Received: 6 March 2023 | Revised: 25 July 2023 | Accepted: 11 August 2023

DOI: 10.1002/aaai.12117

ARTICLE

Visual crowd analysis: Open research problems

Muhammad Asif Khan | **Hamid Menouar** | **Ridha Hamila**

Qatar Mobility Innovations Center (QMIC), Qatar University, Doha, Qatar

Correspondence
Muhammad Asif Khan, Qatar Mobility Innovations Center (QMIC), Qatar University, Doha, Qatar.
Email: muasifkn@gmail.com

Funding information
Qatar National Research Fund, Grant/Award Number: PDRA7-0606-21012

Abstract

Over the last decade, there has been a remarkable surge in interest in automated crowd monitoring within the computer vision community. Modern deep-learning approaches have made it possible to develop fully automated vision-based crowd-monitoring applications. However, despite the magnitude of the issue at hand, the significant technological advancements, and the consistent interest of the research community, there are still numerous challenges that need to be overcome. In this article, we delve into six major areas of visual crowd analysis, emphasizing the key developments in each of these areas. We outline the crucial unresolved issues that must be tackled in future works, in order to ensure that the field of automated crowd monitoring continues to progress and thrive. Several surveys related to this topic have been conducted in the past. Nonetheless, this article thoroughly examines and presents a more intuitive categorization of works, while also depicting the latest breakthroughs within the field, incorporating more recent studies carried out within the last few years in a concise manner. By carefully choosing prominent works with significant contributions in terms of novelty or performance gains, this paper presents a more comprehensive exposition of advancements in the current state-of-the-art.

INTRODUCTION

The increasing population in urban areas often leads to crowded situations in densely populated areas which pose several challenges and threats to public safety. To ensure the safety of the people, crowd management strategies require efficient crowd analysis. While manual analysis of the crowd is a tedious task, automatic crowd analysis is desired in many situations.

Automatic crowd monitoring using visual analysis is a hot topic in computer vision research (Khan, Menouar, and Hamila 2023c) with many interesting applications in city surveillance, social distancing, transportation, sports, wildlife monitoring, and so forth (Thirumalaisamy et al. 2022; Zhang et al. 2022). Over the last decade, advances in deep learning have brought new possibilities to achieve state-of-the-art performances in various visual crowd analysis problems such as crowd counting, object detection, activity recognition, anomaly detection, motion analysis, and so forth. Although numerous research efforts have achieved remarkable performance in visual crowd analysis, there remain several challenges and problems yet to address. The reason for many of the unsolved problems lies in the underlying complexity and challenges related to crowd scenes as compared to other computer vision tasks, for example, severe occlusions, clutters, scale variations, unpredictable motion patterns, complex crown behaviors, the unknown context of crowd activities, and

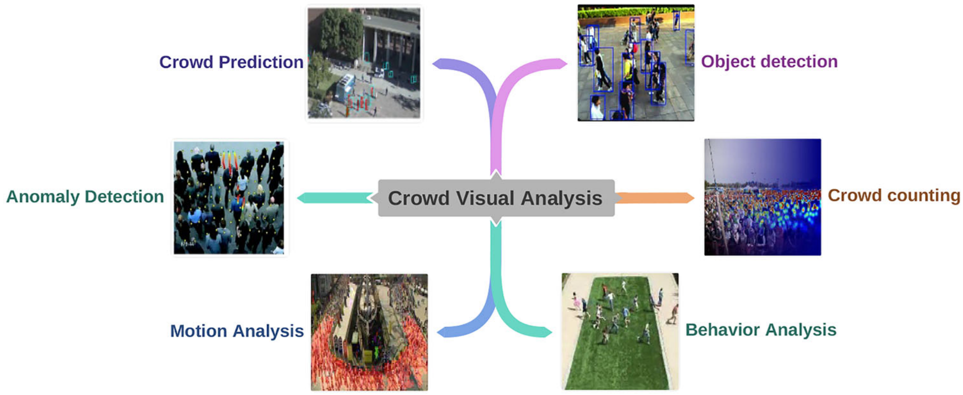

FIGURE 1 Six areas or applications of crowd visual analysis.

so forth. Thus, crowd analysis is often considered more challenging than other computer vision tasks. The complex nature of the problem impacts the development of the crowd visual analysis system and requires more sophisticated algorithms and models, collection of diverse and large-scale datasets, hardware resources for real-time performance, and system-level integration of such algorithms with cameras, sensors, and data storage systems. Algorithmic innovation and data availability are particularly of paramount importance for computer vision researchers.

To understand these challenges, we thoroughly investigated the literature on crowd analysis using computer vision and mostly using deep learning. We also studied different traditional approaches, to understand how the innovations progressed over time and how they created an impact. We greatly focused on research works published in major scientific venues and chose those works that have significantly contributed to the body of knowledge in terms of identifying a real problem, innovation in methodology, or claiming significant performance gains against previous methods. We found existing surveys categorizing works in different ways, however, when studying the common trends and distinct approaches in these works, we found a more intuitive and meaningful categorization of crowd analysis. Our approach is to categorize these works into different types of analysis tasks where each task is completely different and requires different methods to accomplish the task objective. As illustrated in Figure 1, these works on visual crowd analysis are categorized into six major areas, that is, crowd counting, object detection and tracking, motion analysis, behavior recognition, anomaly detection, and crowd prediction.

Crowd counting is to estimate the density of people or the total count in a particular geographical area. Objection detection aims to detect and localize particular objects of interest in the crowd, for example, detecting women only, detecting people holding banners or sticks, and so forth. Motion analysis refers to the collective mobility state of the crowd, for example, if the crowd is stationary/moving along with other motion statistics such as direction, speed, flux analysis, and so forth. Behavior analysis determines the collective attribute of the crowd focusing on the activities performed by crowd members to extract contextual behavior information, for example, if the crowd is calm, violent, and so forth. Anomaly detection focuses on finding unusual and abnormal events and activities at both individual and group levels. Crowd prediction refers to the prediction of proactive accumulation (assembly) or influx and efflux of people in/from a particular region that can lead to a crowd. Generally, there is a logical sequence in the implementation of these applications. For instance, the first information which might interest a user or agency monitoring a crowd is the estimated crowd density, followed by other aspects of crowd members such as age, gender, and so forth, and any detecting objects such as banners, posters, and sticks held by the members. The next information one might be interested in is continuously following up on the crowd movement to know whether the crowd is stationary or moving in a specific direction. The aforementioned aspects are covered in the first three areas of crowd analysis. Then, a sophisticated system may provide a more detailed analysis such as crowd behavior (mood, specific activities performed by crowd members), and detecting specific abnormal events or objects. Lastly, the prediction about crowd formation or dispersion in a specific region or area at a specific time can be a piece of very useful information that one would desire to obtain.

While traditionally, many semi-automated computer vision methods have been proposed (Davies, Yin, and Velastín 1995; Silveira Jacques Junior, Musse, and Jung 2010), the recent advancements in modern deep learning have revolutionized the development of fully automated vision-based crowd-monitoring applications. By leveraging the power of deep neural networks, the accuracy, efficiency, and overall performance of such applications have been significantly improved. Deep learning methods automatically learn and extract meaningful patterns and features from large-scale crowd visual data such as images and videos, thus enabling better crowd analysis. Also, the use of transfer learning allows deploying models trained on one dataset in a different scenario after fine-tuning, speeding up the training process.

Similar and related studies

Several studies have been conducted in the past, which typically focus on individual areas of crowd analysis. For example, Sindagi and Patel (2018), Cenggoro (2019), Ilyas, Shahzad, and Kim (2020), Gao et al. (2020), Luo, Lu, and Zhang (2020), Gouiaa, Akhloufi, and Shahbazi (2021), Fan et al. (2022), and Khan, Menouar, and Hamila (2023c) cover crowd counting research and mainly discuss the advancements in model architectures, benchmarking, and datasets. Hu et al. (2004b), Luca et al. (2020), and Kumar (2021) focus on crowd motion analysis, discussing crowd motion predictions, flow classification, and behavior analysis using motion patterns. Joshi and Patel (2021), Modi and Parikh (2022), and Sharif, Jiao, and Omlin (2022) study anomaly detection with a focus on methods, datasets, and comparisons of results. Unlike the aforementioned studies, few studies also cover the multiple aspects of crowd analysis. For instance, Li et al. (2015) cover motion analysis, behavior recognition, and anomaly detection but does not cover counting, density estimation, object detection, crowd prediction, and so forth. Swathi, Shivakumar, and Mohana (2017) cover density estimation, motion detection, and behavior recognition, but do not cover object detection, and anomaly detection. Similarly, Zhang, Yu, and Yu (2018) cover only physics-inspired methods for crowd analysis and focuses on motion analysis in videos. Table 1 presents a list of recent survey articles on the topic.

Although the studies discussed earlier provide a thorough review of the literature with useful taxonomy of works, our study presents a more comprehensive review consisting of the six intuitive areas of crowd analysis. Our survey is unique from the previous surveys as it does not provide a thorough review covering each and every related work published, but focuses on the most prominent works in each application area. Thus, this review is more concise, allowing even nonexperts and novice researchers in the area of crowd analysis to quickly understand the state-of-the-art in research on the respective topic. Due to the fast-paced research in computer vision, several related works published a few years before may not serve the purpose and yet another new study with fresh perspectives and insights is highly commendable.

Contribution and paper organization

Although several studies exist as stated in the previous section, discussing individual areas of crowd-monitoring applications, we aim to provide a concise study on the topic targeted to a broader audience from various disciplines. We intentionally omitted more technical details and focused more on the crux of the problem for a smooth flow for the reader. For instance, there have been more than a hundred crowd-counting models published in the literature, we chose only those with significant contributions in terms of novelty in model architecture, performance gains, or other aspects of design and evaluation. Similarly, we carefully chose original research works with major contributions (despite if the work is overlooked previously) in each application domain and included them in our survey to summarize the SOTA in each of the six areas of crowd analysis.

This article discusses the six intuitive areas of crowd visual analysis with a concise description of each, provides a brief overview of the state-of-the-art methods, and highlights the open problems in each domain. Each subsequent section describes one of the six areas. The last section draws the conclusion and presents future insights and research directions to advance the SOTA in the respective area.

COUNTING AND DENSITY ESTIMATION

One of the first aspects of crowd monitoring is to estimate the headcount (a scalar value for the whole image) or density across different parts of the scene. As people in the crowd are usually clumped together into groups, density estimates provide more information than just the total count. Headcount or density estimation can provide good situational awareness for the monitoring entities like law enforcement agencies and event managers.

Methods and state-of-the-art

The de facto method for counting people in an image or video frame is *density estimation*. A convolution neural network (CNN) based model is trained to estimate the crowd density. In the density estimation method, each head is detected using a Gaussian blob around the center of the head. This is a pixel-level regression problem and the commonly used Euclidean loss function (or mean squared error MSE) is used to train the model. Figure 2 presents the ground truth density map for a crowd image used by the CNN model.

TABLE 1 Related studies on crowd visual analysis.

Ref.	Topic	Brief scope
Sindagi and Patel (2018)	Crowd counting	Methods, model architectures, datasets, future insights.
Ilyas, Shahzad, and Kim (2020)	Crowd counting	Methods, model architectures, performance evaluation.
Gao et al. (2020)	Crowd counting	Model architectures and categorization, datasets, evaluation metrics, challenges, and insights.
Gouiaa, Akhloufi, and Shahbazi (2021)	Crowd counting	Methods, the taxonomy of CNN models, datasets, and applications.
Fan et al. (2022)	Crowd counting	Methods, applications, and models.
Khan, Menouar, and Hamila (2023c)	Crowd counting	Datasets, model architecture, loss functions, evaluation metrics, insights.
Hu et al. (2004b)	Motion analysis	Motion analysis, behavior analysis, moving object classification, and identification.
Luca et al. (2020)	Motion analysis	Flow prediction, next-location prediction, flow generation, and trajectory generation.
Kumar (2021)	Behavior analysis	Methodologies for the organized crowd and nonorganized crowd, datasets.
Modi and Parikh (2022)	Anomaly detection	Methods for anomaly detection with merits and demerits.
Sharif, Jiao, and Omlin (2022)	Anomaly detection	Methods, algorithms, datasets, and metrics with details analysis and comparisons.
Li et al. (2015)	Crowd analysis	Models, datasets, algorithms, and evaluation of crowd motion analysis, behavior recognition, and anomaly detection.
Swathi, Shivakumar, and Mohana (2017)	Crowd analysis	Density estimation, motion detection and tracking, and behavior recognition.
Zhang, Yu, and Yu (2018)	Crowd analysis	Physics-inspired methods for crowd video analysis.

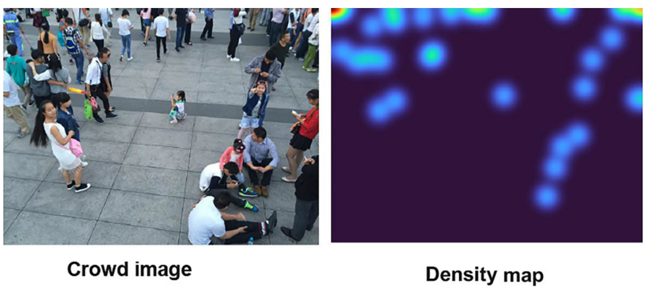

FIGURE 2 Crowd counting using density estimation. The predicted density map shows crowd density across the image scene. The sum of all pixel values of the density map equals the predicted count in the image.

The first CNN-based crowd density estimation model was CrowdCNN (Zhang et al. 2015). The CrowdCNN is a single-column CNN architecture that outperforms traditional non-CNN methods however, still lacks the capability to adapt to the scale variations in head sizes. Scale variation can arise from several reasons such as distance from the camera, camera perspective effects, image resolution, and so forth. To overcome scale variations, multicolumn CNN networks were proposed (Zhang et al. 2016; Sindagi and Patel 2017). Multicolumn models can capture scale variations to some extent (i.e., more columns are needed for images with larger scale variations, which may increase the model size significantly.). However, more efficient architectures were proposed to replace multicolumn architectures, for example, encoder–decoder networks (Jiang et al. 2019; Song et al. 2021; Gao, Wang, and Gao 2019), networks with multiscale modules (Zeng et al. 2017; Wang and Breckon 2022), and so forth. Encoder–decoder models allow hierarchical feature extraction and aggregation at multiple stages in the encoder and decoder modules, respectively. These models are good at preserving the spatial resolution of the predicted density maps using downsampling and upsampling operations, however, take longer times to train and converge. Recently, non-CNN models using vision transformers are also proposed (Liang et al. 2022; Tian, Chu, and Wang 2021). Transformer-based models apply self-attention mechanisms to model global context and capture long-range dependencies. Their disadvantage is that these models require more computational resources compared to CNNs due to the larger number of parameters and self-attention operations.

Crowd counting is the relatively most rigorously investigated area after object detection with the availability of a large number of datasets (Mall, Chen et al. (2012); ShanghaiTech, Zhang et al. (2016); UCF-QNRF, Idrees et al. (2018); JHU-Crowd, Sindagi, Yasarla, and Patel (2020); NWPU-Crowd, Wang et al. (2021); DroneRGBT, Peng, Li, and Zhu (2020); etc.) and a large variety of models

TABLE 2 A summary of research efforts in crowd counting, categorized in distinct categories, that is, model architecture, loss functions, metrics, and training methods.

Research area	Summary of the SOTA
Model architectures	Single-column, Zhang et al. (2015); multicolumn, Zeng et al. (2017); Encoder–decoder, Jiang et al. (2019); pyramid, Vision transformers, Liang et al. (2022).
Loss functions	Euclidean loss, Zhang et al. (2016); combination loss, Zhang et al. (2015); curriculum loss, Wang and Breckon (2022); composite loss, Idrees et al. (2018); AP loss, Jiang et al. (2020); PRA loss, Jiang et al. (2020); SCL loss, Jiang et al. (2019); OT, Wang et al. (2020).
Evaluation metrics	MAE, MSE, GAME, Guerrero-Gómez-Olmedo et al. (2015); SSIM, Li, Zhang, and Chen (2018); PSNR, Li, Zhang, and Chen (2018); PAME, PMSE, MPAE.
Training methods	Supervised/weakly supervised, Yang et al. (2020); Liang et al. (2022); Lei et al. (2021); curriculum learning, Khan, Menouar, and Hamila (2023b).

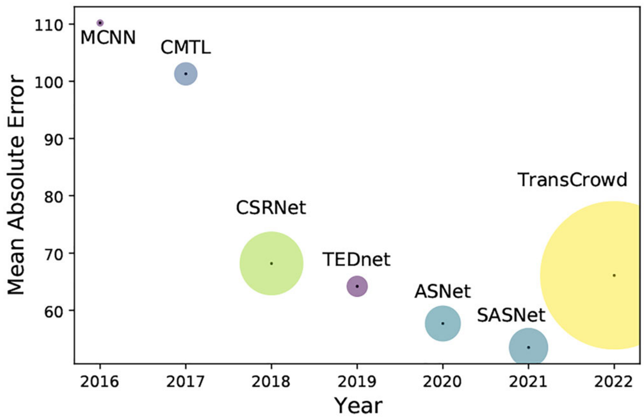

FIGURE 3 Performance of well-known crowd counting models over years.

with reasonably good performance. The performance gains in the counting accuracy have been high initially when single-column shallow models have been replaced by multicolumn CNN, and then pyramids structure models using transfer learning and multiscale modules. However, over the last 1–2 years, the gain in accuracy has been incremental despite major architectural changes and novel loss function. Figure 3 compared some well-known crowd counting models showing the mean absolute error (MAE) over a benchmark dataset and the size of the model parameters. Furthermore, a summary of model architectures, the learning (loss) functions, evaluation metrics, and the training methods (supervised/weakly supervised, etc.) is presented in Table 2.

Challenges and open problems

The accuracy performance (measured using MAE metric) of crowd counting models has significantly improved over time on benchmark datasets. For instance, the MAE over ShanghaiTech Part A dataset for MCNN (Zhang et al. 2016) was 110.2, which is reduced to 66.1 by TransCrowd (Liang et al. 2022). A similar improvement has been achieved on other benchmarking datasets (Khan, Menouar, and Hamila 2023c) as well. Despite the significant improvements in counting accuracy, there still remain several challenges. The problem of scale variations caused due to perspective effects have been overcome by multiscale architectures, other issues such as occlusions and complex backgrounds are still a major challenge in many complex scenes. Although some methods propose background segmentation as a preprocessing step, that can increase the complexity of the task. Second, there is still more room for further improvement in counting accuracy on new benchmark datasets such as UCF-QNRF, JHUCrowd++, and NWPU Crowd datasets due to more challenging scenes, for example, extremely dense crowds, extreme weather and low light conditions, camera blur effects, and so forth. These conditions are common in the real world and the results typically achieved over previous datasets would not be attainable on these new datasets and thus in the real world. Third, the commonly used metric in almost all studies is MAE provides an average performance of crowd-counting models. In practice, a model may produce less accurate predictions on difficult examples (e.g., extremely dense and blurry images), but the MAE (being an average over the entire test set) is compensated to be low due to better predictions on easy examples posing a low value of MAE (e.g., less dense and clear images). Fourth, as the benchmark datasets become larger and more challenging over time, the resulting models to achieve better performance over these benchmarks become deeper. This means more complexity to run these models in real-time, especially on resource-limited edge devices. This will create potential bottlenecks in edge-based crowd-monitoring solutions. Lightweight models are being developed (Khan, Menouar, and Hamila 2023a, 2023b), but these provide limited accuracy in very dense crowds.

We urge the requirement of more scenario-specific datasets to build reliable models for production. Models trained on generic datasets typically require a long training time and yet lack generalization capabilities.

OBJECT DETECTION AND TRACKING

There are many situations in which one is interested to detect an object of interest in a video frame and then track it over time in consecutive frames. Detecting a particular object class (e.g., pedestrians, vehicles, men, women, etc.) in a single image or video frame is called object detection, whereas identifying an individual object or a set of objects in consecutive video frames from a single camera or frames from multiple cameras is called object tracking.

Methods and state-of-the-art

Object detection is a well-researched problem that gained significant attention in computer vision, and several prominent models are developed over time that achieved state-of-the-art performance. These models are typically divided into two categories, that is, anchor-based methods and anchor-free models. Anchor-based methods use a predefined set of anchor boxes placed over the entire image and predict the final set of boxes around the detected objects. These models provide better accuracy and are widely deployed, for example, Faster-RCNN, Girshick (2015); SSD, Liu et al. (2015); and YOLO (v3), Redmon and Farhadi (2018). Anchor-based methods adapt well to various scales and aspect ratios and work well in complex scenes. However, their performance is greatly affected by various factors such as sizes, aspect ratios, the number of anchor boxes, and shape variations (Tian et al. 2019). The predefined anchor boxes require manual design and careful calibration, which can be time-consuming and computationally expensive. However, anchor-based methods have been widely adopted.

Anchor-free methods are relatively new and more efficient than anchor-based methods. They use the keypoint detection approach. For example, in CornerNet Law and Deng (2018), the model predicts the top-left and bottom-right corners around objects to draw the final bounding box. In CenterNet (Duan et al. 2019), a single point at the center of the object is detected to draw the bounding box.

A fully convolutional one-stage (FCOS) detector is another single-stage anchor-free method proposed in Tian et al. (2019) that detects objects using per-pixel prediction that achieve comparable performance to anchor-based methods (Faster-RCNN, YOLOv2, etc.) and outperforms previous anchor-free methods (CornerNet, etc.). Anchor-free methods are simpler in design and implementation, and computationally efficient. However, these models perform poorly in accurately localizing small objects or objects with complex shapes, especially in dense and overlapping instances. They also require more training data to sufficiently train.

FIGURE 4 Object detection sample from EuroCity Persons dataset (Braun et al. 2019).

TABLE 3 A summary of object detection models.

Model Category	Examples
Anchor-based	FasterRCNN, Girshick (2015); SSD, Liu et al. (2015); YOLO.
Anchor-free	CornerNet, Law and Deng (2018); CenterNet, Duan et al. (2019); FCOS, Tian et al. (2019).
Transformer models	Vision transformers (ViT), Chen et al. (2022).

More recently, non-CNN methods are getting attention in detection tasks. The vision transformer (ViT) model is proposed in Chen et al. (2022) as an alternative to CNN-based models. The original ViT and other transformer-based models have shown comparative performance in many tasks compared to several CNN-based models. However, CNNs are still considered as de facto methods for detection tasks due to their fast learning capabilities (training and fine-tuning) as compared to the transformer models. Figure 4 depicts a sample output of an object detection model showing bounding boxes around the detected objects (i.e., persons). Table 3 shows a list of objection detection models in the three categories.

There have been several publicly available datasets for object detection tasks; however, the most popular datasets used for benchmarking are Pascal VOC (Everingham et al. 2012) and MS-COCO (Lin et al. 2014). All mainstream object detection models (e.g., YOLO family, FCOS, CornerNet, etc.) are evaluated over these datasets, making the fair benchmarking of model accuracy.

Figure 5 shows a performance comparison of various object detection models discussed in this section.

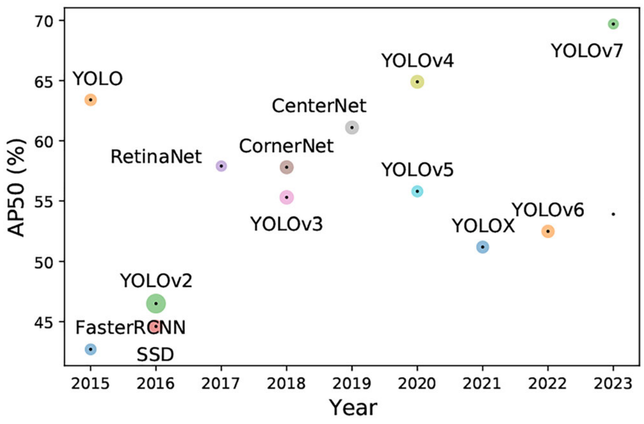

FIGURE 5 Average precision (AP50) of well-known object detection models on MS COCO dataset (except YOLO, which is evaluated on Pascal VOC dataset).

Challenges and open problems

Both anchor-based and anchor-free methods have achieved significant performance gains (in production) in the past few years, but both methods still face some intrinsic limitations in the context of crowd scenes. While some of these challenges are generic to any computer vision tasks, some are more severe in crowded scenes given as follows: Viewpoint variation–when an object looks different when captured from different angles. Object deformation—when an object appears in different shapes in the same frame or in consequent frames of a video (e.g., a person bending down). Severe occlusions—when one or more objects are partially not visible in an image due to overlapping with other objects in front of them. Illuminations—when there are large variations of brightness values of pixels in images. Clutters—when an image contains many or large objects other than objects of interest. The aforementioned issues are very likely to be encountered in many real-world scenarios. Although augmentation techniques play a role to improve the models' performance in learning such challenging environments, there is still no standard method that solves these problems in all scenarios. It is worthy noting that the existing models for object detection are not well-suited for crowd environments and hence despite finetuning produce poor results. This is a serious concern that hinders the adoption of serious surveillance applications (e.g., that can be used by law enforcement agencies).

MOTION ANALYSIS

Understanding crowd dynamics can provide more useful information in addition to the crowd count and density estimates. For instance, one may be interested to know whether the crowd is stationary or moving. For a moving crowd, it will be interesting to understand the crowd flux and other patterns related to the crowd movement including trajectory, direction, velocity, and so forth. It also includes detecting stationary groups in the crowd. Motion analysis has many interesting applications, for example, access control, human identification, congestion analysis, and multicamera interactive surveillance (Hu, Tan, Wang, and Maybank 2004a). Motion analysis may refer to recognizing the movement of body parts of a person (e.g., gestures, actions, etc.), but in the context of the crowd, it is often referred to as the coherent motion of a group of individuals. Crowd motion analysis is of great interest in understanding crowd behavior analysis and scene understanding, for example, categorizing a crowd as a stationary or moving crowd, crowd trajectory predictions, crowd flux analysis, and crowd motion patterns analysis. Motion analysis may also include studying abnormal motion behavior (Gupta, Nunavath, and Roy 2019).

Methods and state-of-the-art

Crowd motion analysis methods aim to detect, track, and analyze motion patterns (e.g., in Figure 6) to infer important insights about the dynamic behavior of a crowd. Technically, motion analysis includes tempo-spatial analysis of the crowd. There have been several manual and end-to-end automated methods and mathematical models for crowd movement statistics. Traditionally, motion analysis used methods such as motion segmentation (pixel-wise separation of moving objects from the background), temporal differencing (pixel-wise differences consecutive frames), and optical flow (using flow vectors of moving objects over time) (Hu et al. 2004b). For instance, a running count of people's trajectories passing through user-defined lines in a scene is used to measure crowd flows. Several flows can be further integrated over multiple spatial and temporal windows. Crowd motion is represented in different ways, for example, optical flow (movement of each pixel from one frame to another), particle flow (moving grid of particles with optical flow through numerical integration), streaklines flows (traces left as line upon injection of colored material in the flow), spatiotemporal features, tracklets (a fragment of trajectory obtains by a tracker with a short period of time), and so forth (Saqib 2019).

Motion segmentation-based methods can be easily implemented using background subtraction or may use methods such as clustering or graph-based approaches. These methods are more sensitive to noise and variations

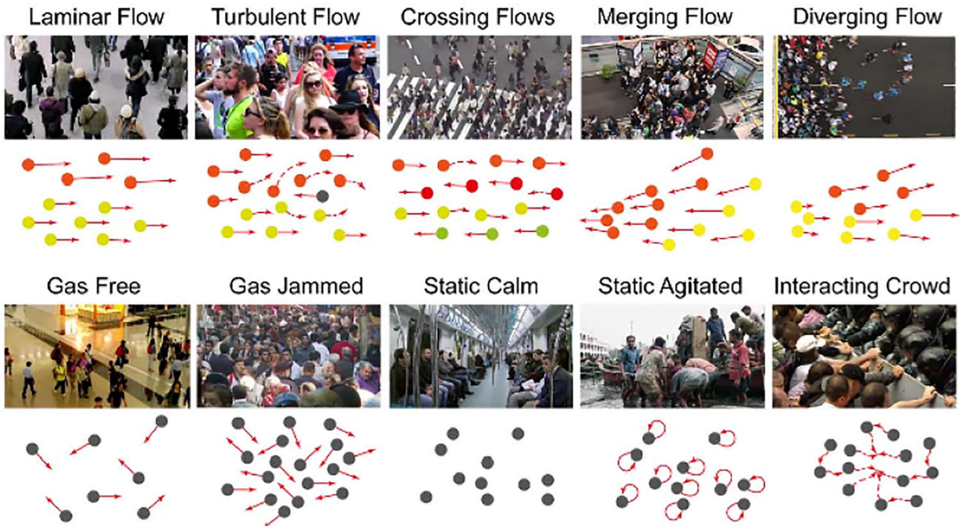

FIGURE 6 Crowd motion patterns in Crowd 11 dataset (Dupont, Tobías, and Luvison 2017).

in lighting conditions and poorly perform in the presence of occlusions and recognizing complex motion patterns. Optical flow-based methods capture the motion of every pixel. The algorithms are computationally efficient allowing real-time performance. However, these methods suffer from inaccuracies to capture small motions between frames, frames with repetitive patterns, and significant scale variation.

Authors in Rabaud and Belongie (2006) focus on counting moving objects in a crowd by detecting independent motions. Similarly, Lin, Grimson, and Fisher (2009) and Gupta, Nunavath, and Roy (2019) study the global motion patterns in crowds. Ali and Shah (2008) study the segmentation of crowd flows whereas Hu, Ali, and Shah (2008) use optical flows to learn the crowd motion patterns.

Challenges and open problems

Most research on crowd motion analysis is based on the representation of pictorial information such as color, texture, and so forth. However, these methods only provide some very basic analysis of motion patterns and do not sufficiently provide semantic information on crowd motion. Automated motion analysis not only requires large-scale video datasets but also efficient methods to extract both micro and macro statistics of crowd motion. The existing methods seriously lack the sophistication required for real-world implementations. To extract more high-level and intuitive motion information, models need to learn automatic semantic features and relationships between low-level pictorial features and high-level semantic features.

BEHAVIOR, ACTIVITY, AND CONTEXT RECOGNITION

Behavior analysis refers to the analysis of the crowd as a whole or a portion of the crowd on a longer time scale (i.e., minutes to hours). It may involve simpler tasks such as identifying the state of the crowd's behavior, for example, calm, active, violent crowd, or may involve complex activity recognition. Activity recognition refers to detecting various grouped activities of crowd members such as protesting, dancing, fighting, and so forth usually on a shorter time scale (i.e., seconds to minutes). It can also refer to actions of an individual object, for example, pose estimation. Some examples of group activities are depicted in Figure 7.

The individuals in a crowd may interact with each other and engage with each other in different activities. While activity recognition detects crowd activities using shape, pose, or motion features, context analysis studies the social interaction among group members using time, location, and other contextual information and their relation to the crowd. Context analysis of crowd activity is a more complex problem than activity recognition due to many other factors, for example, environment and other factors related to psychology and sociology.

Methods and state-of-the-art

Activity recognition involves a framework for defining individual and grouped activities and then assigning unique descriptors to each activity. The next step is then to accurately learn the spatial and temporal profiles of

FIGURE 7 Activity recognition with several crowd activities (Wang et al. 2022).

each activity either using traditional methods such as mathematical models or machine learning or to apply end-to-end learning using deep learning models. Behavior analysis is typically considered as a macroscopic crowd analysis, whereas activity recognition may refer to as microscopic crowd analysis. In Kok, Lim, and Chan (2016), the authors present common attributes of a crowd (i.e., decentralized, collective motion, emergency behavior) and map these attributes to biological and physical entities.

The research on contextual analysis of crowd activities analysis is still very limited due to the inherited complexity of the problem. Authors in Benetka, Krumm, and Bennett (2019) present a qualitative analysis of the context in human activity recognition using several attributes such as time and location to establish a spatiotemporal context in the human activity prediction system. Several works propose the use of low-level feature-based methods such as optical flow. In Tran et al. (2015), authors studied various aspects of the crowd context analysis. First, discovering meaningful groups in a crowd is modeled as a dominant set clustering algorithm. Second, it uses group context activity (GCA) descriptors of a target person and its semantic neighbors and applies conditional random field (CRF) and support vector machines (SVM). The authors use two datasets (i) the Collective Activity dataset that involves simple activities, that is, crossing, waiting, queuing, walking, and talking (with two additional activities, i.e., dancing and jogging), and (ii) UCLA Courtyard dataset having 10 human activities (Riding, Skateboard, Riding Bike, Riding Scooter, Driving Car, Walking, Talking, Waiting, Reading, Eating, and Sitting).

Challenges and open problems

Activity recognition and behavior analysis are generally more complex due to the semantic relationships between the detected activities to the human habits. A single behavior can be generally mapped to multiple semantic concepts. Thus, to infer a meaningful semantic behavior, an accurate relationship between the low-level features and the semantic behavior must be established apriori. It is generally very hard to detect and track individual persons in different crowded scenes (Shao, Loy, and Wang 2017). Some key problems in behavior analysis include applying background knowledge and reasoning theory to correctly define natural language descriptors to semantic behaviors and then learning these behaviors from the transformations of the object in a scene at different levels. It is generally desired to use multiple cameras in crowd surveillance. However, it is also more challenging to apply data fusion from multiple sources (e.g., cameras or other sensors), which involves automatically inferring human activities and behaviors from multiple features (rather than images or frames). Feature extraction and fusion from multiple camera sources also require significant hardware resources and large-scale adoption of such systems can be slower. Context recognition in crowd analysis is of major importance in a fully automatic crowd surveillance application. It mostly involves multimodal data from several sources including camera outputs (images or videos) as well as data from sources such as social media (tweets, Facebook posts, comments, etc.), and real-time check-in/checkout records from venues such as airports, metro stations, event venues, and so forth. The rich information obtained from these diverse sources can provide

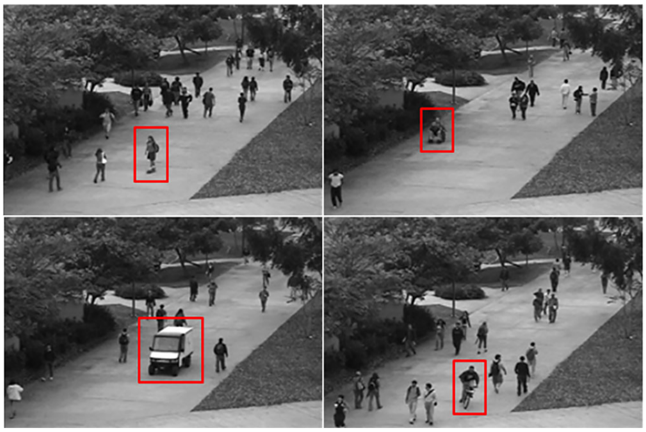

FIGURE 8 Anomaly examples in pedestrian spaces, for example, wheelchair, skater, biker, and cart. Mahadevan et al. (2010b).

more accurate contextual information on crowd activities and more accurate semantic descriptors. However, the research in this area is still very limited, and existing works only touch the surface of this broad and deep area in crowd analysis.

ANOMALY DETECTION

Anomaly detection refers to finding anomalous (aka abnormal) events and has many significant applications, for example, crime detection, traffic violations, abandoned objects detection, weapons detection, and so forth. In the context of crowd monitoring, anomalies are found using spatiotemporal feature analysis of the video frames as well as in a single image. Figure 8 shows example anomalies in an outdoor scene.

Methods and state-of-the-art

Crowd anomaly detection is a challenging task mainly due to the rare occurrence of such events and thus lack of sufficient data on anomalous events. The definition of anomaly is subjective and the same event can be classified as normal and abnormal at different times and places. Anomalies can be segregated into (i) point anomaly (when a single entity or person looks or behaves differently than other entities in the scene), (ii) contextual anomaly (when an entity or object is treated as abnormal in a specific contextual situation or environment), and (iii) collective anomalies (when a group of instances behaves abnormally than the rest of the entities in the scene). Anomaly detection is also classified as a global anomaly (does the scene/frame has anomaly) versus local anomaly (the localization of the anomaly in the frame or video).

The most basic and commonly used approach to detect anomalies is to train a one-class classifier (OCC) trained with data containing normal examples. The OCC model trained with a sufficiently large number of normal training instances can predict abnormal events. However, even collecting data with all normal events is not easy.

Anomaly detection in crowded scenes has been addressed in numerous ways using a variety of algorithms including classical machine learning schemes, for example, k-means and SVMs (Yang et al. 2019), GMM, and so forth, as well as deep learning methods, for example, CNNs (Joshi and Patel 2021; Pang et al. 2020; Wu et al. 2020), LSTM (Esan, Owolawi, and Tu 2020), GANs (Luo, Liu, and Gao 2017a; Chen et al. 2021), and autoencoders (AEs) (Simonyan and Zisserman 2014; Pawar and Attar 2021), bag-of-words (BOW) method, and physics-inspired approaches (Wu, Moore, and Shah 2010), and so forth. These methods are often inter-related and the exact taxonomy is difficult. In BOW method, spatiotemporal visual features are extracted to detect abnormal events (Javan Roshtkhari and Levine 2013). BOW method for anomaly detection may not capture complex anomalies and can be used to train on specific anomalies. Datasets for anomaly detection include UCSD (Mahadevan et al. 2010a), PETS2009, UMN (Bird et al. 2006), CUHK-Avenue (Lu, Shi, and Jia 2013), and ShanghatTech Campus (Luo, Liu, and Gao 2017b).

Challenges and open problems

Anomaly detection is very challenging due to several reasons including data availability, compute power requirement, fairness, and generalization. Some of the open problems and challenges are listed as follows: First, there is no universal definition of an abnormal event, that is, an event that is considered abnormal may be considered abnormal in a different context. For example, a person carrying a gun is abnormal but becomes normal when the person carrying the gun is a police constable. Thus, the context is always significant in anomaly detection, which makes the anomaly detection problem very challenging. This is a serious challenge that must be tackled first to enhance the outcomes of research on anomaly detection. Contribution from government and private law enforcement agencies can play significant in data acquisition and annotation. Without sufficient data and standard definitions of crowd anomalies, the research outcomes will be significantly limited. Second, there is a lack of good datasets for anomaly detection. The existing datasets cover only a small number of anomalies. The methods to label data in these datasets also vary (frame-level anomalies, video-level anomalies, segment-level anomalies, etc.). Third, it is not easy to create large datasets with diverse anomalies because several abnormal events can not be obtained beforehand. Video datasets for anomalies are sparse, and not typically accessible publicly. The publicly available datasets are mostly captured in the same location and capture very few anomalies. Fourth, most real-world anomalies can be better captured in video sequences, thus the anomaly detection datasets are typically of very huge size (as compared to other computer vision tasks such as object classification or detection). As a result, huge computational resources are required to train deep-learning models over tons of videos.

CROWD PREDICTION

Crowd prediction refers to predicting in advance crowd accumulation in a particular region. Predicting crowds ahead of time has key significance in various scenarios and applications such as events, detecting incidents, tourism attractions, and so forth. Crowd prediction typically is complex and often leverages multimodal data from various sources, such as surveillance cameras, social media, mobile phone signals, sensors, and so forth. Machine learning and statistical modeling techniques are employed to learn and make predictions about future crowds. The process may relate to specific crowd events or gatherings, to estimate the duration of the event, forecast peak crowd hours, or anticipate the popularity of certain areas or attractions within the event. A crowd in specified administrative areas can emerge typically in different ways, for example, **Fountainhead**: crowd emerges from a single direction and spread over different directions just like a fountain, **Bottleneck**: crowd emerges from different directions and assembles at a single point, and **Lane**: crowd emerges at a uniform rate and moves in a single direction. Crowd prediction involves forecasting the aggregated flows (incoming and outgoing) in a specified region. Further information such as the origin and destination of crowd flows are also studied as part of the crowd prediction problem.

Methods and state-of-the-art

Crowd prediction can be simply modeled as a time series forecasting problem to regress an increase in crowd density over time starting from a no-crowd condition or can also use spatial statistics of crowds to predict crowds in spatial regions. Traditionally, it uses moving averaging methods such as autoregressive integrated moving average (ARIMA) (Liu et al. 2021). However, these methods fail to capture the complex temporal and spatial dependencies despite several feature engineering techniques. To cope with complex dependencies, deep learning methods including CNN (Song et al. 2020), LSTM (Li et al. 2019), and graph neural networks (GNNs) (Li et al. 2022) have been proposed. Although multimodal data can be used for predicting crowds, a sufficient quantity of such datasets does not exist. The use of LSTM and CNNs together has been a more promising method to predict crowds over shorter time periods. However, over longer time spans, forecasting methods (e.g., ARIMA) are applied.

Challenges and open problems

Crowd prediction can either be logically concluded from the analysis of crowd motion patterns or predicted directly using time series prediction models. However, the research in this area is very limited, and very few works are found often using deep learning methods such as CNN and LSTM. Crowd prediction has both spatial and temporal dependencies and generally, it is quite difficult to predict crowd accumulation in a region over long periods. The prediction in fountainhead and bottleneck is more difficult than a lane pattern. Due to the spatial dependency, crowd flow prediction in irregular-shaped regions can be more challenging than in regular-shaped regions.

Last, Table 4 provides a summary of the aforementioned six crowd analysis areas, some common examples of tasks in each area, and open research challenges.

CONCLUSION AND FUTURE INSIGHTS

This article defines six major areas of visual crowd analysis, which together form a full-fledged automated crowd-monitoring system. Each of these six areas involves a different level of complexity, and thus the state-of-the-art greatly varies in these. For instance, crowd counting and detection are two areas with significantly improved results achieved recently over large benchmark datasets. Typically, existing real-world crowd-monitoring implementations cover these tasks. The benchmarking in counting and detection is also very clear and can be easily compared. However, there are still areas requiring significant attention. In counting and density estimation, the common regression loss function (Euclidean loss or MSE) has been used. However, over time, several other loss functions (e.g., OT loss, AP loss, PRA loss, etc.) have been proposed with reportedly improved performance. However, these loss functions could not be continued in the subsequent works and many recent works consistently used the original MSE loss function. It is encouraged to evaluate these loss functions over different datasets and several mounting

TABLE 4 Crowd analysis tasks and Open research problems.

Crowd Analysis	Example Tasks	Open Challenges
Crowd Counting	Find the total headcount in a frame. crowd localization in a large region.	Handling severe occlusions, scale variations in highly dense crowd images, perspective distortion, multi-view counting.
Object Detection	Detect and localize objects of interest. Track objects in a video.	Viewpoint variations, Object deformation, severe occlusions, clutter, illumination, etc.
Motion Analysis	Observe crowd collectiveness. Is the crowd slow/fast-moving? Where is the crowd heading? at which speed? Find trajectories and motion patterns.	Complex semantic features, complex relationship between low-level pictorial features with high-level semantic features.
Behavior/Activity Analysis	Is the crowd calm or active? Are people protesting, dancing, fighting, etc?	Complex mapping of low-level features to activities, semantic relationship and mapping of detected activities to human behavior, standard activity definitions to semantic behavior.
Anomaly Detection	Detect abandoned objects. Detect any weapons or crimes? Detect intrusion.	Lacks of a unified definition of anomalies, lack of realistic datasets, huge size of video datasets[1].
Crowd prediction	Detect if a crowd is expected in a region? How crowd accumulates over time?	Difficult to capture dual (spatial and temporal) dependencies, complex motion patterns.

1An extensive list of challenges can be found in Sharif, Jiao, and Omlin (2022).

models to conclude their potential benefits in this domain. We also encourage testing the potential performance gain of using curriculum learning (CL) in crowd-counting tasks with appropriate curriculum strategies.

Motion analysis that focuses on crowd-level mobility statistics often uses flow-based and spatiotemporal features. Flow-field models are relatively more studied. However, clustering models are gaining attention due to their performance in more crowded scenes. Motion analysis tasks vary such as detecting crowd source/sink, trajectory finding, speed, and so forth. The benchmarking in this application domain is less coherent due to the nature of the task and multiple objectives being considered in existing studies. Behavior analysis and anomaly detection, which are sometimes overlapping, are the most complex tasks, and the progress in these tasks is still very limited and scattered in terms of methods, approaches, assumptions, and objectives despite their major importance in several use cases. The lack of definition of anomalies, activities, and behavior causes researchers to use different objectives and evaluation metrics, which makes benchmarking unfair. We suggest that future research should focus on developing common definitions of activities and anomalies considering context and enhancing existing datasets as well as creating larger and balanced datasets. We also believe that in many scenarios, local anomalies would be required rather than global anomalies, which shall make the task easier to learn; however, the variations in environment and context will make the cross-domain transfer-learning challenging. Physics-inspired approaches (e.g., energy models) are interesting directions to implement anomaly detection.

We envision major advances in the near future in the under-explored areas due to the recent developments in generative AI, which will be helpful to cope with the need for more training data. Furthermore, crowd analysis in real-time typically requires fast inference. In CCTV-based surveillance, it is more convenient to perform all processing and inference tasks on a local server due to the high-speed wired connectivity option however, in aerial surveillance (using drones) on-device processing and inference may be more convenient in some cases. Thus, lightweight crowd analysis models will be preferred.

ACKNOWLEDGMENTS

This publication was made possible by the PDRA award PDRA7-0606-21012 from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

ORCID

Muhammad Asif Khan https://orcid.org/0000-0003-2925-8841

REFERENCES

Ali, Saad, and Mubarak Shah. 2008. "Floor Fields for Tracking in High Density Crowd Scenes." In *European Conference on Computer Vision*.

Benetka, Jan R., John Krumm, and Paul N. Bennett. 2019. "Understanding Context for Tasks and Activities." In *Proceedings of the*

2019 Conference on Human Information Interaction and Retrieval, CHIIR '19, 133–42. New York, NY, USA: Association for Computing Machinery.

Bird, N., S. Atev, N. Caramelli, R. Martin, O. Masoud, and N. Papanikolopoulos. 2006. "Real-Time, Online Detection of Abandoned Objects in Public Areas." In *Proceedings 2006 IEEE International Conference on Robotics and Automation, ICRA 2006*, 3775–3780.

Braun, Markus, Sebastian Krebs, Fabian B. Flohr, and Dariu M. Gavrila. 2019. "EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41: 1844–61.

Cenggoro, Tjeng Wawan. 2019. "Deep Learning for Crowd Counting: A Survey." *Engineering, MAthematics and Computer Science (EMACS) Journal* 1: 17–28.

Chen, Dongyue, Lingyi Yue, Xingya Chang, Ming Xu, and Tong Jia. 2021. "NM-GAN: Noise-modulated Generative Adversarial Network for Video Anomaly Detection." *Pattern Recognition* 116: 107969.

Chen, Ke, Chen Change Loy, Shaogang Gong, and Tony Xiang. 2012. "Feature Mining for Localised Crowd Counting." In *Proceedings of the British Machine Vision Conference*, 21.1–21.11. BMVA Press.

Chen, Zhe, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Y. Qiao. 2022. "Vision Transformer Adapter for Dense Predictions." *ArXiv* abs/2205.08534.

Davies, Anthony C., J. H. Yin, and Sergio A. Velastín. 1995. "Crowd Monitoring Using Image Processing." *Electronics & Communication Engineering Journal* 7: 37–47.

Duan, Kaiwen, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. "Centernet: Keypoint Triplets for Object Detection." In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6568–77.

Dupont, Camille, Luis Tobías, and Bertrand Luvison. 2017. "Crowd-11: A Dataset for Fine Grained Crowd Behaviour Analysis." In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2184–91.

Oladayo Esan, Dorcas, Pius Adewale Owolawi, and Chuling Tu. 2020. "Anomalous Detection System in Crowded Environment Using Deep Learning." In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, 29–35.

Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. https://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Fan, Zizhu, Hong Zhang, Zheng Zhang, Guangming Lu, Yudong Zhang, and Yaowei Wang. 2022. "A Survey of Crowd Counting and Density Estimation Based on Convolutional Neural Network." *Neurocomputing* 472: 224–51.

Gao, Chenyu, Peng Wang, and Ye Gao. 2019. "MobileCount: An Efficient Encoder-Decoder Framework for Real-Time Crowd Counting." In *Pattern Recognition and Computer Vision: Second Chinese Conference, PRCV 2019, Xi'an, China, November 8–11, 2019, Proceedings, Part II*, 582–95. Springer-Verlag.

Gao, Guangshuai, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. 2020. "CNN-Based Density Estimation and Crowd Counting: A Survey."*ArXiv* abs/2003.12783.

Girshick, Ross B. 2015. "Fast R-CNN." In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1440–8.

Gouiaa, Rafik, Moulay A. Akhloufi, and Mozhdeh Shahbazi. 2021. "Advances in Convolution Neural Networks Based Crowd Counting and Density Estimation." *Big Data and Cognitive Computing* 5: 50.

Guerrero-Gómez-Olmedo, Ricardo, Beatriz Torre-Jiménez, Roberto Javier López-Sastre, Saturnino Maldonado-Bascón, and Daniel Oñoro-Rubio. 2015. "Extremely Overlapping Vehicle Counting." In *IbPRIA*.

Gupta, Tanu, Vimala Nunavath, and Sudip Roy. 2019. "CrowdVAS-Net: A Deep-CNN Based Framework to Detect Abnormal Crowd-Motion Behavior in Videos for Predicting Crowd Disaster." In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2877–82.

Hu, Min, Saad Ali, and Mubarak Shah. 2008. "Learning Motion Patterns in Crowded Scenes Using Motion Flow Field." In *2008 19th International Conference on Pattern Recognition*, 1–5.

Hu, Weiming, Tieniu Tan, Liang Wang, and S. Maybank. 2004a. "A Survey on Visual Surveillance of Object Motion and Behaviors." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34(3): 334–52.

Hu, Weiming, Tieniu Tan, Liang Wang, and Stephen J. Maybank. 2004b. "A Survey on Visual Surveillance of Object Motion and Behaviors." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34: 334–52.

Shivakumar, H. Y.,Swathi, G., and H. S. Mohana. 2017. "Crowd Behavior Analysis: A Survey." In *2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT)*, 169–78.

Idrees, Haroon, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Ali Al-Maadeed, Nasir M. Rajpoot, and Mubarak Shah. 2018. "Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds." *Computer Vision – ECCV 2018*. eds. Ferrari, Vittorio, Hebert, Martial, Sminchisescu, Cristian, Weiss, Yair, pp. 544–559

Ilyas, Naveed, Ahsan Shahzad, and Kiseon Kim. 2020. "Convolutional-neural Network-based Image Crowd Counting: Review, Categorization, Analysis, and Performance Evaluation." *Sensors (Basel, Switzerland)* 20: 43.

Mehrsan Javan, Roshtkhari, and Martin D. Levine. 2013. "An Online, Real-Time Learning Method for Detecting Anomalies in Videos Using Spatio-Temporal Compositions." *Computer Vision and Image Understanding* 117(10): 1436–52.

Jiang, Xiaoheng, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. 2020. "Attention Scaling for Crowd Counting." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4705–14.

Jiang, Xiaolong, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David S. Doermann, and Ling Shao. 2019. "Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6126–35.

Vishnuprasad Joshi, Kinjal, and Narendra M. Patel. 2021. "A CNN Based Approach for Crowd Anomaly Detection." *International Journal of Next-Generation Computing* 12: 1–11.

Asif Khan, Muhammad, Hamid Menouar, and Ridha Hamila. 2023a. "DroneNet: Crowd Density Estimation Using Self-ONNs for Drones." In *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, 455–60.

Asif Khan, Muhammad, Hamid Menouar, and Ridha Hamila. 2023b. "LCDnet: A Lightweight Crowd Density Estimation Model for Real-Time Video Surveillance." *Journal of Real-Time Image Processing* 20: 29.

Khan, Muhammad Asif, Hamid Menouar, and Ridha Hamila. 2023c. "Revisiting Crowd Counting: State-of-the-Art, Trends, and Future Perspectives." *Image Vision Computing* 129(C). https://doi.org/10.1016/j.imavis.2022.104597.

Jyn Kok, Ven, Mei Kuan Lim, and Chee Seng Chan. 2016. "Crowd Behavior Analysis: A Review Where Physics Meets Biology." *Neurocomputing* 177: 342–62.

Kumar, Ajay. 2021. "Crowd Behavior Monitoring and Analysis in Surveillance Applications: A Survey." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12(7): 2322–2336.

Law, Hei, and Jia Deng. 2018. "CornerNet: Detecting Objects As Paired Keypoints." *International Journal of Computer Vision* 128: 642–56.

Lei, Yinjie, Yan Liu, Pingping Zhang, and Lingqiao Liu. 2021. "Towards Using Count-Level Weak Supervision for Crowd Counting." *Pattern Recognition* 109: 107616.

Li, Fuxian, Jie Feng, Huan Yan, Depeng Jin, and Yong Li. 2022. "Crowd Flow Prediction for Irregular Regions With Semantic Graph Attention Network." *ACM Transactions on Intelligent Systems and Technology (TIST)* 13: 1–14.

Li, Teng, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. 2015. "Crowded Scene Analysis: A Survey." *IEEE Transactions on Circuits and Systems for Video Technology* 25: 367–86.

Li, Wei, Wei Tao, Junyang Qiu, Xin Liu, Xingyu Zhou, and Zhisong Pan. 2019. "Densely Connected Convolutional Networks With Attention Lstm for Crowd Flows Prediction." *IEEE Access* 7: 140488–98.

Li, Yuhong, Xiaofan Zhang, and Deming Chen. 2018. "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1091–100.

Liang, Dingkang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. 2022. "Transcrowd: Weakly-Supervised Crowd Counting With Transformers." *Science China Information Sciences* 65(6): 160104.

Lin, Dahua, Eric Grimson, and John Fisher. 2009. "Learning Visual Flows: A Lie Algebraic Approach." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 747–54.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. "Microsoft COCO: Common Objects in Context." *Computer Vision – ECCV 2014*, eds. Fleet, David, Pajdla, Tomas, Schiele, Bernt, Tuytelaars, Tinne, 740–755 Cham. http://arxiv.org/abs/1405.0312.

Liu, Menghang, Luning Li, Qiang Li, Yu Bai, and Cheng Hu. 2021. "Pedestrian Flow Prediction in Open Public Places Using Graph Convolutional Network." *ISPRS International Journal of Geoinformatics* 10: 455.

Liu, W., Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2015. "SSD: Single Shot MultiBox Detector." In *European Conference on Computer Vision*.

Lu, Cewu, Jianping Shi, and Jiaya Jia. 2013. "Abnormal Event Detection at 150 FPS in MATLAB." In *2013 IEEE International Conference on Computer Vision*, 2720–7.

Luca, Massimiliano, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. 2020. "A Survey on Deep Learning For Human Mobility." *ACM Computing Surveys (CSUR)* 55: 1–44.

Luo, Weixin, Wen Liu, and Shenghua Gao. 2017a. "A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework." In *2017 IEEE International Conference on Computer Vision (ICCV)*, 341–9.

Luo, Weixin, Wen Liu, and Shenghua Gao. 2017b. "A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework." In *2017 IEEE International Conference on Computer Vision (ICCV)*, 341–9.

Luo, Ying, Jinhu Lu, and Baochang Zhang. 2020. "Crowd Counting for Static Images: A Survey of Methodology." In *2020 39th Chinese Control Conference (CCC)*, 6602–7.

Mahadevan, Vijay, Wei-Xin LI, Viral Bhalodia, and Nuno Vasconcelos. 2010a. "Anomaly Detection in Crowded Scenes." In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1975–81.

Mahadevan, Vijay, Wei-Xin LI, Viral Bhalodia, and Nuno Vasconcelos. 2010b. "Anomaly Detection in Crowded Scenes." In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1975–81.

Modi, Harshad S., and Dhaval A. Parikh. 2022. "A Survey on Crowd Anomaly Detection." *International Journal of Computing and Digital Systems* 12(1): 1081–1096.

Pang, Guansong, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. 2020. "Self-Trained Deep Ordinal Regression for End-to-End Video Anomaly Detection." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12170–9.

Pawar, Karishma, and Vahida Z. Attar. 2021. "Application of Deep Learning for Crowd Anomaly Detection from Surveillance Videos." In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 506–11.

Peng, Tao, Qing Li, and Pengfei Zhu. 2020. "RGB-T Crowd Counting from Drone: A Benchmark and MMCCN Network." In *Computer Vision – ACCV 2020: 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 –December 4, 2020, Revised Selected Papers, Part VI*, 497–513. Berlin, Heidelberg: Springer-Verlag.

Rabaud, V., and S. Belongie. 2006. "Counting Crowded Moving Objects." In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* 1: 705–11.

Redmon, Joseph, and Ali Farhadi. 2018. "YOLOv3: An Incremental Improvement." *ArXiv* abs/1804.02767.

Saqib, Muhammad. 2019. "Automatic Analysis of Crowd Dynamics Using Computer Vision and Machine Learning Approaches."

Shao, Jing, Chen Change Loy, and Xiaogang Wang. 2017. "Learning Scene-Independent Group Descriptors for Crowd Understanding." *IEEE Transactions on Circuits and Systems for Video Technology* 27(6): 1290–303.

Sharif, Md. Haidar, Lei Jiao, and Christian Walter Peter Omlin. 2022. "Deep Crowd Anomaly Detection: State-of-the-art, Challenges, and Future Research Directions." *ArXiv* abs/2210.13927.

Silveira Jacques Junior, Julio Cezar, Soraia Raupp Musse, and Claudio Rosito Jung. 2010. "Crowd Analysis Using Computer Vision Techniques." *IEEE Signal Processing Magazine* 27(5): 66–77.

Simonyan, Karen, and Andrew Zisserman. 2014. "Two-Stream Convolutional Networks for Action Recognition in Videos." In NIPS.

Sindagi, Vishwanath A., and Vishal M. Patel. 2017. "CNN-Based Cascaded Multi-Task Learning of High-Level Prior and Density Estimation for Crowd Counting." In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6.

Sindagi, Vishwanath A., and Vishal M. Patel. 2018. "A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation." *Pattern Recognition Letters* 107: 3–16.

Sindagi, Vishwanath A, Rajeev Yasarla, and Vishal M Patel. 2020. "Jhu-crowd++: Large-scale Crowd Counting Dataset and A Benchmark Method." Technical report.

Song, Qingyu, Changan Wang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Jian Wu, and Jiayi Ma. 2021. "To Choose Or To Fuse? Scale Selection for Crowd Counting." In *AAAI Conference on Artificial Intelligence*.

Song, Xiao, Kai Chen, Xu Li, Jinghan Sun, Baocun Hou, Yong Cui, Baochang Zhang, Gang Xiong, and Zilie Wang. 2020. "Pedestrian Trajectory Prediction Based on Deep Convolutional Lstm Network." *IEEE Transactions on Intelligent Transportation Systems* 22: 3285–302.

Thirumalaisamy, Manikandan, Shajahan Basheer, Shitharth Selvarajan, Sara A. Althubiti, Fayadh Alenezi, Gautam Srivastava, and Jerry Chun-Wei Lin. 2022. "Interaction of Secure Cloud Network and Crowd Computing for Smart City Data Obfuscation." *Sensors* 22(19). https://www.mdpi.com/1424-8220/22/19/7169.

Tian, Ye, Xiangxiang Chu, and Hongpeng Wang. 2021. "CCTrans: Simplifying and Improving Crowd Counting With Transformer." *ArXiv* abs/2109.14483.

Tian, Zhi, Chunhua Shen, Hao Chen, and Tong He. 2019. "FCOS: Fully Convolutional One-Stage Object Detection." In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9626–35.

Tran, Khai N., Xu Yan, Ioannis A. Kakadiaris, and Shishir K. Shah. 2015. "A Group Contextual Model for Activity Recognition in Crowded Scenes." In *VISAPP 2015 – Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Volume 2, Berlin, Germany, 11–14 March, 2015*, 5–12. SciTePress.

Wang, Boyu, Huidong Liu, Dimitris Samaras, and Minh Hoai. 2020. "Distribution Matching for Crowd Counting." In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada.

Wang, Hao, Junchao Liao, Tianheng Cheng, Zewen Gao, Hao Liu, Bo Ren, Xiang Bai, and Wenyu Liu. 2022. "Knowledge Mining With Scene Text for Fine-Grained Recognition." In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4614–23. IEEE. https://doi.org/10.1109/CVPR52688.2022.00458.

Wang, Qi, Junyu Gao, Wei Lin, and Xuelong Li. 2021. "NWPU-crowd: A Large-Scale Benchmark for Crowd Counting and Localization." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43: 2141–9.

Wang, Qian, and T. Breckon. 2022. "Crowd Counting Via Segmentation Guided Attention Networks and Curriculum Loss." *IEEE Transactions on Intelligent Transportation Systems* 23: 15233–43.

Wu, Chongke, Sicong Shao, Cihan Tunc, and Salim Hariri. 2020. "Video Anomaly Detection Using Pre-trained Deep Convolutional Neural Nets and Context Mining." In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, 1–8.

Wu, Shandong, Brian E. Moore, and Mubarak Shah. 2010. "Chaotic Invariants of Lagrangian Particle Trajectories for Anomaly Detection in Crowded Scenes." In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2054–60.

Yang, Meng, Sutharshan Rajasegarar, Sarah Monazam Erfani, and Christopher Leckie. 2019. "Deep Learning and One-class SVM Based Anomalous Crowd Detection." In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Yang, Yifan, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. 2020. "Weakly-supervised Crowd Counting Learns from Sorting Rather Than Locations." In *Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII*, 1–17. Berlin, Heidelberg: Springer-Verlag.

Zeng, Lingke, Xiangmin Xu, Bolun Cai, Suo Qiu, and Tong Zhang. 2017. "Multi-Scale Convolutional Neural Networks for Crowd Counting." In *2017 IEEE International Conference on Image Processing (ICIP)*, 465–9.

Zhang, Cong, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. 2015. "Cross-scene Crowd Counting Via Deep Convolutional Neural Networks." In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 833–41.

Zhang, Junwei, Zhuzhu Wang, Dandan Wang, Xinglong Zhang, Brij B. Gupta, Ximeng Liu, and Jianfeng Ma. 2022. "A Secure Decentralized Spatial Crowdsourcing Scheme for 6g-Enabled Network in Box." *IEEE Transactions on Industrial Informatics* 18(9): 6160–70.

Zhang, Xuguang, Qinan Yu, and Hui Yu. 2018. "Physics Inspired Methods for Crowd Video Surveillance and Analysis: A Survey." *IEEE Access* 6: 66816–30.

Zhang, Yingying, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. "Single-Image Crowd Counting Via Multi-Column Convolutional Neural Network." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 589–97.

How to cite this article: Khan, M. A., H. Menouar, and R. Hamila. 2023. "Visual crowd analysis: Open research problems." *AI Magazine* 44: 296–311. https://doi.org/10.1002/aaai.12117

AUTHOR BIOGRAPHIES

Muhammad A. Khan is an awarded post-doctoral Research Fellow at Qatar University. He received a Ph.D. degree in Electrical Engineering from Qatar University. His current research interests include visual crowd analysis, distributed machine learning, and edge intelligence. He is a Senior Member of IEEE and a Member of IET. For more detailed information, please visit his homepage: www.asifk.me

Hamid Menouar (SM'16) is a Senior Research and Development Expert at Qatar Mobility Innovations Center (QMIC), Qatar, where he leads the development of innovations in the area of smart cities and smart living. Before moving to Qatar in 2010, Hamid worked for 8 years at Hitachi Research Lab in Sophia Antipolice, France, where he contributed and led research and development activities related to connected vehicles. During his experience with Hitachi, he has heavily contributed to the development of connected vehicles communication standards in Europe through an active participation at ETSI TC ITS and Car to Car Communication Consortium. Hamid received the Engineering degree (M.S.) in Computer Science from the University of Science and Technology Houari Boumediene, Algiers, Algeria, in 2003, the DEA (M.S.) degree in systems and information technologies from the University of Technology of Compiegne, France, 2004, and the Ph.D. degree in Computer Science from Telecom ParisTech, Paris, France, in 2008.

Ridha Hamila received the MSc, LicTech with distinction, and PhD degrees from Tampere University of Technology (TUT), Tampere, Finland, in 1996, 1999, and 2002, respectively. Dr Hamila is currently a Full Professor at the Department of Electrical Engineering, Qatar University, Qatar. From 1994 to 2002, he held various research and teaching positions at TUT within the Department of Information Technology, Finland. From 2002 to 2003, he was a System Specialist at Nokia Research Center and Nokia Networks, Helsinki. From 2004 to 2009, he was with Emirates Telecommunications Corporation, UAE. Also, from 2004 to 2013, he was adjunct Professor at the Department of Communications Engineering, TUT. His current research interests include mobile and broadband wireless communication systems, Mobile Edge Computing, Internet of Everything, and Machine Learning. In these areas, he has published over 200 journal and conference papers, most of them in the peered reviewed IEEE publications, filed seven US patents, and wrote numerous confidential industrial research reports. Dr Hamila has been involved in several past and current industrial projects, Ooreedo, Qatar National Research Fund, Finnish Academy projects, EU research and education programs. He supervised a large number of under/graduate students and postdoctoral fellows. He organized many international workshops and conferences. He is a Senior Member of IEEE.