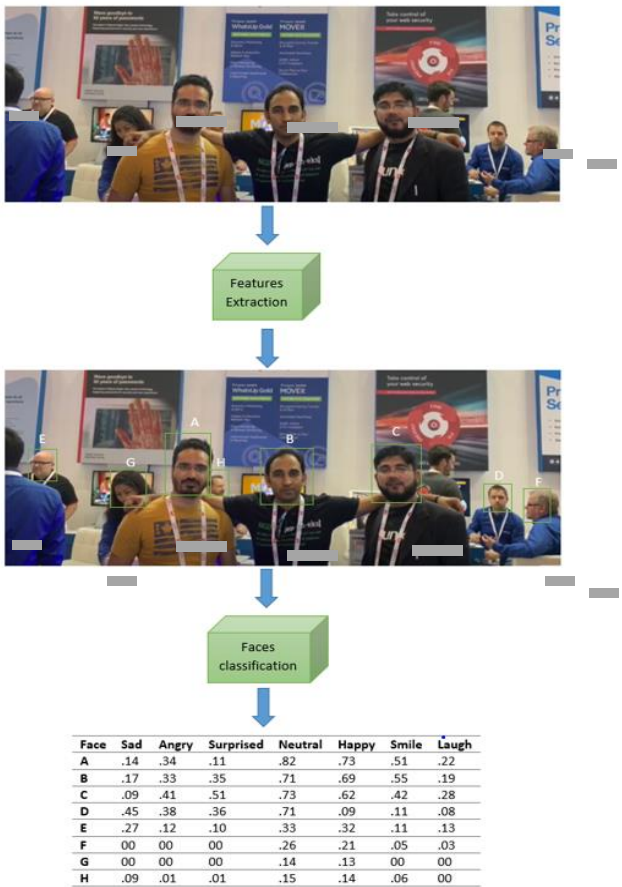
2.1 Facial Emotions' Analysis Approach

Facial expressions' detection is considered as one of the fundamental problems in computer vision, with applications in various domains such as video gaming, call centers, and human-computer interaction. Despite rapid development in the area of machine learning, the problem of facial expressions' detection in not yet well addressed. Facial expressions are typically classified into six basic emotions: Anger, sadness, surprise, laughter, happiness, and smiling, with the neutral expression considered as the seventh emotion. Figure 1 depicts a high-level overview of our proposed facial sentiment analysis approach, which extracts the necessary facial features from an image, and classifies the emotion for each subject appearing in the image.

The main contribution of this work consists of the novel use of convolutional neural networks (CNNs) and advanced features for the accurate classification of facial sentiments and emotions in complete and occluded faces. CNNs can support different input size and the incremental size technique was employed in this work to train the network more quickly. It is work mentioning that our network is trained on datasets that are not target labeled. In comparison to the earlier version of this work [18], we present in this paper a more effective design of CNN to achieve improved speed and performance. More specifically, our new model relies on independent CNNs for learning the responses of different facial regions, from different angles and distances. Figure 2 depicts our facial sentiments' classification approach.

As shown in the figure, the approach used enables the detection and classification of facial emotions from input images that are resized to 48 X 48 pixels for the analysis. The

classification method used consists of two main steps: 1) the automatic extraction of features; and 2) the classification using the trained model. In this method, the image is fed to our CNN model using the input layer. Next, the image is converted into a matrix where each pixel is stored as numbers representing the RGB values (ranging from 0 to 255 as red, blue, green concentration levels. Figure 3 illustrates an example of an image to matrix conversion. Following that step, the system moves on to the convolution process in which the trained model is use to determine familiar patterns in the image, such as curvature, brightness, and intensity [19]. During the convolution process, a series of successive convolutions and pooling operations are performed. Each layer in the CNN consists of a two-dimensional plan called feature map that consists of multiple independent neurons. Each neuron on the feature map receives inputs from a small neighborhood in the previous layer. Different filters are applied to the image at different resolutions to enable accurate classification irrespective of the image quality. Those filters also provide us with a convoluted feature matrix, which consists of the multiplication results that the algorithm found by applying filters over different parts of the image. The application of a filter has two main attributes that are critical for emotions' detection in faces, namely: Size and stride. Zero padding is applied if the filter size differs from the size of the image. Stride represents the number of pixels that can be moved around the image.



Face	Sad	Angry	Surprised	Neutral	Happy	Smile	Laugh
A	.14	.34	.11	.82	.73	.51	.22
B	.17	.33	.35	.71	.69	.55	.19
C	.09	.41	.51	.73	.62	.42	.28
D	.45	.38	.36	.71	.09	.11	.08
E	.27	.12	.10	.33	.32	.11	.13
F	00	00	00	.26	.21	.05	.03
G	00	00	00	.14	.13	00	00
H	.09	.01	.01	.15	.14	.06	00

Fig. 1. Overview of facial sentiment classification framework

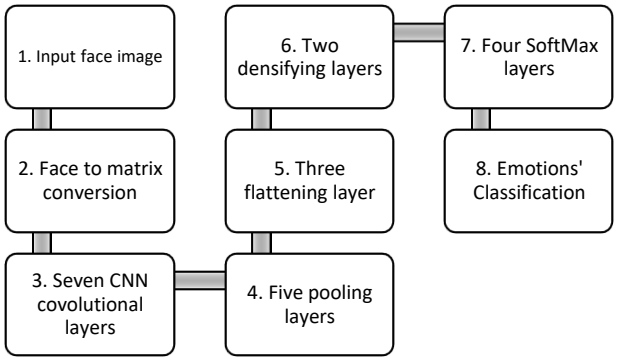


Fig. 2. Image Emotions' Classification Approach

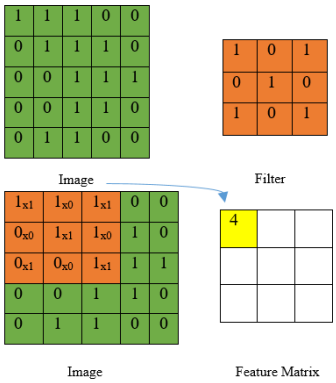


Fig. 3. Image to matrix conversion example

Following the convolution process are the pooling layers. Similar to convolution, pooling involves filters and strides, in addition to a down sampling operation along the dimensions, to that matrices are not multiplied each time a pooling filter is applied. Instead, the max pooling layers take the result from the convolution layers and checks for the highest weight. The pooling layers help ensuring robustness against noise and distortion, thus leading to a reduction in the features' resolution and a more efficient operation. It should be noted that the convolution layers and the pooling layers are very important steps in the CNN feature extractor. The layers from C1 up to C7 are used to extract the image's features. The layers following the pooling step, are the densifying and flattening layers. These layers are responsible of transforming the data from a multidimensional matrix into a single dimension matrix. The one-dimension matrix is used as an input for an artificial neural network layer that will extract the image's features more precisely. With the densifying and flattening layers, the number of parameters the network needs to match is further reduced, thus getting closer to the end goal of classification. The SoftMax layer is the last layer in the model that is responsible for the classification decision about the depicted emotion. This layer returns a probability value (between 0 and 1) of the image belonging to one of the seven emotion classes. Figure 4 depicts an example of the applied process for face detection and expression sentiment classification.

2.2 Dataset Construction and Feature Engineering

The training dataset images used in our system are extracted from the following sources: Facial Expression Recognition

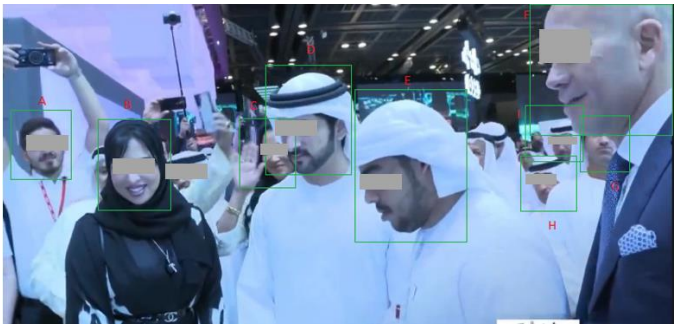


Fig. 4. Result of image sentiment classification process

2013 (FER-2013) database [20], Face Images with Marked Landmark Points [21], and Real and Fake Face Detection [22]. The FER-2013 database was the result of a challenge launched in 2013, in which Google images were obtained using 184 emotions' related keywords. The Face Images with Marked Landmark Points set [21] consist of 10k outdoor face images that were collected from the Internet and manually annotated. Each image was annotated with a bounding box and five landmarks, i.e. centers of the eyes, nose, and corners of the mouth. The Real and Fake Face Detection set [22] contains expert-generated high-quality face images where the images contained different faces.

To complement those datasets, we collected an additional set of 115,200 facial images from the Internet focusing on the occlusion category, using keywords such as partial face, occluded face and side faces. The images were manually annotated based on four features: 1) face location and reference points; 2) mouth and eyes' location and reference points; and 3) Face orientation; and 4) Occlusion area. Those features will be discussed in detail in the coming sub-sections.

A. *Faces' Location & Reference Points*

Similar to the approach used in [23], the detected face is normalized to an image of size 200 × 200 pixels. The location of each face is identified by a square using the 15 facial features and reference points that are illustrated in Table 1. Faces that are difficult to detect due to severe deformation, blurring and unrecognizable mouth or eyes, or a side angle that is less than 32 pixels of its bounding box, are labeled as "unknown".

TABLE I
FACIAL FEATURES AND REFERENCE POINTS

Left-eye-center	Right-eye-outer-corner	Nose-tip
Right-eye-center	Left-eye-brow-inner-end	Mouth - left-corner
Left-eye-inner-corner	Left-eye-brow-outer-end	Mouth - right-corner
Left-eye-outer-corner	Righteye-brow-inner-end	Mouth - center-top-lip
Right-eye-inner-corner	Right-eye-brow-outer-end	Mouth-center-bottom-lip

B. *Mouth and Eyes' Location*

While the facial location is used to delimit that facial region, accurate eyes' detection cannot be obtained using this approach [24]. The eyes' search region is limited to the top half of the detected facial region; hence the eyes always exist at the top half of the face, and the mouth at the bottom down of the face. A features' extractor for eye and mouth location is employed to two select regions – namely: the left and right eyes and the mouth regions. Table 2 presents a description of the features used in relation to the mouth and eyes locations, and their associated emotions.

C. *Face Angle & Orientation*

In this work, five different angles for facial orientation were used, including left and right sides that represent 50% of the face, front face side representing 100% of the face, left-front and right-front representing 55-75% of the face. To achieve this categorization, we use five models of a face (−90◦, +90◦, −120◦, +120◦ and 0◦) [25].

D. *Occlusion Area*

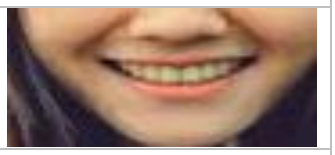
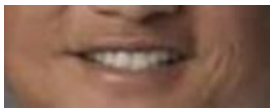
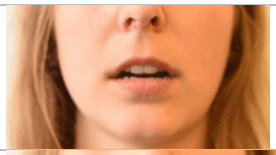
Facial occlusion refers to areas of the face that are obscured or hidden. Occluded faces are a common occurrence in real-life image in which faces may overlap or individual are standing with various angles. To detect facial occlusion, we divide a face into two major regions (eyes and mouth), then each region is sub-divided into two sides Side R (right) and side L(left). This results in the following regions: Right side eye, left side eye, right side mouth, and left side mouth. Based on the number of occluded regions, three occlusion levels were defined: 1) Level one occlusion (one or two sides); 2) Level two occlusion (three sides); and 3) Level three occlusion (four sides) which is complete face. Table 3 depicts the occlusion rules we used in our system.

Figure 5 shows examples of detected faces using the described features. As shown, the annotations consist of three locations (face, eyes, and mouth) and two facial features (orientation and occlusion area). These annotations enable the description of complete and incomplete faces and the classification of their related emotions. Figure 6 provides examples of facial occlusion and incomplete faces.

TABLE III
OCCLUSION PERCENTAGE CALCULATION RULES FOR COMPLETE AND INCOMPLETE FACES

Case	Occlusion %	Face visibility
Four corners visible: Two eyes & Two mouth sides	0%	100% (Complete)
Three corners visible: - Two eyes & one mouth sides - Two mouth sides & one eye side	25%	75% (Incomplete)
Less than three corners visible: - One eye, half of the second eye side & one mouth side - Two eyes & half of one mouth side	35%	65% (Incomplete)
Two corners visible: - One eye & one mouth side - Two eye sides - Two mouth sides	50%	50 (Incomplete)
Less than two corners visible: - One eye side & half of the second eye side - One mouth side and half of the second side	60%	40% (Incomplete)

TABLE II
MOUTH AND EYES' LOCATION AND REFERENCE POINTS

Emotion	Eye Description	Sample Image	Mouth Description	Sample Image
Laughter	Cheek Raiser, Eye Wrinkler, smaller eye		Lip Corner Puller, Upper Lip Raiser, Jaw Upper, and open mouth	
Happiness / Joy	Cheek Raiser, Eye Wrinkler		Lip Corner Puller, Jaw Upper	
Smile	Cheek Raiser, Eye Wrinkler, Upper Lip Raiser		Lip Corner Puller, Upper Lip Raiser	
Sadness	Brow Lower		Lip Corner Depressor	
Surprise	Inner Brow Raiser, Outer Brow Raiser, larger eye		Upper Lid Raiser, Jaw Drop	
Anger	Inner Brow Raiser, and Brow Lower		Upper Lid Raiser, Lid Tightener,	
Neutral	Lip Corner Puller, Dimple		Lip Corner Puller, Dimple	

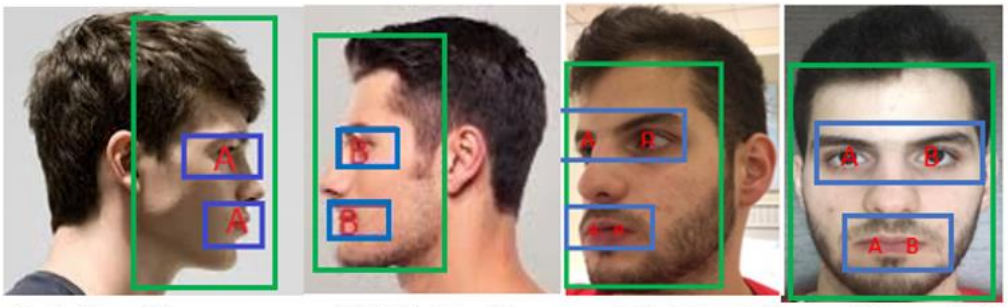


Fig. 5. Facial areas and orientation examples

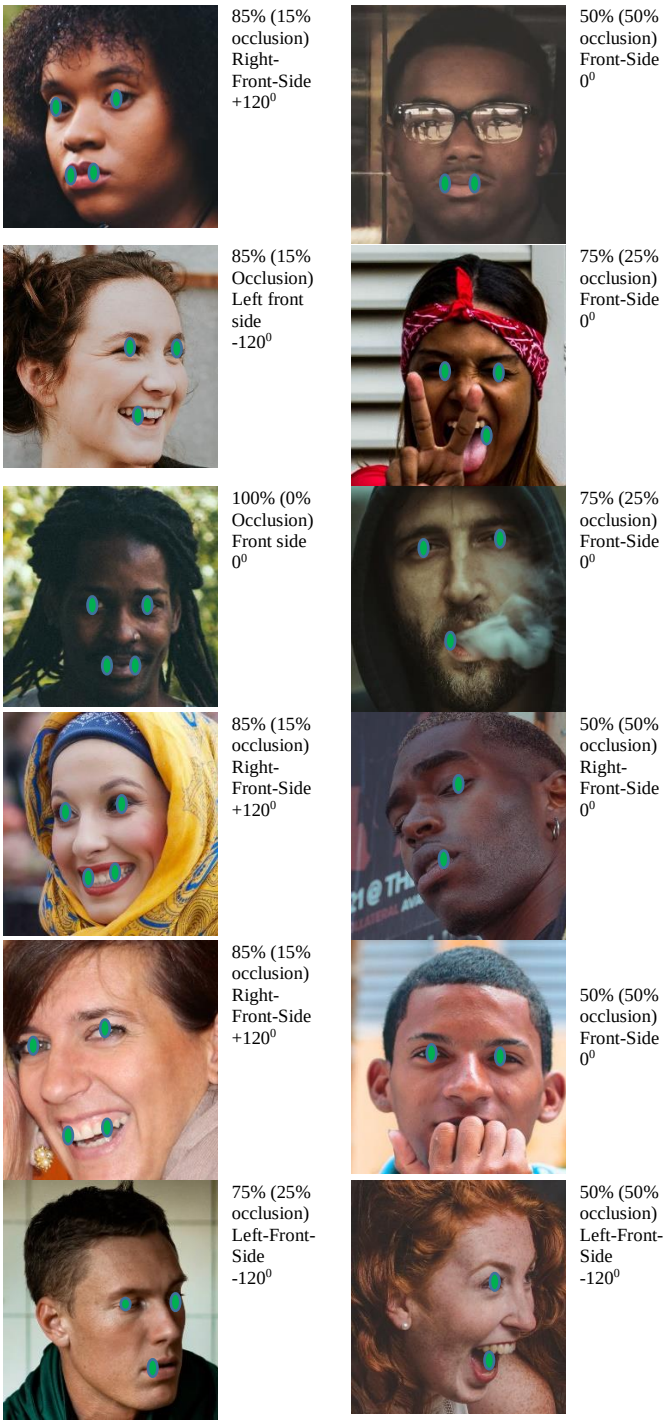
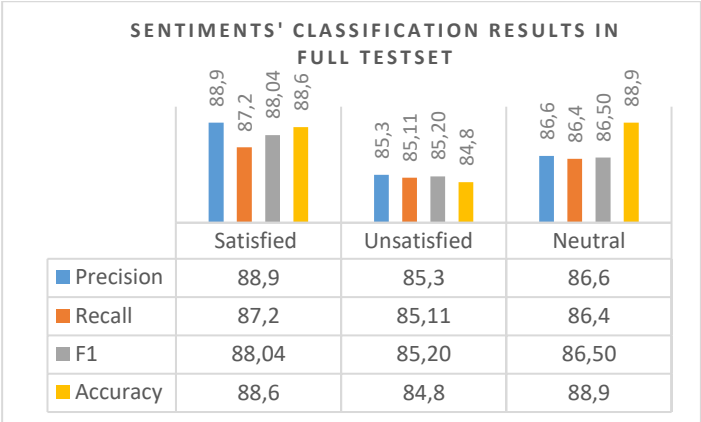


Fig. 6. Facial occlusion & incomplete faces example

III. System's Evaluation

To evaluate our model's facial emotions classification accuracy, we leveraged a video from the GITEX[1] 2019 event – a consumer electronics trade show that takes place annually in Dubai (United Arab Emirates) and attracts more than 100 K attendees. The video chosen to test our system is 7.19 minutes long, divided into 3,750 images including a total of 22,563 faces. Our analysis is restricted to the publicly available video related to that event [26], and is conduced according to ethical standards, without revealing any names or personal identifiers. In the presented image analysis, attendees' eyes were concealed with a grey bar to further protect their identity.

As first step, we tested the system's ability to distinguish between three main sentiment categories in the test dataset: satisfied, neutral, and unsatisfied. Subsequently, the dataset was divided in two subsets: images with complete faces, and images with incomplete faces to evaluate those two categories separately. Figure 7 shows the obtained results for the full test dataset. The best classification accuracy was achieved for the "Neutral" sentiment with 88.9% accuracy, followed by the "Satisfied" sentiment category with 88.6% accuracy. The "unsatisfied" sentiment category yielded an 84.8% classification accuracy.



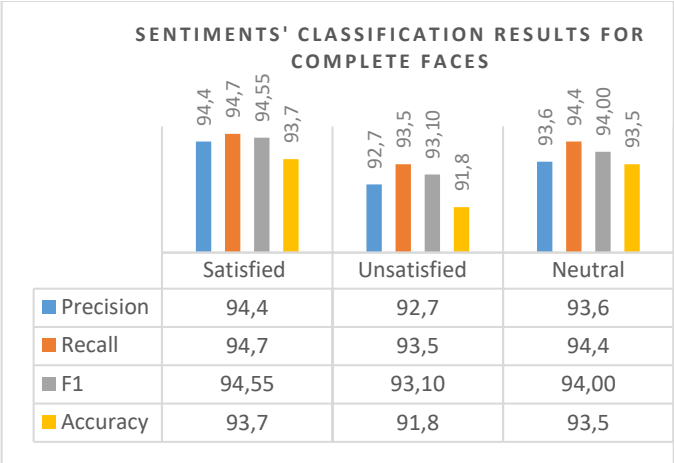
	Satisfied	Unsatisfied	Neutral
Precision	88,9	85,3	86,6
Recall	87,2	85,11	86,4
F1	88,04	85,20	86,50
Accuracy	88,6	84,8	88,9

Fig. 7. Sentiments' classification results on GITEX test dataset

When distinguishing between the sentiment analysis results for complete and incomplete faces, it is observed that higher classification accuracies were obtained for complete vs incomplete faces. This is expected as complete faces have more features/information that inform the classification decision when compared to incomplete faces. The top performing category for complete faces is the satisfied sentiment category (93.7% accuracy), followed by the neutral sentiment category (93.5% accuracy), then the unsatisfied category (91.8% accuracy). For incomplete faces, the top performing category was the unsatisfied sentiment (87.9% accuracy). Followed by satisfied (87.4%), then the neutral category (87.3%). Similar results were obtained for the precision, recall, and F1 scores, indicating that the satisfied and neutral sentiments were better detected than the unsatisfied sentiment. It is interesting to note that face occlusion favored the unsatisfied sentiment detection over the other categories, while for complete faces, the satisfied sentiment category was the favored one.

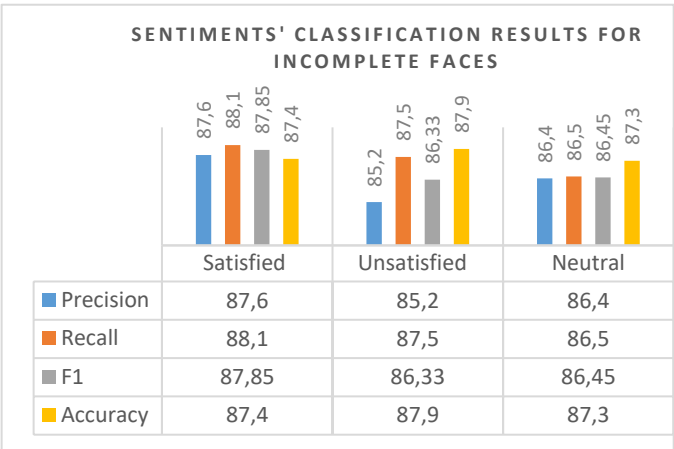
As second testing phase, we evaluated the system's performance in terms of classification accuracy across eight different emotions. The emotions' classification was performed based on the position of the eyes and the alignment of the facial features (e.g., mouth alignment, jaw lines, and smile lines). Those results are depicted in figures 10 and 11.

[1] https://www.dwtc.com/en/events/gitex-shopper-2019-2019



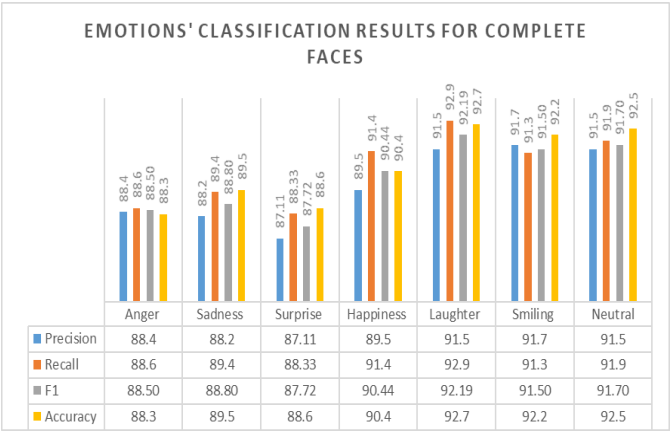
	Satisfied	Unsatisfied	Neutral
Precision	94,4	92,7	93,6
Recall	94,7	93,5	94,4
F1	94,55	93,10	94,00
Accuracy	93,7	91,8	93,5

Fig. 8. Sentiments' classification results – complete faces



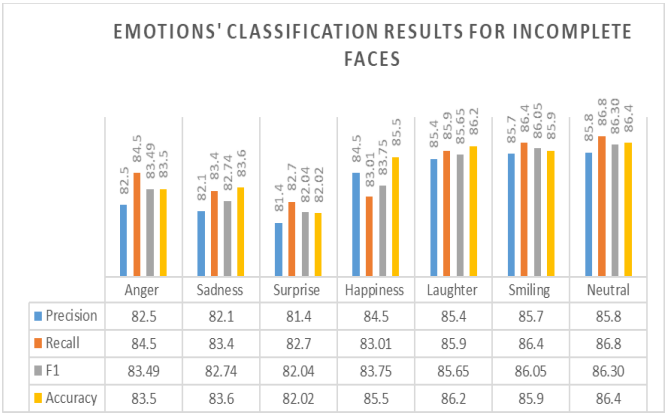
	Satisfied	Unsatisfied	Neutral
Precision	87,6	85,2	86,4
Recall	88,1	87,5	86,5
F1	87,85	86,33	86,45
Accuracy	87,4	87,9	87,3

Fig. 9. Sentiments' classification results – incomplete faces



	Anger	Sadness	Surprise	Happiness	Laughter	Smiling	Neutral
Precision	88.4	88.2	87.11	89.5	91.5	91.7	91.5
Recall	88.6	89.4	88.33	91.4	92.9	91.3	91.9
F1	88.50	88.80	87.72	90.44	92.19	91.50	91.70
Accuracy	88.3	89.5	88.6	90.4	92.7	92.2	92.5

Fig. 10. Emotions' classification results – incomplete faces



	Anger	Sadness	Surprise	Happiness	Laughter	Smiling	Neutral
Precision	82.5	82.1	81.4	84.5	85.4	85.7	85.8
Recall	84.5	83.4	82.7	83.01	85.9	86.4	86.8
F1	83.49	82.74	82.04	83.75	85.65	86.05	86.30
Accuracy	83.5	83.6	82.02	85.5	86.2	85.9	86.4

Fig. 11. Emotions' classification results – incomplete faces

When distinguishing between seven emotions, the system yielded the highest accuracies for the positive emotions related to satisfaction, such as laughter (92.7%), smiling (92.2%), and happiness (90.4%) for the complete facial images. The lowest accuracy was observed for anger (88.3%). Medium accuracies were observed for sadness (89.5%) and surprise (88.3%). For incomplete faces, classification accuracies were lower than for complete faces. The highest accuracies were achieved for laughter (86.2%) and smiling (85.9%), similar to the case of compete faces. The lowest accuracy was observed for surprise (82.02%). Medium accuracy was observed for anger (83.5%) and sadness (83.6%). The neutral emotion was detected with high accuracy for both complete and incomplete faces. This confirms the results obtained from the first testing phase, in which positive and neutral emotions were better detected than negative emotions.

To demonstrate the capabilities of our model in analyzing real-life diverse images related to large scale events, we processed three test images extracted from the GITEX event video. The results of the images' analysis are shown in figures 12, 13, and 14. As shown, a total of 28 faces were detected across the three test images.

Figure 12 is an example of low-resolution image in which nine faces were detected and classified accurately. Five of those nine faces were partially occluded. Furthermore, 4 faces were far away from the camera. Figure 13 is a high-resolution image with clear faces that are mainly facing the camera. Only three faces had angles in this test image and were still correctly classified. Finally, figure 14 is an example of a crowded scene with faces hidden behind others and appearing at different distances and angles from the camera. A total of 11 faces were detected in this image, with correctly classified emotions.

To sum up, the faces detected in those three images had different angles, distance, occlusion areas, facial orientation and resolutions. Yet, all images yielded correct classification, thus making our approach suitable for real-life applications such the analysis of events participants' engagement and satisfaction.

Face	Sad	Angry	Surprised	Neutral	Happy	Smile	Laugh	Face orientation degree	Occlusion Area
A	.19	.33	.25	.87	.56	.41	.10	0◦	Front-side
B	.03	.28	.15	.86	.62	.25	.21	−90◦	Left –side
C	.39	.61	.73	.75	.52	.32	.18	+120◦	Right-Front - side
D	.46	.68	.26	.69	.19	.05	.03	−120◦	Left-Front-Side
E	.42	.47	.11	.53	.22	.04	.03	0◦	Front-side
F	00	00	00	.16	.11	.05	.02	−90◦	Left-side
G	.09	.13	00	.17	.13	00	00	0◦	Front-side
H	.08	.16	.04	.25	.22	.07	00	0◦	Front-side
I	.07	.13	00	.28	.19	.11	00	+90◦	Right-side

Fig. 12. Emotions' classification results in GITEX test image – Low resolution, far away and facial angles case

Face	Sad	Angry	Surprised	Neutral	Happy	Smile	Laugh	Face orientation degree	Occlusion Area
A	.19	.15	.12	.71	.74	.46	.22	−90◦	Left-side
B	.03	.07	.15	.55	.80	.81	.61	0◦	Front – side
C	.19	.26	.33	.25	.71	.85	.68	0◦	Front - side
D	.46	.68	.26	.69	.19	.05	.03	0◦	Front-Side
E	.24	.37	.61	.73	.22	.14	.13	0◦	Front-side
F	.22	.19	0.28	.36	.31	.25	.12	−120◦	Left-Front-side
G	.09	.13	.28	.37	.33	18	.10	+90◦	Right-side
H	.11	.16	.14	.45	.32	.27	.21	0◦	Front-side

Fig. 13. Emotions' classification results in GITEX test image – High resolution, facing camera, clear faces case

IV. Conclusion

Envisioned as the cities of the future, smart cities combine technological advancement with quality of life and sustainability. In the smart cities context, image sentiment analysis is considered as a powerful tool for gaining insights about citizens' opinion, and their engagement and satisfaction with current events and policies. In this work, we proposed a deep learning model for conducting sophisticated sentiment

Face	Sad	Angry	Surprised	Neutral	Happy	Smile	Laugh	Face Orientation Degree	Occlusion Area
A	.09	.05	.32	.37	.84	.89	.22	+120◦	Right-Front-side
B	.13	.17	.35	.85	.70	.62	.37	−90◦	Left –side
C	.18	.06	.20	.24	.24	.15	.08	0◦	Front – side
D	.68	.28	.46	.79	.29	.15	.04	0◦	Front-Side
E	.54	.57	.71	.83	.19	.14	.09	0◦	Front-side
F	.32	.29	.38	.79	.32	.15	.02	0◦	Front-side
G	.04	.03	.08	.17	.13	.08	.02	+90◦	Right-side
H	.19	.36	.44	.75	.22	.17	.11	0◦	Front-side
I	.22	.26	.16	.33	.23	.18	.14	+90◦	Right-side
K	.02	.06	.10	.11	.07	.01	.02	0◦	Front-side
J	.16	.27	.31	.55	.34	.29	.11	0◦	Front-side

Fig. 14. Emotions' classification results in GITEX test image – Crowded scene with faces at different distances and angles case

and emotional analysis of facial images. Our model is aimed at the accurate analysis of a large variety of noisy images that can be observed in natural setting and large-scale events. Unlike existing approaches that requires clear, camera facing, high resolution images to yield accurate classification results, we proposed a sophisticated CNN model that is able to dissect images and conduct accurate emotion analysis of crowded scenes containing complete and partially occluded faces, with different angles, various distances from the camera, and varying resolutions. The proposed model was successfully tested on images collected from a large consumer electronics trade show, yielding an average classification accuracy of 90.6% when distinguishing between seven emotions in complete faces, and an accuracy of 86.16% when distinguishing between the same emotions in incomplete (or occluded) faces. Furthermore, the model was able to accurately analyze a large variety of unconstrained images showing a diversity of human face angles, orientations, occlusions, scales, poses, expressions, blur, and ambient lighting. Such results open the door for sophisticated smart cities' applications such as the analysis of events participants' engagement and satisfaction, and the capturing of public opinion. One of the limitations of our approach lies in the case of images with 60% or more occlusion. Indeed, images of 60% to 64% occlusion rates yield inaccurate emotion classification, while faces with 65% or more occlusion (i.e. missing 2 quadrants and a part of the third quadrant) do not yield any classification. This is due to the severe limitation of information available to accurately classify such largely obstructed faces. Using a hybrid classifier such as CNN with Support Vector Machines may improve the classification accuracy in such cases, although will not solve the problem completely. As future work, we plan to utilize a hybrid model to address such case. In addition, we would like

to investigate other application areas such as monitoring patients' mood in hospitals through video feed analysis, or analyzing job seekers' non-verbal communication cues during interviews to measure anxiety or confidence levels. This will necessitate new training datasets and a focus on appropriate emotion categories for such application areas.

REFERENCES

[1] Zhang, L., Wang, S. and Liu, B., 2018. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), p.e1253, DOI: https://doi.org/10.1002/widm.1253

[2] Rietveld, R., van Dolen, W., Mazloom, M. and Worring, M., 2020. What you feel, is what you like: Influence of message appeals on customer engagement on Instagram. Journal of Interactive Marketing, 49, pp.20-53. DOI: https://doi.org/10.1016/j.intmar.2019.06.003

[3] Liu, J., Gu, Y. and Kamijo, S., 2017. Customer behavior classification using surveillance camera for marketing. Multimedia Tools and Applications, 76(5), pp.6595-6622. DOI: https://doi.org/10.1007/s11042-016-3342-1

[4] Popa, M., Rothkrantz, L., Yang, Z., Wiggers, P., Braspenning, R. and Shan, C., 2010, October. Analysis of shopping behavior based on surveillance system. In 2010 IEEE International Conference on Systems, Man and Cybernetics (pp. 2512-2519). IEEE. DOI: 10.1109/ICSMC.2010.5641928

[5] De-la-Torre, M., Granger, E., Radtke, P.V., Sabourin, R. and Gorodnichy, D.O., 2015. Partially-supervised learning from facial trajectories for face recognition in video surveillance. Information fusion, 24, pp.31-53. DOI: https://doi.org/10.1016/j.inffus.2014.05.006

[6] De Melo, W.C., Granger, E. and Hadid, A., 2019, May. Combining global and local convolutional 3D networks for detecting depression from facial expressions. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-8). IEEE. DOI: https://doi.org/10.1109/FG.2019.8756568

[7] Jia, J., Wu, S., Wang, X., et al.: Can we understand Van Gogh's mood? learning to infer affects from images in social networks. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 857–860. ACM (2012). DOI: https://doi.org/10.1145/2393347.2396330

[8] Yanulevskaya, V., van Gemert, J.C., Roth, K., Herbold, A.K., Sebe, N. and Geusebroek, J.M., 2008, October. Emotional valence categorization using holistic image features. In 2008 15th IEEE international conference on Image Processing (pp. 101-104). IEEE. DOI: 10.1109/ICIP.2008.4711701

[9] Machajdik, J. and Hanbury, A., 2010, October. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM international conference on Multimedia (pp. 83-92). DOI: https://doi.org/10.1145/1873951.1873965

[10] Yuan, J., Mcdonough, S., You, Q., et al.: Sentribute: Image sentiment analysis from a mid-level perspective. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, p. 10. ACM (2013). DOI: 10.1145/2502069.2502079

[11] Borth, D., Ji, R., Chen, T., et al.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 223–232. ACM (2013). DOI: 10.1145/2502081.2502282

[12] You, Q. Luo, J., Jin, H. and Yang, J. 2015. Joint Visual-Textual Sentiment Analysis with Deep Neural Networks, Acm Mm, pp. 1071-1074. DOI: https://doi.org/10.1145/2733373.2806284

[13] Song, K., Yao, T., Ling, Q. and Mei, T., 2018. Boosting image sentiment analysis with visual attention. Neurocomputing, 312, pp.218-228. DOI: https://doi.org/10.1016/j.neucom.2018.05.104

[14] Yu, M., Tang, X., Lin, Y., Schmidt, D., Wang, X., Guo, Y. and Liang, B., 2018. An eye detection method based on convolutional neural networks and support vector machines. Intelligent Data Analysis, 22(2), pp.345-362. DOI: 10.3233/IDA-173361

[15] Zhu, X., Li, L., Zhang, W., Rao, T., Xu, M., Huang, Q. and Xu, D., 2017, August. Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition. In proceedings of the 26th international joint conference on artificial intelligence (pp. 3595-3601). DOI: https://doi.org/10.24963/ijcai.2017/503

[16] Yang, L., Ma, J., Lian, J., Zhang, Y. and Liu, H., Deep representation for partially occluded face verification. EURASIP J. Image Video Process. 2018 (1), 143 (2018). DOI: 10.1186/s13640-018-0379-2

[17] Damer, N., Grebe, J.H., Chen, C., Boutros, F., Kirchbuchner, F. and Kuijper, A., 2020, September. The effect of wearing a mask on face recognition performance: an exploratory study. In 2020 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-6). IEEE.

[18] A. Elabora, M. Alkhatib, S. S. Mathew and M. El Barachi, "Evaluating Citizens' Sentiments in Smart Cities: A Deep Learning Approach," 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 2020, pp. 1-5, DOI: 10.23919/SpliTech49282.2020.9243768.

[19] Mollahosseini, A., Chan, D., and Mahoor, M. H. 2016. Going deeper in facial expression recognition using deep neural networks. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. DOI: 10.1109/WACV.2016.7477450

[20] W. Al Rahal Al Orabi, S. Abdul Rahman, M. El Barachi, and A. Mourad, 2016. Towards On Demand Road Condition Monitoring Using Mobile Phone Sensing as a Service, in Proceedings of the 7th Elsevier's International Conference on Ambient Systems, Networks and Technologies (ANT-2016), Madrid, Spain, pp. 345-352, May 2016. DOI: https://doi.org/10.1016/j.procs.2016.04.135

[21] Koestinger, M., Wohlhart, P., Roth, P.M. and Bischof, H., 2011, November. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In 2011 IEEE international conference on computer vision workshops (ICCV workshops) (pp. 2144-2151). IEEE. DOI: 10.1109/ICCVW.2011.6130513

[22] Li, Y., Meng, J., Luo, Y., Huang, X., Qi, G. and Zhu, Z., 2020, October. Deep Convolutional Neural Network for Real and Fake Face Discrimination. In Chinese Intelligent Systems Conference (pp. 590-598). Springer, Singapore.

[23] H. Nemati, A. Singhvi, N. Kara, and M. El Barachi, 2014. Adaptive SLA-based Elasticity Management Algorithms for a Virtualized IP Multimedia Subsystem, In the IEEE Globecom Workshop on Cloud Computing Systems, Networks and Applications (CCSNA/GLOBECOM 2014), Austin, Texas, USA, December 2014. DOI: 10.1109/GLOCOMW.2014.7063377

[24] Yu, M., Tang, X., Lin, Y., Schmidt, D., Wang, X., Guo, Y. and Liang, B., 2018. An eye detection method based on convolutional neural networks and support vector machines. Intelligent Data Analysis, 22(2), pp.345-362. DOI: 10.3233/IDA-173361

[25] Hulens, D., Van Beeck, K. and Goedemé, T., 2016. Fast and accurate face orientation measurement in low-resolution images on embedded hardware. In Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016), Vol. 4, pp. 538-544. DOI: 10.5220/0005716105380544.

[26] Youtube video, 2019, October. Hamdan bin Mohammed (Fazza) opens GITEX Technology Week 2019 & Dubai. Available online at: https://www.youtube.com/watch?v=XTQYGfDep9k&t=49s

FUNDING

Zayed University RIF grant number R19099 was used to fund this work.

Dr. Sujith Samuel Mathew completed his Ph.D. in Computer Science from the University of Adelaide, South Australia. He has twenty years of experience working both in the IT Industry and in IT Academia. He has held positions as Group Leader, Technical Evangelist, and Software Engineer within the IT industry. In academia, he has been teaching various IT related topics and pursuing his research interests in parallel. Presently, he is an Assistant Professor at the College of Technological Innovations (CTI), Zayed University in Abu Dhabi, UAE.

Dr. Manar Al Khatib holds a Ph.D. in computer science from British university in Dubai, and a Master degree in computer science from Middle East university in Jordan. She is currently a research assistant at Zayed University. Her research interests include natural language processing, machine learning, and smart city applications.

Dr. May El Barachi is a next-generation networking expert holding a Ph.D. and master's degrees in Computer Engineering from Concordia University (Canada). She has 15 years' experience in the field with a strong focus on smart and resource-efficient systems. During those years, she acquired academic experience as an associate professor and has worked as a researcher with Ericsson Research Canada. Presently, she works as Associate Dean of Research at the University of Wollongong Dubai and is involved in several research collaborations with international universities and local industrial partners.