Computers, Materials & Continua
DOI:10.32604/cmc.2022.022147
Article

Tech Science Press

Identification of Anomalous Behavioral Patterns in Crowd Scenes

Muhammad Asif Nauman* **and Muhammad Shoaib**

[1]Department of Computer Engineering, University of Engineering and Technology, Lahore, Pakistan
*Corresponding Author: Muhammad Asif Nauman. Email: asif.nauman.uet@gmail.com

Received: 28 July 2021; Accepted: 03 September 2021

Abstract: Real time crowd anomaly detection and analyses has become an active and challenging area of research in computer vision since the last decade. The emerging need of crowd management and crowd monitoring for public safety has widen the countless paths of deep learning methodologies and architectures. Although, researchers have developed many sophisticated algorithms but still it is a challenging and tedious task to manage and monitor crowd in real time. The proposed research work focuses on detection of local and global anomaly detection of crowd. Fusion of spatial-temporal features assist in differentiation of feature trained using Mask R-CNN with Resnet101 as a backbone architecture for feature extraction. The data from, BIWI Walking Pedestrian dataset and the Crowds-By-Examples (CBE) dataset and Self-Generated dataset has been used for experimentation. The data deals with different situations like one set of data deals with normal situations like people walking and acting individually, in a group or in a dense crowd. The other set of data contains images four unique anomalies like fight, accident, explosion and people behaving normally. The simulated results show that in terms of precision and recall, our system performs well with Self-Generated dataset. Moreover, our system uses an early stopping mechanism, which allows our system to outperform to make our model efficient. That is why, on 89th epoch our system starts generating finest results.

Keywords: Mask R-CNN; crowd management and monitoring; precision and recall

1 Introduction

Anomaly differentiating proof has been one of the fascinating topics that have emerged in recent years. Have provided us with a unique model of a typical line of action Likewise, incredible conduct, particularly in terms of places. Without a doubt, indirect communication in layman's terms, we may argue that any divergence from the norm is abnormal. An abnormality in the system is referred to as rehearsals. Irregularities have existed for a long time. Inconsistencies, offensive observations, surprising insights, rare cases, deviations, shocks, and other terms have been mentioned in various places. In several space applications, quality or contaminations are important. Anomaly distinguishing proof, for example, has a degree in a variety of related subjects. Acceptance of compulsion for Visas, insurance, or clinical benefits, intrusion recognized proof for network security, and flaw area in prosperity are all important adversary buildings and military observation.

In therapeutic settings, since produced proficiently, imaging, irregularity identifiable proof accepts a vital role. Those in charge of locating complex infections are specialists. Requires aid in recognizing the inconsistencies in the typical leader of the difficulties to produce a noticeable result. An unusual MRI image, for example, might reveal the existence of dangerous cancers, inconsistencies in charge card information sharing might indicate charge card or wholesale fraud, as well as a suspect in the case. Test data might be labeled as an exception. Any alterations explicit subjects from normal and regulated conduct of space. It's an odd order. We require a callous approach to dealing with the situation.

Describe a location that addresses common etiquette and make any necessary announcements. Any information that deviates from this defined district is considered irregular. The measure of deviations has been categorized in numerous ways areas. What may be unusual in clinical areas may be normal in a profession such as financial services. Similarly, every space will have a different inconsistency assessing scale that varies by area.

Why is the anomalous area in group assessment so large?

Putting together the pieces In PC vision applications, confirmation has become a crucial field. Examining the data ends up being more time-consuming enticing because of the hidden uses in numerous aspects of daily life while guaranteeing the safety of people in crowded locations such as train stations, farms, and dangerous tourist attractions. The group observation is being robotized using PC vision and AI calculations at the same time. a framework capable of gradually delivering cutting-edge execution the fascinating thing about humans is that their behavior has evolved through time. Been considered, and their potential threat has been decoded in the past.

When the individual is involved, the problem becomes more complicated. Humans enter large groups to form a group. Groups have many perspectives, such as swarm thickness evaluation, swarm Identification of movements, group tracking, swarming techniques, and Checking with the swarm. All the preceding phases connect the collection process to the regular steps of evaluating photos and accounts for legitimate bits of data assessment. Obtaining a conduct evaluation might be a significant step in the right direction. the implementation of a smart transportation system to cook for wealth and security, as well as a never-ending supply/demand situation for the executives in charge of public transportation.

The primary goal of collecting inconsistency ID is to identify inconsistencies. a gathering in an unusual direction, like a car on the highway pedestrian walkway, an incredible depiction of people rushing abruptly because of unforeseen, non-routine congestion areas of interest, a person on a motor vehicle route (jaywalking, etc.) within sight of the truck on the individual by walking in red, the enclosing addresses an inconsistency in the setting environment.

Regardless, recognizing oddities in a crowded setting is an exceedingly difficult undertaking because: (a) the imbalanced lighting situation at the gathered information scenes, (b) the availability of odd event testing is spectacular and frequently unpretentious, and (c) the show's rapid progression weakens it. (d) Displaying a variety of ordinary and uncommon occurrences is an uncomfortable task, and (e) the relevance of conventional and exceptional events after adjusting visual settings, unusual events are dark and considerably reliant. Due to their unimaginable affirmation performance in many PC vision applications, Convolutional Neural Networks (CNNs) are the most sought-after important learning approach.

Contents are a variety of neural connections that may be used to extract many isolating characteristics at different levels of abstraction. Setting up a Convent without any preparation is a computationally demanding operation that necessitates a large pool of checked ready-to-use datasets. This problem is most noticeable in swarm eccentricity ID when the odd event testing is not transparent. The focus of this paper is on Makers who have attempted to resolve this issue by eliminating the need for putting up the Convnets without any preparation and looking into alternative routes for moving CNNs that have been pre-arranged are being learned. We've suggested an ingenious idea.

The present capability of pre-arranged Convnets and a pool of classifiers is impacted by a combination of ensembles (AOE). Aggregation of Ensembles is a word that combines two terms. "Aggregate" and "Outfit" are examples of words that indicate irregularity in the past. The depiction plot and subsequent extraction of high-quality parts imply excellent part extraction.

The suggested method takes advantage of a mash-up of several changes. Models of convolutional neural networks are based on the idea that various CNN architectures acquire different degrees of semantic picture representation. Our social gathering's unique CNNs allow us to isolate certain gathering aspects for characterization. of the varying particular and unnoticed differences between average and strange things that happen in crowded settings. The suggested architecture assists Convnets in replacing traditional components purchased from suppliers.

Typical images that fill space with unmistakable components. The novel the following is a list of this paper's contributions:

- We offer a clever idea for aggregating social events in swarm scenes with high exactness anomaly areas.
- The Aggregation of Ensembles framework is a good proposal.

The first of its kind is the extent to which we could truly be aware of irregularity. acclaim in a swarm of accounts The AOE modifies the learned components through numerous fine-tuned CNN social events. to be more unambiguous for swarm abnormality from usual photos.

As a result, more dazzling portions might be removed at different semantic levels.

- We provide a viewpoint that clarifies the need for planning.

The Convents gathered irregularity identifiable proof without any preparation. The unpredictability of the getting ready exams is terrible.

- In crowd lead assessment, we separate the effects of distinct progression frameworks on adjusting.

One way to control crimes or abnormal behavior is through surveillance cameras. By using surveillance videos, abnormal events can be detected in real time when the crime is happening. Smart video systems have been developed by using machine learning, computer vision, signal processing, data mining and other mining techniques to get information from raw videos and further process them for various detection purposes. A huge amount of data is collected through different devices that require an automatic system, which has the ability of making decisions without human intervention for analysis. Video surveillance systems can understand a situation, detect motion, track and classify objects, discover typical behaviors and identify abnormal events [1]. The systems developed for surveillance should have certain requirements that are Interpretability, Autonomous Decision Making and Real Time Execution.

Moreover, qualifying a certain event as anomalous is highly subjective and mainly depends not only on the proposed application but also on the context. Therefore, the context of the scene is extremely related to anomaly definition. The context includes the concept of scene for example a small market present in a neighborhood. The concept of the scene regards different common behaviors and elements for instance, people selling and buying products, money, cabinets, bags, baskets and people talking. All these things describe the idea of a small market. Moreover, at the same time, the elements can be bounded in some other scene; for example, a small market hardly has vehicles, a cinema, antibiotics, people walking, jumping and running around and a long list of those things that cannot belong to such kind of places.

Furthermore, the main characteristics of usual and unusual activities are quite relative. For example, a certain behavior can become normal in one scenario and abnormal in some other situation [2]. This is why; it is difficult to categorize an activity as an abnormal behavior. Only by measuring the extent of abnormality by comparing its similarity with different examples, the researchers can derive a compatible model. It means, anomaly detection system does not only analyze the normal behavior but also detect various deviations from it [3]. This aspect is very important because it helps in differentiating anomalous detection and event recognition. Mostly, the purpose of models for event detection is to determine the type of an event that is happening by using a list of known scenarios. In anomalous event detection system, the abnormal behaviors either could be previously unknown or could be known events.

Anomaly detection can be classified into two categories [4]:

- Local Abnormal Behavior: Local anomaly is the behavior of a particular group of people or objects present in a localized region that is different from their surroundings in spatio-temporal terms [5].
- Global Abnormal Behavior: While, global anomaly corresponds to the unusual and strange actions of a group of people or objects present in the entire scene.

2 Crowd Anomaly Detection

Automated crowd abnormality detection has gain importance due to rising security implications. Crowded places like hospital, markets, airports and religious buildings etc. are usually open and accessible to everyone, which makes them more sensitive to critical accidents. Any incident like terrorist attack, explosion and fire can result in mass causalities and destruction of infrastructure and surrounding businesses. The Conventional methods of object detection or anomaly detection are usually not appropriate and sometimes they fail in densely crowded scenes that have severe ambiguities, occlusions and are cluttered, where any anomalous activity might lead to terrible and adverse situations. A crowd has the characteristics of both psychological and dynamics that make the behavior analyses a very challenging and complex task.

Human crowds are usually goal oriented and it is not easy to model crowd dynamics at a proper level. Therefore, it is necessary to detect, classify and count the crowd behavior. Crowd anomaly detection is a quite challenging task because of some reasons that are:

- Unavailability of training data related to a certain context
- Difficulties in identifying various activities in the relative scene
- Difficulties to model the spatio-temporal relationship between different activities

To address all these challenges, we have proposed a system, to address these problems by focusing on deep learning-based architecture, given below.

(1) For data analyses, Self-collected crowd dataset, BIWI Walking dataset and CBE dataset have been used for three different categories that are individual, group and dense crowd.
(2) A training framework is built for both usual and unusual activity detection.
(3) A GPU based Mask R-CNN architecture is used for crowd monitoring, behavior analyses and anomaly detection that will help in real-time detection.

3 Related Work

The researchers have proposed multiple different approaches for the detection and location of crowd anomalies. In real time crowd behavior and anomaly detection systems have been discussed by using spatio-temporal texture algorithm, which is based on the concept of spatio-temporal video volume, and has been discussed to analyze the characteristics of crowd pattern. A two-stream-based approach that uses the RGB and two-stream neural associations (Convent's) to remove video characteristics to solve the oddity event recognition problem. The RGB stream recognizes unusual events from video diagrams, whereas the Flow move may detect abnormalities from a development subject to a dense optical stream. The event's development arrangements are based on data from the Flow system. As a result, TAEDM may adequately get proportional information on the RGB stream from still images and development between photos in a single movie. On a scale benchmark illuminating assortment, i.e., UCF-Crime, we will assess the produced approach. Currently, there are a few audits in the observation that incorporate PC vision-based methods [6].

This approach presented in [7] deals with macro as well micro level abnormality detection by using Genetic Algorithms (GA), SIFT (Scale invariant feature Transform) algorithm and filtering and masking algorithm. The framework has been applied on structured, unstructured and semi-structured crowds to evaluate the performance of the system. An individual's unusual behavior does not always indicates a threatening attitude toward other people because some people shows their emotions and excitement in a different way. Therefore, a system is proposed in [8] that has the ability of detecting an abnormal action of an individual automatically by taking into consideration the state of mind in which a person is behaving. Irregular LSTM models may be used to forecast the accompanying concentrations or steps of these tracks, which can be utilized in evaluating the collecting scene, using the tracks removed from the jam-packed scene as time-gathering data. The Social-LSTM model was introduced and is a paradigm for evaluating social human activities in gatherings. By walking the entire path susceptible to change, this effort aimed to anticipate the person [9]. An innovative approach of analyzing features by using optical flow method to detect unusual crowd behavior has been discussed in [10]. The presented features are based on the difference in angle that is calculated between optical flow vectors of the current and previous frames at respective pixel's Location.

In crowd management enormous challenges including crowd study, identification, classification and monitoring of abnormal and dangerous activities. Conventional methods that deal with crowd anomalies are not very efficient and effective because of the huge clutter and occlusions. In crowd management enormous challenges including crowd study, identification, classification and monitoring of abnormal and dangerous activities. Conventional methods that deal with crowd anomalies are not very efficient and effective because of the huge clutter and occlusions. A WCAE-LSTM network is built in [11] that does not only has the ability of reconstructing an input frame but is also able to reconstruct an error between reconstructed and input frame to detect crowd anomalies.

The review of crowd monitoring techniques and algorithms presented in [12] show that DISAM models and SD-CNN (Scale Driven Convolutional Neural Network) are proved as novel techniques for not only crowd counting but also for localization in saturated crowd images with maximum precision on various datasets. The proposed system based on histogram of momentum, magnitude (HoMM) in [13] is verified on datasets like UMN, and UCSD developed for crowd analysis and anomaly detection. It is observed from the experiments that anomalous object has higher distance values as compared to the objects of normal behaviors. A novel approach of Aggregation of Ensembles (AOE) is proposed in [14] to detect anomalies in a video sequence of a crowded scene. AOE shows better and efficient results as compared to the capabilities of multiple classifiers and pertained ConvNets.

A 3D Convolutional Neural Network (C3D) with UCSD dataset has been chosen in [15] for feature extraction and the simulations are performed on MATLAB to achieve patch-based detection and frame-based detection. Moreover, an end-to-end deep model named as Convolutional DLSTM (ConvDLSTM) has been proposed in [16] to understand a crowd scene. ConvDLSTM is the combination of Differential Long Short-term Memory (DLSTM) model and Google Net Inception V3 CNN.

4 Deep Learning Architectures

Deep Learning is considered as a sub-area of machine learning. However, it does not consist of task specific algorithms like machine learning because it works by learning data representations and learning can be unsupervised, supervised or semi-supervised. Moreover, there is also not any specific condition for data type because it can handle data in any form, structured and unstructured, including texts, images and sounds. It is also known as end-to-end learning for its ability of learning directly from input data. Many architectures are being used for deep learning and they can be classified into two categories that are:

- Regression-based detectors
- Classification-based detectors

4.1 Regression Based Detectors

You-Only-Look-Once (YOLO)

YOLO is a globally used process for recognition and detection of objects. YOLO network knows the context and gives less background errors. However, it has certain limitations like it struggles with dealing different aspect ratios and small objects present in an image.

Single Shot Detector (SSD)

SSD uses a single deep network for object detection from an image. In this approach, the output is divided into various default boxes by taking into consideration their aspect ratios, which are used to scale the location of feature. It works by combining multiple features of different resolutions to handle multiple objects of different sizes. This algorithm is easy to integrate into multiple systems that require the detection of objects.

4.2 Classification Based Detectors

Region-Based Convolutional Neural Network (R-CNN)

There were some problems with CNNs that they were computationally expensive and too slow. To solve this issue, the researchers developed R-CNN, which used a different object proposal algorithm known as Selective Search to reduce the total number of bounding boxes. R-CNN uses ConvNet to classify objects for achieving object detection of high accuracy.

Spatial Pyramid Pooling (SPP-Net)

Still R-CNN was slow so, to replace R-CNN, the researchers worked on SPP-Net where CNN representation for whole image is calculated only once. However, still a problem was existed, which was the back propagation through SPP layer. This problem paved the path for Fast R-CNN.

Fast R-CNN

This algorithm was developed to fix the problems of R-CNN and SPP-Net by improving the system's speed and accuracy. In R-CNN, region proposals are cropped to resize them but in Fast R-CNN, the complete image is used to proceed further. The detection rate if Fast R-CNN is higher than R-CNN and SPP-Net because multi-task loss is used to train data in a single stage. Moreover, there is no need of memory in Fast R-CNN to store the features of images.

Faster RCNN

Fast R-CNN and Faster R-CNN are quite similar and the only difference is RPN. After performing ROI pooling, the pooled area is used as an input of CNN and FC layers for bounding box regression and softmax classification. Although, Faster R-CNN improves the drawbacks of Fast R-CNN, but still, it has computational redundancy problem at the last stage.

Mask R-CNN

Mask R-CNN is an extended version of Faster R-CNN. It consists of an additional branch to predict segmentation mask on each ROI in the form of pixel-to-pixel alignment however, this type of alignment is not available in Faster R-CNN. Moreover, Faster R-CNN consists of two outputs known as bounding box offset and class label while on the other hand, Mask RCNN consists of three outputs that are bounding box offset, class label and class mask. It is highly flexible and simple to train. To achieve better performance in accuracy and speed it uses ResNet-FPN model to extract features. In this paper, mask R-CNN is applied for the crowd anomaly detection and the framework of this system can be seen in Fig. 1.

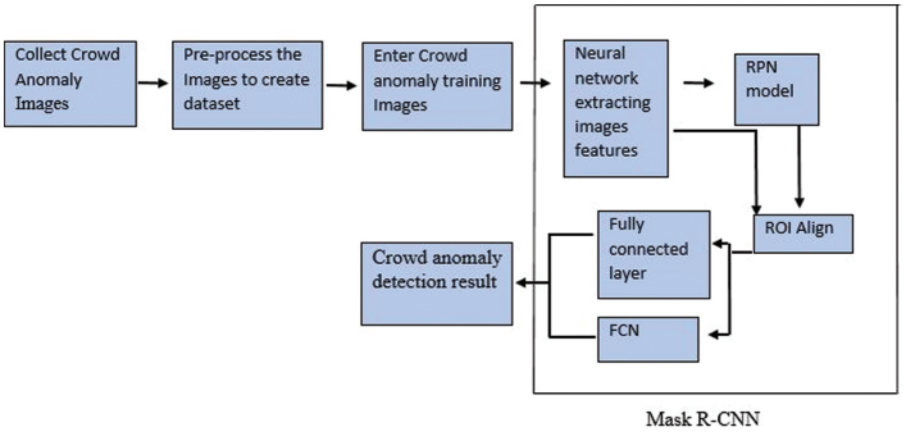

Collect Crowd Anomaly Images
Pre-process the Images to create dataset
Enter Crowd anomaly training Images
Neural network extracting images features
RPN model
ROI Align
Fully connected layer
FCN
Crowd anomaly detection result
Mask R-CNN

Figure 1: Crowd anomaly detection system framework

5 Proposed Methodology

Mask R-CNN model is a state of art architecture that is developed on Faster R-CNN, which operates in two consecutive stages. The first stage is composed of two networks known as region proposal network and backbone network that are Inception, ResNet and VGG etc. Region proposals are the regions of feature map that consist of certain object. After obtaining the proposed regions, the network predicts the object class and bounding box for each region in the second stage. The size of each region can be different however, in order to make predictions the fully connected layers of the network require a fixed size vector, which can be achieved by using RoIAlign method rather than using RoI pool method. In Fig. 2, the network structure of mask R-CNN is shown in terms of a block diagram.

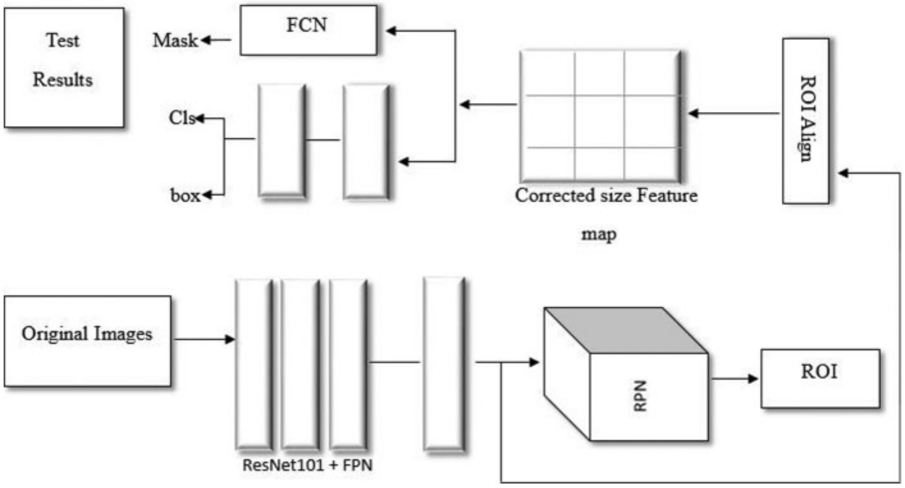


Figure 2: Mask R-CNN network architecture

The algorithm is implemented as follows.

- First, the input image is processed into a model of pre-trained network of ResNet101 + FPN for feature extraction to obtain relevant feature maps.
- The feature map consists of a large number of frames (i.e., ROI), which are obtained from RPN and then frame regression along with a classifier is used to obtain more accurate position information of frames and to filter out irrelevant ROIs.
- After that, the feature map and remaining ROI are directed to ROI Align layer to generate a feature map of fixed size.
- Finally, the output of fixed feature map is divided into two branches. One branch enters to a fully connected layer for frame regression and object classification and other enters to a convolutional network i.e., FCN for pixel segmentation.

All these modules of mask R-CNN network architecture are explained further.

5.1 Feature Extraction and ROI Generation

First, the convolutional layers are employed to extract features by forming a feature vector. By considering a standard kernel size of [3 × 3], we used this feature map to extract the convolutional features. 4000 images of four different classes (i.e., normal situation, explosion, violence and accident) and three other categories of crowd (Individual, group, dense crowd) are collected by using Self-generated Crowd dataset. About 3500 to 4000 frames are extracted using ResNet101 architecture for each class.

After that, the feature output layers of ResNet are fused to attain the strong semantic information with improved accuracy with the help of Feature Pyramid network (FPN). Then, to obtain ROIs, the RPN network is utilized for the prediction of different output layers.

5.2 Region of Interest Alignment

The anchor boxes that are obtained in previous step using RPN are of different sizes. It is not efficient to use these anchor boxes as input to generate feature maps of different sizes. This issue is solved by using ROI align layer that has the ability of transforming these anchor boxes into same sizes. In Faster R-CNN, the layer used was ROI pooling that scales, the sections of a certain feature map corresponding to these ROIs, to a same fixed scale by dividing different region of interests into same and equal number of bins or sections. It is possible that boundaries of region of interest will not align with the boundaries of the bins and feature map. To align all these boundaries, quantization and max pooling is used. However, quantization caused some problems such as reduction in accuracy and misalignment in the generated masks. In order to solve all these issues, ROI align layer is used in Mask R-CNN rather than ROI pooling. In the presented system, each ROI performs the ROI alignment operation with multiple feature layers and then all these ROI alignment layers are fused together for crowd segmentation.

5.3 Mask Acquisition

After ROI alignment, the final steps of bounding box recognition, classification and regression are performed in this last module. The additional branch, which is in parallel to bounding box classification and regression, performs the task of mask prediction for each ROI.

5.4 Loss Function

During training, we formulated a loss function for each ROI as L = Lbox + Lcls + Lmask. For each RoI, the output of mask branch is K∗m∗m dimensional to encode the K binary mask of size m ∗ m. Lmask represents the average cross-entropy loss and is only defined for the k-th mask. Lmask allows the generation of masks for each class without creating any competition among different classes. Morever, regression and classification losses are defined as Lbox and Lcls, respectively.

$$\mathrm{L}_{\mathrm{cls}} = -\log P_u$$

$$\mathrm{L}_{\mathrm{box}} = \sum_{\boldsymbol{i}=1}^{4}(\boldsymbol{Smooth L}_1(\boldsymbol{t_i} - \boldsymbol{v_i}))$$

P is a (k + 1) dimensional vector that is used to represent the pixel's probability belonging to background or k class. For each region of interest, P = P_0, P_1,..., P_k and P_u denotes the probability of corresponding class *u*. Moreover, $t_u = \boldsymbol{t_x^u}, \boldsymbol{t_y^u}, \boldsymbol{t_w^u}, \boldsymbol{t_h^u}$ is used as a translation scaling parameter corresponding to class u. $\boldsymbol{t_h^u}$ and $\boldsymbol{t_w^u}$ denotes the width and height of logarithmic space. v_i denotes the corresponding parameter of the bounding box.

6 Implementation Detail

Our model is implemented on Via tool framework. For our proposed model, we use Mask R-CNN and ResNet-101 as a backbone network. Moreover, to evaluate our system, we collected

Self-Generated Crowd dataset of nearly 4000 images taken from different sources like google images, flicker and YouTube videos. It includes frames with various densities and perspective scales, collected from many different environments, for example streets, shopping malls, university, airports, parks and stadium. Self-Generated Crowd data set description is provided in the Tab. 1.

Table 1: Self-generated crowd dataset description

Dataset	Frames	Groups	Individual/pedestrians	Crowd sets
Dense crowd	500	0–1	0–1	2–3
Sparse crowd	3500	2–5	4–5	2–5

Apart from self-generated dataset, we have also tested our system on two publically available datasets, namely the BIWI Walking Pedestrians dataset [17] and the Crowds-By-Examples (CBE) dataset [18]. We use NVidia 1080 GPU (8GB) to train our deep neural network. Other training parameters are shown in Tab. 2.

Table 2: MASK R-CNN configuration

Training parameters	Values
STEP_PER_EPOCH	1000
EPOCH	100
BACKBONE	resNet 101
LEARNING_RATE	0.001
LEARNING_MOMENTUM	0.9
WEIGHT_DECAY	0.0001
BATCH_SIZE	1
VALIDATION_STEPS	50

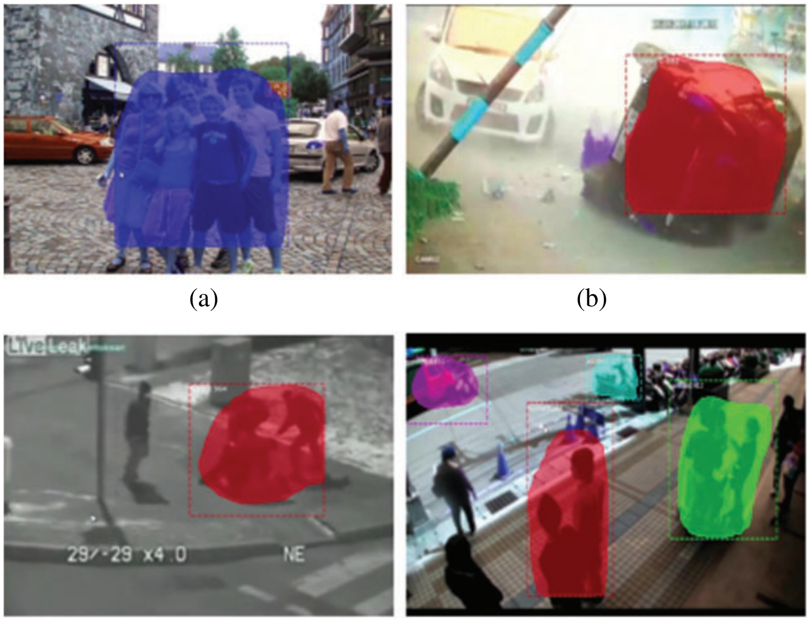
We have divided our results in two cases.

Case 1:

In case 1, we deal with violent group scenario and detect anomalies of four unique categories of groups that are fight, explosion, accident and normal. The detected anomaly is labeled and highlighted by a rectangular box. The Fig. 3 given below show the detected anomalies.

Case 2:

The other case deals with classification of group of people namely identifying dense group, individual humans and people walking. We also annotate the images precisely into group, individual and crowd by using four expert resources. It took fifteen days to execute our experiment and three more days to cross check if there is any mistake. Sample images form annotated dataset with different varieties have been shown in figures below. Fig. 4 represents two images dense crowd, similarly, Fig. 5 depicts individual's images and Fig. 6 provides group image's samples.



(a) (b)

(c) (d)

Figure 3: Detected anomalies (a) normal (b) accident (c) fight (d) explosion

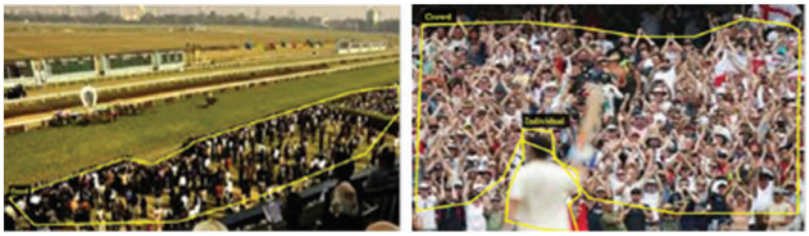


Figure 4: Dense crowd's sample images

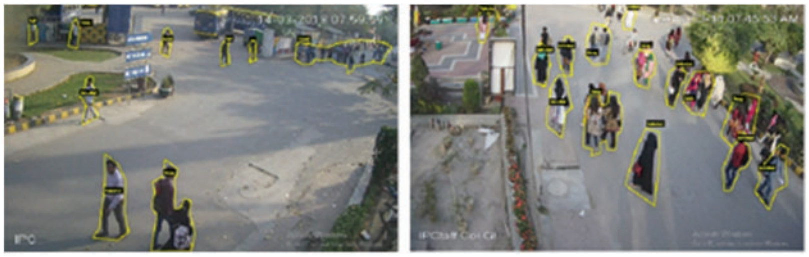


Figure 5: Individual's sample images

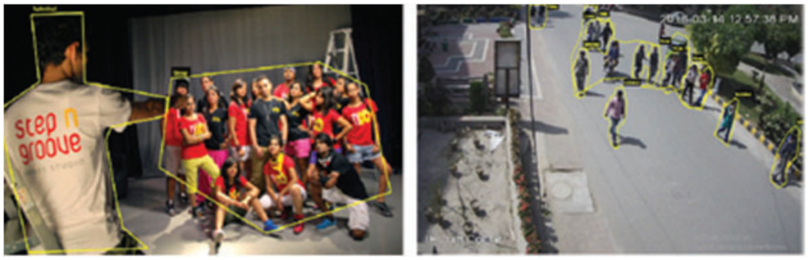


Figure 6: Group's sample images

7 Evaluation and Discussion

To make the model efficient and to minimize error of over-fitting we used early stopping mechanism, which allowed our model to outperform. On 89th epoch, our system started generating finest results. This system stunningly identified and provided segmented portions of group, crowd and individuals. The BIWI dataset records two low crowded scenes, one outside a university, named eth, and one, hotel, at a bus stop, while the CBE dataset records a high-density crowd video outside another university, student003. The precision and recall of testing is described below in the Tab. 3.

Table 3: Precision and recall of system on different datasets

Data set	Group		Crowd		Individual	
	P	R	P	R	P	R
Hotel	77.3	76.47	74.2	75.23	81.23	81.11
Eth	78.12	78.01	69.14	70.32	82.11	80.52
Student	76.41	77.17	82.32	81.33	63.19	65.01
Self-generated	83.12	84.11	84.65	82.12	88.12	89.00

The above table shows that our self-generated dataset shows better performance in terms of precision and recall as compared to other datasets for all three categories (i.e., individual, group and crowd). Also, if we compare the P, R values of all three categories for self-generated dataset, the results show that our system can detect each type of human gathering efficiently.

Moreover, we have also plotted some graphs of losses of different parameters like class prediction, bounding box prediction and mask prediction to evaluate our system's performance. Like Fig. 7 shows the accuracy and loss graph of class prediction with blue line as validation accuracy and red line as training accuracy. Training accuracy is higher than validation accuracy with high gap at initial epochs and small gap at final epochs. As shown in graph, validation accuracy started from 13% and raised up to 83% on 97th epoch. After 97th epoch validation, accuracy became steady with negligible change in accuracy. Similarly, gray line presents training loss and yellow line depicts validation loss. Validation loss started from nearly 1.9 and dropped to 0.0028.

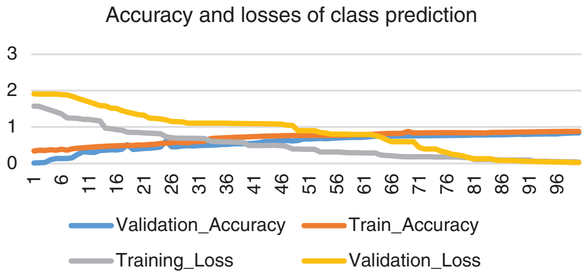


Figure 7: Accuracy and losses of class prediction

Fig. 8 shows the loss graph of bounding box prediction with blue line as validation loss and red line as training loss. As the figure describes the validation loss started from 0.4 and dropped at 0.00142. While the training loss starting at a higher value of 0.7 has reached at a value of 1 at completion of training.

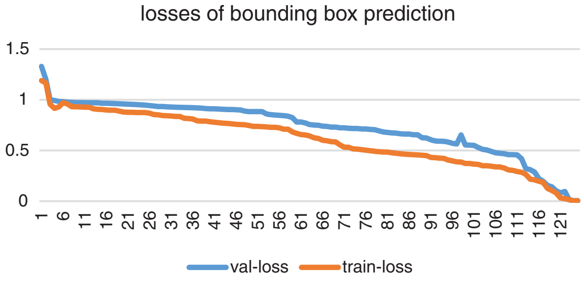


Figure 8: Accuracy and losses of bounding box prediction

Fig. 9 shows the loss graph of mask prediction with blue line as validation loss and red line as training loss. As the figure describes the validation-loss started from 1.8 and dropped at 0.30142. Training and validation losses of bounding box prediction are constantly decreasing with a little fluctuation.

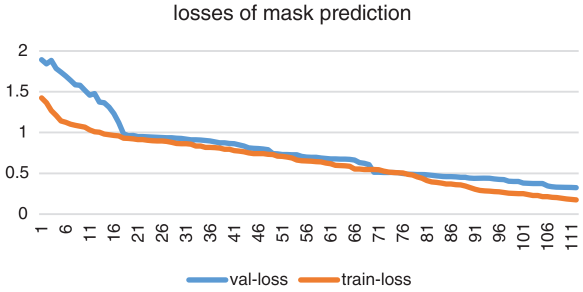


Figure 9: Accuracy and losses of mask prediction

8 Conclusion and Future Work

Effective and real-time crowd monitoring and behavior analyses for public safety has become a challenging task in the real world scenario like hospitals, airports, markets, religious and political gatherings etc. Our research deals with two types of scenarios. One scenario deals with security related violent crime issues like explosion, accident and fighting. The other scenario deals with three types of human gatherings that are dense crowd, people present in a group and people walking alone and behaving normally. We have employed local and global fusion of spatial-temporal features on several images and video sequences to analyze, detect and locate the abnormal events. To evaluate our system, we have used Mask R-CNN with resnet101 as a backbone architecture and three datasets that are Self-Generated dataset, the BIWI-Walking Pedestrians dataset and the Crowds-By-Examples (CBE) dataset have been used to analyze the performance of our system. The simulated results show that our system performs well on Self-Generated dataset and efficiency of the system can further improve by using a large set of data.

In future, we would like to make some improvements in our system based on some suggestions that are:

- Extend our system for more dense crowds
- Evaluate our model on larger real-time scenarios.
- Extend our system to work in occlusion.

Acknowledgement: Authors thank those who contributed to write this article and give some valuable comments.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Liu and N. Kubota, “Intelligent video systems and analytics: A survey,” *IEEE Transactions on Industrial Informatics*, vol. 9, no. 3, pp. 1222–1233, 2013.

[2] Y. Hu, Y. Zhang and L. S. Davis, “Unsupervised abnormal crowd activity detection using semiparametric scan statistic,” in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Washington, DC, United State, pp. 767–774, 2013.

[3] S. Duan, X. Wang and X. Yu, “Crowded abnormal detection based on mixture of kernel dynamic texture,” in *Int. Conf. on Audio, Language and Image Processing*, Columbia, US, pp. 931–936, 2014.

[4] R. Raghavendra, A. Del Bue, M. Cristani and V. Murino, “Abnormal crowd behavior detection by social force optimization,” in *Human Behavior Unterstanding-Second International Workshop*, Amsterdam, The Netherlands, pp. 134–145, 2011.

[5] V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, San Francisco, pp. 1975–1981, 2010.

[6] K. Gupta and K. Varshney, “Crowd anomaly detection in city crime video,” EasyChair 2516-2314, 2021.

[7] N. Ojha and A. Vaish, “Spatio-temporal anomaly detection in crowd movement using SIFT,” in *2nd Int. Conf. on Inventive Systems and Control (ICISC)*, pp. 646–654, 2018.

[8] G. Baliniskite, E. Lavendelis and M. Pudane, “Affective state based anomaly detection in crowd,” *Applied Computer Systems*, vol. 15, pp. 134–140, 2019.

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

[10] C. Direkoglu, M. Sah and N. O'Connor, "Abnormal crowd behavior detection using novel optical flow-based features," in *14th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, Italy, pp. 1–6, 2017.

[11] B. Yang, J. Cao, R. Ni and L. Zou, "Anomaly detection in moving crowds through spatiotemporal auto encoding and additional attention," *Advances in Multimedia*, vol. 2018, pp. 1–8, 2018.

[12] J. A. Shah, A. Khan, K. Kadir, W. Albattah and F. Khan, "Crowd monitoring and localization using deep convolutional neural network: A review," *Applied Sciences (MPDI)*, vol. 20, no. 14, 2020.

[13] S. Bansod and A. Nandedkar, "Crowd anomaly detection and localization using histogram of magnitude and momentum," *The Visual Computer*, vol. 36, pp. 36, 2020.

[14] K. Singh, S. Rajora, D. Vishwakarma, G. Tripathi, S. Kumar *et al.*, "Crowd anomaly detection using aggregation of ensembles of fine-tuned convNets," *Neurocomputing*, vol. 332, pp. 371, 2019.

[15] R. Ramesh, "Abnormality detection with deep learning," *Speech, Audio, Image and Video Technology (SAIVT)*, 2018.

[16] N. Zhuang, J. Ye and K. Hua, "Convolutional DLSTM for crowd scene understanding," in *IEEE Int. Symp. on Multimedia*, Kyoto, Japan, pp. 61–68, 2017.

[17] S. Pellegrini, A. Ess and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Kyoto, Japan, pp. 261–268, 2009.

[18] A. Lerner, Y. Chrysanthou and D. Lischinski, "Crowds by example," *Comput. Graph. Forum*, vol. 26, pp. 655–664, 2007.