Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 1, 42-48 (2024)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Analysis of Emotions and Movements of Asian and European Facial Expressions

Ajla Kulaglic*[,1], Zeynep Örpek[2], Berk Kayı[3], Samet Ozmen[2]

[1]*College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait*

[2]*Research and Development Department, Vakif Katilim Bank, Istanbul 34000, Türkiye*

[3]*Data Engineering Department, Trendyol Group, Istanbul 34000, Türkiye*

ARTICLE INFO

Article history:
Received: 02 October, 2023
Revised: 01 January, 2024
Accepted: 01 January, 2024
Online: 20 January, 2024

Keywords:
Emotion recognition
Movement detection
Asian Facial Expression
European Facial Expression

ABSTRACT

The aim of this study is to develop an advanced framework that not only recognize the dominant facial emotion, but also contains modules for gesture recognition and text-to-speech recognition. Each module is meticulously designed and integrated into unified system. The implemented models have been revised, with the results presented through graphical representations, providing prevalent emotions in facial expressions, body language, and dominant speech/voice analysis. Current research, to identify the dominant facial emotion, involves two distinct approaches that autonomously determine the primary emotional label among seven fundamental emotions found in the input data: anger, disgust, happiness, fear, neutral expressions, sadness and surprise. The dataset utilized comprises over 292680 images sourced from the benchmark datasets FER-2013 and CK, enriched by images sourced from the Google search engine, along with 80 videos obtained during dedicated sessions, used for training and testing purposes. The Residual Masking Network (resmasknet) and CNN architectures are used as pre-trained models in this analysis. Resmasknet and CNN were chosen considering their superior performance compared to other algorithms found in the literature. The CNN architecture comprises 11 blocks, with each block containing a linear operator followed by ReLU or max-pooling layer. Starting with a convolutional layer that uses 32 filters and an 11x11x3 input, followed by a 3x3 max-pooling layer with a step of 2, the next layer includes a convolutional layer that uses 16 filters of size 9x9x16. The Residual Masking Network, contains four residual masking blocks operating on different feature sizes, each consisting a residual layer and masking block. The network initiates with a 3x3 convolution layer, followed by 2x2 max-pooling, effectively downsizing the image to 56x56. Successive transformations within four residual masking blocks generate different maps of sizes 56x56, 28x28, 14x14, and 7x7, culminating in an average pooling layer and a fully connected SoftMax output layer. The significance of this project lies in its focus on a comprehensive analysis of emotions and movements characteristic of Asian and European facial expressions. Showing promising accuracy rates, the proposed solutions achieve 75.2% accuracy for Asian and 86.6% for European individuals. This performance demonstrates the potential of this multidisciplinary framework in understanding and interpreting different facial expressions in different cultural settings.

1. Introduction

Emotional recognition (ER) enabled by artificial intelligence (AI) and machine learning (ML) models is pivotal for detecting and understanding emotions. This study introduces an integrated framework for facial emotion recognition for facial emotion recognition within the EUREKA project [1], exploring facial emotions, gestures, and social communication using AI platform. This paper is an extension of work originally presented in 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET) [2]. Recognizing

*Corresponding Author: Ajla Kulaglic, ajla.kulaglic@aum.edu.kw

https://dx.doi.org/10.25046/aj090105

emotions plays a key role for effective human-AI interaction, as more than 65% of communication is non-verbal, while the remaining 35% is influenced by verbal and physical gestures [1]. Emotional states are represented as a discrete base (the main emotions), complex categories of expression or activation of facial action units (FAC), where each action unit (AU) corresponds to a certain facial muscle movement increased with emotions [3-5]. Video-based emotion recognition is a multidisciplinary area and covers various fields of psychology, affective computing and human-robot interaction. Audio features predict emotion status from audio within video. However, as text recognition can be unexpected and confusion because it refers to the context and the language, we focused only on the implementation and analysis of the dominant emotion in the facial expression, based on detected ethnic group of the person found in the video.

Advancements in real-time applications related to face detection, recognition, face enhancement, reconstruction and facial expression have evolved [6]. The lack of all these tasks can be seen in recognizing unknown faces within test datasets. Early approaches extract geometric features using FAC, but recent focus shifted to ML and deep learning (DL) models to better detect facial movements and expression indicating emotions with extraordinary accuracy [7-10]. For non-temporal features, InceptionNet [11] and DenseNet [12] form an essential part. However, for the tasks where the video is used as input, it is not enough to consider the features of the appearance of the images in the spatial dimension. For that purpose, the information about the movement in the video is also important. The spatial feature generated by pure Deep Neural Network (DNN) is not directly suitable for videos due to the lack of motion modeling. On the other hand, Recurrent Neural Network (RNN) provides an attractive framework for propagating information over a sequence using a continuous value hidden layer representation to capture temporal features [13]. It is efficient for sequence tasks, such as action recognition, speech recognition, natural language processing (NLP), etc. [14]. Long short-term memory (LSTM) [15] developed from RNN has shown its potential in time sequence analysis and is widely used in emotion recognition analysis [16]. In [17], the author combined Convolutional Neural Network (CNN) with RNN to model the spatio-temporal evolution of visual information. In [18], the authors used CNN and bidirectional RNN architecture to learn texture appearance models for video sequences. Unlike the composite CNN-RNN structure, the 3-dimensional CNN presented in [19] uses 3-dimension convolution operations instead of 2-dimensional convolution in the network structure generating multiple channels of information from adjacent video frames. Currently, the 3D-CNN is widely used in solving various tasks of video analysis. In [16], the researchers are also using 3D-CNN proposed hybrid network for encoding information about appearance and movement in face expressions and improved the results.

The wide application of machine learning and deep learning techniques in commercial, health and security environments has created engagement and higher morale among users with an increasing representation of trained models. In this study, the implemented software is intended for the Human Resource (HR) department. The proposed solution offers several benefits in the recruitment process analyzing facial expressions and body language during video interviews, providing insights into candidates' emotional intelligence, confidence and engagement level; offering a more comprehensive understanding of a candidate's suitability for a role. Using the AI approach the bias can be reduced evaluating candidates based on their expressions and responses rather than subjective impressions. It can be used as predictive performance indicator where certain emotions and expressions found in the video might correlate with job performance. For example, a high level of positivity might indicate strong motivation. On the other hand, automated emotion recognition tools can help in the efficient selection of a large number of candidates, prioritizing those who exhibit desired emotional traits. Understanding candidates' emotions during the requirement process enables custom interactions, addressing concern or tailor communication styles based on detected emotional cues, enhancing the overall candidate experience.

The implemented software is divided into five steps. In the first step, we set the performance (personal information) and competence goals. In step 2, the person is evaluated by completing a survey. A huge number of tests are available based on areas of interest and role-based competencies. In Step 3, you have to incorporate your social behavioral cues by taking a video to capture information about your emotions, gestures, and body language using a laptop or smartphone camera. This data is recorded, saved and then processed. In the fourth step, we explore and get real-time insights through the platform. It uses machine intelligence to identify the causes of human behavior. In the final step, the final results are presented. The user can get personalized real-time insight about the gestural articulation patterns and behaviors of your competency displayed on the dashboard.

The main contributions of the proposed work can be summarized as:

1. Ethnicity extraction from facial data based on the input data;

2.Machine learning-based determination of dominant emotions in video;

3. Validation of the proposed module within the Eureka360 project.

The rest of the paper is organized as follows: section 2 discusses emotional face recognition datasets used in this study and data preprocessing performed. Section 3 details implementation specifics - data augmentation, network architectures, parameter settings, and training methods. Section 4 presents experimental validations, with conclusions in Section 5.

2. Dataset and data preprocessing

2.1. Data

The study integrates publicly available datasets, namely the Facial Expression Recognition 2013 (FER-2013) [20] and the Extended Cohn-Kanade Dataset (CK+) [21], together with images and videos gathered during the project implementation. Both, FER-2013 and CK+ datasets contain labeled images categorized into seven emotions (0=Anger, 1=Disgust, 2=Fear, 3=Happiness, 4=Sadness, 5=Surprise, 6=Neutral).

CK+ consists of a total of 981 marked pictures of 123 people. The images are placed with a similar background containing only the front face (Figure 1). FER-2013 (Figure 2) dataset consisting of 28709 labeled images in the training set, 7178 labeled images in the validation and test set together. The validation test contains 3589 images, and the test set consists of another 3589 images. The CK+ dataset does not contain a default test split. In this study, a 70-30 split was made, where 70% of CK+ images were used for model training, and 30% of images were used for testing. Happiness and surprise are the most represented feelings in the FER-2013 and the CK+ data set.

However, reference datasets mainly contain posed facial expressions captured in controlled environments, with consistent lighting and head positioning. Algorithms trained on such a dataset might underperform when faced with different preconditions, especially in natural settings. To address this limitation, additional data were collected from Google searches along with videos of project participants. A total of 291,770 Google images (Figure 3) and 80 videos were collected. The Google images were obtained by searching for expressions related to each emotion and their synonyms, expanding the dataset's diversity beyond the controlled environment. The videos were recorded during meetings and dedicated sessions to capture facial expressions in diverse environments. A total of 80 videos were collected, pre-processed and used in this project. This additional dataset encompasses various conditions, contributing to a more diverse dataset, especially in terms of ethnic representation, with a focus on Asian and European ethnicities.

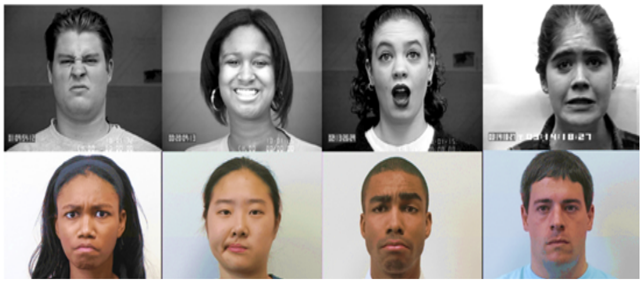


Figure 1: Emotional Expression Samples: CK+ Dataset

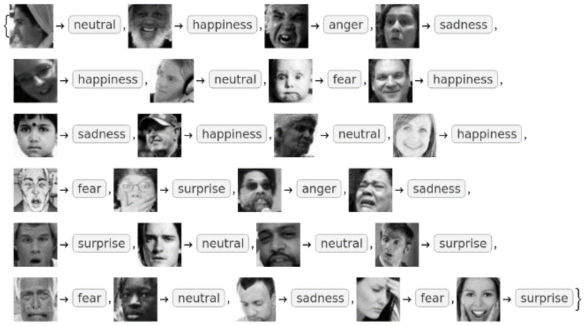


Figure 2: Instances of Facial Expressions: FER-2013 Dataset

In this study, we utilized established benchmark image datasets. However, our input data for analysis encompassed videos rather than individual static images. Video data provides richer source of data, capturing changes of facial expressions that static images might not fully capture. The videos are converted into individual frames, where each frame represents image captured at a specific point in time thought the video. Analyzing these frames sequentially enables us to reconstruct the temporal facial expressions, capturing the subtle changes in emotions displayed through the video. Transforming videos into frames for analysis, we bridge the gap between the continuous nature of video data and the discrete analysis required to interpret facial expressions, allowing us to dive deeply into the temporal dynamics of emotional displays.

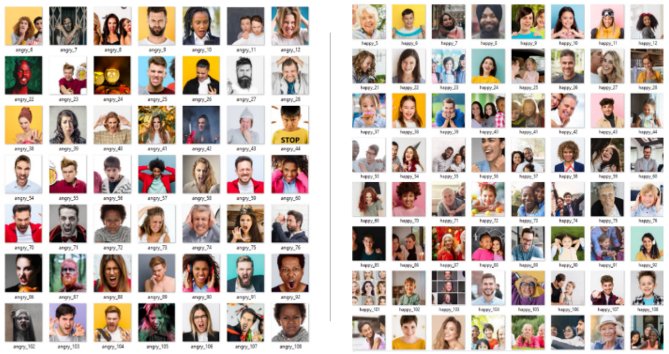


Figure 3: A collection of images obtained through a Google search

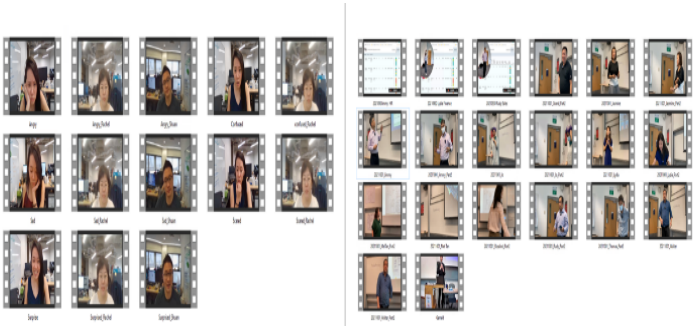


Figure 4: Instances of Collected Videos: Diverse Meetings and Special Sessions

2.2. Data preprocessing

The collected video data consists of many irregularities like changes in illumination, oclussions, perspective shifts, and scale variations. With these anomalies, it is difficult to accurately extract emotional information features from face images in videos. To address this challenge, our study aligns features using the detector to match the expected size of the front faces to the Region of Interest (ROI). We use a detector available in OpenCV based on ResNet-10, which has been shown to be a very efficient CNN model for face detection [22]. Utilizing the ResNet-10 detector, we preprocess video data by analysing it frame by frame. This process involves identifying faces withing bounding boxes, avoiding a separate face detection step. This preprocessing step enables speeding up the training process by isolating faces beforehand, eliminating the need for various data agumentation techniques such as histogram equalization, cropping, or image rotation. By aligning and padding faces withing the frames, we achieve standardized facial presentations. This alignment ensures consistency in facial sizes across varied video conditions, reducing the impact of anomalies like mentioned above, and improving the efficiency of the overall training phase.

3. Methodology

The project is designed to recognize the predominant emotion in a video, emphasizing ethnicity as a crucial factor. The input video is divided into 40 equal parts, and each of those parts have been used for analysis of dominant emotion. Additionally, it analyses signs such as eye movements, blinks, hand gestures and facial touches. Following the ethnicity detection phase, the project employs two distinct neural network models for emotion recognition- one for identifying emotions within Asian ethnicity and the other trained for European ethnicity. These networks are fine-tuned using the training set to enhance their accuracy and performance. Figure 5 illustrates the proposed framework. Bellow, we provide an explanation for each implemented module.

3.1. Asian and European face recognition

The difference in facial expressions for various emotions between Asian and European ethnicities are result of a complex interaction between cultural, societal, and individual factors. Facial recognition algorithms detect the ethnicity by checking facial landmarks, shapes and features unique to different ethnicities. Ethnicity identification poses significant challenges, as evidenced by studies found in the literature, both for humans and computer vision algorithms, especially when restricted to facial attributes [23]. According to the studies and implemented libraries, in this project we use the DeepFace Python library [24] as a solution to estimate the ethnicity in the input data. DeepFace relies on a pre-trained Convolutional Neural Network (CNN) architecture that stands out in determining ethnicity from image or video frames. The mentioned library is trained across a spectrum of data including Asian, White, Middle Eastern, Native American, Latino, and Black races, DeepFace serves as a robust tool. Once the ethnicity is detected based on the input data, the system then selectively activates an emotion recognition model trained for that specific ethnicity. This approach facilitates the accurate assessment of emotions by ethnicity, as shown in Figure 6.

3.2. Dominant emotion recognition

Similar to the ethnicity classification, emotion classification recognizes emotions that are involved in understanding facial expressions, body language, and context. The models analyze facial expressions using feature extraction techniques to capture emotions, which include identifying movements in specific facial muscles, eye expressions associated with different emotions. Upon successfully identifying the ethnicity of the subjects in the video, the project engages different models for recognizing the primary emotion. Two distinctive pre-trained network architectures are used, DeepFace and Python Toolbox for Facial Expression Analysis (Py-Feat) [24, 25]. Within the DeepFace library, an advanced convolutional neural network (CNN) model for emotion recognition is employed (Figure 7). This model has been intentionally selected based on its superior performance compared to other algorithms found in literature. Emotion identification within images and collected videos is done by a CNN architecture consisting of 11 blocks. Each block contains a linear operator followed by at least one nonlinear layer like ReLU or max pooling. The initial phase takes an image input traversing through a convolutional layer with 32 filters, followed by an 11x11x3 and 3x3 max-pooling layer with a stride of 2. Subsequently, another convolutional layer with 16 filters sized at 9x9x16 is employed to extract low-level features from edges and textures within the input image. The subsequent three layers are locally linked layers with different types of filters based on their distinctive feature map. Finally, the last two layers consist of fully connected layers aimed to establish correlation between two distant parts of the face. The network's output is a SoftMax layer, with total of 120 million parameters in this architecture. Each convolutional layer is integrated with a ReLU activation function. Additionally, dropout is used in the initial fully connected layer, and L2-regularization is employed in the final stages [25,26,27].

In contrast, the Py-Feat library introduces the Residual Masking Network (resmasknet) [28]. This network consists of four primary resmasking blocks, each representing residual masking functionality. These blocks operate on diverse feature sizes, containing both residual layer and masking block. The input image will traverse through 3x3 convolutional layer with a stride of 2, followed by a 2x2 max pooling layer, effectively reducing the image dimensions to 56x56. The subsequent feature maps, derived after a 2x2 maximum shrinkage layer, undergo transformation via four residual masking blocks, generating four distinct maps sized at 56x56, 28x28, 14x14, and 7x7, respectively. The network ends with an average pooling layer and a fully connected layer integrated with SoftMax to generate the final output, as depicted in Figure 8.

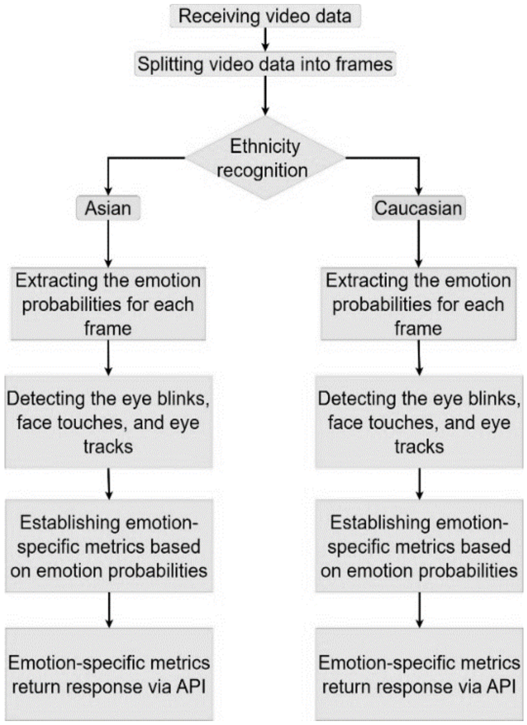
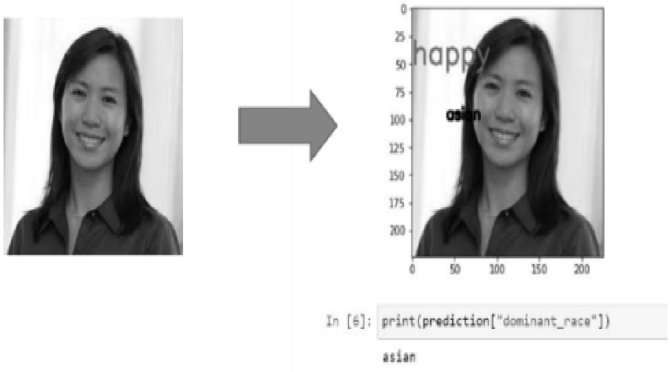
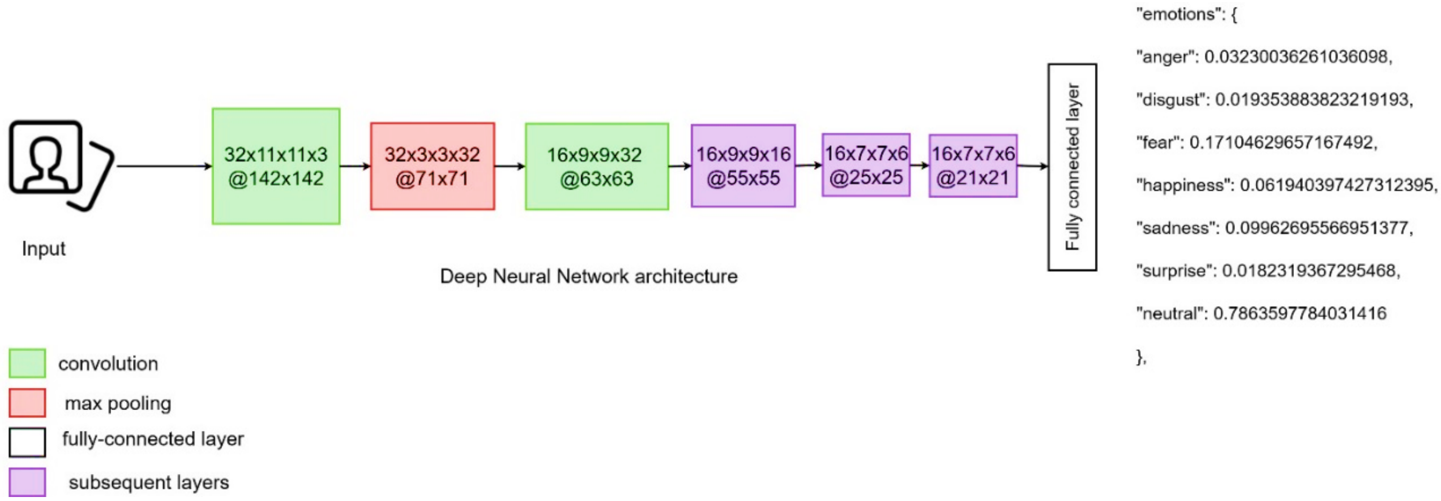

Receiving video data
Splitting video data into frames
Ethnicity recognition
Asian
Caucasian
Extracting the emotion probabilities for each frame
Extracting the emotion probabilities for each frame
Detecting the eye blinks, face touches, and eye tracks
Detecting the eye blinks, face touches, and eye tracks
Establishing emotion-specific metrics based on emotion probabilities
Establishing emotion-specific metrics based on emotion probabilities
Emotion-specific metrics return response via API
Emotion-specific metrics return response via API

Figure 5: Emotion recognition framework for Asian and European facial expressions



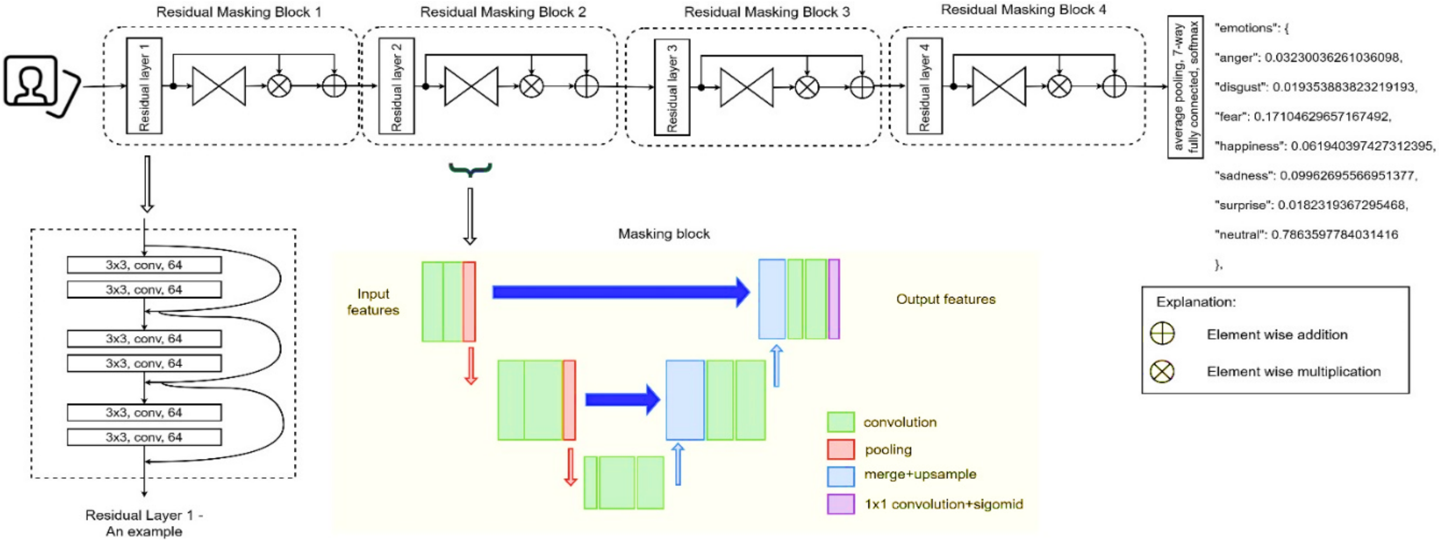
```
In [6]: print(prediction["dominant_race"])

asian
```

Figure 6: Example of dominant race recognition



```
"emotions": {
"anger": 0.03230036261036098,
"disgust": 0.019353883823219193,
"fear": 0.17104629657167492,
"happiness": 0.061940397427312395,
"sadness": 0.09962695566951377,
"surprise": 0.0182319367295468,
"neutral": 0.7863597784031416
},
```

Figure 7: Schematic representation of DeepFace – Deep Neural Network architecture

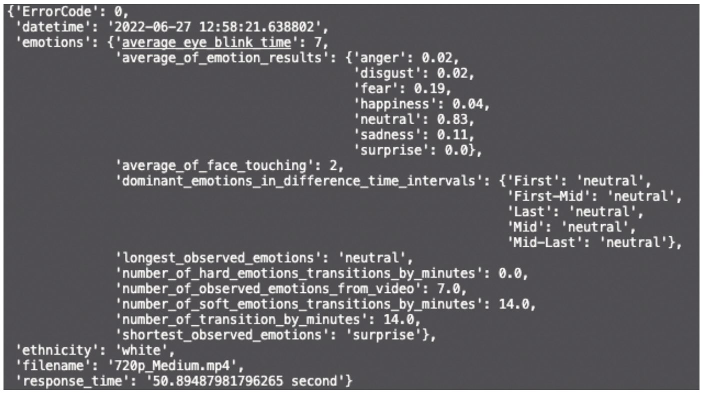


```
"emotions": {
"anger": 0.03230036261036098,
"disgust": 0.019353883823219193,
"fear": 0.17104629657167492,
"happiness": 0.061940397427312395,
"sadness": 0.09962695566951377,
"surprise": 0.0182319367295468,
"neutral": 0.7863597784031416
},
```

Figure 8: Schematic representation of Residual Masking Network (resmasknet) architecture

3.3. Eye tracking, blinking, and face touching detectors

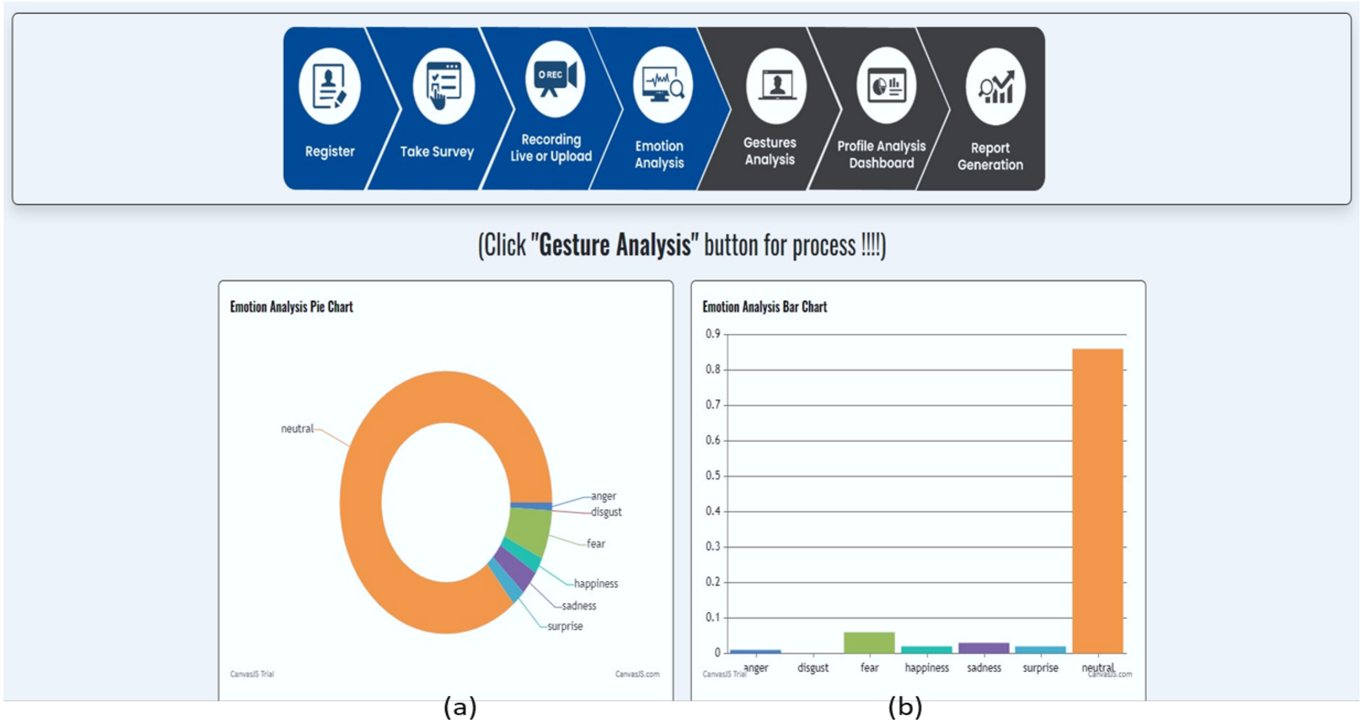
Beyond emotion analysis, this project includes examination of eye movements, blinking frequencies, and facial touches. These non-verbal signs are considered expressive elements that contribute to a better understanding of emotional states. For instance, increased eye blink rates might signify nervousness or stress. Conversely, a reduction in blink frequency might indicate focus or intense concentration. Tracking the movement of eyes can offer valuable information about attention, interest, or cognitive processing related to emotions. People use their hands during conversations to further highlight some of the emotions. Touching face can indicate anxiety, discomfort, or even deception in some context. These non-verbal cues complement facial expressions and body language in emotion recognition. Integrating eye movements, blinking and facial touches into analysis enhance the accuracy and depth of understanding emotional states.



```
{'ErrorCode': 0,
 'datetime': '2022-06-27 12:58:21.638802',
 'emotions': {'average_eye_blink_time': 7,
              'average_of_emotion_results': {'anger': 0.02,
                                             'disgust': 0.02,
                                             'fear': 0.19,
                                             'happiness': 0.04,
                                             'neutral': 0.83,
                                             'sadness': 0.11,
                                             'surprise': 0.0},
              'average_of_face_touching': 2,
              'dominant_emotions_in_difference_time_intervals': {'First': 'neutral',
                                                                 'First-Mid': 'neutral',
                                                                 'Last': 'neutral',
                                                                 'Mid': 'neutral',
                                                                 'Mid-Last': 'neutral'},
              'longest_observed_emotions': 'neutral',
              'number_of_hard_emotions_transitions_by_minutes': 0.0,
              'number_of_observed_emotions_from_video': 7.0,
              'number_of_soft_emotions_transitions_by_minutes': 14.0,
              'number_of_transition_by_minutes': 14.0,
              'shortest_observed_emotions': 'surprise'},
 'ethnicity': 'white',
 'filename': '720p_Medium.mp4',
 'response_time': '50.89487981796265 second'}
```

Figure 9: Probability results of dominant emotions in the analyzed video

The project employs a fixed time window to count occurrences of blinks and facial touches. Moreover, the project extracts diverse features including blink duration, eyelid closure time, blink speed, and hand movement speed. It is important to note that the nearness of individuals to the camera significantly influences the frequency of these observed events, resulting in higher event frequency for closer subjects.



(Click "Gesture Analysis" button for process !!!!)

Figure 10: Visualized Result Representation. Emotion Analysis Pie Chart (a) represents the results in form of pie chart. Emotion Analysis Bar Chart (b) shows seven emotions (0=Anger, 1=Disgust, 2=Fear, 3=Happiness, 4=Sadness, 5=Surprise, 6=Neutral) and the numerical value (probability) of each emotion detected in the video (y - axis).

4. Results

The Eureka Framework comprises several components, including registration, surveys/tests, video recording or uploading, emotional analysis, gesture analysis, a profile analysis dashboard, and a comprehensive report derived from the collected and analyzed data. This paper focuses exclusively on outcomes related to emotional analysis. The initial phase involves preprocessing the collected data and utilizing pre-trained models. Each data set used was analyzed and prepared separately, and at the end all data were used together for training and testing purposes. The reference data sets are recorded in controlled environments, while the collected Google images and videos contain different environments and conditions, which contributes to the diversity of the data set used in addition to ethnic diversity. The 80 videos collected and preprocessed, converting them into data frames. The first step is ethnicity detection within the input frames. A model is deployed to ascertain whether the individual in the input data is Asian or European person. Subsequently, depending on the ethnicity detected, specific models intended for different ethnic groups - Asian or European – are invoked. The accuracy of each model is separately calculated. Based on the input data in this work, the accuracy for Asian individuals stands at 75.2%, while for European individuals it reaches 86.6%. One reason for this may be due to the dataset's imbalance, where significantly more data for European individuals was available for training the models. Notably, the previously trained models from Py-Feat and DeepFace libraries, undergo tuning, involving adjustments in periods and parameters (learning rates, activation functions, etc.) to optimize their accuracy. The results for video recorded on June 27, 2022 are shown in Figures 9 and 10. The person appearing in the video is identified as European, or white. Moreover, the analysis detected 7 instances of eye blinks, 2 facial touches and a dominant neutral emotion. The overall emotion analysis, based on seven emotions (0=Anger, 1=Disgust, 2=Fear, 3=Happiness, 4=Sadness, 5=Surprise, 6=Neutral), reveals on 83% of neutrality in the video. The video segments are analyzed separately, where the dominant emotions are calculated within each segment (first, first-middle, middle, middle-last, and last). Within segments the neutral emotion was detected. The results are graphically presented in the Figure 10, offering a view of dominant emotion through the analyzed video.

5. Conclusion

The presented study demonstrates the implementation of a dominant emotion recognition model applicable across diverse sectors, notably crucial in HR for candidate assessment. Through questionnaires, competency-related tests and recorded or uploaded video introductions, an analysis based on collected data offers insights into candidate suitability. The proposed project empowers individuals to refine their skills and advance in their careers, simultaneously enabling corporations to maximize their investment in employees while exploring new opportunities. The main drawback of the above methods lies in their reliance on detecting overt facial movements, however, individuals might intentionally try to mask their emotions.

Employing AI in candidate interviews ensures unbiased selection. Leveraging unbiased datasets and interview videos, well-trained models exhibit remarkable precision in detecting dominant emotions, eye blinks, facial touches, and eye tracking. Training models with varied datasets and interview videos resulted in high accuracy in recognizing emotions. This study highlights the model's efficacy not only in research but also in practical applications across various fields.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This study was implemented thanks to the support provided by Vakıf Katılım Bank R&D Center. The opportunities offered by Vakıf Katılım R&D Center have added a huge impact on our scientific work. On this occasion, we would like to express our sincere gratitude to Vakıf Katılım Bank R&D Center for their valuable support.

References

[1] https://vimeo.com/775825642, accessed:28.12.2022

[2] B. Kayı, Z. Erbaşı, S. Özmen and A. Kulaglic, "Emotion and Movement Analysis Study from Asian and European Facial Expressions," 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Mohammedia, Morocco, 2023, 1-5, doi: 10.1109/IRASET57153.2023.10152995.

[3] DeepFace - Most Popular Deep Face Recognition in 2022 (Guide) - viso.ai, accessed:19.12.2022

[4] Y. Taigman, M., Yang, M.A, Ranzato, & L. Wolf, L.," Deepface: Closing the gap to human-level performance in face verification", In Proceedings of the IEEE conference on computer vision and pattern recognition, 1701-1708, 2014, doi:10.1109/cvpr.2014.220

[5] G. E. Dahl, T. N. Sainath, and G. E. Hinton. "Improving deep neural networks for LVCSR using rectified linear units and dropout". In ICASSP, 2013., doi:10.1109/ICASSP.2013.6639346

[6] P. M. Ashok Kumar, Jeevan Babu Maddala & K. Martin Sagayam, "Enhanced Facial Emotion Recognition by Optimal Descriptor Selection with Neural Network", IETE Journal of Research, 2021, DOI: 10.1080/03772063.2021.1902868

[7] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," in IEEE Transactions on Affective Computing, **13**(3), 1195-1215, 1 July-Sept. 2022, doi: 10.1109/TAFFC.2020.2981446.

[8] D. Kollias, M. A., Nicolaou, I., Kotsia, G., Zhao, & S. Zafeiriou, "Recognition of affect in the wild using deep neural networks", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (26-33), 2017, DOI: 10.1109/CVPRW.2017.247

[9] H., Yang, U., Ciftci, & L., Yin, "Facial expression recognition by de-expression residue learning". In Proceedings of the IEEE conference on computer vision and pattern recognition (2168-2177)., 2018, DOI: 10.1109/CVPR.2018.00231

[10] S., Xie, & H. Hu," Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks". IEEE Transactions on Multimedia, **21**(1), 211-220., 2018, DOI: 10.1109/TMM.2018.2844085

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, 2818-2826, doi: 10.1109/CVPR.2016.308.

[12] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, 2261-2269, doi: 10.1109/CVPR.2017.243.

[13] P. Scovanner, S. Ali, and M. Shah. "A 3-dimensional sift descriptor and its application to action recognition", In Proceedings of the 15th ACM international conference on Multimedia (MM '07). Association for Computing Machinery, New York, NY, USA, 357–360., 2007 https://doi.org/10.1145/1291233.1291311

[14] T. Mikolov, M. Karafiát, L. Burget, J. H. Černocký and S. Khudanpur. "Recurrent neural network-based language model." Interspeech (2010), doi: 10.21437/Interspeech.2010-343

[15] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". Neural Comput. **9**, 8 (November 15, 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[16] Y. Fan, X. Lu, D. Li, and Y. Liu. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks". In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16). Association for Computing Machinery, New York, NY, USA, 445–450. 2016, https://doi.org/10.1145/2993148.2997632

[17] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. "Recurrent Neural Networks for Emotion Recognition in Video". In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15). Association for Computing Machinery, New York, NY, USA, 467–474. 2015. https://doi.org/10.1145/2818346.2830596

[18] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun. "Multi-clue fusion for emotion recognition in the wild". In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16). Association for Computing Machinery, New York, NY, USA, 458–463. 2016, https://doi.org/10.1145/2993148.2997630

[19] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, 4489-4497, doi: 10.1109/ICCV.2015.510.

[20] FER2013 Dataset | Papers With Code, accessed:24.09.2023

[21] CK+ Dataset | Papers With Code, accessed:24.09.2023

[22] H., Kaushik, T., Kumar, & K. Bhalla, K. "iSecureHome: A deep fusion framework for surveillance of smart homes using real-time emotion recognition". Applied Soft Computing, 122, 108788. 2022. https://doi.org/10.1016/j.asoc.2022.108788.

[23] S. Fu, H. He and Z. -G. Hou, "Learning Race from Face: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, **36**(12), 2483-2509, 1 Dec. 2014, doi: 10.1109/TPAMI.2014.2321570.

[24] https://viso.ai/computer-vision/deepface/, accessed:25.09.2023

[25] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, 1701-1708, doi: 10.1109/CVPR.2014.220.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks". In Proceedings of the 25th International Conference on Neural Information Processing Systems - **1** (NIPS'12). Curran Associates Inc., Red Hook, NY, USA, 1097–1105. 2012. DOI: 10.1109/ACPR.2015.7486599

[27] G. E. Dahl, T. N. Sainath and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, 8609-8613, doi: 10.1109/ICASSP.2013.6639346.

[28] L. Pham, T. H. Vu and T. A. Tran, "Facial Expression Recognition Using Residual Masking Network," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, 4513-4519, doi: 10.1109/ICPR48806.2021.9411919.