Crowd Behaviour Monitoring Using Aerial Surveillance

A Thesis Submitted for the Degree of Master of Computer Science

W. A Wajirasena

University of Colombo School of Computing

2020

DECLARATION

I hereby declare that the thesis is my original work, and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: W. A. Wajirasena

Registration Number: 2018/BA/037

Index Number: 18880374

_____

Signature of the Student & Date

This is to certify that this thesis is based on the work of Mr. /Ms. _____ under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name:

_____

Signature of the Supervisor & Date

I would like to dedicate this thesis to my parents and my teachers

ACKNOWLEDGEMENTS

I would like to acknowledge my supervisor Dr. Kasun Karunanayaka and the staff of the University of Colombo School of Computing for assisting me to do this research and all the researchers that have contributed to the publicly available data sets that I used in this research.

ABSTRACT

With the rise of population and ever-growing megacities, it is essential for the authorities to monitor the crowd movements. Controlling crowd or mass gatherings at special events such as entertainment events or sport events and places that are essential to the daily lifestyle of people such as airports, hospitals is essential in the modern day. In this thesis we present two methods based on human action detection and crowd density prediction to monitor crowd behaviour.

Keywords: Crowd Monitoring, Crowd Density, Crowd Behaviour, Action Detection, Arial Surveillance

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

With the rise of population and ever-growing megacities, it is essential for the authorities to monitor the crowd movements. Controlling crowd or mass gatherings at special events such as sports events or entertainment events and places that are essential to the daily lifestyle of people such as airports, hospitals is essential in the modern day. Monitoring these gatherings is essential to reduce the congestion, ensure the safety of the people, to provide better emergency services in case of congestion related emergencies and to provide optimize services to the crowd (Ilyas, et al., 2020). In the current context of global pandemics, the importance of crowd monitoring in order to maintain social distancing is becoming extremely important (Rezaei & Azarmi, 2020).

In general crowd monitoring and analysis have multiple potential use cases in different domains and applications. These include emergency services, crowd flow management and design in public spaces, safety monitoring, disaster management, analysing group behaviours and analysing and prevention of disease spreading in public places etc. These applications have created a demand for the research and development in analysing and managing crowds, analysing behaviour of individuals and groups of crowd, density and prediction, specific behaviour prediction, flow analysis and mass tracking. Recent COVID19 pandemic has created a wider attention to analysing the crowd behaviour in terms of social distancing and health and safety aspects (Rezaei & Azarmi, 2020).

Crowd gatherings detection and density estimation has proven useful for consequent steps of intelligent analytics and applications. Therefore, real-time crowd monitoring systems have received more attention, especially during the last decade. However, this is still a challenging area specially in unconstrained environments.

1.1 Applications

Several applications are relying exclusively on an efficient and robust crowd management and monitoring system. They are;

People counting in densely populated areas: World population increases daily and maintaining order in certain public areas such as roads, railway stations, parks and airports is

essential., Counting people is an essential factor in a crowd management system. Most importantly an increase in the number of people in smaller areas could create problems such as fatalities or physical injuries. Detecting these scenarios as early as possible is essential. In such scenario of crowd management, counting the number of people provides accurate information about certain conditions such as congestion at some areas. Despite the research this is still a challenging task. The challenges occur due to different illumination conditions, occlusion, cluttering, scale variations and perspectives variations (Rabaud & Belongie, 2006) (Bharti, et al., 2019).

Public Events Management: Events such as mass religious gatherings, concerts, sports events and political gatherings should be managed to avoid unwanted disastrous situations. This is also beneficial in managing all available resources such as security and spatial capacity (Boulos, et al., 2011), (Lv, et al., 2014).

Military Applications: Monitoring movement of armed forces and police is helpful in both conflict and non-conflict situations.

Disaster Management: Overcrowding can happen in public gatherings such as concerts and sport events when a large portion of the crowd moves in random directions. This could cause life-threatening conditions. Better crowd management can be done in such events to avoid accidents (Martani, et al., 2017).

Suspicious-Activity Detection: Crowd monitoring systems can be employed to detect suspicious activities such as terrorist attacks within public gatherings. Traditional machine learning methods fail to perform well in these kind of situations (Chackravarthy, et al., 2018).

Health and Safety Monitoring: Monitoring the maintenance of health and safety protocols is essential in large crowd gatherings. With highly contagious diseases such as COVID 19 it is essential to maintain social distance and other safety measures in the public gathering and also to avoid crowd gatherings as much as possible (Rezaei & Azarmi, 2020).

1.2 Challenges and Motivations

Efficient crowd behaviour monitoring can lead to various useful applications which have further potential in the computer vision paradigm. However, the problem of crowd behaviour monitoring is yet to be solved satisfactorily particularly in the real-world conditions that present multiple different challenges. Some of the challenges that arise can be summarized as follows;

When two or more than two objects in close proximity to each other and as a result merge, in such scenarios, it is hard to recognize each object separately. Consequently, monitoring and measuring accuracy of the system becomes difficult. This is a significant challenge when using remote monitoring methods such as Unmanned Aerial Vehicles (UAV).

Irregular distribution of objects is another complex challenge faced by the monitoring systems. Irregular image distribution happens when the density distribution of a video or image is varying. Crowd monitoring in irregular object distribution is challenging (Zhang, et al., 2015).

Clutter and image noise is also a present challenge in crowd monitoring. Clutter can be described as a non-uniform arrangement of various objects which are close to each other. Especially when methods like UAVs are used, the number of pixels per individual person is very small. In such cases cluster and image noise can present a severe challenge to the behaviour monitoring task (Idrees, et al., 2013).

Another major problem that crowd monitoring systems face is the problem of changing perspective and aspect ratios. To overcome this, we can use a camera mounted at a fixed angle in a drone and the drone can be flown at constant a specific height from the ground.

Availability of data is also a challenge in the case of crowd monitoring. The unavailability of a public dataset is one major problem towards the development. However, in the recent years several comprehensive datasets have been released to the public.

With the recent COVID19 pandemic outbreak there were several challenges in monitoring crowd behaviour and taking necessary actions to prevent the spread of the diseases. In order to monitor large public spaces, the use of UAVs is essential. This gives us a unique sub problem

to monitor crowd behaviour using UAVs and how to apply crowd behaviour monitoring to enforce social distancing in large public areas.

CHAPTER 2

LITRATURE REVIEW

Research done on crowd related area could be categorized into two major domains based on literature. They are crowd management and crowd monitoring. Crowd monitoring than can be splitted to counting, localization and behaviour (Khan, et al., 2020). Since the focus of this research is on crowd monitoring that area will be explored in the coming sections.

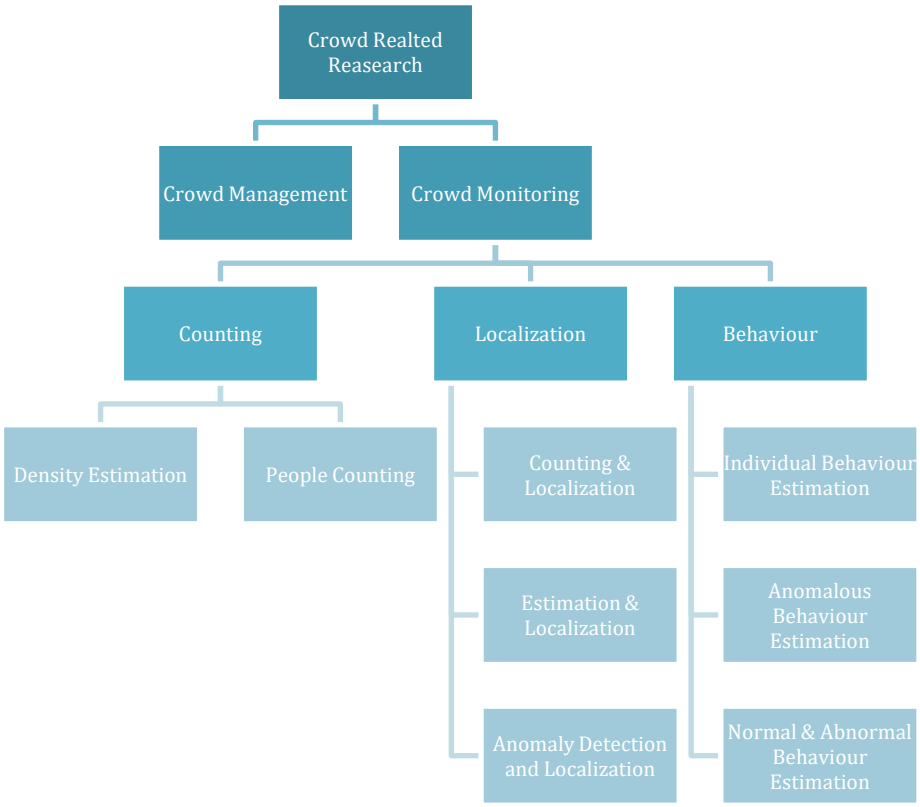

Figure 1 Crowed Related Research

Crowd Counting could be then categorized into two parts. They are crowd density estimation and people counting. Crowd Localization can be categorized to three main subcategories. They are crowd counting and localization, estimation and localization and anomaly detection and localization.

Finally, the crowd behaviour category can be divided into three subcategories. They are individual behaviour estimation, anomalous behaviour detection and normal and abnormal behaviour detection. These categorization levels of crowd monitoring are depicted in Figure 1.

2.1 Crowd Management

Crowd Management has gained significant progress over time. Mohamed et al. propose a Finite state machine to model crowd movement (Mohamed & Parvez, 2019). Nasser, N. et al. have proposed a framework based on weighted round robin to reduce the congestion and overcrowding during hajj pilgrimage (Nasser, et al., 2017). This framework is designed as a proactive system that predicts potential congestion problems using the smart monitoring systems. Also, there are several models built using image processing methods such as optical flow and motion history (Yimin, et al., 2019).

2.2 Crowd Monitoring

Crowd monitoring can be categorized into subcategories, crowd counting, crowd localization and crowd behaviour.

2.2.1 Crowd Counting

Crowd counting can be interpreted simply as counting the number of people in a crowd. Multiple researches have been done in the area of crowd counting. Xu M et al has proposed a crowd density prediction method based on you only look once (YOLO) convolutional neural network (CNN) (Xu, et al., 2019). In (Khan, et al., 2019) a head counting, and localization technique called Density Independent and Scale Aware Model (DISAM) is proposed. The researchers mention that this method performs well for high density crowds where the human head is the only visible part. In here scale aware head proposals are created in the first step using a scale map of the image. Then these proposals are used by a Region Based Convolutional Neural Network (R-CNN) to detect the heads. In (Xu, et al., 2019) two CNN modules are called Scale Preserving Network (SPN) and Learning to Scale Module (L2SM) proposed. The SPN module uses a VGG16 (Simonyan & Zisserman, 2015) CNN as the backbone network to generate an initial density map. Then the L2S module is used to do the re-prediction of highly dense areas. In (Pu, et al., 2017) researchers have used two well-known deep convolutional networks namely GoogleNet and VGGnet to detect crowd densities.

2.2.2. Crowd Localization

Localization of crowds in an image is used to figure out how people have been distributed in an area. It is an important input for crowd management. Information gathered from crowd localization can be used to detect and monitor an individual in dense crowds. In (Lian, et al.,

2019) researchers propose a regression guided detection network (RDNet) that can simultaneously estimate the number of heads and localize those heads by means of bounding boxes. In the same way in (Rodriguez, et al., 2011), a density map has been used to localize the heads in dense images with accurate results. A CNN called LSC-CNN is used in (Sam, et al., 2019) to identify localization. This uses a metric named Mean Localization Error (MLE) to optimize it. In (Xue, et al., 2020) Compressed Sensing based Output Encoding (CSOE) has been proposed. This model has improved the efficiency of localization in highly dense crowded situations.

2.2.3 Crowd Behaviour

Crowd behaviour analysis and detection plates an important role in organizing peaceful events. It is also important in monitoring public spaces for anomalous behaviour of people. The complexity of behaviour identification and abnormal behaviour detection is an important issue in video processing. Several different methods and techniques for the crowd behaviour detection have been proposed.

In (Rohit, et al., 2017) and (Lahiri, et al., 2018) an optical flow image processing and motion history image technique were used to detect the crowd behaviour. In the same way an optical flow method combined with Support Vector Machine (SVM) was roposed to detect abnormal behaviour in (Yimin, et al., 2019). In (Rao, et al., 2016) researchers have presented a fast approach to detect anomalous crowd events nonlinear dimensionality reduction (NDR) using Isometric Mapping (ISOMAP). Wang et al propose a twostep process. A novel descriptor used to capture the multi-frame optical flow information. Then, the feature descriptors of the normal samples are fed into an Auto Encoder based network called cascade deep Auto Encoder (CDA) network for training. Finally, the abnormal samples are distinguished by the reconstruction error of the CDA in the testing procedure (Wang, et al., 2020). In (Gao, et al., 2019) Hybrid Random Matrix (HRM) (which is used to project data from high dimensional space to low dimensional space) and deep neural network was used to the detect the violent behaviour of crowds. SIFT feature extraction technique and Fixed-width clustering algorithm with YOLO were used to detect crowd behaviour in (Yang, et al., 2018).

2.2 Datasets

Due to this being a relatively newly explored area there are not many publicly available datasets. Most datasets that are available have one or two scenes. Therefore, may not be able to be used in generic crowd analysis. In this section we list the available datasets.

2.2.1 NWPU-crowd (Wang, et al., 2020)

Most of the available datasets are small scale and they do not meet the need of large and versatile data of CNN based methods. To overcome this problem the NWPU-Crowd was introduced recently. This dataset consists around 5000 of images with more than 2 million heads annotated. Compared to other datasets this has a large variation of scenes with different crowd densities and illumination conditions. These were collected from the internet and by self-shooting. This dataset has scenes from malls, streets, stations and public areas.

2.2.2 UCF-QNRF (Idrees, et al., 2018)

This dataset consists of 1535 images. It has significant variation in crowd density the resolution of the dataset is also high, which is in the range of (400 * 300 pixels to 9000 * 6000 pixels). The data has been collected from the web which includes a diverse set of locations. This also has a range of scenes containing a very diverse set of crowd densities, perspectives, and lighting conditions. It also contains other physical objects like buildings, sky, roads and vegetation. Therefore, this presents a more realistic set of data.

Figure 2 Sample Images from UCF-QNRF Dataset

2.2.3 Shanghai Tech (Zhang, et al., 2016)

This dataset consists of 1198 images and 330,165 heads. This is intended to use in comparatively large-scale crowd counting. This dataset consists of two parts namely Part A and Part B. Part A consists of 482 images. They were taken from the Internet. Part B consists of 716 images which were captured from the metropolitan street in the city of Shanghai. The researchers have defined training and testing sets for both Part A and B. Part A has 300 training images and 182 testing images whereas Part B has 400 training images and 316 testing images.

2.2.4 JHU-CROWD++ (Sindagi & Yasarla, 2019)

This is a comprehensive dataset with 4,372 images and 1.51 million annotations. It has a diverse set of scenarios and environment conditions such as variation of density, day and night and different weather conditions.

2.2.5 VisDrone (Zhu, et al., 2020)

The VisDrone2020 dataset was compiled by the AISKYEYE team at the Lab of Machine Learning and Data Mining, Tianjin University, China. This consists of 400 video clips and 10,209 static images, captured by various drone-mounted cameras. These consist of multiple environments varying from urban to country and a variation of crowd densities. It has 2.6 million annotated objects including people.

2.6 Discussion

Significant research work has been done in crowd analysis domain in the past few years.

Although many datasets have been introduced, significant portion of these datasets address the crowd counting problem and less focus has been given to the localization and behaviour analysis. UCF-QNRF and NWPU-crowd are two datasets that have sufficient information for behaviour analysis and localization. Therefore, there is a significant space regarding the versatility of publicly available datasets. Most of the time the labelling of these datasets has been carried out by manual annotations. This depends on the subjective perception of a small number of people. Therefore, there is a chance of human error to exist.

Some studies show that using traditional machine learning approaches with had crafted features can perform better than the state-of-the-art deep learning-based methods. However, it is not conclusive that the performance those methods are better than deep learning in general. Therefore, we think that use of a deep learning-based approach is still essential in the area of crowd analysis. In most cases the poor performance of deep learning approaches can be attributed to the limitations of the data sets. This is a significant drawback that is faced by deep learning-based methods.

When applied the performance of the conventional machine learning was acceptable with data collected in simple and controlled environment scenes. However, in a more complex scenario the performance of these methods drops. Unlike the traditional methods, deep learning-based methods performed well by a large margin. They were able to learn comparatively higher levels of abstraction from data. Therefore, these methods can reduce the need for complex feature engineering. However deep learning is also a complicated process that requires various choices from the researcher like network architecture and other inputs. Therefore, researchers mostly take a trial-and-error approach with the deep learning methods. It takes more time to build a solution based on deep learning methods, because of this. Despite these drawbacks, in recent years more focus has been given to the deep learning-based methods in crowd analysis. However, the drawback of the deep learning based methods are, if the data set is not sufficient it can lead to inferior performance.

However Convolutional Neural Networks with relatively more complex structures are still having issues when dealing with multi-scale problems. In the same way these methods are still not performant enough when estimating density distributions.

Most of these CNN methods employ multiple pooling layers that will result in low resolution and feature loss. This could be a significant concern in highly crowded scenes, especially if the images are taken from a UAV. Deeper layers in a CNN extract more high-level features and the shallower layers extract the low-level features such as spatial information Therefore combination of both shallower layers and deeper layers will result in more accuracy in predicting crowd densities.

2.7 Summary and Conclusion

Crowd analysis is an essential task for several particle applications. It can be divided into counting, localization, and behaviour analysis. Crowd analysis is significantly challenging in complex real-world scenarios. However, the recent advancement of deep learning-based methods shows many achievements in this area. There are still challenges in the unavailability of datasets for diverse crowd analysis research.

CHAPTER 3

METHODOLOGY

Our research focuses on crowd behaviour monitoring in the health and safety and social distancing aspects. The need for crowd behaviour analysis in this domain can be categorized into two areas.

- Enforcing isolation and lock down in certain areas.
- Detecting highly dense crowd in public areas

These requirements are diverse and need a larger area to monitor and a normal Closed-circuit television (CCTV) may not be suitable. Therefore, we propose UAVs to acquire footage for these applications.

3.1 Enforcing Isolation and Lockdown

When isolation is enforced in certain areas, we expect people to be in their houses. Therefore, to monitor lock down and isolation we propose to use a people detection algorithm on top of the footage received from the UAV. By detecting people in an outside environment, we can detect whether the isolation is maintained. Also, if two people are in proximity, we can conclude that there is a higher risk of lock down violation. Also, if we can classify the interaction between people, we can detect the serious violations of social distancing from the UAV footage. Stranded deep convolutional networks can be used to detect these scenarios with some transfer learning on the available crowd datasets.

To achieve this goal, we used the publicly available Okutama-Action Dataset (Barekatain, et al., 2017) to train a deep neural network that can detect the people and their actions in a drone footage.

3.1.1 Dataset Preparation

The Okutama-Action Dataset consists of training set consists of 43 video sequences of a total at 30 FPS and 77365 frames in 4K resolution. Out of these 33 videos have been selected as the training set which contains about 60000 frames.

The dataset had the action sequences listed in the following diagram

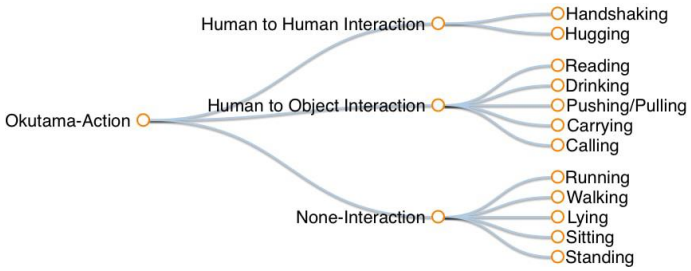

Figure 3 Actions Included in the Okutama Dataset

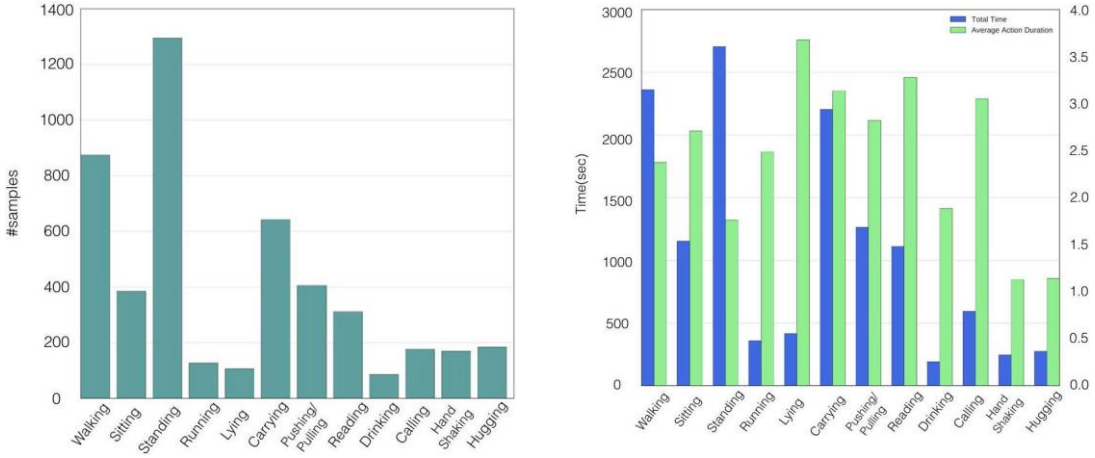

Figure 4 Action Distribution of the Okutama Dataset

For the final training and test set we selected a random sample of 25000 frames for the training set and 12000 frames for the test set considering the time consumed for training and testing. In order to process the data set to the deep neural network this video sequences were split to frames using the python Open CV library. For each of the video sequence these frames were saved in directories with name of the video file.

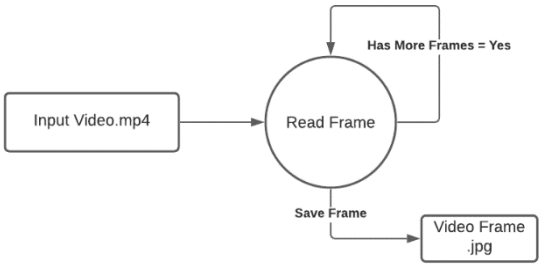

Figure 5 Video Frame Processing

Okutuma dataset annotations were in the following format.

```
0 1475 109 1574 205 0 0 0 0 "Person" "Walking"
0 1473 115 1568 210 1 0 0 1 "Person" "Walking"
0 1473 123 1565 219 2 0 0 1 "Person" "Walking"
0 1470 132 1560 224 3 0 0 1 "Person" "Walking"
```

Figure 6 Sample Annotation

Each line contains 11 columns, delimited by spaces. The definition of these columns are;

1. Track ID – An ID to track the people across the video frames
2. Xmin - The top left x-coordinate of the bounding box.
3. Ymin - The top left y-coordinate of the bounding box.
4. Xmax - The bottom right x-coordinate of the bounding box.
5. Ymax - The bottom right y-coordinate of the bounding box.
6. Frame - The frame that this annotation represents.
7. Lost - If 1, the annotation is outside of the view screen.
8. Occluded - If 1, the annotation is occluded.
9. Generated - If 1, the annotation was automatically interpolated.
10. Label - The label for this annotation, enclosed in quotation marks. This field is always "Person".
11. (+) actions - Each column after this is an action.

Out of these we are interested in Xmin, Ymin, Xmax, Frame and Label. A python code was developed to extract these with the matching frame file name.

Our selected deep learning model You only look once (YOLO) v5 model (Jocher , et al., 2021) accept the annotations in the following format txt files;

- One annotation row per object
- Each row has following columns; class x_center y_center width height.
- Box coordinates must be in normalized xywh format (from 0 - 1). If your boxes are in pixels, divide x_center and width by image width, and y_center and height by image height.
- Class numbers are zero-indexed (start from 0).

Figure 7 YOLO Annotation Format (Jocher , et al., 2021)

Each annotation from the Okutuma dataset was converted to this format using a python code.

3.1.2 Model Development

With the annotated dataset the problem of action detection and the people detection fall into the category of object detection. There have been many researches done on the area of object detection in the recent past. Several prominent models have been emerged such as Fast Region based Convolutional Neural Network (Fast R-CNN) (Girshick, 2015), Faster R-CNN (Ren, et al., 2015), Single Shot Multibox Detector (SSD) (Liu, et al., 2016) and You Only Look Once (YOLO) (Redmon, et al., 2016).

Out of the above Fast R-CNN and Faster R-CNN uses two phases to the object recognition and localization. Therefore R-CNN are not very suitable in detecting object in real-time. On the other hand, SSD and YOLO networks do the object detection and the localisation in one pass. Therefore, they have higher speed in detection with the sacrifice of detection accuracy. (Srivastava, et al., 2021). Because of the detection speed and the application of use we decided not to use R-CNN models in this application.

Out of SSD and YOLO, YOLO v3 model (Redmon & Farhadi, 2018) has better overall performance. (Srivastava, et al., 2021). Therefore, we decided to choose the latest version of the YOLO model YOLO v5 (Jocher , et al., 2021).

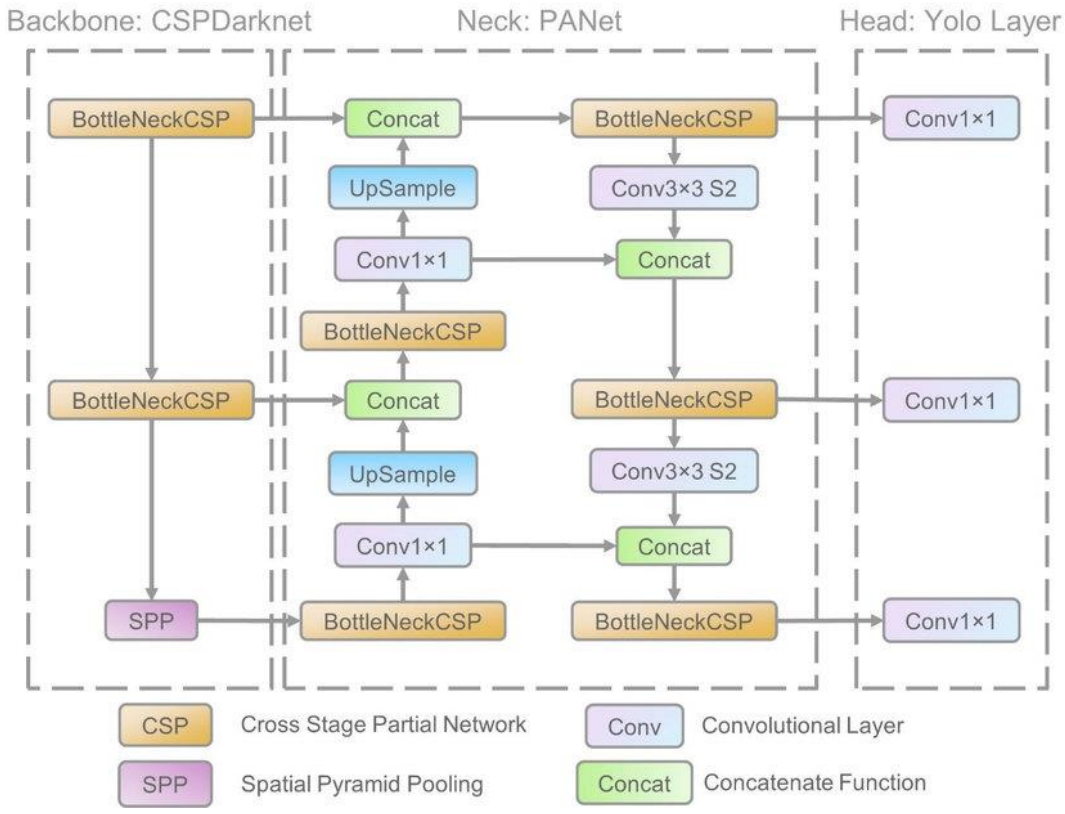

Figure 8 The network architecture of Yolov5 (Xu, et al., 2021)

With the introduction of CSPDarknet53 backbone network in the new YOLO v4 (Bochkovskiy, et al., 2020) the accuracy of the small object detection in YOLO network has been improved (Dwivedi, 2020). This is an important factor to consider in people detection in UAV images due to the nature of the images captured where humans are visible as small objects as shown in Figure 9 People Detection in Okutama-Action Dataset (Barekatain, et al., 2017)

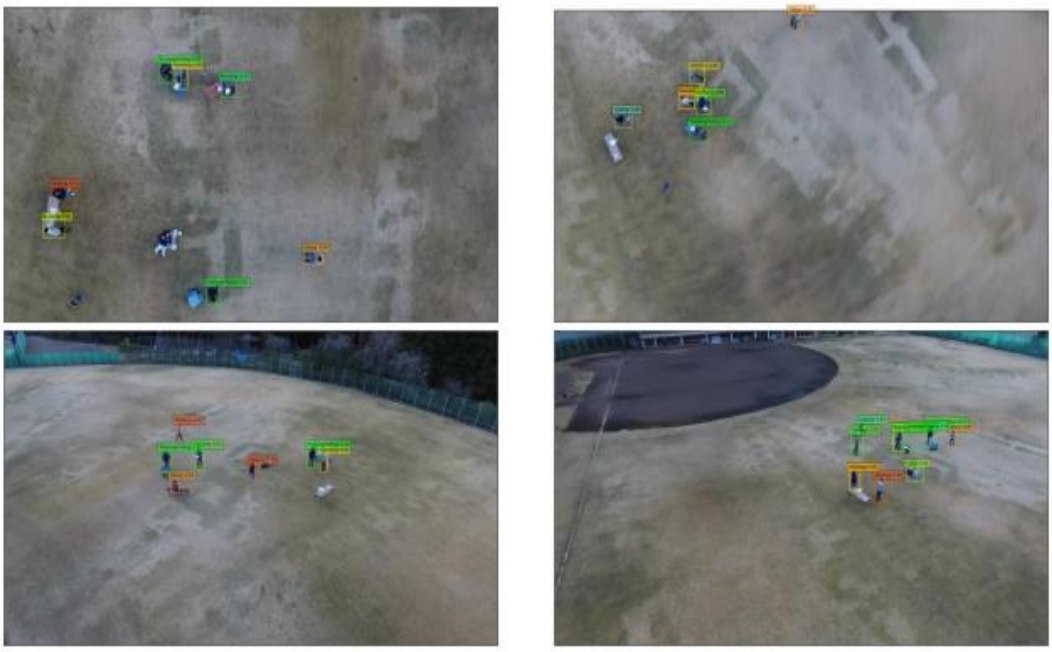

Figure 9 People Detection in Okutama-Action Dataset (Barekatain, et al., 2017)

YOLO v5 provides several model configurations. They are YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. Out of these YOLOv5m and YOLOv5l was used in this application.

3.1.3 Data Augmentation

Following image augmentations were used in order to improve generalisation of the models.

- Image flip up-down.
- Image flip left-right.
- Image rotation.
- Image translation.
- Image scale.
- Image shear.
- HSV augmentation.

3.1.4 Model Training

The model excepts images in the 640 x 640 size. We trained the model with different learning rates 0.01, 0.015, 0.02 and 0.03. Out of them 0.015 gave the best results. We also used `OneCycleLR` in the PyTorch library to achieve faster convergence. We used transfer learning to improve the learning speeds by using a pretrained model and freezing the backbone layers and learning only the other layers.

We trained the model for 200 epochs and the Figure 10 Learning Progress of YOLOshows the learning progress of the neural network.

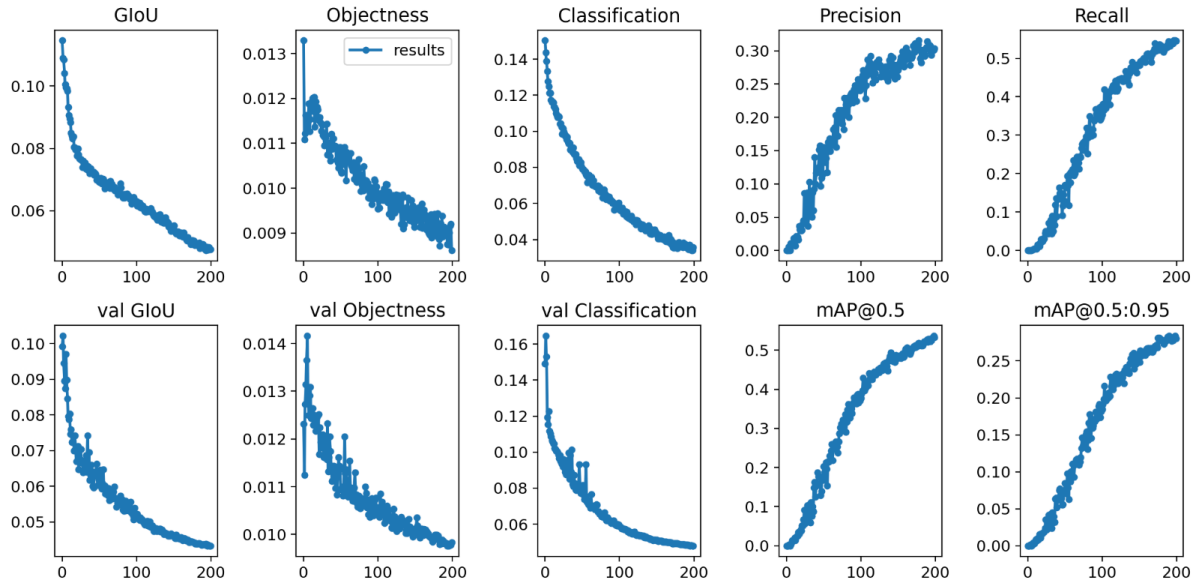

Figure 10 Learning Progress of YOLO

We used an Amazon Web Services GPU instance since the model needs to be trained in a high spec GPU. We trained the model in the g3s.xlarge instance of AWS. It has the following hardware specifications.

Name	GPUs	vCPU	Memory (GiB)	GPU Memory (GiB)
g3s.xlarge	1	4	30.5	8

Table 1 AWS Virtual Machine Specifications

The model training took roughly 12 hours for the YOLOv5m model and 18 hours to the YOLOv5l model.

3.1.5 Model Prediction

Shows the outputs of the predictions the results will be discussed in detail in the evaluation section.

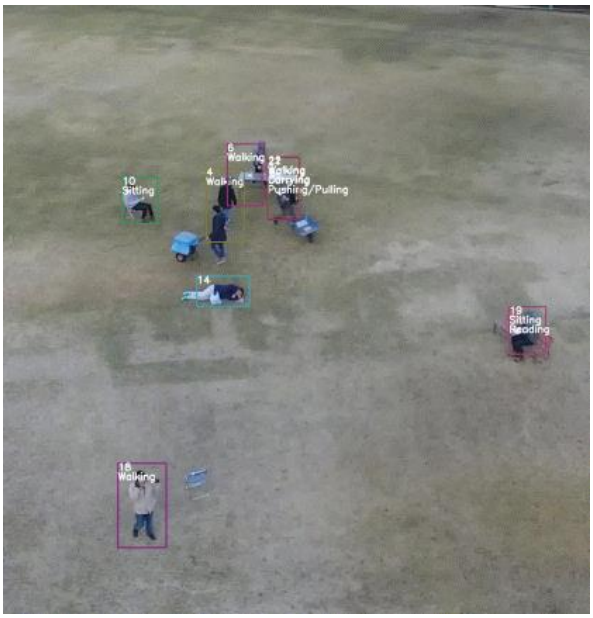

Figure 11 Output of the Model

3.2 Detecting Highly Dense Crowd in Public Areas

Detecting highly dense crowds in public areas is essential to optimise the crowd patterns in such a way that people are having enough social distance. We propose to use deep convolutional neural networks to detect crowd density of UAV aerial images in order to do this. However, since the limitations in the UAV based approach (due to the flight time) we propose to take periodic images at a predefined interval to visualize the crowd density change over time. These outputs can then be used to optimize the crowd flow. To achieve this, we used the deep neural network SFANet proposed by (Zhu, et al., 2019)

3.2.1 Dataset

To do this task we used both UCF-QNRF (Idrees, et al., 2018) dataset. UCF-QNRF dataset contains 1535 images which are divided into train and test sets of 1201 and 334 images respectively.

3.2.2 Model Development

There have been several models proposed to do the crowd density prediction with varying accuracy. However, the SFANet proposed by (Zhu, et al., 2019) in their publication Dual Path

Multi-Scale Fusion Networks with Attention for Crowd Counting has shown a significant improvement in accuracy.

The SFANet neural network has two main components; namely a VGG net backbone convolutional neural network as the feature extractor and a dual path multi scale fusion network as the back end network. This network has two paths. One is for generating attention map by highlighting the crowd regions in images. The other path is for adding multi scale features to the network and to generate the final density maps.

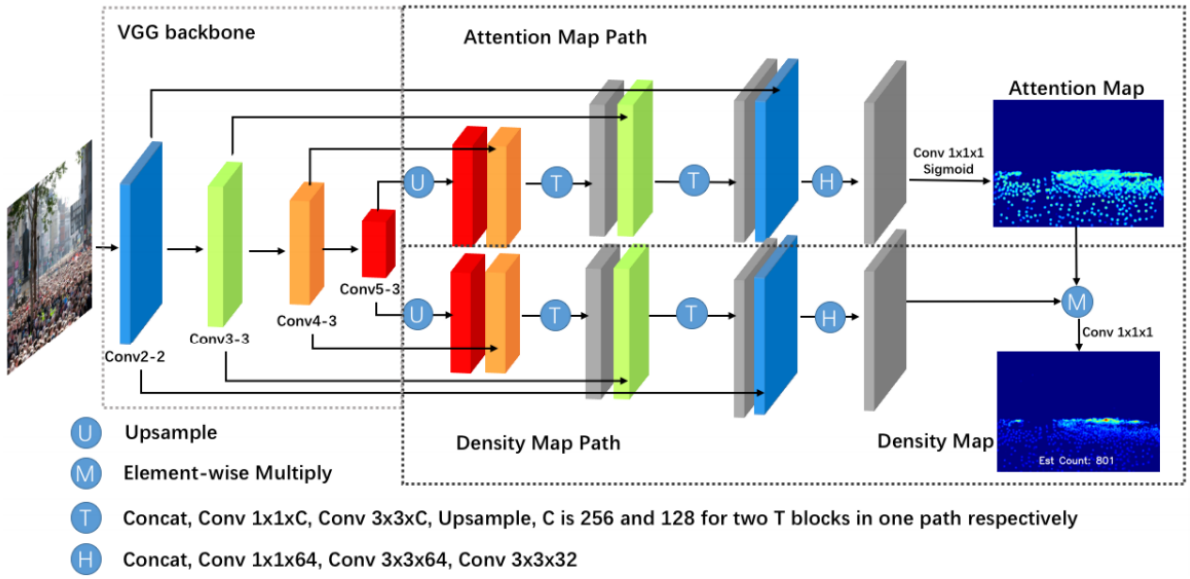

Figure 12 SFANet Architecture

These crowd counting datasets contains annotations where each person is annotated by a point (which is usually the centre of the head). Most of the methods use density map estimations to convert the point annotations into a “ground truth” density map through a Gaussian kernel and use it to train a density map estimator. However, a more robust method to create the density map have been proposed using Bayesian Loss by (Ma, et al., 2019) in their research Bayesian Loss for Crowd Count Estimation with Point Supervision. We have used the Bayesian Loss function proposed by (Ma, et al., 2019) as the loss function in the SFANet.

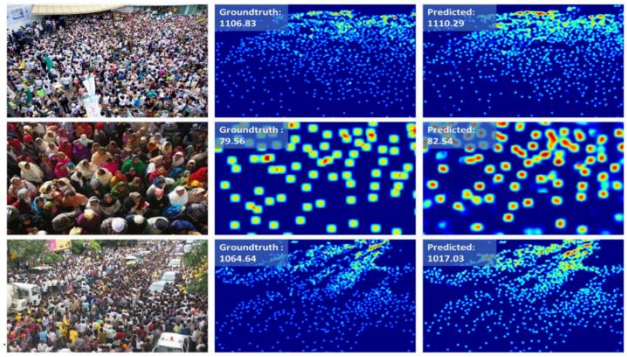

Figure 13 Crowd Density Prediction in (Ma, et al., 2019)

3.2.3 Model Training

We trained the model with different learning rates 0.00005, 0.0001, and 0.0002. Out of them 0. 00005gave the best results. We trained the model for 900 epochs. We used an Amazon Web Services GPU instance since the model needs to be trained in a high spec GPU. We trained the model in the g3s.xlarge instance of AWS. It has the following hardware specifications.

Name	GPUs	vCPU	Memory (GiB)	GPU Memory (GiB)
g3.8xlarge	2	32	244	16

Table 2 AWS Virtual Machine Specifications

The model training took roughly 26 hours.

3.2.4 Model Prediction

Shows the outputs of the predictions the results will be discussed in detail in the evaluation section

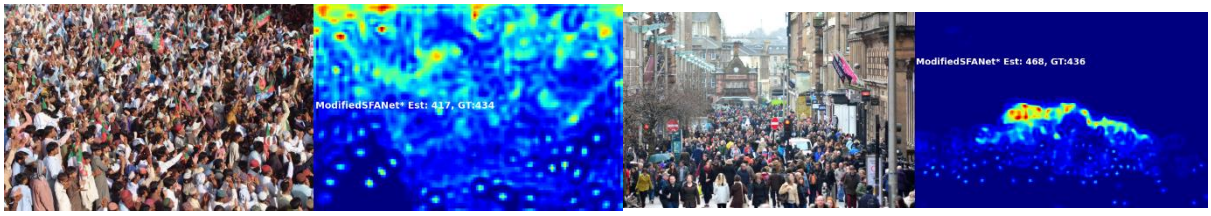

Figure 14 Crowd Density Prediction Output

CHAPTER 4

EVALUATION

We have used several metrics to evaluate the accuracy of the predictions made in the two tasks that we mentioned in the methodology.

4.1 Metrics

4.1.1 Intersection Over Union (IOU)

Intersection Over Union (IOU) is mainly used in applications related to object detection, where we train a model to output a bounding box that fits perfectly around an object.

$$IOU = \frac{Area\ of\ Intersection\ of\ two\ boxes}{Area\ of\ Union\ of\ two\ boxes}$$

The IOU of two bounding boxes can have values between 0 and 1. If the bounding boxes are not intersecting at all the area of their intersection would be 0 and the IOU would also be 0. In the case of complete overlap between the two boxes the IOU will be 1.

4.1.2 Mean Average Precision (mAP)

Average precision combines recall and precision for ranked retrieval results. In the context of object detection if we ranked the detected objects for a single class, the average precision is the mean of the precision scores after each relevant object is detected. Average Precision indicates whether a model can correctly identify all the positive examples without incorrectly identify too many negative examples as positive. Therefore, higher average means the model performs well detecting positive examples.

The Mean Average Precision for object detection is the average of the Average precision calculated for all the classes.

4.1.3 Mean Absolute Error (MAE)

Mean absolute error is a measure of deference between the predicted value of a model and the actual observation. It can be defined as.

$$MAE = \frac{\sum_{i=1}^{n} |Predicted_i - Observed_i|}{n}$$

Where n is the number of observations.

4.1.4 Root-Mean-Square Error (RMSE)

Root-Mean-Square Error is a measure of deference between the predicted value of a model and the actual observation. It can be written as.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Predicted_i - Observed_i)^2}{n}}$$

Where n is the number of observations.

4.2 Enforcing Isolation and Lockdown – Model Evaluation

Table 3 YOLO v5 Prediction Accuracyshows the mAP@0.5 of action detection on the test set of Okutama-Action. mAP@0.5 is commonly used as an evaluation metric for object detection. Where we take the IOU of 0.5 as the threshold of positive object recognition.

Model	Image Size	mAP @0.5 (%)
YOLO v5m	640 x 640	17.38
YOLO v5l	640 x 640	**20.10**

Table 3 YOLO v5 Prediction Accuracy

Figure 15 Per-class Average Precision for the modelsshows the results for each class for the models trained on the images carted from the video frames. We can see that YOLO v5l model works better in all the classes. Also as mentioned in the original research the images that were taken at the 45 degrees angle has better object detection accuracy. Therefore, when the model is used in the real world it is advisable to use a camera angle closer to 45 degrees.

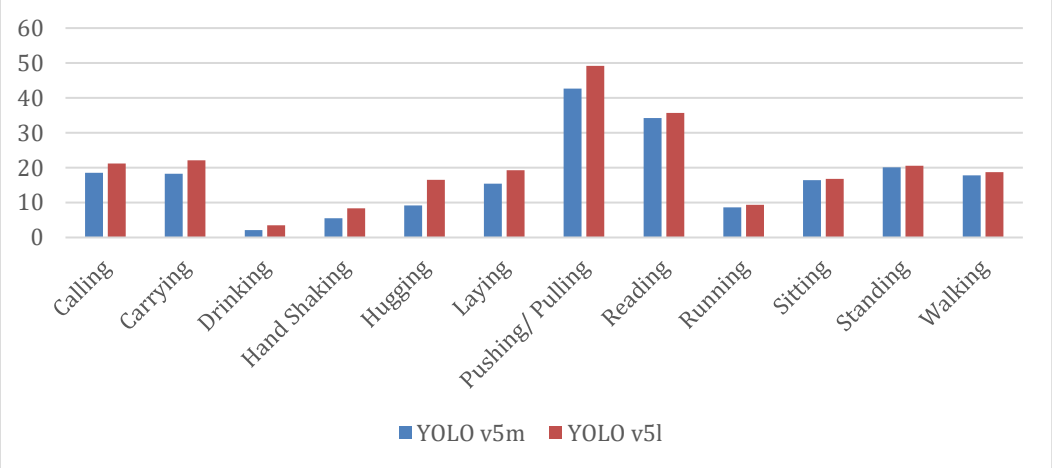

Figure 15 Per-class Average Precision for the models

When we compare the results of the YOLO v5 models with the SSD models presented in the original dataset we can see our YOLO v5l model outperform the SSD model in a smaller margin.

Model	Image Size	mAP @0.5 (%)
SSD	512x512	15.39
SSD	960x540	18.80
YOLO v5m	640 x 640	17.38
YOLO v5l	640 x 640	**20.10**

4.2 Detecting Highly Dense Crowd in Public Areas – Model Evaluation

To evaluate the model developed for crowd density prediction we have used UCF-QNRF dataset. The mean absolute error and mean squared error of detected number of people are give in the Table 4 Crowd Density Prediction Performance

Dataset	MAE	MSE
UCF-QNRF	89.43	156

Table 4 Crowd Density Prediction Performance

Table 5 Crowd Density Prediction Model Comparisonis a comparison between our results with the SFANet model in the (Zhu, et al., 2019) and Bayesian and Bayesian+ models presented by (Ma, et al., 2019). In comparison to the other models our model has a slight improvement to the Bayesian+ model.

Model	MAE	MSE
Bayesian	92.9	163.0
Bayesian+	88.7	154.8
ZFANet	100.8	174.5
ZFANet + Bayesian Loss	**87.2**	**151.3**

Table 5 Crowd Density Prediction Model Comparison

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this research we proposed two methods for crowd behaviour monitoring using arial action detection and automated crowd density predictions. The models that have been proposed has some improvements over the existing state of the art methods. Our YOLO v5l model outperformed the SSD model proposed in the literature by a considerable margin.

However further research is needed to optimise the methods that have been proposed to make them perform in the real time with a limited hardware footprint. In addition to that monitoring distance between people is also an important research area that can be explored.

REFERENCES

Barekatain, M. et al., 2017. *Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection.* s.l., The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Bharti, Y., Saharan, R. & Saxena, A., 2019. *Counting the Number of People in Crowd as a Part of Automatic Crowd Monitoring: A Combined Approach.* Singapore, Information and Communication Technology for Intelligent Systems.

Bochkovskiy, A., Wang, . C.-Y. & Yua, H., 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv.*

Boulos, M. et al., 2011. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health. *Trends, OGC standards and application examples. Int. J. Health Geogr.*, Volume 10.

Chackravarthy, S., Schmitt, S. & Yang, L., 2018. *Intelligent Crime Anomaly Detection in Smart Cities Using Deep Learning.* Philadelphia, PA, USA, IEEE 4th International Conference on Collaboration and Internet Computing (CIC).

Dwivedi, P., 2020. *YOLOv5 compared to Faster RCNN. Who wins?.* [Online]
Available at: https://towardsdatascience.com/yolov5-compared-to-faster-rcnn-who-wins-a771cd6c9fb4
[Accessed September 2021].

Gao, M. et al., 2019. *Violent crowd behavior detection using deep learning and compressive sensing.* Nanchang, China, Chinese Control And Decision Conference (CCDC).

Girshick, R., 2015. Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV),* pp. 1440-1448.

Idrees, H., Saleemi, I., Seibert, C. & Shah, M., 2013. *Multi-source Multi-scale counting in extremely dense crowd images.* Portland, OR, IEEE Conference on Computer Vision and Pattern Recognition.

Idrees, H. et al., 2018. *Composition loss for counting, density map estimation and localization in dense crowds.* Munich, Germany, European Conference on Computer Vision (ECCV).

Ilyas, N., Shahzad, A. & Kim, K., 2020. Convolutional-Neural Network-Based Image Crowd Counting: Review, Categorization, Analysis, and Performance Evaluation. *Sensors,* 20(1).

Jocher , G., Stoken, A. & Borovec, J., 2021. *ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models.* [Online]

Available at: https://github.com/ultralytics/yolov5
[Accessed Sept 2021].
Khan, A. et al., 2020. Crowd Monitoring and Localization Using Deep Convolutional Neural Network: A Review. *Applied Sciences*, 10(14).
Khan, S. et al., 2019. *Disam: Density Independent and Scale Aware Model for Crowd Counting and Localization.* Taipei, Taiwan, IEEE International Conference on Image Processing (ICIP).
Lahiri, S., Jyoti, N., Pyati, S. & Dewan, J., 2018. *Abnormal Crowd Behavior Detection Using Image Processing.* Pune, India, Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).
Lian, D. et al., 2019. *Density map regression guided detection network for rgb-d crowd counting and localization.* Long Beach, CA, USA, IEEE Conference on Computer Vision and Pattern Recognition.
Liu, W. et al., 2016. SSD: Single Shot Multibox Detector. *European conference on computer vision*, pp. 21-37.
Lv, Y. et al., 2014. Traffic flow prediction with big data: A deep learning approach.. *IEEE Trans. Intell. Transp. Syst.*
Martani, C. et al., 2017. Pedestrian monitoring techniques for crowd-flow. *P. I. Civil Eng-Eng. Su.*
Ma, Z., Wei, X., Hong, X. & Gong, Y., 2019. Bayesian Loss for Crowd Count Estimation with Point Supervision. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6142-6151.
Mehta, K. & Valloli, V. K., 2019. W-Net: Reinforced U-Net for Density Map Estimation. *ArXiv.*
Mohamed, S. & Parvez, M., 2019. *Crowd Modeling Based Auto Activated Barriers for Management of Pilgrims in Mataf.* Aswan, Egypt, International Conference on Innovative Trends in Computer Engineering (ITCE).
Nasser, N. et al., 2017. *An expert crowd monitoring and management framework for Hajj*. Rabat, Morocco, International Conference on Wireless Networks and Mobile Communications (WINCOM).
Pu, S., Song, T., Zhang, Y. & Xie, D., 2017. *Estimation of crowd density in surveillance scenes based on deep convolutional neural network.* s.l., Procedia Comput. Sci.

Rabaud, V. & Belongie, S., 2006. Counting crowded moving objects. *In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 1, pp. 705-711.

Rao, A., Gubbi, J. & Palaniswami, M., 2016. Anomalous Crowd Event Analysis Using Isometric Mapping. *Advances in Signal Processing and Intelligent Recognition Systems.*

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788.

Redmon, J. & Farhadi, A., 2018. YOLOv3: An Incremental Improvement. *CoRR.*

Ren, S., He, K., Girshick, R. B. & Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 91-99.

Rezaei, M. & Azarmi, M., 2020. DeepSOCIAL: Social Distancing Monitoring and Infection Risk Assessment in COVID-19 Pandemic. *Applied Sciences*, 10(21).

Rodriguez, M., Laptev, I., Sivic, J. & Audibert, J., 2011. *Density-aware person detection and tracking in crowds.* Barcelona, Spain, International Conference on Computer Vision.

Rohit, K., Mistree, K. & Lavji, J., 2017. *A review on abnormal crowd behavior detection.* Coimbatore, India, International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).

Sam, D. B., Peri, S., Kamath, A. & Babu, R., 2019. *Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection.* s.l., arXiv.

Simonyan, K. & Zisserman, A., 2015. *Very deep convolutional networks for large-scale image recognition.* s.l., ICLR.

Sindagi, V. A. & Yasarla, R. a. P. M., 2019. *Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method.* s.l., IEEE International Conference on Computer Vision.

Srivastava, S., Divekar, A. V. & Ani, C., 2021. Comparative analysis of deep learning image detection algorithms. *Journal of Big Data.*

Wang, Q., Gao, J., Lin, W. & Li, X., 2020. NWPU-crowd: A large-scale benchmark for crowd counting. *arXiv.*

Wang, T. et al., 2020. *Abnormal event detection via the analysis of multi-frame optical flow information.* s.l., Front. Comput. Sci..

Xu, C. et al., 2019. *Learn to Scale: Generating Multipolar Normalized Density Maps for Crowd Counting.* Seoul, Korea, IEEE International Conference on Computer Vision.

Xue, Y., Liu, S., Li, Y. & Qian, X., 2020. *Crowd Scene Analysis by Output Encoding.* s.l., arXiv.

Xu, M. et al., 2019. Depth information guided crowd counting for complex crowd scenes. *Pattern Recognit. Lett.*.

Xu, R., Lin, H., Lu, K. & Cao, L., 2021. A Forest Fire Detection System Based on Ensemble Learning. *Forest,* Volume 12.

Yang, M. et al., 2018. *Cluster-based Crowd Movement Behavior Detection.* Canberra, Australia, Digital Image Computing: Techniques and Applications (DICTA).

Yimin, D., Fudong, C., Jinping, L. & Wei, C., 2019. *Abnormal Behavior Detection Based on Optical Flow Trajectory of Human Joint Points.* Nanchang, China, Chinese Control And Decision Conference (CCDC).

Yimin, D., Fudong, C., Jinping, L. & Wei, C., 2019. *Abnormal Behavior Detection Based on Optical Flow Trajectory of Human Joint Points.* Nanchang, China, Chinese Control And Decision Conference (CCDC).

Zhang, C., Li, H., Wang, X. & Yang, X., 2015. *Cross-scene crowd counting via deep convolutional neural networks.* Boston, MA, USA, IEEE Conference on Computer Vision and Pattern Recognition.

Zhang, Y. et al., 2016. *Single-image crowd counting via multi-column convolutional.* Las Vegas, NV, USA, IEEE Conference on Computer Vision and Pattern Recognition.

Zhu, L. et al., 2019. Dual Path Multi-Scale Fusion Networks with Attention for Crowd Counting. *CoRR.*

Zhu, L. et al., 2019. Dual Path Multi-Scale Fusion Networks with Attention for Crowd Counting. *arXiv.*

Zhu, P. et al., 2020. Vision Meets Drones: Past, Present and Future. *arXiv preprint.*