mathematics

MDPI

Article

Computational Intelligence-Based Harmony Search Algorithm for Real-Time Object Detection and Tracking in Video Surveillance Systems

Maged Faihan Alotaibi [1,2], Mohamed Omri [2], Sayed Abdel-Khalek [3,4], Eied Khalil [3,5] and Romany F. Mansour [6,*]

[1] Department of Physics, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia; malhabrdi@kau.edu.sa
[2] Deanship of Scientific Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia; mnomri@kau.edu.sa
[3] Mathematics Department, Faculty of Science, Taif University, Taif 21944, Saudi Arabia; sabotalb@tu.edu.sa (S.A.-K.); eiedkhalil@tu.edu.sa (E.K.)
[4] Mathematics Department, Faculty of Science, Sohag University, Sohag 82524, Egypt
[5] Mathematics Department, Faculty of Science, Azhar University, Cairo 11884, Egypt
[6] Department of Mathematics, Faculty of Science, New Valley University, El-Kharga 72511, Egypt
* Correspondence: romanyf@sci.nvu.edu.eg

Abstract: Recently, video surveillance systems have gained significant interest in several application areas. The examination of video sequences for the detection and tracking of objects remains a major issue in the field of image processing and computer vision. The object detection and tracking process includes the extraction of moving objects from the frames and continual tracking over time. The latest advances in computation intelligence (CI) techniques have become popular in the field of image processing and computer vision. In this aspect, this study introduces a novel computational intelligence-based harmony search algorithm for real-time object detection and tracking (CIHSA-RTODT) technique on video surveillance systems. The CIHSA-RTODT technique mainly focuses on detecting and tracking the objects that exist in the video frame. The CIHSA-RTODT technique incorporates an improved RefineDet-based object detection module, which can effectually recognize multiple objects in the video frame. In addition, the hyperparameter values of the improved RefineDet model are adjusted by the use of the Adagrad optimizer. Moreover, a harmony search algorithm (HSA) with a twin support vector machine (TWSVM) model is employed for object classification. The design of optimal RefineDet feature extraction with the application of HSA to appropriately adjust the parameters involved in the TWSVM model for object detection and tracking shows the novelty of the work. A wide range of experimental analyses are carried out on an open access dataset, and the results are inspected in several ways. The simulation outcome reported the superiority of the CIHSA-RTODT technique over the other existing techniques.

Keywords: computational intelligence; video surveillance; object detection; object tracking; deep learning; metaheuristics

Citation: Alotaibi, M.F.; Omri, M.; Abdel-Khalek, S.; Khalil, E.; Mansour, R.F. Computational Intelligence-Based Harmony Search Algorithm for Real-Time Object Detection and Tracking in Video Surveillance Systems. *Mathematics* **2022**, *10*, 733. https://doi.org/10.3390/math10050733

Academic Editor: Jakub Nalepa

Received: 22 January 2022
Accepted: 21 February 2022
Published: 25 February 2022

1. Introduction

The rapid development of hardware services such as processing machines, smartphones, and cameras has resulted in an explosion of research in automatic video analysis for tracking and detecting objects [1]. It is a hot research topic in image processing and computer vision (CV). Object tracking and detection in a video sequence is a fundamental method in the expansion of different video analysis applications that endeavors to track and detect objects through a series of images by replacing the conventional method of a surveillance camera with a human operator [2]. Object detection needs precise classification of objects in images and requires the precise location of the object, and is an automated

image detection system based on geometric and statistical features [3]. The accurateness of object location and object classification is a major indicator to evaluate the efficiency of the detection system. Object detection is more commonly employed in military object detection, intelligent monitoring, unmanned vehicles, intelligent transportation, and UAV navigation [4]. However, due to the variety of detected objects, the present system failed to identify objects. Complex backgrounds and changing light increase the complexity of object detection, particularly for objects in challenging conditions [5].

The tracking method works by identifying an object once it originally appears in a frame and forecasting its trajectory [6]. This detection-based algorithm estimates the object's position in all the frames independently. It requires an offline training phase and could not be employed on unknown objects. Many tracking and detection methods have been developed in modern times. However, the numerous problems encountered during this procedure mean that this field requires further study. The problems that complicate tracking and detection include rapid illumination changes, camera jitter, moving cameras, dynamic backgrounds, shadow detection, and so on [7]. These problems cannot be resolved by a simple algorithm due to the improbable factors, complexities, and impreciseness present in the intermediate step. To resolve this, computational intelligence (CI) and deep learning approaches have been initiated [8].

Computational intelligence includes approaches such as artificial neural network (ANN), genetic algorithm (GA), fuzzy logic control (FLC), adaptive neuro-fuzzy inference scheme (ANFS), and particle swarm optimization (PSO) [9]. Research has been conducted to discover methods for reliable and efficient load shedding. CI technology, which shows desirable effectiveness and efficiency in data mining and data processing, is also receiving considerable interest with regards to CV tasks. In recent times, deep convolutional neural networks (DCNN) and their derivates are typical examples. These examples have made considerable achievements in CV tasks, including semantic segmentation and classification, automated image representation generation, data restoration, object detection, and tracking [10]. They have significantly outperformed conventional techniques. Because of the effective performance of CI approaches in data processing and DM in CV tasks, it could be effective and reasonable to explore this CI technology to tackle the problem in a real-time scenario.

This study introduces a novel computational intelligence-based harmony search algorithm for real-time object detection and tracking (CIHSA-RTODT) technique on video surveillance systems. The CIHSA-RTODT technique designs an Adagrad optimizer with an improved RefineDet-based object detection technique. Moreover, a harmony search algorithm (HSA) with a twin support vector machine (TWSVM) model is employed for object classification. The application of HSA helps to appropriately adjust the parameters involved in the TWSVM model and thereby leads to improved classification results. A wide range of experimental analyses is carried out on an open access dataset and the results are inspected in several ways.

2. Literature Review

Elhoseny [11] presented new MODT methods. The presented approach makes use of the optimum Kalman filtering method to track the object moving in the video frame. The video clip was transformed according to the number of frames into a morphological operation with the region growing method. After differentiating the object, Kalman filtering was employed for parameter optimization using the probability-based grasshopper approach. With the optimum parameter, the carefully chosen object was tracked in all the frames by a similarity measure. The authors in [12] developed a multi-object detection and tracking model using background subtraction and the K-means clustering technique. The presented technique has the ability to handle object occlusion, shadows, and camera jitter. Background subtraction removes the unwanted data, and K-means clustering is used to select moving objects from the rest of the data. It is also able to handle the merging and splitting of moving objects via spatial information.

Lyu et al. [13] proposed to increase a highly qualified object detector using effective and efficient class-agnostic convolution regression trackers for object detection tasks. The tracker learns how to track objects by reutilizing the feature from the object detectors that is a lightweight increment to the detectors, with a small speed drop for the object detection process.

Xiong et al. [14] introduced an enhanced active obstacle-separation model that employed push and drag–push operations to separate the problems from the target in three phases. The push and drag vectors were precisely calculated and simplified based on the accurate location of the obstacle. Moreover, in contrast to the system that only "looked" once for the whole picking procedure, the novel scheme utilized a hybrid vision-based control system. Lin et al. [15] introduced a hybrid track association (HTA) approach that models the past appearance distance of a track with an increment Gaussian mixture model (IGMM) and incorporates the derived statistical data into the evaluation of the detection-to-track association cost. Chen et al. [16] proposed an architecture using multiple object tracking (MOT) for detection. Fast RCNN is utilized for obtaining detected objects, and the KCF tracker is utilized for tracking object trajectories. The Hungarian algorithm is utilized for bounding box matching to attain previous data of trajectory to enhance recognition efficiency. Schöller et al. [17] recommended an approach to track an object that is identified by an NN system. The proposed technique is estimated on data attained in Danish near-coastal water. The approach uses a feature that is evaluated in the detection phase, thus ensuring a good feature that is typical for the provided object while saving the time it would take to calculate a novel feature. Shi et al. [18] described a solution to resolve the problem of automated multi-pedestrian counting and tracking. Firstly, the background modeling approach is employed to actively obtain multi-pedestrian candidates; after that is the authorization stage using classification. Next, all the pedestrian patches can be managed by real-time TLD (Tracking–Learning–Detection) to attain a new prediction location based on similarity measures.

3. The Proposed Model

In this study, a new CIHSA-RTODT technique was developed to detect and track objects on video surveillance systems. The CIHSA-RTODT technique designed an Adagrad with an improved RefineDet-based object detector, which can effectually recognize multiple objects in the video frame. Moreover, the HSA with TWSVM model is applied to properly categorize the existence of objects in the video frame and thereby leads to improved classification results.

3.1. Object Detection Module: Adagrad with Improved RefineDet Model

At the initial stage, the improved RefineDet model was applied for the identification of objects that exist in the video frame. The improved RefineDet technique, with VGG16 networks as core networks, creates a series of anchors with distinct scales and various feature ratios from all feature-map cells by utilizing the anchor generation process of RPN [19] and attains a set quantity of object bounding boxes, then two classification and regression boxes, as well as probabilities of the presence of distinct classes under these bounding boxes. Eventually, the last classification and regression outcomes were attained with non-maximum suppression (NMS). The improved RefineDet method is separated into three modules such as the object detection module (ODM), anchor refinement module (ARM), and transfer connection block (TCB). Figure 1 illustrates the structure of RefineDet.

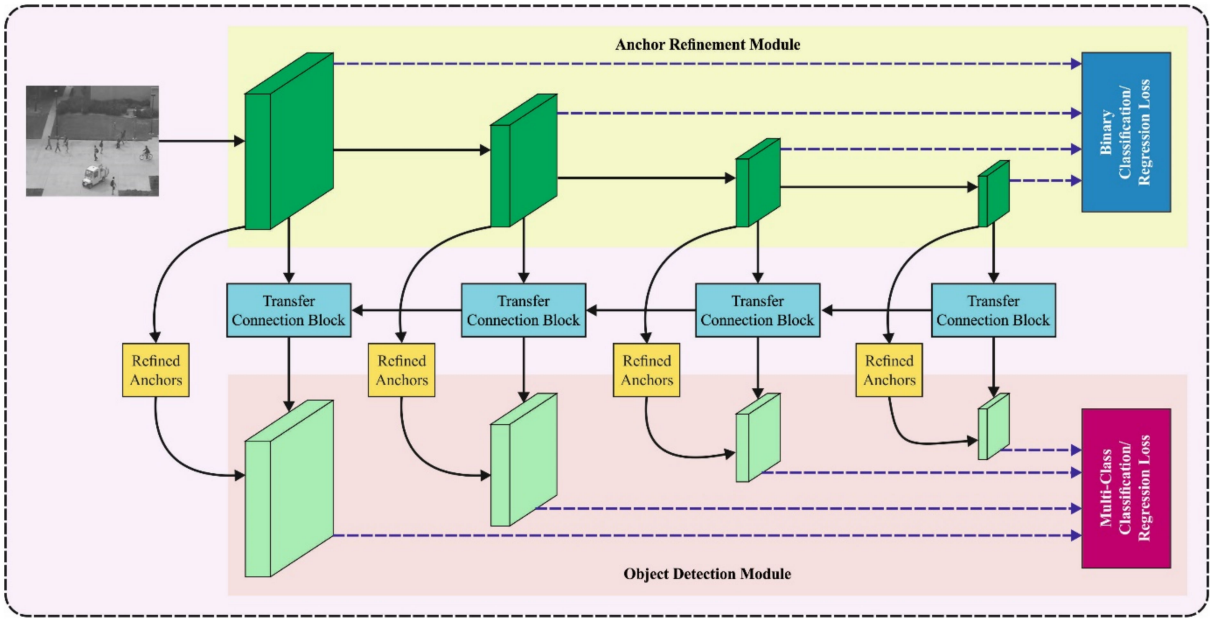

Figure 1. Architecture of RefineDet.

3.1.1. ARM Module

The ARM was mostly collected from the backbone network VGG16 and more convolutional layers. The ARM mostly performs anchor refinement, feature extraction, negative anchor (NA) filtering, and anchor generation. The anchor refinement changes the place and size of the anchor box, and the NA filter implies that, in ARM, if the confidence of the negative instance is greater than 0.99, this technique removes it and does not utilize it in the last detection of the ODM. The NA filter effectually filters out the NA box with classification and alleviates the instance imbalance. During this procedure of feature extraction, two convolutional layers, for instance, ConVo6_1 and ConVo6_2, are at the end of the VGG16 networks. Afterwards, four additional convolutional layers, such as ConVo7_1, ConVo7_2, ConVo8_1, and ConVo8_2, can be added to capture more high-level (HL) semantic data from this technique. Moreover, the HL feature of ConVo8_2 was fused with the low-level (LL) feature of ConVo7_2. Afterwards, the fused feature was transmitted to the LL feature by TCB, but the LL feature map utilized for detection had maximum semantic data and enhanced the detection accuracy of the floating object.

3.1.2. TCB Module

The TCB was mostly utilized for connecting the ARM and ODM and transferring the feature data of the ARM to the ODM. Furthermore, related to the infrastructure of FPN, neighboring TCBs were linked to achieve the feature fusion of high as well as LL features and enhance the semantic data of the LL feature.

3.1.3. ODM Module

The ODM was one of the collection outputs of the TCBs and forecast layer (classification and regression layers, that is, convolutional layers with 3×3 kernel sizes). The result of the forecast layer is a particular type of refined anchor, and the coordinate offset is comparative to the refined anchor box. The refined anchor was utilized as input for more classification and regression, and the last bounding box was chosen based on non-maximum suppression (NMS).

3.1.4. Hyperparameter Optimization

The hyperparameters involved in the object detection model are effectually tuned by the Adagrad optimizer [20]. In the case of the Adagrad optimizer, the gradients (g_τ) and accumulated squared gradients of every variable at the round t are expressed as follows:

$$G_t = \sum_{\tau=1}^{t} g_\tau \odot g_\tau \tag{1}$$

where $\odot$ indicates an element-wise multiplication, and $g_\tau \in \mathbb{R}^{|\theta|}$ indicates the gradient of the present at the τ round. The upgrade value of variables ($\Delta\theta_t$) can be the Adagrad and is represented as follows.

$$\Delta\theta_t = -\frac{\alpha}{\sqrt{G_t + \varepsilon}} \odot g_\tau \tag{2}$$

where α indicates the learning rate and ε represents a smoothing element that avoids division by zero. Since the learning rate can be predefined before training, Equation (3) is rewritten as follows.

$$\theta_t = -\alpha\left(\frac{1}{\sqrt{G_t + \varepsilon}} \odot g_\tau\right) \tag{3}$$

where G_t implies earlier gradient computation, and gradient revision g'_t can be represented as follows:

$$g'_t = \frac{1}{\sqrt{G_t + \varepsilon}} \odot g_\tau \tag{4}$$

Therefore, the Adagrad can be updated by the use of:

$$\Delta\theta_t = -\alpha g'_t \tag{5}$$

where $\Delta\theta_t$ denotes upgrade values of a variable at round t and α indicates the learning rate.

3.2. Object Classification Module: HSA with TWSVM Model

Once the objects were identified, the next stage was to classify them into distinct classes using the TWSVM model [21]. The TWSVM is an enhanced version of the traditional SVM [22]. The SVM is a supervised classification model employed in several real-time applications [23,24]. The TWSVM's purpose is to find two symmetry planes so that all planes have a distance of nearly one data class and are feasible in another data class [21]. Assume a training dataset D, which holds a set of m row vectors in n dimensional space, $D = \{(x_i, y_i)|x_i \in X^m, y_i \in \{-1, +1\}, i = 1, 2, \ldots, N\}$ and $y_i \in \{+1, -1\}$ represents the class to which the *ith* instance belongs. Thereafter, there are d_1 data points from class $+1$ and d_2 data point from class -1 so as $d_1 + d_2 = d$. The procedure $(d_1 \times n)$ matrix A has the data points from class $+1$, and $(d_2 \times n)$ matrix B has the data points from class -1. The two non-parallel hyperplanes are [20]:

$$x^T w_1 + b_1 = 0 \tag{6}$$

$$x^T w_2 + b_2 = 0 \tag{7}$$

where x refers to the data vector, w_1 signifies the weight parameter to the primary hyperplane, b_1 denotes the bias parameter to the initial hyperplane, w_2 represents the weight parameter to the second hyperplane, and b_2 implies the bias parameter to the second hyperplane. The TWSVM technique was attained by resolving the subsequent pair of quadratic programming problems [21]:

TWSVM 1:

$$\min_{w_1, b_1 \xi_2,} \frac{1}{2}\|Aw_1 + e_1 b_1\|^2 + c_1 e_2^T \xi_2 \tag{8}$$

subject to

$$-(Bw_1 + e_2 b_1) \geq e_2 - \xi_2 \tag{9}$$

$$\xi_2 \geq 0 \tag{10}$$

and

TWSVM 2:

$$\min_{w_2, b_2 \xi_1,} \frac{1}{2}\|Bw_2 + e_2 b_2\|^2 + c_2 e_1^T \xi_1 \tag{11}$$

subject to

$$-(Aw_2 + e_2 b_2) \geq e_1 - \xi_1 \tag{12}$$

$$\xi_1 \geq 0 \tag{13}$$

where $c_1 > 0$ and $c_2 > 0$ imply the penalty parameters, ξ_1 and ξ_2 denote the slack variables, and e_1 and e_2 indicate the vectors of ones, for instance, all components are 'one' only [20].

The two hyperplanes of TWSVM with kernel [20]:

$$K\left(x^T, C^T\right)u_1 + b_1 = 0 \tag{14}$$

$$K\left(x^T, C^T\right)u_2 + b_2 = 0 \tag{15}$$

where $C^T = [A, B]^T, u_1, u_2 \in R^d$, and K refer to the kernel matrix equivalent to the suitably selected kernel function. The kernel TWSVM is attained by resolving the optimized problem [21]:

KTWSVM 1:

$$\min_{w_1, b_1 \xi_2} \frac{1}{2}\|K\left(A, C^T\right)u_1 + e_1 b_1\|^2 + c_1 e_2^T \xi_2 \tag{16}$$

subject to

$$-\left(K\left(B, C^T\right)u_1 + e_2 b_1\right) \geq e_2 - \xi_2 \tag{17}$$

$$\xi_2 \geq 0 \tag{18}$$

and

KTWSVM 2:

$$\min_{w_2, b_2 \xi_1} \frac{1}{2}\|K\left(B, C^T\right)u_2 + e_2 b_2\|^2 + c_2 e_1^T \xi_1 \tag{19}$$

subject to

$$-\left(K\left(A, C^T\right)u_2 + e_2 b_2\right) \geq e_1 - \xi_1 \tag{20}$$

$$\xi_1 \geq 0 \tag{21}$$

where $c_1 > 0$ and $c_2 > 0$ define the penalty parameters, ξ_1 and ξ_2 demonstrate the slack variables, e_1 and e_2 stand for the vectors of $'ones'$, that is, all the components are 'one' only, and $C^T = [A, B]^T, u_1, u_2 \in R^d$, and K refer to the kernel matrix equivalent to suitably selected kernel functions.

To attain optimal classification performance, the HSA was applied to TWSVM parameters such as penalty parameters (c_1 and c_2) and slack variables (ξ_1 and ξ_2). The HSA was chosen due to the following benefits: easier implementation, fewer adjustable parameters, and quick convergence. The HSA is a metaheuristic method that searches for optimization issues and generates an accurate state of harmony by improvising the searching procedure. It has wide-ranging applications since it is easy to implement, simple, and involves fewer parameters [25]. The natural musical method is improvised (in terms of pitch adjustment) to produce an optimal state of harmony using HS. It is an optimization approach that is similar to global and local searching methods used to discover an optimal solution. HS is characterized as a set of solution vectors named harmony memory (HM), whereas every individual (a harmony or vector) is analogous to the chromosome of DE or GA and particles in PSO. HM is initialized by an arbitrary solution vector and is upgraded by every improvisation via some parameter adjustment. The control parameter consists of

bandwidth (BW), harmony memory consideration rate (HMCR), and pitch adjustment rate (PAR). The process involved in the HSA is given in Algorithm 1. Optimization with the harmony search approach is given below.

Step 1: Initialize Control Parameter.
Step 2: Initialize Harmony memory.
Step 3: Estimate the efficiency of present harmony.
Step 4: Estimate the efficiency of recently created harmony and improvise harmony.
Step 5: Check ending condition.

Algorithm 1 Pseudocode of the harmony search algorithm (HSA).

```
Begin;
    Determine objective function f(x), x = (x_1, x_2, ..., x_d)^T
    Determine Harmony Memory Considering rate (HMCR)
    Determine Pitch adjusting rate (PAR) and other parameters
    Create Harmony Memory with arbitrary harmonies
    while (t < max number of iterations)
      while (I <= number of variables)
        if (rand < HMCR),
          Select the value in HM for the variable i
            if (rand < PAR),
              Modify the value by adding a particular amount
            end if
        else
          Select an arbitrary value
        end if
      end while
        Take the New Harmony (solution) if better
    end while
  Define the present optimum solution
End
```

The HSA approach derives an FF to reach higher classification performance. It resolves a positive integer to signify an optimum efficacy of the candidate solution. During this analysis, the minimization of the classification error rate, which was regarded as FF, is provided in Equation (22). A better solution is a lower error rate, and the worst solution reaches an enhanced error rate.

$$fitness(x_i) = Classifier\ Error\ Rate(x_i) = \frac{number\ of\ misclassified\ objects}{Total\ number\ of\ objects} * 100 \quad (22)$$

4. Performance Validation

This section includes the experimental result analysis of the CIHSA-RTODT technique carried out on the benchmark UCSD dataset. The proposed model was executed on a Processor—i5-8600k, Graphics Card—GeForce 1050 Ti 4 GB, 16 GB RAM, and OS Storage—250 GB SSD. The proposed model was simulated using the Python 3.6.5 tool. The parameter settings of the proposed model were given as follows: batch size: 500, max. epochs: 15, learning rate: 0.05, dropout rate: 0.2, and momentum: 0.9. The results were inspected using several measures. For experimental validation, the entire dataset was divided into 70% of training data and 30% of testing images.

4.1. Dataset Details

The performance validation of the CIHSA-RTODT technique was performed using the benchmark UCSD dataset [26], which comprises two testbeds, namely the Pedestrian-1 and Pedestrian-2 datasets. The Pedestrian-1 dataset and Pedestrian-2 dataset include 360 frames with a 12 s duration, as shown in Table 1. Figure 2 illustrates the sample test images from the UCSD dataset along with its ground truth images.

Table 1. Description of dataset.

Dataset	Testbed	Frames No.	Time (s)
UCSDped2	Pedestrian-1 Dataset Pedestrian-2 Dataset	360	12

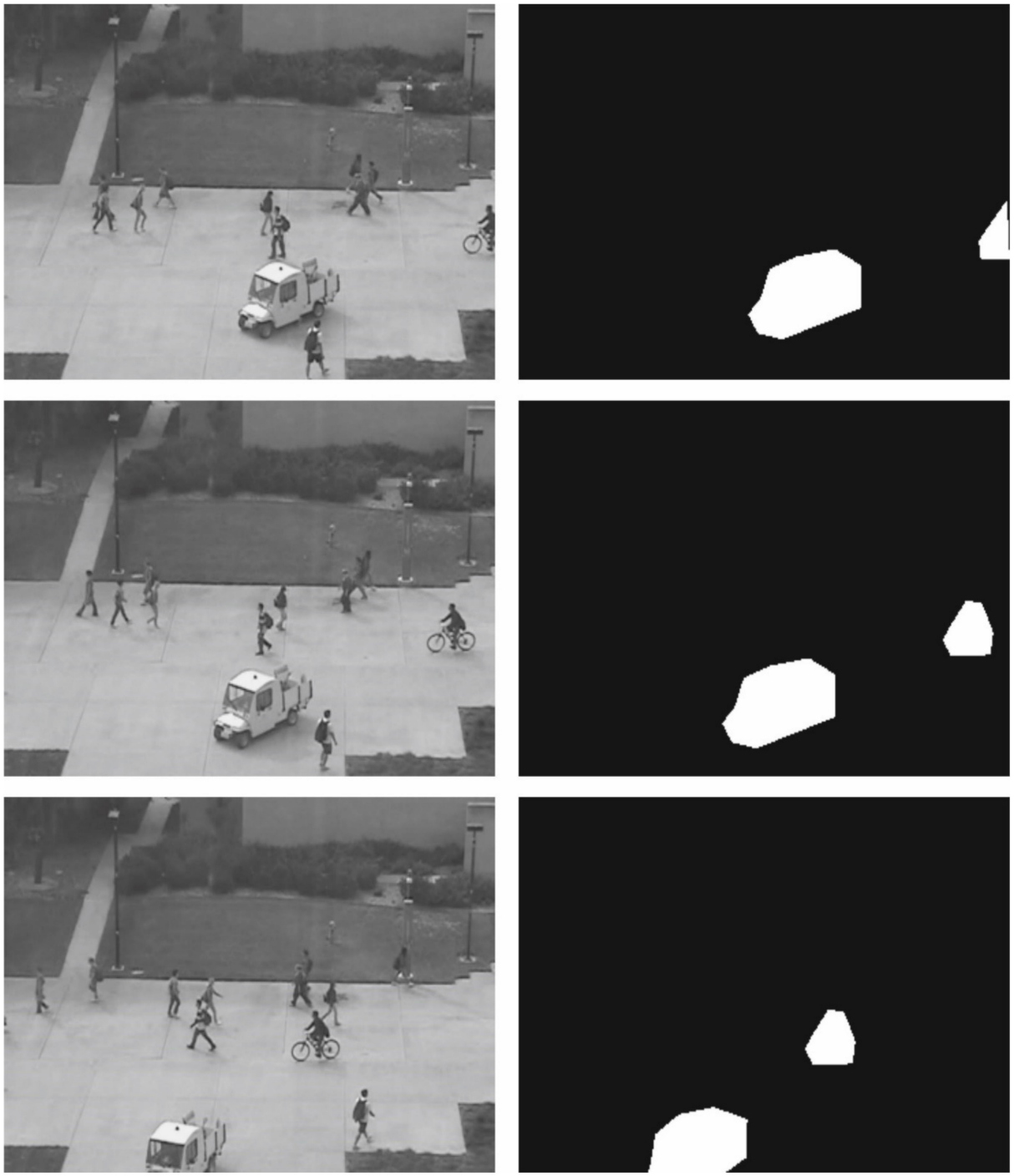

Figure 2. Sample images with ground truth images: First column, original images; second column, ground truth.

4.2. Detection Results of CIHSA-RTODT Technique

A sample visualization result analysis of the CIHSA-RTODT technique is shown in Figure 3. The results indicate that the CIHSA-RTODT technique effectually identified the

objects that exist in all the video frames. From the figure, it is evident that the CIHSA-RTODT technique effectively recognized the objects in the video frames.

Figure 3. Sample test images detection and tracking.

4.3. Running Time Analysis of CIHSA-RTODT Technique

Table 2 and Figure 4 offer a detailed running time examination of the CIHSA-RTODT technique on the Pedestrian-1 and Pedestrian-2 datasets. The results indicate that the CIHSA-RTODT technique has a superior minimum running time over the other methods. For instance, with the Pedestrian-1 dataset, the CIHSA-RTODT technique offered a lower running time of 0.035 min, whereas the MDT, SCLF, AMDN, and ADVAE techniques obtained higher running times of 0.336 min, 0.328 min, 0.188 min, and 0.057 min, respectively. Moreover, with the Pedestrian-2 dataset, the CIHSA-RTODT algorithm offered a lower running time of 0.057 min, whereas the MDT, SCLF, AMDN, and ADVAE algorithms obtained maximum running times of 0.373 min, 0.300 min, 0.207 min, and 0.094 min, respectively.

Table 2. Running Time analysis of CIHSA-RTODT technique.

Models	**Running Time (min)**	
	Pedestrian-1	**Pedestrian-2**
MDT Model	0.336	0.373
SCLF Model	0.328	0.300
AMDN Model	0.188	0.207
ADVAE Model	0.057	0.094
CIHSA-RTODT	0.035	0.057

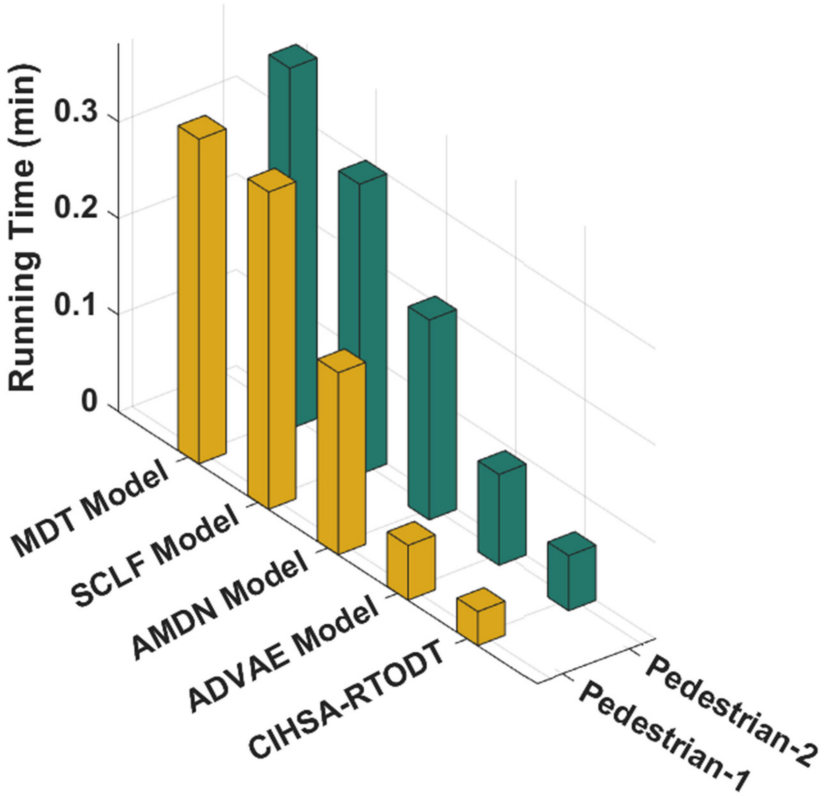

Figure 4. Running time analysis of CIHSA-RTODT technique.

4.4. Comparative Result Analysis of CIHSA-RTODT Technique

Figure 5 offers a brief average accuracy analysis of the CIHSA-RTODT technique with recent methods on the Pedestrian-1 and Pedestrian-2 datasets. The results indicate that the CIHSA-RTODT technique accomplished better results than the existing methods. For instance, with the Pedestrian-1 dataset, the CIHSA-RTODT technique resulted in an increased average accuracy of 98.64%, whereas the DLADT, Region CNN, FR-CNN, MDT-PWD, and MPPCA-PWD techniques had reduced accuracies of 97.44%, 97.17%, 85.2%, 80.18%, and 73.66%, respectively. Furthermore, with the Pedestrian-2 dataset, the CIHSA-RTODT technique attained a maximal average accuracy of 91.23%, whereas the DLADT, Region CNN, FR-CNN, MDT-PWD, and MPPCA-PWD methodologies offered reduced average accuracies of 89.8%, 87.16%, 81.33%, 77.41%, and 70.84%, respectively.

Table 3 and Figure 6 provide a brief AUC analysis of the CIHSA-RTODT technique with recent methods on the Pedestrian-1 and Pedestrian-2 datasets. The results indicate that the CIHSA-RTODT technique accomplished optimum results compared with the existing algorithms. For instance, with the Pedestrian-1 dataset, the CIHSA-RTODT technique resulted in a higher AUC of 97.51%, whereas the DLADT, Region CNN, FR-CNN, MDT-PWD, and MPPCA-PWD techniques had reduced accuracies of 60%, 66.85%, 67.11%, 81.84%, and 92%, respectively. Moreover, with the Pedestrian-2 dataset, the CIHSA-RTODT technique attained a maximum AUC of 94.32%, whereas the DLADT, Region CNN, FR-CNN, MDT-PWD, and MPPCA-PWD algorithms obtained lower AUCs of 69.87%, 55.42%, 61.43%, 82.85%, and 90.75%, respectively.

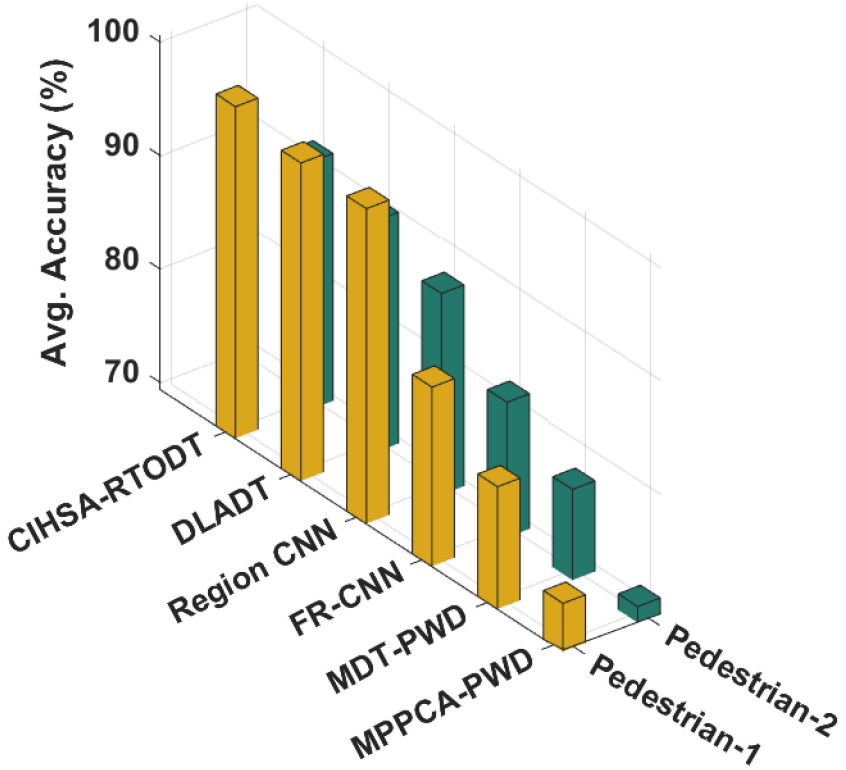

Figure 5. Average accuracy of CIHSA-RTODT technique.

Table 3. AUC analysis of CIHSA-RTODT technique with existing techniques.

Models	AUC (%)	
	Pedestrian-1	**Pedestrian-2**
MPPCA Model	60.00	69.87
SF Model	66.85	55.42
SFMPPCA Model	67.11	61.43
MDT Model	81.84	82.85
AMDN Model	92.00	90.75
ADVAE Model	95.85	92.63
CIHSA-RTODT	97.51	94.32

Table 4 and Figure 7 provide the result analysis of the CIHSA-RTODT technique with the existing techniques on the Pedestrian-1 dataset [27–29]. The results show that the CIHSA-RTODT technique obtained increased TPR over the other methods, with a rise in the false positive rate (FPR). For instance, with an FPR of 10, the CIHSA-RTODT technique attained a higher TPR of 45.50%, whereas the SF, SFMPPCA, AMDN, and ADVAE techniques had lower TPRs of 18.20%, 20.10%, 24.70%, and 20.80%, respectively. Moreover, with an FPR of 30, the CIHSA-RTODT technique attained a higher TPR of 42.50%, whereas the SF, SFMPPCA, AMDN, and ADVAE techniques had lower TPRs of 44%, 64.10%, 68.50%, and 93.80%, respectively. Furthermore, with an FPR of 50, the CIHSA-RTODT technique attained a higher TPR of 63.30%, whereas the SF, SFMPPCA, AMDN, and ADVAE techniques had lower TPRs of 63.30%, 72.40%, 84.70%, and 91.10%, respectively.

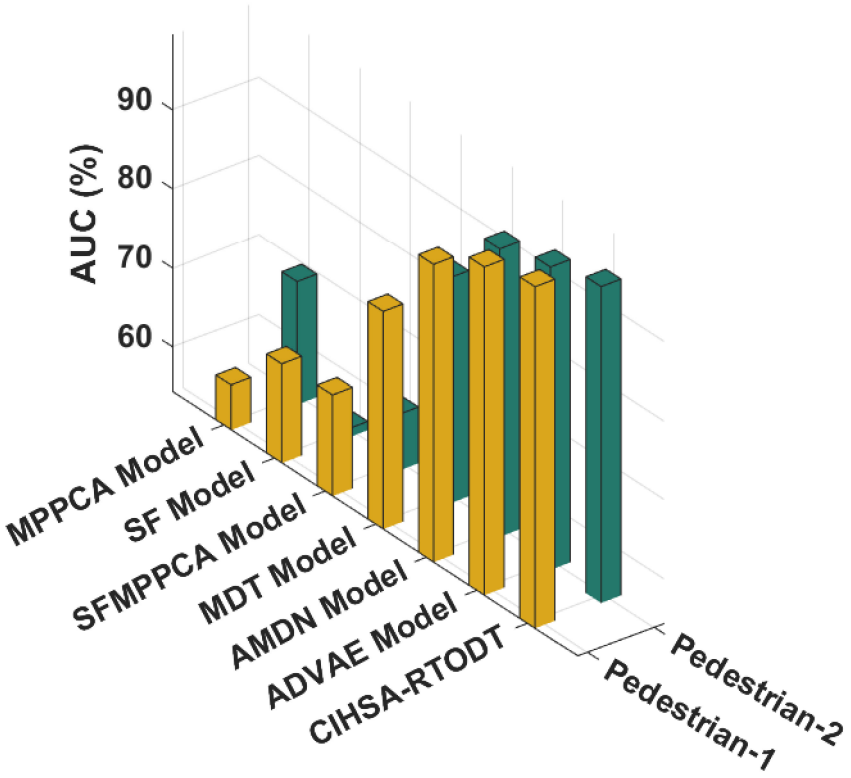

Figure 6. AUC analysis of CIHSA-RTODT technique.

Table 4. Result analysis of CIHSA-RTODT technique with existing techniques on Pedestrian-1 dataset in terms of TPR (in %).

FPR	Methods				
	SF Model	SFMPPCA Model	AMDN Model	ADVAE Model	CIHSA-RTODT
10	18.20	20.10	24.70	20.80	45.50
20	30.20	31.70	46.40	44.50	68.80
30	42.50	44.00	64.10	68.50	93.80
40	53.40	60.90	73.80	79.60	92.00
50	63.30	72.40	84.70	91.10	95.80
60	71.40	81.50	91.30	95.60	98.60
70	88.00	98.00	98.70	98.80	99.40
80	89.30	99.90	99.70	98.80	98.90
90	90.40	96.70	97.30	98.90	99.70
100	91.39	94.30	96.80	98.50	99.60

Table 5 and Figure 8 offer the outcome analysis of the CIHSA-RTODT approach with existing methodologies on the Pedestrian-2 dataset. The outcomes show that the CIHSA-RTODT methodology reached maximum TPR, in contrast to the other approaches, with an increase in FPR. For instance, with an FPR of 10, the CIHSA-RTODT approach reached a superior TPR of 28.20%, whereas the SF, SFMPPCA, AMDN, and ADVAE methodologies had minimum TPRs of 19.50%, 19.30%, 28.10%, and 16.40%, respectively. Moreover, with an FPR of 30, the CIHSA-RTODT methodology gained a TPR of 79.30, whereas the SF, SFMPPCA, AMDN, and ADVAE algorithms resulted in lower TPRs of 40.10%, 56.30%, 57.20%, and 69.10%, respectively. Furthermore, with an FPR of 50, the CIHSA-RTODT technique attained a higher TPR of 99.10%, whereas the SF, SFMPPCA, AMDN, and ADVAE methodologies resulted in lower TPRs of 73.40%, 84.90%, 86.60%, and 87.60%, respectively.

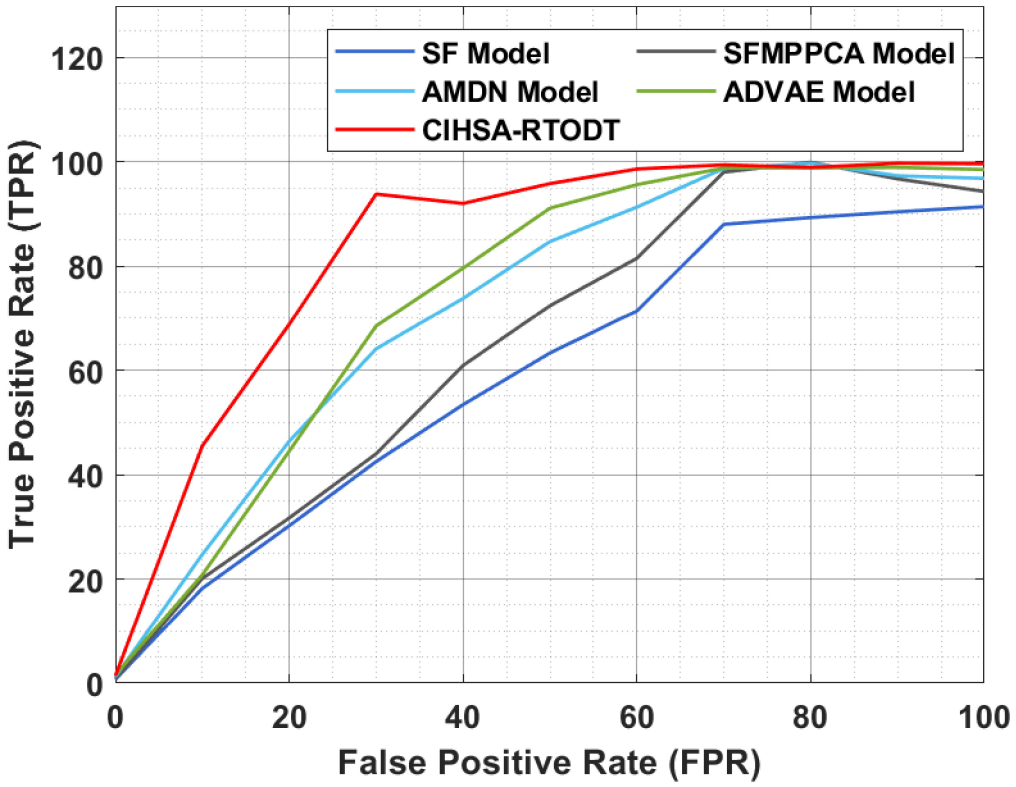

Figure 7. Result analysis of CIHSA-RTODT technique on Pedestrian-1 dataset.

Table 5. Result analysis of CIHSA-RTODT technique with existing techniques on Pedestrian-2 dataset in terms of TPR (in %).

FPR	Methods				
	SF Model	SFMPPCA Model	AMDN Model	ADVAE Model	CIHSA-RTODT
10	19.50	19.30	28.10	16.40	28.20
20	28.10	41.50	48.10	28.50	60.20
30	40.10	56.30	57.20	69.10	79.30
40	55.60	71.10	74.30	81.60	94.30
50	73.40	84.90	86.60	87.60	99.10
60	85.50	92.60	93.40	94.70	99.20
70	99.00	97.30	98.30	99.40	99.90
80	99.60	98.90	98.60	99.60	99.80
90	99.60	99.50	99.40	99.70	99.70
100	99.70	99.20	99.30	99.50	99.80

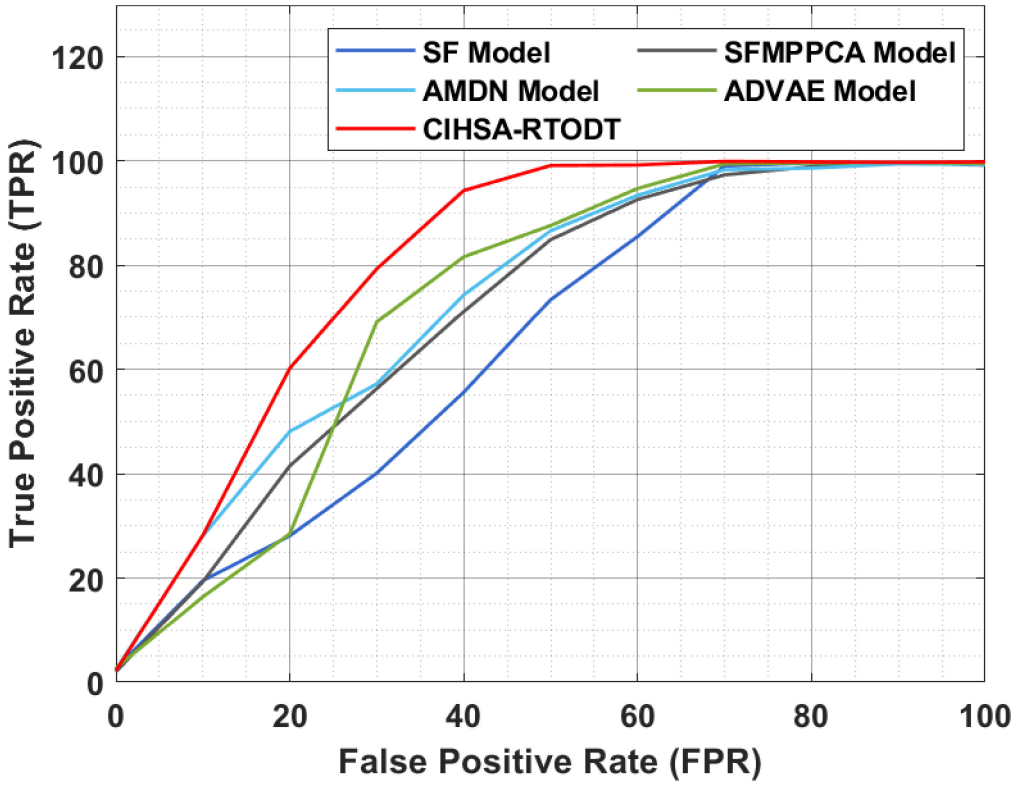

Figure 8. Result analysis of CIHSA-RTODT technique on Pedestrian-2 dataset.

4.5. Discussion

From the above-mentioned results, it is clear that the proposed model accomplished enhanced performance with average accuracies of 97.51% and 94.32% on the test Pedestrian-1 and Pedestrian-2 datasets, respectively. The enhanced performance of the CIHSA-RTODT technique is due to the inclusion of the Adagrad optimizer and HSA for the parameter-tuning process. Therefore, the presented CIHSA-RTODT technique can effectually detect and track objects in the video surveillance system. The proposed CIHSA-RTODT technique can be employed in real-time scenarios such as public places, hospitals, smart cities, etc., to assure security. It can also be utilized for crowd behavior and anomaly events in crowd scenes.

5. Conclusions

In this study, a new CIHSA-RTODT technique was developed to detect and track objects on video surveillance systems. The CIHSA-RTODT technique designed an Adagrad with an improved RefineDet-based object detector which can effectually recognize multiple objects in the video frame. The HSA with TWSVM model was also applied to properly categorize the existence of objects in the video frame and thereby lead to improved classification results. A wide range of experimental analyses were carried out on an open access dataset, and the results were inspected in several ways. The comparative simulation outcome reported the superiority of the CIHSA-RTODT technique over the other existing techniques. Therefore, the CIHSA-RTODT technique appeared to be an effective tool for real-time object detection and tracking. In the future, filtering techniques can be used as a pre-processing step that helps to boost the detection performance.

Author Contributions: Conceptualization, M.F.A.; Data curation, M.O.; Formal analysis, M.O. and S.A.-K.; Investigation, S.A.-K., E.K.; Project administration, E.K.; Resources, E.K.; Supervision, R.F.M.; Visualization, R.F.M.; Writing—original draft, M.F.A.; Writing—review & editing, R.F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Data sharing is not applicable to this article as no datasets were generated during the current study.

Acknowledgments: This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant No. (D-246-130-1442). The authors, therefore, gratefully acknowledge DSR's technical and financial support.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
2. Pal, S.K.; Pramanik, A.; Maiti, J.; Mitra, P. Deep learning in multi-object detection and tracking: State of the art. *Appl. Intell.* **2021**, *51*, 6400–6429. [CrossRef]
3. Połap, D.; Woźniak, M. Meta-heuristic as manager in federated learning approaches for image processing purposes. *Appl. Soft Comput.* **2021**, *113*, 107872. [CrossRef]
4. Hatwar, R.B.; Kamble, S.D.; Thakur, N.V.; Kakde, S. A review on moving object detection and tracking methods in video. *Int. J. Pure Appl. Math.* **2018**, *118*, 511–526.
5. Wieczorek, M.; Sika, J.; Wozniak, M.; Garg, S.; Hassan, M. Lightweight CNN model for human face detection in risk situations. *IEEE Trans. Ind. Inform.* **2021**. early access. [CrossRef]
6. Kaushal, M.; Khehra, B.S.; Sharma, A. Soft Computing based object detection and tracking approaches: State-of-the-Art survey. *Appl. Soft Comput.* **2018**, *70*, 423–464. [CrossRef]
7. Wen, L.; Du, D.; Cai, Z.; Lei, Z.; Chang, M.C.; Qi, H.; Lim, J.; Yang, M.H.; Lyu, S. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Comput. Vis. Image Underst.* **2020**, *193*, 102907. [CrossRef]
8. Połap, D.; Woźniak, M. Image features extractor based on hybridization of fuzzy controller and meta-heuristic. In Proceedings of the 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg, 11–14 July 2021; pp. 1–6.
9. Chouhan, S.S.; Kaul, A.; Singh, U.P. Image segmentation using computational intelligence techniques. *Arch. Comput. Methods Eng.* **2019**, *26*, 533–596. [CrossRef]
10. Połap, D.; Woźniak, M.; Mańdziuk, J. Meta-heuristic Algorithm as Feature Selector For Convolutional Neural Networks. In Proceedings of the 2021 IEEE Congress on Evolutionary Computation (CEC), Krakow, Poland, 28 June–1 July 2021; pp. 666–672.
11. Elhoseny, M. Multi-object detection and tracking (MODT) machine learning model for real-time video surveillance systems. *Circuits Syst. Signal Processing* **2020**, *39*, 611–630. [CrossRef]
12. Supreeth, H.S.G.; Patil, C.M. Efficient multiple moving object detection and tracking using combined background subtraction and clustering. *Signal Image Video Processing* **2018**, *12*, 1097–1105. [CrossRef]
13. Lyu, Y.; Yang, M.Y.; Vosselman, G.; Xia, G.S. Video object detection with a convolutional regression tracker. *ISPRS J. Photogramm. Remote Sens.* **2021**, *176*, 139–150. [CrossRef]
14. Xiong, Y.; Ge, Y.; From, P.J. An improved obstacle separation method using deep learning for object detection and tracking in a hybrid visual control loop for fruit picking in clusters. *Comput. Electron. Agric.* **2021**, *191*, 106508. [CrossRef]
15. Lin, X.; Li, C.T.; Sanchez, V.; Maple, C. On the detection-to-track association for online multi-object tracking. *Pattern Recognit. Lett.* **2021**, *146*, 200–207. [CrossRef]
16. Chen, H.; Cai, W.; Wu, F.; Liu, Q. Vehicle-mounted far-infrared pedestrian detection using multi-object tracking. *Infrared Phys. Technol.* **2021**, *115*, 103697. [CrossRef]
17. Schöller, F.E.T.; Blanke, M.; Plenge-Feidenhans, M.K.; Nalpantidis, L. Vision-based object tracking in marine environments using features from neural network detections. *IFAC-Pap.* **2020**, *53*, 14517–14523. [CrossRef]
18. Shi, J.; Wang, X.; Xiao, H. Real-Time Pedestrian Tracking and Counting with TLD. *J. Adv. Transp.* **2018**, *2018*, 8486906. [CrossRef]
19. Xie, H.; Wu, Z. A robust fabric defect detection method based on improved RefineDet. *Sensors* **2020**, *20*, 4260. [CrossRef]
20. Lydia, A.; Francis, S. Adagrad—An optimizer for stochastic gradient descent. *Int. J. Inf. Comput. Sci.* **2019**, *6*, 566–568.
21. Sadewo, W.; Rustam, Z.; Hamidah, H.; Chusmarsyah, A.R. Pancreatic Cancer Early Detection Using Twin Support Vector Machine Based on Kernel. *Symmetry* **2020**, *12*, 667. [CrossRef]
22. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
23. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189–215. [CrossRef]
24. Nalepa, J.; Kawulok, M. Selecting training sets for support vector machines: A review. *Artif. Intell. Rev.* **2019**, *52*, 857–900. [CrossRef]
25. Gao, X.Z.; Govindasamy, V.; Xu, H.; Wang, X.; Zenger, K. Harmony search method: Theory and applications. *Comput. Intell. Neurosci.* **2015**, *2015*, 39. [CrossRef] [PubMed]
26. Available online: http://www.Svcl.Ucsd.Edu/Projects/Anomaly/Dataset.Htm (accessed on 20 January 2022).

27. Pustokhina, I.V.; Pustokhin, D.A.; Vaiyapuri, T.; Gupta, D.; Kumar, S.; Shankar, K. An automated deep learning based anomaly detection in pedestrian walkways for vulnerable road users safety. *Saf. Sci.* **2021**, *142*, 105356. [CrossRef]
28. Xu, M.; Yu, X.; Chen, D.; Wu, C.; Jiang, Y. An efficient anomaly detection system for crowded scenes using variational autoencoders. *Appl. Sci.* **2019**, *9*, 3337. [CrossRef]
29. Murugan, B.S.; Elhoseny, M.; Shankar, K.; Uthayakumar, J. Region-based scalable smart system for anomaly detection in pedestrian walkways. *Comput. Electr. Eng.* **2019**, *75*, 146–160. [CrossRef]