

Crowd Analysis and Localization Using Deep Convolutional Neural Network

Introduction

Crowd is same or different set of people arranged in one group and motivated by common goals. There are two types of crowd namely structured crowd and unstructured crowd. In the former, the direction of the movement is towards a common point and people are not in scattered form while in the later type the direction of the people is not towards a common point and they are usually in scattered form.

Crowd or mass gatherings at various venues such as entertainment events, airports, hospitals, sports stadiums, theme parks are faced by the individuals on almost a daily basis. The activities are quite diverse and range from social and cultural to religion. Unlike social and sports related events, the crowd situations experienced by the people on important religious events like Hajj and Umrah may not be possible to avoid. It is therefore important to have an intelligent Crowd Monitoring System (CMS) to ensure the safety of the public, maintain high throughput of pedestrian flow to prevent stampedes, provide better emergency services in case of crowd-related emergencies and to optimize the resources for providing good accessibility by avoiding congestion.

In general perspectives, crowd management, monitoring, and analytics have potential for a number of applications. These include but are not limited to the safety domain, emergency

services, traffic flow and management in private and public spaces, people counting and analyzing group behaviors and similarly swarm-based applications. Such integrity of applications provides a natural demand for research and developments in managing and analyzing crowd and the behavior of individuals in crowd for groups analysis, counting and summarizing, density and prediction, flow analysis, specific behavior prediction, and mass tracking. In general, the group detection and density estimation have proven useful for corresponding steps of intelligent analytics and several applications.



Challenges

Efficient crowd monitoring and management contributes to various applications having further potential for computer vision (CV) paradigm; however, crowd management in real time is far from being solved, particularly in the wild conditions and still facing many open challenges.

The task of crowd management is still open for research community. Several factors contribute to a robust real time CMS and also affect the performance of an accurate CMS.

We summarize some of these challenges as follows:

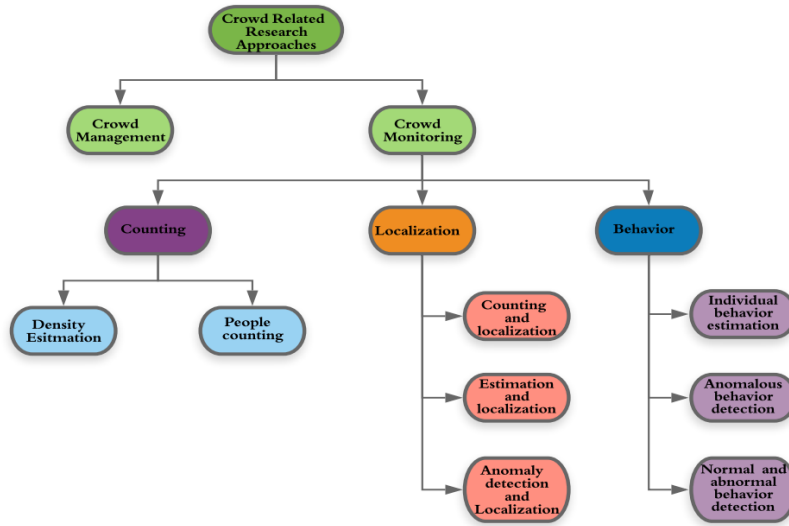
- 1) When two or more than two objects come close to each other and as a result merge, in such scenarios, it is hard to recognize each object individually. Consequently, monitoring and measuring accuracy of the system becomes difficult.
- 2) A non-uniform sort of arrangement of various objects which are close to each other is faced by these systems. This arrangement is called clutter. Clutter is closely related to image noise which makes recognition and monitoring more challenging.
- 3) Irregular object distribution is another serious problem faced by CMS. When density distribution in a video or image is varying, the condition is called irregular image distribution. Crowd monitoring in irregular object distribution is challenging.
- 4) Another main problem faced in real time crowd monitoring systems is aspect ratio. In real time scenarios, normally a camera is attached to a drone which captures videos and images of the crowd under observation. In order to address the aspect ratio problem, the drone is flown at some specific height from the ground surface and installation of the camera is done such that the camera captures the top view of the crowd under observation. The top view results in properly addressing the aforementioned problem of aspect ratio.
- 5) The unavailability of a public dataset is one major problem towards the development of an efficient and mature real time CMS. Although datasets are available for counting purposes, but very few datasets are available for behavior analysis and localization research.

Datasets

Various datasets containing crowd videos and images are publicly available and are being used to validate the experimental results. Some of the publicly available datasets along with its description are shown below:

Taxonomy Level

Database	Year	Task	# of Images	Head Count	Source Obtained
Mecca [45]	2020	crowd monitoring	–	–	surveillance
Kumbh Mela [46]	2020	crowd monitoring	6k	–	surveillance
NWPU-Crowd [47]	2020	crowd counting and localization	5109	2133,375	surveillance and Internet
BRT [48]	2018	crowd monitoring	1280	16795	surveillance
UCF-QNRF [49]	2018	counting in a crowd and localization	1525	1,251,642	surveillance
Shanghai Tech [50]	2016	cross scene crowd counting	482	241,677	surveillance and Internet
WorldExpo'10 [44]	2015	counting in a crowd	3980	199,923	surveillance
WWW [51]	2015	crowd management	10000	8 million	Internet
UCF_CC_50 [43]	2013	Density estimation	50	63,974	surveillance
The Mall [52]	2012	counting in a crowd	2000	62,325	surveillance
PETS [53]	2009	counting in a crowd	8	4000	surveillance
UCSD [54]	2008	counting in a crowd	2000	49,885	Internet



The entire taxonomy level of crowd monitoring has been shown in the form of flow chart. Basically the crowd related research approaches have been categorized into two domains based on the literature review namely crowd management and crowd monitoring. Then we have made categories of crowd monitoring i.e., counting, localization and behavior.

Approaches

Several methods and techniques related to crowd monitoring and localization had been introduced over the past few years. the table below show some of these methods with brief description and conclusion about the method and the result

Process	Frameworks/Methods	Performance	Conclusion/Result
Crowd behavior analysis(Behavior)	Spatio-temporal model	Accuracy 98% and 88%	Visual descriptors have extracted and considered for both individual and interactive behaviors
Crowd evacuation (Evacuation behavior)	Legion Evac software	Correlation scores were positive	Reduced evacuation time
Detection of anomaly (Normal and abnormal)	Optical flow and Horn Schunck algorithm	Computation of distance between centroids	A novel approach of abnormal event detection has proposed
Crowd behavior detection (Identify behavior)	Spatio-Temporal Texture model	The STT method demonstrates comparable results of Spatio-temporal Compositions (STC) and Inference by Composition (IBC)	Crowd anomaly detection framework was introduced
Crowd behavior detection(Behavior)	Approximate median filter and foreground segmentation algorithm	Lower false rate	A robust unsupervised abnormal crowd behavior detection has achieved
Violent behavior detection	Hybrid random matrix (HRM) and deep neural network	Accuracy 90.17% and 91.61%	Combining compressive sensing and deep learning to identify violent crowd behavior
Crowd behavior detection	Holistic approach	The performance of this methods yields better results	Holistic approach for abnormal crowd behavior detection has proposed
Crowd behavior detection (Real time)	Scale-invariant feature transform (SIFT)	Accuracy 95%	The combination of SIFT and genetic algorithm has achieved better simulation results
Crowd behavior monitoring (Event detection)	Fixed-width clustering algorithm and YOLO	Accuracy is between 80%-95.7%	The approach has a superior performance on six videos
Abnormal behavior detection(Abnormality)	Optical flow method and SVM	87.4% accuracy	Higher detection rate for anomaly

Process	Frameworks/Methods	Performance	Conclusion/Result
Crowd monitoring (Counting)	Combination of crowd size estimation and counting	Accuracy 90%	Classification with MSE 0.0081
Crowd monitoring (Counting)	Neural network and regression trees using fisheye camera	BPNN provides the best estimation	BPNN can deal 9 frames in second
Crowd monitoring (Estimation)	ICrowd framework was designed on a three-layer approach, device layer, middleware layer and the application layer	No experimental results	Capable of location updates
Density estimation (Estimation)	Airborne camera systems, support vector machine and Gabor filter	Depends on good training samples and similar images	Gabor filter plays a prominent role in real scenes of images
Crowd monitoring (Controlling crowd movement)	Information management module and decision support system	Expert system module performs well	Real-time crowd density measurements and communications during hajj
Crowd Counting (Normal/abnormal event)	Median filter and Kalman filter	Accuracy 95.5%	Robust smart surveillance system
Estimation and localization(Density map)	Deep CNN networks	Specificity 75.8%	Decrease error rate
Detection and localization (Anomaly detection)	Global and local descriptors, with two classifiers were proposed	Accuracy 99.6%	Achieved good and competing methods with low computational complexity
Counting and Localization (Human heads)	CNN	DISAM outperforms for UCSD and WorldExpo'10 datasets with the lowest MAE of 1.01 and 8.65, respectively	Reduction of classification time
Counting and localization	SD-CNN Model	Average Precision and average recall rate 73.58 and 71.68	Reduction of classification time and improvement in detection accuracy
Crowd monitoring(Behavior)	EHCAF	99.1% accuracy and FNR of 2.8%,	Highly accurate and low FNR
Crowd monitoring (Behavior detection)	Isometric mapping (ISOMAP)	Reduced feature space	Reduction of computation time

Counting of crowd provides an estimate about the number of people or certain objects.

Counting does not provide any information about the location. Density maps are computed at different levels and also provide very weak information about a person's location. On the other hand, localization provides accurate information about the location. However, due to sparse nature, it is comparatively a difficult task. Therefore, the best way is to handle all the three tasks simultaneously, employing the fact that each case is related to the other.

Crowd Localization

Localization of crowds in crowded images received less attention from the research community. With localization information, one can figure out how people are distributed in the area, which is very important for crowd managers. Information about localization can be used to detect and monitor an individual in dense crowds. In order to identify the location of head in an image many proposed methods was developed and in the next section we will consider the main approaches with more interesting ideas.










Leaderboard for Localization

The NWPU-Crowd Dataset:

The NWPU-Crowd Dataset is constructed by Wang et al., from NWPU. It is a large-scale congested crowd counting dataset that consists of 5,109 images crawled from the Internet, elaborately annotating

Leaderboard

Leaderboard

Download History Result											
Name	Extra Data 	O_F1-m(large) 	O_Pre(large) 	O_Rec(large) 	Avg.Rec[B](large) 	Runtime 	Device	Code			
DCST	no	0.775	0.822	0.734	0.609	0	RTX3090	no			
IIM(HRNet)	no	0.762	0.813	0.717	0.613	0	-	yes			
IIM(HRNet) + point an...	no	0.760	0.829	0.702	0.491	0	-	yes			
FIDTM	no	0.755	0.798	0.717	0.475	0.2	V100	yes			
DBMI	no	0.748	0.757	0.739	0.558	0	2080Ti	no			
IIM(VGG16)	no	0.732	0.779	0.692	0.587	0	0	yes			
D2CNet	no	0.700	0.741	0.662	0.583	0.2	1080Ti	no			
TopoCount	no	0.692	0.683	0.701	0.633	0	V100	yes			
SCALNet	no	0.691	0.692	0.690	0.442	0.025	1080 Ti	yes			
GeneralizedLoss	no	0.660	0.800	0.562	0.485	0	2080 Ti	yes			
PDRNet	no	0.653	0.675	0.633	0.470	0.1	P100	no			
Crowd-SDNet	no	0.637	0.651	0.624	0.551	0	0	yes			
LC-Net	no	0.628	0.656	0.603	0.342	0	V100	no			
PSCAL	no	0.623	0.650	0.599	0.423	0	2080ti	no			
AutoScale_localization	no	0.620	0.674	0.574	0.484	0	2080Ti	yes			
RAZ_loc_branch_cb_o...	no	0.599	0.666	0.543	0.424	0	-	yes			
TinyFaces_cb_official	no	0.567	0.529	0.611	0.598	0	-	yes			
VGG+GPR_cb_official	no	0.525	0.558	0.496	0.374	0	-	yes			

2,133,375 instances.

O_F1-m(large):F1-measure (large head area) on overall testing images.

O_Pre(large): Precision (large head area) on overall testing images.

O_Rec(large): Recall (large head area) on overall testing images.

Avg.Rec[B]: Average Recall on Box level.

Revisiting crowd behavior analysis through deep learning

A large number of publications have addressed crowd behavior analysis using Deep Learning techniques in their pipelines. Nevertheless, most of these works are sparse and difficult to compare.

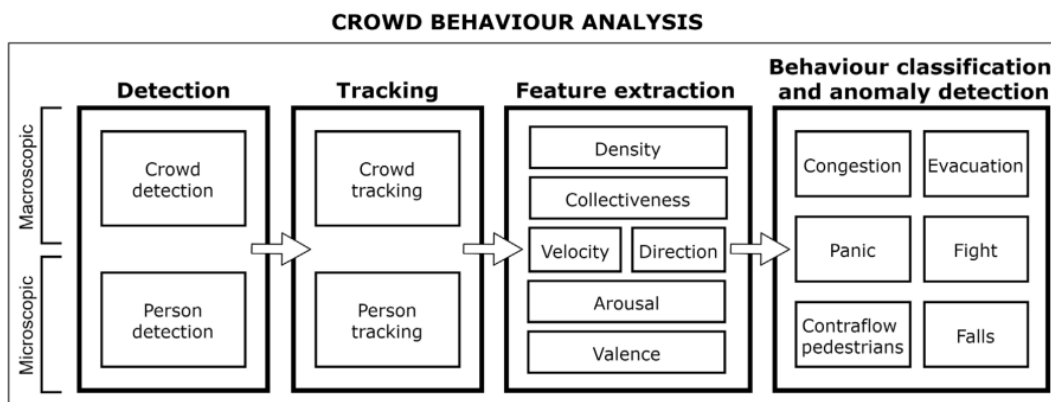
This is particularly critical when developing good solutions, since it is difficult to gather the previously developed knowledge.

This dispersion is due to three different factors:

- (1) There is a lack of consensus on what crowd behavior analysis is.
- (2) the sub-tasks that constitute crowd behavior analysis are not clearly determined
- (3) there is not an available taxonomy on how these sub-tasks should be organized and addressed.

Crowd behavior analysis pipeline

Four main stages of the crowd behavior analysis pipeline:



1. Detection stage: Its objective is to localize the position of individuals and crowds in each frame.

2. Tracking stage: It aims at uniquely identifying the specific persons and crowd trajectories across a sequence of consecutive frames. Frequently, the dominant flows of movement in the crowd are also determined.

3. Feature extraction stage: It computes a set of metrics that describe the dynamics, topological structure and affective state of the crowd. These metrics can be monitored over time, and computed both at the individual level, when the different subjects are studied independently, or at crowd level, when a mass of pedestrians is considered as a unique entity.

4. Crowd behavior classification and anomaly detection stage: On the basis of extracted features, this last stage aims at recognizing particular behaviors and/or abnormal events in video sequences. There are two main approaches for this stage, depending on the type of learning paradigm employed, supervised or unsupervised.

Behavior classification encloses the works that confront the task in a supervised manner. These works previously define a set of behaviors (e.g. people talking, walking together, greeting each other, fighting,

snatching, etc.), and train classification models over them. On the other hand, anomalous behavior detection tries to identify abnormal patterns in the crowd, a priori unknown.

Crowd behavior classification and anomaly detection

Extracted features have to be summarized in order to obtain meaningful information about the behavior of the crowd. There are two main approaches in this stage: crowd behavior classification, where models are trained in a supervised manner over the extracted features; and crowd anomaly detection, when learning is performed in an unsupervised way.

Monitoring crowd features over time opens the door to the detection of abnormal behaviors in crowds, since sudden changes in these features are indicative of strange patterns. For example, a sudden drift in crowd speed values is usually an indicator of alert; unwanted congestion can be characterized in terms of lower speed and higher density; and extreme values of valence and arousal can lead to violent situations between groups of people.

When solving the problem from an anomaly detection perspective, another possible organization emerges from the source of the anomaly itself. As we stated before, the nature of the anomaly may be diverse, and thus the approach to solve the problem may differ slightly. We have identified five types of anomalies.

- **Anomalous position:** The source of this anomaly comes from an atypical position of an object in the scene. This kind of anomaly occurs, e.g., when a non-authorized individual enters a restricted area, or a pedestrian is detected on a dangerous zone. It is considered the easiest kind of anomaly to detect, since it usually involves just a pedestrian detection stage, combined with bounding box overlapping computation.
- **Anomalous movement:** In this case, the anomalous pattern is produced by an unexpected trajectory of one individual or group in the scene. Two different sources of irregularity can be found: speed, when someone moves faster or slower than his/her surroundings; and direction, when predominant flows exist and the movement of an individual deviates from these trends.
- **Anomalous appearance:** This abnormality occurs when a non-recognized object enters the scene. A typical example of this anomaly is the presence of a vehicle in a pedestrian path.
- **Anomalous action:** It is the most difficult anomaly to be identified. It involves the understanding of the usual behavioral patterns of the individuals in the scene, and the detection of non-common ones.

In practice, the detection of these types of anomalies is often combined.

For example, anomalous movement and anomalous appearance are usually tackled together. A clear example of this combination is present in the UCSD Pedestrian dataset, which is the most employed one in the literature. In this dataset, two types of anomalies are present: people



(a) Example of anomaly in the UCSD Pedestrian dataset [30]. Cars are not allowed in the pedestrian path, so the behaviour is anomalous in terms of appearance.



(b) Example of anomaly in the UMN dataset [31]. In this case, an unstructured crowd walks peacefully in the scene, which is considered to be a normal behaviour. Suddenly, every person in the scene starts running, and the anomalous motion pattern begins.



(c) Example of anomaly in the CUHK Avenue dataset [32]. People in this dataset move usually in a parallel direction to the camera plane, and thus a perpendicular direction is considered an anomalous movement pattern.



(d) Example of anomaly in the BOSS dataset [33]. The man steals the woman's phone while she is talking on it. This is an example of anomalous action.

walking in strange directions (motion) and presence of unauthorized vehicles (appearance). Some examples of different types of anomalies are illustrated below:

Datasets for crowd anomaly detection

Due to the complex nature of the crowd behavior anomaly detection problem, many different datasets that focus on solving diverse tasks are publicly available.

In this section, we will categorize these datasets depending on the main task tackled by each one. first, datasets whose main task is motion anomaly detection will be described. second will

Table 1
Public datasets for crowd anomaly detection.

Dataset	# Frames			# Abnormal events	Anomaly type	Description of anomalies
	Total	Training	Testing			
UCSD Peds 1	14000	6800	7200	40	Motion + appearance	Strange directions, speeds, forbidden objects (bikes, cars)
UCSD Peds 2	4560	2550	2010	12	Motion + appearance	Strange directions, speeds, forbidden objects (bikes, cars)
CUHK Avenue	30652	15328	15324	47	Motion + appearance	Abnormal directions, speeds and unexpected objects
ShanghaiTech Campus	317398	274515	42883	130	Motion + appearance	Abnormal directions or speeds, loitering
UMN Dataset	7725	-	-	3	Motion	Whole crowd suddenly changing speed and direction
BEHAVE Interactions	225019	-	-	14 ^a	Action	Fights and chases
CAVIAR	26402	-	-	11 ^b	Action	Abandoned objects, fights, falls
BOSS	48624	-	-	10	Action	Fights, stealing, people falling
UT Interactions	41373	-	-	48	Action	Shake hands push, point kick, punch, hug
UCF Crime	13M	-	-	-	Action	Uncivil behaviours ^c
Películas	4991	-	-	100 ^d	Action	Violence
Hockey Fights	41056	-	-	500 ^d	Action	Violence

^aOnly first video labelled.

^bLabelled for behaviour classification.

^c13 different types of uncivil behaviours reported together with normal videos.

^dShort videos of fights and no-fights.

focus on datasets for action anomaly detection.

1) Datasets for motion anomaly detection

Datasets in this subsection are designed to present different anomalous motion patterns. These anomalies are usually defined by speeds or trajectories that deviate from expected normal motion flows in the scene. The presence of non-authorized elements is also a common trend in these datasets (e.g. vehicles or bicycles on pedestrian paths), and thus abnormal motion and appearance are usually considered together.

The datasets most widely used for motion anomaly detection are the following:

1. UCSD Pedestrian dataset:

UCSD anomaly detection datasets¹ are the most popular in the literature.

There are two different sets of videos, called **Peds1** and **Peds2**. Peds1 contains **34 training and 36 testing video sequences**, and Peds2, **16 training and 12 testing sequences**. Each clip is approximately 200 frames (20 s) long, with a 158 × 238 resolution. **The main difference between Peds1 and Peds2 is the direction of the moving pedestrians**. In Peds1, people walk towards and away from the camera, while in Peds2 individuals move in parallel to the camera plane. No anomalies are present in training videos, which are intended to show

what is considered to be a normal behavior. In testing videos, various abnormal events occur. The frame is considered to be anomalous if there is a non-pedestrian element in the scene (e.g. bikers, skaters, etc.) or if a pedestrian shows an abnormal motion pattern (e.g. somebody running, changing its direction abruptly, and so on). In total, approximately 3400 frames contain anomalies, and 5500 frames are normal.

2. CUHK Avenue dataset:

The CUHK Avenue dataset² **contains 16 training and 21 testing videos, with 15328 frames for training and 15324 testing.** Again, normal samples are formed by people walking in parallel to the camera plane; people moving in other directions, with strange motion patterns or moving vehicles, are considered to be anomalous.

3. UMN dataset:

The UMN dataset³ is a **synthetic** dataset composed by three different scenes, with a total length of **4 min and 17 s (7725 frames)**. In each video, an unstructured crowd is walking in the scene, and suddenly everyone starts running, moment that is marked as an anomaly. The objective on this dataset is to accurately detect the change in the movement of the crowd. It can be seen both as motion and behavior anomaly detection, since the source of anomaly is produced by a sudden change in the speed and direction of people in the scene, but the motion pattern is completely unstructured, in contrast with the structured nature of movement in the previous datasets.

4. ShanghaiTech Campus dataset:

The ShanghaiTech Campus dataset⁴ is divided into **330 videos for training and 107 videos for testing**, taken in 13 different scenarios across the campus.

Anomalous events are produced by strange objects in the scene, pedestrians moving at anomalous speed (running or loitering), and moving in unexpected directions.

2) Datasets for action anomaly detection

The main task of the datasets presented in this subsection is to identify when a person in the scene presents an abnormal behavior. Usually, behaviors considered as abnormal are uncivil behaviors such as stealing, fighting, snatching, etc.

The most relevant datasets for behavior anomaly detection are the following:

1. BEHAVE dataset:

The BEHAVE Interactions dataset⁵ contains **4 video sequences, of a total length of 2 h**. Anomalies are mainly produced by fighting. Only the first sequence is fully annotated. It is divided into 8 fragments, and each fragment is ground truth at the frame-level.

2. CAVIAR dataset:

The CAVIAR Test Case Scenarios dataset⁶ is a **set of videos taken from two different scenes**: the entrance hall of a lab building and a hallway in a shopping center. There are several video sequences for each scenario. In each recording, a person or group of people performs a different action. Most of the anomalies in this dataset are provoked by fighting between pedestrians

3. BOSS dataset:

The BOSS dataset⁷ is a collection of **19 scenes taken inside a moving train**, in which groups of people, ranging from single individuals to crowds of more than 10 pedestrians, interact in different manners, both normally and abnormally. For every scene, the action is recorded from different perspectives, using several cameras. Fights, people falling and group panic are examples of anomalies in this dataset.

4. UT Interactions dataset:

The UT Interactions dataset⁸ is a **collection of 20 videos around 1 min each**, presenting six different classes of **Human–Human interactions: shake-hands, point, hug, push, kick and punch**. All the videos contain several interactions, along with distractor pedestrians.

The aim is to correctly detect and classify the type of interaction between subjects. Ground-truth labels for these interactions are provided, including time intervals and bounding boxes.

5. UCF-Crime dataset:

The UCF-Crime dataset⁹ was produced in 2018, and **contains 1900 videos, 950 of normal events and 950 of abnormal ones, divided into 13 classes**: abuse, arrest, arson, assault, road accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting and vandalism.

Two main tasks are proposed: detection and localization of generic anomalies, at a first stage; and specific anomaly classification, at second stage. This dataset is especially relevant due to its large size (more than 13 million samples) and novelty.

Summary

We will divide our work into three main phases: (up to 4 if needed)

1) Counting Phase: (Part of POC's)

As part of realization of POC's phase, we will start by tackling the crowd counting problem due to its importance and related concepts to our main objectives.

Mainly we will focus on the following :

Crowd counting with deep learning by generating density map:

<https://datduyng.github.io/2019/05/12/crowd-counting-and-localization.html>

CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes

[1802.10062.pdf \(arxiv.org\)](https://arxiv.org/pdf/1802.10062.pdf)

good to know: Improving Object Counting with Heatmap Regulation:

<https://arxiv.org/pdf/1803.05494.pdf>

2) Localization:

Our objective in phase 2 is to apply object localization, given frame/image the model should locate all humans (can be used as a black box for other tasks)

Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization

[Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization \(thevcf.com\)](https://thevcf.com/density-map-regression-guided-detection-network-for-rgb-d-crowd-counting-and-localization/)

good to know : Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds

[1808.01050v1.pdf \(arxiv.org\)](https://arxiv.org/pdf/1808.01050v1.pdf)

3) Behavior Analysis:

Depending on the results and knowledge gain from previous phases, we will select the best approach for behavior analysis.

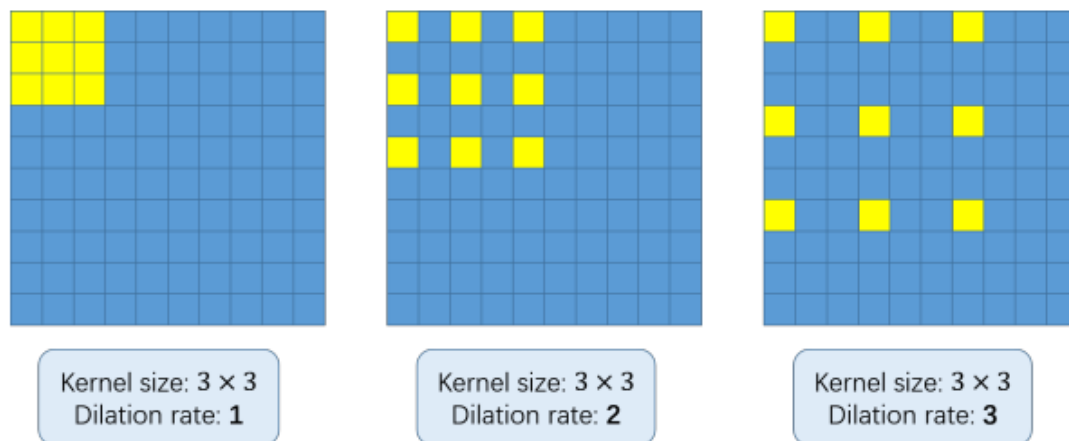
CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes

The fundamental idea of the proposed design is to deploy a deeper CNN for capturing high-level features with larger receptive fields and generating high-quality density maps without brutally expanding network complexity.

we choose VGG16 as the front-end of CSRNet because of its strong transfer learning ability and its flexible architecture for easily concatenating the back-end for density map generation.

In this paper, we first remove the classification part of VGG16 (fully-connected layers) and build the proposed CSRNet with convolutional layers in VGG-16.

The output size of this front-end network is 1/8 of the original input size. If we continue to stack more convolutional layers and pooling layers (basic components in VGG-16), output size would be further shrunk, and it is hard to generate high-quality density maps. we try to deploy dilated



convolutional layers as the back-end for extracting deeper information of saliency as well as maintaining the output resolution.

Dilated convolution:

One of the critical components of our design is the dilated convolutional layer.

Dilated convolutional layers have been demonstrated in segmentation tasks with significant improvement of accuracy and it is a good alternative of pooling layer.

which uses sparse kernels (as shown in Fig below) to alternate the pooling and convolutional layer. This character enlarges the receptive field without increasing the number of parameters or the amount of computation (e.g., adding more convolutional layers can make larger receptive fields but introduce more operations).

Network Configuration

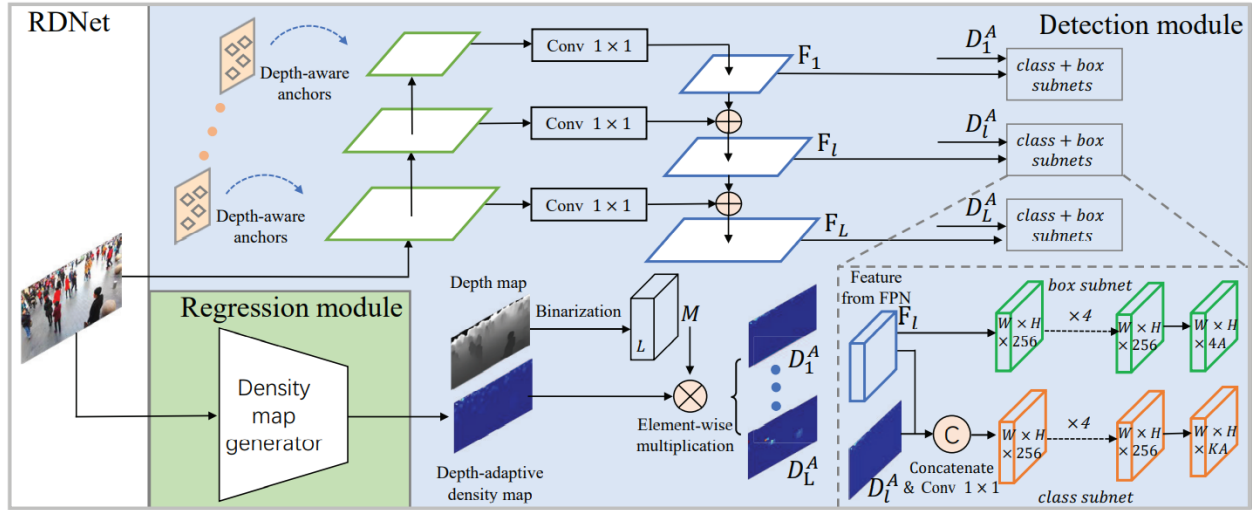
Configurations of CSRNet			
A	B	C	D
input(unfixed-resolution color image)			
front-end (fine-tuned from VGG-16)			
conv3-64-1			
conv3-64-1			
max-pooling			
conv3-128-1			
conv3-128-1			
max-pooling			
conv3-256-1			
conv3-256-1			
conv3-256-1			
max-pooling			
conv3-512-1			
conv3-512-1			
conv3-512-1			
back-end (four different configurations)			
conv3-512-1	conv3-512-2	conv3-512-2	conv3-512-4
conv3-512-1	conv3-512-2	conv3-512-2	conv3-512-4
conv3-512-1	conv3-512-2	conv3-512-2	conv3-512-4
conv3-256-1	conv3-256-2	conv3-256-4	conv3-256-4
conv3-128-1	conv3-128-2	conv3-128-4	conv3-128-4
conv3-64-1	conv3-64-2	conv3-64-4	conv3-64-4
conv1-1-1			

Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization

Method

The overall network architecture of our regression guided detection network (RDNet) for crowd counting is shown in Figure below.

It contains two modules: a density map regression module and a head detection module.



In the density map regression module, depth-adaptive kernel is introduced to generate high-fidelity ground-truth density map.

In the detection module, we leverage a RetinaNet for detection in view of its advantages in both speed and performance. We feed the estimated density map to the classification branch in the detection network to facilitate the classification of heads, meanwhile, the depth-aware anchor strategy is also proposed to initialize appropriate anchor, which also helps the improvement of detection performance.

Density Map Regression Module

Density map regression module takes an image as input and leverages CNN for density map estimation. The most commonly used ground-truth density map generation strategy utilizes a Gaussian with fixed bandwidth for approximating the density map(2D Gaussian kernel with fixed bandwidth).A high-fidelity ground-truth density map is desired. Actually, the sizes of heads vary significantly, even for the heads within an image. Therefore, it is desirable to design different bandwidth for different heads other than using the same bandwidth for all heads. Bounding box annotation can provide such information, but it is time consuming than point annotation and it is also hard to annotate bounding boxes for those tiny or occluded heads. Considering that depth provides information of head sizes within an image under the assumption that all heads are of the same sizes in the real world, we propose a depth-adaptive kernel for density map generation.

Detection Module

Our detection network is based on a RetinaNet because of its advantages in speed and accuracy. Specifically, RetinaNet is based on a feature pyramid network (FPN) and it contains multi-scale encoding and decoding layers. For each decoding layer, it takes features from corresponding encoding layers as well as outputs from its previous decoding layers as inputs. The detection is conducted on every scale feature map, which includes a class subnet for classifying and a box

subnet for regressing bounding boxes.

However, RetinaNet cannot be directly applied for head counting because it fails to detect small/tiny heads, meanwhile crowd counting is only with point based ground-truth annotations rather than bounding boxes. Thus, we propose to use estimated density map from regression module and depth-aware anchor to improve the robustness of RetinaNet for small/tiny heads detection and use depth to generate bounding boxes for training RetinaNet.

Density map guided classification

RetinaNet fails to detect those small/tiny heads because the class subnet fails to classify those anchor boxes as positive. However, such class subnet would benefit from density map. Density map shows the distribution of heads, and its value at each pixel is related to the probability of the pixel being a head. Therefore, we propose to feed the estimated density map into the detection network to boost the performance of small/tiny heads. RetinaNet detects heads of different scales at different decoding layers. The lower layers respond to the detection of smaller heads, and higher layers respond to the detection of larger heads. We thus propose to mask density map based on the depth map.

Then we generate a binarization matrix $M \in \mathbb{B}^{L \times H_d \times W_d}$ based on depth map, where H_d and W_d are the height and width of generated density map, respectively.

In binarization, the depth map is down sampled to the same size as the density map. For each channel l in M , the values of pixels with larger or smaller corresponding depth are set to 0's, and the values of pixels within the range are 1's. We denote this binary mask as M_l , and use it to mask our estimated density map.