IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 17 May 2023, accepted 3 June 2023, date of publication 9 June 2023, date of current version 21 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3284472

RESEARCH ARTICLE

Propounding First Artificial Intelligence Approach for Predicting Robbery Behavior Potential in an Indoor Security Camera

SHIMA POUYAN[1], MOSTAFA CHARMI[1], ALI AZARPEYVAND[2], AND HOSSEIN HASSANPOOR[3,4]

[1]Department of Electrical Engineering, University of Zanjan, Zanjan 38791-45371, Iran
[2]Department of Electrical and Computer Engineering, University of Zanjan, Zanjan 38791-45371, Iran
[3]Medical Engineering Department, Energy Institute of Higher Education, Saveh 67746-39177, Iran
[4]Department of Cognitive Sciences, Dade Pardazi Shenakht Mehvare Atynegar (DSA) Institute, Tehran 19789-76677, Iran

Corresponding author: Mostafa Charmi (charmi.mostafa@znu.ac.ir)

ABSTRACT Crime prediction in video-surveillance systems is required to prevent incident and protect assets. In this sense, our article proposes first artificial intelligence approach for Robbery Behavior Potential (RBP) prediction and detection in an indoor camera. Our method is based on three detection modules including head cover, crowd and loitering detection modules for timely actions and preventing robbery. The two first modules are implemented by retraining YOLOV5 model with our gathered dataset which is annotated manually. In addition, we innovate a novel definition for loitering detection module which is based on DeepSORT algorithm. A fuzzy inference machine renders an expert knowledge as rules and then makes final decision about predicted robbery potential. This is laborious due to: different manner of robber, different angle of surveillance camera and low resolution of video images. We accomplished our experiment on real world video surveillance images and reaching the F1-score of 0.537. Hence, to make an experimental comparison with the other related works, we define threshold value for RBP to evaluate video images as a robbery detection problem. Under this assumption, the experimental results show that the proposed method performs significantly better in detecting the robbery as compared to the robbery detection methods by distinctly report with F1-score of 0.607. We strongly believe that the application of the proposed method could cause reduction of robbery detriment in a control center of surveillance cameras by predicting and preventing incident of robbery. On the other hand, situational awareness of human operator enhances and more cameras can be managed.

INDEX TERMS Surveillance videos, low resolution, RBP prediction, deep learning method, fuzzy inference machine.

I. INTRODUCTION

Today, surveillance cameras are widely used in various places such as stores, banks, airports and homes, to increase public safety and prevent the occurrence of crime. Alternatively, the time and place of the crime and specifically the wrongdoer, can be achieved by analyzing these videos and aiming to identify the delinquent. Meanwhile, someone is needed behind the scene, watching the videos and noticing whenever something anomaly is happening. However, due to very rare occurrence of an anomaly, the person becomes tired and if an anomaly happens, sometimes he cannot realize its occurrence. In other words, he loses the anomaly [1]. Furthermore, the anomaly-detection process is based on human common feeling which is learned during years. On the other hand, skill amount of the person for signs of crime occurrence understanding ability and the cost of employing him are other problems of nonautomated crime prediction and detection systems which are based on watching surveillance videos.

The associate editor coordinating the review of this manuscript and approving it for publication was Varuna de Silva.

To automate anomaly detection, some visual features must be extracted using machine learning and deep learning algorithms [2], [3]. For better performance of these algorithms, specific features for different anomaly classes [4] like vandalism [5], violence detection [6] and robbery [7] can be useful. Predicting the location and time of the crime reducing the destruction. On the other hand, security forces are also present on time such as an experiment, manufactured in Santa Cruz, California, where officers benefit from daily crime forecasts every morning. This forecasting navigates them to patrol determined regions. A Santa Cruz spokesperson declared that thirteen wrongdoers have been stopped in the determined areas during first the six month of experiment [8]. Due to paper [8], [9], and [10], some main symptoms prove that predictive policing is significant to be used for federal financing and security systems including: cost saving and crime reduction. Violent crimes are more dangerous because of their victimization probability and they increased by 20% due to Seattle Police Department (SPD) report during 2021 in Washington, USA [11]. According to statics acquired from Federal Bureau of Investigations-Uniform Crime Reporting System (FBI-UCR), robbery is one of five common crimes in the United States [12]. The detection of robbery is one of the purposes of installing surveillance cameras in many places. Robbery is the crime of taking or attempting to grab any property by force, threat or weapon [13] based on oxford dictionary definition and differentiated from other forms of theft such as shoplifting, pickpocket or burglary, by its intrinsically violent essence [14], [15]. Whiles many lesser types of theft are punished as misdemeanors, robbery is always a felony in jurisdictions. Criminologists distinguish different types of robbery with regards to time and space of occurrence, armed or unarmed robberies, weapon types and force amount. Therefor, one typical scissor is commercial robberies and street robberies [16]. Street robberies usually happens in poor crowded locations with no Closed-circuit televisions (CCTV). Commercial robberies occur in two ways: one where the offender enters the scene dressed up as a customer or conceal his face with normal covers like mask or helmets then suddenly out of the blue pulls a weapon and scares the employee. The other which offenders enter with force, typically in a group and probably conceal their face or head [17]. Both types of commercial robberies occurred in the indoor places which have customarily CCTVs so that detecting offenders or detection and even prediction of commercial robberies can be possible. Additionally, offenders who armed by weapon or knife usually threaten human with force. On the other hand, for offenders bearing any stick or be unarmed, a massive force is more probable [16], [17]. Hereupon, armed or unarmed commercial robberies force causes injury, pain and even death.

Thus, predicting commercial robbery behavior by human, machine or combination of these two, plays an important role in preventing its occurrence and its arisen dangers [18].

In general, there are some methods to automate detection or prediction of crimes based on extracting different crime scenarios and implementing them in different fields. But none of these methods have predicted the potential of robbery behavior. Therefore, there is a need to develop an algorithm for RBP prediction in video images. One could easily notice that, extracting the evidences and features in the surveillance videos is needed for prediction. To do this, the potential of robbery behavior in video images should be investigated. Scenarios of robbery occurrence, vary from one context to another [19] due to different conditions of each place selected for robbing and different cultures of countries. Therefore, robust feature extraction is not accompanied with certainty.

Despite the variety of robbery incidence scenarios and due to scenario-based approaches [20], [21], a common scenario with main points can be considered for commercial robbery videos. Specifically, one or some person choosing a poorly attended place who are usually covering their face or head by helmet, mask, glasses or any garment to not be recognized and they are loitering to get an opportunity for showing their weapon, threat or force. This scenario is completely matched with the knowledge of an expert person and definition of first type of commercial robbery behavior [17], [22]. To implement a system based on this common scenario abstracted from different scenarios inferred from robbery videos, we consider common features found in most robbery cases under three modules including: head cover, crowd and loitering detection. After extracting these features, for modules implementation, an inference machine is needed to conclude on the RBP. The conclusion process must be as competence as a human decision making for potential derivation. Due to the ability of fuzzy set theory to mimic human inference [23], experience could be put in the form of fuzzy rules and according to fuzzy measurement, it facilitates the diagnosis and reasoning of a complex decision [24], [25]. Deep learning methods on the other hand, do not offer such adaptability and may not be able to deal with the nuances and variations of uncertain data well [25]. Owing to these reasons a fuzzy inference machine is proposed in this paper.

To sum up, main contributions of our paper are as below:

1. The proposed algorithm is based on a novel method which can predict RBP and prevent damages resulted by its occurrence in indoor places.To the best of our knowledge, this is the first work focusing on robbery behavior prediction and grounded on three main modules: Head cover, crowd and loitering detection modules.

2. A dataset has been prepared for our system and annotated manually as two states: with or without head cover. For crowd counting, we sum the results of two states reported by head cover detection module. The method dominates the constraints of surveillance videos such as low resolution and single camera videos.

3. The loitering point we have defined, is a novel definition for loitering calculation. A Deep Simple Online Real-time Tracking (DeepSORT) algorithm has used with respect to the

tracking methods to calculate amount of loitering for each person. By analyzing the obtained amount of loitering based on Euclidean distance calculation, a point has assigned to each one.

4. The key contribution of our algorithm is using a fuzzy inference machine with optimized rules, fuzzification and defuzzification steps. Obtained results of these three modules analyzed based on an expert person knowledge about robbery behavior and an inference machine.

The rest of paper is arranged as follows: Section II reviews some literature related to our work including suspicious behavior prediction or detection and also papers related to our modules. Section III explains proposed algorithm and outlines concepts of RBP prediction, the proposed modules and outcome to low-resolution video images by improving YOLOV5. Experimental results are presented and discussed in section IV. The last section concludes the research work and presents future works.

II. RELATED WORKS

Anomalies are infrequent observations, events or behaviors which are suspicious because they are significantly different from normal patterns. Crime is a kind of an anomaly which is any behavior deviating from a normal activity [2]. One could say that the proliferation usage of CCTV has been because of increasing crimes in public places. Crime can be predicted according to suspicious behavior detection. prediction needs defective, vague and unsure information [26]. Our proposed approach concentrates on RBP prediction in indoor places. Robbery is a kind of crime and the proposed algorithm needs loitering, crowd and head cover detection. One important concept of our algorithm is providing a generic RBP prediction framework which is not addressed in any other paper. In this section, we will discuss about some related works relevant for suspicious behavior detection or prediction, crime detection or prediction and articles concerning with the loitering or head cover detection.

Elhamod and Levine [27] proposed a semantics-based suspicious behavior recognition algorithm based on object tracking by blob matching with color histograms and spatial information, for updating objects intended in each frame. For blob and objects similarity specification, intersection of histogram's value is calculated and compared with the defined threshold. Next it assigns appropriate classes contains people for animated and objects for inanimated things. By calculating their 3D motion features and recording it in the form of historical records, behaviors are semantically determined. detected suspicious behaviors include: abandoned luggage by background subtraction methods, fainting by comparing assumed 2D and actual 3D location of person's feet and also head coordinates of that person, fighting by computing merge, split and simultaneously movement of blob's centroid and eventually loitering by aggregating presence time of a person in an area.

Ishikawa and Zin [18], introduced an automated normal system for questionable pedestrian detection by loitering detection. According to [18], a questionable person walks, stops and goes around the location repeatedly for a long time with enhancement of direction changing number. His distance value is greater than the normal person and changing in acceleration is so much. To implement these features, [18] divides the video frames into 25 blocks and counts the frequency of block numbers which feet of person are in that location. if this frequency was more than threshold value, that person descent as suspicious pedestrian. To compute changing of direction, it calculates angles of moving direction. Computing of distance and acceleration changing extracts all needed features. Finally, a decision fusion process detects suspicious pedestrians by aggregating the scores of each step.

Rajapakshe et al. [28] presented an E-police system which contains two components: video surveillance monitoring system and crime prediction.To detect suspicious behaviors such as violent and vandalism, [28] uses human activity recognition methods and classifies them into normal and abnormal categories. They use Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for feature extraction and detect suspicious human who conceal their faces. crime prediction process of [28] is based on public information resources for place and time of crime occurrence prediction with the help of classification algorithms such as SVM and Decision tree.

Arroyo et al. [22] proposed an expert real-time suspicious behaviors detection system in shopping malls. They locate foreground objects by an image segmentation and background subtraction algorithm. Next, a blob fusion algorithm is used to gather the blobs of each segmented parts to detect human. A tracking algorithm is used with the help of a new two-step method: a) using a Kalman filter for detection and tracking human, b) SVM kernels for occlusion management. Then, the obtained trajectories of people are used to analyze human behaviors. The entrance or existence alarm is for the time that too many people enter or a person runs away and it is detected by trajectory analyzing. Moreover, specific risk areas are interiors with more expensive articles and chosen by the human security officers. Loitering of people is evaluated according to their trajectories and the length of time they present in those zones. If the time be more than 30 seconds, which has specified by security experts, their system gives off an alarm. They mounted a camera on the cash desk to protect it. If someone loiters around it, and no shop personnel attended in the cash desk zone, an alarm gives off. Additionally, evaluation process is done on a naturalistic dataset they provided by multi cameras located on entrance, interior and cash-desk of a shop.

Bouma et al. [29] focused on the automatic pickpocket behavior early detection by three steps: detection and tracking pedestrians using multicamera, feature extraction adopted to pickpocket scenario and diagnosis of pickpocket. Based on an expert knowledge, the scenario of pickpocket assumed in [29] includes seven main phases: surroundings observation, looking for a proper opportunity, communicating to accomplice, besieging target person, snatching a stuff, handing it over and evacuating the scene. To implement this scenario, they extract some related features including speed of walking, changing of orientation and crowd merging and splitting. Finally, a Fisher Linear Discriminant Classifier (LDC) used to recognize pickpockets. They used video dataset which has been played by some actors as pickpockets and victims.

Selvi et al. [30] determined enhanced CNN (ECNN) algorithm to detect suspicious actions like shooting and stealing. Their ECNN method have input layer to feed the convolution3D layer for feature extraction. Next, methods of [30] set threshold for LeakyReLU layer for detection of normal or suspicious action.

Roy [31] proposed a snatch theft detection model by using Gaussian Mixture Model (GMM) with regards to scenarios of snatch theft occurrence. Reference [31] use HOF and MBH as feature descriptors to extract dense trajectory feature sets and represented as action-vector. Next, they train GMM by these vectors for snatch theft detection.

Kaur et al. [32] presented a face mask detection algorithm by retraining a CNN with 3832 face dataset images consisting with or without mask. A face mask detection system is developed by retraining YOLOV5 with 685 images that consists of images of people from two categories that are with and without face masks [33]. Proposed method of [33] training their model with different number of epochs and the optimal epoch number they reported was 300. Face mask detection model proposed in [34], retrained MobilenetV2 with the help of data augmentation methods. authors of [35] proposed a method which detects the face mask with the help of deep Neural Networks. They fine-tuned the last layer of RESNET50 by adding five new layers and retrained it with 25876 images including with or without mask faces. Chowdary et al. [36] added 5 more layers to InceptionV3 to finetune it for mask detection. Singh et al. [37] trained both YOLOv3 and faster R-CNN with the images including human with or without face mask. Huang et al. [38] proposed a method to detect helmet with an improved YOLOV3 with the change in feature map size of it for small object detection. Zhou et al. [39] retrained YOLOV5 with 6045 images including people in two categories: with or without helmet for helmet detection. A retrained Single Shot Detectors (SSD) MobileNet V2 deep network used in [40] for helmet detection. YoloV5 has small size and it is faster to train. Furthermore, it has ability of small object detection [18]. Some loitering detection methods are discussed in ensuing part.

Authors of [41] proposed a loitering detection method in which person detected by a YOLOV3 algorithm and tracked by a DeepSORT algorithm. Nayak et al. [41] detected loitering of individual by computing time duration and displacement of them in comparing with a threshold they defined. Nam [42] proposed a loitering detection algorithm based on pedestrian detection by blob detection and comparing spatial and temporal information with a threshold to detect loitering person. The time threshold defined in [42] calculated with regards to time staying in the place and the frame rate. Consequently, our proposed algorithm originated from behavior detection systems like proposed systems of [22] and [29] which used scenarios define suspicious behaviors and different steps accomplished to detect these behaviors. We use a scenario of robbery behavior occurrence to implement our innovative method. In our proposed method, we elaborate on method of [18] and we use DeepSORT algorithm for tracking each person. We also use Euclidean distance for loitering detection based on defining a threshold for distance value. A YoloV5-based algorithm is retrained for mask and helmet detection for head cover detection module.

A quick summary of all these methods is provided in Table 1. Furthermore, none of the methods defined in the above articles, have pondered the RBP prediction with purpose of preventing its occurrence. The potential for robbery behavior can only be calculated in videos with analyzing possibility for a period of time before any force, threat or display of a weapon occurs. That is, the suspicious behavior of the robber can be assessed. Common scenarios of robbery behavior, is choosing a poorly attended stores, covering head to not be recognized, and loitering for an opportunity to threat with a weapon and use force. In our proposed approach, number of human and his head cover is detected using the retrained YOLOV5. Next, the human is tracked using DeepSORT for our novel loitering calculation method which is based on Euclidean distance calculation and our defined displacement thresholds. A human need for inferring a video and decide about RBP of each frame. Fuzzy Logic looks at the world in ambiguous terms, in much the same way that our brain takes in information, then responds with precise actions. The human brain can reason with uncertainties, vagueness, and judgments. Computers can only manipulate precise valuations. Fuzzy logic is an attempt to combine the two techniques. Due to the imprecise terms resulted from three modules, the fuzzy inference machine can, like the human mind, decide on the potential of robbery behavior. So, a fuzzy inference machine inferences RBP. Next section presents our methodology and proposed approach.

III. PROPOSED METHODOLOGY

In this section, the proposed method is introduced in detail for prediction of RBP as illustrated in Fig. 1. At the beginning, three main blocks defined, including:

- Head cover detection module.
- Crowd detection module to check number of humans attended in environments.
- Loitering detection module.

Dataset gathered by our group with regards to intention of proposed robbery scenario, to implement head and crowd detection modules. Next, data prepared by manually annotating and convolving to decrease their resolution. By retraining YOLOV5s to customize it, two first modules are completely provided. An Euclidean method is used to calculate distance traveled by human and the DeepSORT algorithm is employed to track him. By defining our individual thresholds, the label

TABLE 1. Literature survey based on relevance to our proposed approach.

Reference	Proposed Scheme	Implemented Methods
[18]	suspicious persons detection system by loitering detection	To implement this system, they calculate traveled distance, acceleration and direction changes. Finally, they compare calculated parameters with a threshold they have defined to decide if the person is suspicious.
[22]	An expert video-surveillance system is utilized for detection of real-time suspicious behaviors like loitering around interior or cash desk and running, in shopping malls.	The proposed system detects humans by blob fusion algorithm and track them by Kalman filter and SVM kernels. Analysing of behaviours done by obtained trajectories.
[27]	Real-time detection of suspicious behaviors like loitering in public transport areas	The proposed solution recognize behaviour semantically based on object tracking and assigning a label as object or human for each one. Next, they calculate 3D motion features to detect suspicious behavior
[28]	Suspicious behavior detection and spatio-temporal crime prediction	CNN and RNN methods are used for feature extraction to recognize human activity and detect suspicious behavior. they also use public information resources belong to place and time of crime occurrence for crime prediction.
[29]	Automatic early detection of a suspecious behavior with track-based features in a shopping mall. they detect behaviour of pickpocket by it's scenario implementation.	Feature extraction and Fisher linear discriminant classifier is used to early detect pickpocket. Their features are based on an expert knowledge including: speed of walking, changing of orientation and crowd merging and splitting.
[30]	Suspicious action detection	An ECNN algorithm proposed for suspicious action detection.
[31]	Snatch theft detection model	Their model uses GMM based on common theft scenarios. it also uses HOF and MBH to extract dense trajectory feature sets and create action-vectors to train GMM. .
[32] – [37]	Deep learning algorithms are used for face mask detection.	Retraining CNN, YOLOV5, MobilenetV2, RESNET50, InceptionV3, YOLOV3 and faster R-CNN with two-class images: with or without mask.
[38] – [40]	Approaches based on deep learning algorithms are used to detect helmet.	Proposed solutions detect helmet by retraining YOLOV3, YOLOV5 and SSD MobileNet V2.
[41]	Loitering detection method based on human detection and track.	A YOLOV3 based detector is used for human detection and a DeepSORT algorithm for tracking him. By comparing time duration and displacement of individuals with a defined threshold, loitering is detected.
[42]	Spatio-temporal based detection system for loitering.	A YOLOV3 based detector is used for human detection and a DeepSORT algorithm for tracking him. By comparing time duration and displacement of individuals with a defined threshold, loitering is detected.

of loitering allocated to each person and based on this innovating definition, our particular loitering detection module is presented. Finally, a fuzzy inference machine is used for potential prediction of robbery behavior. The RBP prediction can be decomposed into three main parts: feature extraction, feature analysis and RBP prediction.

A. FEATURE EXTRACTION

Features extracted from each video, are organized as three main modules based on common scenario of robbery occurrence, including: head cover, crowd and loitering detection modules. In this part we will explain about each one.

1) HEAD COVER DETECTION

In our proposed method, head cover is defined as any hat, helmet, mask, glasses or clothes that conceals a person's head and face and disturb their identification. Most of robbers prefer not to be recognized when robbing and head covering is their choice. In this paper, we propose a method which can detect head of human intended in a store with or without head cover in low-resolution single frames and label them as masked or no-mask. The proposed algorithm uses a deep learning method for head cover module. You Only Look Once (YOLO) [43] is a deep learning algorithm that uses to detect objects and views entire image as a regression problem [44]. Within training process of YOLO, it looks over the whole image to extract global information of target and divides the image into $S \times S$ grids. This division makes some grid cells and the position of object's center should fall into that grid cell to be labeled as detected object. Resolution of surveillance videos are usually low and head cover detection module should also overcome this challenge.The YOLO algorithms can be retrained by low-resolution images to make an accurate detection result [45]. YOLO v1, v2 and v3 can detect type of an object and its position at an image [46]. YOLOv4 [47] is an improvement on the YOLOv3 in mean average precision(map) and the number of frames per second [48]. YOLOv5 is based on YOLOv4 with higher processing speed and smaller size, even 90% smaller than YOLOV4. Therefore, YOLOv5 is a better choice to be used in an embedded device. Moreover, YOLOv5 has better capability in small objects detection [49]. Due to small size of YOLOV5 and it's ability for small objects detection, we choose YOLOV5 for our proposed method and retrain it with our prepared low resolution images. The YOLOv5 algorithm regulates the width and depth of the backbone network. Beside these reasons, high human and object detection accuracy of YOLOv5, make it a good choice to be used in head cover detection for our proposed method. Therefore, it has four versions of the model, which are YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x. YOLOv5s is the simplest, smallest and fastest version [50] and we use this version in our method.

2) CROWD DETECTION

People attended in the store, either have head cover or not. As a consequence, by considering the results obtained from

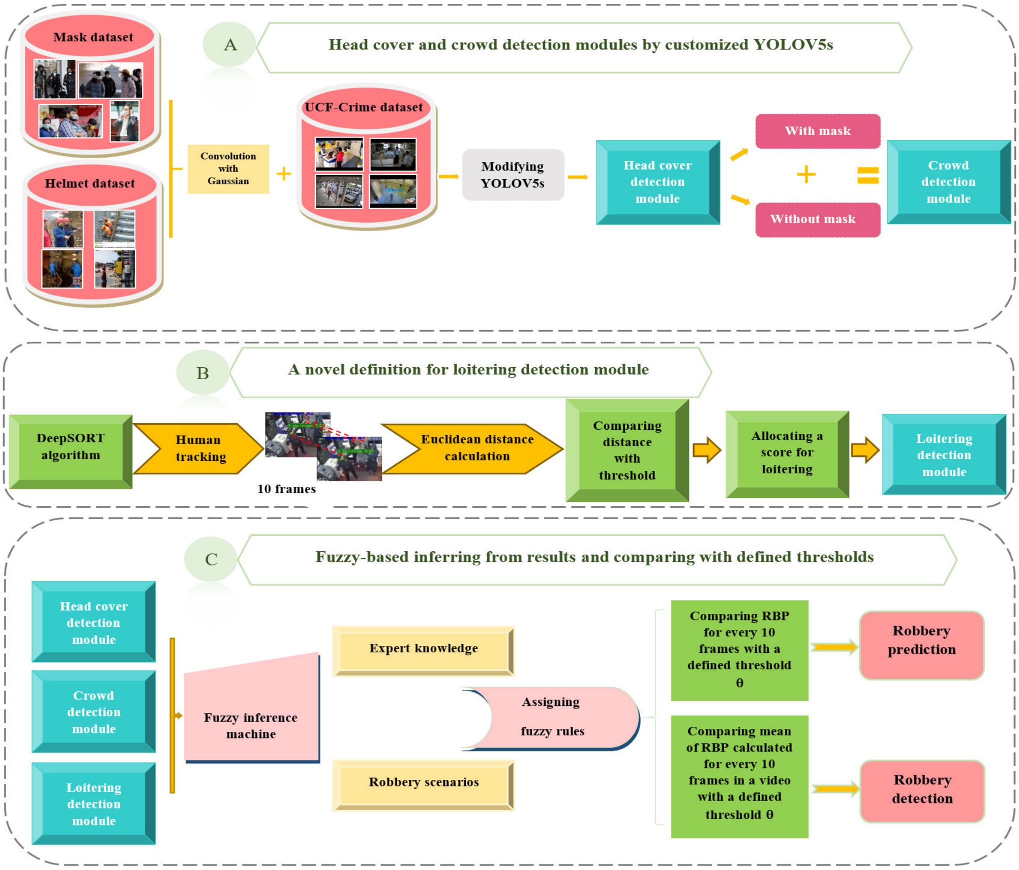

FIGURE 1. **The block diagram of robbery behavior prediction and detection system. After collecting 7622 images from three different datasets and preparing them for YOLOV5s retraining, head cover and crowd detection modules are implemented (block A). Loitering detection module uses an Euclidean method to calculate distance traveled by human and DeepSORT algorithm is employed to track the human.Eventually, the label of loitering allocated to each person by defining our individual thresholds (block B). A fuzzy inference machine concludes modules results with the help of fuzzy rules to compute robbery potential for prediction and detection (block C).**

the head cover detection module, the number of people can be obtained and the amount of crowded environment can be commented on. Therefore, crowd detection module is actually human's head part detection acquired from head cover detection module. Number of humans attended in store changes the potential of robbery in stores. Low attendance makes higher risk for robbery. As human presence increases, the robbery potential decreases as defined in Eq. 1 and 2.

$$C = \left\{ 0.1, \frac{2}{\{N\}} \right\} \tag{1}$$

$$C^s = \begin{cases} 100 \times MinC & \text{N} = 1 \text{ or N} \geq 7 \\ 100 \times MaxC & 1 < N < 7 \end{cases} \tag{2}$$

Let C be crowd corresponding set, N be the number of people and C^s be crowd score. As a result, C^s is between 0-100 and showing the number of people present in the stores. Based on crowd score we defined, the effect of the presence of people can be considered on RBP.

3) LOITERING DETECTION

People intend to buy something in the store, may look for the stuffs they want. In the end, they choose to go to the cash desk and pay the price. But a robber is looking for a good opportunity to take out his weapon. The robber looks more around the counter, and as a result, the amount of loitering caused by the robber around the counter, is more than normal. On the other hand, the sight of view of surveillance cameras is around counter and robber loiters around it as well. To assess the degree of loitering, each person is tracked with the help of DeepSORT algorithm. Next, the sum of the distance traveled for each person is calculated during every 10 frames. A sliding window with 500 frames length termed snippet in this work, is considered with 10 frames called one step. Loitering is acquired by aggregating total Euclidean distance calculated for every snippet. In each video, after about 10 frames, changes in people's position can be considered. Therefore, after passing 10 frames, the distance traveled by people begins to be calculated and after every one step, the amount of distance traveled is updated again in the range of 500 frames. It is also worth mentioning that cameras used for collecting dataset we have use in this paper were not calibrated and due to this reason, calculated distance of our algorithm is up to scale. On the other hand, distance and displacement of human are equal in our algorithm because of low amount of movement during 10 frames.

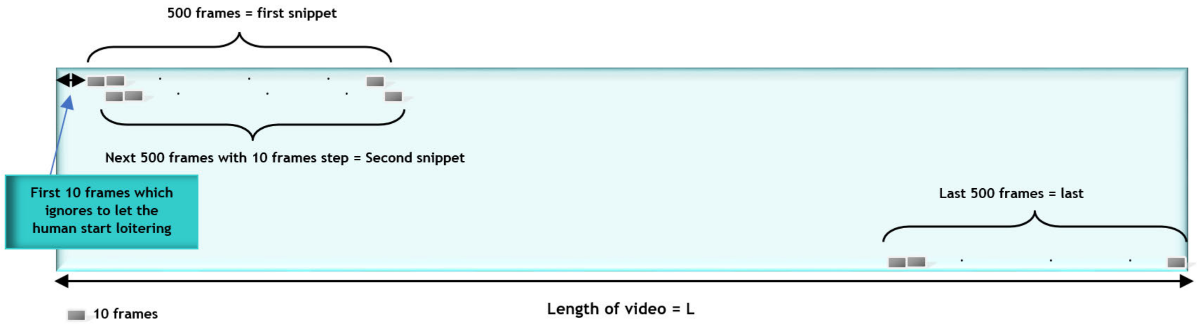

FIGURE 2. Video division into snippets: each 500 frames make a snippet and these snippets slides by one step during the video frames.

Fig. 2, displays procedure of snippet allocation and loitering calculation. As can be seen, a video with length of L is divided into overlapping snippets (S) with 10 frames steps. In fact, if a person is loitering around the counter, the distance traveled by him is more than an appropriate threshold (θ_n), and depending on the amount of movement, each person receives a score between 0 and 100 for loitering. Eq. 3 shows division of a video into 10 to get number of steps. Next, the traveled distance is calculated for each step and aggregated during a snippet. After that, these calculation is done for next snippet with just one step sliding.

$$n = [\frac{L}{10}] \tag{3}$$

Eq. 4 elucidates Euclidean distance calculation and aggregation of them for every snippet.

$$\begin{cases} i = 10j \qquad \mathrm{j = 0 : (n-1)} \\ d_j = \sqrt{(C_{x(i+10)} - C_{x_i})^2 + (C_{y(i+10)} - C_{y_i})^2} \\ D^m = \sum_{i=m}^{50+m} d_j \quad \mathrm{m = 0 : (j-50)} \end{cases} \tag{4}$$

where d_j shows displacement of a human from ith frame to i+10 during one snippet. Besides, t is the number of steps during one snippet and it is equal to 50 because $50 \times 10 =$ 500 frames. Furthermore $C_{x,y}$ shows position of head part for each human and their displacement during 10 frames is shown with i and i+10. d_i is Euclidean distance calculated for every 10 frames and D^m is aggregation of them for every snippets consisting 500 frames (50×10). After calculation of D^m for one snippet, one step slides during the video and the aggregation is calculated again. This exertion is for consideration of the human manner changes during the video. Fig. 3, shows the process of loitering calculation during one snippet.

B. FUZZY INFERENCE MACHINE

The basis of fuzzy set knowledge is awareness of fuzzy logic theory, in which the degree of membership function describes the relationship between a member and the set. In fact, fuzzy logic allows each member to belong to a set in the intermediate state [24]. It can quantify all modules in the form of intermediate values with the appropriate membership rules and functions, to calculate the potential of robbery. In general, the fuzzy inference method has two main types:

- Mamdani approach which follows linguistic fuzzy modeling.
- Takagi-Suegno-Kang (TS) approach which is based on precise fuzzy modeling.

Mamdani approach has high interpretability with low accuracy but TS has high accuracy and low interpretability. Due to the scores obtained from three modules, we need an inference machine which can conclude robbery potential like a human brain. Because of conceptual interpretability of Mamdani approach and interpretable quiddity of our features extracted by three modules, the suitable inference machine for calculating the potential of robbery is Mamdani fuzzy inference machine.

Mamdani fuzzy approach is performed in four steps including: fuzzification, inference, composition and defuzzification. Fuzzification step is for comparing the input variable with membership functions (MF) which defines the feature of fuzzy sets by allocating a corresponding membership value to each element. Inference step is used to combine membership values on the premise step for getting fulfillment command. In composition step, fuzzy or crisp consequent are produced and finally a crisp output is generated by aggregating the consequences and using MFs in defuzzification step. Triangular, trapezoidal and Gaussian function are three main MF used in Mamdani fuzzy inference systems. Defuzzification is the procedure of converting the fuzzified output into crisp value.

In order to provide the algorithm, the modules: head cover, crowd and loitering detection, are considered as input variables and the potential of robbery behavior as output variables. With regard to this issue, a fuzzy inference machine has been designed which uses the rules of Mamdani. Due to simple formula, computational efficiency and an expert knowledge comprising, triangular MF is chosen in this paper. Three parameters a, b and c needed to define triangular MF as Eq. 5.

$$\mu_j = max\left(min\left(\frac{x_i - a}{b - a}, \frac{c - x_i}{c - b}\right), 0\right), \quad (a < b < c) \tag{5}$$

Parameters a,b and c are coordinates of three corners from the triangular MF and they are acquired from experts' knowledge

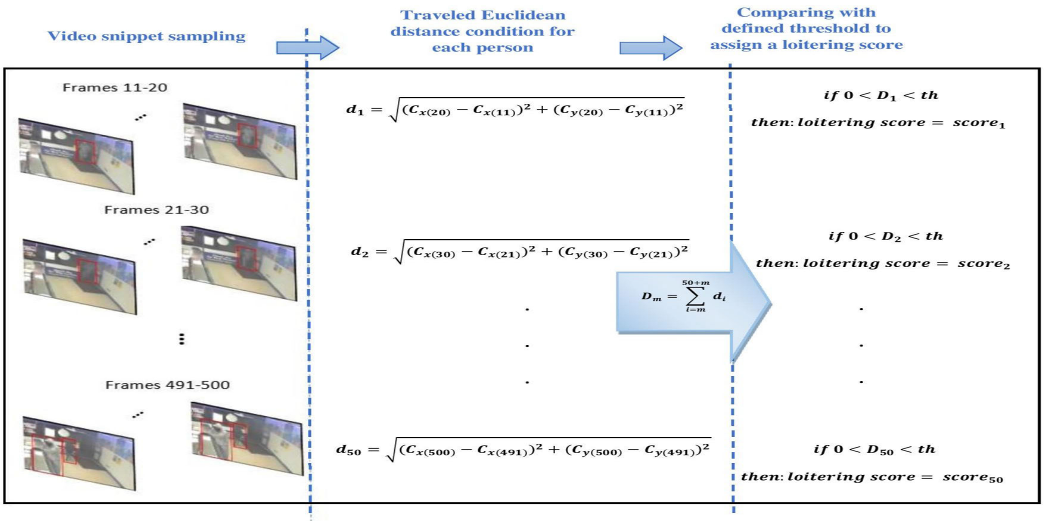

FIGURE 3. Block diagram of loitering detection module. d_i is for Euclidean distance calculation and D_j is aggregation of d_i. For every steps, Euclidean distance calculated and mean distance is computed for every snippets. If loitering carry on, traveled distance accumulated during snippets and loitering score increased.

for all of our detection modules and RBP prediction and detection system.

IV. RESULTS AND DISCUSSION

Our empirical evaluation is on two cases predicting robbery potential and detecting robbery. Preparing proper dataset for retraining deep learning algorithms is for modules created to extract individuals and their head cover beside their loitering. To test our proposed system, 45 videos are eligible from Robbery-UCF-Crime [1] dataset for prediction process. In videos falling into predicting state, there is a period of time that robber loiters for a appropriate moment for robbery. It is crucial to accentuate that prediction signifies for behavior which is not happened yet. For robbery behavior, prediction is feasible up to the moment that the robber shows his weapon, threat or force. While individuals with head cover loiter to get a proper opportunity for threatening or forcing with robbery aim in low crowded shopping malls. On the other hand, 70 videos of Robbery-UCF-Crime dataset are used for robbery detection algorithm. It is noticeable that the detection is the process of finding out the behavior after its occurrence which for robbery behavior it means that the robber shows his weapon, threat or force and maybe leaves the place after robbing something. The 70 videos have condition of potential calculation such as proper camera angle for cash desk sight and visible human. For robbery potential calculation, a fuzzy inference machine is used instead of human brain to allocate a proper potential to each scene of video based on three modules outputs. This part renders experimental results to evaluate our proposed algorithm.

A. DATA PREPARATION

To develop improved YOLOv5s for head cover detection in low-resolution images, we prepared image dataset which gathered from three groups: video image sequences collected from anomaly folders of UCF-Crime dataset except robbery folder [1], Bikes Helmets Dataset,[1] and Mask Dataset.[2]

Thereupon images with proper view angle of CCTV cameras, with more differs in human position and images in which human can be detached from background, are selected. Prepared dataset includes 7621 images comprises 5129 images from Bikes Helmets Dataset, 254 images from Mask Dataset and 2238 images are from UCF-Crime dataset videos except videos from robbery folder, which converted to image sequences. These images divided into training, validation and test sets with approximate ratio of 7:2:1 respectively. Table 2 represents dataset in detail.

Image selection from image sequences obtained from videos of UCF-Crime dataset is done manually with respect to significant changes in background or position of human. Afterward, obtained images annotated with Computer Vision Annotation Tool (CVAT) [54] precisely. CVAT is an annotating tool which can accurately localize the keypoints of bounding boxes and makes a high quality annotation. and we ascertain the head part of humans by a bounding box manually. All images labeled as two classes, masked for human with head cover and no-mask for human without it. The annotation file is in ".txt" format and containing head of human information, one row per head for each image. Each row has information about: "class", "x_c", "y_c", "width" and "height". Where "x_c" and "y_c" includes center information of x and y coordinates for images, respectively. These box coordinates must be normalized by the dimensions of image (from 0 to 1). Normalization of image is based on Eq. 6

[1]https://www.kaggle.com/datasets/andrewmvd/hard-hat-detection
[2]https://www.kaggle.com/datasets/andrewmvd/face-mask-detection

TABLE 2. Images from different dataset used for retraining YOLOV5 in our proposed approach.

Dataset	UCF-Crime	Bikes Helmets	Mask
Number of images for training	1639	3762	187
Number of images for validation	449	1025	51
Number of images for test	150	342	16

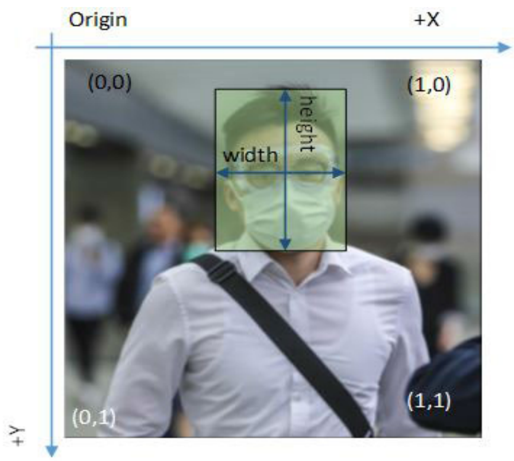

FIGURE 4. Coordinates position belongs to bounding box of human head part.

FIGURE 5. A sample image from prepared dataset, Left: main image - Right: after convolving with Guassian filter of Eq. 8.

and for making all input images uniform.

$$\begin{cases} W = \left(\frac{x_{max} - x_{min}}{w}\right) \\ H = \left(\frac{y_{max} - y_{min}}{h}\right) \\ X = \left(\frac{x_{max} + x_{min}}{2 \cdot w}\right) \\ Y = \left(\frac{y_{max} + y_{min}}{2 \cdot h}\right) \end{cases} \tag{6}$$

where W, H, X and Y are normalized coordinates of width, height, "x_c" and "y_c", respectively. x_{min}, x_{max}, y_{min} and y_{max} are coordinates of bounding box around head part. The w and h are width and height of bounding box, respectively.

Fig. 4, shows each coordinates position in the bounding box which is used for YOLOV5 retraining. Almost 27500 head part specified in our prepared dataset. The resolution of video image sequences from UCF-Crime dataset is low. But we decreased resolution of images belong to Bikes Helmets and Mask datasets to make them adoptable with real world video images of CCTVs mounted within the building.

To prepare appropriate dataset for capable robbery potential detection algorithm, which can be used for videos obtained by CCTV surveillance cameras with low resolution like UCF-Crime dataset, the resolution of images from Bikes Helmets and Mask Dataset should be decreased. Therefore, we convolved these images with Gaussian as Eq. 7 and 8 [51].

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{7}$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{8}$$

In Eq. 7, L(x,y,σ) is scale space of image, G(x,y,σ) is the Gaussian function and I(x,y) is input image. According to resolution of UCF-Crime dataset, we set scale parameter at σ=1.3 for making images smoother and with less details which provides them close to low resolution images of UCF-Crime dataset by blurring. An example of convolving a sample image with Gaussian filter can be seen in Fig. 5. As can be seen, a Region Of Interest (ROI) is magnified to show the edges which becomes smoother. It is notable that, to test our proposed RBP prediction, in addition to videos related to robbery and have condition of prediction, 47 videos from the normal video collection have been reviewed and selected with similar conditions, both in terms of predicting the potential for robbery behavior and also for robbery detection.

B. RETRAINED YOLVv5

To create proposed crowd and head cover detection modules, the YOLOV5s algorithm is used. To customize our gathered low-quality dataset, we modify the YOLOV5s weights by retraining it with our prepared images and suit the YOLOV5s for two classes. This allows us to detect, locate and classify human faces into two classes: with or without head cover. The models complexity performance is determined by the number of parameters and floating-point operations per second (FLOPs). FLOPs measures the complexity of the model by counting the total multiplication and addition operations performed by the model. We demonstrate the lightweight degree of our proposed model by listing the differences in parameters and GFLOPs (giga FLOPs) for original YOLOV5s and our retrained YOLOV5s model in Table 3. As can be seen, the FLOPs reduced more than 25 times under same experimental condition. Low value of GFLOPs shows that the model has lower and faster computation.

It is adopted the torch=1.9.0 framework based on the TensorFlow=1.15.0 implemented in the framework of

TABLE 3. Comparison of different model complexity metrics.

Model	Parameters (M)	FLOPs (G)
YOLOV5s	7.57	15.98
Retrained YOLOV5s	7.22	0.601

TABLE 4. System configuration for retraining YOLOV5s in our proposed method.

Hardware and software	NVIDIA GeForce RTX 2060
Operating system	Windows 10 Pro 64-bit
CPU	Intel® Core™ i7-9750H, CPU@2.60GHz
GPU	CUDA 10.2
RAM	16GB
Deep Learning algorithm	YOLOV5

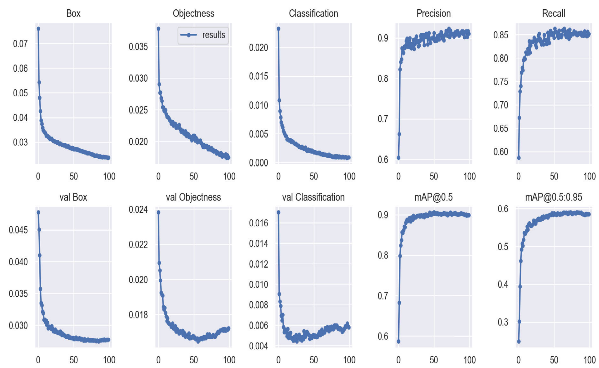

FIGURE 6. Graphs of box loss, objectness loss, classification loss, precision, recall mAP during 100 epoch training process of our method for head cover detection module.

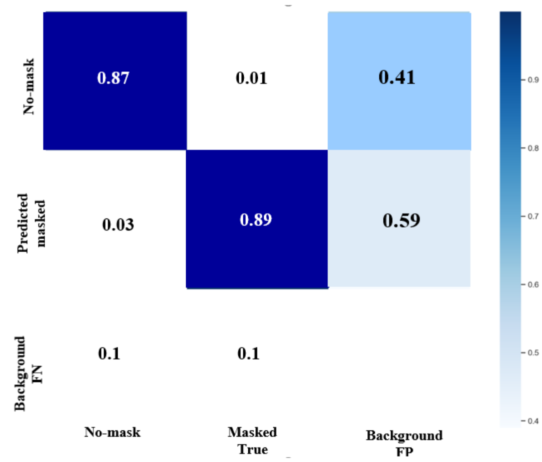

FIGURE 7. Confusion matrix result of head cover detection module for dataset we gathered (Table 2).

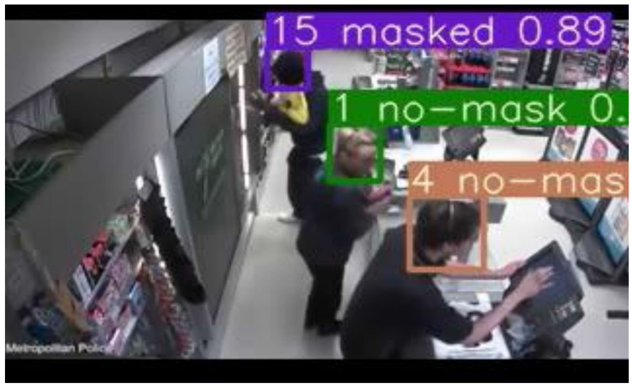

FIGURE 8. Head cover detection results by customized YOLOV5s method for head cover detection module.

PyTorch. To improve human head cover detector, the model was retrained on an NVIDIA GetForce RTX 2060 GPU. Parameters of our model referred to the original YOLOV5 model. Table 4, represents hardware and software configuration of our experimental platform.

This model is retrained by gathered dataset which contains 7622 images (see Table 2). Fig. 6, shows the graphs of the metric curves as training advances contains graphs of Precision, Recall, and Mean Average Precision (mAP) during YOLOv5s training progresses. Eq. 9-11 listed these performance metrics.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$mAP = \frac{1}{C} \times \sum_{c \in C} \frac{|TP_c|}{|TP_c| + |FP_c|} \tag{11}$$

TP shows a positive sample classified rightly by the modules, FP represents a positive sample classified wrongly and FN indicates positive sample which is incorrectly classified [52]. Complementary, C stands for class numbers which is 2 in this paper.

There are three types of loss represented in Fig. 6 containing box loss, objectness loss and classification loss for training and validation sets. Box graph shows loss in covering an object by predicted box. Objectness graph represents loss due to wrong object prediction. Classification loss shows loss in correct prediction of object class. On the other hand, Precision illustrates the ratio of samples number predicted correctly to the total number of positive samples predictions, and Recall represents the positive class is predicted as the number of positive class samples in the total number of positive samples. The mAP represents adaptation of ground-truth bounding box to predicted one and accurate model has higher score in mAP. The mAP@ 0.5 illustrates 50% adaptation and mAP@ 0.5- 0.95 shows 50%- 95% adaptation. Fig. 7, reflects confusion matrices as training progresses of YOLOV5s for two classes, masked and no-masked. For both classes, more than 85% of masked and no-masked human detected correctly during validation process of our proposed model. Further, Fig. 8 represents a sample head cover detection by our customized YOLOV5s method where the head parts are labeled as masked or no-masked which means the person is with head cover or without.

Additionally, as shown in Fig. 9, results for test set images represents more than 90% mAP@ 0.5 for both masked and no-masked classes.

Class	Images	Labels	P	R	mAP@.5	mAP@.5:.95:
all	508	1883	0.914	0.862	0.915	0.6
no-mask	508	701	0.919	0.843	0.902	0.589
masked	508	1182	0.908	0.881	0.927	0.61

FIGURE 9. Detection results of head cover detection module for test set of our gathered dataset (Table 2).

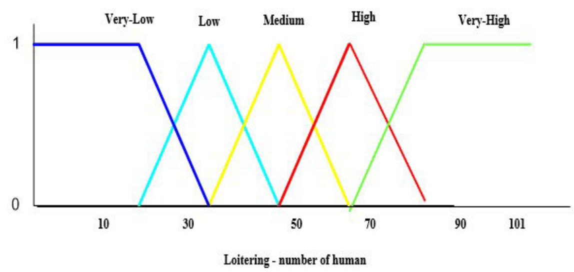

FIGURE 10. Fuzzy MF of crowd and loitering detection modules with 5 linguistic variables containing: Very-Low, Low, Medium, High and Very-High.

C. FUZZY INFERENCE MACHINE

The appropriate inference machine for assembling information of modules and predicting the potential of robbery behavior is fuzzy inference machine. Because it can appraise all modules in the form of intermediate values with the opportune membership rules and functions for RBP computing [53]. Three modules: head cover detection, human detection and loitering computation prepares input variables and the potential of robbery behavior assesses as output variables. As respects to desired output, a fuzzy inference machine is brought forth with Mamdani rules and help of the triangular membership function for the inputs, to deduce the potential of robbery behavior from these inputs. After providing the modules, there are three types of inputs for an inference machine. Due to three covers including glass, mask or hat which causes the label of masked to a person, we assume 3 threshold values for head cover module including "Low", "Medium" and "High". This is because a person with more than two head cover of three is suspicious enough to cover his head. So Low threshold is for one head cover out of these three cover types, Medium is for two out of three covers and High is for three head covers. For human detection and loitering computation there are 5 threshold values, "Very-Low", "Low", "Medium", "High" and "Very-High". These 5 values are for 0-100 quantity describing. As number of individuals increases, potential of robbery decreases. So "Very-High" stands for 2-3 human number because the first person is the seller. Loitering enhancement entails robbery potential increment. Fig. 10- 12, show the definition of linguistic variables as membership functions for three modules and fuzzification of RBP. As can be seen there are 5 linguistic variables for crowd detection module and 5 for loitering detection module as well. Moreover, 3 variables are for head cover detection module.

Therefore, in this fuzzy machine, 75 rules ($5 \times 5 \times 3$) are laid down to implement the relationships between the values extracted by the modules. These rules are based on the

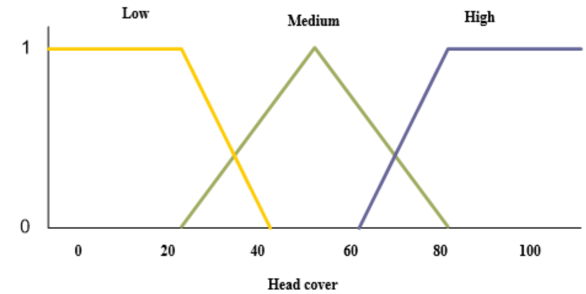

FIGURE 11. Fuzzy MF of head cover detection module with 3 linguistic variables containing: Low, Medium and High.

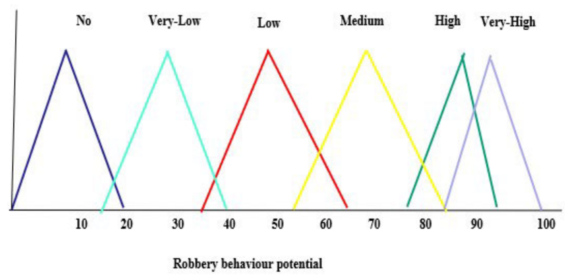

FIGURE 12. Fuzzy MF for RBP prediction and detection. Our defined fuzzy MF has 6 linguistic variables containing: No, Very-Low, Low, Medium, High and Very-High to assign proper RBP for snippets.

effectiveness of each module in increasing or decreasing the robbery potential. The opinion of an expert has been applied in defining these rules. Expert person is the one who has been watching lots of surveillance videos and can understand behavior of robbery immediately. Table 5 presents several rules used in our proposed evaluation of robbery potential prediction. Among 75 rules, rules number 1, 35 and 70 are represented which are from three out of six intervals (see Fig. 12) of robbery potential including: very high, Medium and very low RBP.

Defuzzification method used in this paper is centroid strategy which gives a better representation of the meaning of the three modules because it blends all contributing rules.

D. PREDICTION OF ROBBERY IN VIDEOS

To predict the potential of robbery behavior, we assume only the time period before the moment of displaying weapon, force or threat by the robber. Videos in which robbery occurs and up to the moment of using force, weapons and threats (τ), the potential predicted more than the set threshold (θ), is labeled as data with the potential of robbery behavior and correctly identified (TP). Data with the same conditions that have a potential below θ are also labeled as data with robbery behavior that has not been correctly identified (FN). Fig. 13, represents a diagram of RBP- Time which depicts τ for the time that the robber shows force, threat and weapon. It also shows θ for threshold which is chosen for robbery behavior inferring and happens at the time $t = \tau_\theta$. After t=τ, the robbery behavior has happened definitely and its potential is 100%.

The thresholds with normal and strict conditions are also used to predict the potential for robbery behavior,

TABLE 5. Three sample rules for evaluation of robbery potential prediction.

Rule	Description	Linguistic variable
1	type1-facecover['High']& type2-loitering['Very-High']& type3-crowd['Very-High']	Very-High
35	type1-facecover['Medium'] & type2-loitering['High']& type3-crowd['Very-Low']	Medium
70	type1-facecover['Low']& type2-loitering['Low']& type3-crowd['Very-Low']	Very-Low

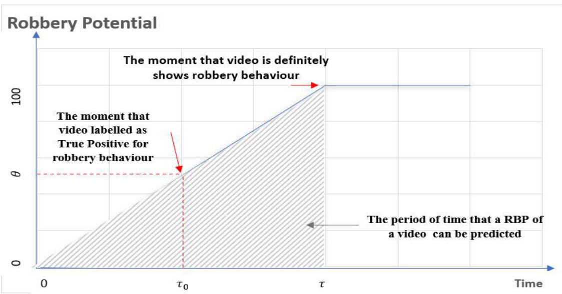

FIGURE 13. Diagram of threshold allocating for RBP prediction. The region(gray hachure) represents period of time that our proposed method can predict RBP.

TABLE 6. Robbery Potential Prediction Results of our proposed method on UCF-Crime dataset.

Threshold	Precision	Recall	F1-score
θ_n	0.474	0.62	0.537
θ_s	0.33	0.33	0.361

respectively as Eq. 12 and 13.

$$\theta_n = 50 \tag{12}$$

$$\theta_s = 60 \tag{13}$$

where θ_n is normal threshold which an expert can conclude robbery occurrence and θ_s is strict one in which a restricted expert person derives robbery will happen. Evaluation considers as binary nature of robbery behavior prediction. we consider a video True-Positive (TP) if predicted RBP is more than θ_n or θ_s and it belongs to Robbery Folder. False-Negative (FN) videos are those belonging to Robbery folder but the estimated RBP is less than θ_n or θ_s. Therefore, videos from normal folder are propounded as True-Negative (TN) with potential assessment of less than θ_n or θ_s and FP with robbery potential appraisement of more than θ_n or θ_s. Recall evaluates the capability of proposed algorithm for predicting robbery potential, whereas Precision determines, high costs of False Positive. Both metrics are in the range from 0 to 1, in which 1 is the optimal value. F1-score is the harmonic means of Recall and Precision and derived by Eq. 14.

$$F1-score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{14}$$

Precision, Recall and the corresponding F1-score give a concept of how well a robbery potential of video is estimated, and how often this assessment is confused with others. Precision, Recall and F1-score is used to appraise the performance of our algorithm. The results are shown in Table 6.

Fig. 15(a-i) shows few frames of a video from Robbery folder of UCF-Crime which is predicted as TP sample and has high robbery potential. As can be seen, there are two humans presented in the scene and the customer has face cover. So, beside a poorly attended store, there is a face covered person who is trying not to be recognized. Moreover, the face covered person, loiters and looks for a proper time to the robber commodity. Thus, the robbery potential should increase by calculating loitering and the reported potential shows likewise. Due to low resolution of videos, sometimes tracking algorithm could not operate correctly and lose person as it is apparent in the Fig. 14-h. frames 14-i and 14-j are those that the robbery behavior has happened and the robber has run.

Fig. 15, shows the curve of RBP computed for each snippet during whole video of the Fig. 14 sample. It starts to be calculated after 10 frames for each video and for last frames it decreases to nearly zero because of ending robbery risk as can be seen in Fig. 14-h 14-i and 14-j. Fluctuation of the curve shows the times modules could not truly calculate values. Final decision of the calculated potential is by averaging the values. Potential computation is as Eq. 15. It is remarkable that snippet numbers are equal to n (see Eq. 3).

$$RP_t = \frac{\sum_{j=1}^{n} \frac{\sum_{k=1}^{50} RP_k^j}{50}}{n} \tag{15}$$

where RP_t is total RBP computed during the video up to the time we want to know the RBP and depicted by τ_0 in Fig. 13. As Eq. 3, n is the number of snippets upto t $= \tau_0$. RP_k^j is robbery potential calculated for each steps with 10 frames.

Fig. 16 shows an instance of FN sample. In lots of frames of this video, like 16-a, 16-b and 16-e, robbery potential has low values due to 1 person appearance but robbery has happened at the end. On the other hand, in some frames like 16-h, no human has been detected to calculate robbery potential and this sample labeled as negative wrongly. Obviously, the angle of camera is not appropriate for tracking people and the head cover of human cannot be detected truly. More precisely, human dimensions are not explicit enough to be detected. So, values of modules calculated incorrectly.

Curve of incorrectly predicted robbery potential for FN sample of Fig. 16 is depicted in Fig. 17. There are lots of steps with less than θ robbery potential. So, calculated RP_t is less than θ and this video labeled as negative wrongly. high frequency of changes in RP_k^j demonstrates weakness of face cover detection due to cameras' angle.

Fig. 18, presents an example video frame of a normal data without any crime occurrence which is labeled as TN. Due to

FIGURE 14. Sample of TP video for robbery potential prediction. In this sample, calculated robbery potential is more than defined threshold and robbery happens at the end. In this video, one person who covers his face and carries weapon, loiters for a proper time to rob something.

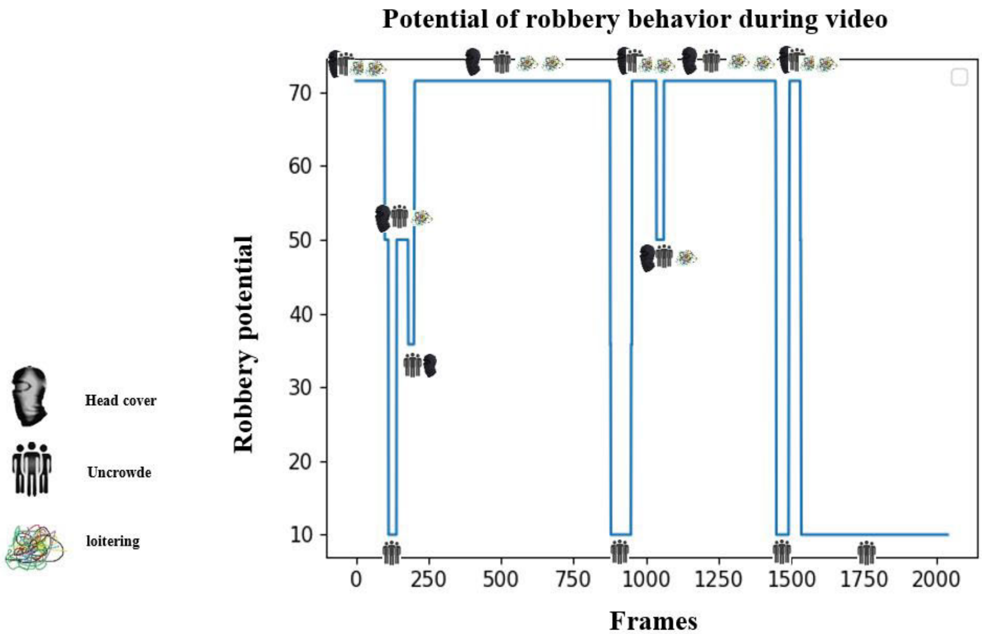

FIGURE 15. Curve of robbery potential for every 10 frames in a TP video of Fig. 14. In many snippets, RBP is more than 70%. In some frames the costumer is not detected because of camera angle and RBP is 50% and less.

numerous humans attended in the scene and no head cover for them, beside low movement of individuals, estimated robbery potential is less than 10%.

Curve of 10 frames robbery potential is shown in Fig. 19. As can be seen, the robbery potential is not change because there is no change in modules' values condition like human number or loitering, that can improve robbery potential. The only variable which affects condition of video is loitering. However, displacement calculated by loitering detection module is not increased regularly for this video. In other words, when a person moves during some frames, he leaves the scene normally and does not repeat the path causeless. Hence, results of loitering module changes for snippets and does not increase robbery potential. As can be seen, it is less than 10% for this TN sample.

A sample video frames of FP data is demonstrated in Fig. 20. Although, due to high values of loitering and poorly attended store, high robbery potential is expected, but the evaluation metric proves the robbery occurrence. Even an expert person would say that the behavior of persons who look around aimlessly in a poorly attended place is not normal. Therefore, values calculated by detection modules increases robbery potential and these kinds of videos classified as FP video.

Next, the predicted RBP is brought as curve of Fig. 21. As can be seen clearly in this figure, the negative video should

FIGURE 16. Sample of FN videos for robbery potential prediction. Calculated robbery potential of this sample is less than defined threshold but robbery happens at the end. Angle of camera can not cover the scene and human can not be detected.

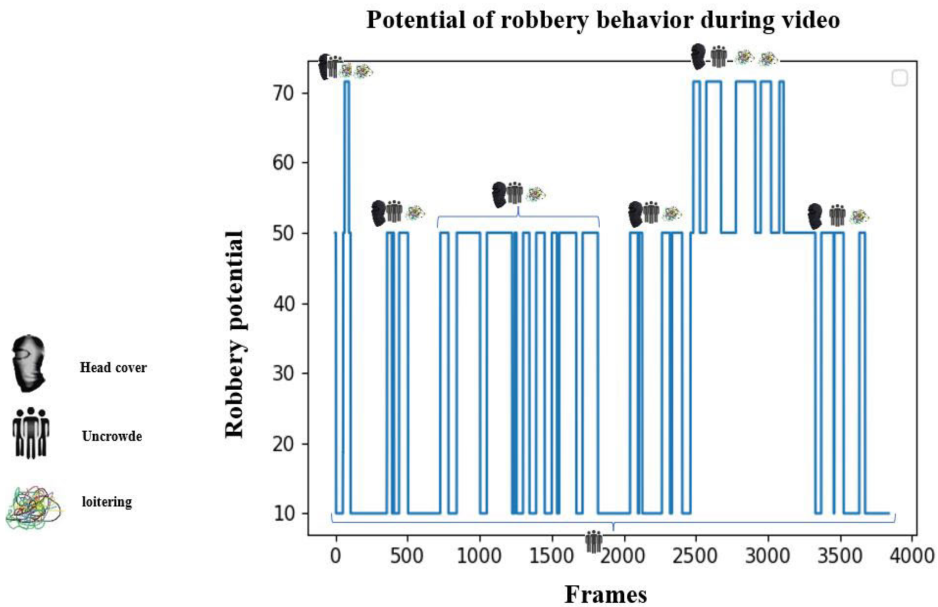

FIGURE 17. Curve of robbery potential for every 10 frames in a FN video of Fig. 16. For snippets with RP_k^j less than 10%, head cover and loitering have not been detected due to inappropriate angle of camera. Snippets with RP_k^j more than 70%, are for frames in which, human are loitering in sight of camera view.

have low robbery potential but, the store has low number of person and they loiter a lot. So higher robbery potential is expected. After nearly 200 snippets, one of the persons leave the scene visible by camera and the robbery potential decreases to 10%.

E. DETECTION OF ROBBERY IN VIDEOS

In this section, we compare the performance of the proposed method with other robbery detection methods for Robbery data of the UCF-Crime data set by transforming the prediction problem to detection. Videos in which the potential of robbery behavior exceeds the θ values are considered as TP and videos with RBP below this value are labeled as FN. Also, in Normal videos, the potential of robbery behavior is calculated. The videos with more robbery-potential calculated by Eq. 15 than θ values are assumed as FP and with less potential are labeled as TN. The results of Precision, Recall and F1-score calculation are given in Table 7. It should be noted that the potential values of robbery that are used to compare with θ values are also obtained by computing truncated mean of snippets. In Table 8, our obtained result is compared with previous methods that have worked in the field of crime detection on UCF-Crime dataset and have reported these results separately for all subdivisions of this dataset. From the previous methods, only the reported results for Robbery data are given. As Table 8 illustrates obviously,

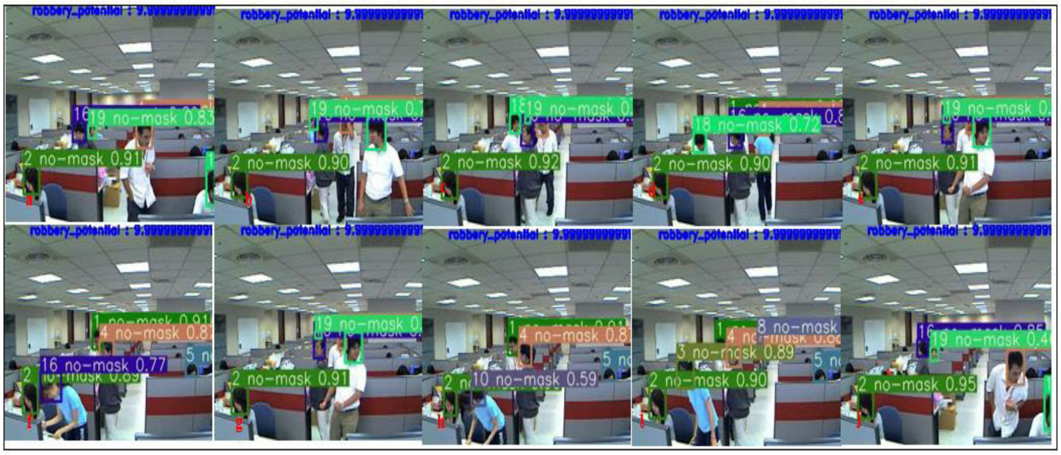

FIGURE 18. Sample of TN videos for robbery potential prediction. In this sample, calculated robbery potential is less than defined threshold due to absence of head covering and low crowd. In this video, robbery does not happen at the end and truly labeled as negative sample.

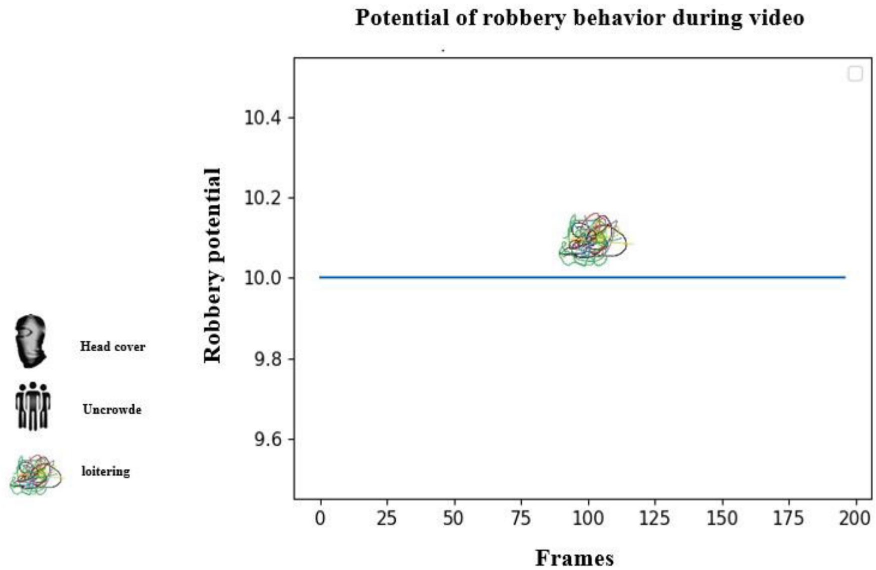

FIGURE 19. Curve of robbery potential for every 10 frames in a TN video of Fig. 18. This video has scores of zero for head cover and crowd detection modules. Additionally, loitering detection module is not increased regularly for it. therefore, the robbery potential is 10% for all snippets.

TABLE 7. Robbery potential detection results of our proposed method on UCF-Crime dataset.

Threshold	Precision	Recall	F1-Score
θ_n	0.77	0.50	0.607
θ_s	0.70	0.33	0.448

our proposed method detects robbery with higher F1-score. It is worth to mention that our method demonstrates the first approach of potential calculation. Method of [55] used a fine-tuning method which is just support videos like UCF-Crime dataset and in [56], the object detection has the key role in robbery detection. It means that if a robber hasn't weapon or fight, detection of robbery will be disturbed. On the other hand, our proposed method is scenario based and can cover different condition of robbery incidence.

F. DISCUSSION

We have a new look at the prediction process of crime. Because of crimes incidence variety, we choose one kind of crime, commercial robbery with the possibility of prediction. Potential calculation of robbery occurrence is crucial to prevent it before any financial or life event falls out. As mentioned before, robbery potential means the amount of robbery occurrence risk in a video. We defined a methodology with crowd, head cover and loitering detection modules and one fuzzy inference machine. For head cover and human detection modules implementation by YOLOV5s, images must gather due to the position of CCTVs cameras based on camera points of view and video images resolution. Number of images in dataset is important for retraining YOLOV5s. Main YOLOV5

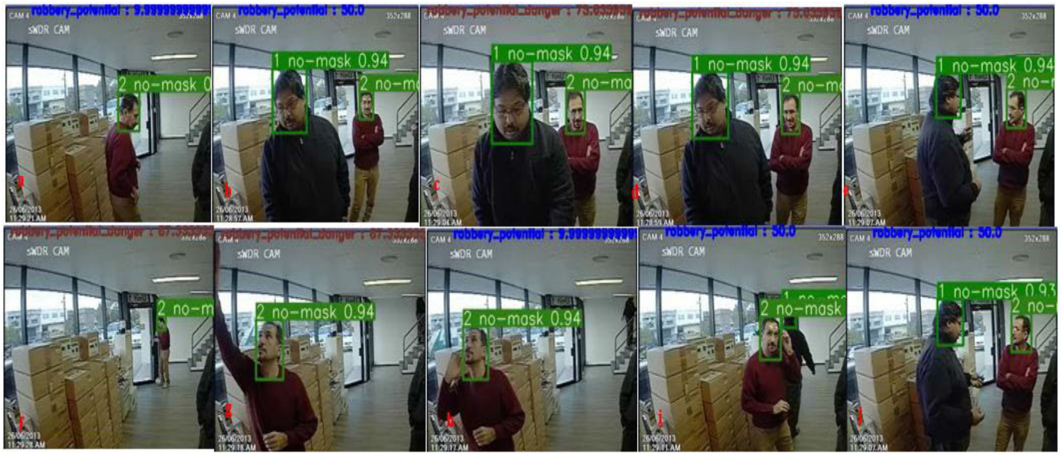

FIGURE 20. Sample of FP videos for robbery potential prediction. Due to high value of loitering and low number of person, calculated robbery potential of this sample is more than defined threshold but robbery does not happen at the end.

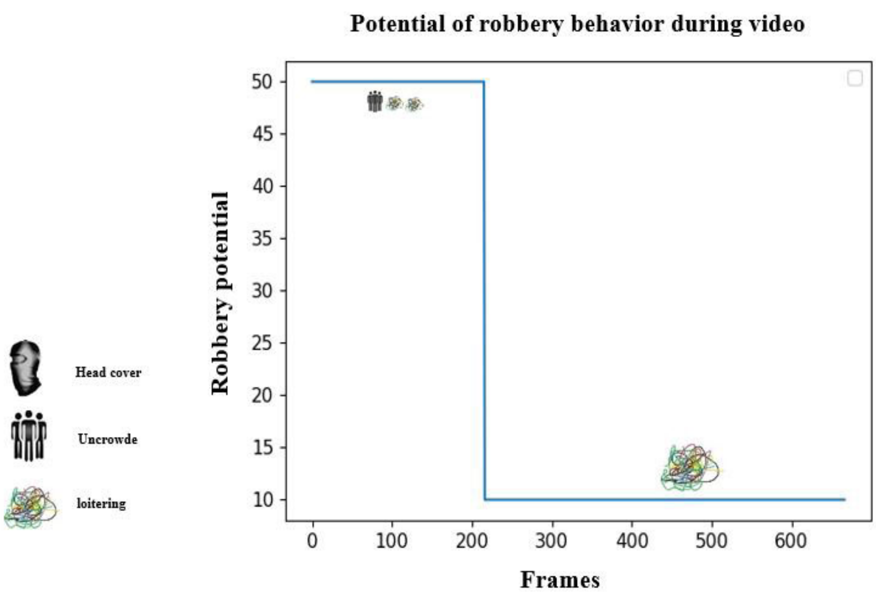

FIGURE 21. Curve of robbery potential for every snippets in a FP video of Fig. 20. The first 200 snippets have high score of low human number and high loitering. but after 200 snippets, one person leave the seen and score of crowd detection module omitted.

TABLE 8. Robbery potential prediction results on UCF-Crime dataset.

Paper	Method	Precision	Recall	F1_score
Maqsood R. [55]	3D ConvNet finetuning by UCF-Crime Dataset	0.35	0.45	0.45
Öztürk, Halil İbrahim (ADOR) [56]	Fusion of object relation information with spatio-temporal features by faster RCNN	–	–	0.448
Öztürk, Halil İbrahim (ADNET) [56]	Spatio-temporal CNN	–	–	0.538
Proposed Method	YOLOV5 finetuning and DeepSORT for feature extraction as three modules, Fuzzy Inference Machine	0.77	0.50	0.607

has been trained by COCO dataset with high resolution images and it could not detect human in UCF-Crime dataset with high accuracy. So that we needed low resolution images to retrain it. On the other hand, surveillance cameras are mostly mounted on top of the counter and the angle of its view point is different for various CCTVs. The difficulty of data collecting increased when we had to look for data with the head cover like the robbers. Also, since helmets are a good

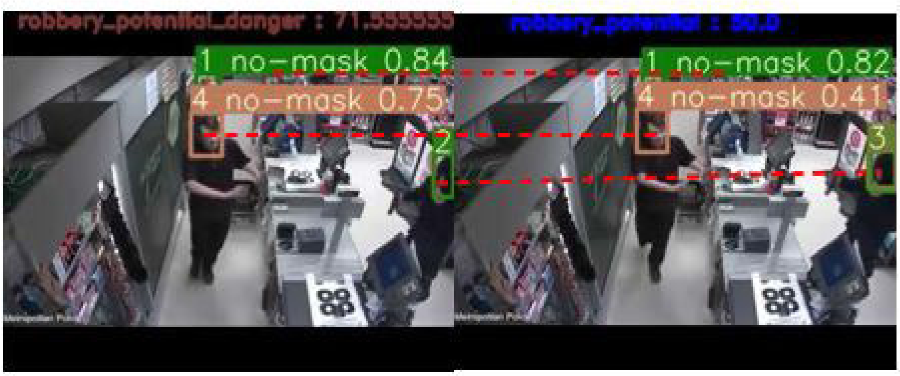

FIGURE 22. Missing ID number sample two frames. ID=2 and ID=3 are for the same person.

choice for covering the whole head, they are highly regarded by robbers. Therefore, after multitude searches, we have gathered considered images with respect to their camera's view point and different kind of head cover. We also decreased resolution of them. Head cover and human detection modules are implemented simultaneously with one dataset. It is worthy to note that according to COVID-19 pandemic, wearing the mask and covering faces is not unusual. Anyhow, face and head occlusion increases probability of suspicious behavior occurrence and a normal person who wears face mask will not do more suspicious behaviors such as loitering. On the other hand, a robber who wears a face mask because of preventing the spread of COVID-19 will enhance his potential of robbery even though without motivation of head covering.

Our loitering detection module tracks human by DEEP-SORT algorithm which is simple, high accurate and has state of the art approaches for multiple object tracking. It uses Faster Region CNN (FrRCNN) detection framework for one class object detection with the parameters learn for the PASCAL VOC high resolution dataset. also, Kalman filter and Hungarian algorithm is used for object tracking [41]. We use Euclidean distance calculation to compute displacement of a person based on our innovated approach. We define our particular thresholds based on prior opinion and by dividing amount of displacements into intervals, we allocated loitering scores to each interval.

This is the first work which is calculating RBP for predicting crime incidence and also detecting it. Previous crime detection approaches are mostly extract common features of the crime. for example, main indicator of robbery is weapon and by detecting it in the videos, they deduce robbery occurrence. But weapon is used in many other crimes such as murdering, kidnapping and fighting. Even by considering skirmish indicator in robbery behavior, it can be seen in other crimes we named. Therefore, by considering a scenario-based procedure for robbery, the approach is especially for robbery behavior. As a consequence, our method has flexibility of involving different cultures and manifestation of various countries for specific robbery occurrence.

Ultimately a fuzzy inference machine infers the robbery potential. The rules of fuzzy inference machine are optimized by comments of an expert person who has been watched lots of CCTV videos to prevent robbery or other crimes. The expert person can devise the fuzzy rules based on any other common scenario and effects of each module can be distinct.

Because of the lack of accurate separation between humans and their tracking due to the low resolution of video images, we have limitation for loitering detection module implementation. With the disappearance of humans, it is difficult to rediscover them. Additionally, in most cases, a new label is assigned to the previous person and the distance traveled by that person is recalculated. Alternatively, overlapping of human in video images increases the error of missing people. Therefore, this module has errors for precise loitering detection in some videos of UCF-Crime dataset. Fig. 22, represents a frame sample that DeepSORT algorithm allocated more than one ID for same person because of missing him.

V. CONCLUSION

This research work proposes an approach for RBP prediction in video surveillance images. There are several challenges of CCTV videos like the various ways for robbery incidence, variety in camera angle mounted in different places and low resolution of video images acquired by CCTVs. Tackling these obstacles ensues timely actions and prevents robbery fully or partially observable from surveillance videos. This work is conducted because based on our extensive literature review, despite significance of preventing robbery occurrence, no RBP prediction has been done before. We extract some common scenarios of robbery occurrence with the help of an expert comments and by watching several robbery videos from CCTVs. We investigate these scenarios to deduce more common features between them and implement a practical approach for RBP prediction. Our study proposes a deep-learning based approach with the help of fuzzy inference machine to calculate potential of robbery. This approach provides a retrained YOLOV5 algorithm by gathering proper dataset of human with or without head cover. This deep-learning based algorithm is used to efficiently implement crowd and head cover detection modules. This paper also executes loitering module by our defined methodology which calculates the Euclidean traveled distance of individuals using DeepSORT method. A fuzzy inference machine is delineated to infer robbery potential of videos for every 10 frames and average them for every snippet based on three module results. The proposed method is applied to the Robbery folder of UCF-Crime dataset and F1-score of proposed system is 0.537. This result shows that our proposed methodology can correctly predict robbery potential for more than half of the videos.

Accordingly, we change the problem of predicting to robbery detection one. Thus, we can compare it with prior literature which have worked on the anomaly-detection specially the robbery detection and their dataset is UCF-Crime. F1-score of detection method is 0.607 and it is utmost among other methods. The result proves that our proposed scenario-based system works correctly with high ability in detecting and also predicting robbery behavior. Our proposed approach can be used by any places which have surveillance cameras and want to prevent robbery crime. They do not need to employ a person to watch the real time videos of theses cameras precisely and infer the robbery potential. However, this person should watch the videos uninterruptedly to not make a mistake. Additionally, any one can make our methodology privately by changing the thresholds value due to particular culture.

We can increase F1-score by improving loitering detection accuracy. As future work, we intend to achieve an improved tracking algorithm for low-resolution video images by improving DeepSORT method. Human of low- resolution videos cannot be detected precisely to track. This is because the detector of DeepSORT algorithm is FrRCNN. Therefore, we will change detection framework of DeepSORT algorithm to retrained YOLOV5 by low-resolution human images. The proposed YOLOV5 will have only one object class, low resolution images.

REFERENCES

[1] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[2] J. James P. Suarez and P. C. Naval Jr., "A survey on deep learning techniques for video anomaly detection," 2020, *arXiv:2009.14146*.

[3] T. Mei and C. Zhang, "Deep learning for intelligent video analysis," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1955–1956.

[4] H. Yan, X. Liu, and R. Hong, "Image classification via fusing the latent deep CNN feature," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, Aug. 2016, pp. 110–113.

[5] M. Ghazal, C. Vazquez, and A. Amer, "Real-time automatic detection of vandalism behavior in video sequences," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2007, pp. 1056–1060.

[6] I. P. Febin, K. Jayasree, and P. T. Joy, "Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm," *Pattern Anal. Appl.*, vol. 23, no. 2, pp. 611–623, May 2020.

[7] W. Lao, J. Han, and P. De With, "Automatic video-based human motion analyzer for consumer surveillance system," *IEEE Trans. Consum. Electron.*, vol. 55, no. 2, pp. 591–598, May 2009.

[8] A. G. Ferguson, "Predictive policing and reasonable suspicion," *Emory Law J.*, vol. 62, no. 2, p. 259, 2012.

[9] C. Beck and C. McCue, "Predictive policing: What can we learn from Wal-Mart and Amazon about fighting crime in a recession?" *Police Chief*, vol. 76, no. 11, p. 18, 2009.

[10] K. J. Bowers and S. D. Johnson, "Who commits near repeats? A test of the boost explanation," *Western Criminol. Rev.*, vol. 5, no. 3, pp. 12–24, 2004.

[11] Seattle Police Department, "SPD 2021 year-end crime report," Seattle, WA, USA, 2021. [Online]. Available: https://www.seattle.gov/documents/Departments/Police/Reports/2021_SPD_CRIME_REPORT_FINAL.pdf

[12] (2019). *FBI*. [Online]. Available: https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/robbery

[13] B. Fawei, J. Z. Pan, M. Kollingbaum, and A. Z. Wyner, "A semi-automated ontology construction for legal question answering," *New Gener. Comput.*, vol. 37, no. 4, pp. 453–478, Dec. 2019.

[14] R. Thompson, "Understanding theft from the person and robbery of personal property victimisation trends in England and Wales," Nottingham Trent Univ., Nottingham, U.K., Tech. Rep. 2010/11, 2014.

[15] P. J. Cook, "Robbery violence," *J. Criminal Law Criminol.*, vol. 78, no. 2, pp. 357–376, 1987.

[16] J. D. McCluskey, "A comparison of Robbers' use of physical coercion in commercial and street robberies," *Crime Delinquency*, vol. 59, no. 3, pp. 419–442, Apr. 2013.

[17] D. F. Luckenbill, "Patterns of force in robbery," *Deviant Behav.*, vol. 1, nos. 3–4, pp. 361–378, Apr. 1980.

[18] T. Ishikawa and T. T. Zin, "A study on detection of suspicious persons for intelligent monitoring system," in *Proc. Int. Conf. Big Data Anal. Deep Learn. Appl.* Singapore: Springer, 2018, pp. 292–301.

[19] J. R. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016, *arXiv:1612.00390*.

[20] M. Shah, O. Javed, and K. Shafique, "Automated visual surveillance in realistic scenarios," *IEEE Multimedia Mag.*, vol. 14, no. 1, pp. 30–39, Jan. 2007.

[21] A. Biswas, S. C. Ria, Z. Ferdous, and S. N. Chowdhury, "Suspicious human-movement detection," Ph.D. dissertation, Dept. Comput. Sci. Eng., BRAC Univ., Dhaka, Bangladesh, 2017.

[22] 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 Syst. Appl.*, vol. 42, no. 21, pp. 7991–8005, Nov. 2015.

[23] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 393–402, Jan. 2020.

[24] F. Wu, G. Jin, M. Gao, Z. HE, and Y. Yang, "Helmet detection based on improved YOLO V3 deep model," in *Proc. IEEE 16th Int. Conf. Netw., Sens. Control (ICNSC)*, May 2019, pp. 363–368.

[25] Y. Liu, X.-K. Wang, W.-H. Hou, H. Liu, and J.-Q. Wang, "A novel hybrid model combining a fuzzy inference system and a deep learning method for short-term traffic flow prediction," *Knowl.-Based Syst.*, vol. 255, Nov. 2022, Art. no. 109760.

[26] V. Vaidehi, S. Monica, S. M. S. Safeer, M. Deepika, and S. Sangeetha, "A prediction system based on fuzzy logic," in *Proc. World Congr. Eng. Comput. Sci.*, 2008, pp. 1–6.

[27] M. Elhamod and M. D. Levine, "Automated real-time detection of potentially suspicious behavior in public transport areas," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 688–699, Jun. 2013.

[28] C. Rajapakshe, S. Balasooriya, H. Dayarathna, N. Ranaweera, N. Walgampaya, and N. Pemadasa, "Using CNNs RNNs and machine learning algorithms for real-time crime prediction," in *Proc. Int. Conf. Advancements Comput. (ICAC)*, Dec. 2019, pp. 310–316.

[29] H. Bouma, J. Baan, G. J. Burghouts, P. T. Eendebak, J. R. van Huis, J. Dijk, and J. H. C. van Rest, "Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall," *Proc. SPIE*, vol. 9253, pp. 112–120, Oct. 2014.

[30] E. Selvi, M. Adimoolam, G. Karthi, K. Thinakaran, N. M. Balamurugan, R. Kannadasan, C. Wechtaisong, and A. A. Khan, "Suspicious actions detection system using enhanced CNN and surveillance video," *Electronics*, vol. 11, no. 24, p. 4210, Dec. 2022.

[31] D. Roy and K. M. C., "Snatch theft detection in unconstrained surveillance videos using action attribute modelling," *Pattern Recognit. Lett.*, vol. 108, pp. 56–61, Jun. 2018.

[32] G. Kaur, R. Sinha, P. K. Tiwari, S. K. Yadav, P. Pandey, R. Raj, A. Vashisth, and M. Rakhra, "Face mask recognition system using CNN model," *Neurosci. Informat.*, vol. 2, no. 3, Sep. 2022, Art. no. 100035.

[33] J. Ieamsaard, S. N. Charoensook, and S. Yammen, "Deep learning-based face mask detection using YOLOV5," in *Proc. 9th Int. Electr. Eng. Congr. (iEECON)*, Mar. 2021, pp. 428–431.

[34] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, and J. Hemanth, "SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2," *Sustain. Cities Soc.*, vol. 66, Mar. 2021, Art. no. 102692.
[35] S. Sethi, M. Kathuria, and T. Kaushik, "Face mask detection using deep learning: An approach to reduce risk of coronavirus spread," *J. Biomed. Informat.*, vol. 120, Aug. 2021, Art. no. 103848.
[36] G. J. Chowdary, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Face mask detection using transfer learning of InceptionV3," in *Proc. Int. Conf. Big Data Anal.* Cham, Switzerland: Springer, 2020, pp. 81–90.
[37] S. Singh, U. Ahuja, M. Kumar, K. Kumar, and M. Sachdeva, "Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment," *Multimedia Tools Appl.*, vol. 80, no. 13, pp. 19753–19768, May 2021.
[38] L. Huang, Q. Fu, M. He, D. Jiang, and Z. Hao, "Detection algorithm of safety helmet wearing based on deep learning," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 13, p. e6234, 2021.
[39] F. Zhou, H. Zhao, and Z. Nie, "Safety helmet detection based on YOLOv5," in *Proc. IEEE Int. Conf. Power Electron., Comput. Appl. (ICPECA)*, Jan. 2021, pp. 6–11.
[40] T. Choudhury, A. Aggarwal, and R. Tomar, "A deep learning approach to helmet detection for road safety," *J. Sci. Ind. Res.*, vol. 79, no. 6, pp. 509–512, 2020.
[41] R. Nayak, M. M. Behera, V. Girish, U. C. Pati, and S. K. Das, "Deep learning based loitering detection system using multi-camera video surveillance network," in *Proc. IEEE Int. Symp. Smart Electron. Syst. (iSES) (Formerly iNiS)*, Dec. 2019, pp. 215–220.
[42] Y. Nam, "Loitering detection using an associating pedestrian tracker in crowded scenes," *Multimedia Tools Appl.*, vol. 74, no. 9, pp. 2939–2961, May 2015.
[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[44] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved YOLO-V3 model," *Comput. Electron. Agricult.*, vol. 157, pp. 417–426, Feb. 2019.
[45] S. Pouyan, M. Charmi, A. Azarpeyvand, and H. Hassanpoor, "Significantly improving human detection in low-resolution images by retraining YOLOv3," in *Proc. 26th Int. Comput. Conf., Comput. Soc. Iran (CSICC)*, Mar. 2021, pp. 1–6.
[46] L. Zhao and S. Li, "Object detection algorithm based on improved YOLOv3," *Electronics*, vol. 9, no. 3, p. 537, Mar. 2020.
[47] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[48] D. Luvizon, H. Tabia, and D. Picard, "SSP-Net: Scalable sequential pyramid networks for real-time 3D human pose regression," 2020, *arXiv:2009.01998*.
[49] Q. Z. Li and M. H. Wang, "Development and prospect of real time fruit grading technique based on computer vision," *Trans. Chin. Soc. Agricult. Machinery*, vol. 30, no. 6, pp. 1–7, 1999.
[50] W. Jia, S. Xu, Z. Liang, Y. Zhao, H. Min, S. Li, and Y. Yu, "Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector," *IET Image Process.*, vol. 15, no. 14, pp. 3623–3637, Dec. 2021.
[51] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[52] Q. Xu, R. Lin, H. Yue, H. Huang, Y. Yang, and Z. Yao, "Research on small target detection in driving scenarios based on improved YOLO network," *IEEE Access*, vol. 8, pp. 27574–27583, 2020.
[53] Y. Chen, "Support vector machines and fuzzy systems," in *Soft Computing for Knowledge Discovery and Data Mining*. Boston, MA, USA: Springer, 2008, pp. 205–223.
[54] D. Roberts, M. Wang, W. T. Calderon, and M. Golparvar-Fard, "An annotation tool for benchmarking methods for automated construction worker pose estimation and activity analysis," in *Proc. Int. Conf. Smart Infrastruct. Construct. (ICSIC)*, Jan. 2019, pp. 307–313.
[55] R. Maqsood, U. I. Bajwa, G. Saleem, R. H. Raza, and M. W. Anwar, "Anomaly recognition from surveillance videos using 3D convolution neural network," *Multimedia Tools Appl.*, vol. 80, no. 12, pp. 18693–18716, May 2021.
[56] H. I. Öztürk, "Temporal anomaly localization," M.S. thesis, Dept. Comput. Eng., Inst. Sci., Hacettepe Univ., Ankara, Turkey, 2021.

SHIMA POUYAN was born in Tehran, Iran, in 1989. She received the B.S. and M.S. degrees in electrical engineering from the University of Zanjan, Zanjan, Iran, in 2015, where she is currently pursuing the Ph.D. degree in electrical engineering. She has authored or coauthored more than five research papers in international journals and international conferences. Her current research interests include image processing, artificial intelligence, and deep learning methods.

MOSTAFA CHARMI was born in Zanjan, Iran, in 1977. He received the M.S. degree in biomedical engineering from the Sharif University of Technology, Iran, and the Ph.D. degree in biomedical engineering from Tarbiat Modares University, Iran, in 2010. Since 2011, he has been an Assistant Professor with the Electrical Engineering Department, Zanjan University, Zanjan. He is the author of more than 30 articles and supervisor of more than 30 projects. His research interests include image processing, computer vision, machine vision, and software development for image processing applications.

ALI AZARPEYVAND was born in Zanjan, Iran, in 1975. He received the B.Sc. degree in computer engineering from the Sharif University of Technology, Tehran, Iran, in 1997, the M.Sc. degree in computer architecture from the University of Tehran, Tehran, in 2000, and the Ph.D. degree from the School of Electrical and Computer Engineering, University of Tehran, in 2013. From 1998 to 2004, he was a Senior Digital Design Engineer with Emad Semicon Company, Tehran, while he was the Pager Receiver. In 2001, he joined the University of Zanjan, Zanjan, as an Instructor, and served as a Head of the Computer Engineering Department and a Research Deputy Director of the Faculty of Engineering. He is currently an Assistant Professor with the University of Zanjan. He was the Founder and the Director of the PORTS Research Laboratory and the Director of the Information Technology Center, University of Zanjan. His current research interests include computer architecture, reliability analysis, soft error mitigation techniques, approximate computing, high-performance computing, and bioinformatics.

HOSSEIN HASSANPOOR received the B.S. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 2006, and the M.S. and Ph.D. degrees in biomedical engineering from the Amirkabir University of Technology, Tehran, Iran, in 2009 and 2015, respectively. His current research interests include data mining, machine learning (cognitive method), and the modeling of biological systems.