TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/



# COMBINATION OF RETENTIVE NETWORKS AND VISIONS TRANSFORMERS FOR FACIAL EMOTION RECOGNITION IN IMAGE AND VIDEO

VELI DEMIR[1], AKUP GENÇ[2]
[1,2]UNIVERSITY: GRADUATED FROM COMPUTER ENGINEERING IN KOCAELI UNIVERSITY, KOCAELI, TURKEY, 2017

**Abstract—** The field of video sentiment analysis has grown significantly with continuous advances in artificial intelligence (AI) and machine learning (ML). In this digital age, understanding and interpreting human emotions in videos is a rapidly developing field that continues to be a matter of deep interest.
The integration of Retentive Network and Vision Transformers has launched a new path in sentiment analysis from videos, showcasing extraordinary capabilities and potential over traditional models. This article discusses the remarkable advantages, groundbreaking results, and promising future that these AI models offer in the field of video sentiment analysis.
An illustrative comparative analysis is presented showing how the combination of Retentive Network and Vision Transformers outperforms other models in terms of accuracy, adaptability and scalability. Although the functionality of these AI models has so far been explored primarily in the context of images, the potential for application to video processing and more nuanced sentiment analysis is vast and exciting.
**Index Terms—** technological models known for their capabilities, primarily used for natural language processing tasks.
**1. Vision Transformers:** Image processing models with improved adaptability to different input resolutions that provide scalability and efficiency.
**2. Video Sentiment Analysis:** The process of examining, understanding and interpreting emotions found in video data using machine learning and artificial intelligence-supported tools.
**3. Natural Language Processing (NLP):** The field of artificial intelligence that involves the interaction between computers and humans through natural language.
**4. Image Processing:** The process of performing some operations on an image to obtain an improved image or extract some useful information from it.
**5. Artificial Intelligence (AI) and Machine Learning (ML):** AI is a branch of computer science that emphasizes the development of intelligent machines that think and work like humans, while ML is a branch of AI where machines are given access to data and use that data to learn on their own.

## I. INTRODUCTION

Retetive Networks have shown promising results in NLP, their comprehensive understanding of context and ability to generate meaningful relationships make them perfectly suited for text analysis. On the other hand, Vision Transformers, which are primarily used for image processing, have proven to be remarkably effective. These models offer an exciting opportunity when integrated using Manhattan Distance, a method of calculating the distance between two points in a grid-based system such as an image or pixel.
This technology has been developed primarily for images. However, the transition to video processing with sentiment analysis holds tremendous potential. This article will show how Retetive Networks and Vision Transformers are useful for sentiment analysis from videos and examine the advantages they offer over competing models. [2], [3]
Recent studies show that integrating Retentive Networks and Vision Transformers outperforms RNNs, Transformers, and CNNs. Retentive Networks provide context understanding, while Vision Transformers offer detailed emotional indicator analysis. Unlike RNNs and CNNs focusing on facial features, the integrated model considers environmental context and details. Retentive Networks distinguish between relevant and irrelevant context, reducing computational workload.
In summary, the combination of Retentive Networks and Vision Transformers via Manhattan Distance is proving to be a promising avenue in the world of video sentiment analysis. As technologies continue to converge and innovate, it will be fascinating to observ how these models evolve and pave the way for more accurate and effective sentiment detection.

## II. WHY FACE RECOGNATION

Imagine a customer satisfaction and staff analysis system integrated into live support, online interviews, meetings, store cameras, and smart tools. This solution processes live interactions and store images, tracking customer positivity, providing representative predictions, and guiding staff. By analyzing user emotions, it enables accurate support and prioritizes customers during crises.

Live support systems will self-assessment and analyze staff through online interviews and meetings. Store cameras will assess customer emotions. Billboards with cameras can gauge reactions to ads, enhancing customer excitement and brand reputation. Smart vehicles can offer mood-based assistance, like playing music when a passenger feels low

Economic gains include real-time SaaS video and photo sentiment analysis, enhancing customer satisfaction and reducing unhappy customer loss. Companies can quickly evaluate feedback and user experiences, improving products and services. Sentiment analysis can be used in retail, healthcare, and more, optimizing business processes and market positioning

Potential commercial success lies in filling gaps in customer experience and staff insights. High-demand companies can use this technology to improve user experience and staff performance. Integration with various platforms broadens its use, making it adaptable for live support, online interviews, meetings, store cameras, billboards, and smart vehicles. Analyzing results helps businesses enhance operations and user experience, ensuring commercial success.

## III. TECHNIQUES

In this article, we will examine some of the deep learning models that are widely used in today's computer vision studies. We will consider the basic principles, Advantages and cons of these methods by considering Convolutional Neural Networks (CNN), ResNet50, EfficientNet, MobileNet, Vision Transformers and Retention Networks and VGG19 models.

**Vision Transformers (ViTs) [2], [3]:** Vision Transformers is a new family of methodologies for image recognition built on the transformer architecture, which was first introduced in NLP. ViTs parse images as sequences of patches, like tokens in text.

**Convolutional Neural Networks (CNN) [13], [14]:** CNNs are deep learning models that have achieved great success in image processing and classification problems. Basically, convolution layers are used to extract features from images.

**ResNet50 [12]:** ResNet50 is a part of the Residual Networks architecture that was introduced to overcome the learning difficulties in deep layer networks. It consists of 50 layers.

**EfficientNet [15]:** EfficientNet is proposed as a balanced way to scale model sizes (depth, width, resolution). The model family strikes a balance between computational efficiency and accuracy.

**MobileNet [16], [18]:** MobileNet is a lightweight CNN model designed to provide energy efficiency in mobile devices and embedded systems.

**VGG19 [14], [21]:** VGG19 is a CNN model that proposes a simple yet deep structure for feature extraction. As the name suggests, it has 19 layers.

Appropriate methods should be selected depending on the application purpose and system requirements. The strengths and weaknesses of each model can provide advantages in certain usage scenarios while limiting in other scenarios...

### 3.1 Distance and Similarity Measures

Choosing using Manhattan Distance with combination of Retenive Networks, and Vision Transformers for enhancing the model performance is an worthwhile method. Here are a few more ways gather statistics, along with rationale, mathematics, and Advantages'n'cons for each of them.

**Manhattan Distance:** Manhattan Distance (also known as L1 norm or taxicab distance) calculates distance between two points on a grid based on sum of the absolute differences of their coordinates. It is so named due to the grid-like data structure used by road maps, where the shortest path between two points is along the grid, unless you're out of storage space in which case it can go around.

- **Advantages:** Computational efficiency, robustness to outliers, sparsity handling, interpretability.
- **Disadvantages:** Less smooth, poor geometric sense, assumption of equal contribution, path dependence

**Euclidean Distance:** Euclidean Distance computes the distance of 2 points (or vectors) in Euclidean Space or N-dimensional space. It is applicable for the comparison of feature embeddings or patch representations.

**Cosine Similarity:** Cosine Similarity is the Cos of the angle between two non-zero vectors. It is commonly used to compare the similarity of two vectors regardless of their scales. Mathematical

**Jaccard Similarity:** Jacard similarites is a similarity measure between finite sample objects, proportion of shared sam- ples (# intersection/# union).

The Manhattan Distance has been a reasonable trade-off between the computational cost, its robustness to pair-wise outliers, and its applicability for some types of data and applications, and is thus a good candidate to connect Retentive Networks and Vision Transformers. But, whether Manhattan Distance is redundant or not, it clearly depends on the circumstances of your dataset and task. If the benefits fit your requirements, Manhattan Distance may be a good method to use.

**3.2 Color Conversion Algorithms**

Color conversion algorithms convert images from one color space to another, for example, from RGB to grayscale or HSV. This step is essential in many computer vision tasks, where some color components can be more relevant than others.

The advantages of simplifying image data with a lot of similarities include reduced redundancy in color data, meaning components do not need to be broken down for analysis, and the ability to focus on certain features like edges in large images, which aids further verification. However, disadvantages include potential loss of color information during conversion, which may be significant for the complete investigation, and its efficiency being mainly dependent on the specific task at hand.

**3.3 Data Augmentation through Noise and Rotation**

This is a technique where you artificially expand a dataset by introducing variations, in the form of noise, etc., and rotation to existing data. This is a standard approach in machine learning to improve model robustness.

The advantages of better generalization include aiding models to generalize more effectively by providing diverse data, enhancing their ability to predict outcomes on unknown data, and improved robustness by incorporating real-world imperfections such as camera noise or changes in object orientation. However, the disadvantages include higher complexity, leading to longer training times, and the risk of overfitting, where models might learn patterns induced by noise rather than significant data features.

**3.4 Normalization**

Normalization is the process of scaling data into a small range of values (e.g., [0, 1]), which can speed up convergence speed and strategies of learning algorithms.

The advantages of consistency include promoting a common scale for various characteristics, improving model training, and enhancing efficiency by speeding up convergence rates in optimization algorithms through minimizing data distribution skewness. However, the disadvantages include the potential loss of context, where important contextual information may be inappropriately lost due to lack of normalization, and challenges in implementation, as it requires careful selection of the appropriate normalization technique (e.g., min-max scaling vs. z-score)

**3.5 Feature Extraction Methods (HOG, MSER, VLFeat, SIFT)**

These techniques will be used to search for engaging patterns or traits in raw data to make models faster and better.

**HOG (Histogram of Oriented Gradients):** HOG is a popular descriptor used in computer vision and image processing aspects for object detection. It investigates how objects look and how they are built by counting appearances of gradient orientation in localized segments of a picture.

**Maximally Stable Extremal Regions (MSER):** MSER is a technique to detect blobs in images and is a generalized method capable of identifying stable regions that remain stable over a broad range of thresholds. This is particularly useful in searching for repetitive patterns in an image.

**VLFeat:** VLFeat is an open-source library featuring numerous computer vision algorithms, such as popular feature extraction algorithms SIFT and HOG.

**Scale-Invariant Feature Transform (SIFT):**SIFT finds and describes local features in images. It can stably produce keypoints invariant of scale, rotation, and affine transformations and can be adapted to multiple types of vision tasks with slight transformations.

SIFT often emerges as a strong choice for situations demanding invariance to scale and rotation, critical for matching and recognition tasks. However, for applications where speed is critical or when lighting variation is a consideration, HOG might be preferred. The choice largely focuses on task-specific needs and available computational power.

**3.6 Optimal Technology for Emotion Analysis**

For emotion analysis, the most beneficial technology might be Feature Extraction Methods like SIFT or HOG. These methods offer a nuanced understanding of visual data by emphasizing critical patterns and features, crucial in recognizing subtle changes in facial expressions. Despite being computationally intensive, their ability to highlight relevant emotional cues makes them a powerful tool in emotion recognition tasks. By accurately identifying key points and features, these methods can significantly enhance the effectiveness of emotion analysis systems.

In our study, these methods were not preferred because they would cause resource insufficiency and lead to feature loss.

**3.7 Other Technical/Technological Uncertainty and Challenges**

Image and video-based emotion analysis focuses on correctly recognizing emotional expressions. However, emotional expressions can be complex and can vary across cultures and individuals. Therefore, there are uncertainties about the accuracy and reliability of models developed to correctly classify and interpret emotions.

As the amount of data increases, inclusiveness will increase and the success rate in unlearned data will also increase, but as this much data increases, both the training time of the artificial intelligence and the response time increase.

As the amount of data and the need for success increase, the system requirement increases. The artificial intelligence learning model should be improved to solve this problem. Microservice architectures will be tried in

the presentation of the technology and it can be tried to run many containers as a single service by dividing the system requirement horizontally.
A technology should be developed where data can be sent and received instantly. The final solution can be developed by trying all the technologies such as sending this data as a rest service with TCP protocol in parts, for example asking 5mb and receiving the answer, or sending two-way packets with Web socket, for example using SignalIR, connecting to our technology live with UDP protocol and returning the analysis without disconnecting the connection.
It is aimed to solve the problem of emotion analysis from these low-quality images where there is more than one face. There are problems such as finding these faces, not having to repeat the finding process during the video, and understanding that it is the same person when they leave the area and come back.

## IV. OUR IMPLEMENTATION

Our method involves connecting retentive networks and vision transformers methods with the manhatten distance method.
**Retentive Network [2], [3],[24]:** A Retentive Network is a neural network designed to maintain context over long sequences, improving tasks like text analysis and language translation. It uses memory units, attention mechanisms, and gate functions to keep relevant information, resulting in better understanding and predictions.
**Vision Transformers (ViTs) [2], [3]:** Vision Transformers is a new family of methodologies for image recognition built on the transformer architecture, which was first introduced in NLP. ViTs parse images as sequences of patches, like tokens in text.
Manhattan Distance [2] is a technique used to measure the distance between two data points. It can be used to balance the different distances between different components of the data, especially as a binding measure. When combining Retention Network and Vision Transformers, Manhattan Distance can be used to compare or integrate different inferences.
We can explain how to use Manhattan Distance to develop the combined model in the following steps [2]:

- **Feature Extraction:**

o First, Vision Transformers and Retention Network are run separately and feature extraction is done from the data. The feature vectors obtained from these two models present different data representations.

- **Feature Comparison:**

o Manhattan Distance can be used to compare two different feature vectors obtained.
o Manhattan Distance calculates the distance between two n-dimensional vectors by summing the absolute values of the differences in each dimension. This is a simple and effective method to measure the total differences between two feature vectors.

- **Fusing and Classifying:**

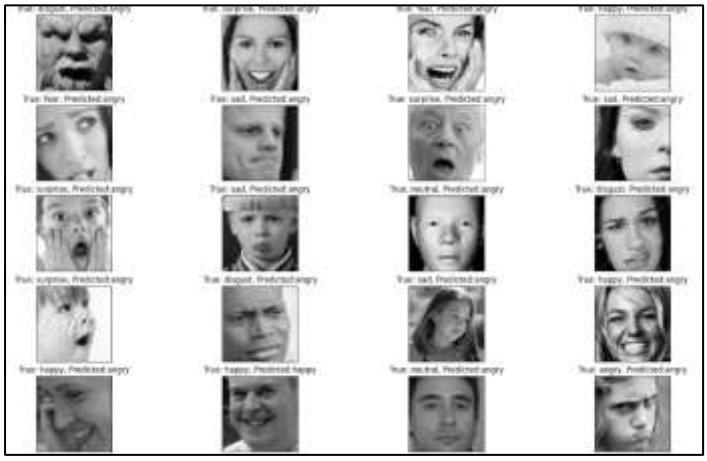
o You can use the Manhattan Distance result as the final combined feature vector or use this distance as an evaluation criterion to determine which features are more dominant.
o The combined or selected features are then fed as input to a final classification layer. The final layer works to classify facial expressions into appropriate classes.

The combination of these methods can provide benefits in image processing, especially in classifying emotions from facial data, in the following ways:

- **Improved Accuracy:** The ability to learn long-range and complex correlations helps classify emotional expressions more accurately.
- **Rich Context:** By integrating both spatial (Vision Transformers) and temporal (Retention Network) information, the entire context is taken into account.
- **Overall Performance Improvement:** Better performing models can be used more effectively and reliably in real-world applications.
- **Integrating Different Perspectives:** Manhattan Distance can help obtain more balanced and meaningful results in both spatial and temporal terms by comparing the features obtained from two models.
- **Robust and Simple:** Easy to calculate and resistant to changes, Manhattan Distance offers the opportunity to integrate complex models in a simple and understandable way.

This integration method can be effective in developing a facial expression classification system based on both historical and spatial information, while keeping the complexity at a reasonable level. Thus, it provides the opportunity to obtain a more accurate and generalizable model.

## V. FER-2013 DATASET



**Picture 1**

FER2013 (Facial Expression Recognition 2013) [6] dataset is a popular dataset created for classifying human facial expressions. It is widely used to train and evaluate facial expression recognition models, especially in the field of computer vision and machine learning.

FER2013 dataset contains a total of 35,887 grayscale images.
The images are 48x48 pixels in size. They are presented in a low resolution and standard format. The dataset has 7 different facial expression classes:

- Angry
- Disgust
- Fear
- Happy
- Sad
- Surprise
- Neutral

The FER2013 dataset has several advantages, including wide coverage of seven different emotional states, a standardized format with small grayscale images that ease data processing, and open access which makes it widely used in academic research. It also has disadvantages such as low resolution, which may be insufficient for applications requiring detailed facial expressions, unbalanced class distribution where some emotions like happiness are overrepresented compared to others like anger, and its limitation for real-world applications as the synthetic or laboratory-collected images may not adequately represent real-world situations.
In conclusion, the FER2013 dataset provides a valuable resource for initial research and model development in facial expression recognition and related fields. However, it may need to be supplemented with additional data sources and preprocessing methods for real-world applications.

## VI. Our Dataset

Our dataset contains a total of 70.000 colored images. There are 10,000 examples for each emotion.
The images are non-standart pixels in size. Since the images are obtained by cutting out the area where the face is located in the videos, the pixel size of each image is different. For example 27x39, 38x49, 51x74 etc. The dataset has same 7 different facial expression classes as FER2013.

## VII. Comparison of results For Fer-2013

Trained and tested using t4 gpu with fer 2013 data.

| Model | Training Time | Test Time (3.589 images) | Loss | Accuracy |
|---|---|---|---|---|
| RetViT (Our) | 01:01:08 | 00:00:18 | 0.75 | 0.73 |
| ViT | 01:01:08 | 00:00:18 | 0.75 | 0.73 |
| VGG19 | 00:23:28 | 00:00:10 | 0.92 | 0.69 |
| CNN | 00:11:04 | 00:00:03 | 1.16 | 0.63 |
| EfficientNet | 00:06:02 | 00:00:05 | 1.10 | 0.61 |
| ResNet 50 | 00:08:34 | 00:00:05 | 1.37 | 0.55 |
| MobileNet | 00:11:46 | 00:00:05 | 1.50 | 0.51 |

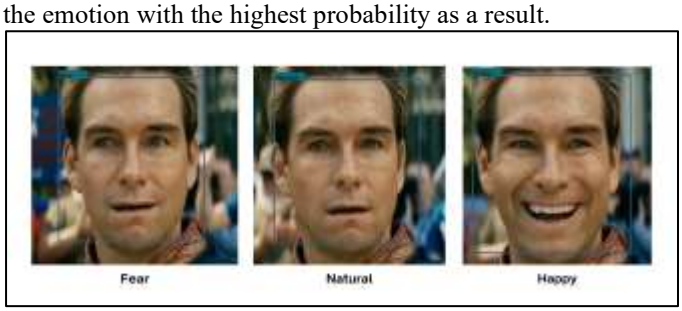
Despite the fact that both the training and testing times of the Vision Transformer and Ret+Vit models are long, they offer much higher success than their competitors.

On the other hand, Ret+Vit, compared to the alone Vision Transformer model,

- The training time requirement did not increase significantly,
- The testing time requirement increased approximately 3 times,
- The success rate increased from 71% to 73%.

## VIII. VIDEO ANALYSIS

For example, a section from the movie Homelander. I aimed to capture the faces of the characters during that section and analyze their emotions. The main purpose was to be able to use this method in online meetings.
1-second video contains 24 frames, but I worked with 7 frames to reduce the resource requirement.
I find consecutive faces for each frame. I accept faces that come approximately to the same screen area as the same face. If it exceeds a certain pixel distance or moves to a different scene, I stop the face selection.
There can be 7 frames in 1 second for a face, but if at least 3 are not found, I ignore it. Instead of waiting 7 for camera shutdown and scene changes, I accept 3 as sufficient.
I analyze 3-7 frames for 1 second faces between 3-7 and average the accuracy value for each category and give the emotion with the highest probability as a result.




**Picture 2**

## IX. FACE DETECTION TECHNIQUES

I tried YOLO [7], SSD, Faster R-CNN, MTCNN, Dlib, Mediapipe, Haar Cascades (OpenCV), RetinaFace, YuNet face detection libraries.
We got the best results with MTCNN, Haar Cascades (OpenCV), Retinaface and SSD methods. We continued with Retinaface because it required fewer resources. I did not perform an accuracy test when comparing these models. I determined it with manual tests.
Below you can see the general advantages and disadvantages of these methods.
**Haar Cascades (OpenCV):** Haar Cascades is a method in the OpenCV library that's great for real-time face detection. It uses a series of classifiers to scan through images and detect faces quickly.
**MTCNN (Multi-task Cascaded CNN):** MTCNN is a deep learning method that not only detects faces but also identifies facial features like eyes, nose, and mouth. It works well in different lighting and poses but might be a bit slow for real-time applications.
**Dlib Library:** The Dlib library excels in accurately marking facial features and handling various expressions and poses. However, it requires quite a bit of processing power, especially if you need it to work in real-time.
**YOLOv8:** YOLOv8 is a cutting-edge model designed for real-time object detection. It's quick and accurate, making it suitable for detecting faces and other objects simultaneously.
**SSD (Single Shot MultiBox Detector):** SSD is another fast method for real-time object detection and integrates well with pre-trained networks like VGG and MobileNet. It's quick but might need some tweaking for specific applications.
**Faster R-CNN:** Faster R-CNN is known for its high accuracy, especially in complex scenarios and with small objects. It's slower than other methods, so it might not be the best choice for real-time use.
**Mediapipe:** Mediapipe is a framework designed for building efficient perception pipelines. It's fast and works well on both mobile and web platforms, particularly Android and iOS.
**RetinaFace:** RetinaFace is highly accurate and can detect multiple faces at once. It requires a lot of processing power and ome expertise to customize effectively.
**YuNet:** YuNet is optimized for fast and efficient face detection, suitable for both mobile and desktop applications. Its performance can vary depending on the complexity of the task and the diversity of the training data.

## X. COMPARISON OF RESULTS FOR OUR AND FER 2013 DATASET

Trained and tested using t4 gpu with our 2 million image data from videos.

| Model | Training Time | Test Time (100.000 images) | Accuracy |
|---|---|---|---|

| | | | |
|---|---|---|---|
| RetViT (Our) | 01:32:06 | 00:00:45 | 0.80 |
| VGG19 | 00:51:49 | 00:00:10 | 0.78 |
| ViT | 02:50:42 | 00:01:32 | 0.77 |
| MobileNet | 00:36:18 | 00:00:12 | 0.71 |
| ResNet 50 | 01:32:22 | 00:00:05 | 0.65 |

The Vision Transformer and Ret+Vit models offer much higher success than their competitors. On the other hand, Ret+Vit alone, compared to the Vision Transformer model,

- The training and testing time requirement decreased by half,
- The success rate increased from 77% to 80%.

## XI. CONCLUSION

In conclusion, the combination of these two powerful methods can open new doors in the classification of emotional expressions, especially in complex datasets, and improve current model performances.
For the video, the retinaface method was preferred, considering both the source requirements and accuracy of face recognition techniques. The advantages and disadvantages of the methods were briefly mentioned in the article.
When our dataset was added to the models, the success rate continued to increase and the training time requirement of the other models did not increase much. In this case, it is predicted that as the dataset for Ret+Vit increases, the training and testing times will not increase much compared to the other models and will require much less time than Vision Transformers alone and will increse the success rate.
An application has been developed to test our models. The application detects the face in the instant camera image and shows the results of the models for this face. The application can be used during the presentation and the codes are available in the drive folder I shared with you.
Different methods may be tried due to reasons such as resource need, training period, and the problem to be solved. For example, problems such as whether the person on camera is wearing company uniform or whether the right person is driving the vehicle could be addressed with different options.
RetViT and retinaface were found to be the best methods for both image (fer-2013) and video analysis for the facial emotion analysis problem.

## REFERENCES

[1] “FER-2013.” [Online]. Available: https://www.kaggle.com/datasets/nicolejyt/facialexpressionrecognition
[2] Q. Fan, H. Huang, M. Chen, H. Liu and R. He, "RMT: Retentive Networks Meet Vision Transformers," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 5641-5651, doi: 10.1109/CVPR52733.2024.00539.
[3] Ali Hatamizadeh, Michael Ranzinger, Jan Kautz “VIR: VISION RETENTION NETWORKS”, 2023, arXiv:2310.19731v1,
[4] P. Shah, D. Ambekar, H. Bodat and S. Kumari, "Multimodal Sentiment Analysis: Techniques, Implementations and Challenges across Diverse Modalities," 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2024, pp. 405-413, doi: 10.23919/INDIACom61295.2024.10498914.
[5] P. Shah, D. Ambekar, H. Bodat and S. Kumari, "Multimodal Sentiment Analysis: Techniques, Implementations and Challenges across Diverse Modalities," 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2024, pp. 405-413, doi: 10.23919/INDIACom61295.2024.10498914.
[6] M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini and A. Lanata, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," in IEEE Access, vol. 12, pp. 45543-45559, 2024, doi: 10.1109/ACCESS.2024.3380847.
[7] S. V. Mohan Dev Vanamoju, M. V. Vineetha, H. Tekchandani, P. Joshi, P. Kumar Shukla and A. Khanna, "Facial Emotion Recognition using YOLO based Deep Learning Classifier," 2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP), New Delhi, India, 2024, pp. 1-5, doi: 10.1109/ICECSP61809.2024.10698173.
[8] M. E. Wibowo, A. Ashari, A. Subiantoro and W. Wahyono, "Human Face Detection and Tracking Using RetinaFace Network for Surveillance Systems," IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society, Toronto, ON, Canada, 2021, pp. 1-5, doi: 10.1109/IECON48115.2021.9589577.
[9] Y. Wang, L. Huang, J. Li and T. Sun, "Research on Face Detection Based on Lightweight MTCNN," 2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2023, pp. 345-348, doi: 10.1109/ITAIC58329.2023.10408812.
[10] L. Cuimei, Q. Zhiliang, J. Nan and W. Jianhua, "Human face detection algorithm via Haar cascade classifier combined with three additional classifiers," 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), Yangzhou, China, 2017, pp. 483-487, doi: 10.1109/ICEMI.2017.8265863.

[11] X. Huang and X. Cao, "Face Detection and Tracking Using Raspberry Pi based on Haar Cascade Classifier," 2022 37th Youth Academic Annual Conference of Chinese Association of Automation (YAC), Beijing, China, 2022, pp. 505-509, doi: 10.1109/YAC57282.2022.10023612.
[12] Z. Zhu and R. Jiao, "Real-time Facial Expression Recognition Research Based on Blazeface Face Detection and Resnet Emotion Classification," 2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL), Zhuhai, China, 2024, pp. 401-408, doi: 10.1109/CVIDL62147.2024.10603598.
[13] P. Asha, A. Vipulendiran, A. Kumaravelu, J. Refonaa, S. L. Jany Shabu and L. K. Joshila Grace, "Emotion Detection by Employing Deep Learning CNN Model," 2024 Second International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2024, pp. 1-6, doi: 10.1109/ICDSIS61070.2024.10594468.
[14] A. S. Negi, A. Arora, S. Bisht, S. Devliyal, B. V. Kumar and G. Kaur, "Facial Emotion Detection using CNN & VGG16 Model," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024, pp. 1-6, doi: 10.1109/I2CT61223.2024.10543618.
[15] P. Utami, R. Hartanto and I. Soesanti, "The EfficientNet Performance for Facial Expressions Recognition," 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2022, pp. 756-762, doi: 10.1109/ISRITI56927.2022.10053007.
[16] A. Nouisser, R. Zouari and M. Kherallah, "Enhanced MobileNet and transfer learning for facial emotion recognition," 2022 International Arab Conference on Information Technology (ACIT), Abu Dhabi, United Arab Emirates, 2022, pp. 1-5, doi: 10.1109/ACIT57182.2022.9994192.
[17] A. V, A. S. Bharadwaj, C. C. Bagan, D. K and S. G, "Eye-Move: An Eye Gaze Typing Application with OpenCV and Dlib Library," 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2022, pp. 952-957, doi: 10.1109/ICACRS55517.2022.10029276.
[18] P. Ranjana, K. Ramesh, S. J. S and M. B, "Face Mask Detection using Single Shot Multibox Detector and Mobile Net," 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2022, pp. 1-4, doi: 10.1109/ICATIECE56365.2022.10047292.
[19] K. Jayanthi, D. Chitradevi, N. Saranya and S. Anbukkarasi, "Group-Scanning by Face Identification and Real time Emotion detection using Faster R-CNN," 2023 12th International Conference on Advanced Computing (ICoAC), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICoAC59537.2023.10249758.
[20] W. Hua and Q. Tong, "Research on Face Expression Detection Based on Improved Faster R-CNN," 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2020, pp. 1189-1193, doi: 10.1109/ICAICA50127.2020.9182525.
[21] N. Yamsani, M. B. Jabar, M. M. Adnan, A. H. A. Hussein and S. Chakraborty, "Facial Emotional Recognition Using Faster Regional Convolutional Neural Network with VGG16 Feature Extraction Model," 2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, India, 2023, pp. 1-6, doi: 10.1109/ICMNWC60182.2023.10435819.
[22] D. Gandhi, K. Shah and M. Chandane, "Dynamic Sign Language Recognition and Emotion Detection using MediaPipe and Deep Learning," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-7, doi: 10.1109/ICCCNT54827.2022.9984592.
[23] S. V. Vasantha, B. Kiranmai, M. A. Hussain, S. S. Hashmi, L. Nelson and S. Hariharan, "Face and Object Detection Algorithms for People Counting Applications," 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2023, pp. 1188-1193, doi: 10.1109/ICACRS58579.2023.10405114.
[24] Sun, Yutao & Dong, Li & Huang, Shaohan & Ma, Shuming & Xia, Yuqing & Xue, Jilong & Wang, Jianyong & Wei, Furu. (2023). Retentive Network: A Successor to Transformer for Large Language Models. 10.48550/arXiv.2307.08621.

**Veli Demir**
I have been working on software as a computer engineer for 8 years and now I am working as a Senior Software Engineer Manager.
Educations;
- University: Graduated from Computer Engineering in Kocaeli University, Kocaeli, Turkey, 2017
- Master: Student of Computer Engineering in Gebze Technical University, Kocaeli, Turkey, [Continues]