H Y B R ID C L A S S I F I E R S F O R S P A T I O - T E M P O R A L R E A L - T I M E A B N O M A L B E H A V I O R S D E T E C T I O N , T R A C K I N G , A N D R E C O G N I T I O N I N M A S S I V E H A J J C R O W D S

A P R EPRINT

T a r ik A l a f i f
Department of Computer Science
Jamoum University College
Umm Al-Qura University
Jamoum, Makkah, Saudi Arabia 25375

A n a s H a d i
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

M a n a l A l l a h y a n i
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

B a n d e r A l z a h r a n i
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

A r e e j A l h o t h a l i
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

R e e m A l o t a i b i
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

A h m e d B a r n a w i
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

A B S T R A C T

Individual abnormal behaviors vary depending on crowd sizes, contexts, and scenes. Challenges such as partial occlusions, blurring, large-number abnormal behavior, and camera viewing occur in large-scale crowds when detecting, tracking, and recognizing individuals with abnormal behaviors. In this paper, our contribution is twofold. First, we introduce an annotated and labeled large-scale crowd abnormal behaviors 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-scales 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 individuals detection method based on the magnitudes and orientations of Horn-Schunck optical flow is used 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, variances, and standard deviations statistical features are computed and fed to the RF to classify individuals with abnormal behaviors in the temporal domain. In large-scale crowds, we fine-tune the ResNet-50 model using YOLOv2 object detection technique to detect individuals with abnormal behaviors in the spatial domain. The proposed method achieves 99.77% 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 1.67%, 6.06%, and 2.85% on UMN, UCSD Ped1, and UCSD Ped2, respectively. It also achieves a satisfactory result in the large-scale crowds.

K e y w o r d s Abnormal behaviors · small-scale crowd · large-scale crowd · convolutional neural network · random forest · detection · tracking · recognition
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 according to crowd contexts and scenes. Therefore, the definition of abnormal behaviors might vary from one place or scenario to another. Similarly, the density and the number of individuals in the crowd often vary significantly, which can be small or large-scale 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 might raise many challenges as a result of many individuals moving or gathering in one location at almost the same time. In large-scale crowds, challenges such as partial and full occlusions, blurring, large-number 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 are 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 wide range of trajectory features (Zhang et al., 2016; Bera et al., 2016a; Zhou et al., 2015; Zhao et al., 2018; Coşar et al., 2016; Piccirelli et al., 2008), dense motion features (Colque et al., 2016; Cho and Kang, 2014; Qasim and Bhatti, 2019; Zhang et al., 2014), spatial-temporal features (Guo et al., 2016; Yuan et al., 2015; Fradi et al., 2016), or deep learning based features and optimization techniques for anomaly recognition (Chan et al., 2015; Sikdar and Chowdhury, 2020; Mehmood, 2021; Bansod and Nandedkar, 2019; Sabokrou et al., 2017). Most of the developed methods perform a binary frame-level for anomaly detection. Several studies considered locating anomalies in crowd surveillance videos (Bansod and Nandedkar, 2019; Chaker et al., 2017; Zhou et al., 2016; Chen and Shaol, 2015; Sabokrou et al., 2017; Bansod and Nandedkar, 2020), and less attention was paid to multi-class anomalies (Sikdar and Chowdhury, 2019). This study proposes a hybrid model that first identifies anomaly at the frame-level and then locates and classifies crowd anomaly 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 research attention. The proposed methods in the field are also often evaluated on datasets of 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 and HAJJ dataset.

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 the 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 for large-scale crowds in surveillance cameras to ensure pilgrims’ safety and smooth flow during Hajj. It also helps security authorities and decision-makers to 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 (LeCun et al., 1995, 1998; He et al., 2016). 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 (Liaw et al., 2002). The contributions of this work are summarized as follows:

- We introduce a manually annotated and labeled large-scale crowd abnormal behaviors dataset in Hajj, HAJJv2.
- We propose two methods of hybrid CNNs and RFs 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 abnormal behaviors public 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 behaviors 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-scale and large-scale crowd videos in Section 4. Experimental implementation, results, and evaluation are provided in Section 5. Then, a discussion on experimental evaluations, limitations, and challenges is provided in Section 6. Finally, we conclude our work 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 behaviors 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 (Mehran et al., 2009; Mahadevan et al., 2010; Zhang et al., 2014; Hasan et al., 2016; Tudor Ionescu et al., 2017; Cong et al., 2011; Alafif et al., 2021).

  (Picciarelli et al., 2008) introduced a normal model by clustering the trajectories of moving objects for anomaly detection. Then, (Mehran et al., 2009) proposed to use 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., 2009) work, (Mahadevan et al., 2010) 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., 2011) applied a sparse reconstruction cost and a dictionary to measure normal and abnormal behaviors.

  After that, (Zhang et al., 2014) 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., 2016) 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 flow were used to extract the Spatio-temporal features from raw video frames to feed the convolutional autoencoder for classification. (Fradi et al., 2016) applied local feature tracking to describe the movement 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., 2016) used the histograms of magnitude, orientation, and entropy of the optical flow with the nearest neighbor search algorithm to detect the anomalous. 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 find the abnormality.

- **Large-scale crowds:** Several research works studied abnormal behaviors on large-scale crowds in (Solmaz et al., 2012; Wang et al., 2013; Alqaysi and Sasi, 2013; Bera et al., 2016b; Zou et al., 2015; Bera et al., 2016b; Pennisi et al., 2016; Fradi et al., 2016; Wu et al., 2017; Alafif et al., 2021).

  First, (Solmaz et al., 2012) introduced a linear approximation using a Jacobian matrix to identify large-scale crowd abnormal behaviors. An optical flow and a particle advection were used. Then, (Wang et al., 2013) started to cluster crowd feature maps to analyze motion patterns. Followed by (Wang et al., 2013), (Alqaysi and Sasi, 2013) 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., 2015) detected large-scale crowd motions and trajectories using tracklets association. Similar to (Zou et al., 2015), (Bera et al., 2016b) computed abnormal behaviors trajectories using Bayesian learning techniques. Then, (Pennisi et al., 2016) segmented extracted features to detect crowd abnormal behaviors. In recent years, (Fradi et al., 2016) and (Wu et al., 2017) worked on analyzing large-scale crowd properties using visual feature descriptors.
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 except the work presented in (Alafif et al., 2021). In comparison with the recent work in (Alafif et al., 2021), the proposed methods don’t require generating individual abnormal behavior images. Compared to the work in (Alafif et al., 2021), the proposed method achieves better accuracy using the HAJJv1 dataset.

3 HAJJv2 Dataset

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 (Alafif et al., 2021). The HAJJv2 dataset consists of 18 manually collected videos from the annual Hajj religious event. All the videos are stored in 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 “Mas- saa” 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 9 short videos. Each video in the training set lasts for 25 seconds, while each video in the testing set lasts for 20 seconds.

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 to large-scale crowd flows. Figure 1 shows examples of these abnormal behaviors included in HAJJv2 dataset. The dataset statistics are provided in Table 2. 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 training and testing.

Individuals’ anomalous behaviors in the videos are manually annotated and labeled for training and testing. 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 behaviors datasets and the Abnormal Behaviors HAJJv2 dataset is shown in Table 1.

4 Proposed Methods

In this section, we present the details of our proposed methods and algorithms. First, we show the individual abnormal behaviors detection and recognition methodology pipeline and algorithm for small-scale crowds. Then, similarly, a methodology pipeline and an algorithm for detecting and recognizing abnormal behaviors are presented for large-scale crowds. Figure 2 shows the methodology pipelines for large-scale and small-scale crowd abnormal behavior detection and recognition.
Table 1: A comparison of existing public abnormal behaviors datasets and the Abnormal Behaviors HAJJv2 dataset.

| Dataset | Abnormal Behaviors                                                                 | Size     | Crowd Scale | Reference                  |
|---------|------------------------------------------------------------------------------------|----------|-------------|-----------------------------|
| UMN     | Escape                                                                            | 24,240 KB | Small-scale | (of Minnesota, 2020)        |
| UCSD    | Non-pedestrian movements                                                          | 1.74 GB  | Small-scale | (Mahadevan et al., 2010)    |
| HAJJv1  | Standing, sitting, sleeping, running, moving in opposite crowd direction, crossing or moving in different crowd direction, and non-pedestrian movements | 831 MB   | Large-scale | (Alaffi et al., 2021)       |
| HAJJv2  | Standing, sitting, sleeping, running, moving in opposite crowd direction, crossing or moving in different crowd direction, and non-pedestrian movements | 831 MB   | Large-scale | -                           |

Table 2: HAJJv2 dataset statistics.

| n | Classes                  | Training | Testing |
|---|--------------------------|----------|---------|
| 1 | Different Crowd Direction| 7,152    | 6,262   |
| 2 | Moving In Opposite       | 36,577   | 18,802  |
| 3 | Moving Non Human Object  | 4,186    | 4,146   |
| 4 | Running                  | 51       | 190     |
| 5 | Sitting                  | 100,633  | 83,644  |
| 6 | Sleeping                 | 2,400    | 2,618   |
| 7 | Standing                 | 19,773   | 14,107  |
| Total |                      | 170,772  | 129,769 |

4.1 Individual Abnormal Behaviors Detection, Tracking, and Recognition in Small-scale Crowds

Figure 2a shows the methodology pipeline for detecting and recognizing abnormal behaviors in small-scale crowded 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 an 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 lot 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 huge 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 [He et al. 2016], to detect anomalies 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. The ResNet uses a deep residual learning framework to solve the degradation problem. This gives the advantage of using a deep network to extract the complex feature patterns in the spatial domain. Therefore, we use the ResNet-50 in our experiment.

The ResNet-50 consists of 49 convolution layers as a features extractor, followed by an average pooling and a fully connected layer as a classifier. Fine-tuning the pre-trained models is made 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. Anomalies detection is a binary classification problem. Thus, the classifier is trained on normal and abnormal frames. Therefore, we fine-tune the 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 2,048 units into 128, followed by an output layer that maps 128 units into 2 units representing the normal and the abnormal probabilities. Since the ResNet-50 processes inputs with a size 224×224×3, we resize the frames to ResNet-50’s input size. A feed-forward and backpropagation algorithm is applied by updating the errors and weights to converge.

Figure 2a shows detected normal and abnormal frames resulted from the ResNet-50 classifier. The detected normal frames appear in green while the detected abnormal ones appear in red.

\[
\begin{align*}
\mu(m^j) &= \frac{\sum_{i=1}^{p} (m_i^j)}{p} \\
\mu(r^j) &= \frac{\sum_{i=1}^{p} (r_i^j)}{p} \\
\sigma^2(m^j) &= \frac{\sum_{i=1}^{p} ((m_i^j - \mu(m^j))^2)}{p} \\
\sigma^2(r^j) &= \frac{\sum_{i=1}^{p} ((r_i^j - \mu(r^j))^2)}{p} \\
\sigma(m^j) &= \sqrt{\frac{\sum_{i=1}^{p} ((m_i^j - \mu(m^j))^2)}{p}} \\
\sigma(r^j) &= \sqrt{\frac{\sum_{i=1}^{p} ((r_i^j - \mu(r^j))^2)}{p}}
\end{align*}
\]
**Temporal domain:** We use the optical flow to detect the anomalies on a pixel level. By analyzing the optical flow, we can observe crowd movements, instantaneous velocity, orientations, and magnitudes. These low-level features are used to recognize individuals’ behaviors. After detecting the anomalies at $i$-th frame ($F_i$), the optical flow of this frame ($O_i$) is computed using Horn-Schunck optical flows (Horn and Schunck, 1981). Then, magnitudes ($m_i$) and orientations ($r_i$) features are automatically extracted from the optical flows. In small-scale crowded videos, binary magnitude-based masks using a threshold ($T$) are initiated to localize and track individuals within the frame (i.e., $\{F_i\}_{j=1}^{z}$ is the $j$-th individual in the $i$-th frame). Figure 2(a) shows the binary magnitude-based mask of the small-scale datasets. After extracting the magnitudes and orientations features, the statistical features including means ($\mu$), variances ($\sigma^2$), and standard deviations ($\sigma$) are computed. They are computed for the total pixels ($\mu$) of the area that represents an individual ($j$), for both the magnitudes ($m_i$) and orientations ($r_i$) as follows:

Then, these statistical features are fed to the RF classifier for training to classify and recognize temporal individual 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.

### 4.2 Individual Abnormal Behaviors Detection and Recognition in Large-scale Crowds

Figure 2(b) shows the methodology pipeline for detecting and recognizing abnormal behaviors in large-scale crowd

---

(a) Small-scale crowds (UMN and UCSD datasets).

(b) Large-scale crowds (HAJJv2 dataset).
After detecting individuals’ abnormal behaviors and extracting their statistical features, we compute the means ($\mu$), variances ($\sigma^2$), and standard deviations ($\sigma$) for each individual similar to the small-scale crowd method. Then, an RF classifier is used as a multi-class classifier to classify and recognize all individuals with abnormal behaviors.

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

Algorithm 1: A hybrid CNN and RF algorithm for Spatio-temporal small-scale crowd abnormal behaviors 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.

1. 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.
2. Compute optical flow $\{O_i\}_{i=1}^{n-1}$ from the original video sequence $F_i: F_1 \rightarrow \{O_1, O_2, \ldots, O_{n-1}\}$.
3. Extract optical flows orientations $\{r_i\}_{i=1}^n$ and magnitudes $\{m_i\}_{i=1}^n$ features from $O_i$.
4. Create the binary magnitude-based mask using a threshold $T$, $mask = m_i > T$.
5. Extract the individuals within the mask $\{I_j^o\}_{j=1}^n$.
6. Compute orientations and magnitudes means $\{\mu_j\}_{j=1}^2$, variances $\{\sigma_j^2\}_{j=1}^2$, and standard deviations $\{\sigma_j\}_{j=1}^2$.
7. Use the statistical features $\{\mu_j\}_{j=1}^2$, $\{\sigma_j^2\}_{j=1}^2$, and $\{\sigma_j\}_{j=1}^2$ of the individuals to train temporal normal and abnormal behaviors using an RF model;
8. Insert a test video frame sequences $\{F_i\}_{i=1}^n$;
9. Compute optical flows $O_i$;
10. If $F_i$ is abnormal then
    1. Use $F_i^o$ to test the fine-tuned CNN model;
    2. Compute optical flows orientations $r_i^o$ and magnitudes $m_i^o$ features from $O_i^o$;
    3. Create binary magnitude-based mask to localize and track individuals $\{I_j^o\}_{j=1}^n$;
    4. Test the statistical features: means $\mu_j^{o,i}$, variances $\sigma_j^{o,i}$, and standard deviations $\sigma_j^{o,i}$ features to classify $I_j^{o,i}$ using the trained RF model;
11. End

5 Experiments

In this section, we first provide details of the implementation of the proposed method. Second, we briefly describe
5.1 Implementation

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

5.2 Datasets

In this section, we use public datasets such as UMN (of Minnesota [2020]), UCSD (Mahadevan et al., 2010), HAJJv1 (Alafif et al. [2021]), 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 used datasets are summarized 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 outdoor while one was recorded indoor. Each UMN scene starts with a normal activity followed by abnormal behavior. Walking, for example, is considered normal activity, while running is an abnormal one. The frame resolution in UMN scenes is 320 x 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 [ illustrates an example of UMN’s frames. The train and test splits are not explicitly specified. Moreover, the annotations are only available at the frame level. Because of 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 x 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 x 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 used both normal and abnormal videos for training and testing. Figure [ illustrates some examples from Ped1 and Ped2 frames.
5.3 Experimental Settings and Hyperparameters

For both small-scale and large-scale crowds experiments, different configurations are evaluated to determine the most effective approach. Details 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 model as a 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 Adam optimizer (Kingma and Ba, 2014) 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 Lucas-Kanade derivative of Gaussian, Lucas-Kanade, Farneback,
and Horn-Schunck, are also employed. The Horn-Schunck (Horn and Schunck [1981]) method is selected since it provides magnitude and orientation features to create a binary magnitude-based mask localize and track individuals. The means, variances, and standard deviations are computed using these features to classify the individuals’ abnormal behaviors.

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

**Large-scale crowds:** The ResNet-50 and SqueezeNet pre-trained CNN models are used as a base network of the YOLOv2 object detection technique. We initialize the weights on ImageNet (Krizhevsky et al., 2012). Then, we fine-tune the model with stochastic gradient descent (SGD) (Bottou et al., 1991) optimizer for 20 epochs with the learning rate of 0.001 and a minibatch of 8 frames.

Similar to the small-scale crowd experiment, the RF classifier is also accountable for the abnormal behavior recognition in the temporal domain. Unlike in the small-scale crowd experiment, we train the RF classifier using only 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, F1 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:

\[
\begin{align*}
\text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\
\text{Precision} &= \frac{TP}{TP + FP} \\
\text{Recall} &= \frac{TP}{TP + FN} \\
F1\text{-score} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\end{align*}
\]

Receiver Operating Characteristics (ROC) curve (Powers 2020) is a plot of 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 99.77% and 93.71% of average AUCs among the scenes on UMN and UCSD datasets, respectively. Table 4 shows the quantitative results in the temporal domain using the RF classifier on small-scale crowd datasets. Figures 8 and 9 illustrate the ROC curves of our experiments using the ResNet-50 and the RF classifiers on UMN and UCSD datasets.

To better illustrate the comparison with existing methods in (Mehran et al., 2009), (Cong et al., 2011) and, (Alafif et al., 2021). Table 4 shows that the proposed method yields better results using the UMN dataset.

**Large-scale crowds:** We evaluate our method in the spatial and temporal domains on large-scale crowds using HAJJv1 and HAJJv2.

### Table 3: The evaluation results on UCSD dataset. The percentages are AUCs.

| Method                        | UCSD Ped1 | UCSD Ped2 |
|-------------------------------|-----------|-----------|
| MPPCA (Kim and Grauman 2009) | 59.0%     | 69.3%     |
| Social Force (Mehran et al. 2009) | 67.5%     | 55.6%     |
| SF+MPPCA (Mahadevan et al. 2010) | 68.8%     | 61.3%     |
| MDT (Mahadevan et al. 2010)   | 81.8%     | 82.9%     |
| Conv-RC (Hasan et al. 2010)   | 75.0%     | 85.0%     |
| Stacked RNN (Luo et al. 2012) | N/A       | 92.2%     |
| Unmasking (Tudor Ionescu et al. 2017) | 68.4%     | 82.2%     |
| Alafif et al. (Alafif et al. 2021) | 82.81%    | 95.7%     |
| **Ours**                      | **99.8**  | **99.8**  |

### Table 4: The evaluation results on UMN dataset. Percentages are AUCs.

| Method                        | UMN (%)   |
|-------------------------------|-----------|
| Optical-flow (Mehran et al. 2009) | 84.0      |
| SFM (Mehran et al. 2009)      | 96.0      |
| Sparse Reconstruction (Cong et al. 2011) | 97.0      |
| Alafif et al. (Alafif et al. 2021) | 98.1      |
| **Ours**                      | **99.8**  |

### Table 5: Our results of the RF classifier on small-scale crowd datasets.

| Dataset | Accuracy (%) | Precision (%) | Recall (%) | F1 (%) | AUC (%) |
|---------|--------------|---------------|------------|--------|---------|
| UMN scene 1 | 89.69        | 95.13         | 86.86      | 93.52  | 97.11   |
| UMN scene 2 | 81.51        | 99.04         | 76.81      | 86.51  | 94.66   |
| UCSD Ped1   | 99.5         | 99.6          | 99.89      | 99.75  | 97.36   |
| UCSD Ped2   | 99.6         | 99.82         | 99.78      | 99.79  | 97.99   |

Table 2 reports a performance comparison of the proposed method with the existing methods (Kim and Grauman 2009, Mehran et al. 2009, Mahadevan et al. 2010, Hasan et al. 2016, Tudor Ionescu et al. 2017, Alafif et al. 2021) using the UMN dataset. It is clearly shown that the proposed method achieves higher AUCs using USCSD Ped1 and Ped2 scenes.
Table 6: The experimental results of track assignment and IOU detections using YOLOv2 on HAJJv2 dataset.

| Video No. | Track Assignment   | IOU          |
|-----------|--------------------|--------------|
|           | 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       | 7.40    | 89.55       | 4.65        | 0.14       | 0.27   |
| 9 and 11 (Jamarat) | 89.26 | 96.44       | 3.85       | 7.40    | 89.53       | 51.66       | 33.65      | 40.75  |
| 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  |

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

| Video No. | Accuracy (%) | Precision (%) | Recall (%) | F1 (%) | AUC (%) |
|-----------|--------------|---------------|------------|--------|---------|
| 10 (Arafat) | 63.65     | 64.81         | 58.30      | 61.38  | 91.18   |
| 12 (Tawaf)  | 62.91     | 64.59         | 52.59      | 57.80  | 75.11   |
| 9 and 11 (Jamarat) | 96.86 | 33.92        | 33.03      | 33.47  | 57.92   |
| 2, 3, 5, 7, and 8 (Masaa) | 56.44 | 37.97        | 32.81      | 35.20  | 80.12   |
| Average    | 75.42     | 50.32         | 44.15      | 46.96  | 76.08   |

Table 8: A comparison table of abnormal behaviors detection performance with the recent existing methods using HAJJv1 dataset against existing methods.

| Model                  | Accuracy (%) | Precision (%) | Recall (%) | F1 (%) | AUC (%) |
|------------------------|--------------|---------------|------------|--------|---------|
| Alafif et al. (2021)   | 65.10        | 61.48         | N/A        | N/A    | 79.63   |
| YOLOv2 (Ours)          | 95.67        | 9.421         | 28.82      | 10.99  | N/A     |
| RF (Ours)              | 34.48        | 9.910         | 11.22      | 10.39  | 66.56   |

Figure 8: The ROC curves for the ResNet50 classifier on small-scale datasets.

Table 9: Our results using the ResNet-50 classifier on public and benchmark small-scale crowd datasets.

| Dataset   | Accuracy (%) | Precision (%) | Recall (%) | F1 (%) | AUC (%) |
|-----------|--------------|---------------|------------|--------|---------|
| UMN scene 1 | 97.93        | 98.14         | 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   |

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 an 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 illumi-
nation. In the large-scale crowds, we still haven’t achieved an excellent performance, yet using the HAJJv2 dataset since the videos in the dataset are very challenging. The challenges are represented by far camera viewing and heavy partial and full occlusions with a huge number of individuals. Figures 7b and 7c show some of the challenges in Tawaf and Jamarat scenes which are considered the hardest scenes for the classifiers to classify the individuals with abnormal behaviors. During annotating and labeling the abnormal behaviors using these scenes, much more human attention and focus were required since they contain a massive number of individuals moving in one spot with heavy partial occlusions and far camera views. On the other hand, the easiest scenes for the annotators, labelers, and classifiers to classify are in the Masaa scenes in which most of the individuals are classified correctly 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 individuals with abnormal behaviors in large-scale and massive crowds.

7 Conclusion

In this research work, we first introduce the annotated and labeled large-scale crowd abnormal behaviors Hajj dataset (HAJJv2). Second, we propose two methods of hybrid CNNs and RFs to detect and recognize Spatio-temporal abnormal behaviors in small-scale and large-scale crowd videos. In small-scale crowd videos, a ResNet-50 pre-trained CNN model is fine-tuned to check every frame, whether it is normal or abnormal in the spatial domain. If abnormal behavioral behaviors are found, a motion-based individual detection using magnitude and orientation features of Horn-Schunck optical flow is employed to create a binary magnitude-based mask to localize and track individuals with abnormal behaviors. In large-scale crowd videos, a Kalman filter is employed to predict and track the detected individuals in the next frames. Then, means, variances, and standard deviations statistical features are computed and fed to the RF to classify individuals with abnormal behaviors in the temporal domain. In the large-scale crowd videos, we fine-tune the ResNet-50 model using the YOLOv2 object detection technique to detect individuals with abnormal behaviors in the spatial domain. Our method using public benchmark small-scale crowd datasets achieves 99.77% and 93.71% of AUCs respectively on the UMN and UCSD datasets, while the method in large-scale crowd achieves 76.08% of average AUC using the HAJJv2 dataset. Our methods outperform 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 achieves a satisfactory result in large-scale crowds. Still, lots of work is needed to increase the effectiveness of abnormal behavior detection and recognition in large-scale crowded scenes due to their challenges. Most of the current research works only uses small-scale crowded scenes in which the abnormal behaviors can be extracted and classified easily. In the future, our work will be focused on large-scale crowds. We will incorporate an attention mechanism and fusion strategies to enhance the performance. This work can potentially help researchers to study and apply it in different contexts of crowd scenes such as in airports, stadiums, and marathons. It can be also extended in the manufacturing industries [Patel et al., 2018] by inspecting and detecting unusual behaviors of defective manufactured goods and products on a production line. Examples and features of the products’ unusual behaviors are required to be collected, extracted, and learned by a classifier to achieve high performance.

8 Acknowledgement

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).

References

Alafif, T., Alzahrani, B., Cao, Y., Alobaib, R., Barnawi, A., Chen, M., 2021. Generative adversarial network based abnormal behavior detection in massive crowd videos: a hajj case study. Journal of Ambient Intelligence and Humanized Computing, 1–12.

Alqaysi, H.H., Sasi, S., 2013. Detection of abnormal behavior in dynamic crowded gatherings, in: 2013 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, pp. 1–6.

Bansod, S.D., Nandedkar, A.V., 2019. Anomalous event detection and localization using stacked autoencoder, in: Interna-
Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, Guo, H., Wu, X., Cai, S., Li, N., Cheng, J., Chen, Y.L., 2016.

Bera, A., Kim, S., Manocha, D., 2016a. Realtime anomaly detection using trajectory-level crowd behavior learning, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1289–1296. doi:10.1109/CVPRW.2016.163

Bera, A., Kim, S., Manocha, D., 2016b. Realtime anomaly detection using trajectory-level crowd behavior learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 50–57.

Bottou, L., et al., 1991. Stochastic gradient learning in neural networks. Proceedings of Neuro-Numes 91, 12.

Chaker, R., Aghbari, Z.A., Junеjo, I.N., 2017. Social network model for crowd anomaly detection and localization. Pattern Recognition 61, 266–281. URL: https://www.sciencedirect.com/science/article/pii/S0031320316301327. doi:https://doi.org/10.1016/j.patcog.2016.06.016

Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y., 2015. Pcanet: A simple deep learning baseline for image classification? IEEE Transactions on Image Processing 24, 5017–5032. doi:10.1109/TIP.2015.2475629

Chen, C.Y., Shao, Y., 2015. Crowd escape behavior detection and localization based on divergent centers. IEEE Sensors Journal 15, 2431–2439. doi:10.1109/JSEN.2014.2381260

Cho, S.H., Kang, H.B., 2014. Abnormal behavior detection using hybrid agents in crowded scenes. Pattern Recognition Letters 44, 64–70. URL: https://www.sciencedirect.com/science/article/pii/S0167865513004613. doi:https://doi.org/10.1016/j.patrec.2013.11.017 pattern Recognition and Crowd Analysis.

Colque, R.V.H.M., Caetano, d., de Andrade, M.T.L., Schwartz, W.R., 2016. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. IEEE Transactions on Circuits and Systems for Video Technology 27, 673–682.

Cong, Y., Yuan, J., Liu, J., 2011. Sparse reconstruction cost for abnormal event detection, in: CVPR 2011, IEEE. pp. 3449–3456.

Coser, S., Donatiello, G., Bogorny, V., Garate, C., Alvarado, L.O., Brémont, F., 2016. Toward abnormal trajectory and event detection in video surveillance. IEEE Transactions on Circuits and Systems for Video Technology 27, 683–695.

Fradi, H., Luvison, B., Pham, Q.C., 2016. Crowd behavior analysis using local mid-level visual descriptors. IEEE Transactions on Circuits and Systems for Video Technology 27, 589–602.

Guo, H., Wu, X., Cai, S., Li, N., Cheng, J., Chen, Y.L., 2016. Quaternion discrete cosine transformation signature analysis in crowd scenes for abnormal event detection. Neurocomput. 204, 106–115. URL: https://doi.org/10.1016/j.neucom.2015.07.153. doi:10.1016/j.neucom.2015.07.153

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S., 2016. Learning temporal regularity in video sequences, in: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 733–742.
Redmon, J., Farhadi, A., 2017. Yolo9000: better, faster, stronger, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263–7271.

Sabokrou, M., Fayyaz, M., Fathy, M., Klette, R., 2017. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing 26, 1992–2004. doi:10.1109/TIP.2017.2670780

Sikdar, A., Chowdhury, A.S., 2019. Multi-level threat analysis in anomalous crowd videos, in: International Conference on Computer Vision and Image Processing, Springer. pp. 495–506.

Sikdar, A., Chowdhury, A.S., 2020. An adaptive training-less framework for anomaly detection in crowd scenes. Neurocomputing 415, 317–331. URL: https://www.sciencedirect.com/science/article/pii/S0925231220311668, doi:https://doi.org/10.1016/j.neucom.2020.07.068

Solmaz, B., Moore, B.E., Shah, M., 2012. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 2064–2070.

Tudor Ionescu, R., Smeureanu, S., Alexe, B., Popescu, M., 2017. Unmasking the abnormal events in video, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2895–2903.

Wang, C., Zhao, X., Wu, Z., Liu, Y., 2013. Motion pattern analysis in crowded scenes based on hybrid generative-discriminative feature maps, in: 2013 IEEE International Conference on Image Processing, IEEE. pp. 2837–2841.

Welch, G., Bishop, G., et al., 1995. An introduction to the kalman filter.

Wu, S., Yang, H., Zheng, S., Su, H., Fan, Y., Yang, M.H., 2017. Crowd behavior analysis via curl and divergence of motion trajectories. International Journal of Computer Vision 123, 499–519.

Yuan, Y., Fang, J., Wang, Q., 2015. Online anomaly detection in crowd scenes via structure analysis. IEEE Transactions on Cybernetics 45, 548–561. doi:10.1109/TCYB.2014.2330853

Zhang, Y., Qin, L., Ji, R., Yao, H., Huang, Q., 2014. Social attribute-aware force model: exploiting richness of interaction for abnormal crowd detection. IEEE Transactions on Circuits and Systems for Video Technology 25, 1231–1245.

Zhang, Y., Qin, L., Ji, R., Zhao, S., Huang, Q., Luo, J., 2016. Exploring coherent motion patterns via structured trajectory learning for crowd mood modeling. IEEE Transactions on Circuits and Systems for Video Technology 27, 635–648.

Zhao, K., Liu, B., Li, W., Yu, N., Liu, Z., 2018. Anomaly detection and localization: A novel two-phase framework based on trajectory-level characteristics, in: 2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW), pp. 1–6. doi:10.1109/ICMEW.2018.8551517

Zhou, S., Shen, W., Zeng, D., Zhang, Z., 2015. Unusual event detection in crowded scenes by trajectory analysis, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1300–1304. doi:10.1109/ICASSP.2015.7178180

Zou, Y., Zhao, X., Liu, Y., 2015. Detect coherent motions in crowd scenes based on tracklets association, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 4456–4460.