ZIBELINE INTERNATIONAL™ PUBLISHING

ISSN: 2948-4901 (Online)
CODEN: MMHJB9

Malaysian Mental Health Journal (MMHJ)

DOI: http://doi.org/10.26480/mmhj.02.2024.60.63


CrossMark

*RESEARCH ARTICLE*

# A SURVEY OF EMOTION IDENTIFICATION TECHNIQUES BASED ON YOLO AND DEEPSORT

**Rajendra Kumar[a*], Masanori Fukui[b], Aman Anand[c]**

*[a] Department of CSE, Sharda University, Greater Noida, India.*
*[b] Iwate Prefectural University, Japan*
*[c] ITS Engineering College, Greater Noida, India*
**Corresponding Author Email: rajendra04@gmail.com*



**ARTICLE DETAILS**

***Article History:***

Received 18 September 2024
Revised 22 October 2024
Accepted 25 November 2024
Available online 09 December 2024

**ABSTRACT**

Emotion recognition or identification is one of the most complicated domains in the field of artificial intelligence and data science. Much important research work has been done on emotion identification. The primary challenge in emotion identification is the different features in the facial image and the appropriate choice of technique. In this review paper, the recent work on emotion identification and different issues related to emotion identification have been compared and analyzed. The study explores the potential of thermal imaging and anatomical parameters to enhance emotion recognition systems. This multi-faceted approach has implications for healthcare system and well-being, marketing and customer service, entertainment and gaming, automotive industry, personal devices and wearable, security, emotion analysis, and education, offering new insights and applications. In this research paper we have provided detail analysis of technique such as YOLOv3, YOLOv5 and CNN for the emotion identification.

**KEYWORDS**

Real-time emotion identification, YOLO, Deep SORT, Facial expressions, Emotion recognition, Thermal imaging, Vision-based smart classroom, Student monitoring, Real-time feedback, CNN

## 1. INTRODUCTION

Emotion identification means recognizing and understanding feelings. It is about figuring out how someone or even yourself is feeling by noticing facial expressions, body language, and tone of voice. It helps us to connect better with others by knowing if they are happy, sad, angry, scared, or surprised. Understanding emotions helps us to connect and communicate better with others. It allows us to recognize how someone feels, fostering empathy and support. Emotion identification is important in relationships, at work, and in everyday life because this allows us to give appropriate responses to the feelings of people, increasing meaningfulness and respect in interactions (Goel et al., 2024). State-of-the-art object detection models come in all shapes and sizes, but they are variants of one single backbone: YOLO divides input images into a grid and assigns each cell responsible for predicting bounding boxes and class probabilities for objects in that cell. These bounding boxes are defined by coordinates and associated with confidence scores and aim to precisely localize objects.

YOLO makes use of SoftMax activation in predicting class probabilities for each bounding box to assign labels to potential objects. It combines box confidence with class probabilities and calculates an overall confidence score to help in filtering low-certainty predictions. DeepSORT, or Simple Real Time Tracker, changed object tracking by incorporating deep features into the simple SORT framework. It integrates the identification of objects, extraction of features, data association, formation of tracks and their updates, track management, suppression of non-maximums, and production of output in a multistep process that empowers it to follow the objects steadily across the video frames. The study surveyor is going to study how the computer will understand people better by tracking them and recognizing their emotions from their faces using high-level technology that can do the job in less time with complete accuracy. The aim is the better interaction of people with computers, or to make it more sensitive to human feelings.

It is studying how computers track people in real time, hence understanding the movements of people even when the view is partly occluded by objects. It also focuses on the teaching of emotions to computers by observing a person's face. Although this technology is not perfect yet, it is getting really good at understanding emotions better than humans in some cases. It also explores using heat-sensing cameras to see how people's emotions change by measuring the temperature of their faces. This could be a better way to understand how people feel without needing them to say anything. As technology becomes more important in our lives, this research aims to make sure computers understand our emotions better. It wants to create a future where computers interact with us more effectively, helping society in various ways like improving security, education, and our emotional well-being. The study is organized into sections that look at what others have studied, explain the advanced techniques in simple terms, describe what this research does, show the results, and then give conclusions. Overall, it is about making computers smarter such that emotions can be identified easily of the people and making life easier.

## 2. LITERATURE REVIEW

Emotion identification is one of the most complicated domains in the field of computer vision and deep learning. In this paper emotion identification can be used in areas such as emotion identification from facial expression, emotion detection from speech and voice and multimodal emotion recognition. Emotion identification finds application in numerous areas which include healthcare and mental health, education market research and advertising, human-computer interaction and the entertainment

| Quick Response Code | Access this article online | |
|---|---|---|
| | **Website:** www.mmhj.com.my | **DOI:** 10.26480/mmhj.02.2024.60.63 |

Cite The Article: Rajendra Kumar, Masanori Fukui, Aman Anand (2024). A Survey of Emotion Identification Techniques Based on Yolo and Deepsort. *Malaysian Mental Health Journal, 3(2): 60-63.*

industry (Kumar et al., 2005; Kumar et al., 2023, Ng et al., 2024). It also helps in augmenting the image in several images from the same image.

In recent years the work done in emotion identification techniques such as YOLO V3, YOLO V4, YOLO V5, DeepSORT, and CNN are mentioned in below table 1.

**Table 1:** Recent contribution in the field of emotion identification

| Contribution | Dataset | Techniques | Major findings |
|---|---|---|---|
| Parambil et al., 2022 | Self-created dataset of 4250 images | YOLO v5 and DeepSort Algorithm. | Monitoring the presence of students in the classroom and analyse the students' gesture. |
| Bharathi et al., 2022 | FER2013 Containing 36087 samples of images | YOLO, Shallow Convolutional Neural Network | Identification of person's expressions with an accuracy 95.57%. |
| Gunasekar et al., 2022 | JAFFE and FER2013 Containing 36087 images | YOLO, DeepFace | Detection of face at a faster rate and analyzing the emotions of a person |
| Wu et al., 2021 | CARLA | YOLO | YOLOv5s-Ghost detects vehicles and the distance of the vehicle in real-time in the CARLA simulation environment. |
| Broussard et al., 2021 | Self-created dataset of 11 persons (8 male and 3 females, aged 22-40) | VR Technologies | Accuracy of prediction 91% |
| Chaitanya et al., 2020 | NVIE Containing 2340 facial images. | YOLO | YOLO and the Darknet framework predicted human emotions with an accuracy of 95%. |
| Zheng et al., 2020 | EIDB dataset of 10393 images for training and 1164 for testing | Combination of CNN and attention mechanism | With the verification set of EIDB-13, the prediction accuracy was observed 78%. |
| Azhar et al., 2020 | YOLOv3, YOLOv3 (tiny and custom) | System use You Only Look Once (YOLO) and DeepSORT | Successful detection and tracking of the person's movement at an average of 2.59 frames per second. |
| Pranav et al., 2020 | Self-created dataset of images of size 1920×2560. | Deep Convolutional Neural Network | The model has an accuracy of 78.04% |
| Hou et al., 2019 | A self-generated dataset named UA-DETRAC having 122234 images and 1809 features | YOLO, DeepSORT | Experimenting using UA- DETRAC test dataset achieved promising accuracy |
| Luo et al., 2019 | Self-created dataset having inspection images of the robots and the drones | Contextual- YOLOv3 | Contextual-YOLOv3 combined with YOLOV3 observed better accuracy than then original classification. |
| Zhang et al., 2017 | Self-created dataset | J48 decision tree, Random Forest, and (SVM) | Promising results in capturing and analyzing student's facial emotions. |
| Basu et al., 2017 | Datasets namely elicited emotional speech, Actor-based speech, and Natural speech. | SVM, GMM, MLP, RNN, KNN, HMM | Performed well on speech samples of different size |

## 3. Analysis and Discussion on Recent Studies

A group researcher proposed a smart class that utilized AI techniques, such as YOLOv5 for object detection and DeepSORT for object tracking, to continuously monitor students' actions and emotions during class sessions (Parambil et al., 2022). It categorized the actions of students into high and low attention categories and recognized emotions like happy, sad, angry, neutral and surprise. The system employs facial recognition technology to identify students in the classroom and maintain attendance. The system employs around 5600 images for training and testing to identify students' actions and emotions accurately. The YOLOv5 object detection model is used for training and testing, with an evaluation metric of mean average precision (mAP). The training results show a mAP of 0.734. The system is tested with eight students in a classroom setup. Each student's attention level and actions (e.g., using a phone, or raising hands) are tracked in real-time. The system provides live graphical feedback to the instructor, enabling easy tracking of students' attention levels. The system demonstrates successful real-time monitoring of student's attention and emotions.

In other study, authors proposed an approach for Human facial expression recognition that integrates YOLOv5 for people detection in real-time once a person is identified, their face is detected using Haar cascade face detection and a CNN model is specifically trained for facial expression recognition (Bharathi et al., 2022). The researchers collected a diverse dataset that included images from different sources and focused on various facial expressions, covering different age groups and skin tones, especially concentrating on Indian skin tones. It includes diverse facial expressions like Happy, Surprise, Neutral, Disgust, Fear, Sad, and Angry, focusing on Indian skin tone to enhance model efficiency. A group researcher presented a facial recognition system using EEG signals (Kumar et al., 2024). The researchers experimented with different learning rates, epochs, and batch sizes to optimize model training. This model achieves an accuracy of 95.57% for recognizing seven different facial expressions.

Some researchers explored emotion recognition through facial expression using YOLO and DeepFace algorithm (Gunasekar et al., 2022). It emphasizes face detection and emotion classification in real time and DeepFace output determines the music recommended in an application based on the predicted emotions. The YOLO algorithm segments images into cells, employing bounding boxes to detect objects and eliminate unnecessary overlaps through intersection over union evaluation. Applied to real-time face detection, it analyses images, removing irrelevant boxes to isolate faces accurately. On the other hand, DeepFace uses facial landmarks to create a 2D and 3D representation, employing convolution layers to extract facial features. It then identifies emotions by analyzing facial expressions. This method enables multi-class face recognition and analyses attributes like age, gender, mood and race. The result in this paper shows that the accuracy of YOLO is better than Haar cascade model.

In some research paper, authors focuses on enhancing automatic driving technology through testing in virtual environments, utilizing a modified YOLOv5- Ghost neural network structure (Wu et al., 2021). By adapting YOLOv5s to a more suitable form for embedded devices, the study achieved increased detection speed without significant loss in accuracy. The paper also implemented a monocular camera image-based distance detection system using the CARLA virtual environment. This approach involved generating object recognition data into distance-formula curves and achieved an average distance detection error of about 5%. In actual practice, the achieved detection accuracy was 80.76% with the best speed of 47.62 FPS against the YOLOv5s, which had 83.36% at 28.57 FPS. It also demonstrates the possibility of real-time detection and measurement of distance in simulated driving conditions, although work is still in progress to achieve higher accuracy without losing any detection speed (Broussard et al. 2021).

Other researcher mostly worked on thermal imaging for emotion recognition. This method is based on the variation of facial skin temperature due to different emotions such as stress, anxiety, and happiness (Chaitanya et al., 2020). It utilizes the YOLO algorithm for object detection in real-time situations, specifically for detecting facial features associated with various emotions. The paper describes an algorithm for eliminating the background from thermal images. Preprocessing steps involve generating landmark files in XML format, followed by training the YOLO algorithm using preprocessed images. The model achieved an accuracy of around 65% for most emotions, with exceptions noted for fear and neutral emotions.

A group researcher propose an intensity-based facial expression dataset (EIDB-13) and a recognition system that integrates a Convolutional Neural Network (CNN) and attention mechanisms in order to overcome the difficulty of subjective and qualitative evaluations in traditional teaching quality assessments (Zheng et al., 2020). The authors apply InceptionResNetV2 and introduce CBAM, an attention method, to improve feature extraction and use migration learning to alleviate overfitting in the training of deep networks on small sample datasets. They highlight the influence of emotional expression on student engagement and learning outcomes by integrating face detection and expression recognition to assess teachers' facial emotions during instruction. Their method works well, as seen by the high accuracy rates they achieve in their experiments on the public RAF-DB dataset as well as the suggested dataset. The result in this paper shows the accuracy of the model as 78%. It contributes to the field by providing a systematic method of fine-grained objective data collection in teaching evaluations and underlines face expression recognition technology for studying teaching quality through the intensity of the teachers' expressions.

A group researcher presented a people-tracking system using the DeepSORT framework while considering most of the challenges real-time tracking might face in crowded scenarios with occlusions (Azhar et al., 2020). It turns out that, implementing the YOLO algorithm to perform person detection with Deep SORT for trajectory tracking, is quite effective in surveillance and security applications. On the other hand, issues on occlusion were pointed out by the authors, as well, along with the importance of diversity in a dataset if deep learning is going to work properly. This approach consists of three major parts: user input, video processing, and visual tracking. Results of experiments include frame rates and accuracy based on different YOLOv3 datasets. It generates unique tracking IDs for every individual in the system and handles occlusion nicely, hence making it a robust system. Thus, the system detected and tracked the person's movement path with an average of 2.59 frames per second. The paper concludes by noting that integrating these into the security infrastructure will have potential impacts on society and further advance the wide frontier in applications of artificial intelligence.

Most of researchers made one of the important contributions to facial emotion recognition using a Deep Convolutional Neural Network model (Pranav et al., 2020). Underlining the accentuation of the rising role of artificial intelligence, the authors express the inability of conventional algorithms and underline the success of machine learning and deep learning in many areas, to which belongs emotion recognition. In this context, the proposed DCNN model has been designed very carefully for five different human facial emotions, which might find application fields in analysis related to customer feedback and face unlocking. It presents in-depth architectural details comprising convolution layers, dropout mechanisms, and fully connected layers. The training and validation of the model using a manually collected image dataset have been presented by the authors, showing a very impressive result with an accuracy of 78.04%. Of note is the use of an Adam optimizer and loss function of categorical cross-entropy, which this paper has iterated to be crucial in the model optimization process. The conclusion presents some avenues for future research, such as extending this model in analyzing video sequences to change emotions and integrating it with electronic devices for real-time applications. In summary, the paper contributes much to emotion recognition systems; hence, it is not devoid of theoretical underpinning but rather points at practical implications relevant to real life.

In other researcher study, author proposed a new approach to improving the Deep SORT algorithm through vehicle tracking (Hou et al., 2019). The contribution is three-fold: first, the development of a mechanism for track filtering with low confidence in order to alleviate the problems related to false-positive tracks due to unreliable detections. The authors try to improve the original Deep SORT algorithm by addressing difficult real-world situations through this extension, namely DSLCF (Deep SORT with Low Confidence track Filtering) (Luo et al., 2019). A very specific methodology has been given here, including how they have generated a self-created vehicle reidentification dataset for training the CNN of DeepSORT. The proposed method is thoroughly evaluated on the UA-DETRAC dataset, demonstrating a substantial improvement over the baseline DeepSORT algorithm.

Comparative analyses with state-of-the-art trackers underscore the effectiveness of DSLCF, showcasing its potential for applications in video surveillance and traffic monitoring. As a result, the approach not only enhances tracking accuracy but also effectively reduces false positives, contributing to the robustness of the tracking system in complex environments. YOLO, DeepSORT, and machine learning are not only used in emotion detection but also many other mental conditions like seizure detection, also in pattern recognition, mental health, etc. (Kumar at al., 2020; Kumar et al., 2021; Maisea et al., 2023; Tiwari et al., 2023; Sharma et al., 2024). A combination of all these approaches can be used in future research for batter predictions.

## 4. Conclusion and Future Scope

In this paper, the research has explored the various domains of automatic attention assessment, facial emotion detection, and real-time people tracking and it has revealed a wide range of possibilities at the heart of deep learning and technology. Expression Recognition using YOLO and Shallow CNN model identifies a person's expressions with an accuracy of 95.57%. With the help of DeepFace technology, we got faster and very good accuracy eliminating Haar cascade model. These cutting-edge methods have the potential to transform emotional health, education, and security and build a more emotionally aware and secure society. These initiatives open the door to a future where human-computer interactions are more fluid, perceptive and sensitive to the complex web of human emotions and behaviors as we continue to progress in the age of artificial intelligence.

Future works aim at augmenting the image such that the accuracy increases. The quality of the dataset can be improved using data augmentation. The system can be enhanced with the help of using a more powerful GPU such that processing can be faster and improve frame rates. Future research can also fuse two or more technologies such as YOLO and DeepSORT for emotion identification which can result in higher accuracy.

## References


Azhar, Muhamad, I.H., 2020. People tracking system using DeepSORT." 2020 10th IEEE international conference on control system, computing and engineering (ICCSCE). IEEE.

Basu, S., 2017. A review on emotion recognition using speech. 2017 International conference on inventive communication and computational technologies (ICICCT). IEEE.

Bharathi, S., Hari, K., and Senthilarasi, M., 2022. Expression Recognition using YOLO and Shallow CNN Model. 2022 Smart Technologies, Communication and Robotics (STCR). IEEE.

Broussard, D.M., 2021. An interface for enhanced teacher awareness of student actions and attention in a vr classroom. IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE.

Chaitanya, S.S., Prasanna, M., and Karthik, P., 2020. Human Emotions Recognition from Thermal Images using Yolo Algorithm. 2020 International Conference on Communication and Signal Processing (ICCSP).

Goel, A., Karim, R., Singh U., and Kumar, R., 2024. A Review On Emotion Identification Using YOLO and DeepSORT. *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, India, Pp. 430-434. doi: 10.1109/IC2PCT60090.2024.10486695.

Gunasekar, M., 2022. Improved Facial Emotion Recognition using Yolo and DeepFace for Music suggestion. 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE.

Hou, X., Yi, W., and Lap-Pui, C., 2019. Vehicle tracking using deep sort with low confidence track filtering." 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE.

Kumar, R., 2005. Human Computer Interaction. Firewall Media, ISBN: 978-81-318-0280-9

Kumar, R., Jain, V., Han, G.T., Touzene, A., 2023. Immersive Virtual and Augmented Reality in Healthcare: An IoT and Blockchain Perspective. Taylor & Francis, CRC Press, ISBN: 9781032372617

Kumar, R., Singh, R.C., Kant, S., 2020. Dorsal Hand Vein-Biometric Recognition using Convolution Neural Network, Adv. in Intell. Sys. and Comp. DOI: 10.1007/978-981-15-5113-0_92

Kumar, R., Singh, R.C., Kant, S., 2021. Dorsal Hand Vein Recognition Using Very Deep Learning, Macromolecular Symposia, 397, 2000244, pp. 1-13.

Kumar, T., Singh, R.C., and Kumar, R., 2023. Emotions Recognition Based on Physiological Signals Using Machine Learning Techniques. 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, Pp. 823-827, doi: 10.1109/ICTACS59847.2023.10390266.

Luo, H.W., 2019. Contextual-YOLOV3: Implement better small object detection based deep learning." 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE.

Maisea, S., Kumar, R., and Ali, J., 2023. A Review on CNN-based Approaches for Diabetic Retinopathy," *2023 Global Conference on Information Technologies and Communications (GCITC)*, Bangalore, India, Pp. 1-9. doi: 10.1109/GCITC60406.2023.10426348.

Parambil, M.M.A., 2022. Smart classroom: A deep learning approach towards attention assessment through class behavior detection. Advances in Science and Engineering Technology International Conferences (ASET). IEEE.

Pranav, E., 2020. Facial emotion recognition using deep convolutional neural network. 2020 6th International conference on advanced computing and communication Systems (ICACCS). IEEE.

Sharma, A., Khan, S., Chhabra, R., and Kumar, R., 2024. A Way to Detect Seizure through ML for Enhancing the Epilepsy Way of Management. *IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, India, 2024, Pp. 424-429. doi: 10.1109/IC2PCT60090.2024.10486613.

Tiwari, U., Suri, G., and Kumar, R., 2023. Analysis of Effects of E-Games on Human Cognitive and Physical Behavior. Indian Journal of Natural Sciences, 14 (78), Pp. 57752- 57761.

Wu, T.H., Tong-Wen, W., and Ya-Qi, L., 2021. Real-time vehicle and distance detection based on improved Yolo v5 network. 3rd World Symposium on Artificial Intelligence (WSAI). IEEE.

Zhang, X., 2017. Analyzing students' attention in class using wearable devices." 2017 IEEE 18th International Symposium on a world of wireless, mobile and multimedia networks (WoWMoM). IEEE.

Zheng, K., 2020. Recognition of teachers' facial expression intensity based on convolutional neural network and attention mechanism." IEEE Access 8: 226437-226444.