Deep Contextual Analysis for Enhanced Suspiciousness Estimation

Kuldeep Singh Yadav[1], Sonalika Singh[1], and Lalan Kumar[1]

[1]Affiliation not available

July 29, 2024

Abstract

Suspiciousness detection is crucial for anticipating potential security threats and facilitating timely intervention and risk mitigation. Enhanced by deep learning, vision-based systems can play a significant role in this area. This paper presents a computationally efficient vision-based suspiciousness estimation system cascading detection, classification, and analysis modules. It incorporates factors like suspicious objects, facial expressions, and abnormal body language. A suspicious object detector (SOD) is designed to precisely locate the objects invariant to the size, scale, rotation, translation, and occlusion. A deep convolutional neural network is implemented for body language and facial expression recognition with the image and landmark features as its input. To estimate the final suspiciousness, an algorithm (USE-riskometer) is proposed in this work by considering all the detected and analyzed factors. All the modules are separately trained with the corresponding object detection and facial expression datasets. In addition to this, a novel dataset for suspiciousness estimation is proposed in this work. This dataset includes various suspicious elements, such as weapons, fire, crowd presence, facial expressions, and body language in an uncontrolled environment. Each module, along with the dataset, is evaluated against state-of-the-art methods, demonstrating robustness. This work represents a significant advancement in preemptive security measures, leveraging advanced technologies for improved recognition of suspicious activities.

Deep Contextual Analysis for Enhanced Suspiciousness Estimation

Kuldeep Singh Yadav, *Member, IEEE*,Sonalika Singh, *Student Member, IEEE*, and Lalan Kumar, *Member, IEEE*

***Abstract*—Suspiciousness detection is crucial for anticipating potential security threats and facilitating timely intervention and risk mitigation. Enhanced by deep learning, vision-based systems can play a significant role in this area. This paper presents a computationally efficient vision-based suspiciousness estimation system cascading detection, classification, and analysis modules. It incorporates factors like suspicious objects, facial expressions, and abnormal body language.**

A suspicious object detector (SOD) is designed to precisely locate the objects invariant to the size, scale, rotation, translation, and occlusion. A deep convolutional neural network is implemented for body language and facial expression recognition with the image and landmark features as its input. To estimate the final suspiciousness, an algorithm (USE-riskometer) is proposed in this work by considering all the detected and analyzed factors. All the modules are separately trained with the corresponding object detection and facial expression datasets.

In addition to this, a novel dataset for suspiciousness estimation is proposed in this work. This dataset includes various suspicious elements, such as weapons, fire, crowd presence, facial expressions, and body language in an uncontrolled environment. Each module, along with the dataset, is evaluated against state-of-the-art methods, demonstrating robustness. This work represents a significant advancement in preemptive security measures, leveraging advanced technologies for improved recognition of suspicious activities.

***Index Terms*—Suspicious activity recognition, object detection, FER, body language analysis, deep learning.**

I. Introduction

A. Background

IN In the modern era of rising crime rates, various public places, including airports, banks, hospitals, and railway stations, are subjected to extensive video surveillance. This surveillance not only captures human activities but also serves to identify objects that may pose potential threats. Vigilance for suspicious behavior, such as unauthorized activities, loitering, murders, kidnapping, road accidents, and fire incidents, is paramount for ensuring the safety of individuals and their environments.

Due to the sheer volume of image and video data, manual processing and analysis become impractical. Computer vision, an artificial intelligence discipline, empowers computers to interpret and analyze visual input. It examines and understands digital images, extracting high-dimensional data from the real world and prompts appropriate actions. Leveraging algorithms, computer vision can scrutinize image and video data, making informed decisions about the surrounding environment. It enables a more secure and efficient visual content analysis and monitoring.

Various researchers have explored computer vision and deep learning in designing systems across several domains, such as human activity recognition (HAR) [1]–[4], image and video captioning [5], fall detection [6], audio-based activity recognition [7], traffic management [8], object localization [9], region segmentation [10], and data classification [11]. However, relatively few research efforts have focused on suspicious activity recognition (SAR) [12]–[15], crime detection [16], and early threat detection [17]. These AR systems involve identifying and classifying human actions or activities from various data sources, such as video, audio, or sensor data. Acquiring audio or sensor data in uncontrolled environments is particularly complex. Vision-based acquisition devices offer greater flexibility to capture relevant information, which can then be applied in these systems for analysis. The existing activity recognition (AR) systems consist of feature extraction and classification modules. The feature extraction module incorporates various methods for feature extraction, including traditional feature descriptors like the histogram of oriented gradients (HOG), local binary patterns (LBP), and spatio-temporal interest points (STIP). Classification techniques include support vector machines (SVM), multilayer perceptrons (MLP), artificial neural networks (ANN), convolutional neural networks (CNN), long short-term memory (LSTM), and bidirectional LSTM (BLSTM). These systems have been deployed and evaluated in controlled environments for specific applications.

However, there is still a need for comprehensive analysis and system design for suspiciousness estimation with the involvement of suspicious objects, human body language, and emotions in uncontrolled environments.

B. Related Work

Nowadays, HAR is a key area to explore in computer vision and machine learning that focuses on specific problems. Over the years, it has seen significant advancements, transitioning from traditional handcrafted features engineering to advanced deep convolutional feature engineering [10], [13]. Additionally, HAR leverages multiple data modalities such as RGB video, depth sensors, infrared, and inertial measurement units (IMUs) to enhance performance and robustness.

Kuldeep Singh Yadav is with the MSP LAB, Electrical Engineering, Indian Institute of Technology Delhi, India. (e-mail: aksyadav@ee.iitd.ac.in).

Sonalika Singh is with the MSP LAB, Electrical Engineering, Indian Institute of Technology Delhi, India. (e-mail: sonalika@ee.iitd.ac.in).

Lalan Kumar is with the Department of Electrical Engineering, Bharti School of Telecommunication, Yardi School of Artificial Intelligence, IIT Delhi, New Delhi, India (e-mail: lkumar@ee.iitd.ac.in).

Manuscript received July 22, 2024

Sun et al. [1] provided a comprehensive analysis on HAR approaches, categorizing them into handcrafted feature-based and deep learning-based approaches. Handcrafted features, such as HOG, scale-invariant feature transform (SIFT), and STIP, have been benchmarked but are limited in capturing complex spatiotemporal patterns and generalizing across uncontrolled environments. In contrast, deep learning-based methods leverage CNNs, RNNs, and 3D CNNs to directly extract and learn more prominent features from the images. It helps to better generalization and learning of the network to handle variations like translation, rotation, and scaling at a certain level. Despite these advancements, challenges such as handling variations in human actions, pose, scaling, orientation, occlusions, illumination, dealing with low-resolution and blurred images, integrating multimodal data, and achieving real-time performance persist.

Song et al. [2] introduced the modality compensation network (MCN), which enhances action recognition by compensating for lost or corrupted data modalities. It utilizes adversarial learning to align features from different input modalities, such as RGB, depth, and skeletal data. The most important feature of this approach is that it ensures consistent performance even when some modalities are unavailable or noisy. The architecture of the network includes a feature extractor, modality-specific generators, and a discriminator to enforce cross-modal feature alignment. The primary challenges addressed include handling missing or noisy modalities, ensuring the generalization of the network, and maintaining computational efficiency.

Wang et al. [4] presented a novel approach using depth video to capture spatial and temporal information for action recognition. This method converts depth video frames into voxel grid representations, capturing the 3D structure and motion over time. 3D CNNs then process these voxel grids to extract spatiotemporal features, which are used for action classification. This technique leverages depth information to handle challenges such as occlusions and varying viewpoints. However, it faces significant challenges, including efficiently processing high-dimensional voxel grids, ensuring robustness to noise in-depth data, and achieving real-time performance. The method also needs to handle variations in action execution and differences in physical environments.

Researchers [15] introduced an innovative framework that preprocesses the data to manage missing values and integrates diverse data, i.e., text, image, video, and audio. They tried to address the pervasive challenge of analyzing multimodal data that is often incomplete or partially paired, a common issue in real-world scenarios such as security surveillance, financial fraud detection, and cybersecurity. Their algorithm combines statistical and machine learning techniques to identify anomalies by detecting deviations from normal patterns.

Wu et al. [11] introduced an algorithm to handle suspicious data (in the form of text) and code that could indicate potential security threats, emphasizing dynamic defense mechanisms. While their dynamic defense model adapts to emerging threats, it primarily focuses on detecting anomalies in system activities rather than incorporating human-centric indicators of suspiciousness.

Similarly, Jiang et al. [14] proposed a system to recognize suspicious behaviors in social networks, financial transactions, and network security by identifying anomalies in interaction patterns and transaction amounts. Although their general metric and algorithms are versatile, they primarily rely on quantitative data analysis and may overlook more subtle indicators of suspicious activities.

Researchers [17] contribute to this field by proposing an algorithm for early threat detection through suspicious behavior representation. While their approach emphasizes real-time monitoring and pattern recognition, it primarily focuses on data-driven behavioral analysis rather than incorporating non-traditional indicators of suspiciousness.

In study [18], a deep-CNN framework was introduced to detect abnormal human behavior from standard RGB images. To distinguish suspicious object entities from detected objects, a detection and classification head was designed. The spatial features of abnormal behavior were extracted through a posture classification module connected to LSTM for effective abnormal behavior detection. Nevertheless, the LSTM-based strategy demands substantial computational resources.

In a separate study [19], researchers employed an image segmentation-based method for real-time detection of potentially suspicious behaviors within a shopping mall setting. This involved applying a blob fusion technique to segmented subjects, facilitating the detection of suspicious objects. Subsequently, a feature-based tracking algorithm was deployed to monitor the suspicious target. However, this approach was assessed under limited scenarios and could face challenges if the target object disappeared in the intermediate frames.

To address computational efficiency and design a real-time suspicious activity detection model, another study [12] leveraged the YOLOv3 model. This research focused on evaluating three anomalies—lock-breaking, bag-snatching, and wallet-stealing—across five different subjects in diverse backgrounds. A comparable methodology was adopted by [9] to develop an abnormal behavior detector supporting smart surveillance. Their algorithm addressed walking and running scenarios by segregating the video background from its internal objects and eliminating noise through morphological operations. In all such abnormal behavior detection, a quantitative and objective measure for a given image is a missing factor. Additionally, the evaluation of this work requires further scrutiny and exploration in uncontrolled scenarios with a larger dataset.

It is observed that most of the existing literature on suspiciousness primarily focuses on identifying abnormal patterns in the data to flag potential threats. However, there is a notable gap in utilizing more effective parameters of suspiciousness, such as the availability of suspicious incidents, objects, specific body language, and emotions. These factors can enhance the effectiveness of the system and reduce the computational complexity. Localizing suspicious objects and giving more attention to that region improves the accuracy of the overall system.

C. Objectives and Contributions

To address these challenges, we propose a novel algorithm (USE-Riskometer) to identify suspiciousness in a given scene

by incorporating two cascaded modules. The output of the cascaded modules serves as input for the algorithm, quantifying suspiciousness in the surroundings. The first module, an object detector, identifies factors contributing to suspiciousness that include the count of individuals, weapons, fire, and facial expressions. Recognizing an individual's emotions can significantly impact threat levels, the second module incorporates a facial emotion and body language classification model. This model, employing a residual network strategy, discerns emotions such as anger or sadness, crucial for threat assessment. To train these modules effectively, we introduce a robust database tailored for suspiciousness estimation in uncontrolled environments. This algorithmic approach offers a promising advancement in the automated detection of suspicious activities, enhancing security measures in diverse scenarios.

II. Proposed Research

A complete system for uncontrolled suspiciousness estimation based on computer vision and deep learning (DeepUSEvision) is presented in this research work. It is comprised of four major modules, i.e., input acquisition, detection, analyzer, and discriminator network. The overall pipeline of this system is presented in Fig.1.

A. Data Acquisition

1) Data Collection: To integrate the findings of this research, we introduce a novel dataset called IITD-USE (Indian Institute of Technology Delhi - Uncontrolled Suspiciousness Estimation) dataset, sourced from diverse platforms including social media and manual collection. The dataset includes objects such as persons, faces, fire, and weapons, allowing for a comprehensive analysis of the visual elements associated with suspicious content. It also involves body language and facial expressions providing valuable insights into the behavioral cues that may contribute to the overall assessment of suspiciousness. It comprises approximately 15,000 images, introducing variations in scale, resolution, and illumination to simulate real-world conditions. This diversity ensures the robustness of the models trained on this data, allowing them to generalize effectively across a spectrum of environmental settings. Additionally, we utilize the FER20E (Facial Expression Recognition) database [20] having around 100,000 images from 20 facial expressions. This specialized dataset is designed to train FER models robustly.

2) Data Annotation using Deep Automated Annotation Tool: Annotating large datasets is a labor-intensive task that requires significant manual effort and substantial computational resources. It becomes more complex in computer vision for object localization and classification due to several challenges. Manually assigning the location of each object takes a large amount of system memory. As the number of iterations increases, the system becomes slower, impacting overall efficiency. Additionally, the process demands continuous visual focus from human annotators. This constant need to keep eyes fixed on the screen leads to eye strain and potential health issues over prolonged periods. Furthermore, the complexity of accurately annotating large volumes of data results in substantial processing time.

To overcome these challenges, a deep automated annotation (DAA) tool is proposed. It involves the customization of the Python library LabelImg, transitioning it from a manual to an automated process through the integration of a lightweight backend network altered for the specific database. This backend network, comprised of a streamlined architecture with eight layers, is initially trained to recognize key elements such as people, faces, weapons, and fire, forming the foundation for comprehensive image annotation. This strategic training minimizes the need for extensive manual input during annotation. To ensure high precision in the annotations, a meticulous monitoring system was implemented for the incorrect annotations and rectified the misannotated images, maintaining the integrity and accuracy of the annotated dataset. This hybrid approach, merging automated annotation with manual oversight, strikes a balance between efficiency and

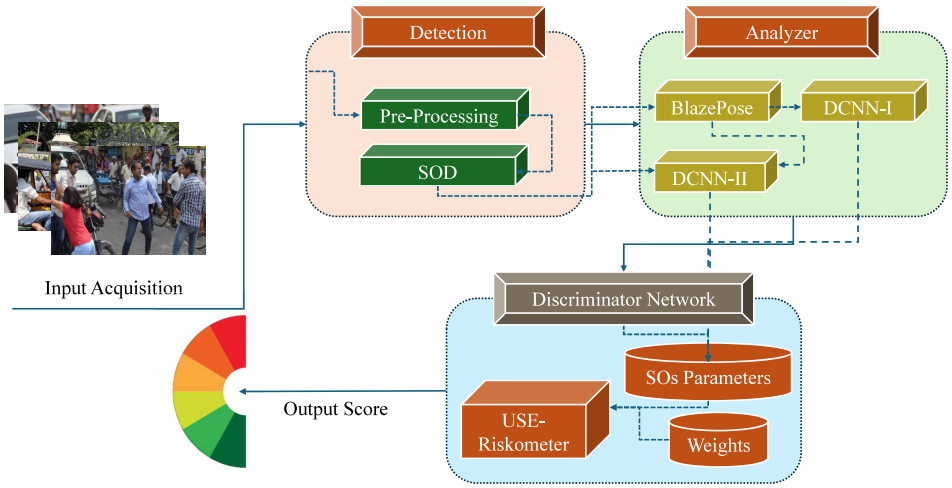

Fig. 1. Block diagram of the proposed DeepUSEvision system

precision, optimizing the annotation process for large and complex datasets. The GitHub repository of DAA tool will be made available at [21].

3) Data Augmentation: Data augmentation stands out as an essential strategy to minimize both underfitting and overfitting in neural networks. In this research work, various augmentation techniques were employed to improve the resilience and capabilities of the model. It includes diverse transformations, such as translation (ranging from 0 to 10 pixels along both axes), scaling (varying from 0 to 10 pixels along both axes), flipping, and the introduction of noise (gaussian noise with intensities ranging from 0.01 to 0.03). By incorporating translation, we enable the model to adapt to variations in spatial positioning, simulating real-world scenarios where the content's placement may vary. Scaling variations further contribute to the ability of the model to handle diverse sizes of visual elements within the images. The inclusion of flipping operations enhances the robustness of the model by exposing it to mirrored representations of the data. Additionally, noise injection, achieved through the introduction of Gaussian noise with intensities ranging from 0.01 to 0.03, imparts a layer of randomness to the data.

B. Detection

The detection module contains a pre-processing block, which improves the quality of input data using traditional image processing techniques such as contrast stretching, histogram equalization, and morphology. In addition, the image is resized as per the input layer of the suspicious objects detector (SOD). To design the SOD model, an approach was adopted from an advanced object detection modality, YOLOv8 [22]. Four classes, i.e., person, face, fire, and weapon are considered in this work. We have divided the SOD model into three distinct phases the backbone, neck, and head. Each segment has a specific role, e.g., the backbone is responsible for feature extraction, the neck processes and refines these features, and the detection head predicts bounding boxes, object classes, and confidence scores. The block diagram of the SOD is presented in Fig.2.

The backbone of SOD is comprised of convolutional layers to capture spatial hierarchies, residual blocks to mitigate the vanishing gradient, and spatial pyramid pooling fast (SPF) to enhance the detection of the objects by combining feature maps from different layers and overcoming the scale variation. Also, it utilizes a bottleneck with 2 convolutions (c2f) for robust feature extraction. For an input feature map X with dimensions $C \times H \times W$, where C is the number of channels, and H and W are the height and width, respectively, the c2f feature map is calculated as

$$Z = g(\text{concat}(f(X_1), X_2)) \tag{1}$$

where $X_1, X_2 = \text{split}(X)$, $f(\cdot)$ represents a series of convolutional operations, and $g(\cdot)$ represents the final transformation.

This structure allows for efficient feature extraction and processing, leading to faster inference and maintaining accuracy in the SOD model. The neck of the SOD is the interconnected module between the feature extractor and the detection head.

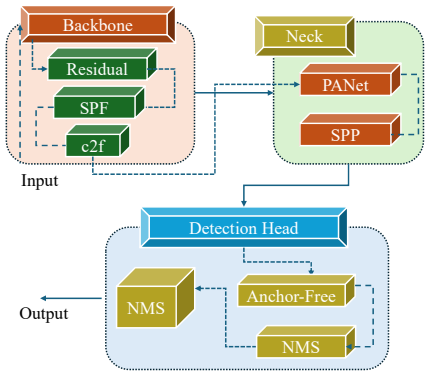

Fig. 2. Block diagram of the SOD model

It comprises a path aggregation network (PANet) and spatial pyramid pooling (SPP). For the combination and selection of more prominent features, PANet employed adaptive feature pooling through a series of pooling operations that aggregate contextual information at multiple scales. In addition, it utilizes a bottom-up path to extract robust features and preserve the spatial information unlike traditional feature pyramid networks (FPNs). SPP layers pool these features at different scales, ensuring the network is invariant to the size. For a given level l with $k_l \times k_l$ bins, the bin size for pooling is calculated as:

$$b_h = \left\lceil \frac{H}{k_l} \right\rceil, \quad b_w = \left\lceil \frac{W}{k_l} \right\rceil$$

The pooled value for each bin is:

$$P_{i,j}^{(l)} = \max_{m,n} F\left(i \cdot b_h + m, j \cdot b_w + n, d\right) \tag{2}$$

where $0 \leq i < k_l$, $0 \leq j < k_l$, $0 \leq m < b_h$, $0 \leq n < b_w$, and $0 \leq d < D$. The pooled features from all levels are concatenated to form a single feature vector. The total number of pooled values is:

$$\sum_{l=1}^{L} k_l^2$$

The final feature vector V has the size:

$$\left(\sum_{l=1}^{L} k_l^2\right) \times D$$

where D represents the depth (number of channels) of the feature map.

The detection head in the SOD model predicts object locations and classifications directly without relying on predefined anchor boxes. After extracting features through the backbone and SPP layers, an enhanced feature pyramid network (FPN) processes these features across multiple scales to ensure effective object detection. The FPN outputs a heatmap where each pixel represents the probability of being the center of an object:

$$P_{\text{center}} = \sigma(\text{Conv}(F_{\text{FPN}})) \tag{3}$$

where, F_{FPN} denotes the feature map from the FPN, and σ is the sigmoid function.

For each predicted center point, the network regresses the bounding box coordinates relative to that point, using predicted offsets. These offsets compute the actual bounding box coordinates as shown in equation 4.

$$(b_x, b_y, b_w, b_h) = W_{\text{bbox}} \cdot F_{\text{FPN}} + b_{\text{bbox}} \tag{4}$$

where W_{bbox} and b_{bbox} are the weights and biases for bounding box regression.

For each bounding box, the model predicts class scores C. The class scores are converted into probabilities using the softmax function:

$$C = \text{Softmax}(W_{\text{cls}} \cdot F_{\text{FPN}} + b_{\text{cls}}) \tag{5}$$

where W_{cls} and b_{cls} are the weights and biases for class prediction. The confidence score s for each bounding box, indicates the likelihood that the box contains an object. It is measured by:

$$s = \sigma(W_{\text{conf}} \cdot F_{\text{FPN}} + b_{\text{conf}}) \tag{6}$$

where W_{conf} and b_{conf} are the weights and biases for confidence prediction.

Finally, the output consolidates the bounding box coordinates, class probabilities, and confidence scores for each detected center point:

$$(b_x, b_y, b_w, b_h, C, s)$$

C. Facial Expression and Body Language Analyser

The approach for analyzing body language and facial expressions initiates with the capture of 33 landmarks of body and face using the BlazePose landmark detection model [23]. The initial ten landmarks belong to facial features, while the remaining landmarks define the body. These precise landmarks serve as input vectors for subsequent backend networks, shaping the foundation of the DeepUSEvision system. As the backend network for body language analysis and facial expression recognition, a lightweight deep network is employed to extract the features by adopting the design principles of the MobileNetV3 architecture. This innovative network is structured with a stem convolutional layer, succeeded by five inverted residual blocks with input specifications of 56x56x16, 28x28x24, 28x28x24, 14x14x48, and 7x7x96.

The inverted residual blocks utilize depthwise separable convolutions and linear bottlenecks, complemented by squeeze-and-excitation (SE) blocks. The incorporation of SE Blocks enhances the network's learning efficiency by capturing channel-wise dependencies and adaptively recalibrating features. It helps to reduce the computational complexity and maintains the superior performance. The final layer of the network is dedicated to classification tasks. This network is leveraged for two primary tasks in this research: facial expression recognition (DCNN-I) and body language analysis (DCNN-II). This backend network is separately trained using the FER20E and IITD-USE databases for each task. In the case of facial expression recognition, the input for the DCNN-I is a fusion of features extracted from both the face image and facial landmarks (0-10). This fusion of landmark features significantly augments the efficacy of the network in discerning emotional states. It includes seven basic facial expressions, i.e., Anger, Disgust, Fear, Happy, Neutral, Sad, and surprised (Shocked and Amazed as per FER20E).

Similarly, for body language analysis, DCNN-II is trained for two classes - normal and abnormal body language. The input vector comprises all 33 landmarks, providing a holistic representation of body language cues for robust classification. This training approach ensures the proficiency of the backend network in capturing nuanced aspects of facial expressions and body language, contributing to the effectiveness of our overall scene analysis model.

D. Discriminator Network

The discriminator network is the final module for suspiciousness estimation. A USE-Riskometer algorithm is presented within this. It evaluates the level of suspiciousness in a scene by measuring factors like the number of individuals, detected facial expressions, presence of weapons, fire detection, and observed body language. The process begins with normalization, which adjusts the input parameters, i.e., number of people, emotion scores, number of weapons, fire detection, and body language into a common scale. This ensures consistency and comparability across different data types. For instance, the presence of weapons or fire is normalized into binary values, highlighting their significance.

Following normalization, weights are assigned to each parameter based on their relative importance. For example, abnormal body language and combined emotions of anger and fear are given higher weights due to their strong correlation with suspicious behavior.

After the weight assignment, this algorithm calculates a comprehensive risk score by combining the normalized parameters with their respective weights. This involves multiplying each parameter by its weight and summing the results, producing a single risk score that integrates all factors. Finally, this suspiciousness score S_{USE} is scaled to a 0-10 range, providing a standardized measure of suspiciousness that is easy to interpret. The steps involved in computing S_{USE} are shown in the proposed algorithm 1 called USE-Riskometer. Each step in this process ensures that significant risk indicators are accurately captured and balanced, resulting in a reliable assessment of crowd suspiciousness. This structured approach allows for quick and effective decision-making in response to potential risks, enhancing security and safety measures.

This adaptable algorithm allows for adjustments in parameters and weights, ensuring relevance across diverse situations. Continuous refinement based on real-world feedback is encouraged for improved accuracy.

III. Results and Discussion

A. Experimental Setup

The experiments are conducted on MATLAB and Python platforms, utilizing a robust system configuration comprising an Intel i7-13650HX processor, 64 GB RAM, and an NVIDIA Quadro P5000 GPU equipped with 16 GB of memory. This high-performance computational setup ensured the efficiency

Algorithm 1 USE-Riskometer

Input:

Persons (N_{person}) ▷ Number of persons in the crowd
Emotions (E_{ccore}) ▷ Emotion scores
Weapons (N_{weapons}) ▷ Number of detected weapons
Fire ($F_{\text{detection}}$) ▷ Fire detection flag
Body Language (B_{L}) ▷ Normal / Abnormal

Output:

Suspiciousness (S_{USE}) ▷ Score on a scale of 0 to 10

Step 1: Normalize Parameters

$$Nm_P = \text{norm}(N_{\text{person}})$$
$$Nm_{Em} = \text{norm}(E_{\text{score}})$$
$$Nm_{Wp} = \begin{cases} 1 & \text{if } N_{\text{weapons}} > 0 \\ 0 & \text{otherwise} \end{cases}$$
$$Nm_F = \begin{cases} 1 & \text{if } F_{\text{detected}} = \text{True} \\ 0 & \text{otherwise} \end{cases}$$
$$Nm_{BL} = \begin{cases} 1 & \text{if body language is abnormal} \\ 0 & \text{if body language is normal} \end{cases}$$

Step 2: Assign Weights

$$w_P = 0.3$$
$$w_{Em} = \begin{cases} 0.6 & \text{if both (Anger \& Fear) detected} \\ 0.4 & \text{if Anger or Fear detected} \\ 0 & \text{otherwise} \end{cases}$$
$$w_{Wp} = \begin{cases} 0.3 & \text{if weapons detected} \\ 0 & \text{otherwise} \end{cases}$$
$$w_F = \begin{cases} 0.1 & \text{if fire detected} \\ 0 & \text{otherwise} \end{cases}$$
$$w_{BL} = \begin{cases} 0.5 & \text{if body language is abnormal} \\ 0 & \text{if body language is normal} \end{cases}$$

Step 3: Calculate Risk Scores

RiskScore = $(Nm_{\text{P}} \times w_{\text{P}})$ + $(Nm_{\text{Em}} \times w_{\text{Em}})$ + $(Nm_{\text{Wp}} \times w_{\text{WP}})$ + $(Nm_{\text{F}} \times w_{\text{F}})$ + $Nm_{\text{BL}} \times w_{\text{BL}})$

Scale to 0-10:

$S_{\text{USE}} = \left(\frac{\text{Risk Score}}{\text{Max Risk Score}}\right) \times 10$

and reliability of extensive data analysis and comprehensive model evaluation.

All the proposed models, including suspicious object detection, facial expression recognition, body language detection, and suspiciousness estimation are evaluated separately to show the efficacy of the individuals. For object detection, some of the publicly available benchmark datasets, such as multi-person pose (MPII) [24], Market-1501 [25], and Adience [26]are also considered with the proposed IITD-USE dataset. Similarly, FER20E [20], FER2013, The LNMIIT, and Affectnet are utilized to evaluate the proposed FER model. These databases are systematically organized into distinct sets, i.e., training, validation, and test aligning with the specific requirements of the designed models. Further details about the dataset organization are provided in Table I, offering a comprehensive understanding of the datasets employed for our experimentation and model evaluation. This diversified dataset selection ensures a thorough examination of our models across various scenarios, enhancing the robustness and generalizability of the research outcomes.

TABLE I
DATASET DISTRIBUTION FOR TRAINING, VALIDATION, AND TESTING

Database	Train	Val.	Test	Model
Datasets for SOD				
MPII [24]	15,000	3,000	7,000	Person, Face
Market [25]	13,000	3,000	6,000	Person, face
Adience [26]	15,000	3,000	7,000	Person
IITD-USE	9,000	1,500	3,500	All
Datasets for SOD				
The LNMIIT [27]	800	300	400	Expression
FER2013 [28]	20,000	5,000	10,000	Expression
FER20E	20,000	5,000	8,000	Expression
AffectNet [29]	200,000	100,000	110,000	Expression

B. Suspicious Object Detection

1) Evaluation of the SOD model over the Proposed and Publicly Available Datasets: This experiment focuses on evaluating the performance of the SOD with initial training on the IITD-USE dataset, encompassing four classes, i.e., face, person, fire, and weapons. The model is trained on the training set with hyperparameter tuning, utilizing a mini-batch size of 8 and training for 100 epochs. Increasing the mini-batch size yielded marginal improvements but incurred higher computational costs. Moreover, the batch loss stabilized after 90 epochs.

During training, various data augmentation techniques such as scaling, cropping, rotation, jittering, flipping, random erasing, and mosaic are incorporated to improve the robustness and performance of the SOD. These augmentations are applied in various combinations to create a diverse set of training samples, enhancing the ability of the model to generalize to new unseen data. The training and validation performance is shown in Fig. 3 and Fig. 4.

While applying the SOD model on the test set of IITD-USE, it achieves a mean average precision (mAP) of 0.877 at a 50% intersection over union (IoU), and 0.0599 across a range of IoU thresholds from 50% to 95%. The model achieved this level of performance on the IITD-USE dataset with an inference time of 6.2ms. The MPII and Market-1501 datasets contain only face and person as the objects in the controlled environment. The SOD model provides higher mAP (0.905 and 0.933) on these datasets compared to the IITD-USE dataset. However, the Adience dataset consists of one class only, i.e., person and the SOD model provides a mAP of 0.94. All these datasets characterized by unidirectional and limited variations in comparison to the IITD-USE dataset, resulted in higher mAP values, surpassing 0.9, as indicated

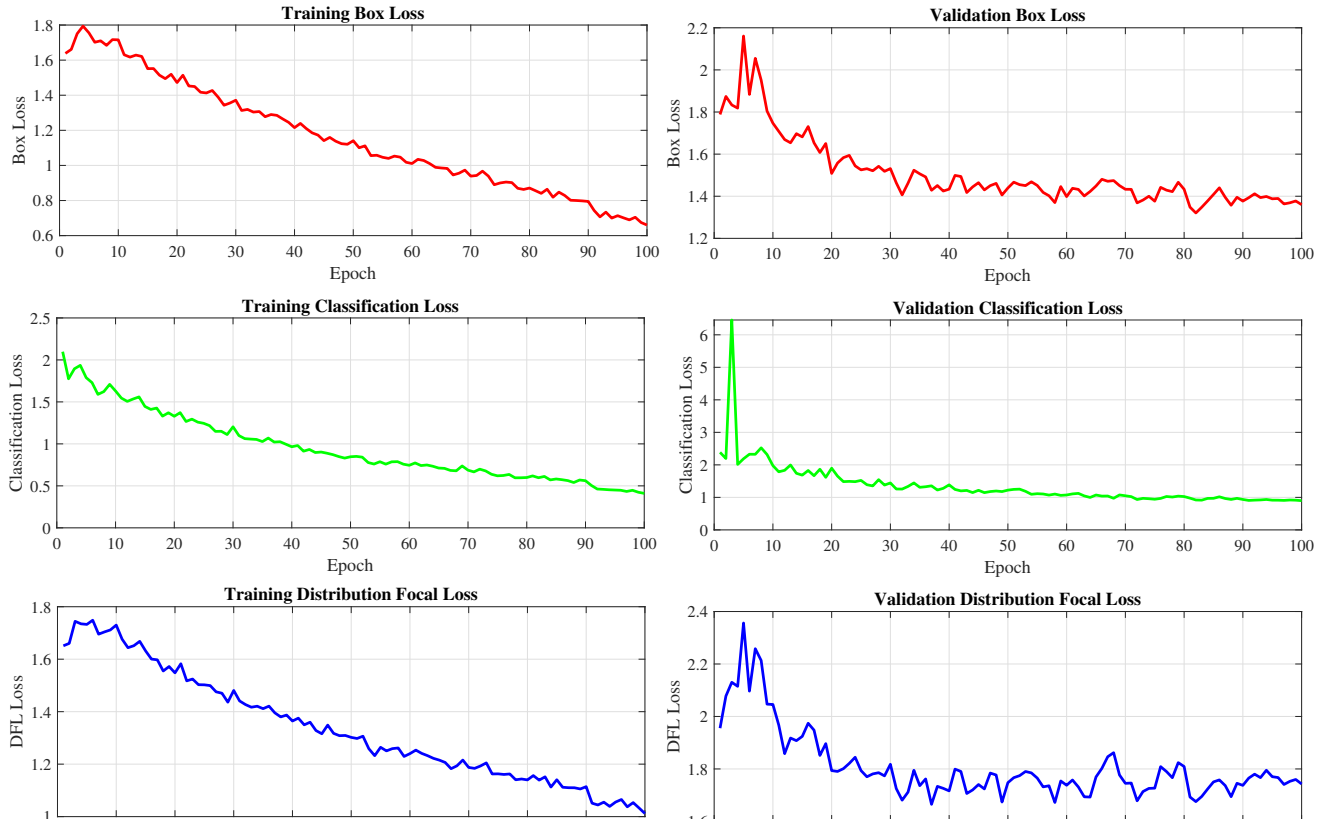

Fig. 3. Various losses of the SOD over epochs

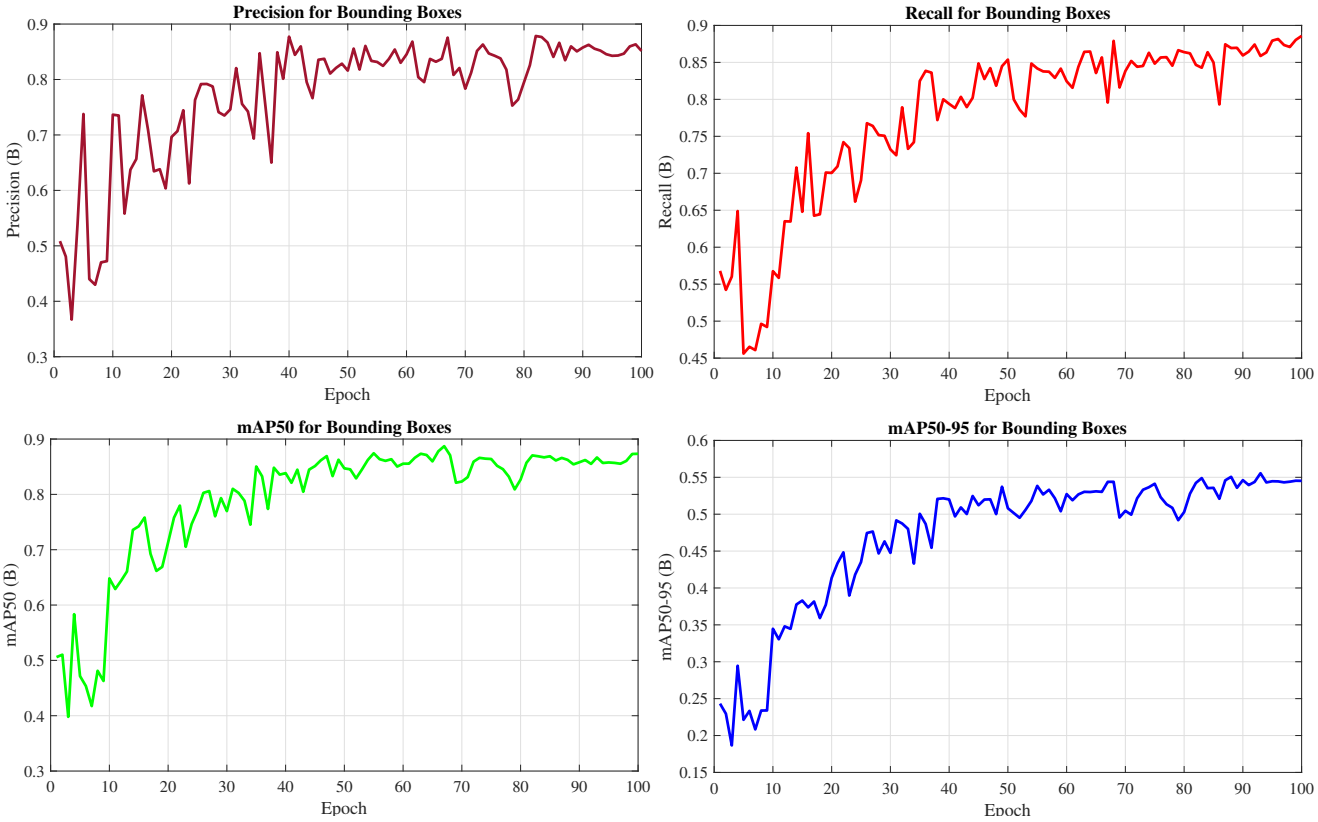

Fig. 4. Bounding boxes estimation of the SOD

TABLE II
EVALUATION OF THE OBJECT DETECTION ON VARIOUS DATASETS

Database	mAP50	mAP50-95	Classes	Inference (ms)
MPII	0.905	0.689	2	5.9
Market	0.933	0.723	2	5.7
Adience	0.940	0.778	1	5.9
IITD-USE	0.877	0.599	4	6.2

in Table II. This discrepancy highlights the adaptability of the model to datasets with distinct characteristics and underscores its efficacy in diverse scenarios. Some sample images of the SOD detection are shown in Fig. 10.

2) Coprehensive and Comparative Analysis: This performance of the proposed SOD model is assessed against state-of-the-art (SOTA) object detection approaches shown in Table III. Most of the existing models evaluated the data after categorizing it into small and big instances. It provides a comprehensive analysis of the data complexity and robustness of the model. Therefore, this experiment is also designed in a similar manner using the IITD-USE dataset. This dataset, specifically designed to encompass suspicious scenarios, is segregated into two sets to capture varying complexities. Set-1 comprises images from less challenging environments, while Set-2 includes more intricate scenes featuring small objects, low resolutions, crowded places, and similar challenges.

Notably, approaches such as Tiny Face, extremely tiny face detector (EXTD), single stage headless (SSH) face detector, deep residual network (DRN), Face-MagNet, and RetinaFace demonstrated superior mAP exceeding 0.72 on Set-2. However, the region-based convolutional network (Face-RCNN) faced challenges adapting to the complex environment's intricacies. Furthermore, these advanced models exhibited remarkable accuracy, surpassing 0.77, when applied to Set-1, portraying their proficiency in less complex scenarios. Detailed class-wise accuracies can be found in Table III for a comprehensive understanding of the performance of the model across different classes.

C. Facial Expressions and Body Language Analysis

1) Training and Evaluation of the DCNNs: This experiment analyzes the efficacy of the proposed DCNNs for the recognition of facial expressions and body language. These models are trained with the specific training sets of IITD-USE and FER20E datasets. The notable performance improvement can be attributed to the utilization of well-generalized data, focusing on specific applications. Therefore, the data augmentation approach is employed during training. Introducing substantial variations in the dataset, this approach prevents the model from being adversely affected by pattern variations and overcomes the overfitting and underfitting issues.

The hyperparameters of the backend network, i.e., minibatch size (16, 32, 64, 128, and 256), learning rate (0.1, 0.01, 0.03, 0.001, 0.003, 0.0001, and 0.00001), and optimizers are varied and fine-tuned. The models achieve higher performance with a minibatch size of 128 and a learning rate of 0.0001. The validation frequency of the experiment is kept the total number of samples in an epoch. It is observed that the loss and accuracy of the models are saturated among 90-100 epochs. Therefore, an early stopping is imposed during training. The experiment adopted the exploration of different optimizers, including stochastic gradient descent with momentum (SGDM), root mean square propagation (rmsprop), and adaptive moment (Adam). Notably, the Adam optimizer emerges as the most effective, yielding higher mean Average Precision (mAP) values of 0.970 for emotion classification and 0.744 for body language classification, as detailed in Table IV.

TABLE IV
PERFORMANCE ANALYSIS OF DCNNS OVER THE VARIOUS OPTIMIZERS

Approach	DCNN-I		DCNN-II	
	mAP	recall	mAP	recall
SGDM	0.964	0.930	0.669	0.645
rmsprop	0.966	0.945	0.723	0.711
Adam	0.970	0.956	0.744	0.691

2) Comparative Analysis with the SOTA: The robustness of the DCNN-I is analyzed by evaluating it on the benchmark FER datasets with a similar training and hypertuning setup. It reveals the average accuracies of 97%, 97.63%, 72.20%, 95.60%, and 64.50% for The LNMIIT, CK+, FER2013, FER20E, and AffectNet, respectively. Notably, FER2013 and AffectNet datasets present a richer spectrum of variations, especially in terms of intermediate nuances, compared to the LNMIIT and FER20E datasets. This increased variability contributes to slightly lower accuracy. The performance metrics have been shown in Fig. 5.

An extensive evaluation of facial expression recognition was conducted by incorporating various state-of-the-art (SOTA) approaches, as detailed in Table V. These models leverage a range of machine learning and deep learning techniques. The AffectNet dataset, which includes discrete and compound

TABLE III
COMPREHENSIVE ANALYSIS OF THE OBJECT DETECTOR IN TERMS OF MAP WITH THE SOTA APPROACHES

Network	Backbone	Approach	Person		Face		Fire		Weapon	
			Set_1	Set_2	Set_1	Set_2	Set_1	Set_2	Set_1	Set_2
Tiny Face [30]	ResNet101	Context reasoning	0.78	0.75	0.77	0.69	0.73	0.66	0.78	0.74
EXTD [31]	Mobilenet	Feature fusion	0.774	0.729	0.765	0.658	0.723	0.644	0.778	0.739
SSH [32]	VGG16	Context reasoning	0.78	0.76	0.8	0.73	0.68	0.61	0.73	0.66
DRN [33]	ResNet50	Anchor Matching	0.79	0.76	0.79	0.74	0.73	0.66	0.77	0.73
Face-MagNet [34]	VGG16	Context reasoning	0.776	0.731	0.771	0.684	0.698	0.667	0.779	0.736
Face-RCNN [35]	VGG19	Anchor matching	0.786	0.691	0.788	0.625	0.708	0.618	0.77	0.699
RetinaFace [36]	RetinaNet	Context reasoning	0.801	0.766	0.815	0.776	0.786	0.721	0.8	0.726
Proposed	CSPDarknet53	Fusion of features	0.829	0.776	0.833	0.790	0.806	0.728	0.836	0.750

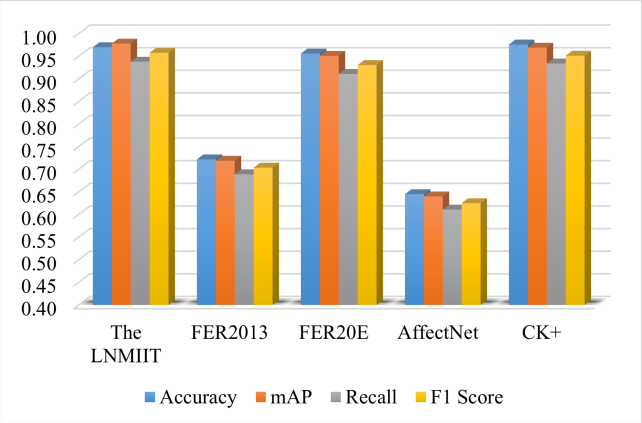

Fig. 5. Performance analysis of DCNN-I on various benchmark datasets

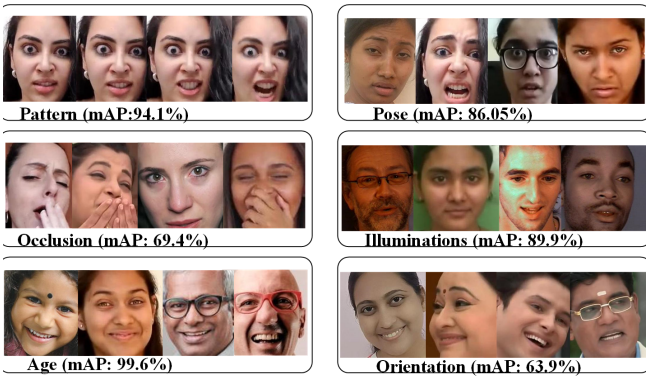

Fig. 6. Variations-wise evaluation of the FER20E

facial expressions captured in uncontrolled environments, is among the most recent and challenging datasets. Researchers have utilized several convolutional networks (ConvNets) such as ResNet, Inception, MobileNet, VGG16, Xception, and BReG-Net, reporting a mean Average Precision (mAP) of 0.60 ± 0.05. Similarly, the FER2013 dataset, despite its complexity and issues like data imbalance, achieved an mAP of 0.58 ± 0.05. Other datasets, such as CK+ and LNMIIT, are less complex as they were collected in controlled environments. The proposed DCNN-I model demonstrates a 1-2% improvement in these datasets. A detailed performance analysis is presented in Table V.

TABLE V
STATE OF THE ART COMPARISON (%)

Data	Algorithms	mAP
AffectNet	ResNet [37]	0.607
	InceptionV2 [38]	0.599
	MobileNet [39]	0.617
	Vgg16 [40]	0.612
	Xception [41]	0.595
	Inception_ResNet [42]	0.595
	BReG-Net [43]	0.638
	Ours	**0.648**
FER2013	Going deeper [44]	0.664
	FER2013 winner [45]	0.712
	Multiple deep network learning [46]	0.720
	Adaptive Weighting [47]	0.726
	Hierarchical committee of DCNNs [48]	0.727
	Multi-scale CNNs [49]	0.728
	Custom CNN [[50]	0.664
	ZFER-FCNN [51]	0.650
	SML [52]	0.728
	Ours	**0.730**
CK+	Custom CNN [50]	0.932
	CNN + Op. loss [53]	0.955
	FCNN [54]	0.844
	SBN-CNN [55]	0.968
	ZFER-FCNN [51]	0.970
	Ours	**0.985**
LNMIIT	KNN with fusion of features [27]	0.962
	KNN + distance features [56]	0.960
	CNN + distance features [57]	0.965
	Ours	**0.978**
FER20E	**Ours**	**0.920**

To show the efficacy of the DCNN-I, a variation-wise evaluation is also done in this work including the variation in pattern, pose, occlusion, illumination, age, and orientation. It is observed that the occluded and oriented faces become more challenging in the recognition of expressions. DCNN-I provides an average accuracy of 69.4% on occlusion and 63.9% on orientation as shown in Fig. 6. An in-depth analysis of cross-cultural variation can be performed in future research.

D. Evaluation of the Discriminator Network

Within the discriminator network, crucial parameters are precisely computed, including the count of individuals, identified facial expressions, distinguished body language, recognized weapons, and detected instances of fire. During this evaluation, 3,500 images of the IITD-USE dataset are considered by manually discarding blurred and distorted images. The ground truth of these images is in the form of a risk score distributed from 0 to 10.

Various performance metrics such as mean absolute error (MAE), root mean squared error (RMSE), mean squared error (MSE), correlation coefficient (CRC), and R-squared are considered to evaluate the proposed USE-Riskometer. The MAE measures the average magnitude of the errors between predicted and actual risk scores, suggesting a straightforward interpretation of prediction accuracy. The RMSE and MSE provide additional information by emphasizing larger errors, which can be crucial for understanding model performance in cases with significant outliers. Similarly, the CRC evaluates the strength and direction of the linear relationship between the predicted and actual values, emphasizing the ability of the model. The R-squared (R^2) score as the coefficient of determination measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It ranges from 0 to 1 and is calculated using the Equation 7.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{7}$$

where y_i tends the actual values, $\hat{y}_i$ represents the predicted values, and $\bar{y}$ is the mean of the actual values.

The numerator $\sum(y_i - \hat{y}_i)^2$ is the residual sum of squares, representing the unexplained variance by the model. The denominator $\sum(y_i - \bar{y})^2$ is the total sum of squares, representing the total variance in the data. By comparing the unexplained

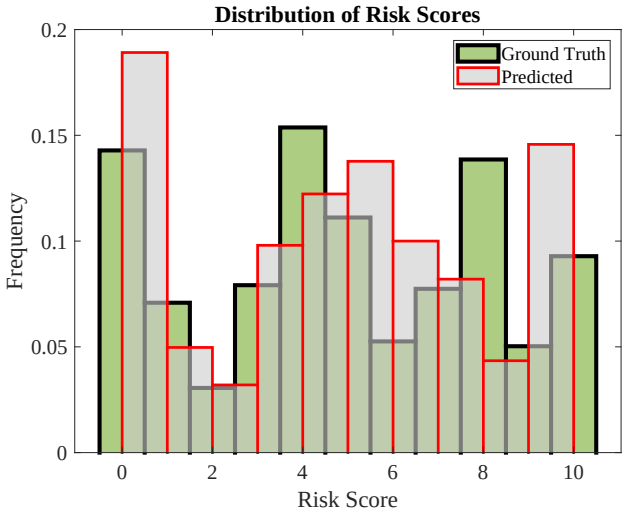

Fig. 7. Distribution of the Risk Score

variance to the total variance, R^2 provides a measure of how well the model accounts for the variability in the data.

A graphical representation is shown in Fig. 7, which provides a comparative analysis of the risk score's frequency distribution between the ground truth and the estimated data. It helps to understand how well the estimated risk scores match the actual (ground truth) risk scores. For the comprehensive analysis of the risk estimation, two kinds of assessments have been done in this experiment, i.e., original data having a 0 to 10 risk scale, and categorized data with the quantized risk scale. The quantized risk scale categorizes predicted scores into three classes such as low (0-3), medium (4-6), and high (7-10). It improved the system's performance and reduced complexity as shown in Fig. 8.

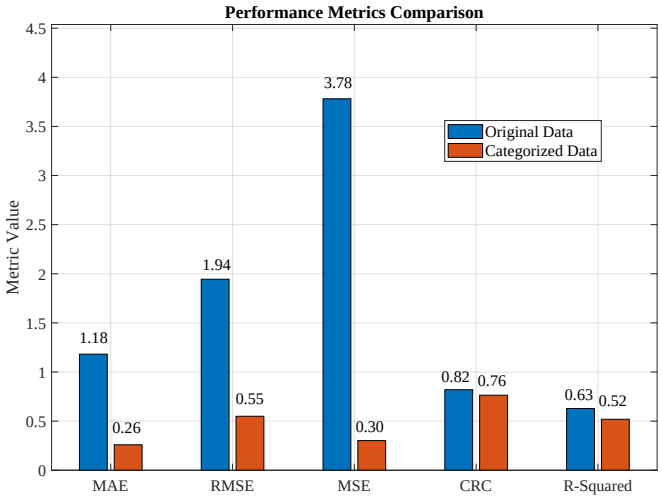

Fig. 8. Performance analysis of the USE-Riskometer on the original and categorized data

The confusion matrix of the USE-Riskometer on this categorized data is shown in Fig. 9.

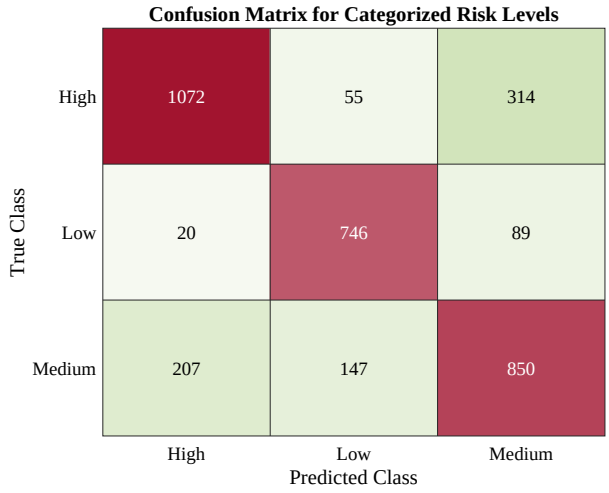

Fig. 9. Confusion matrix of the USE-Riskometer on the categorized risk data

IV. Conclusion and Future Work

In this work, a comprehensive system for suspiciousness estimation is presented, exploring various domains such as computer vision, image processing, and deep learning. To design this system, three major modules, i.e., a suspicious objects detector (SOD), a facial expressions classifier (DCNN-I), and a body-language analyzer (DCNN-II) are developed and evaluated against state-of-the-art (SOTA) models. The SOD model shows an improvement of 1-2% in mAP over the SOTA models. For facial expression recognition, our model demonstrates a 1-5% improvement in mAP on benchmark datasets including AffectNet, FER2013, CK+, The LNMIIT, and FER20E. Evaluations on diverse datasets confirm the effectiveness of these models.

The overall system performance has been assessed using performance metrics such as MAE, MSE, RMSE, correlation coefficient, and R-squared score. This performance is evaluated by 1) predicting the score on a scale of 0-10, and 2) categorizing levels of risk as low, medium, and high. The system achieves higher performance in categorized levels compared to the score scale due to reduced complexity.

Notably, the creation of the IITD-USE dataset enhances our work's reliability. It has been validated by comparing existing models and datasets, establishing its effectiveness in real-world scenarios. This paper marks a significant advancement in image-based suspiciousness estimation, promising practical applications in computer vision and security. Future research may focus on refining, instances, orientation, distortion, and implementing real-time solutions for increased impact.

Acknowledgments

We would like to acknowledge the Department of Electrical Engineering, Indian Institute of Technology Delhi for their invaluable guidance and computational resources.

References

[1] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3200–3225, 2023.

Fig. 10. Some samples from IITD-USE evaluated through the SOD model: green color bounding boxes represent the $Person$, blue color bounding boxes highlight the detected $Face$, red color bounding boxes represent $Weapons$, while yellow color bounding boxes highlight $Fire$.

[2] S. Song, J. Liu, Y. Li, and Z. Guo, "Modality compensation network: Cross-modal adaptation for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 3957–3969, 2020.

[3] Z. Jiang, V. Rozgic, and S. Adali, "Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[4] Y. Wang, Y. Xiao, F. Xiong, W. Jiang, Z. Cao, J. T. Zhou, and J. Yuan, "3dv: 3d dynamic voxel for action recognition in depth video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] S. Ding, S. Qu, Y. Xi, A. K. Sangaiah, and S. Wan, "Image caption generation with high-level image features," *Pattern Recognition Letters*, vol. 123, pp. 89–95, 2019.

[6] H. Wang, D. Zhang, Y. Wang, J. Ma, Y. Wang, and S. Li, "Rt-fall: A real-time and contactless fall detection system with commodity wifi devices," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 511–526, 2017.

[7] D. Liang and E. Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 1, 2019.

[8] C. Vajiac, D. H. Chau, A. Olligschlaeger, R. Mackenzie, P. Nair, M.-C. Lee, Y. Li, N. Park, R. Rabbany, and C. Faloutsos, "Trafficvis: Visualizing organized activity and spatio-temporal patterns for detecting and labeling human trafficking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 53–62, 2023.

[9] F. G. Ibrahim Salem, R. Hassanpour, A. A. Ahmed, and A. Douma, "Detection of suspicious activities of human from surveillance videos," in *2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA*, 2021, pp. 794–801.

[10] K. S. Yadav, R. H. Laskar, N. Ahmad *et al.*, "Exploration of deep learning models for localizing bare-hand in the practical environment," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106253, 2023.

[11] Z. Wu, X. Chen, Z. Yang, and X. Du, "Reducing security risks of suspicious data and codes through a novel dynamic defense model," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2427–2440, 2019.

[12] N. Bordoloi, A. K. Talukdar, and K. K. Sarma, "Suspicious activity detection from videos using yolov3," in *2020 IEEE 17th India Council International Conference (INDICON)*. IEEE, 2020, pp. 1–5.

[13] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 1168–1181, 2007.

[14] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos, "Spotting suspicious behaviors in multimodal data: A general metric and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2187–2200, 2016.

[15] C. Chiu, J. Zhan, and F. Zhan, "Uncovering suspicious activity from

partially paired and incomplete multimodal data," *IEEE Access*, vol. 5, pp. 13 689–13 698, 2017.
[16] A. Wahrstätter, J. Gomes, S. Khan, and D. Svetinovic, "Improving cryptocurrency crime detection: Coinjoin community detection approach," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 6, pp. 4946–4956, 2023.
[17] D. Martínez, H. Loaiza, and E. Caicedo, "Algorithm for early threat detection by suspicious behavior representation," *IEEE Latin America Transactions*, vol. 18, no. 05, pp. 825–832, 2020.
[18] K.-E. Ko and K.-B. Sim, "Deep convolutional framework for abnormal behavior detection in a smart surveillance system," *Engineering Applications of Artificial Intelligence*, vol. 67, pp. 226–234, 2018.
[19] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazán, "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7991–8005, 2015.
[20] K. S. Yadav and L. Kumar, "Enhanced dictionary and cultural adaptability in facial expression recognition using self-attention transformer recognizer," 2024, indian Patent application filed.
[21] Kuldeep Singh Yadav Sonalika, Ajeet Kumar, and Lalan Kumar, "labelImg-automated: An automated image data annotation tool for labelImg," https://github.com/akki01133/labelImg-automated, 2024.
[22] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[23] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *CoRR*, vol. abs/2006.10204, 2020.
[24] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Computer Vision, IEEE International Conference on*, 2015.
[26] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, pp. 2170–2179, 2014.
[27] K. S. Yadav and J. Singha, "Facial expression recognition using modified viola-john's algorithm and knn classifier," *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 13 089–13 107, 2020.
[28] "FER-2013 Dataset," https://www.kaggle.com/datasets/msambare/fer2013, 2013, accessed: [15.05.2023].
[29] A. Mollahosseini, B. Hasani, and M. Mahoor, "Affectnet: A Database For Facial Expression, Valence, And Arousal Computing In The Wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.
[30] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 797–813.
[31] Y. Yoo, D. Han, and S. Yun, "Extd: Extremely tiny face detector via iterative filter reuse," *arXiv preprint arXiv:1906.06579*, 2019.
[32] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4875–4884.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
[34] P. Samangouei, R. Chellappa, M. Najibi, and L. S. Davis, "Face-MagNet: Magnifying feature maps to detect small faces," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 122–130.
[35] J. Yan, H. Wang, M. Yan, W. Diao, X. Sun, and H. Li, "IoU-Adaptive Deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery," *Remote Sensing*, vol. 11, no. 3, 2019.
[36] Y. Li and F. Ren, "Light-weight RetinaNet for object detection," *CoRR*, vol. abs/1905.10011, 2019.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.
[39] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.
[41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.
[42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." AAAI Press, 2017.
[43] B. Hasani, P. S. Negi, and M. H. Mahoor, "Bounded residual gradient networks (breg-net) for facial affect computing," in *Automatic Face & Gesture Recognition*. IEEE, 2019, pp. 1–7.
[44] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.
[45] Y. Tang, "Deep learning using linear support vector machines," 2015. [Online]. Available: https://arxiv.org/abs/1306.0239
[46] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning." Association for Computing Machinery, 2015.
[47] W. Xie, L. Shen, and J. Duan, "Adaptive weighting of handcrafted feature losses for facial expression recognition," *IEEE Transactions on Cybernetics*, vol. 51, no. 5, pp. 2787–2800, 2021.
[48] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, vol. 10, pp. 173–189, 2016.
[49] H.-D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I.-S. Na, and S.-H. Kim, "Facial emotion recognition using an ensemble of multi-level convolutional neural networks," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 11, p. 1940015, 2019.
[50] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.
[51] T. Shahzad, K. Iqbal, M. Khan, and N. Iqbal, "Role of zoning in facial expression using deep learning," *IEEE Access*, vol. 11, pp. 16 493–16 508, 2023.
[52] W. Hayale, P. S. Negi, and M. H. Mahoor, "Deep siamese neural networks for facial expression recognition in the wild," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1148–1158, 2023.
[53] A.T.Lopes, E. D. Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern recognition*, vol. 61, pp. 610–628, 2017.
[54] T. Shahzad, K. Iqbal, M. Khan, and N. Iqbal, "Role of zoning in facial expression using deep learning," *IEEE Access*, vol. 11, pp. 16 493–16 508, 2023.
[55] J. Cai, O. Chang, X. L. Tang, C. Xue, and C. Wei, "Facial expression recognition method based on sparse batch normalization cnn," in *Chinese control conference (CCC)*. IEEE, 2018, pp. 9608–9613.
[56] K. S. Yadav, J. Singha, and R. H. Laskar, "Facial expression recognition using facial features detection using the fusion of classifiers: In a real-time scenario," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, 2019, pp. 262–267.
[57] K. S. Yadav, N. Ahmad, A. M. K., S. Alom Barlaskar, N. Saidulu, and R. H. Laskar, "A holistic approach towards detection, tracking, and recognition of face," in *2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, vol. 1, 2022, pp. 384–388.