# Suspicious Action Detection and Recognition in Remote Areas Using AI and ML Techniques

Chaitra B M[1], Vijay Kumar H R[2], Pavan G S[3]

[1]PG Scholar, Department of ECE, Akshaya institute of technology, Tumkuru
[2,3]Assistant Professor, Department of ECE, Akshaya institute of technology, Tumkuru

## ABSTRACT

*Detection of suspicious activities in public transport areas using video surveillance has attracted an increasing level of attention. In general, automated offline video processing systems have been used for post-event analysis, such as forensics and riot investigations. However, very little has been achieved regarding real-time event recognition. In this paper, we introduce a frame- work that processes raw video data received from a fixed color camera installed at a particular location, which makes real- time inferences about the observed activities. First, the proposed framework obtains 3-D object-level information by detecting and tracking people and luggage in the scene using a real-time blob matching technique. Based on the temporal properties of these blobs, behaviors and events are semantically recognized by em- ploying object and interobject motion features. These supervised machine learning techniques are used for detection and tracking of social distancing between one or more person's movements in public places and these observations can be done by the CCTV videos. A number of types of behavior that are relevant to security in public transport areas have been selected to demonstrate the capabilities of this approach. Examples of these are abandoned and stolen objects, fighting, fainting, and loitering. Using standard public data sets, the experimental results presented here demonstrate the out- standing performance and low computational complexity of this approach.*

**Keyword**: *blob matching, fainting, fighting, loitering, meeting, object tracking,*

## 1. INTRODUCTION

INCREASINGLY, police and security staff rely on video surveillance systems to facilitate their work. This practice is most evident in large public transportation areas such as metro stations and airports. However, these systems remain largely labor intensive, and the personnel monitoring the video displays find it extremely difficult to be attentive to randomly occurring incidents [1], [2]. Although automated video surveil- lance systems do exist, they have been used mainly for offline video analysis after an event has occurred, most notably in the case of riot investigations and forensics. At present, these surveillance systems are of marginal help for real-time alerts. Moreover, contrary to the false image created by the media and film industry, research in this young but promising field has made little advancement so far.

The function of an automated surveillance system is to draw the attention of monitoring personnel to the occurrence of a user-defined suspicious behavior when it happens. Two challenges stand in the face of developing fully automated behavior recognition. First, objects of interest, such as people and luggage in a scene, must be found robustly, classified, and tracked through time. Second, a stable means of describing events must be found. This is particularly an issue for complex types of events having many different possible variations, such as fighting. Undeniably, in many cases, they are extremely difficult to describe. What are the contributions of this paper? The majority of researchers to date have invoked machine learning to detect suspicious behavior. To our knowledge, we uniquely propose here a complete semantics-based solution to the behavior de- tection problem that addresses the whole process from pixel to behavior level. Furthermore, the processing is achieved in real time. Although much of the lower level processing stages in this paper are not original, part of our contribution in this regard was to carefully select and integrate them. This proved to be critical for ultimately making correct high-level inferences, which is an issue seldom addressed in the field.

The primary disadvantage of machine learning is that the learned classifiers depend on having reliable standard data sets for training and testing. These are extremely difficult to obtain, particularly for anomalous types of behaviors. This issue is of utmost importance when determining classifier parameters and thresholds. In contrast, the semantic approach replaces this need for training with a more straightforward process based on human reasoning and logic. We claim that this is a more feasible and viable method. For example, it eliminates the specification of complex learning parameters such as decision-tree-pruning thresholds, which are not intuitive to tune and require the intervention of experts in the field. In the semantic approach, more intuitive and meaningful parameters replace these. This paper assumes that foreground blobs are extracted in each frame using a conventional background subtraction method. These blobs represent the silhouettes of animate (e.g., people) and inanimate (e.g., luggage) objects in the scene, which are the semantic entities associated with the events described. How- ever, in practice, we note that a single blob will often represent multiple objects occluding or standing next to each other. After all blobs have been extracted, inferences are made to segment, track, and classify the objects that they represent. Finally, the anomalous events must be labeled.

**Aim:** Analyzing the performance in the human behavior in public places and video of person are mapping and testing for HAR system.

**Objectives:**

- Analysis of Human Action Recognition paves a way to develop a video analytics system which helps to recognize the Human suspicious behavioural expression.

- The objective of the project is to recognize the Real Time Human Actions using CCTV datasets by extracting required features.

- These feature extraction techniques are also used for identifying the social distancing between one or more person’s movements in public places and these observations can be done by the CCTV videos.

## 2. LITERATURE SURVEY

Following is the work done carried out in various article by the authors .

**Multi-View Fusion for Action Recognition in Child-Robot Interaction:** In 2018, Niki Efthymiou, Petros Koutras, Panagiotis Paraskevas Filntisis, Gerasimos Potamianos, Petros Maragos took the challenge of leveraging computer vision methods in order to enhance Human Robot Interaction (HRI) experience, this work explores methods that can expand the capabilities of an action recognition system in such tasks. A multi-view action recognition system is proposed for integration in HRI scenarios with special users, such as children, in which there is limited data for training and many state-of-the-art techniques face difficulties. Different feature extraction approaches, encoding methods and fusion techniques are combined and tested in order to create an efficient system that recognizes children pantomime actions. This effort culminates in the integration of a robotic platform and is evaluated under an alluring Children Robot Interaction scenario.

**Deep Learning Fusion Conceptual Frameworks for Complex Human Activity Recognition Using Mobile and Wearable Sensors:** In 2018, Nweke Henry Friday, Ghulam Mujtaba, Mohammed Ali Al-garadi, Uzoma Rita Alo, analysed to recognize activities using mobile or wearable sensor, data are collected using appropriate sensors, segmented, needed features extracted and activities categories using discriminative models (SVM, HMM, MLP etc.). Feature extraction is an important stage as it helps to reduce computation time and ensure enhanced recognition accuracy.

**Improving human action recognition with two-stream 3D convolutional neural network:** In 2018, Van-Minh Khong, Thanh-Hai Tran, They have proposed a method that exploits both RGB and optical flow for human action recognition. Specifically, we deploy a two stream convolutional neural network that takes RGB and optical flow computed from RGB stream as inputs. Each stream has architecture of an existing 3D convolutional neural network (C3D) which has been shown to be compact but efficient for the task of action recognition from video. Each stream works independently then is combined by early fusion or late fusion to output the recognition results.

**Information Fusion for Human Action Recognition via Biset/Multiset Globality Locality Preserving Canonical Correlation Analysis:** In NOVEMBER 2018, Nour El Din Elmadany , Student Member, IEEE, Yifeng He, Member, IEEE, and Ling Guan , Fellow, IEEE proposed two novel information fusion techniques for fusing the information from multisets. The first technique is biset globality locality preserving canonical correlation analysis (BGLPCCA), which aims to learn the common feature subspace between two sets. The second technique is multiset globality locality preserving canonical correlation analysis (MGLPCCA), which aims to deal with three or more sets. The proposed BGLPCCA and MGLPCCA are able to learn a low-dimensional common subspace that preserves the local and global structures of data samples.

**A survey of depth and inertial sensor fusion for human action recognition:** In 2018, Roozbeh Jafari, Nasser Kehtarnavaz took a number of review or survey articles have previously appeared on human action recognition where either vision sensors or inertial sensors are used individually. Considering that each sensor modality has its own limitations, in a number of previously published papers, it has been shown that the fusion of vision and inertial sensor data improves the accuracy of recognition. This survey article provides an overview of the recent investigations where both vision and inertial sensors are used together and simultaneously to perform human action recognition more effectively. The thrust of this survey is on the utilization of depth cameras and inertial sensors as these two types of sensors are cost-effective, commercially available, and more significantly they both provide 3D human action data.

**Local Feature Extraction from RGB and Depth Videos for Human Action Recognition:** In June 2018, Rawya Al-Akam and Dietrich Paulus studied a novel system to analyze human body actions for recognizing human actions by using 3D videos (RGB and depth data). We apply the Bag-of-Features techniques for recognizing human actions by extracting local-spatial temporal features from all video frames. K-means clustering and multi-class Support Vector Machines are used for the action classification task. This system is invariant to scale, rotation and illumination. This new features combination method is help to reach recognition rates superior to other publications on the dataset.

**Data Fusion and Multiple Classifier Systems for Human Activity Detection and Health Monitoring: Review and Open Research Directions:** In 2018, Henry Friday Nweke, The Ying Wah, Ghulam Mujtaba, they have focused of this review is to provide in-depth and comprehensive analysis of data fusion and multiple classifier systems techniques for human activity recognition with emphasis on mobile and wearable devices. First, data fusion methods and modalities were presented and also feature fusion, including deep learning fusion for human activity recognition were critically analysed, and their applications, strengths and issues were identified. Furthermore, the review presents different multiple classifier system design and fusion methods that were recently proposed in literature. Finally, open research problems that require further research and improvements are identified and discussed.

## 3. EXISTING AND PROPOSED SYSTEM

**EXISTING SYSTEM:** Behavior recognition is a broad term that covers a number of categories of activities, which require different means of detection to describe suspicious behaviors and activities in public transportation systems. Most of the existing systems focuses on detecting a single type of behavior rather than providing a generic framework. For example, “abandoned luggage detection” is usually handled using low-level background subtraction methods. These methods are useful for detecting stationary foreground objects but hardly so for other complex types of behavior such as loitering or fighting, which requires experimentation and fine-tuning. This is because such behaviors manifest a broad range of variations and are very difficult to model, even if based on reasoning. The former are robust to non-rigid motion of the kind observed in surveillance videos, but are susceptible to clutter, noise, and fast pose changes. Further, the most published researches only describes behavior detection, without delving into the details of object detection and tracking. Examples are loitering, abandoned objects, and fighting. These types of behavior may occur over a significant period of time. They often involve more than one object; therefore, such matters as finding trajectories, identity tracking, and object classification must be addressed. Ultimately, machine-learning approaches in the current literature seem to be unable to generalize and systems based on semantics.

**PROPOSED SYSTEM:** The proposed system focuses on automatically flagging suspicious behavior in public transportation systems. First, the proposed framework obtains 3-D object-level information by detecting and tracking people and luggage in the scene using a real-time blob matching technique. Based on the temporal properties of these blobs, behaviors and events are semantically recognized by employing object and inter-object motion features. A number of types of behavior that are relevant to security in public transport areas have been selected to demonstrate the capabilities of this approach. However, in practice, we note that a single blob will often represent multiple objects occluding or standing next to each other. After all blobs have been extracted, inferences are made to segment, track, and classify the objects that they represent. Finally, the anomalous events must be labeled. Examples of these are abandoned and stolen objects, fighting, fainting, and loitering. Using standard public data sets, the experimental results presented here demonstrate the outstanding performance and low computational complexity of this approach. Our framework performs object tracking in an average time of 11 ms per object per frame, whereas behavior recognition averages just about 1 ms per frame. These components, along with background subtraction, constitute the total processing time required per frame, which is approximately 200 ms. Further, the color histograms we use seem to provide low complexity while simultaneously dealing with constantly occluding patterns. In addition to the single-object features, the inter-object features between every combination of two objects are also stored in historical sequence.

## 4. DESIGN METHODOLOGY

The system consists of four major parts; speech acquisition, feature extraction at each timescale level, machine learning for each feature set, and information fusion to merge the information. Fig. 1 illustrates the basic concept of the system.

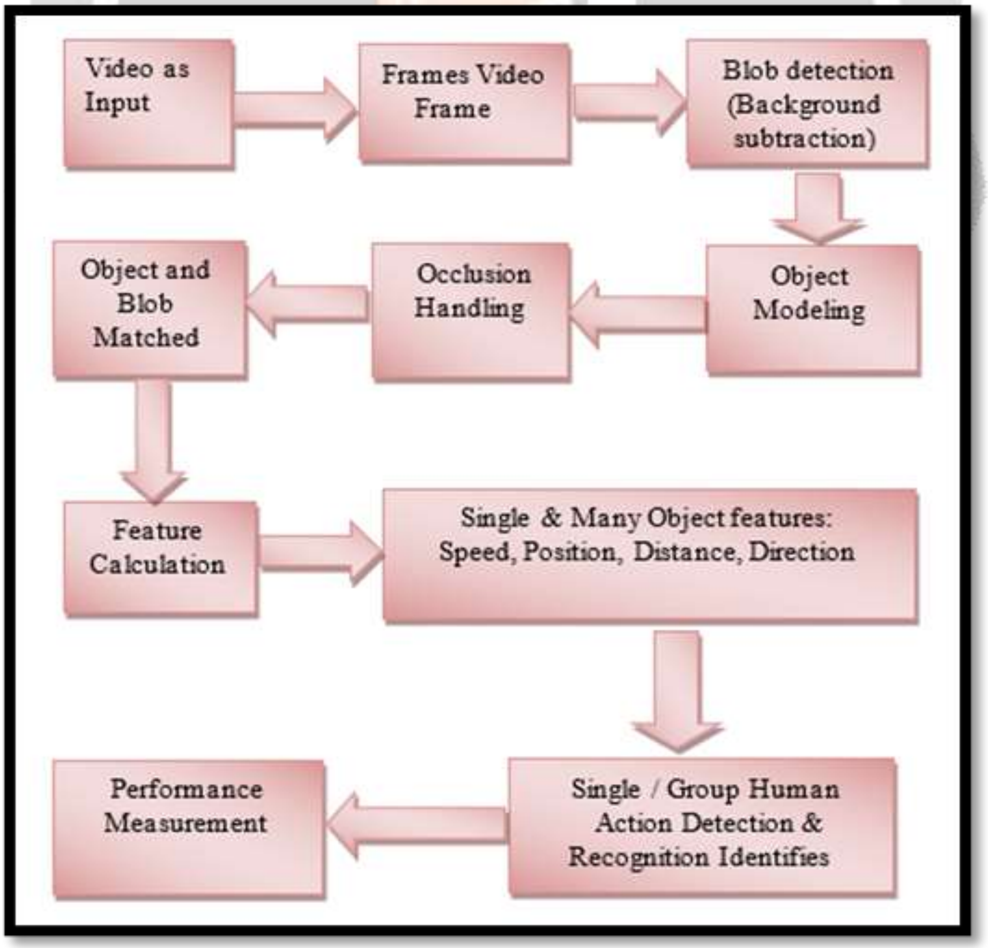

Video as Input
Frames Video Frame
Blob detection (Background subtraction)
Object Modeling
Occlusion Handling
Object and Blob Matched
Feature Calculation
Single & Many Object features: Speed, Position, Distance, Direction
Single / Group Human Action Detection & Recognition Identifies
Performance Measurement

Fig 1: System Architecture

The proposed system architecture is as shown in the Fig.1. The main action detection and recognition is based on the occlusion handling stage.

The common localization and target representation algorithms are as shown below:

**Blob Tracking:** This algorithm describes the object interior segmentation (e.g. Blob detection, optical flow / block-based correlation)
**Kernel Based Tracking**: KBT is also called as mean-shift tracking, which is a procedure of iterative localization and also depends on the similarity measure maximization of all actions.
**Contour Tracking:** CT detection is used to identify of object boundary (e.g. Condensation algorithm/active contours)
**Visual Feature Matching:** VFM techniques is used to identify the matching of visual actions of single or group human behaviors.

**Occlusion Handling:** Occlusion handling is a critical task because it bears on the robustness of object tracking and coherence. If occlusion is resolved incorrectly, inferences following from this will most likely lead to a false understanding of the scene. In concordance with [30] and [33], we argue that finding the exact location of objects participating in occlusion within a single blob is an exhaustive search that is computationally expensive and actually unnecessary. This is because localization at the blob level provides sufficient spatial information for determining the object location. Thus, we consider the location of a blob to be the actual location of all its constituent objects. In this paper, the issue of which objects are occluding which is completely ignored, and we adopt the position that all merged objects form a pool (the blob) with no particular occluding/occluded relationships being noted. We also create a dummy object for the pool that exhibits the adaptive appearance model necessary for blob matching. In a nutshell, we render the phenomenon of occlusion into a split/merge problem. In addition, we adopt the concept of potential occlusion [28], which permits an object that has not yet been conclusively associated with any of the splitting blobs to be associated with all the accompanying split- ting blobs until such time that resolution becomes conclusively possible.

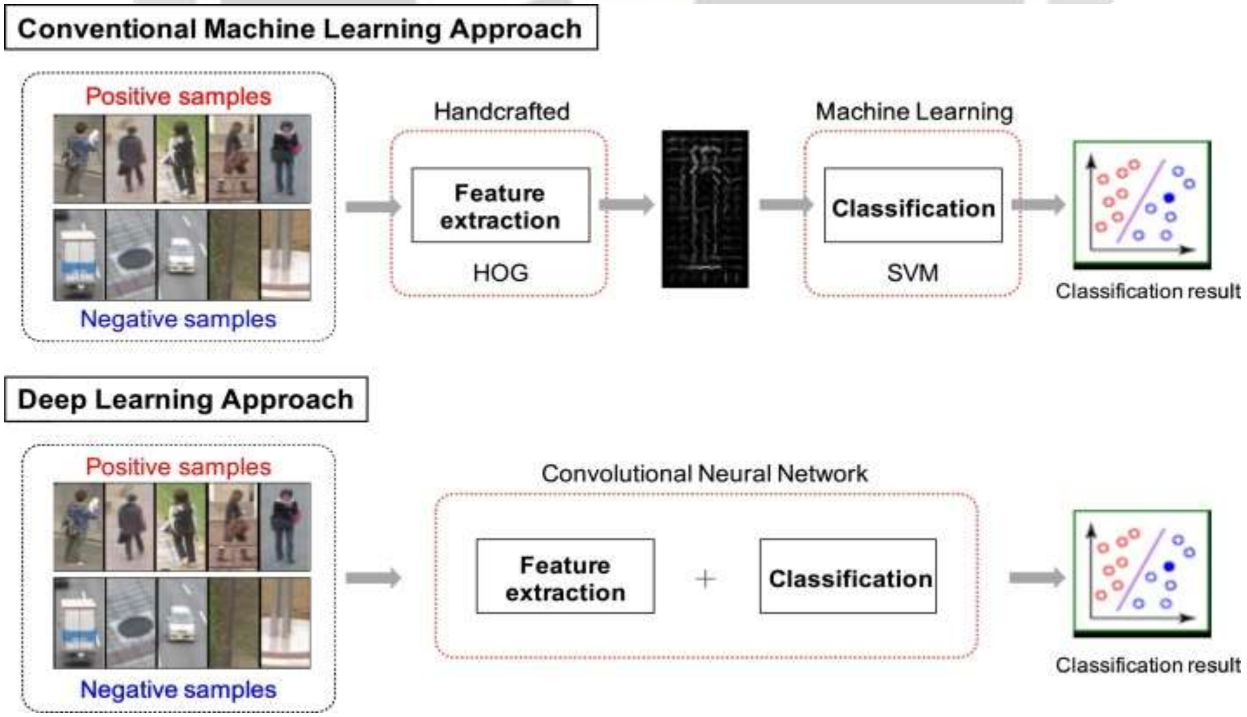


Fig2. The conventional Machine Learning and Deep Learning approaches for single or group human action detection and recognitions

**Evaluation:** Using the aforementioned techniques, our object tracking system was found to be highly reliable. This is evident in a number of tests that were performed on public data sets. References [44] and [46] are examples that demonstrate smooth tracking and occlusion handling. Videos cited in Section VII also demonstrate that our system yields reliable behavior recognition, although it is not based on using learning methods. It is worth noting that, in spite of the robustness of our approach, failures such as lost tracks and object confusion are inevitable. However, in the majority of tests performed on a number of standard data sets, this approach was able to successfully track people and their luggage, even in circumstances that involved three or four occluding objects. After determining the objects of interest in the video, their 3-D motion features are calculated, and an historical record is created. Based on this record, objects are classified as being either animate (persons) or inanimate. This classification process is important because it is an integral component of the definition of semantic behavior. There are many potential features discussed in the literature [14], [15], [48]. These can be split into single-object features, such as position, and inter object features, such as the alignment between two objects.

These features are measured in real-world 3-D spatial coordinates, which can be calculated from the image (pixel) coordinates by means of any traditional camera calibration method. The position of an object, in terms of which almost all the rest of its features are calculated, is obtained by applying the transformation to the pixel locations of the feet. These are simply designated as the lowest pixels of the 2-D blob to which the object belongs.

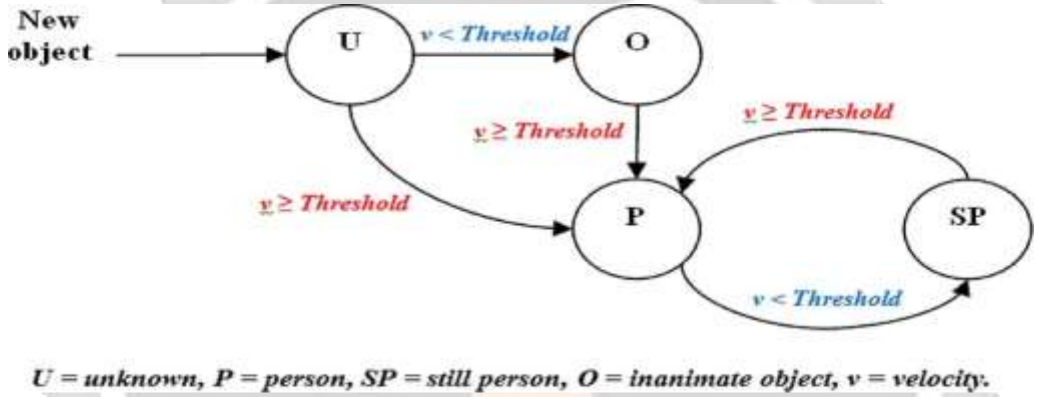


Fig. 3. Object classification state diagram.

When a new object occurs in the scene, it is classified as unknown. When its motion features are sampled, the velocity is used to determine whether it is a person or an inanimate object. Using this transition model ensures that a still person is not misclassified as luggage. Fig. 3 shows the state diagram for the implemented algorithm, which is largely adapted from [12] with a few minor modifications. The algorithm distinguishes between an inanimate object and a still person, a subtlety highly important for its consequences in understanding the scene.

**Abandoned and Stolen Objects:** A major concern in the literature to date has been the detection of abandoned luggage. Generally, detection has been performed using only background subtraction methods, such as[16] and [17], without other forms of reasoning such as object classification and tracking. The problem with this is that such an approach cannot discriminate between a stationary person and an abandoned object. Other methods use features such as color, edges, shape completeness, and histogram contrast [51]. In our experience, none of these was found to be sufficiently robust to noise and pose changes. Moreover, the issue of finding the object’s owner is still inadequately addressed. This is crucial, for example, to the distinction between stolen and retrieved luggage. This paper addresses the aforementioned shortcomings using a semantic definition. We use the definition in [12], which defines an abandoned object as “a stationary object that has not been touched by a person for some time threshold.” Integrating the object ownership into this statement

**Meeting and Walking Together:** Although generally not considered to be suspicious, meeting and walking together may be useful in certain surveillance scenarios. This would be particularly the case were face recognition included as a feature. For example, it might be pertinent for security purposes to flag individuals that meet with a suspicious individual. Table III defines both events semantically in terms of each person’s speed, the distance between them, and their alignment.

## 5. RESULTS AND DISCUSSION

The evaluation of behavior recognition experiments is challenged by a number of difficulties at several levels [60]. First, most activities of interest are of high complexity, which becomes an issue in the presence of clutter in the test scenario. Another issue is the inadequacy of professional and challenging high-quality data sets currently available for testing. Moreover, criteria for performance evaluation, such as a standard metric, hit-and-miss weighting, and the construction of the ground truth, are still subject to controversy. These challenges lead to inconsistencies among the experimental results in different papers in the literature.

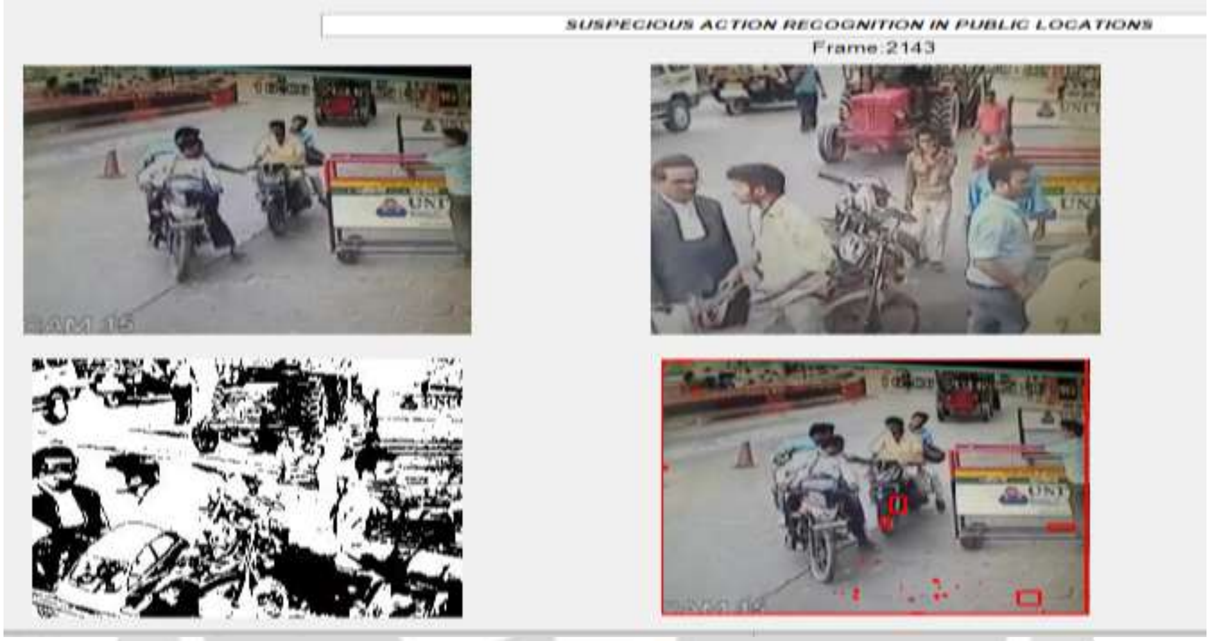


Fig 4. Illustrating the Real time fighting video in public area using image processing

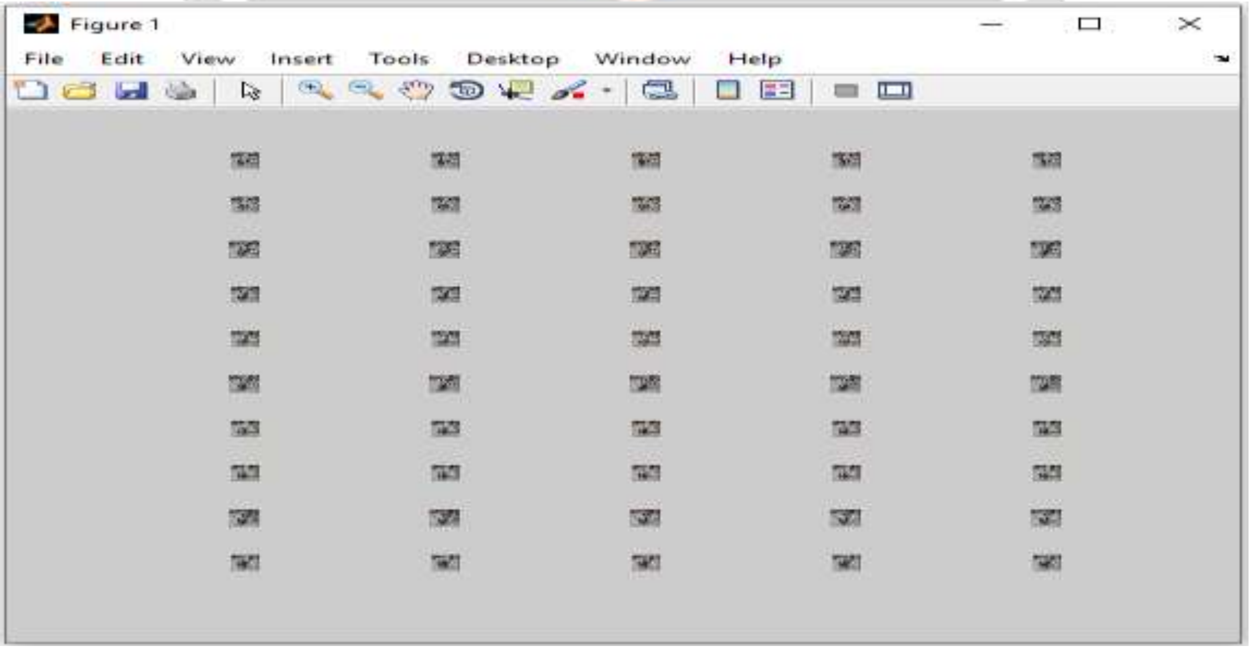
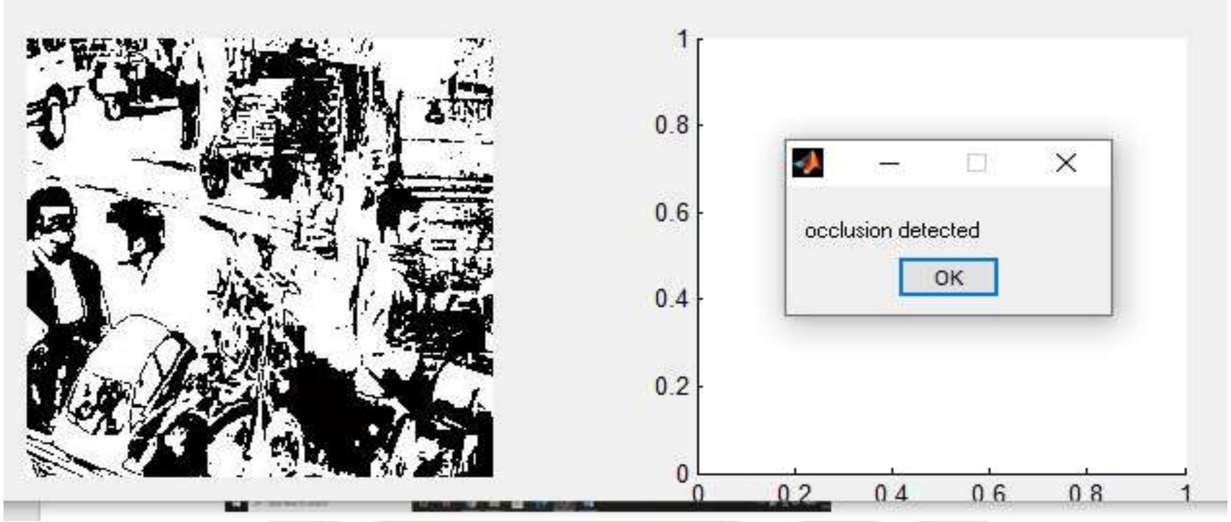


Fig 5. Illustrating the process of converting video into frames of Real time fighting video in public area using image processing



occlusion detected

Fig 6. Illustrating the blobs detection and Occlusion detection outputs of Real time fighting video in public area using image processing.

## 6. CONCLUSION AND FUTURE SCOPE

**Conclusion:** We observe that parameter tuning can be interpreted as being analogous to the problem of undertraining in machine learning since both represent a certain deficiency of knowledge. However, the semantic approach has the advantage of permitting human reasoning to easily model parameter values (e.g., speeds, distances, and angles). This is contrasted to the difficulty of finding sufficiently large and meaningful data sets for training machine learning systems. Of course, learning also requires fine-tuning of parameters, such as neural network size, connections, as well as learning parameters. Ultimately, machine-learning approaches in the current literature seem to be unable to generalize and systems based on semantics.

In this paper, a complete semantics-based behavior recognition approach that depends on object tracking has been introduced and extensively investigated. Our approach begins by translating the objects obtained by background segmentation into semantic entities in the scene. These objects are tracked in 2-D and classified as being either animate (people) or inanimate (objects). This approach ensures real-time performance, adaptability, robustness against clutter and camera nonlineari- ties, ease of interfacing with human operators, and elimination of the training required by machine-learning-based methods. Experimentation was carried out on multiple standard publicly available data sets that varied in terms of crowd density, camera angle, and illumination conditions. The experimental results demonstrated successful detection of the various activities of interest.

**Future Scope:** The future applications of real time human action recognition system using artificial intelligence and machine learning techniques can also be made functional using novel concept data fusion techniques. In this study the fusion of multiple actions of a single person or multiple human actions can be detected and recognized to obtain the best output performance to the given input videos.

## REFERENCES

[1] Dr. H S Mohan and Mahanthesha U, “Human action Recognition using STIP Techniques”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-7, May 2020

[2] J. F. Allen, “Maintaining knowledge about temporal intervals,” Commun. ACM, vol. 26, no. 11, pp. 832–843, Nov. 1983.

[3] C. Fernandez, P. Baiget, X. Roca, and J. Gonzalez, “Interpretation of complex situations in a semantic-based surveillance framework,” Image Commun., vol. 23, no. 7, pp. 554–569, Aug. 2008.

[4] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, “Understanding transit scenes: A survey on human behavior-recognition algorithms,” IEEE Trans. Intell. Transp. Syst., vol. 11, no. 1, pp. 206–224, Mar. 2010.

[5] Y. Changjiang, R. Duraiswami, and L. Davis, “Fast multiple object tracking via a hierarchical particle filter,” in Proc. 10th IEEE ICCV, 2005, vol. 1, pp. 212–219.

[6] A. Loza, W. Fanglin, Y. Jie, and L. Mihaylova, “Video object tracking with differential Structural SIMilarity index,” in Proc. IEEE ICASSP, 2011, pp. 1405–1408.

[7] D. Comaniciu, V. Ramesh, and P. Meer, “Kernel-based object tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 5, pp. 564–577, May 2003.

[8] V. Papadourakis and A. Argyros, “Multiple objects tracking in the presence of long-term occlusions,” Comput. Vis. Image Underst., vol. 114, no. 7, pp. 835–846, Jul. 2010.

[9] Mahanthesh U, Dr. H S Mohana “Identification of Human Facial Expression Signal Classification Using Spatial Temporal Algorithm” International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE) Vol 2, Issue 5, May 2016

[10] NikiEfthymiou, Petros Koutras, Panagiotis, Paraskevas, Filntisis, Gerasimos Potamianos, Petros Maragos “Multi-View Fusion for Action Recognition in Child-Robot Interaction”: 978-1-4799-7061-2/18/$31.00 ©2018 IEEE.

[11] Nweke Henry Friday, Ghulam Mujtaba, Mohammed Ali Al-garadi, Uzoma Rita Alo, analysed “Deep Learning Fusion Conceptual Frameworks for Complex Human Activity Recognition Using Mobile and Wearable Sensors”: 978-1-5386-1370-2/18/$31.00 ©2018 IEEE.

[12] Van-Minh Khong, Thanh-Hai Tran, ”Improving human action recognition with two-stream 3D convolutional neural network”, 978-1-5386-4180-4/18/$31.00 ©2018 IEEE.

[13] Nour El Din Elmadany , Student Member, IEEE, Yifeng He, Member, IEEE, and Ling Guan, Fellow, IEEE ,”Information Fusion for Human Action Recognition via Biset /Multiset Globality Locality Preserving Canonical Correlation Analysis” IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 27, NO. 11, NOVEMBER 2018.

[14] Pavithra S, Mahanthesh U, Stafford Michahial, Dr. M Shivakumar, “Human Motion Detection and Tracking for Real-Time Security System”, International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 12, December 2016.

[15] Lalitha. K, Deepika T V, Sowjanya M N, Stafford Michahial, “Human Identification Based On Iris Recognition Using Support Vector Machines”, International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE) Vol 2, Issue 5, May 2016

[16] RoozbehJafari, Nasser Kehtarnavaz “A survey of depth and inertial sensor fusion for human action recognition”, https://link.springer.com/article/10.1007/s11042-015-3177-1, 07/12/2018.

[17] Rawya Al-Akam and Dietrich Paulus, ”Local Feature Extraction from RGB and Depth Videos for Human Action Recognition”, International Journal of Machine Learning and Computing, Vol. 8, No. 3, June 2018

[18] V. D. Ambeth Kumar, V. D. Ashok Kumar, S. Malathi, K. Vengatesan and M. Ramakrishnan, “Facial Recognition System for Suspect Identification Using a Surveillance Camera”, ISSN 1054-6618, Pattern Recognition and Image Analysis, 2018, Vol. 28, No. 3, pp. 410–420. © Pleiades Publishing, Ltd., 2018.