electronics

MDPI

Article

Hybrid Classifiers for Spatio-Temporal Abnormal Behavior Detection, Tracking, and Recognition in Massive Hajj Crowds

Tarik Alafif [1], Anas Hadi [2], Manal Allahyani [2], Bander Alzahrani [2], Areej Alhothali [2,*], Reem Alotaibi [2] and Ahmed Barnawi [2]

[1] Department of Computer Science, Jamoum University College, Umm Al-Qura University, Makkah 25375, Saudi Arabia
[2] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
* Correspondence: aalhothali@kau.edu.sa

Abstract: Individual abnormal behaviors vary depending on crowd sizes, contexts, and scenes. Challenges such as partial occlusions, blurring, a large number of abnormal behaviors, and camera viewing occur in large-scale crowds when detecting, tracking, and recognizing individuals with abnormalities. In this paper, our contribution is two-fold. First, we introduce an annotated and labeled large-scale crowd abnormal behavior Hajj dataset, HAJJv2. Second, we propose two methods of hybrid convolutional neural networks (CNNs) and random forests (RFs) to detect and recognize spatio-temporal abnormal behaviors in small and large-scale crowd videos. In small-scale crowd videos, a ResNet-50 pre-trained CNN model is fine-tuned to verify whether every frame is normal or abnormal in the spatial domain. If anomalous behaviors are observed, a motion-based individual detection method based on the magnitudes and orientations of Horn–Schunck optical flow is proposed to locate and track individuals with abnormal behaviors. A Kalman filter is employed in large-scale crowd videos to predict and track the detected individuals in the subsequent frames. Then, means and variances as statistical features are computed and fed to the RF classifier to classify individuals with abnormal behaviors in the temporal domain. In large-scale crowds, we fine-tune the ResNet-50 model using a YOLOv2 object detection technique to detect individuals with abnormal behaviors in the spatial domain. The proposed method achieves 99.76% and 93.71% of average area under the curves (AUCs) on two public benchmark small-scale crowd datasets, UMN and UCSD, respectively, while the large-scale crowd method achieves 76.08% average AUC using the HAJJv2 dataset. Our method outperforms state-of-the-art methods using the small-scale crowd datasets with a margin of 1.66%, 6.06%, and 2.85% on UMN, UCSD Ped1, and UCSD Ped2, respectively. It also produces an acceptable result in large-scale crowds.

Keywords: abnormal behaviors; small-scale crowd; large-scale crowd; convolutional neural network; random forest; detection; tracking; recognition

Citation: Alafif, T.; Hadi, A.; Allahyani, M.; Alzahrani, B.; Alhothali, A.; Alotaibi, R.; Barnawi, A. Hybrid Classifiers for Spatio-Temporal Abnormal Behavior Detection, Tracking, and Recognition in Massive Hajj Crowds. *Electronics* **2023**, *12*, 1165. https://doi.org/10.3390/electronics12051165

Academic Editor: Silvia Liberata Ullo

Received: 28 January 2023
Revised: 22 February 2023
Accepted: 25 February 2023
Published: 28 February 2023

1. Introduction

Abnormal behavior detection in videos has been receiving lots of attention. This research area has been widely examined in the past two decades due to its importance and challenging nature in the computer vision domain. Generally, abnormal behavior is described as the unusual act of an individual in an event such as running, walking in the opposite direction, jumping, etc. Individual abnormal behaviors can be perceived differently in different contexts and scenes. Therefore, the definition of abnormal behaviors may vary from one place or scenario to another. Similarly, the density and the number of individuals in the crowd often vary significantly, which can result in small or large crowds according to the context of the scene. A small-scale crowd often contains approximately tens of individuals gathering or moving in the same location, while a large-scale crowd contains hundreds or thousands of individuals in the same place. Therefore, the large-scale crowd scene may raise many challenges as a result of many individuals moving to

or gathering in one location at the same time. In large-scale crowds, challenges such as partial and full occlusions, blurring, a large number of abnormal behaviors, and low scaling usually occur when detecting, tracking, and recognizing abnormal behaviors. As a result, detecting, tracking, and recognizing anomalous actions in large crowds is difficult, whereas performing comparable tasks in small crowds is easier.

To ensure safety in public places, many studies have tackled the problem of abnormality detection in crowd scenes. These studies have exploited a wide range of trajectory features [1–6], dense motion features [7–10], spatial–temporal features [11–13], or deep learning-based features and optimization techniques for anomaly recognition [14–18]. Most of the developed methods perform a binary frame-level for anomaly detection. Several studies considered locating anomalies in crowd surveillance videos [17–22], and less attention was paid to multi-class anomalies [23]. This study proposes a hybrid model that first identifies anomalies at the frame-level and then locates and classifies crowd anomalies into one of multiple classes. Distinguishing between different types of abnormal behaviors (e.g., running and walking against the crowd) raises many challenges that are worth researching. The current proposed methods in the field are also often evaluated on datasets with low-to-moderate crowd density levels. In this research, we evaluate the proposed methods on both moderate and very high crowd density levels of benchmark datasets.

Hajj is an annual religious pilgrimage that takes place in Makkah, Saudi Arabia. It is considered a large-scale event because it regularly attracts over two million pilgrims from various countries and continents who congregate in one location. The diversity and cultural differences of pilgrims reduce our ability to understand their abnormal behaviors. However, we define, annotate, and label a set of abnormal behaviors based on the context of the Hajj. The definition of abnormal behaviors has been studied thoroughly in this research and is associated with the causes of potential obstacles or dangers to large-scale crowd flows. This analysis aims to help automate the detection, tracking, and recognition of abnormal behaviors in large-scale crowds using surveillance cameras to ensure pilgrims' safety in a smooth flow during Hajj. It also helps security authorities and decision-makers visualize and anticipate potential risks.

Our work is inspired by the power of convolutional neural networks (CNNs) and transfer learning in many computer vision tasks [24–26]. In addition to the success of CNNs, the work is also motivated by the success of random forests (RFs) in the classification of unstructured data [27]. The contributions of this work are summarized as follows:

- We introduce a manually annotated and labeled large-scale crowd abnormal behaviors dataset for Hajj, HAJJv2;
- We propose two methods of hybrid CNN and RF classifiers to detect, track, and recognize spatio-temporal abnormal behaviors in small-scale and large-scale crowd videos;
- We evaluate the first proposed method on two common benchmark small-scale crowd video datasets, UMN and UCSD, against the currently published methods. Then, we evaluate the second proposed method on the HAJJv2 dataset and compare it with the previously existing method.

The remainder of this paper is organized as follows. We provide a literature review for abnormal behavior detection and recognition in Section 2. In Section 3, we briefly describe the abnormal behavior HAJJv2 dataset. Then, we present our proposed methods to detect, track, and recognize spatio-temporal abnormal behaviors in small and large crowd videos in Section 4. Experimental implementation, results, and evaluation are provided in Section 5. Then, a discussion on experimental evaluations, limitations, and challenges are provided in Section 6. Finally, we conclude our work and present some future directions in Section 7.

2. Related Work

Many research works have been proposed to detect abnormal behaviors in crowds in the past two decades. In this section, we provide the most recent related work. Current abnormal behavior detection and recognition methods can be briefly overviewed in two scales of crowds as follows:

- Small-scale crowds: Many recent studies have proposed and evaluated their methods on small-scale and common benchmark crowd public datasets, including UMN and UCSD [10,28–33].
  Piciarelli et al. [6] introduced a normal model by clustering the trajectories of moving objects for anomaly detection. Then, Mehran et al. [28] proposed to use an optical flows-based social force model to detect abnormal behaviors. A grid of particles was computed over the frames. Then, a bag of words method was applied to classify normal and abnormal behaviors.
  After Mehran et al. [28]'s work, Mahadevan et al. [29] applied learned mixtures of dynamic textures based on optical flow with salient location identification to detect abnormalities in the spatial domain. In the temporal domain, the learned mixtures of dynamic textures based on optical flow with negative log-likelihood were applied to detect abnormalities. Then, Cong et al. [32] applied a sparse reconstruction cost and a dictionary to measure normal and abnormal behaviors.
  After that, Zhang et al. [10] introduced a social attribute-aware force model. Using an online fusion algorithm, the social attribute-aware force maps are computed. Then, global abnormal events are detected with a bag-of-words representation and local abnormal events with an abnormal map.
  Later, Hasan et al. [30] learned semi-supervised spatio-temporal local hand-crafted features on a convolutional autoencoder to detect abnormal patterns. Histograms of oriented gradients and histograms of optical lows were used to extract the spatio-temporal features from raw video frames to feed the convolutional autoencoder for classification. Fradi et al. [13] applied local feature tracking to describe the movements of the crowd. They represented the crowd as an evolving graph. To analyze the crowd scene for an abnormal event, mid-level features are extracted from the graph.
  Colque et al. [7] used the histograms of magnitude, orientation, and entropy of the optical flow with the nearest neighbor search algorithm to detect the anomalies. In the training phase, they stored the histograms of each moving object as normal patterns. In the testing phase, they used the nearest neighbor search to find normal patterns to decide the abnormality.
  Coşar et al. [5] employed trajectory features and motion features. They used a bag-of-words representation to describe the actions. Then, they applied a clustering algorithm to perform abnormal detection in an unsupervised manner.
  Followed by [5,7,13,30], Tudor Ionescu et al. [31] used a sliding window technique to obtain partial video frames. The motions and appearance features were extracted from the frames and fed to a linear binary classifier to detect normality and abnormality in behaviors.
  Recently, Alafif et al. [33] applied a FlowNet and UNet framework to generate normal and abnormal optical flows to detect abnormalities. However, most current existing abnormal behavior detection methods are computationally expensive since they require modeling the appearance of the frames [29], particles advection [28], sliding windows [30,31], dictionaries [32], hand-crafted features extractors [30], and generating images [33]. In addition to the computational efficiency drawbacks, the effectiveness of their approach may decrease in large-scale crowds since they have many challenges, including partial and full occlusions, different scales, blurring, and a large number of abnormal behaviors.
- Large-scale crowds: Several research works studied abnormal behaviors on large-scale crowds [2,13,33–41].

First, Solmaz et al. [34] introduced a linear approximation using a Jacobian matrix to identify large-scale crowd abnormal behaviors. An optical flow and particle advection were used. Then, Wang et al. [35] started to cluster crowd feature maps to analyze motion patterns. Followed by [35], Alqaysi and Sasi [36] applied motion history image and segmented optical flow to extracted features. Then, a histogram was used for the motion direction and magnitude to detect crowd abnormal behaviors.

Later, Zou et al. [37] detected large-scale crowd motions and trajectories using tracklets association. Similar to [37], Bera et al. [2] computed abnormal behavior trajectories using Bayesian learning techniques. Then, Pennisi et al. [38] segmented the extracted features to detect crowd abnormal behaviors. In recent years, Fradi et al. [13] and Wu et al. [39] worked on analyzing large-scale crowd properties using visual feature descriptors. Then, Luo et al. [42] proposed a large-scale crowd motion framework for abnormal behavior detection. However, they focused on a crowd level rather than an individual level in their study. Finally, Miao et al. [40,41] leveraged unmanned aerial vehicles, airborne LiDAR, and computer vision technologies to continuously analyze individual abnormal behaviors in large-scale crowds.

However, existing methods are only confined to detecting and analyzing large-scale crowds as a mass. To the best of our knowledge, no existing works have detected individuals' abnormal behaviors in large-scale crowds, with the exception of the work presented in [33]. In comparison with the recent work in [33], the proposed methods do not require generating individual abnormal behavior images. Compared to the work in [33], the proposed method achieves better accuracy using the HAJJv1 dataset.

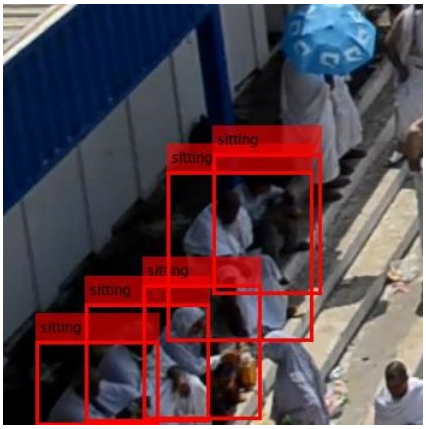
3. HAJJv2 Dataset

The HAJJv2 dataset is introduced due to the imbalance of training examples in each class and the absence of many annotations and labeling for individuals with abnormal behaviors in the HAJJv1 dataset [33]. The HAJJv2 dataset consists of nine manually collected videos from the annual Hajj religious event. All the videos are stored with an mp4 extension. The collected videos include individuals' abnormal behaviors in massive crowds. The videos are captured from different scenes and places in the wild during the Hajj event. Five videos are captured in the "Massaa" scene while other videos are captured in "Jamarat", "Arafat", and "Tawaf". These videos were recorded using high-resolution cameras. Then, the videos are cropped and split into training and testing sets. Each set contains nine short videos. Each video in the training set lasts for 25 s, while each video in the testing set lasts for 20 s.

In these videos, individuals' abnormal behaviors include standing, sitting, sleeping, running, moving in opposite or different crowd directions, and non-pedestrian entities such as cars and wheelchairs. These behaviors can be potentially dangerous for large-scale crowd flows. Figure 1 shows examples of these abnormal behaviors in the HAJJv2 dataset. The dataset statistics are provided in Table 1. As seen in the table, the dataset is imbalanced. The sitting class has the largest number of training and testing examples, while the running class has the smallest number of examples in the training and testing sets.

Table 1. HAJJv2 dataset statistics.

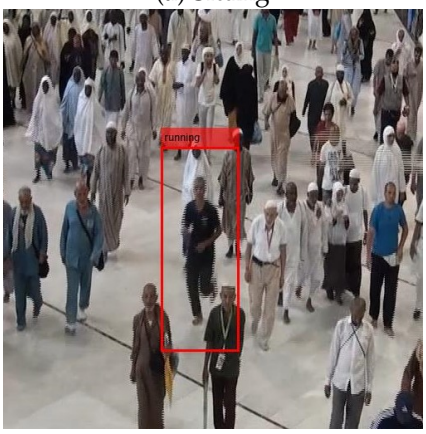
n	Classes	Training	Testing
1	Different Crowd Direction	7152	6262
2	Moving In Opposite	36,577	18,802
3	Moving Non Human Object	4186	4146
4	Running	51	190
5	Sitting	100,633	83,644
6	Sleeping	2400	2618
7	Standing	19,773	14,107
	Total	170,772	129,769



(**a**) Sitting

(**b**) Standing

(**c**) Sleeping



(**d**) Running

(**e**) Moving non-pedestrian

(**f**) Moving in different direction

(**g**) Moving in opposite direction

Figure 1. Abnormal behavior examples in HAJJv2 dataset.

Individuals' anomalous behaviors in the videos are manually annotated and labeled for the training and testing sets. The annotations and labeling are stored in two CSV files. The training CSV file contains 170,772 annotated and labeled individuals' abnormal behaviors, while the testing CSV file contains 129,769 annotated and labeled individuals' abnormal behaviors. A comparison of existing public abnormal behavior datasets and the abnormal behavior HAJJv2 dataset is shown in Table 2. The videos and HAJJv2 dataset are publicly available for research and non-commercial use only. The videos and HAJJv2 annotations and labeling files can be downloaded from https://github.com/KAU-Smart-Crowd/HAJJv2_dataset, accessed on 10 February 2023.

Table 2. A comparison of existing public abnormal behavior datasets and the abnormal behavior HAJJv2 dataset.

Dataset	Abnormal Behaviors	Size	Crowd Scale	Reference
UMN	Escape	24,240 KB	Small-scale	[43]
UCSD	Non-pedestrian movements	1.74 GB	Small-scale	[33]
HAJJv1	Standing, sitting, sleeping, running, moving in the opposite crowd direction, crossing or moving in different crowd direction, and non-pedestrian movements	831 MB	Large-scale	[43]
HAJJv2	Standing, sitting, sleeping, running, moving in the opposite crowd direction, crossing or moving in different crowd direction, and non-pedestrian movements	831 MB	Large-scale	–

4. Proposed Methods

In this section, we present the details of our proposed methods and algorithms. First, we show the individual abnormal behavior detection and recognition pipeline and algorithm for small-scale crowds. Then, similarly, a pipeline and an algorithm for detecting and recognizing abnormal behaviors are presented for large-scale crowds. Figure 2 shows the detection and recognition methodology pipelines for abnormal behaviors in small-scale and large-scale crowds.

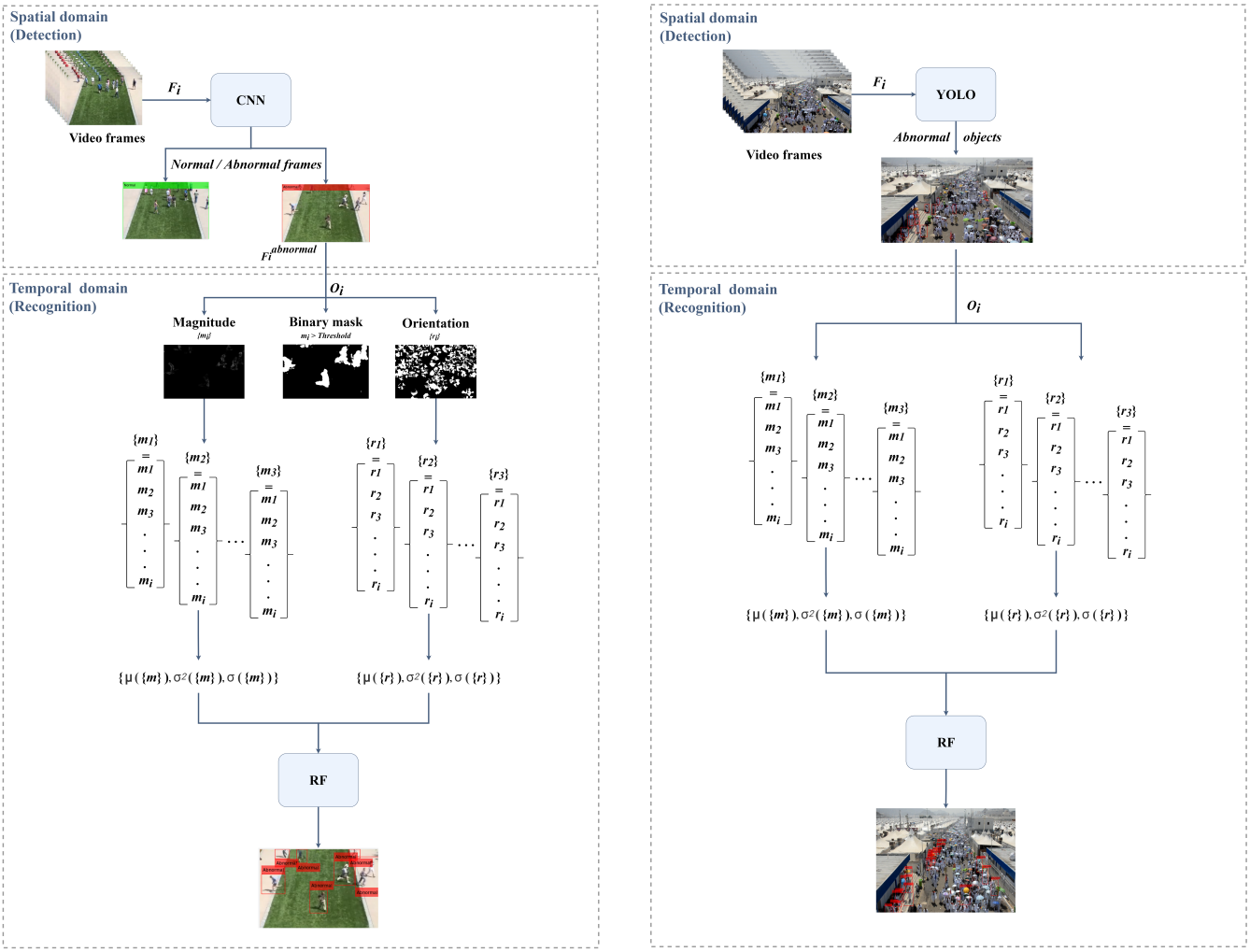
4.1. Individual Abnormal Behavior Detection, Tracking, and Recognition in Small-Scale Crowds

Figure 2a shows the pipeline for detecting and recognizing abnormal behaviors in small-scale crowd videos. The pipeline consists of spatial and temporal domains and hybrid classifiers. The spatial domain includes a pre-trained CNN classifier which focuses on classifying and detecting the abnormal behaviors generally on a frame level. On the other hand, the temporal domain includes the RF classifier that aims to classify and recognize individuals' behaviors at an object level within the frames.

Spatial domain: Training a specialized deep model from scratch requires a vast amount of data, a significant amount of resources, and a long training time. Transfer learning overcomes these challenges by utilizing pre-trained deep learning models that have been trained on a significant amount of labeled data and using the previously optimized weights to perform other predictive tasks. Due to the lack of sufficient abnormal training datasets, we utilize transfer learning in the spatial domain. We fine-tune the pre-trained model, ResNet-50 [26], to detect abnormalities at the frame level. Deeper networks are capable of extracting more complex feature patterns; however, they may cause a degradation problem, which degrades the detection performance. ResNet generally uses a deep residual learning framework to solve the degradation problem. This gives the advantage of using a deep neural network to extract the complex feature patterns in the spatial domain. Therefore, we use ResNet-50 in our experiments.

ResNet-50 consists of 49 convolutional layers as a feature extractor, followed by average pooling and a fully connected layer as a classifier. Fine-tuning the pre-trained models is performed by modifying the previous weights of the model such that they work with a new classification task. The classification layers of the pre-trained model are replaced by a fully connected layer and an output layer that outputs values equal to the number of classes. Anomaly detection is a binary classification problem. Thus, the classifier is trained on normal and abnormal frames. Therefore, we fine-tune ResNet-50 as a binary classifier using video frames from small-scale crowd datasets. We replace the last layer with a fully connected layer that maps 2048 units into 128, followed by an output layer

that maps 128 units into 2 units, representing the normal and abnormal probabilities. Since the ResNet-50 model processes inputs with a size of 224 × 224 × 3, we resize the frames to input size. A feed-forward and back-propagation algorithm is applied by updating the errors and weights to converge. Figure 2a shows the detected normal and abnormal frames resulted from the ResNet-50 classifier. The detected normal frames appear in green, while the detected abnormal frames appear in red.



(**a**) (**b**)

Figure 2. The proposed pipelines for individuals' abnormal behaviors detection, tracking, and recognition. (**a**) Small-scale crowds such as in UMN and UCSD datasets; (**b**) Large-scale crowds in HAJJv2 dataset.

Temporal domain: We use the optical flow to detect the anomalies at the pixel level. By analyzing the optical flow, we can observe crowd movements, instantaneous velocities, orientations, and magnitudes. These low-level features are used to recognize individuals' behaviors. After detecting the anomalies at the i-th frame (F_i), the optical flow of this frame (O_i) is computed using Horn–Schunck optical flows [44]. Then, magnitude (m_i) and orientation (r_i) features are automatically extracted from the optical flows. In small-scale crowd videos, binary magnitude-based masks using a threshold (T) are initiated to localize and track individuals within the frame (i.e., $\{I_j^i\}_{j=1}^z$ is the j-th individual in the i-th frame). Figure 3 shows the proposed binary magnitude-based mask using the small-scale crowd datasets. After extracting the magnitude and orientation features, the statistical features including means (μ) and variances (σ^2) are computed. They are computed for the total

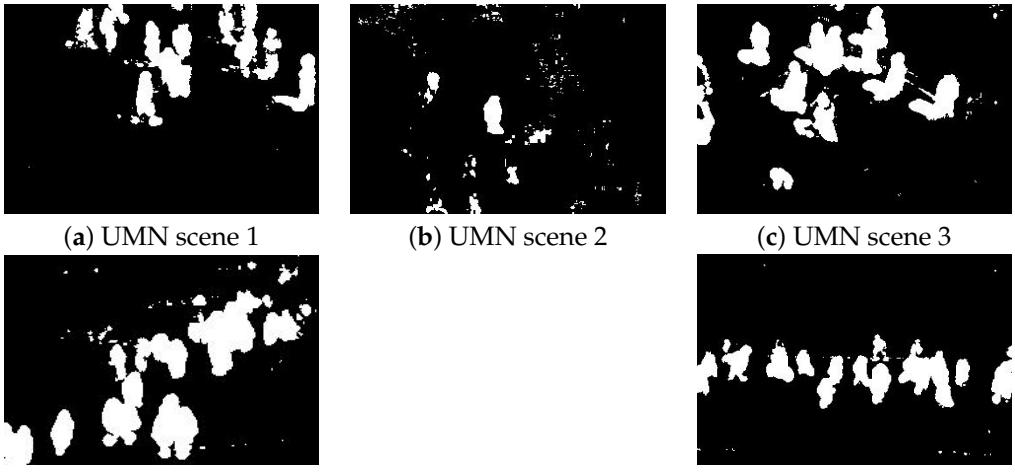
pixels (p) of the area that represent an individual (j), for both the magnitudes (m^j) and orientations (r^j), as follows:

$$\begin{aligned} \mu(m^j) &= \frac{\sum_{i=1}^{p}(m_i^j)}{p} \\ \mu(r^j) &= \frac{\sum_{i=1}^{p}(r_i^j)}{p} \\ \sigma^2(m^j) &= \frac{\sum_{i=1}^{p}((m_i^j - \mu(m^j))^2)}{p} \\ \sigma^2(r^j) &= \frac{\sum_{i=1}^{p}((r_i^j - \mu(r^j))^2)}{p} \end{aligned} \tag{1}$$

Then, these statistical features are fed to the RF classifier for training to classify and recognize individual temporal abnormal behaviors. Algorithm 1 shows the computational steps of the proposed method in the small-scale crowds. The algorithm runs $O(n^2)$ in the worst case.

Algorithm 1: A hybrid CNN and RF algorithm for spatio-temporal small-scale crowd abnormal behavior detection, tracking, and recognition in a video.

Input : Video frame sequences $\{F_i\}_{i=1}^{n}$, where F_i consists of a number of frames f such that $F_i = \{f_1, f_2, \ldots, f_n\}$.
Output: Abnormal behavior frames and objects.
Use F_i to fine-tune a pre-trained CNN model in the spatial domain using feed-forward and back-propagation algorithm until convergence and update the weights;
Compute optical flow $\{O_i\}_{i=1}^{n-1}$ from the original video sequence F_i : $F_i \longrightarrow \{O_1, O_2, \ldots, O_{n-1}\}$;
Extract optical flow orientations $\{r_i\}_{i=1}^{n}$ and magnitudes $\{m_i\}_{i=1}^{n}$ features from O_i : $O_i = \{(r_1, m_1), (r_2, m_2), \ldots, (r_{n-1}, m_{n-1})\}$;
Create the binary magnitude-based mask using a threshold T, $mask = m_i > T$;
Extract the individuals within the mask $\{I_j^i\}_{j=1}^{z}$;
Compute orientations and magnitudes means $\{\mu_j^i\}_{j=1}^{z}$ and variances $\{\sigma^{2i}_j\}_{j=1}^{z}$;
Use the statistical features $\{\mu_j^i\}_{j=1}^{z}$ and $\{\sigma^{2i}_j\}_{j=1}^{z}$ of the individuals of temporal normal and abnormal behaviors to train the RF model;
Insert test video frame sequences $\{F_i^t\}_{i=1}^{n}$;
while $F_i^t \neq empty$ **do**
    Use F_i^t to test the fine-tuned CNN model;
    if F_i^t *is abnormal* **then**
        Compute optical flows O_i^t;
        Extract optical flows orientations r_i^t and magnitudes m_i^t features from O_i^t;
        Create binary magnitude-based mask to localize and track individuals $\{I_j^{t,i}\}_{j=1}^{z}$;
        while $\{I_j\}^{t,i} \neq empty$ **do**
            Test the statistical features: means $\mu_j^{t,i}$ and variances $\sigma^{2t,i}_j$ features to classify $I_j^{t,i}$ using the trained RF model;
        end
    end
end



(**a**) UMN scene 1 (**b**) UMN scene 2 (**c**) UMN scene 3

(**d**) UCSD Ped1 (**e**) UCSD Ped2

Figure 3. Examples of our proposed binary magnitude-based masks on the small-scale crowd datasets.

4.2. Individual Abnormal Behavior Detection and Recognition in Large-Scale Crowds

Figure 2b depicts the pipeline for detecting and recognizing abnormal behaviors in large-scale crowded scenes in the HAJJv2 dataset. Similar to the method presented for small-scale crowded scenes, the method for large-scale crowded scenes also consists of hybrid classifiers. The first classifier is accountable for detecting the abnormal behavior frames in the spatial domain, while the second classifier is employed for recognizing the individuals' abnormal behaviors in the temporal domain.

Spatial domain:Similar to the previous detection method for small-scale crowds, we fine-tune another pre-trained CNN model, ResNet-50. We train ResNet-50 as a one-class classifier using all abnormal behaviors in the training set of the HAJJv2 dataset. The main goal of the ResNet-50 model is to only detect individuals' abnormal behaviors in frames if they exist. To address the problem of the overlapped white areas when a large number of individuals and a large number of partial occlusions occur, the YOLOv2 [45] technique is employed to locate individuals with the abnormal behaviors in the spatial domain. We use the back-propagation algorithm to update the errors and weights in the ResNet-50 model until convergence.

Temporal domain: After detecting all individuals with abnormal behaviors, we employ Horn–Schunck optical flows (O_i) on the detected individuals. This approach is different from the previous method applied for small-scale crowds. We extract the m_i and the r_i features from the resulted optical flows. To track individuals with abnormal behaviors, a Kalman filter [46] is used directly with a YOLOv2 detector. The Kalman filter predicts individuals' locations in the next frames. We avoid using our binary magnitude-based masks since they mainly cause overlapping contiguous groups of white pixels due to heavy partial occlusions in large-scale crowded scenes.

After detecting individuals' abnormal behaviors and extracting their statistical features, we compute the means (μ) and variances (σ^2) for each individual, similar to the small-scale crowd method. Then, another RF classifier is used as a multi-class classifier to classify and recognize all individuals with abnormal behaviors.

Algorithm 2 shows the sequence of our implementable method in large-scale crowd videos. Similar to Algorithm 1, Algorithm 2 also runs $O(n^2)$ in the worst case.

Algorithm 2: A hybrid CNN and RF algorithm for spatio-temporal large-scale crowd abnormal behavior detection, tracking, and recognition in a video.

Input : Video frame sequence $\{AB_i\}_{i=1}^{n}$, where AB_i consists of a number of abnormal behavior examples ab such that $AB_i = \{ab_1, ab_2, \ldots, ab_n\}$.
Output: Abnormal behavior objects.
Use AB_i as one class to fine-tune a pre-trained CNN model in the spatial domain using feed-forward and back-propagation algorithm until convergence and update the weights;
Compute the optical flow $\{O_i\}_{i=1}^{n-1}$ from the original video sequence AB_i : $AB_i \longrightarrow \{O_1, O_2, \ldots, O_{n-1}\}$;
Extract optical flow orientations $\{r_i\}_{i=1}^{n-1}$ and magnitudes $\{m_i\}_{i=1}^{n-1}$ from O_i : $O_i = \{(r_1, m_1), (r_2, m_2), \ldots, (r_{n-1}, m_{n-1})\}$;
Compute the orientations and magnitudes means $\{\mu_j^i\}_{j=1}^{z}$, and variances $\{{\sigma^2}_j^i\}_{j=1}^{z}$ for each abnormal behavior example ab_j within the frame AB_i;
Train multi-class temporal abnormal behavior features using an RF model;
Insert test video frame sequences $\{AB_i^t\}_{i=1}^{n}$;
while $\{AB_i^t\} \neq empty$ **do**
    Use AB_i^t to test the fine-tuned CNN model;
    Compute the optical flow O_i^t;
    while $\{ab_j^{i,t}\}_{j=1}^{z} \neq empty$ **do**
        Compute the orientations and magnitudes means $\mu_j^{i,t}$ and variances ${\sigma^2}_j^{i,t}$ for $ab_j^{i,t}$;
        Use $\{\mu_j^{i,t}, {\sigma^2}_j^{i,t}\}$ to test the trained RF model;
        Use Kalman filter to track $ab_j^{i,t}$;
    end
end

5. Experiments

In this section, we first provide details of the implementation of the proposed methods. Second, we briefly describe the benchmark datasets used in the experiments. Third, we show the results of our abnormal behavior detection and recognition qualitative and quantitative experiments. Then, we compare the results with the existing and the most recent methods for abnormal behavior detection in small and large crowds.

5.1. Implementation

We implemented the proposed methods in MATLAB R2020b. The ResNet-50 and the RF models were trained using NVIDIA Tesla V100S GPU server with 32GB of RAM.

5.2. Datasets

In this section, we use the most common and public benchmark datasets such as the UMN [43], UCSD [29], HAJJv1 [33], and HAJJv2 datasets to evaluate the proposed method on small-scale and large-scale crowds. HAJJv2 is described in Section 3. The UMN and USCD datasets are briefly described as follows:

- The University of Minnesota (UMN) dataset. The UMN dataset is a small-scale crowd dataset that contains three different unrealistic scenes. Two scenes were recorded outdoors, while one was recorded indoors. Each UMN scene starts with a normal activity followed by an abnormal behavior. Walking, for example, is considered a normal activity, while running is an abnormal one. The frame resolution in UMN scenes is 320 × 240 pixels. The abnormal frames contain a short description at the top of the frames. Thus, we apply a pre-processing technique on the frames to remove the pixels that contain these descriptions to avoid biases in training and testing the model

in the experiment. Figure 4 illustrates an example of UMN's frames. The training and testing splits are not explicitly specified. Moreover, the annotations are only available at the frame level. Due to these ambiguities, we use 70% of the frames for training and the rest for testing. To address the lack of pixel-level annotations, we consider all objects in the abnormal frames as abnormal individuals and all objects in the normal frames as normal individuals. The UMN scenes are evaluated separately since they have illumination and background variations.

- The University of California, San Diego (UCSD) dataset. The UCSD dataset is also a small-scale crowd dataset that consists of two subsets, namely Pedestrian 1 (Ped1) and Pedestrian 2 (Ped2). The dataset contains clips from independent static cameras viewing pedestrian walkways. It includes abnormal behaviors such as bicycles, cars, carts, skateboards, and wheelchairs as non-pedestrian objects. Ped1 contains 34 normal behavior videos and 16 abnormal behavior videos. Each video contains 200 frames with a resolution of 238 $\times$ 158 pixels. Ped2 contains 16 normal behavior videos and 12 abnormal behavior videos. The videos have different numbers of frames with a resolution of 360 $\times$ 240 pixels. Both temporal and spatial annotations are provided. Thus, the UCSD is appropriate for locating and tracking abnormal objects in small-scale crowds. In our experiment, we use both normal and abnormal videos for training and testing. Figure 5 illustrates some examples from Ped1 and Ped2 frames.

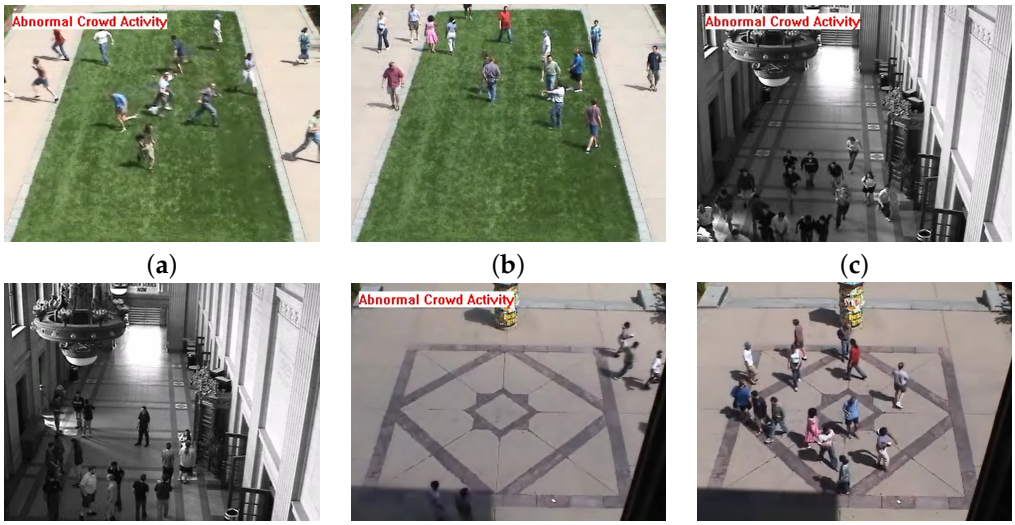


Figure 4. Examples of normal and abnormal behaviors in UMN dataset from three different indoor and outdoor scenes. (**a**) Outdoor abnormal behaviors in scene 1; (**b**) Outdoor normal behaviors in scene 1; (**c**) Indoor abnormal behaviors in scene 2; (**d**) Indoor normal behaviors in scene 2; (**e**) Outdoor abnormal behaviors in scene 3; (**f**) Outdoor normal behaviors in scene 3.

Figure 5. *Cont.*

(**c**) (**d**)

Figure 5. Examples of normal and abnormal behaviors in UCSD dataset from two different outdoor scenes. (**a**) Abnormal Ped1; (**b**) Normal Ped1; (**c**) Abnormal Ped2; (**d**) Normal Ped2.

5.3. Experimental Settings and Hyperparameters

For both small-scale and large-scale crowd experiments, different configurations are evaluated to determine the most effective approach, the details of which are described in the following.

Small-scale crowds: Different pre-trained CNN models such as ResNet-50, VGG-16, VGG-19, AlexNet, and SqueezeNet were examined in the spatial domain as part of the proposed method. According to our preliminary experiments, the ResNet-50 model achieves better performance on small-scale crowd datasets. We fine-tune the ResNet-50 model using the Adam optimizer [47] with a learning rate of 0.0001 for 15 epochs and 128 normal and abnormal frames per batch of each dataset.

Many methods to estimate optical flow, such as the Lucas–Kanade derivative of Gaussian, Lucas–Kanade, Farneback, and Horn–Schunck, are employed. The Horn–Schunck [44] method is selected since it provides magnitude and orientation features to create a binary magnitude-based mask to localize and track individuals. The means and variances are computed using these features to classify the individuals' abnormal behaviors in small-scale crowds.

In addition to using different pre-trained CNN models and optical flow estimators, different classifiers are examined, such as the linear classifier, decision tree, and RF with cross-validation. The RF classifier is selected since it achieves better results compared to the other classifiers.

Large-scale crowds: The ResNet-50 and SqueezeNet pre-trained CNN models are used as the base network of the YOLOv2 object detection technique. We initialize the weights on ImageNet [48]. Then, we fine-tune the model with a stochastic gradient descent (SGD) [49] optimizer for 20 epochs with a learning rate of 0.001 and a mini-batch of eight frames.

Similar to the small-scale crowd experiment, the RF classifier is also accountable for the recognition of abnormal behavior in the temporal domain. Unlike in the small-scale crowd experiment, we train the RF classifier using only the statistical features of the detected individuals with abnormal behaviors.

5.4. Effectiveness Evaluation

To evaluate the proposed methods, we evaluate them in both spatial and temporal domains. In the spatial domain, we consider the accuracy, precision, recall, *F*1 score, and area under the curve (AUC) metrics as performance measures. The accuracy, precision, and

recall metrics are defined in terms of true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*) as the following:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ F1score &= \frac{2 \times Precision \times Recall}{Precision + Recall} \end{aligned} \tag{2}$$

The receiver operating characteristics (ROCs) curve [50] is a plot of the true positive rate (TPR) and false positive rate (FPR). The ROC curve represents the change of TPR and FPR over different thresholds. Thus, it is a powerful metric to evaluate a classifier. However, it is difficult to compare different classifiers using the ROC curve. Therefore, the AUC is used to compute the area under the ROC curve and compare the performance of the classifiers. The AUC scores range from zero to one. Stronger classifiers have higher AUC scores.

Small-scale crowds: Table 3 shows the frame detection quantitative results in the spatial domain using the small-scale crowd datasets. The ResNet-50 classifier achieves 99.76% and 93.71% of average AUCs among the scenes on the UMN and UCSD datasets, respectively. Table 4 shows the quantitative results in the temporal domain using the RF classifier on small-scale crowd datasets. Figure 6a,b illustrate the ROC curves of our experiments using the ResNet-50 and the RF classifiers, respectively, on UMN and UCSD datasets. In Figure 7, samples of our qualitative results using the UMN and UCSD datasets are shown. One can notice that the proposed method detects and recognizes the abnormality correctly in the datasets' testing samples.

To better illustrate the comparison with existing methods in [28,32,33], Table 5 shows that the proposed method yields better results using the UMN dataset.

Table 6 reports a performance comparison of the proposed method with the existing methods [28–31,33,51] using the UCSD dataset. It is clearly shown that the proposed method achieves higher AUCs using UCSD Ped1 and Ped2 scenes.

Table 3. Our results using the ResNet-50 classifier on the public and benchmark small-scale crowd datasets.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
UMN scene 1	97.93	98.31	99.15	98.73	99.73
UMN scene 2	98.49	99.36	98.82	99.09	99.79
UMN scene 3	98.07	99.46	98.41	98.93	99.77
UCSD Ped1	75.72	64.72	89.31	75.05	88.87
UCSD Ped2	94.14	96.29	92.11	94.15	98.55

Table 4. Our results of the RF classifier on small-scale crowd datasets.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
UMN scene 1	88.85	99.34	87.35	92.96	97.00
UMN scene 2	81.07	99.06	76.23	86.16	94.45
UMN scene 3	93.33	99.40	93.32	96.26	97.38
UCSD Ped1	99.49	99.60	99.88	99.74	97.66
UCSD Ped2	99.62	99.76	99.86	99.81	97.43

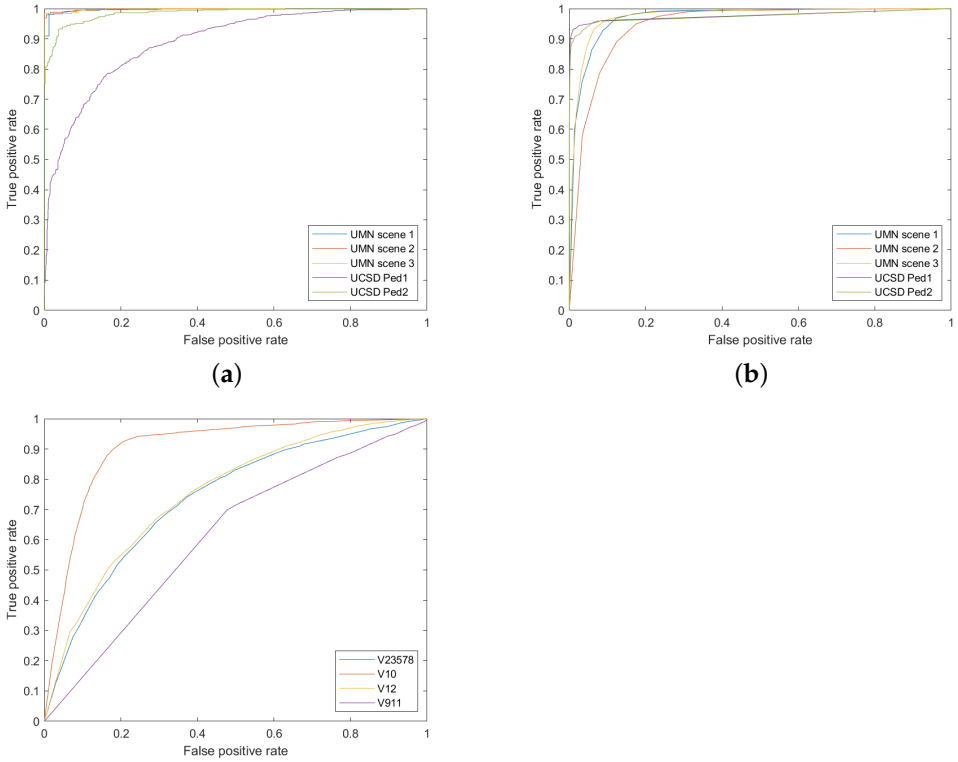


Figure 6. The ROC curves for ResNet-50 and RF classifiers. (**a**) ResNet-50 classifier on the small-scale crowd datasets; (**b**) RF classifier on the small-scale crowd datasets; (**c**) RF classifier on the HAJJv2 dataset.

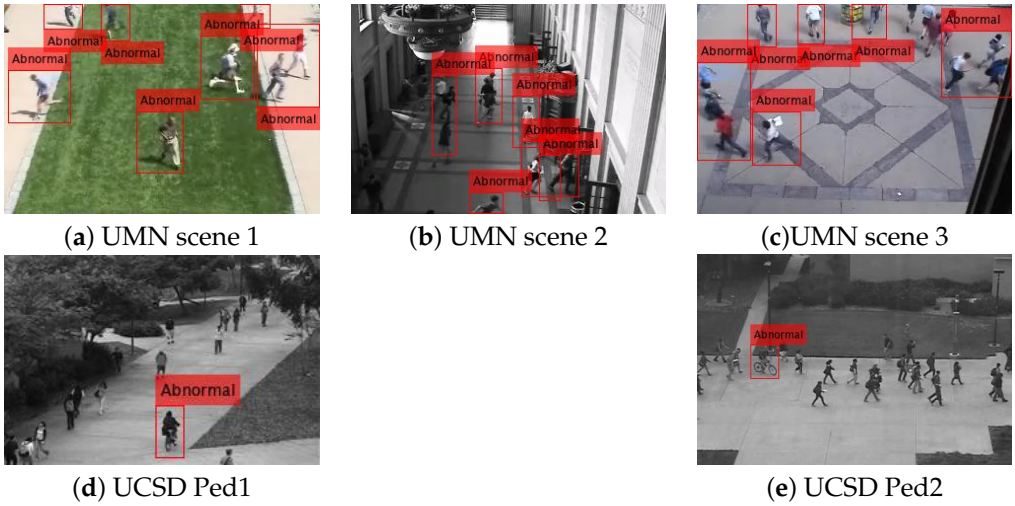


Figure 7. Our qualitative results on small-scale crowd datasets.

Table 5. The evaluation results on the UMN dataset. Percentages are AUCs.

Method	UMN (%)
Optical flow [28]	84.0
SFM [28]	96.0
Sparse Reconstruction [32]	97.0
Alafif et al. [33]	98.1
Ours	**99.76**

Large-scale crowds: We evaluate our method in the spatial and temporal domains on large-scale crowds using the HAJJv1 and HAJJv2 datasets. The spatial domain is evaluated using two criteria, track assignment and intersection over union (IOU). The results of track assignment are computed using Kalman filter assignment results for each detected object. Nevertheless, it is not important whether the detected pixels match most of the labeled pixels exactly. Thus, we use the IOU to evaluate the YOLOv2 detector. The IOU is a powerful evaluation metric to evaluate the detection of objects, as it is commonly used in the computer vision community. It computes the overlap ratio between the ground-truth and detected boxes. Then, using a 50% threshold of the overlapping boxes, we compute the TP, FP, and FN. The accuracy of the YOLOv2 is computed at the pixel level. A pixel is considered a TN if no TP, FP, and FN pixels are detected by the detector. It is observable that the accuracy cannot report the performance well. Since YOLOv2 does not detect any anomalies at most of the frames' pixels, and since the majority of the frames' pixels do not contain abnormal behaviors, the TN number is increased. This affects the accuracy computation and neglects the values of TP, FP, and FN.

Tables 7 and 8 show our quantitative results in the spatial and temporal domains using the HAJJv2 dataset. Our fine-tuned pre-trained ResNet-50 model with the YOLOv2 detector achieves 91.77%, 92.47%, 27.99%, and 36.05% of average accuracy, average precision, average recall, and average F1 score, respectively, using the track assignment technique. Meanwhile, the same model achieves 92.72%, 31.68%, 16.49%, and 20.62% of average accuracy, average precision, average recall, and average F1 score, respectively, using the IOU technique. In the temporal domain, the RF classifier achieves 75.18% of AUC for abnormal behavior recognition using the HAJJv2 dataset. Figure 8 shows the qualitative results for the proposed method on the HAJJv2 dataset. Figure 6c shows the ROC curves for the RF classifier. A quantitative comparison with the work in [33] using the HAJJv1 dataset is also provided in Table 9.

Table 6. The evaluation results on the UCSD dataset. The percentages are AUCs.

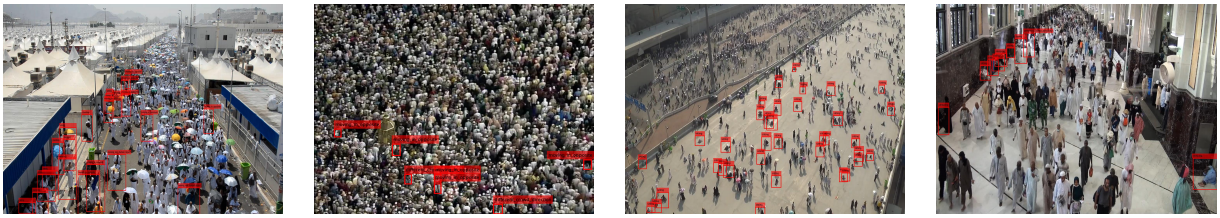
Method	UCSD Ped1	UCSD Ped2
MPPCA [51]	59.0%	69.3%
Social Force[SF] [28]	67.5%	55.6%
SF+MPPCA [29]	68.8%	61.3%
MDT [29]	81.8%	82.9%
Conv-AE [30]	75.0%	85.0%
Stacked RNN [52]	N/A	92.2%
Unmasking [31]	68.4%	82.2%
Alafif et al. [33]	82.81%	95.7%
Ours	**88.87%**	**98.55%**

Table 7. Abnormal behavior recognition results using the RF classifier on the HAJJv2 dataset.

Video No.	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
10 (Arafat)	60.73	62.01	55.95	58.83	90.38
12 (Tawaf)	63.62	64.95	53.75	57.80	75.25
9 and 11 (Jamarat)	96.83	33.47	33.02	33.24	61.19
2, 3, 5, 7, and 8 (Masaa)	51.90	33.51	28.87	31.02	73.89
Average	68.27	48.49	42.90	45.22	75.18

Table 8. The experimental results of track assignment and IOU detections using YOLOv2 on the HAJJv2 dataset.

	Track Assignment				IOU			
Video No.	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
10 (Arafat)	89.86	96.41	54.27	69.44	91.25	56.16	31.61	40.45
12 (Tawaf)	96.66	98.84	2.92	5.67	96.87	4.65	0.14	0.27
9 and 11 (Jamarat)	89.26	96.44	3.85	7.40	89.53	14.24	0.57	1.09
2, 3, 5, 7, and 8 (Masaa)	91.30	78.19	50.93	61.69	93.23	51.66	33.65	40.75
Average	91.77	92.47	27.99	36.05	92.72	31.68	16.49	20.62



(**a**) (**b**) (**c**) (**d**)

Figure 8. Our qualitative results on large-scale crowds dataset using HAJJv2. (**a**) Video No. 10 from Arafat scene; (**b**) Video No. 12 from Tawaf scene; (**c**) Video No. 9 from Jamarat scene; (**d**) Video No. 5 from Masaa scene.

Table 9. A comparison table of abnormal behavior detection performance with the recent existing methods using the HAJJv1 dataset against existing methods.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1(%)**	**AUC (%)**
Alafif et al. [33]	65.10	61.48	80.30	N/A	79.63
YOLOv2 (Ours)	**95.67**	9.42	28.82	10.99	N/A
RF (Ours)	41.81	9.69	10.23	9.96	54.40

6. Discussion

The proposed methods are robust in detecting and recognizing individuals with abnormal behaviors in small-scale and large-scale crowd videos. The results show that the small-scale crowd method achieves a great performance in comparison with the state-of-the-art techniques. Although the small-scale method outperforms other existing techniques, it shows unsatisfactory performance when using the UCSD Ped1 dataset. Several factors contributed to this, including the low resolution of the frame, the camera viewing, the shadows cast by trees, and the low illumination. In the large-scale crowds, we still have not achieved an excellent performance using the HAJJv2 dataset since, the videos in the dataset are very challenging to analyze. The challenges are represented by a far-away camera viewing as well as heavy, partial, and full occlusions with a significant number of individuals. Figure 8b,c shows some of the challenges in the Tawaf and Jamarat scenes, which are considered the hardest scenes for the classifiers to classify the individuals with abnormal behaviors. Due to the fact that these scenes contain a large number of individuals moving in one spot with heavy partial occlusions and far camera views, much more human attention and focus were required when annotating and labeling the abnormal behaviors. On the other hand, the easiest scenes for the annotators and labelers are in the Masaa scenes, since these videos are captured from a closed camera view and have a moderate number of partial occlusions. Therefore, these factors definitely contribute to the performance of the abnormal behavior detection and recognition classifiers. Much more future work is required to better detect and recognize the individuals with abnormal behaviors in large-scale and massive crowds.

7. Conclusions

In this research work, we first introduced the annotated and labeled large-scale crowd abnormal behavior dataset, HAJJv2. Second, we proposed two methods of hybrid CNNs and RFs to detect and recognize spatio-temporal abnormal behaviors in small-scale and large-scale crowd videos. In the small-scale crowd videos, a ResNet-50 pre-trained CNN model was fine-tuned to verify every frame, determining whether it is normal or abnormal in the spatial domain. If abnormal behaviors were found, a motion-based individual detection using the magnitude and orientation features of Horn–Schunck optical flow was employed to create a binary magnitude-based mask to localize and track individuals with abnormal behaviors. In large-scale crowd videos, a Kalman filter was employed to predict and track the detected individuals in the next frame. Then, means and variances as statistical features were computed and fed to the RF classifier to classify individuals with abnormal behaviors in the temporal domain. In the large-scale crowd videos, we fine-tuned the ResNet-50 model using the YOLOv2 object detection technique to detect individuals with abnormal behaviors in the spatial domain. The proposed method in a small-scale crowd achieved 99.76% and 93.71% average AUCs on the UMN and UCSD datasets, respectively, while the method in a large-scale crowd achieved 76.08% average AUC on the HAJJv2 dataset. Our method outperformed state-of-the-art methods using the small-scale crowd datasets with a margin of 1.67%, 6.06%, and 2.85% on the UMN, UCSD Ped1, and UCSD Ped2 datasets, respectively. It also achieved a satisfactory result for large crowds.

Still, a significant amount of work is needed to increase the effectiveness of abnormal behavior detection and recognition in large-scale crowded scenes due to their challenges. The majority of current research only uses small-scale crowded scenes in which abnormal behaviors can be easily extracted and classified. In the future, our work will be more focused on large-scale crowds. We will incorporate an attention mechanism and fusion strategies to enhance the performance. This work can potentially help researchers study and apply it in different contexts of crowded scenes, such as in airports, stadiums, and marathons. It can also be used in the manufacturing industry to inspect and detect abnormal behaviors of defective manufactured goods and products on a production line [53]. Examples and features of the products' unusual behaviors are required to be collected, extracted, and learned by a classifier to achieve high performance.

Author Contributions: Conceptualization, T.A. and A.H.; methodology, T.A.; software, T.A. and A.H.; validation, T.A., A.H., M.A. and A.A.; formal analysis, T.A.; investigation, T.A., A.H. and A.B.; resources, B.A.; data curation, B.A., T.A. and A.H.; writing—original draft preparation, T.A., A.H., M.A. and A.A.; writing—review and editing, T.A., M.A., A.A., B.A., R.A. and A.B.; visualization, T.A., A.H. and M.A.; project administration, B.A.; funding acquisition, B.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Deputyship for Research and Innovation, Ministry of Education, Saudi Arabia, project number (227).

Data Availability Statement: The HAJJv2 dataset is available on https://github.com/KAU-Smart-Crowd/HAJJv2_dataset (accessed on 25 April 2020).

Acknowledgments: The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number (227).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, Y.; Qin, L.; Ji, R.; Zhao, S.; Huang, Q.; Luo, J. Exploring coherent motion patterns via structured trajectory learning for crowd mood modeling. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 635–648. [CrossRef]
2. Bera, A.; Kim, S.; Manocha, D. Realtime Anomaly Detection Using Trajectory-Level Crowd Behavior Learning. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 27–30 June 2016; pp. 1289–1296. [CrossRef]

3. Zhou, S.; Shen, W.; Zeng, D.; Zhang, Z. Unusual event detection in crowded scenes by trajectory analysis. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 1300–1304. [CrossRef]
4. Zhao, K.; Liu, B.; Li, W.; Yu, N.; Liu, Z. Anomaly Detection and Localization: A Novel Two-Phase Framework Based on Trajectory-Level Characteristics. In Proceedings of the 2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW), San Diego, CA, USA, 23–27 July 2018; pp. 1–6. [CrossRef]
5. Coşar, S.; Donatiello, G.; Bogorny, V.; Garate, C.; Alvares, L.O.; Brémond, F. Toward abnormal trajectory and event detection in video surveillance. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 683–695. [CrossRef]
6. Piciarelli, C.; Micheloni, C.; Foresti, G.L. Trajectory-based anomalous event detection. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1544–1554. [CrossRef]
7. Colque, R.V.H.M.; Caetano, C.; de Andrade, M.T.L.; Schwartz, W.R. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 673–682. [CrossRef]
8. Cho, S.H.; Kang, H.B. Abnormal behavior detection using hybrid agents in crowded scenes. *Pattern Recognit. Lett.* **2014**, *44*, 64–70. . [CrossRef]
9. Qasim, T.; Bhatti, N. A hybrid swarm intelligence based approach for abnormal event detection in crowded environments. *Pattern Recognit. Lett.* **2019**, *128*, 220–225. [CrossRef]
10. Zhang, Y.; Qin, L.; Ji, R.; Yao, H.; Huang, Q. Social attribute-aware force model: Exploiting richness of interaction for abnormal crowd detection. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *25*, 1231–1245. [CrossRef]
11. Guo, H.; Wu, X.; Cai, S.; Li, N.; Cheng, J.; Chen, Y.L. Quaternion Discrete Cosine Transformation Signature Analysis in Crowd Scenes for Abnormal Event Detection. *Neurocomputing* **2016**, *204*, 106–115. [CrossRef]
12. Yuan, Y.; Fang, J.; Wang, Q. Online Anomaly Detection in Crowd Scenes via Structure Analysis. *IEEE Trans. Cybern.* **2015**, *45*, 548–561. [CrossRef] [PubMed]
13. Fradi, H.; Luvison, B.; Pham, Q.C. Crowd behavior analysis using local mid-level visual descriptors. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 589–602. [CrossRef]
14. Chan, T.H.; Jia, K.; Gao, S.; Lu, J.; Zeng, Z.; Ma, Y. PCANet: A Simple Deep Learning Baseline for Image Classification? *IEEE Trans. Image Process.* **2015**, *24*, 5017–5032. [CrossRef] [PubMed]
15. Sikdar, A.; Chowdhury, A.S. An adaptive training-less framework for anomaly detection in crowd scenes. *Neurocomputing* **2020**, *415*, 317–331. [CrossRef]
16. Mehmood, A. Efficient Anomaly Detection in Crowd Videos Using Pre-Trained 2D Convolutional Neural Networks. *IEEE Access* **2021**, *9*, 138283–138295. [CrossRef]
17. Bansod, S.D.; Nandedkar, A.V. Anomalous Event Detection and Localization Using Stacked Autoencoder. In Proceedings of the International Conference on Computer Vision and Image Processing, Jaipur, India, 27–29 September 2019; Springer: New York, NY, USA, 2019; pp. 117–129.
18. Sabokrou, M.; Fayyaz, M.; Fathy, M.; Klette, R. Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes. *IEEE Trans. Image Process.* **2017**, *26*, 1992–2004. [CrossRef]
19. Chaker, R.; Aghbari, Z.A.; Junejo, I.N. Social network model for crowd anomaly detection and localization. *Pattern Recognit.* **2017**, *61*, 266–281. [CrossRef]
20. Zhou, S.; Shen, W.; Zeng, D.; Fang, M.; Wei, Y.; Zhang, Z. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Process. Image Commun.* **2016**, *47*, 358–368. [CrossRef]
21. Chen, C.Y.; Shao, Y. Crowd Escape Behavior Detection and Localization Based on Divergent Centers. *IEEE Sens. J.* **2015**, *15*, 2431–2439. [CrossRef]
22. Bansod, S.D.; Nandedkar, A.V. Crowd anomaly detection and localization using histogram of magnitude and momentum. *Vis. Comput.* **2020**, *36*, 609–620. [CrossRef]
23. Sikdar, A.; Chowdhury, A.S. Multi-level Threat Analysis in Anomalous Crowd Videos. In Proceedings of the International Conference on Computer Vision and Image Processing, Jaipur, India, 27–29 September 2019; Springer: New York, NY, USA, 2019; pp. 495–506.
24. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1995.
25. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
28. Mehran, R.; Oyama, A.; Shah, M. Abnormal crowd behavior detection using social force model. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 935–942.
29. Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly detection in crowded scenes. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1975–1981.

30. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.
31. Tudor Ionescu, R.; Smeureanu, S.; Alexe, B.; Popescu, M. Unmasking the abnormal events in video. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2895–2903.
32. Cong, Y.; Yuan, J.; Liu, J. Sparse reconstruction cost for abnormal event detection. In Proceedings of the CVPR 2011, Providence, RI, USA, 20–25 June 2011; pp. 3449–3456.
33. Alafif, T.; Alzahrani, B.; Cao, Y.; Alotaibi, R.; Barnawi, A.; Chen, M. Generative adversarial network based abnormal behavior detection in massive crowd videos: A Hajj case study. *J. Ambient. Intell. Humaniz. Comput.* **2022**, *13*, 4077–4088. [CrossRef]
34. Solmaz, B.; Moore, B.E.; Shah, M. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2064–2070. [CrossRef] [PubMed]
35. Wang, C.; Zhao, X.; Wu, Z.; Liu, Y. Motion pattern analysis in crowded scenes based on hybrid generative-discriminative feature maps. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–18 September 2013; pp. 2837–2841.
36. Alqaysi, H.H.; Sasi, S. Detection of abnormal behavior in dynamic crowded gatherings. In Proceedings of the 2013 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 23–25 October 2013; pp. 1–6.
37. Zou, Y.; Zhao, X.; Liu, Y. Detect coherent motions in crowd scenes based on tracklets association. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 4456–4460.
38. Pennisi, A.; Bloisi, D.D.; Iocchi, L. Online real-time crowd behavior detection in video sequences. *Comput. Vis. Image Underst.* **2016**, *144*, 166–176. [CrossRef]
39. Wu, S.; Yang, H.; Zheng, S.; Su, H.; Fan, Y.; Yang, M.H. Crowd behavior analysis via curl and divergence of motion trajectories. *Int. J. Comput. Vis.* **2017**, *123*, 499–519. [CrossRef]
40. Miao, Y.; Yang, J.; Alzahrani, B.; Lv, G.; Alafif, T.; Barnawi, A.; Chen, M. Abnormal Behavior Learning Based on Edge Computing toward a Crowd Monitoring System. *IEEE Netw.* **2022**, *36*, 90–96. [CrossRef]
41. Miao, Y.; Tang, Y.; Alzahrani, B.A.; Barnawi, A.; Alafif, T.; Hu, L. Airborne LiDAR assisted obstacle recognition and intrusion detection towards unmanned aerial vehicle: Architecture, modeling and evaluation. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 4531–4540. [CrossRef]
42. Luo, L.; Li, Y.; Yin, H.; Xie, S.; Hu, R.; Cai, W. Crowd-level Abnormal Behavior Detection via Multi-scale Motion Consistency Learning. *arXiv* **2022**, arXiv:2212.00501.
43. of Minnesota, U. Unusual Crowd Activity Dataset of University of Minnesota. 2020. Available online: http://mha.cs.umn.edu/movies/crowdactivity-all.avi (accessed on 25 April 2020).
44. Horn, B.K.; Schunck, B.G. Determining optical flow. In *Artificial Intelligence*; Elsevier: Amsterdam, The Netherlands, 1981; Volume 17, pp. 185–203.
45. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
46. Bishop, G.; Welch, G. An introduction to the kalman filter. *Proc. Siggraph Course* **2001**, *8*, 41.
47. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
48. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
49. Bottou, L. Stochastic gradient learning in neural networks. *Proc. Neuro-Nimes* **1991**, *91*, 12.
50. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
51. Kim, J.; Grauman, K. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2921–2928.
52. Luo, W.; Liu, W.; Gao, S. A revisit of sparse coding based anomaly detection in stacked rnn framework. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 341–349.
53. Patel, N.; Mukherjee, S.; Ying, L. Erel-net: A remedy for industrial bottle defect detection. In Proceedings of the International Conference on Smart Multimedia, Toulon, France, 24–26 August 2018; Springer: New York, NY, USA, 2018; pp. 448–456.