IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 May 2023, accepted 30 June 2023, date of publication 10 July 2023, date of current version 26 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3293537

RESEARCH ARTICLE

Context Aware Crowd Tracking and Anomaly Detection via Deep Learning and Social Force Model

FAISAL ABDULLAH[1], MAHA ABDELHAQ[2], (Member, IEEE), RAED ALSAQOUR[3], (Member, IEEE), MOHAMMED HAMAD ALATIYYAH[4], KHALED ALNOWAISER[5], SAUD S. ALOTAIBI[6], AND JEONGMIN PARK[7]

[1]Department of Computer Science, Air University, Islamabad 44000, Pakistan
[2]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[3]Department of Information Technology, College of Computing and Informatics, Saudi Electronic University, Riyadh 93499, Saudi Arabia
[4]Department of Computer and Information, Prince Sultan University, Riyadh 12435, Saudi Arabia
[5]Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[6]Information Systems Department, Umm Al-Qura University, Makkah 24382, Saudi Arabia
[7]Department of Computer Engineering, Tech University of Korea, Siheung-si, Gyeonggi-do 15073, South Korea

Corresponding author: Jeongmin Park (jmpark@tukorea.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) (2021R1F1A1063634) funded by the Ministry of Science and Information & Communications Technology (MSIT), Republic of Korea.

ABSTRACT The world's expanding populace, the variety of human social factors, and the densely populated environment make humans feel uncertain. Individuals need a safety officer who generally deals with security viewpoints for this frailty. Currently, human monitoring techniques are time-consuming, work concentrated, and incapable. Therefore, autonomous surveillance frameworks are necessary for the modern day since they are able to address these problems. Nevertheless, hardships persist. The central concerns incorporate the detachment of the foreground from the scene and the understanding of the contextual structure of the environment for efficiently identifying unusual objects. In our work, we introduced a novel framework to tackle these difficulties by presenting a semantic segmentation technique for separating a foreground object. In our work, Super-pixels are generated using an improved watershed transform and then a conditional random field is implemented to obtain multi-object segmented frames by performing pixel-level labeling. Next, the Social Force model is introduced to extract the contextual structure of the environment via the fusion of a novel chosen particular histogram of an optical stream and inner force model. After using the computed social force, multi-people tracking is performed via three-dimensional template association using percentile rank and non-maximal suppression. Next, multi-object categorization is performed via deep learning Feature Pyramid Network. Finally, by considering the contextual structure of the environment, Jaccard similarity is utilized to make the decision for abnormality detection and identify the unusual objects from the scene. The invented framework is verified through rigorous investigations, and it obtained multi-people tracking efficiency of 92.2% and 89.1% over the UCSD and CUHK Avenue datasets. However, 95.2% and 93.7% abnormality detection efficiency is accomplished over UCSD and CUHK Avenue datasets, respectively.

INDEX TERMS Conditional random field, feature pyramid network, improved watershed transform, Jaccard similarity, multi-object association, social force model.

The associate editor coordinating the review of this manuscript and approving it for publication was Zeev Zalevsky.

I. INTRODUCTION

Crowd tracking and abnormality detection are demanding topics in today's crowded and complex environment [1], [2]. With the accessibility of video streams from open-air spots, there has been a flood in research yields on video examination and oddity discovery [3], [4]. Ordinarily, oddity recognition strategies gain proficiency in the normal environment through training. Anything veering off altogether from the normal environment can be named irregular [5], [6]. Some abnormalities include the appearance of automobiles on sidewalks, an unexpected dispersal of individuals inside a social occasion, and an individual falling out of nowhere while strolling. Crowd irregularity location is vital for robotized crowd environment examination [7], [8].

To regulate, secure, and control the pedestrian crowd, multi-people tracking is an essential video-outline examining process as it gives fundamental portrayals of group status [9], [10]. In any case, tracking in packed scenes is a difficult issue as a result of prompt enlightenment changes, various viewpoints and ways of behaving, halfway or full impediments, confounded foundation, indoor and open-air scenarios, and per human pixel reduction with the expansion of crowd [11], [12]. Then again, challenges involved in crowd behavior detection or detection of an unusual object include poor quality with moving surroundings, displaying the group conduct, impediment between people, and irregular fluctuations of a crowd. Consequently, detecting unusual anomalous objects in a real-world environment demands robust knowledge of surrounding environments in a context, structural behavior, and the identification of every object present in the scene [13], [14].

This paper proposes a new robust approach for multi-people tracking and abnormality detection by identifying unusual objects via conditional random field (CRF) and deep learning by understanding surveillance videos. In our work, the extracted frames are initially passed through necessary prepossessing steps. Super-pixels-based multi-object segmentation is performed using an improved watershed transform (IWT) technique and conditional random field (CRF). Next, the Social force model is computed via irregularity strength. However, irregularity strength is calculated Using a fusion of a novel inner force model and Chosen Particular Histogram of Optical Stream (CPHOS). After that, using the computed social force, tracking is performed via three-dimensional association using percentile rank and non-maximal suppression. In the fourth block, multi-object categorization is performed via Feature Pyramid Network, a deep learning model. Finally, Jaccard similarity highlights the unusual objects from the scene and performs anomaly detection.

The significant contributions and features of our thesis are presented below.

- We proposed a robust Improved Watershed Transform (IWT) technique for super-pixel generation. Using a semantic segmentation approach, we used the Conditional Random Field (CRF) to perform pixel-labeling for object segmentation.
- The Social Force model is introduced for extracting the contextual structure of the scene via irregularity strength using a fusion of the inner force model and movement variation utilizing the novel Chosen Particular Histogram of Optical Stream (CPHOS).
- Multi-people tracking technique is introduced via three-dimensional temporal association by utilizing percentile rank and non-maximal suppression.
- A real-time anomaly detection framework is proposed to make the manual monitoring systems as an intelligent automatic surveillance system.
- In our work, we integrate the crowd tracking system and crowd abnormality system in a single model and track multiple individuals by identifying the unusual objects present in the scenario.
- A Jaccard similarity is used to make the decision for anomaly detection and spot the unusual objects based on the contextual structure of the environment. A comparative analysis is conducted on openly accessible benchmark datasets: UCSD Ped 1, Ped 2, and CUHK Avenue dataset for multi-people tracking and abnormality detection.

The remainder of the article is structured into VI sections as: Section II presents related work in the field. Section III contains a definite prologue to the proposed technique and examines all means and levels for multi-people tracking and anomaly detection. Section IV obliges the graphical and tabular results for our experimentations and illustrates our system's comparative study with currently established state-of-the-art systems. Section V demonstrates the discussion part. Finally, Section VI outlines the future direction and conclusion of our proposed methodology.

II. RELATED WORK

In current history, various scholars have put forth alternative techniques for crowd-tracking and anomaly identification [15], [16]. We have discussed current research in this area and the efforts of many scientists who put their efforts into improving crowd-tracking and anomaly-detection systems. We segregate our literature review into two sections. In the first section, we threw light on recently developed crowd-tracking methodologies. At the same time, the second section describes current systems that have been developed for crowd anomaly detection.

A. MULTI-PEOPLE TRACKING SYSTEMS

Various researchers [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] have developed a series of models and procedures in recent years to track multiple people in crowded environments. For instance, In [27] Ristani et al. detects peoples using an off-the-shelf person detector. Appearance and motion features are extracted using a convolutional neural network. Finally, multiple targets are tracked by eliminating

missing detection through post-processing. However, systems accuracy decreases in significant motion, occlusion, and pose changes. Liu et al. in [28] designed a fog computing module where the algorithm for object tracking is performed. The objects are forwarded to the fog node for tracking. Optical flow is utilized to measure the target's speed and a template update strategy is used to store the alternate template to track the objects. Due to fog computing, related technologies and resources are required to implement the system.

In [29], Le et al. detect multiple people using a convolutional neural network. A filter was initialized and correlated on the next sequence frame to track detected people. Also, the creation of a tracking path for every individual in the data association is utilized. However, the system has an expensive detection module that may lead to misdetection because of the long processing time. In [30], Chahyati et al. performed object detection via retinaNet and then tracked the individuals using the Hungarian algorithm. After detection, the people are associated in a sequence of frames using the Hungarian algorithm. They did not fine-tune and train the detector, limiting the system's Accuracy.

Gochoo et al. in [31] present a system for tracking individuals via head and shoulder detection using hough circular gradient transform and HOG-based symmetry. Finally, a one-dimensional convolutional neural network is utilized for tracking the multiple detected people. In any case, the framework proficiency declines with light changes likewise in complex impediments, particularly in line position the framework might create bogus identifications. In [32], Pervaiz et al. used thresholding for foreground extraction. The detected humans are first counted using the centroid of each human and then the Jaccard similarity index is used to track multiple humans. The system removed the humans close to border areas during the foreground extraction process by considering them as part of the background. Also, in occlusion, the humans are not detected accurately, decreasing the system's efficiency.

Ren et al. in [33] present multi-people tracking by counting model. Density maps are utilized for object detection. They incorporate flow constraints and data association cost with the object counter constraints and track the humans by counting. The system's Accuracy on large-scale datasets can be further improved. In [34], Pervaiz et al. performed template matching on detected objects for verification of humans. A self-organizing map is utilized to group the particle flows. Motion trajectories' are then used to track the multiple pedestrians. However, the system is inefficient in complex, crowded scenes due to size-based object detection in foreground extraction.

B. METHODS FOR ANOMLAY IDENTIFICATION IN CROWDS

Various scholars have devoted their endeavors to foster frameworks for anomaly detection utilizing various innovations [1], [19], [35], [36], [37], [38], [39], [40], [41]. For instance, In [42], Zhou et al. designed a sparse-coding-based anomaly detection model by using deep learning. They detect anomalies by extracting features and motion trajectories using motion and appearance connections. At the same time, long short-term memory of sparse coding is utilized as a coding block. However, the system requires large data for training and is computationally complex. In [43], Nawaratne et al. present an unsupervised deep learning approach using incremental spatio-temporal learning for localization and detecting anomalies. They utilized fuzzy aggregation and active learning for the detection of new anomalies that occurred with the passage of time. Sparse evaluation is overcome using temporal and anomaly thresholds depending on the context.

Moustafa et al. in [44] introduced two crowd behavior detection model streams. First, scene-dominant trajectories are extracted by using a meta-tracking procedure. Secondly, long short-term memory is utilized to define a predictive model for each super region. Finally, an anomaly matrix is formed with the predictive model to detect anomalies. In [45], Hassanein et al. identified the motion pathways using tracklet clustering by utilizing a distance-dependent Chinese restaurant process strategy, also used in two hierarchical levels to identify motion pathways. These pathways are further examined to detect abnormal behavior in both frame and tracklet levels.

In [46], Rehman et al. detect anomalies using both acoustic and visual features. An inference system for detecting anomalies is built using audio and visual features and a support vector machine makes a final decision. The system accuracy decreases for indoor scenes. In [47] Sun et al. used a deep one-class learning model to detect the anomalies. The loss function is established by utilizing one-class SVM to optimize model parameters. Finally, combining a convolutional neural network and one class SVM is used to differentiate between normal and abnormal scenes. However, the system is computationally complex and requires extensive training.

Bansod et al. in [14] first remove the background using the thresholding technique on optical flow. Position features via the appearance of objects are utilized for the localization of anomalies. The final decision is made by k-means clustering algorithm. The use of a threshold in background removal may affect the performance in case of illumination changes and dense crowds. In [48] Zhang et al. grouped the crowd consistency via scene perception clustering and segments the moving pedestrians by designing line integral convolution. A one-class support vector machine makes a Final decision for the detection of abnormal crowd behavior. The adaptability of the method needs to be further improved in different environments. In [49], Wu et al. extract the Spatio-temporal and appearance features. These features are used to train the denoising autoencoder for detection. The concatenation of shapely additive explanation and autoencoder is used to detect abnormal crowd behavior.

III. PROPOSED SYSTEM FRAMEWORK

This paper introduces a novel approach for multi-people tracking and detecting anomalous objects among multi-pedestrians via conditional random field (CRF) and deep learning by understanding surveillance videos. The proposed system subsists of the following main steps. Initially, the pre-possessing step is performed in which video data is converted into frames, and filtration is performed to efficiently reduce noise by protecting edges. After that contrast adjustment is implemented. In the second block, super-pixels-based multi-object segmentation is performed. Super-pixels are generated using an improved watershed transform (IWT) technique. After that, a conditional random field (CRF) is implemented for pixel-level labeling. In the third block, the Social Force model is introduced via irregularity strength for extraction of the contextual structure of the environment. However, irregularity strength is computed using a fusion of newly introduced inner force between particles and the movement variation of people computed using Chosen Particular Histogram of Optical Stream (CPHOS). After that, computed social force tracking is performed via three-dimensional association using percentile rank and non-maximal suppression. In the fourth block, multi-object categorization is performed via Feature Pyramid Network, a deep learning model. Finally, in the last block, Jaccard similarity is utilized to highlight the unusual objects from the scene and perform anomaly detection. The schematic diagrams of our proposed framework are illustrated in Fig. 1. The next subsections discuss each of the aforementioned modules in further detail.

A. PRE-PROCESSING

In pre-processing [50], [51], [52], [53], captured videos are initially transformed into frames $[f_1, f_2, f_3, ..., f_z]$. where z is the total number of frames. All the extracted frames are passed through an adaptive median filter to effectively remove noise, blurriness, and distortion by protecting edges. The principal benefit of using an AMF is that it produces superior results because the size of the kernel around a distorted frame changes. Another major benefit of an AMF is that, in contrast to a median filter, it refrains from converting every pixel value to its median value. Corrupted and uncorrupted pixels are segregated in the filtering window in an adaptive median filter. After that, the filtering technique is implemented in the window upon corrupted pixels. All the noisy pixels in the filtering window are traded with the average worth of the pixels. The adaptive median filter works in two phases, considers each pixel in the image in relation to its neighbor and organizes the pixels into a cacophony via spatial handling. Boisterous pixels are displaced by the center pixel worth of the pixels nearby, which has easily passed a clamor marking assessment, while uncorrupted pixels remain the same. By performing this, one may guarantee that just the pixels containing impulse noise are altered and all other pixels remain unchanged. Algorithm 1 is described the process of an AMF.

Algorithm 1 Frames Filtering via AMF

Input: Input frames

Output: Frames after removal of noise, blurriness, and distortion by protecting edges.

F_{uv} = filter window
I_s=minimum intensity value in F_{uv}
I_l= maximum intensity value in F_{uv}
I_d= median of intensity value in F_{uv}
I_{uv}= intensity value at coordinate (u, v)
F_l = maximum allowed size of F_{uv}.

Stage 1:

$Y_1 = I_d - I_s$
$Y_2 = I_d - I_l$
If $Y_1 > 0$ AND $Y_2 < 0$, go to stage 2
Else increase the window size
If window size $> F_l$ Output I_d
Else repeat stage 1

Stage 2:

$Z_1 = I_{uv} - I_s$
$Z_2 = I_{uv} - I_l$
If $Z_1 > 0$ AND $Z_2 < 0$, output I_{uv}
Else output I_d

Return: Filtered frames

The filtered frames are subjected to histogram equalization for contrast adjustment following filtering. Histogram equalization spreads the very recurrent intensity values stretching out the range of intensity in an image. The global contrast of an image is usually enhanced by this technique when usable data is in close contrast values. After histogram equalization, frame areas with lower local contrast will gain a higher contrast. We apply the histogram equalization on the filtered frames using Eq. 1, as illustrated in Fig. 2.

$$R_l = S(C_l) = (T-1)\sum_{i=0}^{l} p_C(C_i) \tag{1}$$

where C signifies the forces of an input picture to be handled, and $l = 0, 1, 2, \ldots, (T-1)$. While R represents the output intensity level after intensity mapping for every pixel in the input image, having intensity C. However, $p_C(C)$ is the probability density function (PDF) of C.

B. SEGMENTATION OF MULTIPLE OBJECTS

The pre-processing step is followed by the deployment of conditional random field-based semantic segmentation for multi-object segmentation [54], [55], [56], [57], [58], [59]. Object segmentation is a technique for breaking up a computerized image into a smaller grouping called image segments, diminishing the frames' intricacy and empowering further handling or investigation of each frame fragment. Segmentation is giving labels to pixels to recognize items, individuals, or other significant components in the frames. In any case, preceding segmentation, we first build an improved watershed transform (IWT) technique for producing the super-pixel image. Super-pixel creation is a preprocessing

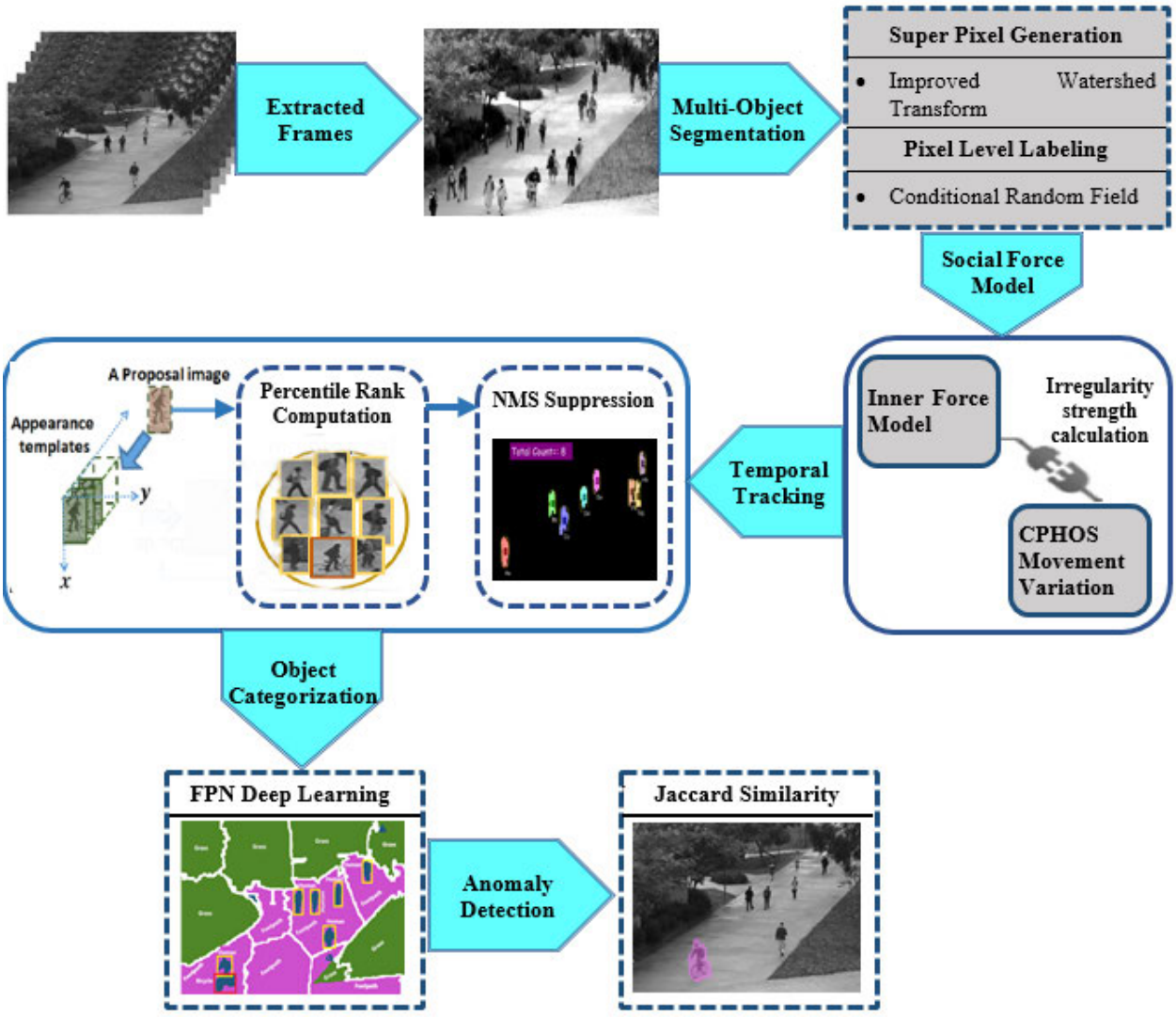


FIGURE 1. The overall architecture of the invented crowd anomaly identification and pedestrians tracking model.

method for semantic segmentation that divides a frame into numerous small sections.

Traditional WT has low computational intricacy and delivers a sporadic super pixel region that is more powerful when contrasted with the hexagonal zones created by the Simple linear iterative clustering (SLIC) [60], [61], [62]. Nonetheless, as WT is susceptible to noise and may lead to over-segmentation, we further improve it by applying morphological processes delineated in Eq. 2 to conquer this issue.

$$\begin{cases} M^{O}(r) = M^{d}(M^{e}) \\ M^{c}(r) = M^{e}(M^{d}) \end{cases} \tag{2}$$

where r signifies structural element and M^{O} and M^{c} stand for morphological opening and closing, M^{e} and M^{d} addresses morphological erosion and dilatation, respectively. We apply morphological closing and opening characterized in Eq. 2 by utilizing multi-scale structural elements to get numerous reconstructed images by reconstructing the gradient image. M^{O} and M^{c} can decrease over-segmentation by eliminating region minima in gradient images. Pointwise maximums are calculated from reconstructed pictures to create a gradient picture that disposes of futile neighborhood minima by saving edges. Fig. 3 illustrates the sample frames for super pixel generation.

Following super-pixels' creation, a conditional random field (CRF) is carried out in this stage for doling out labels to every pixel [63], [64], [65], [66]. A CRF is a discriminative factual demonstrating strategy that is utilized when the class labels for various variables are not independent. For instance, during picture segmentation, the class label for a pixel is also influenced by the labels of the pixels nearby. The four types of semantic features are extracted from each super pixel region for efficient semantic segmentation: Color Features, Texture Features, SIFT Features, and Adjoint Marking, as expressed in Algorithm 2. Maximum Likelihood is applied to the data

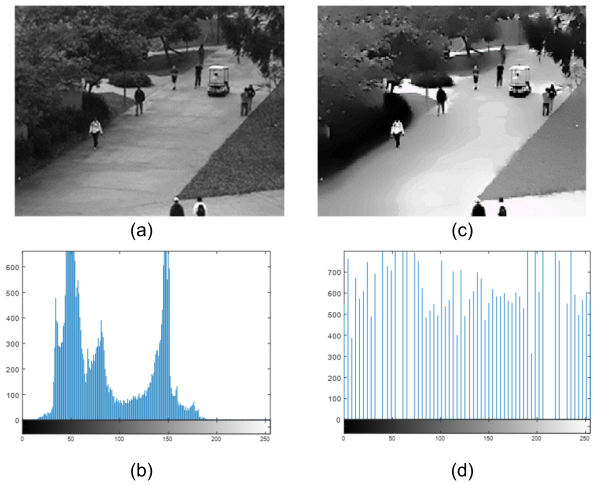


FIGURE 2. Pre-processing of input frames. (a) AMF based filtered frame, (b) Filtered frame histogram, (c) Pre-processed frame, and (d) Equalized histogram of pre-processed frame.

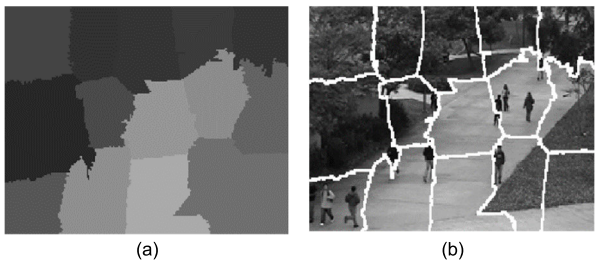


FIGURE 3. Super pixel generation steps. (a) Morphological-based WT segmentation, and (b) IWT-based Super-pixel generation.

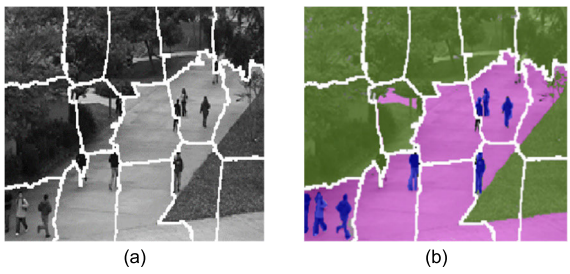


FIGURE 4. Semantic segmentation. (a) Super-pixel generation via IWT, and (b) CRF-based semantic segmentation.

for training purposes. The model's variables that generate the training data and maximum probability, are selected. To predict the right label for each object, a CRF model is trained on a training sample with conditional log-likelihood maximized. That makes the CRF model anticipate the right label for each object. Some samples are shown in Fig. 4.

C. CONTEXTUAL STRUCTURE EXTRACTION VIA SOCIAL FORCE MODEL

An innovative Social Force Model (SFM) is introduced to reflect the structure of every person in a scene. Our underlying premise for the social force modeling technique is that people

Algorithm 2 Extraction of Semantic Features

Input: Super pixel generated frames
Output: Semantic Feature Vectors
$\mathbf{S_1}$= ExtractColorFeatures ()
$\mathbf{S_2}$= ExtractTextureFeatures ()
$\mathbf{S_3}$= ExtractSIFTFeatures ()
$\mathbf{S_4}$= ExtractAdjointMarkingFeatures ()
$\mathbf{V_i}$= EstablishedSemanticFeatureVector s_i
For every s_i in feature, **do**
{
Merged = (s_i, s_{i+1})
}
return Feature vector containing semantic features

exhibiting significant behavioral divergences with their environmental elements are profoundly likely to be abnormal. An irregularity strength estimates the disparity between the inspected object and its environmental elements. The heavier the strength, the more inspected object behavior will be more obvious with their environmental elements. For calculating the social force of all the individuals present in this scene via irregularity strength computation, the fusion of inner force model and movement variation via Chosen particular histogram of an optical stream is utilized. In the accompanying, the inner force of particles is introduced first. Afterward, the strategy of Chosen particular histogram of an optical stream is given, trailed by the social force calculation.

In strong physical science, the potential energy of two particles addresses their connecting strength. We initially turn each segmented object into a particle before presenting a reliable Inner Force Model (IFM). Each segmented object turns into an assortment of numerous particles since every pixel was seen by us as a fluid element. We determine the potential energy of particles on object forms for ease of use, and from that main, those two particles having the greatest possible energy are considered. For figuring the force among two particles, connecting strength of two particles is characterized using Eq. 3.

$$C\,(\mathrm{k}) = \frac{1}{|F\,(k)|}/(\int_{\sqrt{2}}^{k} \frac{1}{|F\,(k)\,|}dk), k \in [\sqrt{2}, \infty] \quad (3)$$

where $C(k)$ indicates the connecting strength of two particles and k is the Euclidean distance among two particles. While F (k) represents the joining power. Fig. 5 illustrates the visual representation of the inner force model.

The movement variation of people ought to be effectively processed. Concerning the movement property, optical stream is registered to feature the movement of each and every person. Histogram of the stream is determined as the movement measurements where each receptacle of the histogram addresses the direction of the stream and the worth in each container corresponds to the magnitude of a stream. We find that the movement direction and the magnitude keep up reliability for a particular person. Hence in our work,

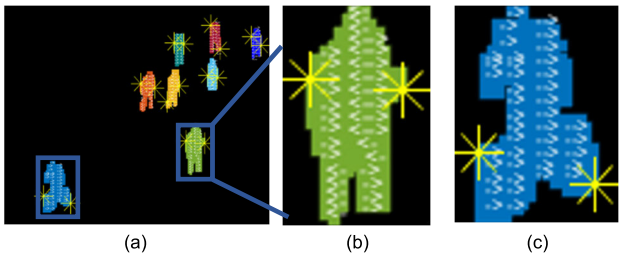


FIGURE 5. Model of inner force. (a) Inner force acting among two particles, (b & c) Amplified perspective on Internal force.

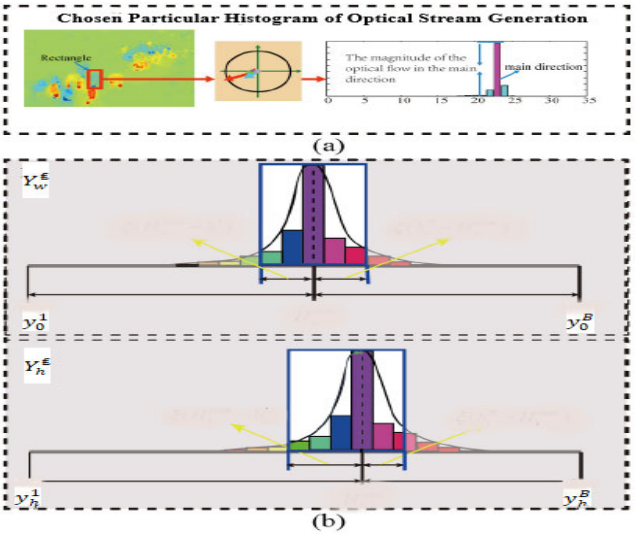


FIGURE 6. Chosen particular histogram of optical stream generation steps. (a) Extraction of the optical stream histogram, and (b) Particular histogram of an optical stream that is chosen by the system.

we chose a particular histogram of the optical stream to address the movement properties of people. Chosen particular histogram of the optical stream is a restricted histogram of an optical stream by a boundary £ which decides the scope of the histogram choices that should be utilized for movement variation calculation. The ideal £ is achieved by the movement variation of people in the normal frames and figured as in Eq. 4.

$$£' = \arg\min_{£}[diff\ (\dot{\mathrm{d}}S)] \tag{4}$$

where $diff\ (\dot{\mathrm{d}}S)$ is the fluctuation estimation. The addressing of above equation can be satisfied by looking through the entire reach [0,1] of £ from 1 to 0 with a 0.1 span. Since the start of the video succession is typical with ordinary circumstances, hence £′ can be achieved in the beginning, once £′ is created, it can keep up unaltered for the remaining video. Fig. 6 illustrates the generation of the Chosen particular histogram of an optical stream model.

After the inner force model and calculation of the movement variation of every individual, the social force is computed via irregularity strength as in Eq. 5 using a fusion of the inner force model and movement variation chosen particular histogram of an optical stream. Hence, in our work, while figuring out the social force of the analyzed individual, the sensible thought is that every one of the others in the environmental factors together adds to its entire Social Force. Consequently, the standardization interaction ought to be based on the quantity of the encompassing people, and the itemized articulation is reconsidered.

$$IW(\dot{\mathrm{d}}S_h) = \frac{1}{s\left(B\dot{\mathrm{d}}S_h\right)}/\left(\sum_{h=1}^{N}\frac{1}{s\left(B\dot{\mathrm{d}}S_h\right)}\right) \tag{5}$$

where B is consistent with expanding $\dot{\mathrm{d}}S$ to a sufficient boundary and $IW(\dot{\mathrm{d}}S_h)$ signifies the irregularity strength among the inspected person and the i^{th} surrounding persons, and N is the quantity of surrounding individuals. Hence the social force for every individual is computed to extract the contextual structure of the environment via the irregularity strength.

D. CROWD TRACKING VIA TEMPORAL ASSOCIATION

The next step is to track several people by studying the spatial-temporal Social Force fluctuations using the calculated Social Force for every person inside the crowd, which also aids in the discovery of abnormalities. Hence in this section, we performed temporal tracking through association. Subsequently, for this reason, we initially created the template database containing three-dimensional volume templates of individuals in different stances from the prior frames. A separate template database is built sequentially for each individual in a video. To preserve each individual's precise appearance and gain great associates for coming frames, all template databases must be refreshed repeatedly and periodically while using a restricted stack capacity. We processed the percentile rank utilizing Eq. 6 for a precise multi-object temporal association by using probability and semantic data.

$$S_{i,j} = \max_{j}\{K_{i,j} * \rho(£\,(q_i, ϩ))\} \tag{6}$$

where i^{th} template associates with the j^{th} template database if the percentile rank S is greatest, while K is the semantic distance, £ is the probability capability, ρ is the sigmoid function and ϩ is the scaling parameter.

Nonetheless, several of the created suggestions are overlapping and iterative. Therefore to maintain satisfactory pairings between conflicting overlapping templates, we applied non-maximal suppression over the percentile rank. The ultimate detections are the templates that were still present after the non-maximal suppression. we ultimately eliminate the template database that continues to be not associated with any individual within $\dot{\mathrm{d}}S$ time. Along these lines, we adjust to dependability safeguarding of undoubted appearance of an object. In our tracking module, we highlight every individual with a box with an associated ID at the base right, as shown in Fig. 7.

E. OBJECT CATEGORIZATION

For categorizing objects, we apply a Feature Pyramid Network (FPN), a deep learning model for multi-object categorization [67], [68], [69]. FPN distinguish objects with

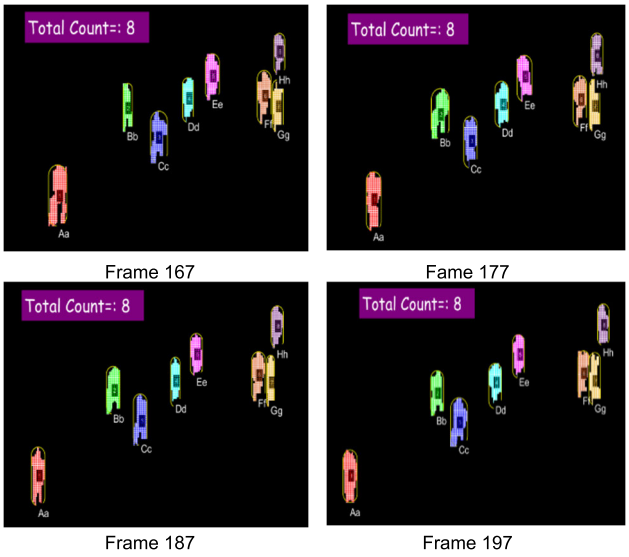


FIGURE 7. Example frames of pedestrians tracking outcomes at various time stretches over the UCSD Ped 1 dataset.

different scale utilizing hierarchical and base-up features with literal links. In contrast to inserting a layer within FPN, we employed a concatenation layer in our work. Initially, characteristics are extracted at various scales in FPN. Five categorization strands, all with two neurons, first highlights are separated at various scales is utilized in the FPN which applies relu enactment capability. Eventually, we linked these layers to make a thick layer and associated the last characterization layer, which utilizes the softmax capability. In our work, we create a feature pyramid with four levels. We signify the result of the backbone as (S_2, S_3, S_4, S_5), which have steps of {4,8,16,32} pixels for the source picture, the characteristics (C_2, C_3, C_4) have identical decreased channels of 256 after 1×1 convolution. The feature pyramid (N_2, N_3, N_4) is created by the top-down route. We eliminate the terminals C_5 and N_5, the basic greatest features having semantic information for FPN.

We presented Sub-pixel omit Combination to straightforwardly up-sampling the minimum quality picture without diverting decrease. The detailed channel data of S_4, S_5 are utilized and blended with C_3, characterized as shown in Eq. (7).

$$C_x = \begin{cases} \mathfrak{Q}(S_x) + BL(\mathfrak{Q}'(S_{x+1}), & x = 3, 4 \\ \mathfrak{Q}(S_x), & x = 2 \end{cases} \tag{7}$$

where $\mathfrak{Q}$ means 1×1 convolution to decrease channels, and x demonstrates the file of pyramid levels, $\mathfrak{Q}'$ signifies channel change and performs character mapping. It embraces 1×1 convolution or division activity to modify the channel aspects for twofold sub-pixel up-sampling. Furthermore, let channel aspects meet the criteria. Then, at that point, the featured pyramid N_x are delivered by C_x through component-wise summation and closest neighbor up-sampling. Sub-pixel omit Combination should be visible as an additional two associations from S_x to C_3 and S_5 to C_4. It processed upsampling

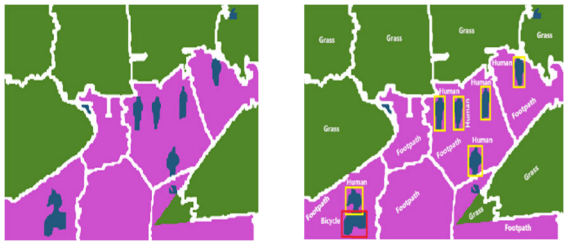


FIGURE 8. Object categorization. (a) CRF-based semantic segmentation, (b) deep learning-based object categorization.

and a combination of the channel at the same time, which uses the huge channel data of up-level features (S_4, S_5) to improve the portrayal capacity of the feature pyramid.

The channel-focused directed mechanism utilizes the integration map I to derive channel strengths. And afterward, every resultant characteristic is amplified by the channel strengths. Initially, we separately use worldwide normal pooling and worldwide maximum pooling to combine two distinct geographic contexts. Afterward, the two descriptors are sent to completely associated layers individually. At last, the resultant selected characteristics are converged through component-wise summation and a sigmoid capability. The sole purpose of channel-focused directed mechanism is to lessen the deceptive effects of ghosting characteristics and to upgrade more discriminative capacities of characteristics. The results of multi-object categorization using the Feature Pyramid Network are shown in Fig. 8.

F. ANOMALY DETECTION VIA JACCARD SIMILARITY

After categorizing the objects and highlighting the areas of interest, we use Jaccard Similarity (JS) to assess the similarities among the classed objects in light of context to identify anomalies in the frame [70], [71]. As for our scenario, it will be recognized as an abnormal scene if the individual is riding a bicycle, scooter, automobile, wheelchair, skater, or any other transport among a crowd of walkers on a pathway. In the object categorization phase, the recently updated information in the log record is gathered as the detection information. In order to evaluate whether the recent actions are unusual or regular, the feature vectors of the observed behavior are examined to those of the normal and unusual behavior using the detection data of every super pixel.

The generalized Jaccard similarity coefficients are generated and utilized to contrast the ongoing movement characteristics with the ordinary and unusual movement characteristics to identify if an anomaly has happened. In our framework, some of the unusual example sets incorporate (pedestrians, bikes, pathways), (walkers, bikes, roads), and so on. If the ongoing way of behaving is normal, then to guarantee the continuous and exactness of the ordinary conduct model, the typical highlight vector ought to be refreshed; on the off chance that the ongoing way of behaving is unusual,

(a) (b)

FIGURE 9. Example frames of anomaly identification at various time stretches over the UCSD dataset.

FIGURE 10. Sample frames of different scenarios of the UCSD Ped 1 dataset.

the strange handling will be completed to produce an alert and highlight the anomalous object.

During the database training process, the vectors of anomalous behavior are created. The mechanism used in the training process is also utilized in the anomaly detection phase, which processes the discovered stream of data first. At long last, the vector of the identified information is framed. Then, at that point, the vector of the identified information is contrasted and that of the unusual behavior utilizing the generalized Jaccard similarity. And the resulting comparison outcome $Sim = (x, y)$ is utilized as a foundation for determining whether the ongoing behavior is abnormal or not. This Jaccard index mirrors the comparability between two selected sets. Let A be the main set and B the subsequent set, the comparability among A and B as indicated by the Jaccard index, is registered as shown in Eq. (8).

$$J(A, B) = |A \cup B| / |A \cap B| \tag{8}$$

If the list is near 1, the two sets are practically the same, and if it is near 0, they are considered extremely different. Fig. 9 depicts the sample frames for anomaly detection.

IV. EXPERIMENTAL SETTINGS AND ANALYSIS

This part expounds on all trials performed to approve the proposed framework. All experimentations and processing is done by utilizing a Leave One Out Cross Validation (LOOCV) technique on two openly accessible benchmark datasets: UCSD Ped 1 dataset and the CHUK Avenue dataset. Each dataset is partitioned into N subgroups, each containing k number of pictures. To start with, every one of the subgroups is utilized to prepare the framework, and afterward, one subgroup is utilized for testing. The framework is then verified by utilizing one more subgroup for testing and the leftover subgroup for preparation. The pictures of the subgroup that are utilized for preparing are excluded from the testing set. The trials and evaluation are carried out entirely using Google Colab (Python) and MATLAB resources. The equipment framework utilized was Intel core i5-6200U with a processor clocked at 2.40 gigahertz, having 8 gigabytes of random access memory, and a two giga bytes Nvidia dedicated graphics card.

In this section, the investigations are divided into four categories. We assess the effectiveness of object segmentation in the initial segment. In the subsequent segment, the effectiveness of multiple tracking is assessed. Ultimately, the robustness of the crowd abnormality detection is assessed in the third part. Lastly, we contrast our proposed framework and other already introduced state-of-the-art systems. Initially, in this section, we briefly first through light on datasets that we utilized for evaluation purposes and then we discussed efficiency matrices along with outcomes that we obtained for object segmentation, multi-people tracking, and anomaly detection.

A. DATASETS DESCRIPTION

We employed two publicly viewable benchmark datasets for crowd-tracking and anomaly detection: the UCSD Pad1 and the CUHK Avenue datasets. Details of each dataset are mentioned in the following subsections.

1) UCSD PED 1 DATASET

The UCSD Ped 1 dataset gives videos of individuals on common walkways at the University of California San Diego. There are 34 practice videos and 36 test videos in the Ped1 dataset. Each video is captured with a resolution of 238 × 158 pixels at a 30 FPS frame rate, having 200 frames in total. The density of the crowd ranges from sparse to crowded. Figure 4.1 illustrates some sample images of the UCSD Ped1 dataset. The ground truth information of the UCSD pedestrian dataset is available at [72].

2) CUHK AVENUE DATASET

The CUHK Avenue dataset records pedestrian movements at the Chinese University of Hong Kong (CUHK) [73]. The dataset was obtained utilizing a fixed camcorder with a quality of 360 × 640 pixels and a 25 FPS frame rate. There are 15 fragments in the dataset, so each fragment lasts approximately two minutes. There are 16 practice videos and 21 test videos captured at the CUHK campus and included in the dataset, which has been switched over completely into 15,328 training frames and 15,324 practice frames. There are 30,652 frames in all. Fig. 11 illustrates some sample images of the CUHK Avenue dataset.

B. PERFORMANCE METRICS AND RESULTS

We utilized five assessment measurements to gauge how well our presented method performed. The efficiency of object

FIGURE 11. Sample frames of different scenarios of the CUHK Avenue dataset.

TABLE 1. Confusion matrix for the proposed object segmentation method accuracy over the CUHK avenue dataset.

Class	hm	gr	bc	rd	cr	tr	fp	sk	wc	pl	bl
hm	**0.99**	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
gr	0.00	**0.96**	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00
bc	0.02	0.00	**0.98**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rd	0.00	0.01	0.00	**0.96**	0.00	0.00	0.02	0.01	0.00	0.00	0.00
cr	0.00	0.00	0.01	0.00	**0.97**	0.00	0.01	0.00	0.01	0.00	0.00
tr	0.00	0.02	0.00	0.00	0.00	**0.96**	0.00	0.01	0.00	0.01	0.00
fp	0.00	0.02	0.00	0.02	0.00	0.00	**0.95**	0.01	0.00	0.00	0.00
sk	0.00	0.01	0.00	0.01	0.00	0.02	0.00	**0.96**	0.00	0.00	0.00
wc	0.02	0.00	0.02	0.00	0.03	0.00	0.00	0.00	**0.93**	0.00	0.00
pl	0.02	0.00	0.01	0.00	0.00	0.03	0.00	0.00	0.00	**0.92**	0.02
bl	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	**0.98**
Mean Accuracy = 96.03 %											

*hm = humans, gr = grass, bc = bicycle, rd = road, cr = car tr = trees, fp = footpath, sk = sky, wc = wheelchair, pl = pillar/poles and bl = buildings

segmentation is presented using a confusion matrix. However, four assessment measurements are utilized to assess the efficiency of multi-people tracking and anomaly detection: Accuracy, precision, recall, and F_1 score [74], [75] expressed as.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{F}_1\text{score} = 2\times(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}) \tag{12}$$

In this context, *TN* addresses true negative, *TP* stands for true positive, *FN* stands for false negative, and *FP* is a false positive. In any case, the F1 score is the extent of test exactness. While a level of the total count of the care plan is implied as Accuracy, the level of genuine positives that are delegated anomalous is alluded to as recall. However, precision alludes to the closeness of the estimations to one another.

TABLE 2. Confusion matrix for the proposed object segmentation method accuracy over the CUHK avenue dataset.

Class	hm	gr	bc	rd	cr	tr	fp	sk	wc	pl	bl
hm	**0.95**	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.02	0.00
gr	0.00	**0.96**	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00
bc	0.02	0.00	**0.92**	0.00	0.02	0.01	0.00	0.00	0.02	0.01	0.00
rd	0.00	0.02	0.00	**0.94**	0.00	0.00	0.03	0.01	0.00	0.00	0.00
cr	0.00	0.00	0.01	0.00	**0.93**	0.01	0.00	0.00	0.03	0.01	0.01
tr	0.00	0.02	0.00	0.00	0.00	**0.96**	0.00	0.01	0.00	0.01	0.00
fp	0.00	0.02	0.00	0.03	0.00	0.00	**0.94**	0.01	0.00	0.00	0.00
sk	0.00	0.01	0.00	0.01	0.00	0.02	0.00	**0.95**	0.00	0.00	0.01
wc	0.03	0.00	0.02	0.00	0.03	0.01	0.00	0.00	**0.91**	0.00	0.00
pl	0.02	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.00	**0.94**	0.01
bl	0.00	0.01	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.01	**0.94**
Mean Accuracy = 94 %											

*hm = humans, gr = grass, bc = bicycle, rd = road, cr = car tr = trees, fp = footpath, sk = sky, wc = wheelchair, pl = pillar/poles and bl = buildings

Inside this part, we initially assess the exhibition precision of semantic segmentation for the segmentation of multiple objects over two openly accessible datasets. Table 1 illustrates the Accuracy of our proposed multi-object segmentation method over the UCSD Ped 1 dataset via confusion matrix as a color function with rows showing the predicted tags and columns showing the true tags.

Table 2 depicts the Accuracy of our proposed multi-object segmentation method over the CUHK Avenue dataset using a confusion matrix via color function with rows showing the predicted tags and columns showing the true tags.

1) EXPERIMENT 1: PERFORMANCE ACCURACY FOR TRACKING OF PEDESTRIANS

In this experiment, we assess the performance of our proposed multi-object tracking model using efficiency measurements i.e., Accuracy, precision, recall, and F_1 score on two openly accessible datasets, i.e., UCSD Ped 1, Ped 2, and CUHK Avenue datasets. Table 3 addresses the mean Accuracy for tracking pedestrians alongside recall and F_1 score across 70 sessions of the UCSD Ped 1 dataset, with each session consisting of 200 frames.

Table 4 demonstrates the efficiency of our proposed multi-people tracking model by presenting recall and F_1 score along with mean Accuracy across 70 sessions of the CUHK Avenue dataset, with each session consisting of 200 frames.

2) EXPERIMENT 2: PERFORMANCE ACCURACY FOR IDENTIFICATION OF ANOMALOUS OBJECTS

In this analysis, we survey the vigor of our crowd abnormality identification framework by utilizing four assessment measurements above over UCSD Ped1, and CUHK Avenue

TABLE 3. UCSD Ped 1 dataset readings of accuracy, recall, and F1 score for tracking of pedestrians.

Session No (200 frames)	Ground Truth	TP	FP	FN	Accuracy	Recall	F_1 score
14	8	8	0	0	1	1	1
28	13	12	0	01	0.923	0.923	0.959
42	17	16	0	01	0.941	0.941	0.969
56	20	18	01	01	0.90	0.947	0.946
70	26	22	01	03	0.846	0.880	0.916
Mean Accuracy= 92.2 %							

TABLE 4. CUHK avenue dataset readings of accuracy, recall, and F1 score for tracking of pedestrians.

Session No (200 frame)	Ground Truth	TP	FP	FN	Accuracy	Recall	F_1 score
14	07	07	0	0	1	1	1
28	12	11	0	01	0.916	0.916	0.956
42	15	13	01	01	0.866	0.928	0.927
56	19	16	01	02	0.842	0.888	0.913
70	24	20	02	02	0.833	0.909	0.908
Mean Accuracy= 89.1 %							

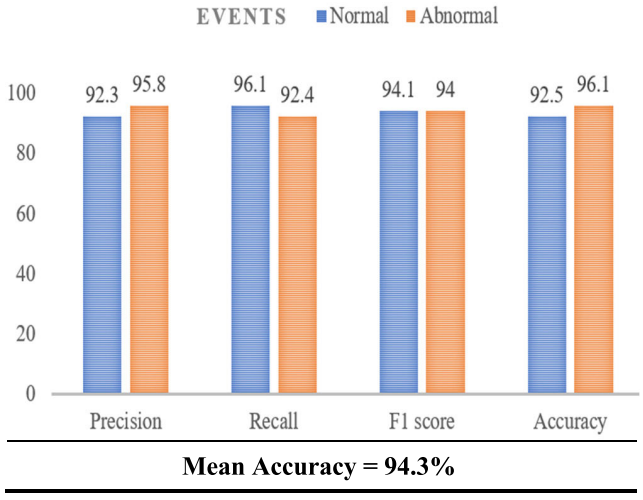


FIGURE 12. UCSD Ped1 dataset Precision, Recall, F_1 score and Accuracy for identification of anomalous objects in crowd.

benchmark datasets. The exhibition estimations for the identification of anomalous objects in the crowded environment over the UCSD Ped1 dataset is illustrated in Fig. 12.

Fig. 13 demonstrates the productivity of our invented framework for identifying anomalous objects in a crowded scene along with precision, recall, F_1 score and accuracy efficiency measurements upon the CUHK Avenue dataset.

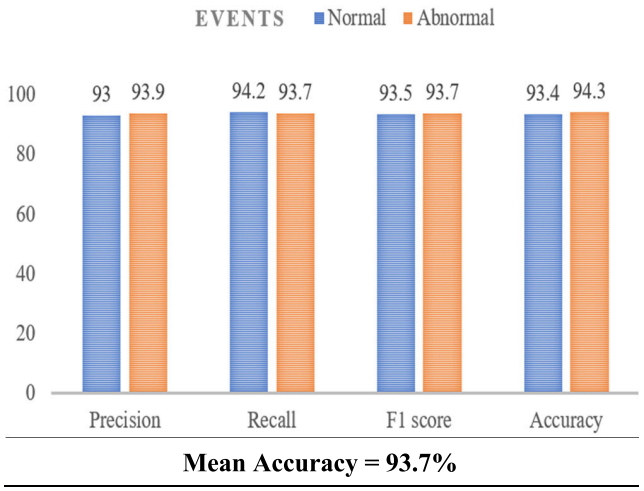


FIGURE 13. CUHK Avenue dataset Precision, recall, F_1 score and Accuracy for identification of anomalous objects in crowd.

TABLE 5. Anomaly detection comparison of the invented framework with best in class techniques upon UCSD Ped1 dataset.

Methods	Average Accuracy (%)
W. Luo [76]	75.5 %
J.T. Zhou [42]	83.5 %
W. Chou [77]	90.9 %
B. Yang [78]	89.1 %
R.T. Ionescu [79]	68.4 %
S.D. Bansod [14]	82.3 %
X. Zhang [48]	90.0 %
C. Wu [49]	85.9 %
Proposed Model	**94.3 %**

TABLE 6. Anomaly detection comparison of the invented framework with best in class techniques upon CUHK avenue dataset.

Methods	Average Accuracy (%)
W. Luo [76]	77.0%
J.T. Zhou [42]	86.1%
W. Chou [77]	82.1%
B. Yang [78]	87.2%
R.T. Ionescu [79]	80.6%
Proposed Model	**93.7 %**

3) EXPERIMENT 3: COMPARISON ANALAYSIS OF OUR INVENTED FRAMEWORK WITH WELL KNOWN EXISTING METHODS

We broke down our suggested framework during this examination by contrasting it with other notable existing strategies. Table 5 demonstrates the comparison between our proposed crowd anomaly detection system with

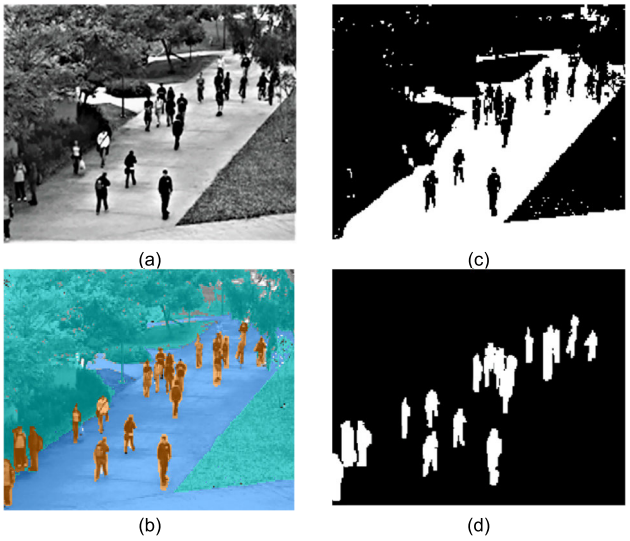


FIGURE 14. Foreground Extraction. (a) original image from UCSD dataset, (b) Foreground extraction using multilevel thresholding by A. Shehzad et al. [76], (c) semantic segmentation class labeling, and (d) Extracted foreground via semantic segmentation.

currently established state-of-the-art methods over UCSD Ped 1 dataset. As depicted, our approach outperformed a best-known current model in terms of accuracy rate.

Table 6 illustrates the comparison analysis of our invented framework with currently established best in class methods upon CUHK Avenue datasets.

V. DISCUSSION

In this study, we presented a sustainable solution with outstanding consistency and conformity against diverse performance difficulties. The difficulty of extracting the foreground from a complicated backdrop in a congested setting presents the problem of object segmentation, as shown in Fig. 14 (a). In preliminary studies, specialists either involved head location or thresholding methods, as shown in Fig. 14 (b) for foreground extraction [80], [81], [82], which restricts the framework precision in light changes and dynamic foundation, hence we performed semantic segmentation for multi-object segmentation as shown in Fig. 14 (c) and extract foreground human silhouettes as depicted in Fig. 14 (d), from a complex background. Object segmentation is a technique for categorizing a computerized image into a smaller grouping called image segments. Segmentation is giving labels to pixels to recognize items, individuals, or other significant components in the frames. In any case, preceding segmentation, we first build an improved watershed transform (IWT) technique. For producing the super-pixel image. Super-pixel creation is a preprocessing method for semantic segmentation that divides a frame into numerous small sections. Following super-pixels' creation, a conditional random field (CRF) is carried out in this stage for doling out labels to every pixel and performing multi-object segmentation.

To analyze the contextual structure of the extracted foreground, an innovative Social Force Model (SFM) is introduced. Our underlying premise for the social force modeling technique is that people exhibiting significant behavioral divergences with their environmental elements are profoundly likely to be abnormal. An irregularity strength estimates the disparity among the inspected object and its environmental elements. For calculating the social force of all the individuals present in this scene via irregularity strength computation, the fusion of the inner force model and movement variation via Chosen particular histogram of the optical stream is utilized. After Social Force computation for every individual, the individuals are tracked temporally in a series of frames. Through temporal tracking of every individual, the discrepancy between the examined target and its surrounding is measured by an irregularity strength.

We performed temporal tracking through three-dimensional associations by creating the template database containing three-dimensional volume templates of individuals in different stances from the prior frames. In track, there are some frames where the silhouettes of one pedestrian either partially or completely overlap the silhouettes of another pedestrian, hence for handling such situations, a percentile rank is introduced such that i^{th} template associates with the j^{th} template database if the percentile rank S is greatest. We processed the percentile rank using probability and semantic data. Nonetheless, several of the created suggestions are overlapping and iterative. Therefore to maintain satisfactory pairings between conflicting overlapping templates, we applied non-maximal suppression over the percentile rank. The ultimate detections are the templates still present after the non-maximal suppression. However, in a dense crowd and severe occlusion, the tracking performance gets slightly lower as compared to other cases.

Moreover, for the final detection of anomalous objects in the environment, object categorization is performed via deep learning feature pyramid network, and then categorized objects along with contextual information are fed to Jaccard similarity to assess the similarities among the classed objects in light of context to identify anomalies in the frame. Our scenario will be recognized as abnormal if the individual is riding a bicycle, scooter, automobile, wheelchair, skater, or any other transport among a crowd of walkers on a pathway. In the object categorization phase, the recently updated information in the log record is gathered as the detection information. In order to evaluate whether the recent actions are unusual or regular, the feature vectors of the observed behavior are examined to those of the normal and unusual behavior using the detection data of every super pixel.

VI. CONCLUSION

A real-time crowd-tracking and abnormality discovery framework via semantic segmentation and deep learning is presented in this article. It enables the identification of strange entities in the field by analyzing environmental context. This framework offers healthcare and life-care administrations

in open-air spaces such as colleges, stadiums, walkways for pedestrians, residential streets, crisis facilities, and retail plazas. This framework combines the two most significant certifiable frameworks—multi-person tracking and crowd abnormality detection systems. The efficiency of our model goes down lightly in case of little anomalous objects in pedestrian walkways like a skateboard or rolling wheels. This is usually due to the complete overlapping of pedestrians in test samples. The robustness of our model is evaluated on two openly accessible benchmark datasets i.e., UCSD Ped 1 dataset and CUHK Avenue dataset. The mean accuracy rate for crowd tracking is 92.2% and 89.1% with UCSD Ped 1 and CUHK Avenue datasets, respectively. However, an accuracy rate of 94.3% with UCSD Ped 1 and 93.7% with the CUHK Avenue dataset is accomplished for crowd anomaly detection. We successfully demonstrated the capability of our proposed framework in a busy environment through careful experimentation. A comparative study of a developed framework with different existing frameworks is additionally given, demonstrating the proposed system's superior performance.

In the upcoming work, we aim to work on increasingly trickier crowd scenarios and introduce new occlusion techniques to address the occlusion issue. Furthermore, we stretch out our work to acknowledge various scenes like mobs or tumultuous demonstrations, battles, sports, theft, and street mishap scenes.

ACKNOWLEDGEMENT

Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2023R97), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

[1] A. Bahamid and A. M. Ibrahim, "A review on crowd analysis of evacuation and abnormality detection based on machine learning systems," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21641–21655, 2022.
[2] S. Tariq, H. Farooq, A. Jaleel, and S. M. Wasif, "Anomaly detection with particle filtering for online video surveillance," *IEEE Access*, vol. 9, pp. 19457–19468, 2021.
[3] A. Kumar, "Crowd behavior monitoring and analysis in surveillance applications: A survey," *Turkish J. Comput. Math. Educ.*, vol. 12, pp. 2322–2336, Apr. 2021.
[4] A. Jalal, M. A. K. Quaid, and M. A. Sidduqi, "A triaxial acceleration-based human motion detection for ambient smart home system," in *Proc. 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2019, pp. 353–358.
[5] M. Abdelghafour, M. Elbery, and Z. Taha, "Comparative study for anomaly detection in crowded scenes," *Int. J. Intell. Comput. Inf. Sci.*, vol. 21, pp. 84–94, Nov. 2021.
[6] K. Chidananda and A. P. S. Kumar, "Human anomaly detection in surveillance videos: A review," in *Proc. ICTCS*, 2022, pp. 791–802.
[7] M. Mahmood, A. Jalal, and K. Kim, "WHITE STAG model: Wise human interaction tracking and estimation (WHITE) using spatio-temporal and angular-geometric (STAG) descriptors," *Multimedia Tools Appl.*, vol. 79, pp. 6919–6950, 2020.
[8] A. Jalal, A. Ahmed, A. A. Rafique, and K. Kim, "Scene semantic recognition based on modified fuzzy C-mean and maximum entropy using object-to-object relations," *IEEE Access*, vol. 9, pp. 27758–27772, 2021.
[9] K. Boekhoudt, A. Matei, M. Aghaei, and E. Talavera, "HR-Crime: Human-related anomaly detection in surveillance videos," in *Proc. CAIP*. Cham, Switzerland: Springer, 2021, pp. 164–174.
[10] A. Arif and A. Jalal, "Automated body parts estimation and detection using salient maps and Gaussian matrix model," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2021, pp. 667–672.
[11] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Pers. Ubiquitous Comput.*, vol. 16, pp. 1–17, Jun. 2021.
[12] H. Ashfaq and A. Jalal, "3D shape estimation from RGB data using 2.5D features and deep learning," in *Proc. 4th Int. Conf. Advancements Comput. Sci. (ICACS)*, Feb. 2023, pp. 1–7.
[13] A. Bamaqa and B. B. Bastaki, "Anomaly detection using hierarchical temporal memory (HTM) in crowd management," in *Proc. CBDC*, 2020, pp. 37–42.
[14] S. D. Bansod and A. V. Nandedkar, "Crowd anomaly detection and localization using histogram of magnitude and momentum," *Vis. Comput.*, vol. 36, no. 3, pp. 609–620, Mar. 2020.
[15] U. Azmat and A. Jalal, "Smartphone inertial sensors for human locomotion activity recognition based on template matching and codebook generation," in *Proc. Int. Conf. Commun. Technol. (ComTech)*, Sep. 2021, pp. 109–114.
[16] K. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly detection in road traffic using visual surveillance: A survey," *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–26, Nov. 2021.
[17] Y. Hao and A. Mahmood, "An end-to-end human abnormal behavior detection framework for crowd with mental disorders," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 3618–3625, Aug. 2022.
[18] A. Jalal, M. Mahmood, and A. S. Hasan, "Multi-features descriptors for human activity tracking and recognition in indoor-outdoor environments," in *Proc. 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2019, pp. 371–376.
[19] F. Yang, Z. Yu, Q. Li, and B. Guo, "Human-machine cooperative video anomaly detection," in *Proc. CSCW*, 2021, pp. 1–18.
[20] W. Raad, A. Hussein, M. Mohandes, B. Liu, and A. Al-Shaikhi, "Crowd anomaly detection systems using RFID and WSN review," in *Proc. 4th Int. Symp. Adv. Electr. Commun. Technol. (ISAECT)*, Dec. 2021, pp. 1–5.
[21] S. Vahora, K. Galiya, H. Sapariya, and S. Varshney, "Comprehensive analysis of crowd behavior techniques: A thorough exploration," *Int. J. Comput. Digit. Syst.*, vol. 11, no. 1, pp. 991–1007, Mar. 2022.
[22] Y. C. Yoon, D. Y. Kim, Y. M. Song, K. Yoon, and M. Jeon, "Online multiple pedestrians tracking using deep temporal appearance matching association," *Inf. Sci.*, vol. 561, pp. 35–326, Jun. 2021.
[23] W. Ren, D. Kang, Y. Tang, and A. B. Chan, "Fusing crowd density maps and visual object trackers for people tracking in crowd scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5353–5362.
[24] X. Zhang and L. Zhang, "Real time crowd counting with human detection and human tracking," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2014, pp. 1–8.
[25] D. Merad, K.-E. Aziz, R. Iguernaissi, B. Fertil, and P. Drap, "Tracking multiple persons under partial and global occlusions: Application to customers' behavior analysis," *Pattern Recognit. Lett.*, vol. 81, pp. 11–20, Oct. 2016.
[26] Z. Li, Y. Li, and Z. Gao, "Spatiotemporal representation learning for video anomaly detection," *IEEE Access*, vol. 8, pp. 25531–25542, 2020.
[27] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6036–6046.
[28] G. Liu, S. Liu, K. Muhammad, A. K. Sangaiah, and F. Doctor, "Object tracking in vary lighting conditions for fog based intelligent surveillance of public spaces," *IEEE Access*, vol. 6, pp. 29283–29296, 2018.
[29] M. C. Le, M.-H. Le, and M.-T. Duong, "Vision-based people counting for attendance monitoring system," in *Proc. 5th Int. Conf. Green Technol. Sustain. Develop. (GTSD)*, Nov. 2020, pp. 349–352.
[30] D. Chahyati and A. M. Arymurthy, "Multiple human tracking using retinanet features, Siamese neural network, and Hungarian algorithm," *Int. J. Mech. Eng. Technol. (IJMET)*, vol. 10, pp. 6340–6359, Jan. 2020.
[31] M. Javeed and A. Jalal, "Body-worn hybrid-sensors based motion patterns detection via bag-of-features and fuzzy logic optimization," in *Proc. Int. Conf. Innov. Comput. (ICIC)*, Nov. 2021, pp. 1–7.
[32] M. Pervaiz, A. Jalal, and K. Kim, "Hybrid algorithm for multi people counting and tracking for smart surveillance," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2021, pp. 530–535.
[33] W. Ren, X. Wang, J. Tian, Y. Tang, and A. B. Chan, "Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets," *IEEE Trans. Image Process.*, vol. 30, pp. 1439–1452, 2021.

[34] M. Waheed, M. Javeed, and A. Jalal, "A novel deep learning model for understanding two-person interactions using depth sensors," in *Proc. Int. Conf. Innov. Comput. (ICIC)*, Nov. 2021, pp. 1–8.

[35] A. Aldayri and W. Albattah, "Taxonomy of anomaly detection techniques in crowd scenes," *Sensors*, vol. 22, no. 16, p. 6080, Aug. 2022.

[36] A. Nadeem, A. Jalal, and K. Kim, "Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy Markov model," *Multimedia Tools Appl.*, vol. 80, no. 14, pp. 21465–21498, Jun. 2021.

[37] A. Al-Dhamari, R. Sudirman, and N. H. Mahmood, "Transfer deep learning along with binary support vector machine for abnormal behavior detection," *IEEE Access*, vol. 8, pp. 61085–61095, 2020.

[38] M. Javeed, M. Shorfuzzaman, N. Alsufyani, S. A. Chelloug, A. Jalal, and J. Park, "Physical human locomotion prediction using manifold regularization," *PeerJ Comput. Sci.*, vol. 8, p. e1105, Oct. 2022.

[39] A. J. Alzahrani, S. D. Khan, and H. Ullah, "Anomaly detection in crowds by fusion of novel feature descriptors," *Int. Trans. J. Eng., Manag., Appl. Sci. Technol.*, vol. 11, no. 16, 2020, Art. no. 11A16B.

[40] E. S. Sezer and A. B. Can, "Anomaly detection in crowded scenes using log-Euclidean covariance matrix," in *Proc. 13th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2018, pp. 279–286.

[41] Y. Dou, C. Fudong, J. Li, and C. Wei, "Abnormal behavior detection based on optical flow trajectory of human joint points," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2019, pp. 653–658.

[42] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "AnomalyNet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

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

[44] A. N. Moustafa and W. Gomaa, "Gate and common pathway detection in crowd scenes and anomaly detection using motion units and LSTM predictive models," *Multimedia Tools Appl.*, vol. 79, nos. 29–30, pp. 20689–20728, Aug. 2020.

[45] A. S. Hassanein, M. E. Hussein, W. Gomaa, Y. Makihara, and Y. Yagi, "Identifying motion pathways in highly crowded scenes: A non-parametric tracklet clustering approach," *Comput. Vis. Image Understand.*, vol. 191, Feb. 2020, Art. no. 102710.

[46] A.-U. Rehman, H. S. Ullah, H. Farooq, M. S. Khan, T. Mahmood, and H. O. A. Khan, "Multi-modal anomaly detection by using audio and visual cues," *IEEE Access*, vol. 9, pp. 30587–30603, 2021.

[47] J. Sun, J. Shao, and C. He, "Abnormal event detection for video surveillance using deep one-class learning," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3633–3647, Feb. 2019.

[48] X. Zhang, D. Ma, H. Yu, Y. Huang, P. Howell, and B. Stevens, "Scene perception guided crowd anomaly detection," *Neurocomputing*, vol. 414, pp. 291–302, Nov. 2020.

[49] C. Wu, S. Shao, C. Tunc, P. Satam, and S. Hariri, "An explainable and efficient deep learning framework for video anomaly detection," *Cluster Comput.*, vol. 25, pp. 2715–2737, Nov. 2021.

[50] J. Chaki and N. Dey, *A Beginner's Guide to Image Preprocessing Techniques*. Boca Raton, FL, USA: CRC Press, 2018, p. 114.

[51] M. A. K. Quaid and A. Jalal, "Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm," *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 6061–6083, Mar. 2020.

[52] T. Li, J. Wang, and K. Yao, "Visibility enhancement of underwater images based on active polarized illumination and average filtering technology," *Alexandria Eng. J.*, vol. 61, no. 1, pp. 701–708, Jan. 2022.

[53] K. Gromada, "Unsupervised SAR imagery feature learning with median filter-based loss value," *Sensors*, vol. 22, no. 17, p. 6519, Aug. 2022.

[54] M. Thoma, "A survey of semantic segmentation," 2016, *arXiv:1602.06541*.

[55] A. A. Rafique, A. Jalal, and K. Kim, "Statistical multi-objects segmentation for indoor/outdoor scene detection and classification via depth images," in *Proc. 17th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2020, pp. 271–276.

[56] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019.

[57] A. Ahmed, A. Jalal, and K. Kim, "Multi-objects detection and segmentation for scene understanding based on Texton forest and kernel sliding perceptron," *J. Elect. Eng. Technol.*, vol. 16, pp. 1143–1150, Mar. 2020.

[58] K. Kim, A. Jalal, and M. Mahmood, "Vision-based human activity recognition system using depth silhouettes: A smart home system for monitoring the residents," *J. Electr. Eng. Technol.*, vol. 14, no. 6, pp. 2567–2573, Nov. 2019.

[59] B. Kang and T. Q. Nguyen, "Random forest with learned representations for semantic segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3542–3555, Jul. 2019.

[60] J. Yin, T. Wang, Y. Du, X. Liu, L. Zhou, and J. Yang, "SLIC superpixel segmentation for polarimetric SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5201317.

[61] A. Jalal, M. A. K. Quaid, and A. S. Hasan, "Wearable sensor-based human behavior understanding and recognition in daily life for smart environments," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2018, pp. 105–110.

[62] M. Angulakshmi and G. G. L. Priya, "Walsh Hadamard transform for simple linear iterative clustering (SLIC) superpixel based spectral clustering of multimodal MRI brain tumor segmentation," *IRBM*, vol. 40, no. 5, pp. 253–262, Oct. 2019.

[63] A. Jalal, M. Batool, and K. Kim, "Sustainable wearable system: Human behavior modeling for life-logging activities using k-ary tree hashing classifier," *Sustainability*, vol. 12, no. 24, p. 10324, 2020.

[64] M. G. Uzunbas, C. Chen, and D. Metaxas, "An efficient conditional random field approach for automatic and interactive neuron segmentation," *Med. Image Anal.*, vol. 27, pp. 31–44, Jan. 2016.

[65] L. Zhou, K. Fu, Z. Liu, F. Zhang, Z. Yin, and J. Zheng, "Superpixel based continuous conditional random field neural network for semantic segmentation," *Neurocomputing*, vol. 340, pp. 196–210, May 2019.

[66] A. Jalal, N. Khalid, and K. Kim, "Automatic recognition of human interaction via hybrid descriptors and maximum entropy Markov model using depth sensors," *Entropy*, vol. 22, no. 8, p. 817, Jul. 2020.

[67] S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, "Feature pyramid network for multi-class land segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 272–2723.

[68] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: One proposal, multiple predictions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12211–12220.

[69] A. A. Rafique, A. Jalal, and K. Kim, "Automated sustainable multi-object segmentation and recognition via modified sampling consensus and kernel sliding perceptron," *Symmetry*, vol. 12, no. 11, p. 1928, Nov. 2020.

[70] B. Fernando and S. Herath, "Anticipating human actions by correlating past with the future with Jaccard similarity measures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13219–13228.

[71] R. Pradhan, M. Aggarwal, D. Maheshwari, A. Chaturvedi, and D. K. Sharma, "Diabetes mellitus prediction and classifier comparitive study," in *Proc. Int. Conf. Power Electron. IoT Appl. Renew. Energy Control (PARC)*, Feb. 2020, pp. 133–139.

[72] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909–926, May 2008.

[73] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.

[74] S. H. Abdulhussain, A. R. Ramli, B. M. Mahmmod, M. I. Saripan, S. A. R. Al-Haddad, T. Baker, W. N. Flayyih, and W. A. Jassim, "A fast feature extraction algorithm for image and video processing," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[75] A. Chriki, H. Touati, H. Snoussi, and F. Kamoun, "Deep learning and handcrafted features for one-class anomaly detection in UAV video," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2599–2620, Jan. 2021.

[76] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.

[77] W. Chu, H. Xue, C. Yao, and D. Cai, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 246–255, Jan. 2019.

[78] B. Yang, J. Cao, N. Wang, and X. Liu, "Anomalous behaviors detection in moving crowds based on a weighted convolutional autoencoder-long short-term memory network," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 4, pp. 473–482, Dec. 2019.

[79] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2914–2922.

[80] A. Shehzed, A. Jalal, and K. Kim, "Multi-person tracking in smart surveillance system for crowd counting and normal/abnormal events detection," in *Proc. Int. Conf. Appl. Eng. Math. (ICAEM)*, Aug. 2019, pp. 163–168.
[81] Y. Zhang, "Detection and tracking of human motion targets in video images based on camshift algorithms," *IEEE Sensors J.*, vol. 20, no. 20, pp. 11887–11893, Oct. 2020.
[82] A. R. Shahzad and A. Jalal, "A smart surveillance system for pedestrian tracking and counting using template matching," in *Proc. Int. Conf. Robot. Autom. Ind. (ICRAI)*, Oct. 2021, pp. 1–6.

FAISAL ABDULLAH received the M.S. degree in computer science from Air University, Islamabad, Pakistan. He is currently a Research Assistant with Air University. He is also being a reviewer in different international conferences and journals. His research interests include machine learning, deep learning, artificial intelligence, pattern recognition, scene understanding, and crowd analysis.

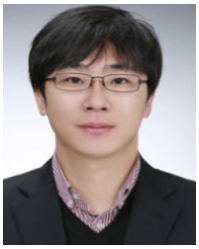
MAHA ABDELHAQ (Member, IEEE) received the B.Sc. degree in computer science and the M.Sc. degree in securing wireless communications from the University of Jordan, Jordan, in 2006 and 2009, respectively, and the Ph.D. degree from the Faculty of Information Science and Technology, National University of Malaysia, Malaysia, in 2014. She is currently an Associate Professor with the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Saudi Arabia. She is a member of ACM and International Association of Engineers (IAENG). Her research interests include vehicular networks, MANET routing protocols, artificial immune systems, network security, and intelligent computational.

RAED ALSAQOUR (Member, IEEE) received the B.Sc. degree in computer science from Mutah University, Jordan, in 1997, the M.Sc. degree in distributed systems from University Putra Malaysia, Malaysia, in 2003, and the Ph.D. degree in wireless communication systems from the National University of Malaysia, Malaysia, in 2008. He is currently an Associate Professor with the College of Computation and Informatics, Saudi Electronic University, Riyadh Branch, Saudi Arabia. He is a member of ACM and IAENG. His research interests include wireless ad hoc and vehicular networks, routing protocols, simulation, and network performance evaluation.

MOHAMMED HAMAD ALATIYYAH is currently a Full Professor in computer science and information with Prince Sultan University. His research interest includes image processing.

KHALED ALNOWAISER received the Ph.D. degree in computer science from Glasgow University, Scotland. He is currently an Assistant Professor with the Computer Engineering Department, Prince Sattam Bin Abdulaziz University, Saudi Arabia. His research interests include computer vision, optimization techniques, and performance enhancement.

SAUD S. ALOTAIBI received the Bachelor of Computer Science degree from King Abdul Aziz University, in 2000, the master's degree in computer science from King Fahd University, Dhahran, in May 2008, and the Ph.D. degree in computer science from Colorado State University, Fort Collins, USA, in August 2015, under the Supervision of Dr. Charles Anderson. He started his career as an Assistant Lecturer with Umm Al-Qura University, Makkah, Saudi Arabia, in July 2001, where he was a Deputy of the IT-Center for E-Government and Application Services, in January 2009. From 2015 to 2018, he was with the Deanship of Information technology to improve the IT services that are provided to Umm Al-Qura University. He is currently an Assistant Professor in computer science with Umm Al-Qura University. He is also the Vice Dean for Academic Affairs with the Computer and Information College. His current research interests include AI, machine learning, natural language processing, neural computing IoT, knowledge representation, smart cities, wireless, and sensors.



JEONGMIN PARK received the Ph.D. degree from the College of Information and Communication Engineering, Sungkyunkwan University, South Korea, in 2009. He is currently an Associate Professor with the Department of Computer Engineering, Tech University of Korea, South Korea. Before joining Tech University of Korea, in 2014, he was a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI) and a Research Professor with Sungkyunkwan University. His research interests include high-reliable autonomic computing mechanism and human-oriented interaction systems.