IET Image Processing

Research Article

Novel framework for automatic localisation of gun carrying by moving person using various indoor and outdoor mimic and real-time views/Scenes

ISSN 1751-9659
Received on 12th May 2020
Revised 29th November 2020
Accepted on 5th January 2021
E-First on 3rd March 2021
doi: 10.1049/iet-ipr.2020.0706
www.ietdl.org

Rajib Debnath[1], *Mrinal Kanti Bhowmik*[1] ✉

[1]*Department of Computer Science and Engineering, Tripura University (A Central University), Suryamaninagar 799022, India*

✉ *E-mail: mrinalkantibhowmik@tripurauniv.in*

Abstract: Hand held gun detection has an important application in both the field of video forensic and surveillance, because, gun is operative by hand only while committing any crime with it. The significant application encompasses the vulnerable places, such as around airport, marketplace, shopping malls, etc. In view of non-availability of relevant public data set, this study provides a newly created mimicked video data set for detection of gun carried by a person and entitled as Tripura University Video Data set for Crime-Scene-Analysis (TUVD-CSA). Effects of illumination, occlusion, rotation, pan, tilt, scaling of gun are effectively demonstrated in it. Moreover, the authors proposed an Iterative Model Generation Framework (IMGF) for gun detection, which is immune to scaling and rotation. Instead of locating the best matched object (gun) in the whole reference image to a query model via exhaustive search, IMGF searches only where the moving person carrying gun appears, which drastically reduces the computational overhead associated with a general template matching scheme. This has been employed by the background subtraction algorithm. Experimental results demonstrate that the proposed IMGF performs efficiently in gun detection with lesser number of true-negatives compared with the state-of-the-art methods.

1 Introduction

Of late video surveillance is extensively used as a monitoring tool for ensuring the security of particular area [1, 2]. Closed Circuit Television (CCTV) footage of the crime area and its analysis are used in forensic for discovering a clue to detect suspect [3]. Security systems are already installed at the entrance of important areas such as airports, office areas, places of worships, shopping mall, banks, building premises, etc. [2]. Along with these security issues, video monitoring systems are used to reduce other crimes and social offenses in public areas. CCTV footage is also accepted as evidence in courts for prosecution [3, 4]. In the video monitoring system there are two main components, one is a remote camera, and another is an operator. The operator monitors the videos transmitted from the remote camera to a screen of the base station. Therefore, the operator has to perform simultaneous work to monitor all the video feeds and along with detecting suspicious activities of any objects carrying guns. Thereby they successfully able to collect evidence followed by informing the appropriate authorities to handle the situation [5]. It is a challenging task for an operator to pay attention to all the videos. One study suggests that detection rates for operators engaged in monitoring 4, 9 and 16 screens oscillate around 83, 84 and 64%, respectively, which drop significantly after an hour [6]. According to Velastin *et al.* [7], a CCTV operator suffers from video blindness after 20–40 min of active monitoring due to which he or she becomes unable to recognise objects in video feeds. According to another study as available in [8], an operator may often miss up to 45% of screen activity after 12 min of continuous watching, and this miss rate goes as high as 95% after elapse of 22 min duration. Therefore, automation of suspicious object detection with a gun or without a gun becomes imperative for achieving comprehensive security and surveillance system. Such an automated system is liable to raise the alarm or indication whenever any abnormal activity is encountered under CCTV surveillance, due to which the operator will prioritise his attention on the video feed and will initiate appropriate action thereon [5].

Forensic team also analyses the CCTV footage of the crime scene at the first stage of an investigation. They collect different crime footage from different sources as much as possible. The analysis of these video footage's is helpful in detecting suspect, crime gun and other pieces of evidence. Detection of the suspicious object with a gun is an imperative task for the investigators. For person(s) carrying gun and/or engaged in the shooting may be identified from the CCTV footage's as formidable legal evidence. Detection of a specific position of the gun held by the moving object as available from the video scene from CCTV footage is extremely important. The aim of this work is the detection of the object as well as the gun being carried by the object. There are several reporting related to moving object detection, object recognition, object classification, object tracking, etc. from surveillance videos [9–12]. However, few works are available in the existing literature engrossing the simultaneous detection of object and gun [5, 13–17]. According to a survey by the US Department of Justice, on an average, about 75% of the homicide take place during 1994–2017 involves the use of handgun [18, 19]. The situation is even worse presently [20].

In the last few decades, a very few data sets are designed to meet the increasing demands in developing and benchmarking new models for gun detection. There are some reporting of visual data sets used by the researchers tabulated in Table 1.

We mentioned all the key features of these data sets. It can be observed that Gun Movies data set [13] is the only data set that has been captured and the data is collected in only indoor conditions with a uniform background, and other data sets [21, 22] are collected from the movie clips. Each of these data sets is extensive in terms of amount or features. However, there is still a lack of video data set for gun detection that can provide a balanced coverage in terms of collecting data in both indoor and outdoor conditions, different complex background, mass firing scenario, illumination change, pan, tilt, scaling and rotation of guns etc.

Problems indicated in the present context [5, 13–17] are sudden illumination change, moving background and shadows in videos. The crowded indoor and outdoor regions are particularly subject of concern with respect to these syndromes. So far our knowledge the number of relevant data sets is scanty. Therefore, we proposed a video data set addressing the described problem entitled as 'TUVD-CSA' (the data set is available for research community in (http://www.mkbhowmik.in). The video data set comprises of

Table 1 Key features of existing gun detection data sets

Database name	Key features	Videos/ images	Data set details				
			Environment condition	Format	Database Type	Pixel resolution	Ground truth
Gun movies data set [13]	Captured in indoor conditions, background is constant, One object carrying gun at one time	videos	indoor	.mp4	visual	640×480	no
IMFDB [21]	Consist of images of different guns from movies, available freely, large variety of data, dynamic background, occlusion and shadow, appearance change, Motion Blur	images	indoor and outdoor	.jpg	visual	different	no
data set of Olmos *et al.* [22]	Collected from internet, labelled manually, consist only images of pistols, Camouflage Foreground Object, complex background, shadow	images	indoor and outdoor	.jpg	visual	different	no

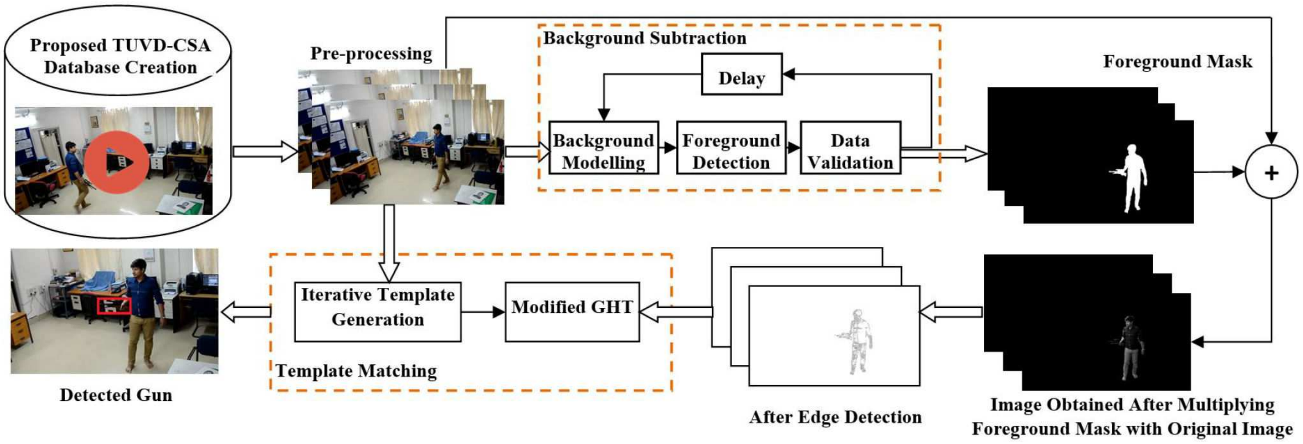


Fig. 1 *Proposed system flow*

persons carrying gun, persons without carrying a gun. Different features, such as rotation, scaling, illumination change make the data set as versatile as possible. We also propose, an iterative model generation framework (IMGF) for gun detection. Object detection and modified generalised Hough transform (M-GHT)-based template matching is the primary step of IMGF. In IMGF, a general and discriminative model or template is generated using an iterative training. The performance of the proposed method is compared with the state-of-the-art methods and it attains fair accuracy compared with state-of-the-art methods. The performance of methods is calculated by comparing the results of methods with the corresponding ground truths (GTs). Fig. 1 depicts the flow of the proposed system. Main contributions of this paper are summarised as follows:

- *We proposed a newly created TUVD-CSA*: The data set comprises 150 video clips (>500,000 frames) created in indoor/outdoor environment (**65 and 60 mimicked videos in indoor and outdoor, respectively**) of moving persons carrying gun in hand and also **25 movie clips** are collected to include the real-time scenario with different challenges (in Section 3).
- Designing of a novel automatic gun detection framework based on the template matching using M-GHT, which is immune to scaling and rotation (in Section 4).
- GTs are also annotated (in Section 3.4).
- A detailed comparison with existing state-of-the-art methods exhibiting substantial performance superiority over its competition is presented in this work (in Section 5).

The paper is organised as follows. Section 2 describes the related literature in the field of gun detection. Section 3 describes the designing issues, features of proposed data set (TUVD-CSA) along with the GT generation. Section 4 presents the proposed methodology in detail and Section 5 provides the analysis of the results and finally, Section 6 concludes the paper.

2 Related work

In the literature, there are a very few amount of work has been reported in the field of visual gun detection. In [14] they concentrate on colour and point descriptor of the gun. Harris point detector and FREAK descriptor are used for the gun detection. In their next work [5] they used SURF and attains better results. In [17], initially cascade classifier is trained with HOG features using a global threshold to classify the positive images. Gun detection using GBVS map has been used in [23], with an assumption that gun will be held by a person. It fails, if any person tries to take the gun from floor, shoes, even if a group of persons dealing with guns. Bag of word Surveillance System is proposed in [24], they extracts SIFT descriptor and clustered the features set using K-means algorithm and trained the network using SVM, finally SVM classifies positive images. There are some other works reported that used CNN architecture for gun detection. In [25] they used sliding window based techniques and their method is able to find only one type of gun in given images. In [22], they fine tuned the parameters of F-RCNN for a gun detection. The method fails when they tried with a huge variety different types of guns and more number of false positives. By experimenting faster RCNN along with minimisation of false negatives they achieve a fair accuracy. In [26] they also proposed OAOD algorithm based on the CNN architecture. Firstly, it predicts orientation of the object which is used to rotate the object proposal, than maximum area rectangles are cropped from the rotated object proposals which are again classified and localised finally. Their method achieves fair results.

Works discussed above are not able to handle all the challenges like, scaling, rotation, illumination change, viewing angle variations and occlusions by the gun's carrier and the surrounding people. Moreover, the existing object detectors process rectangular areas, though a thin and long rifle may actually cover only a small percentage of that area and the rest may contain irrelevant details suppressing the required object signatures. In this scope of the

Table 2 Recent state-of-the-art based on gun detection on visual images

Author	Algorithm used	Positive aspects	Data set used	Reported accuracy	Limitations
Grega *et al.* [13]	neural network, background subtraction	detect object with gun from video, then identify the gun using classification.	Tr Set: 4500, Ts Set: 1369	99.32%	no features included to deal with illumination, occlusion and rotation problems.
Tiwari and Verma [14]	SURF features, K-means clustering	immune to occlusion, rotation, scaling.	P Img: 65 N Img: 24	84.27%	works on images not on video frame, only gun is located not the object carrying it
Tiwari and Verma [5]	Harris point detector	rotation, scaling and shape invariant.	P Img: 65 N Img:24	88.67%	works on images not on video frame
Asnani *et al.* [17]	HOG features and classification	provide promising results.	–	NP	cannot deal with the shadow, illumination problem
Ardizzone *et al.* [23]	graph-based visual saliency	provided the information that gun is in the hand of human	2000	NP	used a small set of online available data set with medium quality
Halima and Hosam [24]	BoWSS, SIFT descriptors and SVM classifier	works well in cluttered scenes.	1000	F-measure: 0.35	will not work on real-time surveillance system and time consuming
Gelana and Yadav [25]	sliding window, CNN	classify each image into two classes, gun present and gun not present	P Img: 4000 N Img: 1869	97.78%	it can detect only one type of gun
Olmos *et al.* [22]	faster RCNN	design a data set and fine tuning of faster RCNN by minimising false positives.	Tr Set: 28635 Ts Set: 608	Pre:94.17%, Rec:31.91%	the results are database dependent and design of data set is also based on minimisation of false positives
Iqbal *et al.* [26]	faster RCNN	proposed orientation aware object detector for detection of gun	Tr Set: 8872 Ts. Set: 2101	Pre: 0.888	it has huge time complexity

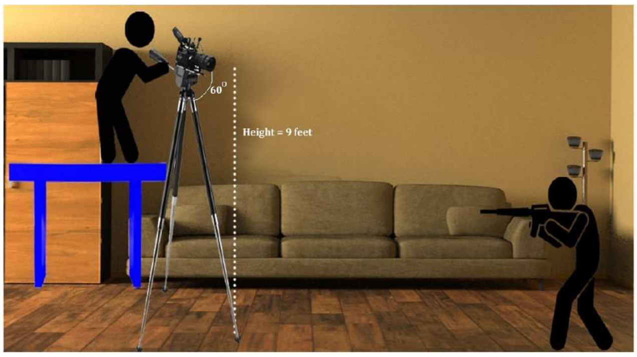
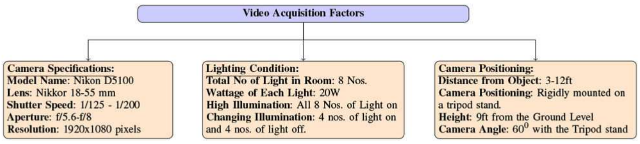


Fig. 2 *Demo of the camera setup of the created data set*


Video Acquisition Factors

Camera Specifications:
Model Name: Nikon D5100
Lens: Nikkor 18-55 mm
Shutter Speed: 1/125 - 1/200
Aperture: f/5.6-f/8
Resolution: 1920x1080 pixels

Lighting Condition:
Total No of Light in Room: 8 Nos.
Wattage of Each Light: 20W
High Illumination: All 8 Nos. of Light on
Changing Illumination: 4 nos. of light on and 4 nos. of light off.

Camera Positioning:
Distance from Object: 3-12ft
Camera Positioning: Rigidly mounted on a tripod stand.
Height: 9ft from the Ground Level
Camera Angle: 60^0 with the Tripod stand

Fig. 3 *Camera setup and acquisition factors of the created data set*

work, we try to compensate with these challenges. Moreover, methods discussed in Table 2 mostly detect gun in a still image, not on the whole video sequence.

If we consider detection of object(s) as well as carrying gun(s) in the whole video sequence, it will also improve the existing security system.

3 TUVD-CSA description

Literature work described that there are a para amount of work on the designing of real-time video data set for gun detection. In this work, we focused on the designing of data set for gun detection and incorporated different challenges that are related to the real-time application. Demo of the capturing setup is shown in Fig. 2.

Till date we have created and collected **150 video clips**. Among these **125 videos are created and collected in indoor/ outdoor** conditions and **25 video clips** are collected from few **movie**. These clips are collected to include some real-time scenario, which will make the data set as versatile as possible. Other 125 video clips are collected in parking places, building premises, corridors, garden, open fields, different cross-ings (3 way, 4 way), lobby, laboratory, classroom etc. of Tripura University. We ask a few students to annotate actions of a suspected person with guns in these areas. We also include challenges like, more number of false positives like object(s) carrying mobile, bottle, pen etc. Toy guns are used here in designing of this data set, as it is very difficult to have several real guns. Gun detection is based upon the shape of the guns, therefore, incorporation of real guns is not necessary. The videos are collected with 30 fps in this work. All the acquisition protocols like camera model, shutter speed, aperture, the resolution of collected videos etc. are presented in Fig. 3. The data set statistics is given in Table 3.

Table 3 Statistics of TUVD-CSA

		Total videos			
Image type	Video/ image format	Indoor condition	Outdoor condition	Movie clips	Total frames
visual	.MOV/.JPG	65	60	25	>500,000

3.1 Applications of TUVD-CSA

The proposed data set has two-fold applications; one is in the area of forensics, and another one is in security & surveillance. Forensics includes the analysis of the crime scene before crime, during crime and after crime [27]. Captured videos are divided into three categories, (a) before crime, (b) during crime and (c) after crime. These features are included to provide the data set a forensic insight. In addition, this data set can be used for **Human Activity Recognition, Human Behaviour Recognition, Re-identification of Suspect, Recognition of Different Hand Position etc**. Fig. 4 represents the data set with respect to the forensic analysis. Regarding the security and surveillance, an automated system is liable to raise the alarm or indication whenever any abnormal activity is encountered, due to which the operator will prioritise his attention on the video feed and will initiate appropriate action. Incorporate different real-time problems open up a different way in research work. One well established benchmark data set is used for validation of different methods under one parameter.

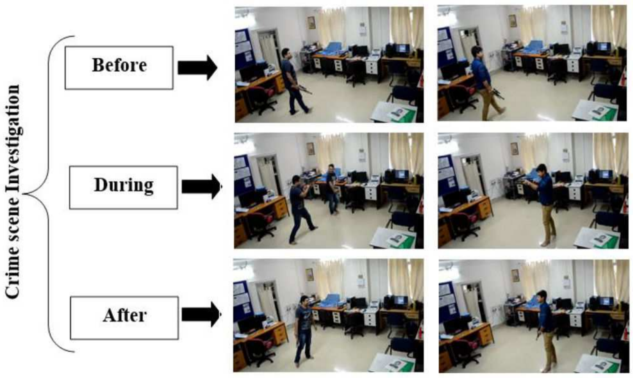


Fig. 4 *Samples of crime scene analysis*

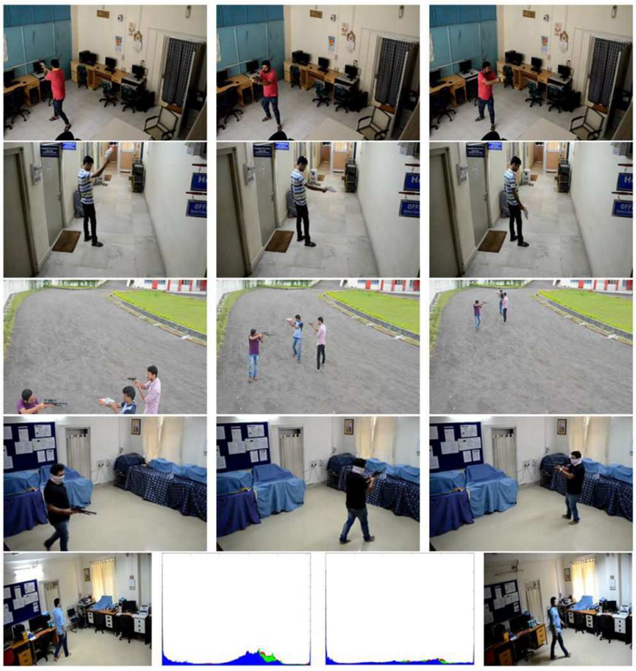


Fig. 5 *Different features of the TUVD-CSA like panning of gun, tilting of gun, scaling of guns, rotation of gun along with subject and also shows the subject wearing the mask shown in the first, second, third and the fourth rows, respectively, finally, the fifth row represents the illumination change condition and that has been examined by the histogram*

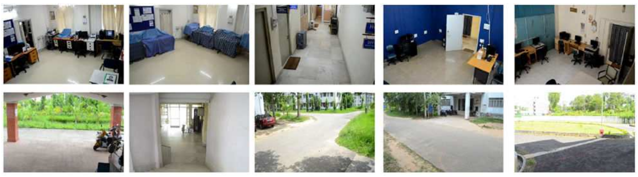


Fig. 6 *Few different backgrounds used in TUVD-CSA: The first and second rows represent the background of indoor and outdoor conditions, respectively*

Fig. 7 *Some example of mass firing scenarios*

3.2 Features of TUVD-CSA

The proposed data set able to serve the needs required to design a real-world system. Along with this, it will help in identifying the constraints that can be used to improve automatic gun detection systems. The contributory features of the data sets are as follows:

I. The primary feature of the data set is, it's the only video data set that contain moving object(s) with and without gun(s) in hand, collected in indoor and outdoor conditions.

II. False positive images like object(s) carrying mobile, bottle, pen etc. are also collected to enrich non-gun class.

III. Panning, tilting of guns with respect to the stationary objects are present.

IV. As the videos are captured from a certain distance and objects are moving horizontally and vertically hence, effect of scaling, rotation and occlusion are also present in the data set.

V. Changing illumination is a major challenge which is also incorporated using the lights in rooms. Guns and objects are affected by the illumination variation.

VI. Simple and complex backgrounds in indoor and outdoor conditions are considered in this data set.

VII. The data set have an application in forensics analysis. It incorporates frames of before, during and after crime are occurred.

VIII. Videos of people with mask on their face are also considered.

Few features of the database are shown graphically in Fig. 5. Here, Pan referred to the rotation of gun horizontally, and tilt of gun referred to the rotation of gun perpendicularly. When the gun is perpendicular to the object, we consider the gun is in 90°, whenever it goes down parallel to the object then gun is in 0° and when it goes up parallel to the object then, the gun is in 180°. Likewise tilt, pan of the guns are also considered from 0° to 180°. Along with these above mentioned features the data set also has following challenges:

3.2.1 Different complex background: We collect videos in a different background, as background considered as an important factor in video data set. In simple cases the gun is visible and can be easily identified, but in case of complex background the colour of gun same as the colour of background or background objects. Therefore, identifying guns in that scenario will be bit challenging task. Including these types of real life challenges also increase the complexity of the data set. Fig. 6 shows some examples of different background that are used in the data set.

3.2.2 Mass firing: In TUVD-CSA, we also included some scenario of mass firing. In recent times, mass firing is an issue regarding the security. There are several cases of mass firing reported in recent years worldwide. Considering these things in mind, we have included some mass firing scenarios, in which a group of people are firing with the gun. Identifying all the persons and all the guns automatically will really be a very challenging task, but have a great importance. Having these real life challenging scenarios are also labelling up the complexity of the data set. Fig. 7 shows some mass firing scenarios.

3.2.3 Different types of pistol and rifle: In TUVD-CSA, we have used different types of toy guns. In real world, there are numerous number of pistols and rifles are available, but it is very difficult to have all kinds of real guns. In this database, we have used some toy guns, which are replicas or imitation of real guns. These type of features of this data set are very much needful for the research community and also label up the usefulness of this created data set. We also include challenges like more number of false positives, objects carrying mobile, water bottle, pen etc., to enrich the non-gun class. Fig. 8 shows some examples of toy guns used in TUVD-CSA and also some samples of non-gun class. The statistics of the features of the data set are briefly noted in Table 4.

3.3 Naming convention

Naming of TUVD-CSA has been done in understanding the category of the data set during analysis. To make naming convention meaningful different codes have been used for different conditions of videos, like indoor/outdoor, whether there is any sudden change of illumination condition, firearm present or not. Likewise, different codes have been used for multiple firearms, the presence or the absence of objects with mask on their face, panning, tilting, scaling, rotation of firearm along with subject and different mass firing scenario.

Using the above codes, every image in the database acquires a distinctive identity. All the assigned codes for different parameters are given in Table 5. Based on that codes the image names **I_V001_IC_G(M)_M_P-T-R-S_MF_0001.jpg** indicate that the video has been captured in indoor condition with video ID 001, there is an effect of sudden change of illumination, multiple types of guns has been used, subject with mask on their face, panning, tilting, scaling of guns and rotation of guns along with subject is presented and mass firing scenarios are also presented in this video and 0001.jpg means that is the frame number of this video. If any of the feature is not present in any video then it's corresponding code will not be mentioned in that particular case.

Fig. 8 *Few example of Toy Guns and non-gun class that are used in proposed database*
(a) Toy guns, ***(b)*** Non-gun glass

Table 4 Statistics of different features of TUVD-CSA

Variations	No. of frames (approx.)	With guns	Illumination change	Multiple object
pan	70,000	P	P	NP
tilt	68,500	P	P	NP
rotation	69,600	P	P	P
scaling	78,500	P	P	P
occlusion	36,050	P	P	P
guns(M)	2,20,500	P	P	P
mask	57,800	P	P	P
crime scene	51,700	P	P	P
mass firing	80,650	P	NP	P

Table 5 Codes used for naming TUVD-CSA

Feature	Code	Feature	Code	Feature	Code
Indoor	I	Outdoor	O	Illumination change	IC
Video ID	V	Gun	G	Multiple gun	G(M)
Pan	P	Tilt	T	Mass firing	MF
Mask	M	Rotation	R	Scaling	S

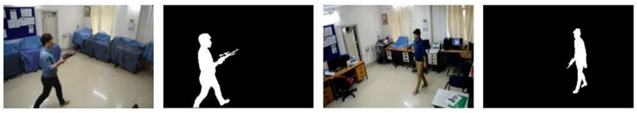


Fig. 9 *Samples of few GT along with original frame*

3.4 GT generation of TUVD-CSA

GTs are required for evaluation of methods performed on the data set. We performed three label annotations, in the first label moving objects are annotated, in second label shadows are annotated, in the third label, we annotate the gun. The GT annotations have the following attribute [28]: **Bounding Box**: An axis-aligned bounding box surrounding the extent of the object visible in the image. **Class**: we include three classes, such as moving object, shadows and gun.

To maintain the consistency in the annotation of GT, we rely on a single annotation party. The GT annotation structure is based on the two-label tree structure. At the root label, an annotation party comprises a certain number of annotators. They provided a list of guidelines, which consist what to annotate, how to annotate, what are the labels, how to label and more importantly how to handle occlusions and rotations. During the annotations, the annotators are observed periodically to ensure that they follow the given guidelines. They are trained and co-located to maintain consistency in GT annotation. After root label annotations, the parent label annotators have verified the annotated GTs. They check whether any false annotations are there or not. They also ensure the occlusions are correctly handled or not, and all objects are annotated or not.The GT annotation employed in this scope is consistent, correct (few annotation errors) and exhaustive.

TSLAB [29] annotation tool is used to verify the proposed annotation framework. It is widely used for object annotation. TSLAB software is believed to handle occlusion and other challenges correctly. The data set is annotated with the TSLAB tool and compared the annotated data produced by both methods. Similarity score of 95% is enough to ensure the effectiveness of this annotation framework used in this work. The annotation framework is more effective as we overcome the disadvantage of the bounding box by employing the accuracy of the human brain. Fig. 9 shows few samples of GT along with the original frame.

3.5 Feature-based comparison with existing datasets

There exist only one video data set for gun detection [13], which is collected in a very restricted environment, so, it lacks different features required to be established as a benchmark data set. Reported literature regarding detection of gun used image data set to validate their method for detection or classification. These data sets are designed by collecting images from IMFDB as per their requirement. Although, the data set have different types of real guns, different backgrounds and a variety of objects and environment, the data set cannot provide challenges that applicable for real-time application as it is an image data set. Moreover, these data sets does not have the implications in forensic application. A detailed feature-based comparison is shown in Table 6.

4 Methodology

The proposed methodology is based on the machine learning idea of template matching. During the training phase, the machine had learned to correctly match template in the input image. The authors in [5, 13, 30] proposed the template matching-based procedure for handheld gun detection. In 2009, Christos Grecos *et al.* proposed a gun detection method to overcome the time complexity of template matching [30]. Firstly, they performed background subtraction to detect objects in the incoming frame with shadow removal and

Table 6 Feature-based comparison of existing gun detection data sets with the proposed TUVD-CSA

DB name	Type of DB	Sudden Il change	Mult type of G	Mult type of BG	Occlusion of G	Mult object	Mass firing	Panning of G	Tilting of G	Scaling of G	Forensic App	Real-time App
gun movie data set	video	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
IMFDB	image	✗	✓	✓	✗	✓	NP	✗	✗	✓	✗	✓
data set of R. Olmos	image	✗	✓	✓	✗	✓	✗	NP	NP	NP	✗	✓
TUVD-CSA	video	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

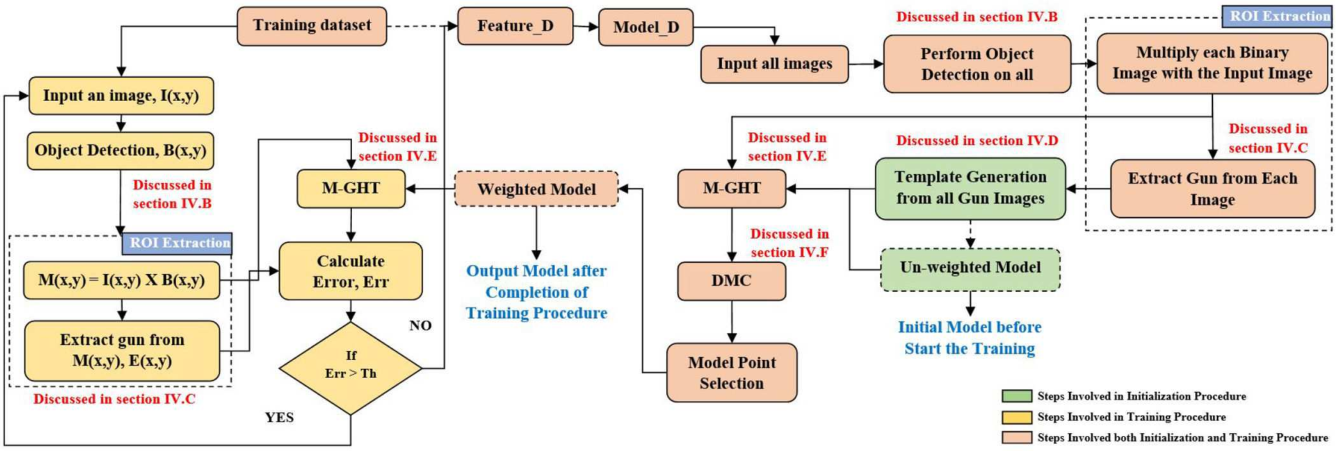

Training dataset
Feature_D
Model_D
Input all images
Discussed in section IV.B
Perform Object Detection on all
ROI Extraction
Multiply each Binary Image with the Input Image
Input an image, I(x,y)
Object Detection, B(x,y)
Discussed in section IV.B
ROI Extraction
M(x,y) = I(x,y) X B(x,y)
Extract gun from M(x,y), E(x,y)
Discussed in section IV.C
Discussed in section IV.E
M-GHT
Calculate Error, Err
If Err > Th
NO
YES
Weighted Model
Output Model after Completion of Training Procedure
Model Point Selection
Discussed in section IV.E
M-GHT
Discussed in section IV.F
DMC
Discussed in section IV.D
Template Generation from all Gun Images
Un-weighted Model
Initial Model before Start the Training
Discussed in section IV.C
Extract Gun from Each Image
🟩 Steps Involved in Initialization Procedure
🟨 Steps Involved in Training Procedure
🟧 Steps Involved both Initialization and Training Procedure

Fig. 10 *Training procedure of the proposed method. The diagram represent one iteration of one training step along with the initialisation step. The colours are represented which steps are followed in initialisation, in training and in both initialisation and training. Initialisation step is used to design the first unweighted model. Afterwards, training procedure has started, which will be continued until localisation error obtained for each image is lower than a pre-determined threshold*

afterwards they used morphological operation to obtain the final region of interest (ROI). They stored SIFT feature of 100 gun templates and during detection SIFT feature of the ROI is compared to the stored features to confirm the presence of the gun in the frame. This work is effective in deduction of time and space complexity of template matching algorithm. The proposed method motivated from [30] and employed an iterative training procedure for generation of a discriminative and general gun template. In this work, we compared existing state-of-the-art background subtraction-based object detection algorithm to indicate the method that performed well on the data set. After obtaining a binary image from the object detection method, we multiplied the binary image with it's corresponding original frame, which specifically provide us the colour and texture information of the person with gun, thereafter template matching is performed on that image. The template should be more general, that can detect the object in any input image. Therefore, an iterative training procedure is used that periodically update the initial template. Furthermore, M-GHT-based template matching has been performed here. Instead of a shape model, the proposed work is based on model points. Here, model points referred to the edge points that further referred to each pixel presents in the given edge. To reduce the number of model points a thinning procedure is applied. In the thinning procedure, a weight has imposed to each model points. The weights of model points defined the importance of the model point, and a final model with optimal weights are used for detection of template. Discriminative model combination (DMC) has been used to find the optimal weights. The weights are learned through a training procedure. The training procedure is represented diagrammatically in Fig. 10. Different steps of the proposed framework are elaborately described in the following subsections.

4.1 Data preparation

As proposed method is based on the machine learning procedure, therefore, it is obvious that data set is to be divided into two major subsets, such as **training** data set and **testing** data set. Here, the training data set should contain representative characteristics of the gun, as the aim of training is to produce a general model of gun. In this work, the training set is further divided into **featured** data set (*Feature_D*) and **model** data set (*Model_D*). Initially, three frames from the training set is used to initialise the *Feature_D*. The *Model_D* is initialised by one randomly selected frame from the *Feature_D*. More specifically, *Model_D* is a subset of *Feature_D*. To avoid the over-fitting, we are not including all frames of *Feature_D* to *Model_D*. *Feature_D* and *Model_D*, are updated periodically during training. During training, after adding of three images in *Feature_D*, update *Model_D* with one among these three. After completion of the training phase, $Feature_D$ will represent a sub data set that reflect the major features of the data set. In addition, $Model_D$ will stored the frames which have been used to update the model at the time of training.

4.2 Object detection

After initialisation of the data sets, the next step of the proposed framework is object detection. The primary concern of this step is to reduce the search space for matching of the template. In addition, background subtraction eliminates unwanted background information and also ensures the detection of multiple guns by identifying multiple moving objects. Here, we used existing state-of-the-art object detection algorithm for this purpose. There are different background subtraction methods present in the literature, that are able to handle shadow, cluttered background, sudden illumination change etc. Illumination variation increases difficulty in gun detection, as guns are prone to change its colour and texture due to changes in the illumination. The proposed data set includes videos in outdoor conditions too. To compensate with these mentioned problems, there are different algorithms in the literature that are claimed to handle these challenges properly, such as multiple temporal difference (MTD) [10], Gaussian mixture model (GMM) [9], ViBe [11], illumination sensitive background subtraction (ISBS) [12] 3dSOBS+ [31], PAWCS [32] and O-SSR [33]. The comparison between the existing methods is shown in the result section. Mathematically, we can consider an image frame $I_{(x,y)} \in Tr$, where Tr define the training data set and (x,y) defines a two-dimensional image with the $x - y$ co-ordinate. $I_{(x,y)}$ is a RGB colour image.

Therefore, after background subtraction we will obtain a binary image $B_{(x,y)}$, where the foreground objects represented by 1 and the background is represented by 0. The binary image $B_{(x,y)}$ is passed on to the next step of the system.

4.3 Extraction of ROI

The binary image $B_{(x,y)}$ obtained from the background subtraction is further processed to eliminate unwanted small objects. Morphological erosion and dilation have been carried out to eliminate small objects and closed the shape of an object. Afterwards, the obtained binary image $B_{(x,y)}$ is multiplied with the original image $I_{(x,y)}$ to obtain the texture and colour information of the foreground object. Suppose, $O_{x,y}$ is an output image after the multiplication operation. During the training GT information for $O_{x,y}$, i.e. the bounding box information is used. Using the corresponding bounding box information, the gun has been extracted from the image $O_{x,y}$ is, $G_{i,j}$, where $i \neq x$ and $j \neq y$. Therefore, guns image $G_{i,j}$, required to generate the model is obtained in this step. More specifically, edge points extracted from the $G_{i,j}$, is the model $T_{i,j}$ for this framework. More than one ROI can be extracted depend on the number of detected moving objects.

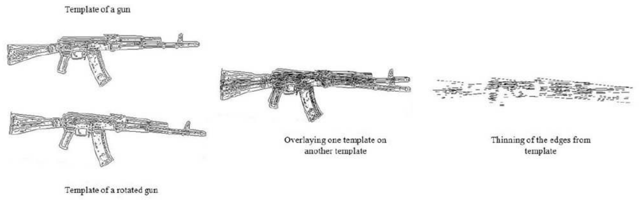

Template of a gun
Overlaying one template on another template
Thinning of the edges from template
Template of a rotated gun

Fig. 11 *Template generation procedure*

Table 7 Illustration of the R-table with bin size $\triangle \phi$

i	φ_i	PSF_{φ_i}	$D1_{\varphi_i}$	$D2_{\varphi_i}$
0:	0	$\mathrm{PSF}\,\vert\,\varphi(x) = 0$	$D1\,\vert\,\varphi(x) = 0$	$D2\,\vert\,\varphi(x) = 0$
1:	$\Delta\varphi_i$	$\mathrm{PSF}\,\vert\,\varphi(x) = \Delta\varphi_i$	$D1\,\vert\,\varphi(x) = \Delta\varphi_i$	$D2\,\vert\,\varphi(x) = \Delta\varphi_i$
2:	$2\Delta\varphi_i$	$\mathrm{PSF}\,\vert\,\varphi(x) = 2\Delta\varphi_i$	$D1\,\vert\,\varphi(x) = 2\Delta\varphi_i$	$D2\,\vert\,\varphi(x) = 2\Delta\varphi_i$
⋮	⋮	⋮	⋮	⋮

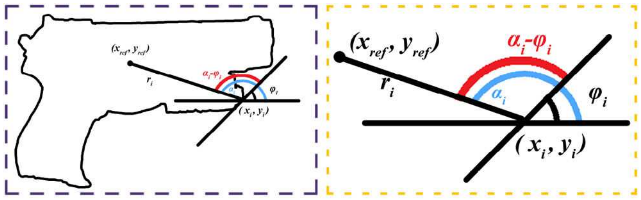


Fig. 12 *Parameters of GHT*

All the ROIs are considered during the template generation. The manual annotation of the bounding boxes is a tedious job, though it is a necessary step of any supervised detection algorithm.

4.4 Template generation

Concerning the generation of optimal gun shape templates we rely on the edge points rather than on the segmentation of the gun from the images. We have already discussed about the model data set, *Model_D*, a subset of training data set. In *Model_D*, the frames that are used to generate and update the template have been stored eventually. Initially, the *Model_D* is populated with one training image and from that image an initial meaningful model has been generated. Template generation will perform on an ROI image ($G_{i,j}$) extracted in the previous step (refer to Section 4.3). The initial template, $T_{i,j}$ is created by sampling the edge points from a gun image $G_{i,j}$. After the initial iteration, fusion of templates has been performed for updation of template $T_{i,j}$. After each iteration, an automatic checking procedure will be employed to assess the accuracy of the template for each training image, if exist template fails in some training images, these images are included in *Feature_D*. After inclusion of three images, one among these three images (randomly selected) is included in the *Model_D*. When there is any new entry in *Model_D*, then a new model points have been generated. Model point generation using more than one image has been shown in Fig. 11. In the case of fusion of two models, edge points from each image are fused, more specifically imposed on one another. Thus the resultant template will consist of edge points required to identify different size of the gun that are present in the training images. Afterwards to reduce the size of the model by a thinning procedure. The thin template has been used for the next iteration of the training procedure. The same procedure has been followed for the next training iterations. Templates created from different images are fused to captured variability in the template.

In the last step of model, generation procedure is thinning to reduce the model size by discarding the points which are not carries exclusive shape information, as shown in Fig. 11. The discriminating weights are estimated for model points. Unnecessary model points, which do not aid in the localisation procedure can be identified and excluded based on their weights. Therefore, the larger size of ROI does not do any harm. The considered data set possesses several problems in detection of guns, such as number of false positives. Objects similar to the gun present in the background are referred as false positives. Therefore, including the surroundings of the gun can prevent the confusion of the gun with objects of a similar shape. Nevertheless, taking a large size of surroundings increased the complexity of the method, therefore, a large ROI is not advisable.

The Canny edge detector is used for generation of the edge points. Afterwards, thinning of the edge points is implemented to eliminate the similar points, which is referred to the points having the similar values in the Hough space. Therefore, after each training a model point selection is performed to obtained a generalised template for the gun. The weighting procedure is described in Section-4.6.

4.5 Proposed M-GHT

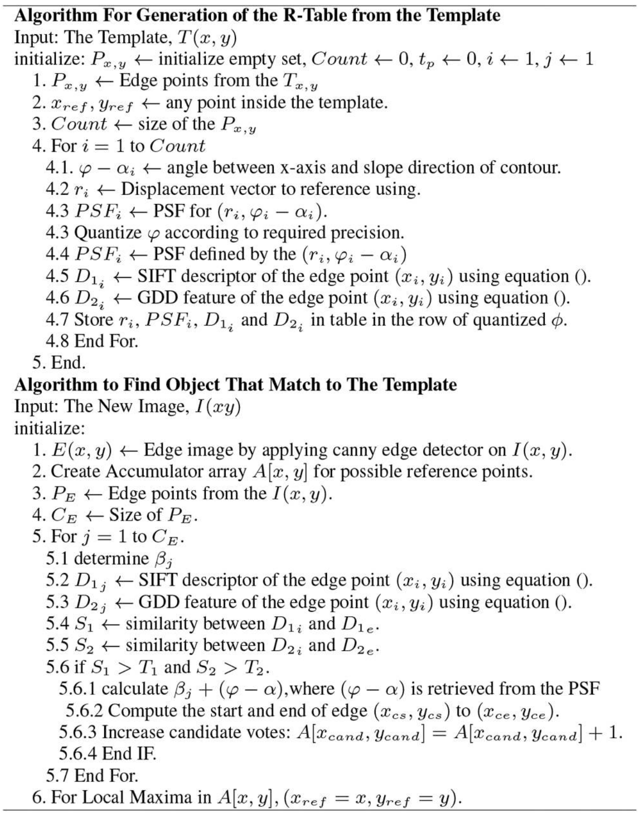
Generalised Hough transform [34], is a method generally deployed for localisation of a target shape in the image using a template, T. The template T mostly represents the edge map of the target shape. GHT finds the highest correspondence of the template points in the image using a voting policy. Specifically, GHT maps the Image I into a voting space using the template T. The voting space is known as a Hough space. A reference point is considered in the model, $(x_{\mathrm{ref}}, y_{ref})$ such that $(x_{\mathrm{ref}}, y_{\mathrm{ref}}) \in T$, usually centre point in the model. A line is drawn from this reference point, usually the object centroid, to the boundary of the object. At each edge points (x_{ti}, y_{tj}), such that $(x_{ti}, y_{tj}) \in T$, the displacement vectors in relation to a set of axes is placed in a look-up table called an R-table. The displacement vectors for each edge pixel are defined in relation to a set of axes that are rotated in such a way that the vertical axis is aligned with the gradient direction at that pixel. Instead of constructing Table 7 relating a single value of the displacement vector $\boldsymbol{R}$ to each value of α, a set of vectors $\boldsymbol{R}_j$, which describe the shape of the point spread function (PSF) is entered. The PSF is defined by the template image using the magnitude of its position vector $\boldsymbol{r}_i$ and the relative angle between its gradient and displacement angle $(\alpha_i - \varphi_i)$ as illustrated in Fig. 12.

For the test objects, the gradient angle β and thus the index of Table 7 are obtained for each edge point (u_i, v_j). Next, the respective entries in Table 7 are used to compute possible locations of reference points in the parameter space using the following equation. The value of $(\alpha - \varphi)$ is retrieved from the PSF.

$$
\begin{aligned}
(x_{tj}, y_{tj}) = u_j + r\boldsymbol{R}\cos[\beta_j + (\alpha - \varphi)],\\
v_j + r\boldsymbol{R}\sin[\beta_j + (\alpha - \varphi)]
\end{aligned}
\tag{1}
$$

Whereas (x_{tj}, y_{tj}) is the probable position of the reference point and $\boldsymbol{R}$ is the range of scale within which the test image is to be detected. The $\boldsymbol{R}$ is used here to handling the scale variation. In this GHT, instead of accumulating points in the parameter space, lines corresponding to a range of defined scale are accumulated. The peak in the accumulator array is the point where most of the lines intersect.

The aim of this work is correct detection of gun, therefore, we are adding two parameters that increase the dimension of the parameter space. The two parameters are gradient distance descriptor (GDD) and SIFT descriptors represented by $D1$ and $D2$, respectively. $D1$ and $D2$ for each edge point are also calculated and are stored in Table 7 against each index β along with the PSF. The use of these two descriptors are two-folds. Firstly, these two descriptors are scale invariant as well as they also invariant to illumination. Secondly, GDD and SIFT descriptor handle the limitations of PSF. The problem with the PSF is that different shapes can have the same PSF and one PSF can show peaks for two different shapes. During the generation of Table 7, instead of storing the PSF, we are storing PSF along with the GDD and SIFT descriptors for the edge points. Therefore, at the time of test not only the PSF is compared but both the descriptors are compared. Similarity measures are used for this comparison. Therefore, the proposed transformation is not dependent solely on the PSF and able to obtained more accurate possible position of gun in the test image. We proposed, the **M-GHT**, which also uses displacement



```
Algorithm For Generation of the R-Table from the Template
Input: The Template, $T(x, y)$
initialize: $P_{x,y} \leftarrow$ initialize empty set, $Count \leftarrow 0, t_p \leftarrow 0, i \leftarrow 1, j \leftarrow 1$
  1. $P_{x,y} \leftarrow$ Edge points from the $T_{x,y}$
  2. $x_{ref}, y_{ref} \leftarrow$ any point inside the template.
  3. $Count \leftarrow$ size of the $P_{x,y}$
  4. For $i = 1$ to $Count$
    4.1. $\varphi - \alpha_i \leftarrow$ angle between x-axis and slope direction of contour.
    4.2 $r_i \leftarrow$ Displacement vector to reference using.
    4.3 $PSF_i \leftarrow$ PSF for $(r_i, \varphi_i - \alpha_i)$.
    4.3 Quantize $\varphi$ according to required precision.
    4.4 $PSF_i \leftarrow$ PSF defined by the $(r_i, \varphi_i - \alpha_i)$
    4.5 $D_{1_i} \leftarrow$ SIFT descriptor of the edge point $(x_i, y_i)$ using equation ().
    4.6 $D_{2_i} \leftarrow$ GDD feature of the edge point $(x_i, y_i)$ using equation ().
    4.7 Store $r_i$, $PSF_i$, $D_{1_i}$ and $D_{2_i}$ in table in the row of quantized $\phi$.
    4.8 End For.
  5. End.
Algorithm to Find Object That Match to The Template
Input: The New Image, $I(xy)$
initialize:
  1. $E(x, y) \leftarrow$ Edge image by applying canny edge detector on $I(x, y)$.
  2. Create Accumulator array $A[x, y]$ for possible reference points.
  3. $P_E \leftarrow$ Edge points from the $I(x, y)$.
  4. $C_E \leftarrow$ Size of $P_E$.
  5. For $j = 1$ to $C_E$.
    5.1 determine $\beta_j$
    5.2 $D_{1_j} \leftarrow$ SIFT descriptor of the edge point $(x_i, y_i)$ using equation ().
    5.3 $D_{2_j} \leftarrow$ GDD feature of the edge point $(x_i, y_i)$ using equation ().
    5.4 $S_1 \leftarrow$ similarity between $D_{1_i}$ and $D_{1_e}$.
    5.5 $S_2 \leftarrow$ similarity between $D_{2_i}$ and $D_{2_e}$.
    5.6 if $S_1 > T_1$ and $S_2 > T_2$.
      5.6.1 calculate $\beta_j + (\varphi - \alpha)$,where $(\varphi - \alpha)$ is retrieved from the PSF
       5.6.2 Compute the start and end of edge $(x_{cs}, y_{cs})$ to $(x_{ce}, y_{ce})$.
      5.6.3 Increase candidate votes: $A[x_{cand}, y_{cand}] = A[x_{cand}, y_{cand}] + 1$.
      5.6.4 End IF.
    5.7 End For.
  6. For Local Maxima in $A[x, y]$, $(x_{ref} = x, y_{ref} = y)$.
```

Fig. 13 *Algorithm 1: Algorithm for Hough-based transformation and matching*

vectors but includes scale invariant features. These feature added dimension to the parameter space. SIFT and GDD operators are well-known scale invariant descriptors. For each edge pixel, a fixed sized window is defined. Based on the pixel of that windows GDD and SIFT are calculated for that edge pixel. The algorithm to find GDD and SIFT descriptors can be find in this literature's [35, 36]. The comparison between the descriptors are carried out matrix cosine similarity measures. During detection process difference between the similarity measure value is compared with the predefined threshold values T_1 and T_2. If the difference is higher than the threshold value then the pixels are considered as similar. 0.85 is considered as value of T_1 and T_2 in implementation.

The proposed M-GHT confirmed a fair accuracy in detection of firearm. Mathematically, GHT able to detect any analytic curve in an image. Therefore, GHT able to detect the edges of gun by using an edge template of the gun. GHT inherently can handle the rotation and scaling to a some extent. In conventional GHT, 4D parameter space is used to locate the reference point in the input image. That increase the computational complexity and memory requirements. The computational complexity of conventional GHT can be represented by $(n_p/R_q)n_tS_q\theta_q$ and memory requirement is $N^2S_q\theta_q$. Where, n_p is the number of edge pixels in the prototype object, R_q is the resolution of the R-table index, n_t is the number of edge pixels in the test image, S_q is the resolution of the scale parameter, θ_q is the resolution of the rotation parameter. And N^2 is the size of the accumulator array. The proposed M-GHT has less computational complexity than the conventional GHT by eliminating one degree of freedom in the parameter space. As the proposed M-GHT establishes a direct relationship between the displacement vectors and the edge gradient direction by considering $\alpha_i - \varphi_i$. Therefore, the memory requirements are just N^2, the size of the accumulator array. While its computational complexity is $n_tn_pk\bar{r}(S_u - S_l)$. $(S_u - S_l)r$ is the scale parameter, along which accumulation in the parameter space occurs for each edge pixel. Where S_u and S_l are the upper and lower limits of the scale, respectively, $\bar{r}$ defines the average magnitude of the position vector. The factor $k(\leq 1)$ is included because the line incrementation procedure is limited by the size of the parameter space. Both the algorithm for Hough R-table generation and matching has shown in Algorithm 1 (see Fig. 13).

4.6 Weight learning of model points using DMC

For using the weighted model, a point model has been generated and weights are imposed to each point of the model. Ideally, we would like to train our model parameters such that the end-to-end performance in template matching is optimal. We use basic log-likelihood optimising algorithm to optimise model weights to obtains final weights of the model.

Let us assume, that we are given an image, *I*, which is to be translated into a Hough space *H*. *H* is nothing but are the cells consisting votes of that location, to have the target object. Among all possible target cells, we will choose the cell with the highest votes. The Hough space can be represented by a probability mass function. Suppose, each Hough cell, h_i consist V_i number of votes and total number of votes is V. Now the posterior probability for each possible object location (represented by the Hough cells) given an image *I* can be written as follows:

$$\Pr(h_i \Big| I) = \frac{V_i}{V} \tag{2}$$

Therefore, in this case, instead of searching the cell with the highest vote, Hough cell with the highest posterior probability will be the result of the GHT. From (2), we can determine model point dependent posterior probability for each model point m_j as follows:

$$\Pr_j(h_i \Big| I) = \frac{V_{ij}}{V_j} \tag{3}$$

The log-linear model can directly used to model the given posterior probability (3) as $\Pr^{(}h_i \big| I)$. Where ∧ is the parameter value or weights of each model point. In this framework, the training problem amounts to obtaining suitable parameter values ∧. A standard criterion for log-linear models is the MMI (maximum mutual information) criterion, which can be derived from the maximum entropy principle, as follows:

$$\mathrm{pr}^{(}h_i \Big| I) = \frac{\exp(\sum_j \gamma_j \,.\, \log\mathrm{pr}_j(h_i \big| I))}{\sum_k \exp(\sum_j (\gamma_j \,.\, \log\mathrm{pr}_j(h_k \big| I)))} \tag{4}$$

The coefficients $\wedge = \gamma_{jj}$ regulate the posterior probability of each model point, i.e. $\mathrm{pr}_j(h_i \big| I)$ on the determining equation $\mathrm{pr} \wedge (h_i | I)$. In this framework, γ_j represents the importance of the model point m_j and used as a weights of model point.

There are different methods to find the parameter ∧, which aim to minimise the empirical localisation error rate on given training data. From (4), an empirical error rate *E* can be defined as follows:

$$D(\wedge) = \sum_n \sum_i L(h_n, h_i) \frac{\Pr \wedge (h_i | I_n)_\eta}{\sum_k \Pr \wedge (h_k | I_n)\eta} \tag{5}$$

Since (5) describes the model recognition, it can be argued reasonable to compute values γ_j which minimises the Levenshtein distance between the correct object position h_n and Hough cell h_i. The error measure is weighted with an indicator function comprised of the posterior probabilities. The exponent η in the indicator function regulates the influence of the rivaling hypotheses h_k on the distance error measure.

Gradient decent algorithm is deployed here for optimisation of the error function $D(\wedge)$. Find an optimal set of ∧ is a difficult task as optimisation cannot be guaranteed to provide optimal solution for a given problem. In spite of that weighting of the template points using a sub-optimal ∧ also improves localisation accuracy and robustness. Therefore, the cell values are not only incremented but 1 but by the weight of the template point. It can also be shown that the direct usage of the trained template point weights in the GHT is valid since the position of the maximum h_n is not affected.

DMC generates negative weights to repel the template from competing objects that resemble the target and might lead to false-positive localisation results. The training stops if an error of less than 5 Hough cells is obtained on all images of the training data set or if no further improvement is achieved.

4.7 Training procedure

The training procedure is for the generation of a generalised template for detection of gun in video frames. The training procedure consists of two phases one phase is the generation of the template and another phase is the checking phase. In the checking phase, the current template has been applied on each training image, and the accuracy of the template for each image has been noted. Three images, on which accuracy lower than a pre-defined threshold is obtained are stored in the data set $Feature_D$. Afterwards, randomly one image among the new three images is chosen for the creation of the template and stored in the data set $Model_D$. The images of $Model_D$ are used to generate an updated template and using DCM, weights of each template point are decided. Hence, a weighted template is generated and then the updated template applied on each of the training images. The training procedure goes on likewise till the updated template obtains fair accuracy on each of the training images. Fair accuracy can be defined by maintaining a pre-defined threshold. By using a high threshold value more generalised template can be obtained. The accuracy is determined by IOU (intersection over union) parameter. IOU can easily be realised by comparing the output with the corresponding ground truth using the following formula:

$$\text{IOU} = \frac{\text{area of overlap}}{\text{area of union}}, \tag{6}$$

where area of overlap defines the overlapping area or common area between the predicted bounding box and the GT bounding box and area of union is the area encompassed by both the predicted bounding box and the GT bounding box. The evaluated value of the IOU determines the localisation error and the accuracy of the detection through the template. Higher the value of IOU means high 'detection accuracy'. By setting an acceptable threshold we automatise the accuracy checking procedure for each training image. Higher threshold value will increase the accuracy of the algorithm, but it takes time to converge. By considering each factor, we set the threshold value as 0.7.

After the initialisation or updation of the template, using the template we try to detect gun from each training image and calculate IOU for each training image. The procedure of detection using template matching has already been discussed in Section 4.5. Training images where we observe low IOU are included in the data set, $Feature_D$. After addition of three images in the $Feature_D$ data set, one image among the three is added to the data set namely $Model_D$. Afterwards, the template has updated or generated by the images present in the data set $Model_D$. The procedure of the template generation has already been discussed in Section 4.4.

Table 8 Quantitative measurement of the existing state-of-the-art detection of object with gun

Gun movie data set							
Method	Rec	Sp	FPR	FNR	PWC	Prec	F1
GMM	0.7599	0.9769	0.0231	0.2401	5.0607	0.8127	0.7597
MTD	0.8089	0.9559	0.0441	0.1911	1.0858	0.5981	0.6637
3dSOBS+	0.8362	0.9921	0.0079	0.1638	1.6947	0.8868	0.8531
ViBe	0.8400	0.9988	0.0012	0.1600	0.9779	0.9681	0.8938
PAWCS	**0.9876**	0.9914	0.0086	**0.0124**	1.2047	0.9667	0.9777
O-SSR	0.9537	0.9859	0.0141	0.0463	1.1939	0.9421	0.9512
ISBS	0.9852	**0.9973**	**0.0027**	0.8148	**0.8277**	**0.9894**	**0.9877**
TUVD-CSA							
GMM	0.8388	0.9766	0.0234	0.1612	4.7901	0.8127	0.7788
MTD	0.8615	0.9732	0.0268	0.1385	3.2555	0.8103	0.8215
3dSOBS+	0.8435	0.9820	0.0180	0.1565	2.9015	0.8304	0.8255
ViBe	0.9037	0.9948	0.0052	0.0963	1.2723	0.9477	0.9213
PAWCS	0.9505	0.9912	0.0088	0.0495	1.1477	0.9324	0.9393
O-SSR	0.9353	0.9888	0.0112	0.0647	1.5127	0.9143	0.9219
ISBS	**0.9906**	**0.9982**	**0.0018**	**0.0094**	**0.2544**	**0.9806**	**0.9855**

Bold values represent the significant results of different object detection methods in two datasets.

5 Experimental results and discussion

This section is divided into two parts. In first, we seperately compared the performance of existing background subtraction-based moving object detection techniques on the newly proposed data set TUVD-CSA and only publicly available video data set for gun detection, Gun Movies Data set [13]. These state-of-the-art object detection methods are GMM [9], MTD [10], ViBE [11], ISBS [12], 3dSOBS+ [31], PAWCS [32], and O-SSR [33]. In second, we compared proposed method with existing state-of-the-art gun detection methods for TUVD-CSA data set and also we compared proposed method with other detection methods for publicly available data sets. Noted that 60% data of the data set is used for training purpose and remaining for testing.

5.1 Object detection

To demonstrate the existing background subtraction-based object detection results using TUVD-CSA data set and gun Movies data set [13], we present a performance evaluation in Table 8 through seven quantitative measurement matrices. The usual way of evaluating the performance of background subtraction algorithms for moving object detection in videos is to pixel-wise compared the computed foreground masks with the corresponding GT foreground masks and compute suitable metrics. We used different accuracy markers for noted the results; recall (Rec), precision (Prec), percentage of wrong classification (PWC), specificity (SP), false positive rate (FPR), false negative rate (FNR), and F-measure (F1) for comparison of the seven existing background subtraction methods of object detection [37]. Note that all metric-attained values range from 0 to 1, with higher values representing better accuracy and in some cases the lower value representing the better accuracy.

With the overall results of Table 8 shows ISBS method outperforms other state-of-the-art object detection methods, recall, precision is high compared to other methods, specificity is also reasonable for the data set, FPR and FNR are maintaining the essential ratio. We evaluated the performance in overcoming **sudden change of illumination** effect presented in TUVD-CSA. Compared with other approaches, the ISBS approach is more adaptable to the sudden illumination change because of its determination of light and dark background candidates, thus leading to the superior detection results.

Videos of Gun Movies Data set [13] is collected in a very restricted environment. The challenges like complex background, illumination change etc. are not present in this data set. The overall result is shown in Table 8, there is not much difference in the values of recall, precision and specificity.

In addition to quantitative evaluation, qualitative results of state-of-the-art methods are also shown in Fig. 14. For TUVD-CSA rather than ISBS, other approaches fail to properly segment salient object areas. Their PWC is also very low for this data set. For gun movies data set, only GMM method partially fails to segment the salient object area, but other approaches showing substantial performance over its composition.

5.2 Gun detection

We evaluate the proposed gun detection method on the proposed database. Also reported the results of the proposed method on the other data sets too. The proposed method is devised for video sequences as detection of a moving object is the first task to be performed. Except the gun movies data set, other available data set are image data set, therefore, only the proposed gun detection method is implemented on these data sets. State-of-the-art methods

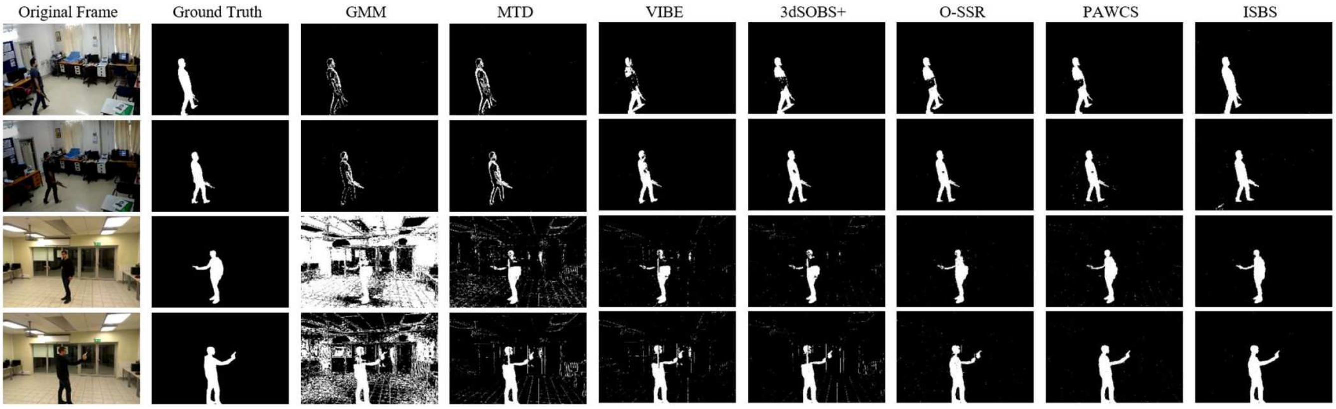


Fig. 14 *First 2 row and last 2 two row represents TUVD-CSA and gun movies data set video sequence, respectively. The first and second column represent original frames and its corresponding GTs, columns third to ninth represent the binary mask of moving objects generated by GMM, MTD, ViBe, 3dSOBS+, O-SSR, PAWCS, and ISBS approaches*

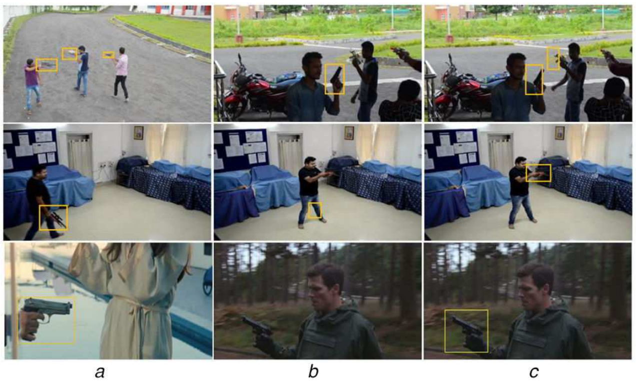


Fig. 15 *Some sample data set from our data set and IMFDB*
(a) Frame used to generate initial template for gun, ***(b)*** Result of proposed method using initial template without an iterative training, ***(c)*** Result of proposed method using template generated by an iterative training

Table 9 Results of the proposed method on the proposed data set

Conditions	Key challenges	Recall	Precision	F1	MCC
indoor	different rotations of gun	0.9207	0.9334	0.9270	0.6779
	complex background	0.9257	0.9507	0.9380	0.5878
	partial occlusion	0.7829	0.7625	0.7726	0.5010
	different types of guns	0.9027	0.9123	0.9075	0.6122
	multiple guns	0.9123	0.9118	0.9120	0.5922
	guns with scaling	0.9362	0.8980	0.9167	0.6336
	guns with panning	0.9433	0.8772	0.9091	0.6489
	guns with tilting	0.9149	0.8776	0.8958	0.6000
	mass Firing	0.8936	0.8571	0.8750	0.5663
	human with mask	0.9574	0.8491	0.9000	0.6522
outdoor	different rotations of gun	0.9091	0.9132	0.9111	0.6111
	complex background	0.9055	0.9271	0.9162	0.5213
	partial occlusion	0.7519	0.7157	0.7334	0.4766
	different types of guns	0.8809	0.8937	0.8873	0.6343
	multiple guns	0.8831	0.8972	0.8901	0.6505
	guns with scaling	0.9067	0.9600	0.9320	0.5977
	guns with panning	0.9245	0.8909	0.9074	0.6043
	guns with tilting	0.8867	0.9215	0.9039	0.5847
	mass firing	0.8679	0.8679	0.8679	0.5777
	human with mask	0.9245	0.9074	0.9158	0.6355

are also implemented on these data sets and the performance of the proposed method is compared with them. For fair comparison, the state-of-the-art methods are included that are implemented to detect guns from image or video sequence [5, 13, 14, 22, 23, 38–40]. The performance of methods are discussed for each data set as follows:

5.2.1 Data set 1: proposed data set: The proposed data set can be think of a first and only data set designed for detection of gun incorporating at most challenges. Existing methods including the F-RCNN fails to handle false positives in video frames. Therefore, results in higher number of false positive rates. The proposed method able to give satisfactory performance in this data set. The iterative refinement of the initial template make the proposed method efficient in detection of handgun in several situation or condition. Fig. 15 shows the effect of iterative training in localisation of gun and Table 9 represents the performance of the proposed method on the proposed data set. In outdoor scenarios, illumination problem is common and illumination effect able to change the colour of the gun and the data set also include rotation, scaling, pan and tilt of gun. To handle all these situations, the proposed method used SIFT, GDD operator along with the conventional GHT. No other existing algorithm, including F-RCNN able to handle these situation efficiently and results in higher values of False negatives and false positives. The results of each method experimented on the proposed data set are shown in Table 10.

5.2.2 Data set 2: gun movies data set: As mentioned earlier, this data set has been collected in very restricted environment with a single type of gun. The video sequences of the data set are very simple with respect to detection of gun. Uniform background is used here that intuitively reduce the number of false positives. The colour contrast between the gun and the background is distinct into some extent and only one object is present in the video. Rotation and scaling of guns also not noticeable in these sequences. Therefore, detection of the guns would be a simple task in this video data set. Some samples of this data set with the detected ROIs are shownn in Fig. 16.

The proposed method performed very well and the reason is simplicity of the video sequences. The initial template is enough to represent the properties of all the training images as there are no variability notices in the gun type, color and position. State-of-the-art methods also attain good results in this video data set. state-of-the-art methods need to work with the whole image. Whereas, the proposed method process only the result of the object detection method. As only one object is present in the video, therefore the proposed method needs to process a very small part of the image to localise gun. The quantitative measures of each method are tabulated in Table 11. From the view of the quantitative measure, it is evident that the proposed method is able to achieve best results compared to the state-of-the-art methods. F-RCNN also shows good result, but it is notable that F-RCNN has lacking in detection

Table 10 Comparison of the proposed method with state-of-the-art methods on the proposed data set

State-of-the-art methods	Different Rotations of gun		Complex background		Partial occlusion		Different types of gun		Multiple gun	
	recall	precision	recall	precision	recall	precision	recall	precision	recall	precision
proposed method	**0.9149**	**0.9233**	**0.9156**	**0.9389**	*0.7674*	*0.7391*	**0.8918**	**0.9030**	**0.8977**	**0.9045**
sliding window+CNN [13]	0.8545	0.8631	0.5080	0.5333	0.5447	0.5710	0.6970	0.7100	0.7543	0.7321
blob+SURF feature [14]	0.7688	0.7231	0.4976	0.4534	0.4868	0.4900	0.7699	0.7850	0.5000	0.5546
blob+Harris point features [5]	0.7455	0.7312	0.4829	0.4920	0.4776	0.4502	0.7218	0.7040	0.5823	0.5714
saliency map based [23]	0.5211	0.5439	0.8231	0.8541	**0.8632**	**0.8543**	0.8763	0.8923	0.7430	0.7752
RCNN [38]	0.8210	0.7934	0.9178	0.9343	0.7374	0.7321	0.8518	0.8830	0.8549	0.8045
FRCNN [22]	0.8319	0.8000	0.9276	0.9441	0.7479	0.7461	0.8798	0.8920	0.8667	0.8243
CNN detection model [39]	0.8912	0.9129	0.9217	0.9427	0.7555	0.7288	0.8850	0.8930	*0.8893*	*0.8935*
CNN detection model [40]	*0.9000*	*0.9033*	*0.9155*	*0.9342*	0.7324	0.7100	0.8534	0.8210	0.8015	0.8513

Highest accuracies are highlighted in bold and second highest accuracies are highlighted in italic.

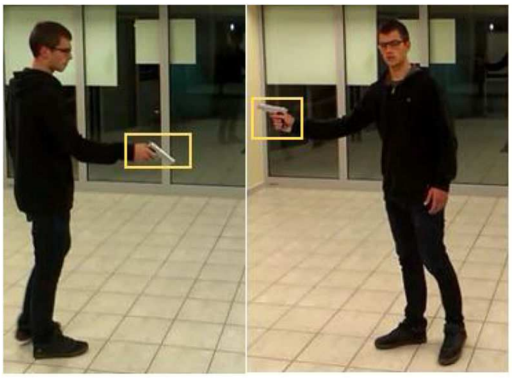


Fig. 16 *Few samples output of this data set with the detected ROIs*

of the gun. The reason is F-RCNN is designed to detect objects like persons, cars etc. In comparison to these objects, gun can be differentiated by its shape and size. Mostly, the size of handguns is very small compared to the object for which the F-RCNN is designed. Many literatures pointed out that the F-RCNN is not able to detect object captured from a large distance.

5.2.3 Data set 3: data set of Olmos et al. [22]: Olmos *et al.* [22] used online image repository for implementing deep learning-based methods for the gun detection. They collected total of **29,296 images** from online and grouped them into five categories. The **five categories** of the images are named as databases 1, 2, 3, 4 and 5. By experimenting fine tuned F-RCNN on each data set, proposed the different issues in designing data set for gun detection. Data set-1 designed with the images having different types of gun. F-RCNN fails in this data set 1 miserably and they concentrate on detection of one type of gun (Pistols). Database-2, 3, 4 and 5 are designed with images having only pistols. F-RCNN also fails in database-2, as there are only one class and F-RCNN consider the white background as a part of gun. Whereas, database-3, 4 and 5 contains a large number of classes. Although F-RCNN performed well in these three data sets, it attains highest accuracy on the data set 4 which contains the maximum number of classes.

Whereas feature matching based state-of-the-art methods [5, 14] are failed to detect gun effectively in all these data sets. The reason is images are of low quality image as the images are collected from the Internet. Furthermore, complex background and movement of objects are other difficulties, state-of-the-art methods cannot handle. Whereas the proposed method performs very well in Database-1 and Database-2, on which FRCNN fails as the proposed method is based on the edges of gun. The proposed method results in a high number of false positives when implementing on Database-4 because of a high number of classes. The average performance of the proposed method and other methods on the data set are shown in Table 11.

In this data set, proposed method attain 86% accuracy without DMC and iterative training module and above 90% accuracy with DMC and iterated training module in Database-1 and 2. A significant difference in this two results can be noticed. Some example results of both FRCNN and proposed method are shown in Fig. 17.

5.2.4 Database-4: IMFDB: IMFDB, a benchmark data set, with a huge online repository of gun images from movies. The data set contains a large variability of background, objects, environments and guns. We use a subset of the data set for implementing the proposed method and state-of-the-art methods. All state-of-the-art methods fail to detect guns in this data set for a variability of scenes. F-RCNN also unable to learn the features of gun effectively as the data set contains different type of objects, and different backgrounds. If the number of classes is increased then an improvement of performance of F-RCNN can be noticed. DMC and iterated training model able to improve the result, but compared to the other data set, the proposed method attains lower accuracy in this data set. The proposed method attains better results in this data set compared to the other methods but the accuracy is low. Low accuracy with respect to the accuracy obtained in the other data set. Fig. 18 shows some sample images IMFDB, and results of the proposed method.

5.3 Discussion

Above subsection describes that the proposed method performed very well in all the existing methods in comparison with the other existing methods. For further analysis of the parameters of the proposed method, we experimented proposed method for different experimental setup. At first, we implemented proposed method without the DMC and iterated training procedure. The aim of this experiment is to find out the performance of DMC and iterated training module. Without DMC and iterated training module proposed method fails due to the presence of false positives, partial occlusion and variability of shape of guns. When we experimented proposed method with DMC and iterated training on the data set, it improves the accuracy of the proposed method and handle the false positives very effectively. Fig. 15 describes the performance of the proposed method with and without DMC and iterated training.

Next experimental setup describes the performance of the proposed method in handling illumination and scaling. We mentioned that to handle illumination and scaling of gun we modify the R-table. Therefore, we first experiment on the proposed data set using conventional GHT and afterwards repeat the experiment using proposed M-GHT. In both cases we used the DMC and iterative module. The existing GHT fails to detect gun in different illumination but in the same location shown in Fig. 19. Conventional GHT attain ∼26% in the video frames with illumination. The accuracy is calculated frame by frame and then we averaged the each accuracy. We also experiment proposed method on all of these indoor and outdoor videos. M-GHT improve the performance of conventional GHT by at most 60%. The proposed method is able to detect almost every type of handgun due to the unique iterative generation procedure of template. Also, in mass firing scenarios proposed method able to detect all the

Table 11 Comparison of proposed method with state-of-the-art in other data set

Data set	Method	Detection result		
		Recall	Precision	F1
Gun movie data set [13]	Sliding window + CNN [13]	0.5630	0.5870	0.5747
	blob + SURF frature [14]	0.5821	0.6780	0.6264
	blob + Harris point frature [5]	0.5710	0.6790	0.6203
	saliency map based [23]	0.6440	0.6308	0.6373
	RCNN [38]	0.8967	0.8785	0.8875
	faster RCNN [22]	*0.9473*	*0.9322*	*0.9482*
	YOLO V3 [41]	0.9380	0.8730	0.9043
	CNN detection model [39]	0.9443	0.9328	0.9385
	CNN detection model [40]	0.9432	0.9217	0.9323
	proposed method	**0.9540**	**0.9550**	**0.9545**
IMFDB [21]	sliding window + CNN [13]	0.5421	0.5749	0.5580
	blob + SURF frature [14]	0.3951	0.4356	0.4143
	blob + Harris point frature [5]	0.4546	0.4837	0.4686
	saliency map based [23]	0.6784	0.7308	0.7036
	RCNN [38]	0.9067	0.9085	0.9075
	faster RCNN [22]	**0.9580**	*0.9430*	*0.9504*
	YOLO V3 [41]	0.9473	0.9473	0.9464
	CNN detection model [39]	*0.9576*	**0.9601**	**0.9588**
	CNN detection model [40]	0.8769	0.8892	0.8830
	proposed method	0.9478	**0.9279**	0.9377
Data set of Olmos *et al.* [22]	sliding window + CNN [13]	0.5064	0.5336	0.5196
	blob + SURF frature [14]	0.3554	0.3931	0.3733
	blob + Harris point frature [5]	0.4112	0.3952	0.4030
	saliency map based [23]	0.7553	0.7608	0.7580
	RCNN [38]	0.9131	0.9006	0.9068
	faster RCNN [22]	0.9421	0.9304	0.9362
	YOLO V3 [41]	0.9299	0.9464	0.9386
	CNN detection model [39]	*0.9569*	**0.9583**	*0.9575*
	CNN detection model [40]	0.8532	0.8906	0.8714
	proposed method	**0.9619**	*0.9582*	**0.9600**

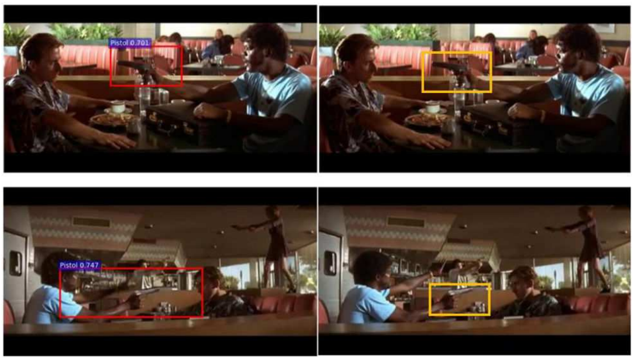


Fig. 17 *Column 1: output of RCNN and Column 2: proposed method on some sample images of database-4*

handguns. The proposed method fails in two cases when gun is mostly occluded and when the gun is pointed to the camera. For the second reason only the front hole of the gun is visible in the image. There is another situation, when the handgun is present in a large distance from the camera.

6 Conclusion

This paper presents a newly developed data set with an efficient gun detection algorithm which features low time complexity and high performance. The data set aims to provide the research community with a facility for testing and ranking of existing and new algorithms. The proposed algorithm used an iterative training procedure for the generation of novel gun model. To reduce the time complexity during model matching we employed an moving object detection algorithm. Moving object detection algorithm removed unwanted background object and reduce the search space for generated model. Experimentation showed that the proposed method performed averagely at the rate of 24 FPS in each environment. The speed of the algorithm is quite satisfactory for real-time implementation. In future, we enhance scope of this framework in detection of any object in video sequence.

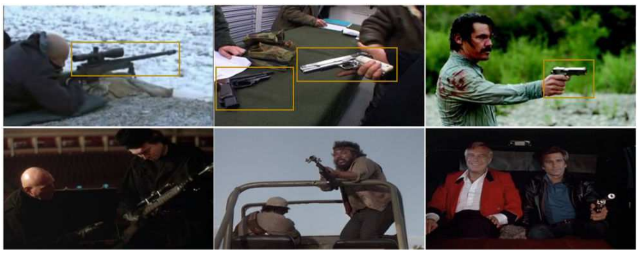


Fig. 18 *First row shown sample images where the proposed method succeed. Second row shown the where proposed method fails*

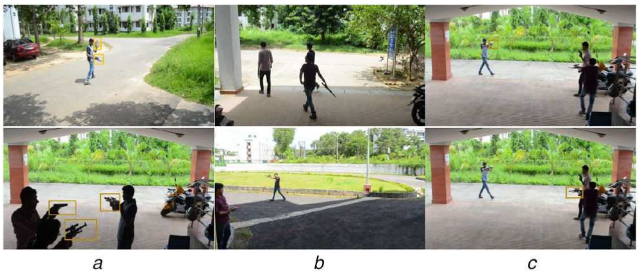


Fig. 19 *Few example images where proposed has failed to correctly detect the gun*
(a) Shown the scenario on which template is generated iteratively, ***(b)*** Generated template failed on the images which have different illumination from the images used to generate template, ***(c)*** Provide different result in the same frame because of the template

7 Acknowledgments

The work presented here is being conducted in the Biometrics Laboratory of Department of Computer Science and Engineering of Tripura University (A central university), Tripura, Suryamaninagar-799022, India.

8 References

[1] Welsh, B.C., Farrington, D.P.: ‘Public area CCTV and crime prevention: an updated systematic review and meta–analysis’, *Justice Q.*, 2009, **26**, (4), pp. 716–745

[2] Gill, M., Spriggs, A.: 2005 ‘Assessing the impact of CCTV’. Home Office Research', Development and Statistics Directorate, London

[3] Murphy, T.: ‘The admissibility of CCTV evidence in criminal proceedings’, *Int. Rev. Law Comput. Technol.*, 1999, **13**, (3), pp. 383–404

[4] Ojha, S., Sakhare, S.: ‘Image processing techniques for object tracking in video surveillance-A survey’. 2015 Int. Conf. on Pervasive Computing (ICPC), Pune, India, 2015, pp. 1–6

[5] Tiwari, R.K., Verma, G.K.: ‘A computer vision based framework for visual gun detection using harris interest point detector’, *Proc. Comput. Sci.*, 2015, **54**, pp. 703–712

[6] Glowacz, A., Kmiec, M., Dziech, A.: ‘Visual detection of knives in security applications using active appearance models’, *Multimedia Tools Appl.*, 2015, **74**, (12), pp. 4253–4267

[7] Velastin, S.A., Boghossian, B.A., Vicencio-Silva, M.A.: ‘A motion-based image processing system for detecting potentially dangerous situations in underground railway stations’, *Transp. Res. C, Emerg. Technol.*, 2006, **14**, (2), pp. 96–113

[8] Ainsworth, T.: ‘Buyer Beware’. Security Oz, 2002, vol. 19, pp. 18–26

[9] Stauffer, C., Grimson, W.E.: ‘Adaptive background mixture models for real-time tracking’. Proc. 1999 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 23 June 1999, vol. 2, pp. 246–252

[10] Ha, J.E, ., Lee, W.: ‘Foreground objects detection using multiple difference images’, *Opt. Eng.*, 2010, **49**, (4), p. 047201

[11] Barnich, O., Van, D.M.: 'ViBe: a universal background subtraction algorithm for video sequences', *IEEE Trans. Image Process.*, 2010, **20**, (6), pp. 1709–1724

[12] Cheng, F.C., Huang, S.C., Ruan, S.J.: 'Illumination-sensitive background modeling approach for accurate moving object detection', *IEEE Trans. Broadcast.*, 2011, **57**, (4), pp. 794–801

[13] Grega, M., Matiolanski, A., Guzik, P., *et al.*: 'Automated detection of firearms and knives in a CCTV image', *Sensors*, 2016, **16**, (1), p. 47

[14] Tiwari, R.K., Verma, G.K.: 'A computer vision based framework for visual gun detection using SURF'. 2015 Int. Conf. on Electrical, Electronics, Signals, Communication and Optimization (EESCO), Andhra Pradesh, India, January 2015, pp. 1–5

[15] Kmiec, M., Glowacz, A., Dziech, A.: 'Towards robust visual knife detection in images: active appearance models initialised with shape-specific interest points'. Int. Conf. on Multimedia Communications, Services and Security, Berlin, Germany, 31 May 2012, pp. 148–158

[16] Maksimova, A.: 'Knife detection scheme based on possibilistic shell clustering'. Int. Conf. on Multimedia Communications, Services and Security, Berlin, Germany, 6 June 2013, pp. 144–152

[17] Asnani, S., Ahmed, A., Manjotho, A.A.: 'Bank security system based on weapon detection using HOG features', *Asian J. Eng. Sci. Technol.*, 2014, **4**, (1), pp. 23–29

[18] Planty, M, Truman, J.L.: '*Firearm violence, 1993–2011*' (U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, Washington, DC, 2013)

[19] Perkins, C.A.: '*firearm use and violent crime', 1993–2001*' (U.S. Department of Justice, Washington, 2003)

[20] 'UPDATE: CPRC Original Research: All but two of the 25 ..'. [Online], Available: https://crimeresearch.org/2018/02/with-39-killed-in-tunisia-attack-the-top-three-mass-public-shootings-are-outside-the-united-states/, Accessed: 01-May-2019

[21] 'Main Page, Internet Movie Firearms Database – Guns in Movies, TV and Video Games'. Available at http://www.imfdb.org/wiki/Main_Page., accessed: 13-May-2019

[22] Olmos, R., Tabik, S., Herrera, F.: 'Automatic handgun detection alarm in videos using deep learning', *Neurocomputing*, 2018, **275**, pp. 66–72

[23] Ardizzone, E., Gallea, R., La Cascia, M., *et al.*: 'Combining top-down and bottom-up visual saliency for firearms localization'. 2014 Int. Conf. on Signal Processing and Multimedia Applications (SIGMAP), Vienna, Austria, 28 August 2014, pp. 25–32

[24] Halima, N.B., Hosam, O.: 'Bag of words based surveillance system using support vector machines', *Int. J. Secur. Appl.*, 2016, **10**, (4), pp. 331–346

[25] Gelana, F., Yadav, A.: 'Firearm detection from surveillance cameras using image processing and machine learning techniques'. Smart Innovations in Communication and Computational Sciences, Singapore, 2019, pp. 25–34

[26] Iqbal, J., Munir, M.A., Mahmood, A., *et al.*: 'Orientation Aware Object Detection with Application to Firearms'. arXiv preprint arXiv:1904.10032. 22 April 2019

[27] Milne, R.: '*Forensic intelligence*' (CRC Press, Boca Raton, FL, 2013)

[28] 'The PASCAL Visual Object Classes Homepage', available: http://host.robots.ox.ac.uk/pascal/VOC/, accessed: 13-May-2019

[29] Cuevas, C., Yanez, E.M., Garcia, N.: 'Tool for semiautomatic labeling of moving objects in video sequences: TSLAB', *Sensors*, 2015, **15**, (7), pp. 15159–15178

[30] Darker, I.T., Kuo, P., Yang, M.Y., *et al.*: 'Automation of the CCTV-mediated detection of individuals illegally carrying firearms: combining psychological and technological approaches'. Visual Information Processing XVIII, Orlando, FL, USA, 27 April 2009, vol. 7341, pp. 73410p

[31] Maddalena, L., Petrosino, A.: 'The 3dSOBS+ algorithm for moving object detection', *Comput. Vis. Image Underst.*, 2014, **122**, pp. 65–73

[32] St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: 'Universal background subtraction using word consensus models', *IEEE Trans. Image Process.*, 2016, **25**, (10), pp. 4768–4781

[33] Javed, S., Mahmood, A, ., Al-Maadeed, S., *et al.*: 'Moving object detection in complex scene using spatiotemporal structured-sparse RPCA', *IEEE Trans. Image Process.*, 2018, **28**, (2), pp. 1007–1022

[34] Ballard, D.: 'Generalizing the Hough transform to detect arbitrary shapes', *Pattern Recognit.*, 1981, **13**, (2), pp. 111–122

[35] Capar, A., Kurt, B., Gokmen, M.: 'Gradient-based shape descriptors', *Mach. Vis. Appl.*, 2009, **20**, (6), pp. 365–378

[36] Lowe, D.G.: 'Object recognition from local scale-invariant features'. Proc. of the seventh IEEE Int. Conf. on Computer Vision, Kerkyra, Greece, 20 September 1999, vol. 2, pp. 1150–1157

[37] Maddalena, L., Petrosino, A.: 'Background subtraction for moving object detection in RGBD data: a survey', *J. Imaging*, 2018, **4**, (5), p. 71

[38] Dubey, S.: 'Building a Gun Detection Model Using Deep Learning'. Program Chair & Proceedings Editor: M. Afzal Upal, PhD Chair of Computing & Information Science Department Mercyhurst University 501 E 38th St, Erie, PA, USA, 16546

[39] Egiazarov, A.: 'Firearm detection through component decomposition'. Master's thesis

[40] Romero, D., Salamea, C.: 'Design and proposal of a database for firearms detection'. The Int. Conf. on Advances in Emerging Trends and Technologies, 27 March 2019, pp. 348–360

[41] Redmon, J., Farhadi, A.: 'YOLO9000: better, faster, stronger'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 7263–727