Recent trends in crowd analysis: A review

Mounir Bendali-Braham, Jonathan Weber, Germain Forestier, Lhassane Idoumghar, Pierre-Alain Muller

Machine Learning with Applications 4 (2021) 100023

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Recent trends in crowd analysis: A review

Mounir Bendali-Braham *, Jonathan Weber, Germain Forestier, Lhassane Idoumghar, Pierre-Alain Muller

IRIMAS, Université de Haute Alsace, 12 Rue des Frères Lumière, 68093 Mulhouse, France

ARTICLE INFO

Keywords:
Crowd analysis
Crowd behavior analysis
Group behavior analysis
Abnormal behavior detection
Deep Learning
Video-surveillance

ABSTRACT

When overpopulated cities face frequent crowded events like strikes, demonstrations, parades or other sorts of people gatherings, they are confronted to multiple security issues. To mitigate these issues, security forces are often involved to monitor the gatherings and to ensure the security of their participants. However, when access to technology is limited, the security forces can quickly become overwhelmed. Fortunately, more and more important smart cities are adopting the concept of intelligent surveillance systems. In these situations, intelligent surveillance systems require the most advanced techniques of crowd analysis to monitor crowd events properly. In this review, we explore various studies related to crowd analysis. Crowd analysis is commonly broken down into two major branches: crowd statistics and crowd behavior analysis. When crowd statistics determines the Level Of Service (LoS) of a crowded scene, crowd behavior analysis describes the motion patterns and the activities that are observed in a scene. One of the hottest topics of crowd analysis is anomaly detection. Although a unanimous definition of anomaly has not yet been met, each of crowd analysis subtopics can be subjected to abnormality. The purpose of our review is to find subareas, in crowd analysis, that are still unexplored or that seem to be rarely addressed through the prism of Deep Learning.

1. Introduction

Nowadays, the world overpopulation leads to multiple crowded situations in plenty of cities. These crowded situations stem from parades, stations' exits and entrances, political demonstrations, strikes. These situations imply a multiplication of security issues (Krausz & Bauckhage, 2012). At the same time, more and more cities are setting up surveillance systems based on video-protection cameras (Porikli et al., 2013). For some while, these surveillance systems were monitored by human agents. But this solution quickly turned out to be inefficient, error-prone, and overwhelming.

In the last decades, we have witnessed the emergence of smart cities. A smart city implies the use of technology to enhance the well-being of urban citizens. The rise of this concept is on par with the use of intelligent surveillance systems that substitutes the massive intervention of human agents with algorithms. These algorithms are part of crowd analysis. Within the field of computer vision, crowd analysis is gaining more and more interest. Understanding the crowd mechanisms, that explain what could endanger massive gatherings is of utmost concern for security forces. Many studies have been conducted to understand human and crowd behavior within a crowded scene (Walia & Kapoor, 2016). According to many surveys (Grant & Flynn, 2017; Lamba & Nain, 2017; Zhan et al., 2008), crowd analysis is subdivided into two research axes: crowd statistics and crowd behavior analysis. The purpose of crowd statistics is to estimate crowd density by the means of crowd counting methods. The most suitable metric used to evaluate crowd density is a metric taken from vehicular traffic flow domain: the Level of Service (LOS) of a crowd (Grant & Flynn, 2017). The purpose of crowd behavior analysis is to study the behavior of a crowd. This field is commonly subdivided into two main sub-fields: crowd tracking and activity analysis (Grant & Flynn, 2017; Lamba & Nain, 2017).

Recent works seized the importance of shifting the interest from crowd behavior analysis to crowd motion and behavior detection and forecasting (Li, Chang, Wang, Ni, & Hong, 2015; Thida, Yong, Climent-Pérez, Eng, & Remagnino, 2013). Hence, dividing crowd behavior analysis into two subtopics: trajectory analysis, and crowd action recognition. With the recent upsurge in the use of Deep Learning methods in computer vision and natural language processing, we have witnessed the prediction capabilities that are offered by this category of Machine Learning methods. Before delving into the presentation of the different categories of methods developed in the field of crowd analysis, we are going to establish a comparative study of some previous reviews on crowd analysis.

In Section 4, we mention mostly recent works in crowd analysis, that were published in these last 4 years. Furthermore, we focus on the following subjects:

* Corresponding author.
E-mail addresses: mounir.bendali-braham@uha.fr (M. Bendali-Braham), jonathan.weber@uha.fr (J. Weber), germain.forestier@uha.fr (G. Forestier), lhassane.idoumghar@uha.fr (L. Idoumghar), pierre-alain.muller@uha.fr (P.-A. Muller).

https://doi.org/10.1016/j.mlwa.2021.100023
Received 30 October 2020; Received in revised form 11 January 2021; Accepted 7 February 2021
Available online 27 February 2021

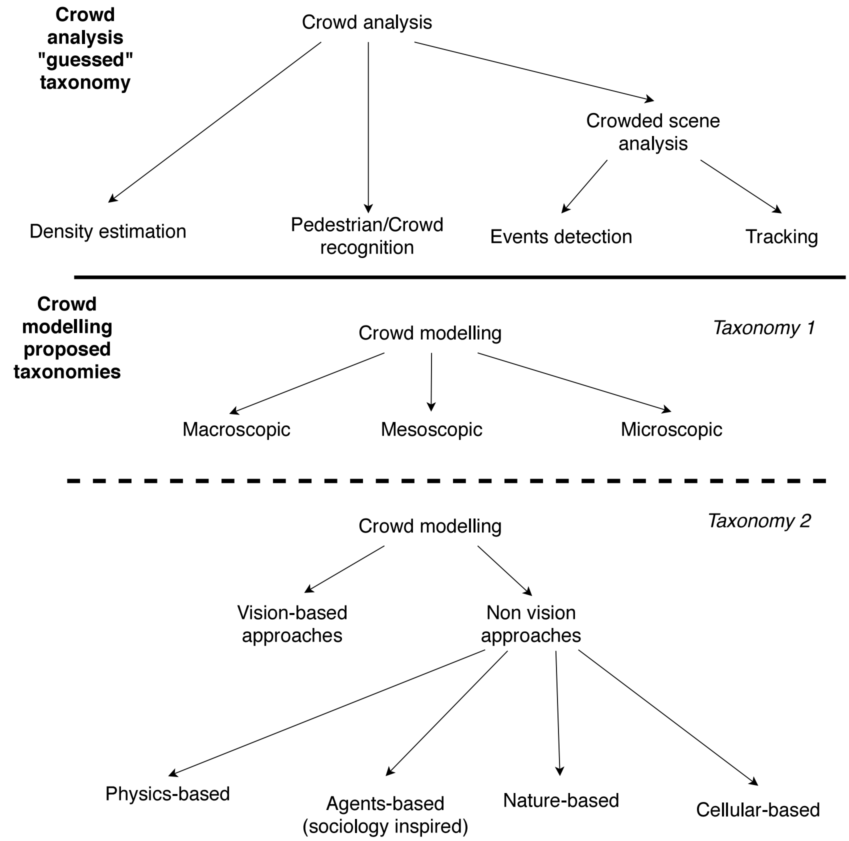


Fig. 1. Taxonomies guessed from Zhan, Monekosso, Remagnino, Velastin, and Xu (2008)'s review.

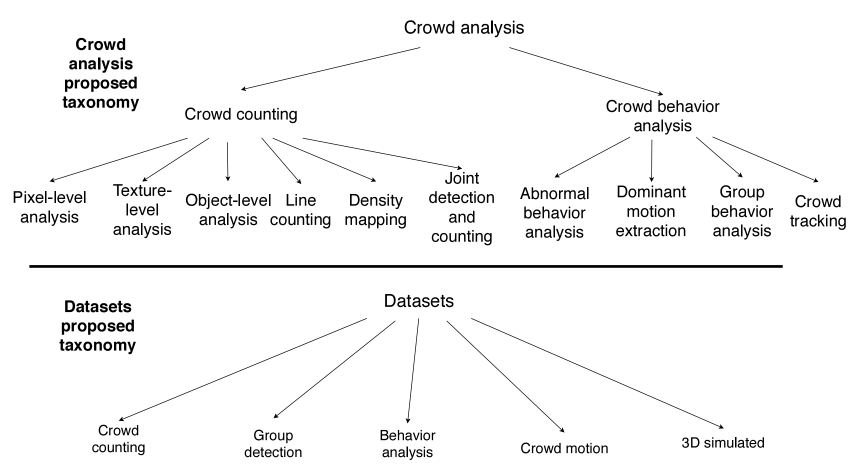


Fig. 2. Taxonomies proposed by Grant and Flynn (2017).

- The use of deep neural networks in crowd analysis, without ignoring some very recent works that still rely on the extraction of hand-crafted features and do not use deep learning methods;
- As Deep Learning is ubiquitous in computer vision, we do not dedicate a section to feature engineering, but we talk about the detection of frequent important scene elements like pedestrians and groups;
- We point out the subtopics of crowd analysis that lack attention in the literature;
- Exploration of multiple sources of data: may they be from live video-surveillance or from private/public datasets;
- And finally, the use of annotators. Their contribution is required to enrich existing datasets and to create new ones. Many subareas of crowd analysis are still at their genesis due to data scarcity.

The remainder of this paper is organized as follows: Section 2 gathers previous reviews on crowd analysis. Many of them adopt a particular taxonomy to describe the research axis. Some reviews do not present crowd analysis as a whole but analyze one of its subtopics. Hence, we divided the section into subsections according to the topic/subtopic that is presented. Section 3 is dedicated to pedestrian and group detection. We did not include these studies within crowd analysis (Section 4), because we consider pedestrian and group detection as essential tools for crowd analysis but not a subtopic itself. Section 4 presents the part of the recent literature of crowd analysis. The section is broken down into two major subtopics: crowd statistics and crowd behavior analysis. Many of the mentioned studies are Deep Learning-based. Section 5 presents all the sources of data we came

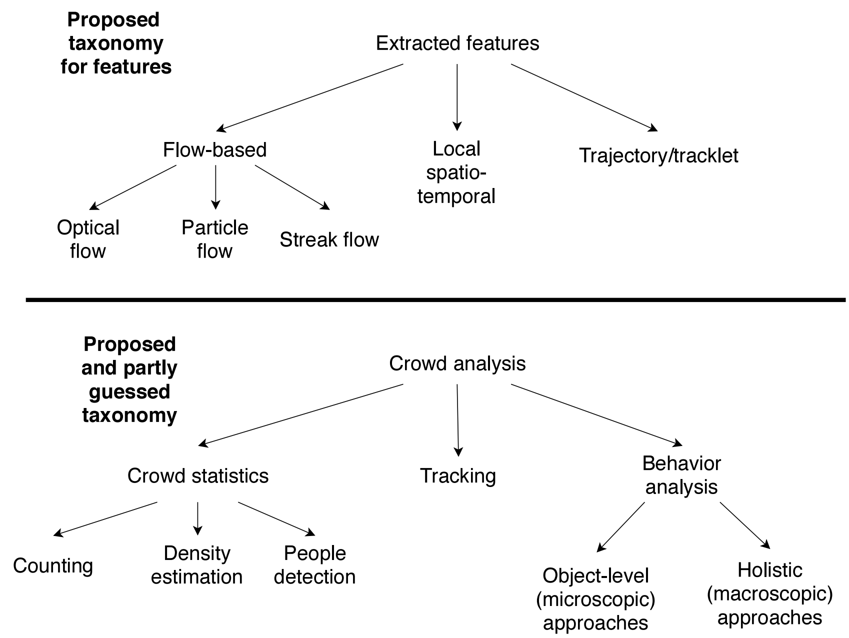


Fig. 3. Taxonomies proposed and guessed from Lamba and Nain (2017)'s review.

across. Some of these datasets are used for pedestrian and group detection. Section 6 is dedicated to the description of some annotators.

2. Previous reviews on crowd analysis

Since the last decade, many surveys have been written on crowd analysis. Some of them study the research axis as a whole, and others focus on one of its subtopics. In this section, we present the most influential reviews we came across and those which attracted our attention. To establish a solid bedrock for our future work, we have studied a dozen of them and organized them into subsections. Each of the following surveys adopts from the beginning a taxonomy to categorize their reviewed studies. Some taxonomies are redundant, and others are unique.

2.1. Crowd analysis as a whole

The reviews we mention in this subsection aim to present a panoramic view of the crowd analysis domain.

Zhan et al. (2008)'s survey on crowd analysis offers an interesting overview of the shape of the literature by the end of 2008. The survey provides a general idea of what was done in crowd analysis ranging from crowd statistics to crowd behavior analysis. It starts by providing an idea about feature engineering for each of crowd density measurement, crowd counting, action recognition and tracking. After that, it gives an insight about three different taxonomies adopted in crowd analysis: the dichotomy between crowd statistics and crowd behavior analysis, the division into macroscopic, microscopic and mesoscopic studies (proposed by the Federal Highway Administration, in the first place (FHWA, 2004), and the division into computer vision-based studies, physics-based, sociology-inspired and biology-based approaches. A visualization of these taxonomies is proposed in Fig. 1. Zhan et al. (2008)'s survey does not mention the datasets that were used for crowd analysis.

Grant and Flynn (2017) study crowd analysis and divide it into two broad categories: crowd counting and crowd behavior analysis. A visualization of the taxonomy, that they propose, can be observed in Fig. 2. The authors show that former studies on human activity recognition focused on individual scenes. The interest for group actions or actions within a crowd came later. For crowd behavior analysis, the review mentions works on group analysis, the detection of abnormal events, and crowd motion. For crowd statistics, the authors evoke the use of a measure used in traffic flow (TRB, 2000), to estimate crowd's density, the Level of Service (LoS). As highlighted by the authors, the reviewed works do not tackle many challenging problems, for example:

- they focus mainly on small crowds;
- the images on which they work are easy to process;
- the studies do not tackle demographics-related issues;
- and, these studies do not intend to tackle activity or behavior-related questions.

One finding of this review is that few research is held on identifying dangerous crowd environments because of data scarcity. A mention is given to some popular datasets used for crowd analysis. However, as stressed by Tripathi, Singh, and Vishwakarma (2018), despite the ubiquity of Deep Learning, few of the mentioned studies are Deep Learning-based.

Lamba and Nain (2017) start by highlighting the inability of hand-crafted methods to model crowd dynamics because of occlusion and cluttered scenes. They evoke in their review the features that are used for crowded scene analysis and categorized them in flow-based, local spatio-temporal features, and trajectory/tracklets. They classified the reviewed studies in crowd counting, people tracking and crowd behavior analysis. We can visualize the categories of the studies that were discussed in this review in Fig. 3.

Tripathi et al. (2018) regret that recent reviews in crowd analysis have neglected Deep Learning-based works despite their ubiquity in every computer vision sub-domain. The review includes an analysis on over one hundred studies involving Convolutional Neural Networks (CNNs). They split these studies into four categories:

1. Works relying on the variation of the number of layers and the input fed to the Neural Network;
2. Works relying on cascading a variety of CNNs and the fusion of the classification decisions;
3. Works using CNNs for feature extraction and applying state-of-the-art classifiers;
4. Works using CNNs incorporated with other deep learning architectures in order to increase the overall performance.

Following the shape of previous taxonomies, the authors split the literature of crowd analysis into four subareas, as we can observe from

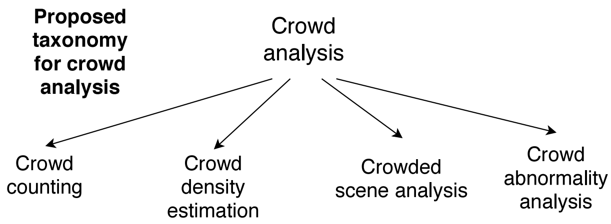


Fig. 4. Taxonomy proposed by Tripathi et al. (2018).

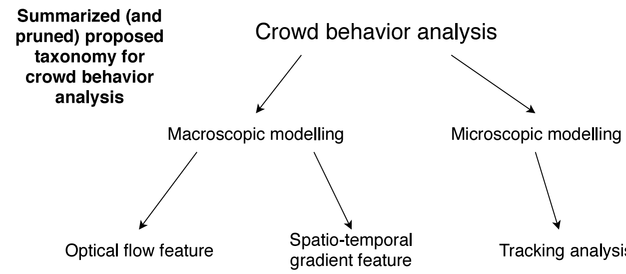


Fig. 5. Taxonomies found in Thida et al. (2013)'s review.

Fig. 4. The authors highlight the challenges that are still faced by Deep Learning-based methods. Among these challenges are the lack of labeled data and the need for powerful Graphics Processing Units (GPUs) to train models. As a workaround to this issue, they suggest the use of transfer learning. However, the authors did not mention any perspective on how the literature should evolve.

In this subsection, we have seen four reviews that tackle crowd analysis while evoking other sub-parts of the field. Contrary to Zhan et al. (2008), the other reviews were written very recently and after that Deep Learning had been used in many fields of computer vision, including crowd analysis. However, only Tripathi et al. (2018) focused its review on the application of Deep Learning to crowd analysis. Contrary to others, Tripathi et al. do not propose many taxonomies. For instance, Zhan et al. propose three taxonomies to categorize each work connected with crowd analysis which allow us to observe the field from several points of view. Moreover, while Grant and Flynn (2017) propose a taxonomy for crowd analysis and for its datasets, Lamba et al. propose a taxonomy for crowd analysis and feature extraction.

2.2. Crowd behavior analysis

In this subsection, the reviews analyze the part of the literature dedicated to crowd behavior analysis, a sub-domain of crowd analysis.

Contrary to one of Zhan et al. (2008)'s taxonomies, Thida et al. (2013) split the literature into two broad categories: microscopic and macroscopic approaches. The authors show how the crowd is observed through these two points of view.

- From a microscopic point of view, bottom-up approaches are privileged. We start by pedestrian detection, continue with tracking, and finish with activity analysis. The difficulties it faces are: occlusions, events' complexity due to the multiplication of interactions, etc.
- From a macroscopic point of view, top-down approaches are privileged. These latter consider a crowd as a single entity. Top-down approaches can often face obstacles when the crowd is unstructured, which means that people move anarchically. In this context, it is difficult to find regular patterns.

The authors point out that a big part of the studies carried out in crowd behavior analysis is intended for events detection, and especially abnormal events detection. However, the definition of abnormality is not unanimous, sometimes associated to rarity, some definitions link it to unobserved events. Macroscopic approaches rely on the holistic properties of a scene. Macroscopic modeling either uses optical flow features or spatio-temporal features. Conversely, microscopic approaches, are commonly agent-based, and analyze moving entities in a scene. The authors mention several tracking bottom-up approaches linked with the Particle Filter framework. This framework is mostly based on color cues, but the authors mention several studies where it is paired with other cues. They evoke the possibility to enhance the quality of tracking using crowd-level cues such as contextual information and social interactions. As a solution to occlusion, the authors propose the use of multiple cameras. We can visualize the taxonomy proposed for crowd behavior analysis in Fig. 5. The review ends on a section dedicated to event detection in crowded scenes, where anomaly detection is briefly evoked. However, we regret that this section was not included in the proposed taxonomy.

Li et al. (2015) start by summarizing the basic concepts of crowd behavior analysis. They show how the crowd is perceived by Crowd Dynamics and how it is by Computer Vision. While Crowd Dynamics consider a crowd as either a fluid by proposing continuum-based approaches inspired from statistical mechanics and thermodynamics, or as a set of individuals satisfying agent-based approaches like the Social Force Model (SFM) (Helbing & Molnar, 1995), Computer Vision adopts the macroscopic and microscopic scales to observe a crowd. Some mesoscopic approaches rely on the features provided by the two scales to yield a better analysis. The authors highlight the importance of feature engineering to model crowd dynamics. They rank the features into three levels based on their degree of expressiveness: 1. low-level flow-based features, 2. mid-level local spatio-temporal features, and, 3. high-level trajectory/tracklet features. As observed from Fig. 6, the reviewed studies are split into three categories: motion pattern segmentation, behavior recognition, and anomaly detection. This review does not evoke pedestrian and group detection, and does not tackle group analysis.

2.3. Abnormal behavior analysis

Abnormal behavior analysis is a bubbling subtopic in crowd analysis.

Chong and Tay (2015) highly recommend the use of Deep Learning to find anomalies in videos. Anomaly detection in videos is an arduous task. The difficulties of this task come from the video's resolution and from the variety of changes that can happen within the video ranging from human movements to environmental variations.

Traditionally, handcrafted methods for anomaly detection are pipelines that include several steps: video pre-processing, feature engineering, context modeling, and finally classification/clustering. The engineered features are flow-based or trajectory-based. The engineered features of these handcrafted methods need to be manually tuned each time the environment changes. Deep Learning methods alleviate this need, especially when it is possible to use transfer learning. The authors highlight the fact that videos cannot be directly fed to a classifier because of action pattern variations, environmental variations and clutter. Due to these inconveniences, the authors discuss how to model a good representation for videos. They start by providing an overview about conventional feature descriptors before talking about Deep Learning-based features extractors. They classify these features extractors into three groups: 1. Conditional restricted Boltzmann machine and Space–time deep belief network, 2. Independent component analysis and its variants, and 3. Deeply learned slow feature analysis and gated models.

The authors mention the works of Taylor et al. and Chen to introduce respectively Convolutional Restricted Boltzmann machine (CRBM) and Space–Time Deep Belief Network (ST-DBN). Restricted Boltzmann machine (RBM), from which the CRBM are derived, are used to model multivariate time-series data. ST-DBN comprises multiple stacks of two

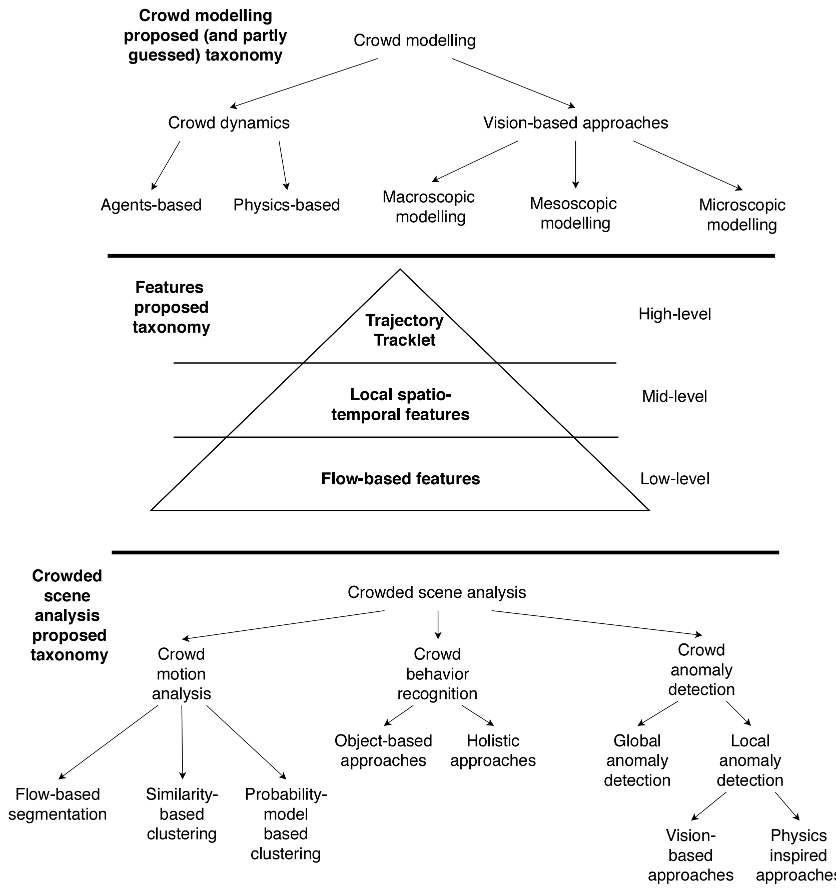


Fig. 6. Taxonomies found and guessed from Li et al. (2015)'s review.

layers of CRBMs: a layer of spatial CRBMs and a layer of temporal CRBMs. Both of these networks are resilient to spatial transformations and thus achieve good invariance. ST-DBN has also the particularity to fill in missing video data. Because of the necessity to determine the number of latent features representing a video and the fact that such a task can sometimes become burdensome, authors highlighted the advantages of using ICA-like methods (ICA stands for Independent Component Analysis), such as the one illustrated in Chatzis and Kosmopoulos (2015)'s work and which proposes a convolutional model using ICA as building blocks.

Last but not least, the authors highlight the similar performances of Deeply Learned Slow Feature Analysis (DL-SFA) (Sun et al., 2014) and Hierarchical Independent Subspace Analysis (ISA) (Le, Zou, Yeung, & Ng, 2011). They found a slight superiority for ISA, however amortized because this latter leverages dense sampling. They mention the gated models such as Gated-Restricted Boltzmann machine (GRBM) (Taylor, Fergus, LeCun, & Bregler, 2010), that are good for capturing image transformation and spatial information. When they compare the methods on KTH and Hollywood2 datasets, the authors used as baseline handcrafted methods: HOG3D (Histogram of Oriented Gradients 3D) and HOG/HOF (Histogram of Oriented Gradients/Histogram of Optical Flow). It is worth mentioning that HOG/HOF+Mining and dense trajectories are state-of-the-art feature extractors for both datasets. We can visualize the categories of feature engineering and feature extraction explored in this review in Fig. 7.

Apart from KTH and Hollywood2, the review does not propose a set of datasets that can be used for anomaly detection in videos of crowded scenes.

By putting forth the fact that anomaly detection is a challenging domain within the unsupervised learning research area, Kiran, Thomas, and Parakkal (2018) talk mostly about anomaly detection and evoke anomaly prediction. They highlight the issue arising from the paucity of annotated data despite the availability of raw video data. They define anomaly as the detection of unseen objects and infrequent events.

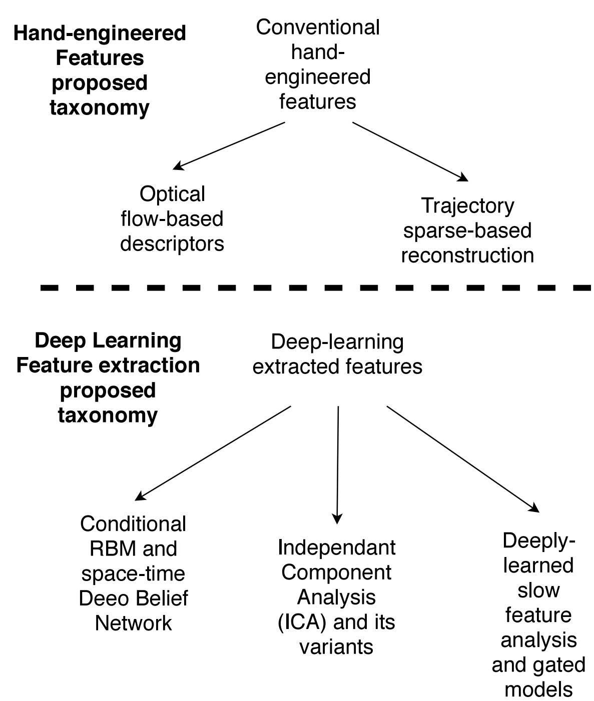


Fig. 7. Taxonomies proposed by Chong and Tay (2015).

The survey mentions six datasets among the frequently uses ones: UCSD Anomaly detection, CUHK Crowd, and UMN Social Force that we describe later in the datasets Section 5.2. It introduces less used datasets, that we do not speak about later, like the Subway entry and exit datasets, The Train dataset, The Queen Mary University of London U-turn, and a newly introduced dataset, the lv dataset, by Leyva, Sanchez, and Li (2017). Starting from the premise that normality is defined by a static background, normal crowd appearance, no change in trajectory, etc., the mentioned datasets comprise videos where the background is still.

The survey reviews deep learning methods used for unsupervised and semi-supervised anomaly detection in videos. It classifies them in terms of the model's type and detection criteria. More precisely, it explores three types of statistical models: reconstruction models, predictive models, and deep generative models.

- Reconstruction models comprise Principal Component Analysis (PCA), Autoencoders, Convolutional AutoEncoders, Contractive Autoencoders and other deep models like SDAEs (Stacked Denoising AutoEncoders) and DBNs (Deep Belief Nets).
- Predictive models comprise composite LSTM model that performs reconstruction and prediction, convolutional LSTM which is also a composite LSTM, 3D-AutoEncoder and Predictor, Slow Feature Analysis (SFA) that is calculated using batch PCA iterated twice.
- Deep generative models comprise Variational AutoEncoders (VAEs), Generative Adversarial Networks (GANs), Adversarial AutoEncoders (AAEs).

We depict the categories of the statistical models explored by this review in Fig. 8.

The authors discuss the experiments they undertook on CUHK Crowd and UCSD datasets to compare the models. The performance metrics used are Precision–Recall (PR) and Receiver-Operator-Characteristics (ROC) curves. We regret that the chosen datasets are not intended for anomaly detection in massively crowded scenes.

2.4. Motion tracking

Motion tracking is a subtopic of crowd behavior analysis. In this subsection, we mention the comprehensive review of Walia and Kapoor (2016).

Walia et al. sheds light on the drawbacks of single cue tracking methods. These methods are not suited for real-world applications. Their alternatives are multi-cue tracking methods. The review classifies these latter in compliance with the source of their cues: single-modal sensors, or multi-modal sensors. A visualization of this classification can be observed in Fig. 9.

For single-modal multi-cue object tracking methods, there are vision and InfraRed/thermal-based methods. Vision and IR/thermal sensors are a rich source of information that can yield complementary cues. Among the cues that can result from these two sources are shape, texture, color, intensity, position, motion and orientation. However, most of the studies reviewed are vision-based. Part of them are deterministic, and the other part is about stochastic methods. Despite the adaptability of Thermal/IR sensors for night vision applications, they frequently need completion from other sensors, especially when they cannot distinguish targets with similar thermal profile. Moreover, data from Thermal/IR sensors do not often come from video-surveillance cameras.

For multi-modal multi-cue object tracking methods, one of the widespread combinations that the authors highlight is vision camera with an additional sensor. The following is a list of possible combinations:

- Vision sensor paired with Thermal/IR sensor. This combination is required when tracking is done during nighttime. However data fusion and multi-sensors' calibration are still challenging.
- Combination of vision and audio sensors. This combination usually happens during lectures and meetings. However, most of the reviewed studies are limited to coherent audio sources. Chen et al. (2014) showed the negative impact of non coherent audio sources on image features, and consequently recommend evaluating both types of audio sources in real world tracking situations.
- Combination of vision and laser sensors. The use of this kind of combination is found in tracking for video-surveillance applications. The use of laser scanner is motivated by the advantages it offers such as low computational requirement, insensitivity to environmental changes and the easy laser data projection to rectangular coordinates. However, according to Cui, Zha, Zhao, and Shibasaki (2008)'s and Song et al. (2013)'s works, laser scanner suffers from some drawbacks due to clutter and occlusion.
- The combination of radio and vision sensors. Low cost radar are more and more available, and radial information provides accurate object identification.
- Stereo vision was also investigated for its ability to capture object gesture and its resilience to illumination changes.

As for single modal methods, the authors have pointed out the necessity to enhance the quality of information fusion and sensors and cues calibration. In the study, the authors presented a series of datasets that match with the previously introduced works.

The authors discuss the performance measures used to measure the robustness and efficiency of several trackers. They mention two cases: the presence of Ground Truth (GT) data and its absence and how to deal with it. Measuring the tracker's performance while GT is missing is not a cakewalk, but the authors mention several workarounds previous studies had used to overcome this obstacle. They evoke two ways to deal with the absence of GT data: 1. the detection of abnormality raising from normality (Chau, Bremond, & Thonnat, 2009; Spampinato, Palazzo, & Giordano, 2012), 2. or by using prior knowledge about an object trajectory (Wu, Sankaranarayanan, & Chellappa, 2010). Although these measures are easy to compute in real time, they are errors' prone and their reliability decreases when it faces context-dependent features. In parallel, performance measures using GT data benefit from the advent of PETS workshops.

The authors mention both qualitative and quantitative measures proposed from several previous works. These measures show good precision and are independent from the used tracker. However, despite the widespread use of Deep Learning (DL) methods for multi-object tracking, Walia et al. do not mention any DL-related work.

2.5. Group behavior analysis

Group behavior analysis is a subtopic of crowd behavior analysis. We highlight the recent trends in group behavior analysis through the review of Borja-Borja, Saval-Calvo, and Azorin-Lopez (2017).

Borja et al. associate the nature of a human action to the number of individuals that perform it. They mention the classification used in other works (Azorin-Lopez, Saval-Calvo, Fuster-Guillo, Garcia-Rodriguez, & Orts-Escolano, 2015; Chaaraoui, Climent-Pérez, & Flórez-Revuelta, 2012) that identify and classify human actions in function of their duration. Both of these classifications are depicted into two level pyramids in Fig. 10.

- The first classification breaks down the behaviors into four levels depending on their duration and the number of persons involved in them: gestures, actions, interaction, and group activity.
- The second classification divides the behaviors into four levels depending on their duration only: Motion, Action, Activity and Behavior.

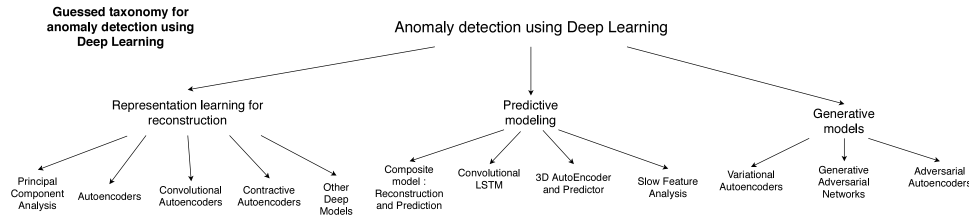


Fig. 8. Taxonomy guessed from Kiran et al. (2018)'s review.

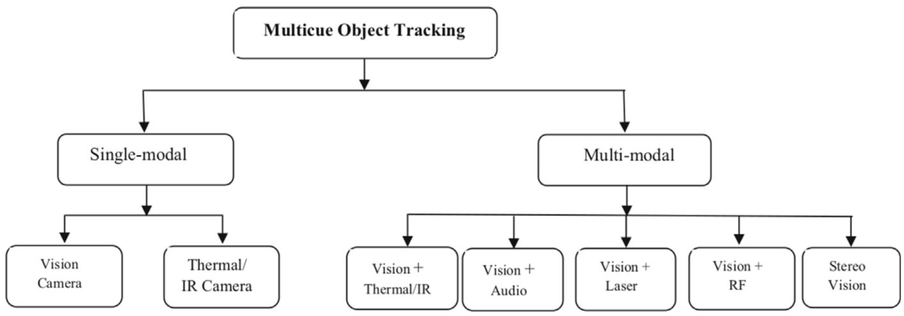


Fig. 9. Illustration of the taxonomy of Multicue Object Tracking proposed by Walia and Kapoor (2016) and taken from its review paper.

In the second part of the review, the authors describe datasets used for group activity recognition such as: BEHAVE, CAVIAR, CVBASE, ETISEO, ETH, UHD, HMDB, SportsVU, PETS, ViF. Using the Group Activity Descriptor Vector (GADV), proposed by Azorin-Lopez et al. (2016), the authors classify the group behaviors by taking into account the number of individuals involved in each of them. This descriptor is obtained after extracting features from trajectories using Neural Networks. Afterwards, the authors present the features used for several group behavior recognition tasks: anomaly detection when this anomaly is caused by a small group within a crowd; the distinction between a normal and an abnormal behavior based on the assumption that normal behaviors last longer; group's characteristics such as the distance between its members, and the velocity of each member. Although the review offers a general view of what is done for group behavior analysis, no mention is given to group detection tasks.

2.6. Conclusion and discussion

To sum up the taxonomies proposed in the previous reviews, we came up with a synthesized taxonomy illustrated in Fig. 11. Through this diagram, that is mainly inspired from the conclusions of the previous reviews, we describe our perception of the shape of the literature. As stated before, crowd statistics is divided into crowd counting and density estimation. Even if these two subjects are intertwined, density estimation is frequently related to crowd management and can be used by security forces to guess when a place is overcrowded and may represent a danger for its population, while crowd counting may be useful for entities that need to measure their audience for statistical purposes. Crowd behavior analysis can be divided into three broad subtopics:

- Crowd behavior recognition/classification, a topic that is slightly linked with Action Recognition in individual scenes, as we will see subsequently in Section 4.2.1;
- Motion analysis, that includes trajectory analysis and prediction;
- and Group behavior analysis, that often require group detection and is related to action recognition for groups within a crowded scene.

In our taxonomy, we do not mention straightforwardly the division into microscopic (Lagrangian)-mesoscopic-macroscopic (Eulerian) approaches evoked by the literature (Allain, Courty, & Corpetti, 2012; Thida et al., 2013; Wang, Cheng, & Wang, 2018; Zhan et al., 2008), because on one hand this division maps the branches of the taxonomy we propose of crowded scenes analysis:

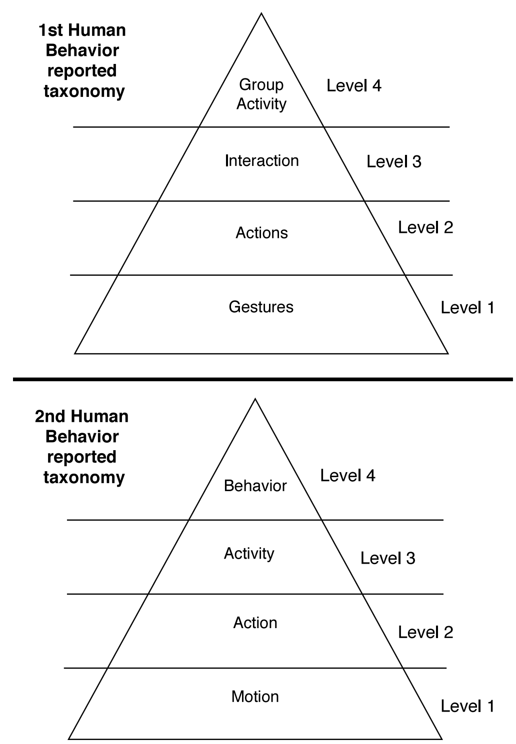


Fig. 10. Taxonomies about levels of Human Behavior found in Borja-Borja et al. (2017)'s review. First upper taxonomy taken from Vishwakarma and Agrawal (2013)'s survey, and second lower taxonomy taken from Chaaraoui et al. (2012)'s review.

- Motion Analysis is frequently associated to object-based microscopic approaches,
- while Crowd Behavior Recognition stems from a holistic analysis of a crowd.
- Group Behavior Analysis covers the in-between mesoscopic approaches.

On an other hand, we observe that crowd statistics can obey to the dichotomy Micro/Macro. If we consider crowd counting as a microscopic approach for crowd statistics, because it is object-based, we can consider density estimation as a macroscopic approach, because it sees the crowd as a whole.

In the taxonomy we propose, we prefer to avoid considering anomaly detection as a sub-branch of crowd analysis, because each sub-field of crowd analysis, that is included in our taxonomy, can cover works related to anomaly detection.

Since 2012, following the prowess of Deep Learning and more precisely Convolutional Neural Networks in Image Classification, with the development of the AlexNet model (Krizhevsky, Sutskever, & Hinton, 2012), the use of Deep Learning in computer vision has skyrocketed. Hence, more and more approaches are adopting Deep Learning in crowd analysis. The unique review which stated the need to focus solely on reviewing Deep Learning methods for crowd analysis is the recent Tripathi et al. (2018)'s review. However, because of the lack of datasets, especially for group behavior analysis and crowd behavior analysis from massively crowded scenes, literature in these sub-domains still needs generic models generated from recent Deep Learning architectures.

3. Pedestrian and group detection

Before analyzing crowd motion or behavior, it is often necessary to detect the components of a crowd. Those components are pedestrians, and groups formed by those pedestrians.

Before the advent of Deep-Learning based methods in crowd analysis, a lot of works used hand-crafted methods to extract features and fed those features to Machine-Learning approaches to detect pedestrians and group. In this section, we start by mentioning a less recent, still very well cited, work that uses hand-crafted methods for pedestrian and group detection. After that, we review and compare the works that have been done more recently in these two fields.

Ge et al. (2012) propose a bottom-up agglomerative hierarchical clustering approach to detect small groups of individuals. They handle two types of video-recorded crowded scenes. Scenes recorded from an elevated camera where individuals appear tiny, and scenes recorded from a video-surveillance camera level with a higher resolution, where individuals can be clearly identified.

To detect pedestrians from elevated cameras, the authors use the Reversible Jump Markov Chain Monte Carlo (RJMCMC) method. For higher resolution videos, they use a modified version of the HoG detector (Dalal & Triggs, 2005). After the detection process, they get tracklets using the Sampling Importance Resampling (SIR) particle Filter (Doucet & Johansen, 2009). The tracklets are assembled in sets of tracklets. Sliding window is used to find tracklets that overlap and that constitutes potential candidates for longer trajectories, which allows the authors to yield sets of trajectories. New tracklets are assigned to the right trajectory using the Hungarian Algorithm (Kuhn, 1955). If in a trajectory the authors face missed locations, they infer them through linear interpolation. From a set of trajectories, they combine the sliding window strategy to hierarchical clustering to find small groups of people. After that, to measure the proximity between small groups, they use the symmetric Hausdorff distance (a distance already used for trajectory analysis (Wang, Tieu, & Grimson, 2006). The authors undertook their experiments on a self-made dataset of 5 video sequences: SU1, SU2, Artefest, Stadium1, Stadium2. In their experiments, they outperform, in terms of groups detection accuracy, the methods of Corner Clustering (Rabaud & Belongie, 2006), and another category of methods: (Brostow & Cipolla, 2006; Sugimura, Kitani, Okabe, Sato, & Sugimoto, 2009).

3.1. Pedestrian detection

In real world situations frequently observed by video-surveillance applications, pedestrian detection is not a simple task. The scenes are often cluttered and detectors face several types of occlusions. While the hand-crafted methods often fail to detect pedestrians in challenging situations, Deep Learning models achieve frequently impressive results. When we develop a pedestrian detector, the utmost purpose is to create a highly accurate and quick detector that can run on systems with affordable computing power (Angelova et al., 2015).

Li, Wu et al. (2016) train a Region proposal multi-layered Convolutional Neural Network (RCNN), on Pedestrian Detection. Instead of the classic sliding window, the authors use the Edge Boxes algorithm (Zitnick & Dollár, 2014) for region proposals. After that, they use a Support Vector Machine (SVM) to classify the obtained features to identify pedestrians. The RCNN is trained and tested on the INRIA pedestrian dataset. During the training procedure, they fine-tuned the AlexNet model (Krizhevsky et al., 2012) that was trained on the ILSVRC2012 datasets. To avoid over-fitting, the authors expanded the INRIA dataset with comparable pedestrian data from Wang, Shi, Song, and Shen (2007)'s work. During the testing phase, they evaluated a model trained on the non-extended INRIA dataset, and a model relying on the Selective Search (Uijlings, Van De Sande, Gevers, & Smeulders, 2013). The authors' model outperforms both of them. Moreover, it realizes better results than the Histogram of Diagrams (HOG) features and Viola Jones methods. In other words, the model achieves a performance of 10% of false detection rate and 23% of missing rate, which is 23% higher than the HOG features method. The developed method and its variants were compared to handcrafted methods. However, despite the ubiquity of Deep Learning methods in this field, we do not see a comparison of the authors' method with other Deep Learning models.

Tian et al. (2015) propose DeepParts, a part-based pedestrian detector that is presented as an occlusion handler. The core idea of the algorithm is the extensive part pool that contains several scales and positions of pedestrians' body parts, and the data-driven parts selection for occlusion handling. The following properties characterize the algorithm: ability to be trained on weakly labeled data; handles low Intersection over Union (IoU) positive proposals that shift away from ground truth; each part detector can detect a pedestrian by visualizing only a part of a proposal. Training and testing were held on the Caltech dataset. During experiments held on the same dataset, an overall model of DeepParts reached a miss rate score of 11.89%. It was also tested on KITTI dataset, despite being trained on Caltech, and it achieved an Average Precision (AP) score of 58.67%, on a moderate subset, but seems outperformed by Regionlets. As a perspective, the authors propose to treat DeepParts as a cascade stack over other detectors to increase performance. They also suggest model compression, by including all part detectors within a unique Convolutional Neural Network (CNN).

Inspired by the successes brought by the development of Deep Learning-based accurate detectors (Luo, Tian, Wang, & Tang, 2014) and by the remarkable high speed of very fast cascades (Benenson, Mathias, Timofte, & Van Gool, 2012) when processing image patches, Angelova et al. (2015) built DeepCascade, a combination of very fast cascades with deep neural networks. Specifically, they did not completely copy Benenson's cascades, but modified their constitution to keep only 10% of their stages. To obtain a good model they used a pretrained Deep Neural Network (DNN) on ImageNet dataset. Besides, they used data augmentation. Moreover, to select the best model, they trained their algorithm on three datasets producing three different models: the Caltech Pedestrian dataset, an independently collected dataset, and an extra dataset containing Caltech, ETH and Daimler. The most well performing model resulted from the training on the extra dataset. Tested on Nvidia Tesla K20 GPU, the runtime of their cascade is of 15 Frames per Second (FPS) on 640×480 pixels images. Furthermore, it achieved very good results regarding accuracy; they reached the average miss rate score of 26.2% on the Caltech Dataset. To improve their model,

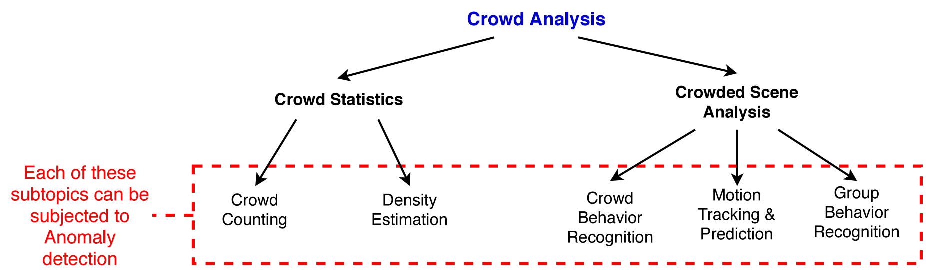


Fig. 11. Proposed taxonomy for Crowd Analysis.

Table 1
Summarized presentation of studies in pedestrian detection and group detection. **"?DL" column:** DL stands for Deep Learning, this column means if the paper used Deep Learning approaches (DL) or not (¬DL).

Reference	Date	Research axis	Used datasets	?DL	Source code
Ge, Collins, and Ruback (2012)	2012	Pedestrian & Group detection	Self-made	¬DL	Unavailable
Li, Wu, and Zhang (2016)	2016	Pedestrian detection	INRIA Person	DL	Unavailable
Tian, Luo, Wang, and Tang (2015)	2015	Pedestrian detection	Caltech, KITTI	DL	Unavailable
Angelova, Krizhevsky, Vanhoucke, Ogale, and Ferguson (2015)	2015	Pedestrian detection	Caltech	DL	Available
Shao, Dong, and Zhao (2018)	2018	Group detection	Student003, GVEII, MPT-20X100	¬DL	Unavailable
Wang et al. (2018)	2018	Group detection	CUHK Crowd	¬DL	Unavailable
Voon, Mustapha, Affendey, and Khalid (2019)	2019	Group detection	CUHK Crowd	¬DL	Unavailable

they encourage future research to include motion information from images and increase the depth of DeepCascade by adding small deep nets and investigating efficiency–accuracy compromises.

We saw recently that although Faster RCNN reached excellent performances for object detection (Ren, He, Girshick, & Sun, 2015), it is not a perfect pedestrian detector (Zhang, Lin, Liang, & He, 2016). When we choose a pedestrian detector, we have to make a difficult trade-off between speed and accuracy. Some tried to reduce the side-effects of this compromise to the extreme extent (Liu et al., 2016).

3.2. Group detection

Two kinds of approaches can be found in the literature of group detection: top-down approaches, which start from a crowded scene, and then segment it into small groups (Chen, Wang, & Li, 2017a, 2017b; Li, Chen, Nie, & Wang, 2017; Wang et al., 2018). Bottom-up approaches, which start by detecting pedestrians and then create clusters of them (Shao et al., 2018; Vascon & Bazzani, 2017; Yuan, Lu, & Wang, 2017).

Shao et al. (2018) propose a real-time bottom-up approach for small group detection that performs well on low and medium density crowded scenes. The key point of their method is the combination between the Social Force Model (SFM) (Helbing & Molnar, 1995) for goal prediction with a goal-based coherent filtering algorithm for group detection. They start by extracting trajectories based on two methods (Dollár, Appel, Belongie, & Perona, 2014; Milan, Roth, & Schindler, 2014). After that, they construct a SFM-based collision avoidance model (Karamouzas, Heil, Van Beek, & Overmars, 2009). Then, they apply the K Nearest Neighbors algorithm to find the K neighbors of a pedestrian. Finally, they use the goal-based coherent filtering to cluster pedestrians sharing the same goal. They compare their method to that of Solera, Calderara, and Cucchiara (2016), and they achieve, most of the time, a better precision with frequently comparable recall scores in a reduced computation time. They tested their method on the following datasets: student003, GVEII, and MPT-20X100.

Wang et al. (2018) propose a top-down approach to detect groups of pedestrians. It consists in crowd segmentation. The method is broken down into two steps: detection of feature points using Tomasi and Kanade (1991)'s method and the computation of their attribute value, and then, the segmentation of the crowd into many groups using the computed attribute values. The authors tested their method on the CUHK Crowd dataset on various situations: such as structured/unstructured dense/less dense crowds. The authors compare their method to that of Shao, Change Loy, and Wang (2014), where they beat them in computation time. However, they did not compare their method to other detectors on CUHK Crowd.

Voon et al. (2019) propose Collective Interaction Filtering (CIF), a top-down clustering approach based on the Expectation–Maximization algorithm that uses trajectories to detect groups. First, they start by extracting pedestrians' tracklets using the Kanade–Lucas–Tomasi (KLT) tracker (Tomasi & Kanade, 1991). Tracklets are used as inputs to CIF to find tracklets' clusters. Within a cluster, they select the pedestrian that has the longest duration and whose trajectory has a small variance to be the key person. Second, they compute the degree of connectivity of each person with the key persons. To do so, distances like Distance Connectivity (DC), Occurrence connectivity (OC), Speed correlation (SC), are computed, from the first frame to the last frame, between the key person and the other persons and are stored within an adjacency matrix. After that, they use the EM algorithm to find the persons that are close to the key persons. Finally, they use a group refinement threshold to handle crowds with various densities, structures, and degrees of occlusion. The method is evaluated on the CUHK Crowd dataset. Compared to Collective Transition (Shao, Loy, & Wang, 2017) and Coherent Filtering (Zhou, Tang, & Wang, 2012), their method obtains better results in terms of the Normalized Mutual Information (NMI) and the Rand Index (RI).

Although there is a considerable quantity of works dedicated to group detection, the topic is still neglected in the crowd analysis research area. Besides, we have not witnessed any recent work using Deep Learning in any stage of group detection. The lack of these studies may be due to data scarcity. All the explored studies from this section are briefly presented in Table 1.

4. Crowd analysis

Like it is portrayed by the tree structure in Fig. 11, the literature commonly divides crowd analysis into two major branches: crowd statistics and crowd behavior analysis (Grant & Flynn, 2017; Lamba & Nain, 2017; Zhan et al., 2008). In this section, we explore studies published in each of these two major branches. All the explored works from this section are briefly presented in Table 2.

Public areas covered by video-surveillance cameras can be the theater of various levels of people gatherings: different degrees of LoS (Level of Service). Those gatherings may be structured or unstructured. As pointed out by Thida et al. (2013), it is very easy to identify abnormal events within structured crowds. However, the task becomes

Table 2
Summarized presentation of studies in crowd analysis. **"?DL" column:** DL stands for Deep Learning, this column means if the paper used Deep Learning approaches (DL) or not (¬DL).

Reference	Date	Research axis	Used datasets	?DL	Source code
Chan, Liang, and Vasconcelos (2008)	2008	Crowd statistics, crowd counting	UCSD Anomaly Detection	¬DL	Unavailable
Shao et al. (2014)	2014	Group behavior analysis, group detection	CUHK Crowd	¬DL	Unavailable
Ali and Shah (2008)	2008	Motion tracking and prediction	Marathon-1, Marathon-1, Marathon-1, Train Station	¬DL	Unavailable
Ali and Shah (2007)	2007	Motion analysis, anomaly detection	Self-made and Inside Mecca documentary	¬DL	Available
Baccouche, Mamalet, Wolf, Garcia, and Baskurt (2011)	2011	Action recognition	KTH dataset	DL	Unavailable
Siva and Xiang (2010)	2010	Action detection	CMU and i-LIDS datasets	¬DL	Unavailable
Hassner, Itcher, and Kliper-Gross (2012)	2012	Anomaly detection	Violent Flows (ViF)	¬DL	Available
Mehran, Oyama, and Shah (2009)	2009	Anomaly detection	UMN and Web datasets	¬DL	Unavailable
Mahadevan, Li, Bhalodia, and Vasconcelos (2010)	2010	Anomaly detection	UCSD anomaly dataset	¬DL	Unavailable
Marsden, McGuinness, Little, and O'Connor (2017)	2017	Crowd statistics, anomaly detection	UMN, UCF CC 50, WWW Crowd	DL	Unavailable
Sindagi and Patel (2017)	2017	Crowd counting	UCF CC 50, ShanghaiTech	DL	Available
Tran et al. (2018)	2018	Individual-scene action recognition	Sports-1 m, Kinetics, HMDB-51, UCF-101	DL	Available
Carreira and Zisserman (2017)	2017	Individual-scene action recognition	HMDB-51, UCF-101	DL	Available
You and Jiang (2018)	2018	Crowded-scene action recognition	Self-made	DL	Unavailable
Bewley, Ge, Ott, Ramos, and Upcroft (2016)	2016	Pedestrian tracking	MOT Challenge 2015	¬DL	Available
Wojke, Bewley, and Paulus (2017)	2017	Pedestrian tracking, re-identification	MOT Challenge 2016	DL	Available
Singh, Patil, and Omkar (2018)	2018	Abnormal behavior detection	Aerial Violent Individual (AVI) "Self-made"	DL	Unavailable
Ravanbakhsh, Nabi, Mousavi, Sangineto, and Sebe (2016)	2016	Abnormal behavior detection	UCSD, UMN	DL	Unavailable
Ramos, Nedjah, de Macedo Mourelle, and Gupta (2017)	2017	Abnormal behavior detection	UMN	¬DL	Unavailable
Vahora and Chauhan (2018)	2018	Group behavior analysis	Collective activity	DL	Unavailable
Shu, Todorovic, and Zhu (2017)	2017	Group behavior analysis	Collective activity, Volleyball	DL	Unavailable
Wei et al. (2020)	2020	Crowd action recognition	CUHK Crowd, normal–abnormal crowd	DL	Unavailable
Wang and O'Sullivan (2016)	2016	Crowd action recognition, Motion analysis	Synthetic, Edinburgh (Forum), MIT Carpark, Train Station	¬DL	Unavailable
Yan, Zhu, and Yu (2019)	2019	Crowd action recognition	WorldExpo'10	DL	Unavailable
Ullah, Ullah, Conci, and De Natale (2016)	2016	Crowd action recognition	WorldExpo'10	¬DL	Unavailable
Ullah, Khan, Ullah, Cheikh, and Uzair (2019)	2019	Crowd action recognition	CUHK Crowd	DL	Unavailable
Liu, Salzmann, and Fua (2019)	2019	Crowd statistics	ShanghaiTech, WorldExpo'10, UCF Crowd Counting, UCF QNRF, Venice dataset	DL	Unavailable
Wan and Chan (2019)	2019	Crowd statistics	ShanghaiTech A/B, WorldExpo'10, UCF QNRF	DL	Unavailable
Wang, Gao, Lin, and Yuan (2019)	2019	Crowd statistics	ShanghaiTech A/B, WorldExpo'10, UCF QNRF, Grand Theft Auto 5 dataset	DL	available
Lamba and Nain (2019)	2019	Motion analysis	UCF Crowd, Collective motion, Violent Flows	¬DL	Unavailable
Li, Liu, Zheng, Han, and Li (2019)	2019	Motion analysis	Self-made (canteen scene)	¬DL	Unavailable
Wu et al. (2017)	2017	Motion analysis	UCF Crowd, CUHK Crowd	¬DL	available
Alahi et al. (2016)	2016	Motion prediction	UCY and ETH	DL	available
Bartoli, Lisanti, Seidenari, and Del Bimbo (2017)	2017	Motion prediction	UCY and Museum dataset	DL	Unavailable
Zitouni, Sluzek, and Bhaskar (2020)	2020	Group behavior analysis	PETS dataset, Parking Lot, Town Center	¬DL	Unavailable
Bisagno, Zhang, and Conci (2018)	2018	Group behavior analysis	UCY and ETH	DL	Unavailable
Singh et al. (2020)	2020	Abnormal behavior detection	UCSD Ped 1, UCSD Ped 2, and the Avenue dataset	DL	Unavailable
Qasim and Bhatti (2019)	2019	Abnormal behavior detection	UMN	¬DL	Unavailable
Hao, Xu, Liu, Wang, and Fan (2019)	2019	Abnormal behavior detection	UMN	¬DL	Unavailable
Lin, Yang, Tang, Shi, and Chen (2019)	2019	Abnormal behavior detection	UCF Crime dataset	DL	Unavailable
Xie, Zhang, and Cai (2019)	2019	Abnormal behavior detection	Self-made	¬DL	Unavailable

Table 3
Summarized presentation of live video-surveillance streams. FPS means number of frames per second. ∼ symbol used to mean at around.

Live video-surveillance	Quality	Camera recording	Can be used for	Website
UK road traffic	Good resolution, 1 frame/minute	Static, 50 views	Vehicle detection/tracking	https://trafficcameras.uk/roads/
Earthcam	Good resolution, realtime frame rate	Static	All types of crowd analysis	https://www.earthcam.com/
Live Mecca	720p	Static/dynamic	Massive crowded scene analysis	https://makkahlive.net/tv.aspx?r=14
Live Vatican	480p	Static, 1 view	Few situations of crowd scene analysis	https://www.youtube.com/watch?v=q5wc5aIpjk4
Monthey Place Centrale	Medium resolution, ∼5 FPS	Static, 2 views	Pedestrian/vehicle detection/tracking	http://www.idelec.ch/webcam/monthey/cam1 (cam2)

Table 4
Summarized presentation of datasets (Part 1). **"Type" column precisions:** Vid/Img are abbreviations of respectively Video and Image. **"Sensor" column precisions:** MonoCam/MultiCam are contractions of respectively mono-camera and multi-camera setups. Several means that there are several types of sensors involved in the recording. **"Representative quantities" column precisions:** ped is an abbreviation of pedestrian. GT are the initials of Ground Truth, and can refer to annotations. BBxs are the initials of Bounding-Boxes. 1 m means 1 million. 1K means 1000. ~ means approximately or at average. OF stands for Optical Flow.

Dataset	Release date	Type	Sensor	Representative quantities
Aerial Violent Individual (AVI)	2018	Img	Drone	2000 Img's, 5124 violent actions, 10863 ped's, 2–10 ped's/frame
Town Center	2009	Vid	MonoCam	1 Vid, ~16 ped's/frame, 71500 labeled head locations
I-Lids AVSS 2007 (Task 1)	2007	Vid	MonoCam	3 Vid's, 9718 annotated regions
UCSD Anomaly Detection	(Updated 2013)	Vid	MonoCam	98 Vid's, each vid contains ~200 frames
UCF Crowd 50	2013	Img	MultiCam	50 Img's, ~1280 ped's per picture, 63705 GT
UMN SocialForce	2009	Vid	MonoCam	11 Vid's on 3 scenes.
Web Dataset	2009	Vid	MonoCam	20 Vid's, 8 for abnormal events and 12 for normal ones
MOT Challenge	2014 (Updated 2016)	Vid	MonoCam	14 Vid's
Multi Task Crowd	2017	Img	MonoCam	100 Img's, 50/50 violent/non-violent crowds, 0 to +150 ped's/img
Agoraset (Simulated)	2012	Vid	MonoCam	7 scenarios subjected to different simulations
CUHK Crowd	(Updated 2014)	Vid	MonoCam	474 videos on 215 crowded scenes
KITTI	2013	Vid	MonoCam	180 GB data. 6 h recording
Toulouse Campus Surveillance	2018	Vid	MultiCam	50 vid's of 2 scenarios. GT for 1 scenario.
UCF-101	2012	Vid	MonoCam	13320 vid's, 27 h recording depicting 101 classes of actions
HMDB-51	2011	Vid	MonoCam	6766 annotated vid's, 51 classes of actions
PathTrack MOT	2017	Vid	MonoCam	720 sequences, 16,287 GT
INRIA Person	2005	Img	MonoCam	1805 Img's
Kinetics	2017	Vid	Several	400 actions, 400–1150 vid's for each action, ~10s duration each
Caltech Pedestrian	2009 (Updated 2018)	Vid	MonoCam	10 h recording, appearance of ~2300 ped's, 350,000 GT BBxs
Violent Flows (ViF)	2012	Vid	MonoCam	246 vid's, ~3.6s duration each
Sports-1m	2014	Vid	MonoCam	1 m vid's, ~5min 36s duration, 487 action classes, 1K–3K vid's/class
Caviar	2003 (Updated 2004)	Vid	Several	28 vid's from MonoCam, 26 vid's from MultiCam
Behave	2010	Vid	MultiCam	4 vid's, 76800 frames, ~125 ped's, 83545 GT
The Friends Meet (FM)	2012	Vid	MonoCam	53 vid's, 16286 frames, 3–16 ped's involved
MuseumVisitors	2015	Vid	MultiCam	2 scenarios
SALSA	2016	Vid	MonoCam	2 scenarios, 18 ped's, vid's lasting for 60 min
COCO	2015	Img	MonoCam	330K img's, 80 classes of detectable objects, 200K GT
Crowd-11	2017	Vid	MonoCam	6,272 clips, 11 classes, all clips labeled
Motion Emotion	2016	Vid	MonoCam	31 vids; 44.000 frames; classes: 5 motions, 6 emotions; GT provided
UCF Crime	2018	Vid	MonoCam	1900 vids, 13 classes of crimes, GT provided
CCTV Fights	2019	Vid	MonoCam	1000 vids, the 2 classes normal/abnormal, GT provided
Crowd-Flow	2018	Vid	MonoCam	10 vids, OF and trajectories GT provided
UCF crowd tracking	2008	Vid	MonoCam	3 vids, trajectories GT provided
GTA5 Crowd Counting (GCC)	2019	Images	MonoCam	15,212 images, GT provided, very diverse scenes

Table 5
Summarized presentation of datasets (Part 2). **"Quality" column precisions:** px is the abbreviation of pixels. FPS means the number of frames per second.

Dataset	Quality	Used for	Availability	Reference
Aerial Violent Individual (AVI)	Not mentioned	Violent group behavior detection/recognition	Private	Singh et al. (2018)
Town Center	1920 × 1080 px, 25 FPS	Pedestrian detection and tracking	Public	Benfold and Reid (2011)
I-Lids AVSS 2007 (Task 1)	720 × 576 px, 25 Hz	Pedestrian detection/tracking, anomaly detection	Public	AVSS (2007)
UCSD Anomaly Detection	Not mentioned	Anomaly detection	Public	UCSD (2013)
UCF Crowd 50	Not mentioned	Crowd statistics	Public	Idrees, Saleemi, Seibert, and Shah (2013)
UMN SocialForce	Not mentioned	Abnormal behavior detection/recognition	Public	Mehran et al. (2009)
Web Dataset	Not mentioned	Abnormal behavior detection/recognition	Public	Mehran et al. (2009)
MOT Challenge	Not mentioned	Group/Pedestrian detection and tracking	Public	Milan, Leal-Taixé, Reid, Roth, and Schindler (2016)
Multi Task Crowd	Not mentioned	Abnormal behavior recognition/Crowd Statistics	Private	Marsden et al. (2017)
Agoraset (Simulated)	Not mentioned	Group/Pedestrian detection/tracking/behavior analysis	Public	Allain et al. (2012)
CUHK Crowd	Not mentioned	Group/anomaly detection	Unavailable	Shao et al. (2014)
KITTI	8-bit PNG frames	Object detection/tracking	Public	Geiger, Lenz, Stiller, and Urtasun (2013)
Toulouse Campus Surveillance	1920 × 1080 px	Object detection/tracking	Public	Malon et al. (2018)
UCF-101	320 × 240 px, 25 FPS	Action detection and recognition	Public	Soomro, Zamir, and Shah (2012)
HMDB-51	240 px height, 30 FPS	Action detection and recognition	Public	Kuehne, Jhuang, Garrote, Poggio, and Serre (2011)
PathTrack MOT	Not mentioned	Pedestrian detection and tracking	Public	Manen, Gygli, Dai, and Van Gool (2017)
INRIA Person	64 × 128 px	Pedestrian detection, re-identification	Public	Dalal and Triggs (2005)
Kinetics	Not mentioned	Action detection and recognition	Public	Kay et al. (2017)
Caltech Pedestrian	640 × 480 px	Pedestrian detection and tracking	Public	Dollar, Wojek, Schiele, and Perona (2012)
Violent Flows (ViF)	320 × 240 px	Violent action recognition	Public	Hassner et al. (2012)
Sports-1m	Not mentioned	Sports action detection/recognition	Public	Karpathy et al. (2014)
Caviar	Not mentioned	Group behavior/anomaly recognition	Public	ED (2003)
Behave	640 × 480 px, 25 FPS	Group behavior/anomaly recognition	Public	Blunsden and Fisher (2010)
The Friends Meet (FM)	Not mentioned	Joint pedestrian/group tracking	Public	Bazzani, Cristani, and Murino (2012)
MuseumVisitors	1280 × 800 px, 5 FPS	Joint pedestrian/group tracking, behavior analysis	Public	Bartoli, Lisanti, Seidenari, Karaman, and Del Bimbo (2015)
SALSA	Not mentioned	Group detection/activity recognition	Public	Alameda-Pineda et al. (2016)
COCO	Not mentioned	Object and keypoint detection, and image segmentation	Public	Lin et al. (2014)
Crowd-11	220 × 400 to 700 × 1250, variable FPS	Crowded scenes classification and anomaly detection	Partly	Dupont, Tobías, and Luvison (2017)
Motion Emotion	554 × 235, 30 FPS	Crowded scenes classification and anomaly detection	Public	Rabiee, Haddadnia, Mousavi, Kalantarzadeh et al. (2016), Rabiee, Haddadnia, Mousavi, Nabi et al. (2016)
UCF Crime	240 × 320, 30 FPS	Anomaly detection and classification	Public	Sultani, Chen, and Shah (2018)
CCTV Fights	Variable resolution and FPS	Anomaly detection	Public	Perez, Kot, and Rocha (2019)
Crowd-Flow	HD resolution, 300–450 frames/vid	Crowd tracking	Public	Schröder, Senst, Bochinski, and Sikora (2018)
UCF crowd tracking	Low resolution, 333–492 frames	Crowd tracking	Public	Ali and Shah (2008)
GTA5 Crowd Counting (GCC)	1080 × 1920 resolution	Crowd statistics	Public	Wang et al. (2019)

tiresome with unstructured crowds. This statement lays out the idea that each of the subtopics of crowd analysis can be subjected to anomaly detection. Hence, anomaly detection and/or forecasting can be done for each of crowd statistics and crowd behavior analysis.

4.1. Crowd statistics

Crowd Statistics consists in determining the quantity of people present in a scene. This can be done either by computing the density of the scene through what is commonly known as crowd density, or by calculating the number of pedestrians appearing in a scene following the application of a pedestrian detection method. Sindagi and Patel (2018)'s review shows that many works, involving Deep Learning methods, have been dedicated to crowd statistics.

In this section, we start by mentioning a work that employs hand-crafted methods from the pre-Deep Learning era. After that, we mostly present works that employ the joint learning of crowd counting and density estimation.

Chan et al. (2008) develops a top-down crowd counting approach that does not rely on pedestrian detection and tracking but leverages holistic features. The authors' purpose is to estimate the size of inhomogeneous crowds. The authors segment the crowd using the Dynamic Textures Mixture (DTM) (Doretto, Chiuso, Wu, & Soatto, 2003) They extract the following holistic features from the segmented regions: segment features, internal edge features (via the Canny edge detector (Canny, 1986), texture features via the Gray-level Co-occurrence Matrix (GLCM) (Haralick, Shanmugam, & Dinstein, 1973). After that, the Gaussian process regression is used to find the number of pedestrians in each crowd segment.

The authors validate their approach on the UCSD Anomaly dataset, which they created. The Receiver Operating Characteristic (ROC) curve had been used to validate the segmentation process of the DTM. During the experiments, the DTM outperforms the NCuts method (Shi & Malik, 1998). The Mean Squared Error (MSE) and the absolute error was used to test the Gaussian process that estimates the pedestrians count. In the UCSD Anomaly detection context, the authors show the superiority of their method and the features they extracted over those chosen by Davies, Yin, and Velastin (1995) and Kong, Gray, and Tao (2005).

Marsden et al. (2017) propose ResnetCrowd, a Residual Network (ResNet) architecture to learn many tasks related to crowd analysis. The architecture is intended for Multi Task Learning purposes: learning simultaneously crowd counting and crowd density estimation as well as violent behavior classification. The architecture is based on ResNet18 (He, Zhang, Ren, & Sun, 2016), which is originally pretrained on the ImageNet dataset. Globally, the architecture is the same. Still, the authors modified it slightly to get accurate task-specific outcomes. The architecture is trained, validated and tested on the self-developed dataset, Multi Task Crowd dataset, that we present later on in this review in Section 5.2. The computations were undertaken using the Nvidia Gefore GTX 970 GPU. The architecture was tested, in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Area Under the Curve (AUC), and compared to state-of-the-art methods on other datasets, such as UCF Crowd 50, WWW Crowd test set (Shao, Kang, Change Loy, & Wang, 2015) and the UMN Anomaly dataset. Their experiments show that architectures that are trained to perform many close-related tasks, yield better results for violence recognition and density estimation than those trained solely on a single task. However, single task trained architectures outperform ResnetCrowd on crowd counting. It is worth noting that the authors' method is massively outperformed by Zhang, Zhou, Chen, Gao, and Ma (2016)'s and Marsden, McGuinness, Little, and O'Connor (2016a)'s in crowd counting on the UCF Crowd 50, slightly by Shao et al. (2015)'s in violence identification on the WWW Crowd test set, and slightly in anomaly detection by Li, Mahadevan, and Vasconcelos (2014)'s and Marsden, McGuinness, Little, and O'Connor (2016b)'s on the UMN dataset.

In the same line, Sindagi and Patel (2017) propose a Cascade of CNNs to jointly learn end-to-end crowd counting and density estimation. The purpose of this work is to yield models that are resilient to scale and appearance variations, and that can classify crowd counts based on the density estimation of the scene. To do so, the network learns relevant discriminative global features to estimate density maps. The architecture of their model consists into two stages. It contains shared convolutional layers, and a high-level prior stage. Training and evaluation were performed on Nvidia GTX TITAN-X GPU, on two publicly available datasets: ShanghaiTech (Zhang, Zhou et al., 2016) and UCF Crowd 50 (Idrees et al., 2013). Compared to other methods on these datasets, the method outperforms state-of-the-art methods in terms of the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) on ShanghaiTech. However, it is outperformed by Onoro-Rubio and López-Sastre (2016)'s and Walach and Wolf (2016)'s in terms of MSE, on UCF CROWD 50.

Liu et al. (2019) propose an end-to-end multi-scale trainable deep architecture which relies on density estimation for crowd counting. Their method extracts features from different unit areas on the image using various receptive field sizes depending on the image perspective to take into account scale variations. Based on the context, they do not predefine image patches suiting every scale, but rather weight each extracted feature to anticipate scale variations, and then fuse the multi-scale information. Their method works well on uncalibrated cameras, but leverages the presence of calibration information. They use a pretrained VGG-16 network to extract features. They obtain scale-aware features using Spatial Pyramid Pooling (He, Zhang, Ren, & Sun, 2015) by taking into consideration 4 different scales. The extracted features from the different scales are then concatenated. The performance of the network is reinforced when the calibration information is available. To leverage the existence of the calibration information, the authors use a supplementary branch, which consists in a truncated VGG network, dedicated to extract features from a perspective map that encodes the number of pixels per meter. They held their experiments on the following datasets: ShanghaiTech,[1] WorldExpo'10, UCF Crowd Counting 50, UCF QNRF,[2] and the ad hoc Venice dataset. To evaluate their methods, they use the mean absolute error (MAE) and the root mean squared error (RMSE). Compared to state-of-the-art methods, they found that their method perform well on densely crowded scenes but has comparable results and is sometimes outperformed in less dense scenes.

Wang et al. (2019) propose a data collector and labeler that generates crowded scenes and automatically annotates them. This helps them to create a synthetic dataset named GTA 5 Crowd Counting (GCC). Second, they propose a crowd counting network, named Spatial Fully Convolutional Network (SFCN), which they pretrain on the GCC dataset and then finetune it on real data. They obtain their best results using the ResNet101 backbone for the crowd counter on UCF-QNRF, ShanghaiTech A and B, UCF Crowd Counting 50. Finally, they propose a domain adaptive crowd counter to convert the synthetic data to real data adapted for each target real dataset. After that, they train the SFCN on the converted data and test it on the real data. The converter they use is SSIM Embedding Cycle Generative Adversarial Network, SE Cycle GAN. It is equipped with the Structural Similarity Index (SSIM) loss. This loss measures the similarity between two images by comparing their local patterns. They evaluated their method using the Mean Squared Error (MSE), Mean Absolute Error (MAE), SSIM measure, and the Peak Signal to Noise Ratio (PSNR) measure on the following datasets: UCF-QNRF, UCF_CC_50, ShanghaiTech A/B, and on WorldExpo'10. Their method outperforms Cycle GAN (Zhu, Park, Isola, & Efros, 2017) and SFCN (without domain adaptation) in every dataset.

Wan and Chan (2019) jointly learn in an end-to-end framework a counter (density map generator) and a density map refiner. Their

[1] ShanghaiTech: https://www.kaggle.com/tthien/shanghaitech.

[2] UCF QNRF: https://www.crcv.ucf.edu/research/data-sets/ucf-qnrf/.

framework is made up of a refinement network and a counting network. The counter network generates a density map from a crowded scene. The refiner receives as input a ground-truth dot map which consists in annotations of individuals within a crowded scene. The refiner yields a better version of the density map and compares it to the ground truth. The refinement network applies a preliminary convolution that passes the dot map through various Gaussian kernels. This step yields blurred density maps that are masked using a self-attention module. These masked density maps are fused to yield a final density map. The produced density map is used as the ground truth to train the counter. They use compare different existing counter networks: MCNN (Zhang, Zhou et al., 2016), FCN-7c (Kang & Chan, 2018), SFCN (Wang et al., 2019) and CSRNet (Li, Zhang, & Chen, 2018). A combined loss for refinement and counting is computed during the training process to jointly train the counter and the refiner. Experiments were undertaken on the following datasets: ShanghaiTech A and B, WorldExpo'10, UCF-QNRF, and the methods were evaluated using the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE).

As we can see from the above-mentioned works (Marsden et al., 2017; Sindagi & Patel, 2017), Multi Task Learning is a good idea for learning/fine-tuning simultaneously several tasks of crowd statistics. However, the training should be undertaken on various types of datasets to yield a good-performing model. We observe that recent works on Crowd Statistics tend to associate Crowd Counting with Density estimation. More precisely, density estimation is undertaken in order to perform crowd counting. Competition in this field is very harsh. However, a lot of work still needs to be done. One of the challenges of Crowd statistics is to accurately identify the heads positions in a crowded scene as illustrated in Sam, Peri, Sundararaman, Kamath, and Babu (2019). Recently, Wan, Kumar, and Chan (2020) counted the number of persons performing a certain action in a crowded scene. It is worth noting that we have not come across such works. Wan et al.'s work reduces the gap between crowd statistics and behavior recognition.

4.2. Crowded scene analysis

Crowded scene analysis should be seen as a different task from human behavior analysis. Contrary to the former, this latter focuses on a unique subject (Marsden et al., 2017). Throughout the last years, crowded scene analysis have gained a certain interest within the computer vision community. Consequently, many methods have been developed. According to recent reviews (Chong & Tay, 2015; Tripathi et al., 2018), we can categorize crowded scene analysis methods into two main classes: conventional methods pertaining to the pre-Deep Learning era, and Deep Learning-based methods. However, as we do not intend to perform a comparative study of Deep Learning methods used in crowd behavior analysis, we will not adopt this categorization to structure this subsection. We will rather follow the taxonomy illustrated in Fig. 11.

4.2.1. Action recognition

Hassner et al. (2012) propose a taxonomy for individual-scene Human Action Recognition methods, and they categorize them can as either local, interest-point based, frame-based, or global. Studying individual-scene Human Action Recognition/Detection is not in the scope of this review, but we will mention some works that might be of interest for action detection and recognition in crowded scenes.

Siva and Xiang (2010) use 3D sliding-window fashion to capture actions within 3D cuboids in a video-recorded crowded scene. The authors extract salient points and track them using the scale-invariant feature transform (SIFT) descriptor and describe motion using the Trajectory Transition descriptor (TTD). These descriptors are then used to construct the Bag of Words (BoW) representation for each action. They split the video into 24 channels for each descriptor (which sums up to 48 for both of them). They use the channel selection routine (comparable to that of (Laptev, Marszalek, Schmid, & Rozenfeld, 2008) to elect 5 clip centers from the 48 channels by relying on cross validation on the training data. An action cuboid is represented by the multi-channel BoW. The 3D cuboid is sled through space and time (3D Sliding-window fashion), and Support Vector Machine (SVM) are used to learn each 3D action cuboid on an annotated sequence. The authors posit the action detection problem using Multiple Instance Learning (MIL) by considering clips containing a specific action as positive bags and clips not containing it as negative bags. They annotate only one positive clip with an action cuboid. They solve the MIL problem by using a greedy K Nearest Neighbor (KNN) approach. Their method is validated on the CMU (Ke, Sukthankar, & Hebert, 2007) and the i-LIDS (Branch, 2006) datasets. The evaluation metrics they work with are the precision–recall curve and the mean average precision (MAP).

In their pioneering work, Baccouche et al. (2011) propose one of the first uses of 3D CNNs in the classification of video clips for human action recognition. They learn spatio-temporal features from action clips using 3D Convolutional Neural Networks (CNN), after extending the 2D convolutions to 3D. After that, they use Recurrent Neural Networks (RNN) with Long-short-term memory (LSTM) units to classify each clip by making use of the temporal evolution of its features. The features extraction architecture (3D CNNs) contains 10 layers alternating convolutions, rectification, and sub-sampling in 8 layers and ending by 2 fully connected network. The architecture contains 17,169 trainable parameters. It is trained by online Backpropagation with momentum (LeCun, Bottou, Bengio, & Haffner, 1998) "adapted to weight sharing". The classifier architecture (RNN) contains 50 LSTM units. It contains 25,000 trainable parameters and is trained using online Backpropagation through time with momentum (Gers, Schraudolph, & Schmidhuber, 2002). The approach had been tested on the KTH dataset (Schuldt, Laptev, & Caputo, 2004) which contains 6 classes of video clips and on which they obtain good results compared to existing methods. They follow the evaluation protocol proposed by Gao, Chen, Hauptmann, and Cai (2010) and relied on cross-validation with 5 runs.

Tran et al. (2018) use a factorized version of 3D Residual Nets (ResNets) to model separately spatial and temporal components of a video clip. They use R(2+1)D blocks, which are convolutional spatio-temporal blocks, trained from scratch on two popular datasets intended for individual-scene action recognition: Sports-1 m and Kinetics. After that, they fine-tuned their blocks on two other well-known datasets: UCF-101 and HMDB-51. Amongst other models, the authors compared their model variants to several I3D variants (Carreira & Zisserman, 2017). The best authors' model outperforms, in terms of accuracy, all the methods, except one, the I3D two-stream. This latter slightly outperforms R(2+1)D two-stream on UCF-101 and HMBD-51.

The conclusions of the precedent work bring us to Carreira et al.'s work on I3D, or namely Inflated 3D ConvNet (Carreira & Zisserman, 2017). Their model relies on 2D ConvNet inflations. When tested on UCF-101 and HMDB-51, they outperformed, in terms of accuracy, all the recent models they evaluated. Their Two-Streams I3D, that was pre-trained on Kinetics, reached 80.9% of accuracy, and the Two-Streams I3D, pre-trained on both of ImageNet and Kinetics, reached 98% of accuracy.

Both of the previous works are about action detection/recognition in individual scenes. In crowd analysis, we are more interested in detecting and recognizing actions in crowded environments. You and Jiang (2018) propose Action4D to detect actions in crowded scenes. They start by detecting and tracking each person in the scene and then use their Action4D-Net to recognize the action he/she is performing. They trained their model on a self-made 4D Action dataset depicting a scene viewed from multiple angles. Although their method can be used in real-time in a multi-cameras setup, the use of RGBD cameras is not enough widespread to make their algorithm applicable in every situation. They tested their method in terms of accuracy, revised accuracy (RAcc), and the confusion matrix.

Wei et al. (2020) support that the crowd type stems from the crowd mood and the crowd behavior. Because of this, they propose a model representation of a crowd, summarized into the following triad : Behavior–Mood-Organized (BMO). They perceive three crowd types, each with a number of characteristics to which they associate a rule-based alert system: 1. heterogeneous crowds which originate for ordinary reasons in common crowded places such as railway stations and supermarkerts, should not be continuously controlled, 2. homogeneous crowds which rise from demonstrations, urban parades or during sports events, should be under control but with a certain restraint, 3. violent crowds which are an exacerbation of a political or a social demonstration must be controlled. To learn this model representation they propose the Crowd Type Recognition Network (CTRN). The CTRN is a two stream network architecture. Each stream is a Visual Geometry Group network of 5 convolutional layers (VGG16) that was pretrained on the ImageNet dataset. The first stream is fed with a static red–green–blue (RGB) image of a scene, and the second stream is fed with a motion map of a scene that contains trajectory features. The CTRN method is included within an emergency alert system that send alerts when the BMO triad satisfies certain conditions. The method is trained and tested onto the CUHK crowd and the normal–abnormal datasets. CUHK Crowd is divided into homogeneous and heterogeneous crowds, and the normal–abnormal crowds, is a local dataset, that contains homogeneous and heterogeneous crowds, but essentially violent crowds. The BMO representation of the crowd, which is a good start to differentiate multiple crowd types situations, had been reduced to a three classes classifier. Hence, the CTRN and the alert system will not be able to catch special cases scenarios where heterogeneous crowds would be endangered by the abnormal behavior of individual elements. The evaluation metrics that were used are the Receiver Operating Characteristic (ROC) curve, the Area Under the Curve (AUC), and the Mean Accuracy.

Wang and O'Sullivan (2016) propose a Spatio-temporal Hierarchical Dirichlet Process (STHDP) that is a non-parametric Bayesian method to learn spatial and temporal patterns in order to detect activities and anomalies in video data. Time is modeled in a non-Markovian continuous fashion to handle the varying time duration of each activity. The STHDP is an unsupervised clustering method that does not require much prior knowledge about the crowd dynamics. In this situation, the number of clusters is not predetermined. The type of clusters are trajectory clusters. The authors do not group trajectories using a distance metric but cluster individual observations of trajectories from frame to frame. The method detects the occurrence of an activity, its duration, its fading, and its disappearance. The method had been trained and tested on a synthetic data and on real data. Three datasets were used for the real data: 1. Edinburgh dataset (Forum), 2. MIT Carpark (Carpark), 3. New York Central Terminal (Train Station). On these datasets, STHDP was compared to two non-recent methods: MOTIF (Emonet, Varadarajan, & Odobez, 2011) and Dual Hierarchical Dirichlet Processes (DHDP) (Wang, Ma, Ng, & Grimson, 2011). Although, it is slower than MOTIF and as fast as DHDP, it performs slightly better than DHDP and is highly more accurate than MOTIF.

Yan et al. (2019) propose a crowd video captioning approach applied on off-site spectators: a type of crowds that they consider as neglected by research studies. The purpose is to generate 8 different comments describing the density and the behavior of a crowd: someone walk in, someone run in, someone walk out, someone run out, many people walk in, many people walk out, many people run in, many people run out. As the possible classes suggest, these off-site spectators crowds occur at entrances or exits of stadiums, theaters, rallies, etc. They apply their methods on the WorldExpo'10 dataset, and they evaluate them using the following evaluation metrics for captioning: Cider, Meteor, Bleu, Rouge. They propose a pipeline of an encoder–decoder made up of a feature extractor (the encoder) that feeds a sequence-to-sequence network that converts video to text (the decoder) S2VT. The feature extractor is either a 3D ConvNets (C3D) network pretrained on the Sports-1 m dataset, or ResNet-152, Inception V3/V4, each of them trained on ImageNet. The S2VT is a 2-layer Recurrent Neural Network (RNN) whose cells are either LSTM or GRU. They evaluate all the possible combinations, in terms of accuracy and the above-mentioned metrics, to select the best combination which ends up to be Inception V3 as feature extractor coupled with GRU as cells for the S2VT to generate comments.

Ullah et al. (2016) work on crowd behavior identification. They intend to identify 5 types of behaviors which are: lane, arch/ring, bottleneck, blocking, fountainhead. They propose an approach based on the extraction of optical flow from a video clip, using the Farnebäck technique (Farnebäck, 2003), on which they apply the Thermal Diffusion Process (TDP) (Wang et al., 2014) to make it more coherent. After that, the moving particles are restricted to individuals on which they apply a modified version of the Social Force Model (M-SFM) to understand the interactions between individuals. At the end of this process, they end up with a continuous dynamic system that describes the motion flow field from which they extract the first order and the second order derivatives to identify the crowd behavior. Their method was validated on the benchmark dataset (Solmaz, Moore, & Shah, 2012) and the UCD dataset (Ullah & Conci, 2012), evaluated using the average F1 score metric, and compared to Solmaz et al. (2012)'s method where it has difficulties with the lane class.

Ullah et al. (2019) work on crowd video classification. They propose a two-stream convolutional architecture to achieve this. A stream extracts spatial features from an RGB frame and a second stream extracts temporal features from a trajectory-based descriptors (Wang, Kläser, Schmid, & Liu, 2011) computed on a motion flow field of a certain number of consecutive frames. The streams are initialized with weights pretrained on the ImageNet dataset. Their method is evaluated on the CUHK crowd dataset and compared to reference methods on terms of average accuracy and confusion matrices: tensor learning classification (TLC) (Zhang, Liu, & Jiang, 2018), spatio-temporal classification (STC) (Li, Liang, & Jin, 2016), and energy-based features (Zhang, Zhang, Hu, Guo, & Yu, 2018).

We observe that even if action recognition in individual scenes is a hot topic, action or behavior recognition in highly crowded scenes is not enough explored in recent works, despite few interesting recent projects (Dupont et al., 2017; Ullah et al., 2019, 2016; Yan et al., 2019). Applying Deep Learning models developed for action detection and recognition and applying them to class video clips of highly crowded scenes is challenging and still a niche due to data scarcity.

4.2.2. Motion tracking, analysis, and prediction

Before the massive application of Deep Learning methods in Computer Vision starting from 2012, research in trajectory and motion analysis was split between flow analysis (top-down approaches) and pedestrians trajectory analysis (bottom-up approaches).

Ali and Shah (2007) use Lagrangian particle dynamics to segment a crowd flow and detect instabilities within the crowd. Their approach is at the crossroads of motion analysis, crowd behavior analysis, and anomaly detection. The authors consider the moving crowd as an aperiodic dynamical system. They start by computing an optical flow field that illustrates the interactions of the individuals within the crowd with each other and with the physical environment. They lay a grid of particles on this flow field and advect it to create Flow Maps. After that, they compute the Finite Time Lyapunov Exponent (FLTE) field from the spatial gradient tensor of the Flow Map. This operation yields the Lagrangian Coherent Structures (LCS) which allows to divide the crowd into several groups. The approach is tested on clips taken from Getty-Images, Photo-Search, Google Videos and Inside Mecca a National Geographic documentary. However, the approach was neither tested using proper metrics nor compared to other methods. The code source of this approach is publicly available in github.[3]

In another work, Ali and Shah (2008) propose a probabilistic approach to track individuals in a highly dense crowded scene based on the scene structure force model. The constraints of this force model

[3] https://github.com/saadali37/Crowd-Flow-Segmentation.

are determined by three floor fields inspired from the evacuation dynamics domain. The three fields are (1) the Static Floor Field (SFF): which represents the locations the crowd is heading to, (2) Dynamic Floor Field (DFF): which represents the interacting forces between the individual and the surrounding crowd. Here, the optical flow, the use of particles advection and the social force model is involved, (3) Boundary Floor Field (BFF): which represents the limits of the crowded environment that the crowd cannot infringe and that are found through the computation of the Finite Time Lyapunov Exponent (FLTE) field. The authors divide the scene image into a grid of cells where each cell represents a particle or a pedestrian. On each particle they calculate the three constraints: DFF, SFF, BFF. These constraints help to predict the future locations of the pedestrians. Ali and Shah undertook their experiments on the three sequences of the Marathon dataset. Their method was compared in terms of accuracy to the mean-shift tracker. The experiments show that the authors met some issues with tracking pedestrians in crowded scenes when their approach faces severe illumination changes and occlusion. However, it could handle partial occlusions.

With the emergence of Deep Learning, the development of trackers caught its second wind during the last years (Alahi, Ramanathan, & Fei-Fei, 2017; Bera, Kim, & Manocha, 2018; Sadeghian, Alahi, & Savarese, 2017). The emphasis is put on a satisfying trade-off between speed and accuracy (Bewley et al., 2016). In what follows, we list out the trackers we came across.

Bewley et al. (2016) implemented SORT, an online and real-time tracker, that relies on the Kalman Filter (Kalman, 1960) and the Hungarian algorithm (Kuhn, 1955). Bewley et al. point out the necessity to choose an excellent pedestrian detector. For their tracker, they used the Faster Region CNN (FrRCNN) (Ren et al., 2015). More precisely, after comparing the results of FrRCNN(Zeiler Fergus (ZF)) (Zeiler & Fergus, 2014) and FrRCNN(Visual Geometry Group (VGG) 16) (Simonyan & Zisserman, 2014), they opted for the FrRCNN(VGG16). The detected pedestrians are represented by Bounding Boxes. The Kalman filter is used to compute the velocity components of a target when a new detection is associated to it. The Hungarian algorithm is used to assign a new detection to a target relying on the Intersection-over-Union IoU distance. The authors point out that this policy helps to handle short term occlusion. Tested on 11 sequences taken from the Multiple Objects Tracking (MOT) benchmark (Leal-Taixé, Milan, Reid, Roth, & Schindler, 2015), and compared with 9 other trackers, SORT is 20 times faster than the state-of-the-art. It achieves a good accuracy score, making it close to state-of-the-art trackers. However, SORT considers long-term occlusions as an unimportant issue, making it failing to perform in person re-identification which worsens the score of identity switches.

In the same line, Wojke et al. (2017) developed deepSORT, an improved version of SORT that includes a deep pre-trained association metric augmented with appearance and motion information. This makes it facing better long-term occlusions while maintaining a similar speed. The association metric is pre-learned on the MARS person re-identification dataset (Zheng et al., 2016) using a Convolutional Neural Network. Specifically, the key changes occur at the assignment stage. More precisely, the SORT's formulation of the assignment problem is kept: the Kalman filter predicts states. These states are correctly associated to newly arrived measurements by using the Hungarian algorithm. However, within deepSORT, the motion and the appearance features are integrated to this formulation. On the one hand, the motion feature is incorporated following the computation of the squared Mahalanobis distance between the Kalman predicted states and the newly arrived measurements. On the other hand, the appearance feature is incorporated following the computation of the cosine distance between the appearance descriptor of the last detection and the appearance descriptors of many tracks. The bounding box appearance descriptors are yielded using a pre-trained CNN model. However, despite the use of these features, the assignment procedure still suffer from some issues that are caused by longer occlusions. To solve this, the authors introduce a matching cascade procedure. This latter prioritizes the young tracks related to frequently observed pedestrians. Compared to SORT and to other batch and online trackers on MOT16 (Milan et al., 2016), deepSORT achieves competitive performance scores. Furthermore, it reduces the identity switches of SORT by 45%. But it suffers from a high number of false positives. The sequences on which the method was evaluated contain top-down surveillance setups and front-view scenes with a moving camera. However, during these experiments, the hyperparameter weighting the intervention of each additional information had been nullified, implying the non-involvement of the motion information. Despite the arguments explaining this nullification for moving camera setups, we do not see how this can be justified for static cameras. It would have been interesting to see the results obtained if different hyper-parameter values were considered for these sequences.

Lamba and Nain (2019) propose a contour-based trajectory clustering method for crowd flow segmentation to detect overcrowdedness within a crowd to issue alerts in the attempt to avoid crowd disasters. Their method starts by separating the foreground region occupied by the crowd from the rest of the image occupied by the environment (walls, trees, etc.). After that, block-level interest points are extracted from the crowd, which are tracked over the frames using Kanade–Lucas–Tomasi algorithm. The trajectories are clustering based on several parameters: position, density, shape, direction, using the Jaccard distance. Finally, the crowd flow is segmented using the DB-SCAN algorithm. The density of each flow is analyzed to detect local overcrowdedness. Their method is trained and tested on UCF crowd dataset, Collective Motion,[4] and Violent Flows. However, their results were compared in terms of Jaccard similarity, F-score, and Mean Absolute Error (MAE) to non-recent methods, which makes the proposed method perform well.

Li et al. (2019) propose a top-bottom clustering algorithm based on pedestrians trajectories to detect groups. First, they extract the pedestrians trajectories. After that, they leverage distance and velocity difference between two trajectories in order to measure their similarity. Secondly, they use a combination between density peak clustering and a greedy algorithm to perform top-level coarse-grained clustering that relies on the mean distance between trajectories based on the Euclidean distance. Here the authors do not need to specify the number of clusters. After that, they use the improved Hausdorff mean distance to perform the bottom-level fine-grained clustering on the output of the top-level clustering. The proposed method is evaluated on a unique scenario of a real-world scene taken by an HD camera of an Unmanned Aerial Vehicle (UAV) that was capturing a crowd at an intersection, where 67 pedestrians were tracked. They use the silhouette coefficient and the consistency rate to evaluate the quality of the bottom fine-grained clustering. They compare their method with the fundamental diagram (FD) (Favaretto, Dihl, & Musse, 2016), Ge, Collins, and Ruback (2009)'s, and Ge et al. (2012)'s methods, and based on their results they obtain a better accuracy.

Wu et al. (2017) propose to extract Curl and Divergence of motion trajectories (CDT) to describe motion patterns and classify them within a crowded scene. Their method is able to classify 5 types of crowd behaviors: lane, clockwise arch, counterclockwise arch, bottleneck, fountainhead. They start by extracting the optical flow from the video clip using the Lucas–Kanade method (Lucas, Kanade, et al., 1981), and then apply a temporal clustering on the optical flow to obtain the motion vector field. After that, they apply particle advection (Solmaz et al., 2012) to decompose the motion field into sub-motion fields. From these fields, they extract the CDT descriptors that describe curl along the tangential paths, and divergence along the radial paths. After applying a feature pooling on the CDT descriptors, they obtain motion features that are used as input to a Support Vector Machine (SVM)

[4] Collective motion: http://mmlab.ie.cuhk.edu.hk/projects/collectiveness/dataset.htm.

classifier to detect crowd behaviors using one-against-all strategy. They evaluate their method in terms of Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) curve on the UCF crowd dataset, CUHK Crowd dataset, and a combination of UCF and CUHK crowded scenes. They compared to other baseline methods, where they showed that it performs well on all the classes, except on the bottleneck class.

Alahi et al. (2016) propose Social-LSTM a Recurrent Neural Network which contains Long Short-Term Memory units, to learn pedestrians movements and jointly predict their future locations by taking into account the social context of each pedestrian's trajectory. From a video clip of a crowded scene, the authors obtain the positions of the pedestrians on all the frames. After that, they use a separate LSTM model to learn each pedestrian's trajectory. The LSTMs are connected to each other via a social pooling layer, so that spatially close LSTMs share information about each other and inputs it to the next step. They train their model following a leave-one-out validation strategy and evaluate it on UCY (Lerner, Chrysanthou, & Lischinski, 2007) and ETH (Pellegrini, Ess, Schindler, & Van Gool, 2009) datasets. Social-LSTM is tested in terms of Average Displacement Error (ADE), Final Displacement Error (FDE), Average non-linear displacement error (ANLDE). They compare their method to the Kalman Filter, a Collision avoidance based on the Social force model (Yamaguchi, Berg, Ortiz, & Berg, 2011), the social force model, the iterative Gaussian process (IGP) (Peter, Richard, Murray, & Krause, 2013), a vanilla LSTM model without the social pooling layer, and a simpler version of their model that contains only the occupancy maps (O-LSTM) which captures the positions of the neighbors at time t without taking into account the previous positions. The methods learn the trajectories for 3, 2 s and try to predict the 4, 8 following seconds. The overall results show that Social-LSTM outperforms other methods on the test sequences of both of the datasets. However, it is sometimes outperformed by IGP because it knows the ground truth final destination of each pedestrian contrary to the other methods, and O-LSTM because it performs well on less crowded scenes.

Bartoli, Lisanti, Ballan, and Del Bimbo (2018) propose a context-aware Recurrent Neural Network that uses Long Short-Term Memory units to learn and predict pedestrians trajectories. They extend the Social-LSTM and the O-LSTM models proposed by Alahi et al. (2016) by incorporating to them a context-aware pooling that takes into account the human–human interactions as well as the static objects that are located in the vicinity of pedestrians. This approach needs a prior knowledge of the positions of the static objects. They evaluate their method, in terms of Average Displacement Error (ADE), on the UCY dataset (Lerner et al., 2007) and the MuseumVisits dataset (Bartoli et al., 2018). As in Alahi et al. (2016), the methods learn the trajectories for 3, 2 s and try to predict the 4, 8 following seconds. The authors found out that context-aware O-LSTM performs well on the MuseumVisits because the persons move in group and are only interested in artworks, but context-aware Social-LSTM performs well in the UCY sequences because there the persons move alone and the prior knowledge of entry and exit points are fed to the model.

A lot of research has been done for motion tracking (Bewley et al., 2016; Wojke et al., 2017) and motion analysis (Lamba & Nain, 2019; Li et al., 2019; Wu et al., 2017). Contribution in these two sub-fields has become very hard due to harsh competition. However, in the recent years, we denote a burgeoning interest for motion prediction (Coscia et al., 2018; Tang, Ma, Liu, & Zheng, 2018) that we can illustrate in the papers we mentioned (Alahi et al., 2016; Bartoli et al., 2018).

4.2.3. Group behavior analysis

Here, we present some approaches used for group behavior analysis. Today, Deep learning is widespread in Group behavior analysis. However, the use of these methods started years after the first successes of Deep Learning in Computer Vision.

Shao et al. (2014) propose an approach to detect groups within a crowded scene and to recognize their behavior. They detect the groups using the Collective Transition (CT) prior. This prior also helps to find visual group descriptors and properties that are scene-dependent and impervious to crowd density variations, which are: stability, collectiveness, uniformity, and conflict. These extracted descriptors help to determine the groups' inner states and behaviors into one of these four classes: Gas, Solid, Pure fluid, Impure fluid. The Collective Transition approach is a clustering algorithm that is used for group detection. The authors start by extracting the tracklets of each pedestrian using the KLT tracker. The initial clusters of tracklets are found by using the Coherent Filtering method (Zhou et al., 2012). After that, anchor tracklets are elected to represent the centers of the clusters (groups) or individuals and tracklets are grouped using the Expectation–Maximization (EM) algorithm. Intra-group and inter-group properties which are stability, collectiveness, uniformity, and conflict, are verified through the use of graph K-Nearest Neighbors algorithm. Experiments had been undertaken on the self-made CUHK Crowd dataset. The authors' method was compared to a mixture of dynamic textures (DTM) (Chan & Vasconcelos, 2008), hierarchical clustering (HC) (Ge et al., 2012), coherent filtering (CF) (Zhou et al., 2012). The measures used to evaluate the performance of their clustering approach are the Normalized Mutual Information (NMI) (Wu & Schölkopf, 2007), purity (Aggarwal, 2004), and Rand Index (RI) (Rand, 1971). For the classification task of the group behavior, the evaluation metrics that were used are accuracy and the confusion matrix.

Vahora and Chauhan (2018) propose a Deep Learning-based bottom-up approach to identify group activities. Their method relies on contextual and human–human interactions. They use Convolutional Neural Networks (CNN) to capture action-pose features and scene-related cues, and Recurrent Neural Networks to unveil group changes. They developed two approaches: one based on Long Short Term Memory (LSTM) and another one based on Gated Recurrent Units (GRU). They evaluated their method on the collective activity dataset (Choi, Shahid, & Savarese, 2009), and compared it to six other state-of-the-art approaches. The author's approach that relies on GRU units outperforms all the others on terms of accuracy.

Shu et al. (2017) propose a Confidence-Energy Recurrent Neural Network (CERN) using Long Short-Term Memory (LSTM) units, to recognize individual human actions, interactions and group activities. They developed two variants of their algorithm: CERN-1 and CERN-2. They tested these two variants on the Collective Activity (Choi et al., 2009) and Volleyball (Ibrahim, Muralidharan, Deng, Vahdat, & Mori, 2016) datasets. Compared to other methods, CERN-2 outperforms all the methods in terms of accuracy.

Zitouni et al. (2020) adopt a mesoscopic approach to crowd behavior analysis, where they perceive the crowd as a compound of two types of elements individuals and groups. In this context, groups are divided into simple groups which demonstrate a homogeneous behaviors and compound groups which demonstrate heterogeneous behaviors. They propose a pipeline where pedestrians, heads, and groups are detected and then tracked. Afterwards, a Gaussian Mixture Model of Dynamic Textures detection technique (GMM-of-DT) (Zitouni, Bhaskar, & Al-Mualla, 2016) based on a Kalman filter is utilized to categorize crowd behaviors into 4 classes: individual, group, leader–follower, and social interaction. The method was validated on 6 sequences from the PETS dataset (Ferryman & Shahrokni, 2009), and tested on sequences from Parking Lot (Dehghan, Modiri Assari, & Shah, 2015) and Town Center (Benfold & Reid, 2011). The evaluation were undertaken using the F-measure.

Bisagno et al. (2018) propose Group LSTM, a method to detect groups and predict their trajectories. First, they use the coherent filtering approach (Zhou et al., 2012) to cluster trajectories and form groups based on the individuals trajectories, and by taking into account their surroundings: pedestrians walking in the same direction belong to the same group. After that, they use an extended version of Social-LSTM (Alahi et al., 2016) to predict the group trajectories. While Social-LSTM predicts solely pedestrians' trajectories by associating an

LSTM for each of them, Bisagno et al. introduce a social pooling hidden layer to extend the interest for all the group of pedestrians, and then predict the trajectory of the group. They evaluate their method on UCY and ETH datasets, using the Average Displacement Error (ADE) and the Final Displacement Error (FDE). Their method is compared to Social-LSTM (Alahi et al., 2016), to its variant (Gupta, Johnson, Fei-Fei, Savarese, & Alahi, 2018), and to the Kalman Filter. In terms of ADE, they obtain the best results, but in terms of FDE, they sometimes fail to find the final position of a group, contrary to other methods.

Research in group behavior analysis does not attract the attention of researchers as it does for motion tracking or action recognition in individual scenes. This may be due to the lack of datasets where group activities can be found. However, we can observe during the recent years a rising interest for this research axis. Recent papers (Bisagno et al., 2018; Zitouni et al., 2020) bypass data paucity and perform group behavior analysis on datasets that are not originally meant for this task such as ETH and UCY, or PETS dataset.

4.3. Anomaly detection

Anomaly detection and forecasting can be done for any subtopic of crowd analysis. This research axis is getting more and more attention in crowd analysis because of its various applications in video surveillance and crowd monitoring (Zhan et al., 2008).

Inspired by the effectiveness of the Mixture of Dynamic Textures (MDT) in video modeling and video clustering (Chan & Vasconcelos, 2008), Mahadevan et al. (2010) propose an unsupervised framework based on MDT to model normalcy in crowded scenes. This model considers temporal and spatial outliers as anomalies. Temporal anomalies represent events that are unlikely to happen or that happen rarely, and spatial anomalies are detected using discriminant saliency. Metrics used to evaluate the model on the UCSD anomaly dataset are the ROC curve when the comparison between ground-truth and the detected anomaly is done at the frame level and the pixel level. Another metric inferred from the ROC curve that had been used is the EER (Equal Error Rate) which computes the percentage of mis-classified frames. The MDT outperforms the Social Force method (Mehran et al., 2009), an optical flow monitoring method (Adam, Rivlin, Shimshoni, & Reinitz, 2008), the mixture of optical flow (Kim & Grauman, 2009), and a combination between (Kim & Grauman, 2009; Mehran et al., 2009). The major inconvenience of this approach is its high dependence on the examples it was trained on to consider as normal events.

Mehran et al. (2009) employ the Social Force model to detect anomalies in crowded scenes. The authors lay a grid of particles on the first frame of a crowded scene, and then they apply particles advection following the fluctuations of the optical flow field. A Force Flow vector field is extracted, relying on the laws of the Social Force model which is based on the interactions that occur between the particles in the scene. From this vector field, a bag of words is constructed to represent the different behaviors of a crowd. The Latent Dirichlet Allocation (LDA), a Natural Language Processing technique, is trained to recognize normal behaviors. Finally, the Expectation–Maximization (EM) method is used to differentiate between normal and abnormal behaviors. Abnormal behavior is found in regions where high force flow occurs. The method is tested on the UMN and the Web datasets. The ROC curve metric is used to test the method. The authors show that the Social force-based approach outperforms methods based solely on Optical Flow. The major drawback of this approach is its inability to recognize new examples taken from a different camera angle or position than the camera angle and position of the examples the model was trained on.

Hassner et al. (2012) propose violent flows descriptors (ViF) to detect violence in crowded scenes. These descriptors are extracted from the fluctuations in the magnitudes of optical flow vectors. The authors propose to use a Bag-of-Features representation of the scenes with the ViF descriptor and then to train Support Vector Machines (SVM) to classify the scenes in either violent or non-violent. For the classification task, the metrics that were used are the mean prediction accuracy (ACC) ±standard deviation (SD) and the area under the ROC curve (AUC). The authors propose a violence detection dataset on which they test their approach and compare it to other state-of-the-art features extraction techniques. Their descriptor is efficient in classifying violent and non-violent scenes. However, as shown by the authors, it is not as well performing on the Hockey Fights dataset where it is outperformed by other Space–Time Interest Points (STIP)-based descriptors (Laptev, 2005).

Singh et al. (2018) developed the Drone Surveillance System (DSS) to identify violent individuals in real-time using the cloud computation. First, the DSS detects humans using the feature pyramid network (FPN) (Lin et al., 2017). Then, a proposed ScatterNet Hybrid Deep Learning (SHDL) network is used for pose estimation. Brought back to the table from previous works (Singh, Hazarika, & Bhattacharya, 2017; Singh & Kingsbury, 2017b, 2017c, 2018), the SHDL network is composed of a combination of a front-end hand-crafted ScatterNet (Singh & Kingsbury, 2017a) and a back-end Regression Network. The SHDL yields a 14 key-points skeleton structure for each individual summarizing his/her pose. This structure is then leveraged to distinguish five violent activities from one normal action using a Support Vector Machine (SVM). The DSS outperforms a State-of-the-Art approach (Penmetsa, Minhuj, Singh, & Omkar, 2014) by over 10% accuracy on the proposed Aerial Violent Individual (AVI) dataset. Following these results, it would be interesting to see the performance of this method on other publicly available datasets.

Ravanbakhsh et al. (2016) developed a measure-based unsupervised approach that detects local abnormality by combining motion information with appearance. They get motion and appearance information using a Convolutional Neural Network (CNN). Their approach consists in three steps: after feeding a Binary Fully Convolutional Network (BFCN) with input frames, they extract from them binary maps. They use these binary maps to compute a measure, inspired from the commotion measure (Mousavi, Nabi, Kiani, Perina, & Murino, 2015), which they called the Temporal CNN Pattern (TCP). Mixed with Optical Flow, they yield refined motion segments. The Binary Fully Convolutional Network is composed of a Fully Convolutional Network and a Binary Quantization Layer (BQL). Although the weights of the FCN are obtained from a pre-trained AlexNet model (Krizhevsky et al., 2012), the weights of the BQL are obtained from a hashing method. They tested their methods in terms of frame-level anomaly detection and pixel-level anomaly detection. When the former assesses the ability of the model to identify correctly an anomalous frame, the latter ensures that it localizes it accurately. The experiments were undertaken on the UMN SocialForce, and UCSD's Ped1 and Ped2 datasets. The evaluation metrics are the Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC). Most of the time, the method performs well. However, sometimes, the developed model cannot detect anomaly when the abnormal object is tiny or partly occluded and/or has not a usual motion "(i.e., a car moves the same speed of normally moving pedestrians in the scene)". To enhance the approach, the authors propose two improvements: plugging a TCP measure layer and fine-tune it with back-propagation; based on two of their former works on Generative Adversarial Nets (Ravanbakhsh,Nabi et al., 2017; Ravanbakhsh, Sangineto et al., 2017), they propose them as a replacement of the BFCN.

Ramos et al. (2017) propose a meta-heuristic based approach to detect global anomalies in crowded scenes that has a low computational cost. To do so, they use Optical Flow to extract motion layers between each pair of consecutive frames. Then, they use the Artificial Bacteria Colony (ABC) meta-heuristic to optimize the extracted layers and results in a coverage of regions of interest (ROIs) depicting high movement. After that, they train a Self-Organizing Map (SOM) on the ABC's population, food storage and centroids to find particular events relying on behavior patterns similarity. ABC's population, stock and centroids represent respectively movement's spread, density and center. During the process, it is noteworthy to mention that ABC's population is subjected to the Darwinian natural selection method, and that the fitness function relies on the food stock. Using Optical Flow to extract motion layers reinforces the method's resilience to noise and light changes. Experiments were undertaken on the UMN SocialForce dataset. For decent values of bacteria population and neurons number for the SOM map, the approach outperforms other runs quite quickly, processing each frame for 0.033 s on average. It outperforms Mehran et al. (2009)'s method AUC (Area Under Curve) score by 18%. It would have been interesting to observe the performance of the author's method compared to more recent approaches. As a future work, the authors propose to use Mehran et al. (2009)'s Social Force Model instead of Optical Flow, and to extrapolate the improvements to the use of ABC to optimize spatio-temporal volumes.

Singh et al. (2020) propose Aggregation of Ensembles (AOE), an aggregation of four classifiers over sub-ensembles of three fine-tuned Convolutional Neural Nets (CNNs) on crowd datasets to detect anomalies in videos of crowded scenes using the majority vote. One classifier is applied for each sub-ensemble. A sub-ensemble of CNNs is made up of the following 3 pretrained models: AlexNet (Krizhevsky et al., 2012) on CIFAR-10, GoogleNet (Hu, Huang, Gao, Luo, & Duan, 2018), VGGNet (Simonyan & Zisserman, 2014), both on ImageNet. The CNNs were used as feature extractors that feed three variants of Support Vector Machines (SVM): a Linear SVM, a Quadratic SVM, a Cubic SVM ; and a Softmax classifier. The features are extracted from a batch of frames selected from a video. If more than 10% of the frames of the batch are classified as anomalous, the video is considered as anomalous. Models finetuning, AOE training and evaluation were undertaken on the following datasets: UCSD Ped 1, UCSD Ped 2, and the Avenue dataset.[5]

Qasim and Bhatti (2019) propose a three-dimensional descriptor made up of three features, based on Optical Flow (OF) extracted from videos. Most of the time, 7 consecutive frames are selected from a video. The 3 features are: 1. sum of thresholded OF magnitude, 2. joint entropy of the OF magnitude of 2 consecutive frames. The joint entropy helps to detect sudden changes related to a rapid dispersion (for example), 3. variance of a space–time cuboid obtained using the history of OF field magnitude. On top of this descriptor a Support Vector Machine (SVM) classifier is used to detect anomalies in videos. Qasim and Bhatti's purpose is to propose a good trade-off between accuracy and real time performance. Their method is evaluated in terms of accuracy on the UMN dataset.

Hao et al. (2019) propose an abnormal behavior detector based on a Gabor-filtered extracted spatio-temporal texture. On a raw video clip, a spatio-temporal volume (STV) (Adelson & Bergen, 1985) is constructed, from which a Spatio-temporal textures (STT) are extracted. STTs are vertical or horizontal slices of STVs along the time axis. Gabor filtering is applied on the STT for background subtraction and noise removal. Among the filtered STTs, the STT that maximizes the information entropy is selected. Second, crowd features (signatures) are obtained by applying a gray-level co-occurrence matrix (GLCM) (Haralick et al., 1973) on the selected STTs. The selected STTs are first converted from RGB to gray images, a raw GLCM is applied on it, and then four features are extracted: contrast, orderliness (angular second moment and entropy), descriptive features (variance). These features help to model a STT signature which is used to feed a behavior classifier. These signature features are compared to TAMURA texture patterns (Ranjan & Agrawal, 2016). Each of these features are fed to a series of classifiers: K Nearest Neighbors, Naïve Bayes, discriminative analysis classifier (DAC), random forest, Support Vector Machine (SVM). The comparison between TAMURA features and GLCM features, held on the UMN dataset, show that Hao et al.'s GLCM strategy is good for detecting panic situations, however TAMURA is good for describing congestion situations.

Lin et al. (2019) propose Social Multiple Instance Learning (Social MIL) framework coupled with a dual branch network to detect anomalies in crowded scenes. Their approach is made up of a Two-Stream Neural Network: 1. a spatio-temporal stream, and 2. an interactive dynamic stream. The first stream is a spatio-temporal branch, that is fed with RGB video clips. The video clip is converted to a video segment then fed through a 3D ConvNets (C3D) pre-trained model on Kinetics and UCF-101 for features extraction. The features are fed to a one dimensional dependency attention capturing module which output is inputted to a fully connected network. The second stream is fed with social force maps. Social forces maps are obtained using the social force model (Mehran et al., 2009) that describe the interactions that occur within a scene. Their method is trained using a MIL ranking loss function, and evaluated on the UCF Crime dataset (Sultani et al., 2018), using the receiver operating characteristic (ROC) curve and area under the curve (AUC) metrics. During the evaluation process, it is compared to 4 other methods.

Xie et al. (2019) propose crowd abnormal behavior recognition algorithm based on the computation of optical flow and the use of the social force model. First, they compute optical flow from a video clip using Lucas–Kanade method. After that, optical flow is mapped to 2D geospatial space using the camera parameters. This operation helps them to extract particle points from a video clip. Second, they use these particle points to compute the social interaction force between the particles. They sum the interaction forces of each frame and they compare it to an empirically set threshold to decide whether a frame is anomalous or not.

Anomaly detection has always caught the interest of researchers in crowd analysis. However, as highlighted in Section 2.3, research has not come up yet with a unanimous definition of anomaly. This situation makes few previously published works generalizable to every situation of anomaly recognition.

5. The sources of data

We can find two sources of video data: 1. public or private video datasets that are used for training and testing purposes, 2. live video-surveillance from which we can create video datasets but that require video annotation. In this section, we will enumerate some sources of live video-surveillance data.

5.1. Live video-surveillance

Several sources of live video-surveillance are available and freely accessible throughout the internet. The proposed scenes can be used for many interesting tasks such as pedestrian/vehicle detection and tracking, crowd counting or crowd behavior analysis. However, their use may be subjected to prior authorization that need to be obtained from the providers. We list below some live video-surveillance sources we came across during our research. Table 3 summarizes the characteristics of the live video-surveillance we came across.

- **United Kingdom road traffic video-surveillance:** The live video-surveillance gives you access to mono-cameras observing fifty roads. The provided frames can be used mainly for vehicle detection. Tracking may be a possibility, however the cameras produce one frame per minute. In the same line to this live video-surveillance, but with a better frame-rate alongside with videos already stored in a repository, UA-DETRAC,[6] can also be used for vehicle detection and tracking (Lyu et al., 2017).

[5] http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html.

[6] UA-DETRAC dataset: https://detrac-db.rit.albany.edu/.

- **Earthcam:** This repository of live fixed mono-cameras offer a gamut of testing possibilities, ranging from crowd behavior analysis, pedestrian detection, tracking and other crowd statistics related tasks.
- **Live Mecca:** The use of the provided videos can be precious for several situations pertaining to crowd analysis ranging from crowd statistics and massive crowd motion, notably during important occasions such as the yearly hajj period. Moreover, the cameras offer a lot of views, and they are not always fixed, offering the challenging possibility to perform crowd analysis with a moving camera. The high-resolution produced frames are at a real-time frame-rate.
- **Live Vatican:** Fixed pedestrian-level mono-camera directed to the St. Peter's square, that is subjected to frequent medium density people gatherings.
- **Monthey Place Centrale:** These two cameras emit mid-to-low resolution images at around 1 frame per second. The images are subjected to illumination changes and to various other challenges such as occlusion, clutter, and different weather situations. They can be used for pedestrian and vehicle detection and tracking.

5.2. Datasets

We do not intend to be exhaustive about the existing datasets.[7] Most of the datasets we introduce, in this subsection, are frequently used for tracking and pedestrian detection such as KITTI and MotChallenge, action recognition such as UCF-101 and HMDB-51, solely pedestrian detection such as Inria Person and Caltech Pedestrian. The datasets that tend more to be used for crowd analysis-related tasks are rare to find. We listed the most relevant and popular ones. Tables 4 and 5 summarize the characteristics of these datasets.

5.2.1. Crowd statistics

- **Multi Task Crowd:** This dataset was developed by Marsden et al. (2017). They used it to train their ResnetCrowd architecture to perform three tasks: crowd counting, density estimation and violence detection. The dataset is composed of a set of images. These images are obtained from the WWW Crowd videos dataset.[8]
- **UCF CROWD 50:** Used for Crowd Counting, UCF Crowd 50[9] dataset consists of 50 images of densely crowded scenes. The pictures that were collected from FLICKR[10] contain a population ranging from 94 to 4543 individuals.
- **GTA 5 Crowd Counting (GCC) dataset:** Proposed by Wang et al. (2019), this synthetic dataset is created using the Script Hook V[11] C++ library applied on the Grand Theft Auto 5 (GTA 5) Rockstar game. The scenes come from 100 different indoor and outdoor locations in Los Santos (fictional city inspired by Los Angeles). The people in the scenes are diverse and generated from 265 different person models. However, the scenes in GTA 5 are limited to 256 individuals. The GCC dataset has 15,212 images. The resolution of each image is 1080 × 1920. The scenes of the dataset have got 7 different weather conditions: clear, clouds, rain, foggy, thunder, overcast and extra sunny. They are captured at any time of the day.

5.2.2. Tracking and motion analysis

- **Town Center:** The Town Center Dataset, proposed by Benfold and Reid (2011), is a video recorded from a video-surveillance mono-camera recording the pedestrians walking in a city center. The dataset is mainly used for tracking. At each frame, there is at around 16 pedestrians. The video quality is good, 1920 × 1080 pixels for 25 FPS.
- **i-Lids for AVSS 2007:** The i-Lids dataset[12] contains two types of videos: one recording a train station (Task 1), and an other recording road vehicle traffic (Task 2). The sequences we are interested in are those of the train station. They can be used for Pedestrian detection, tracking, Abnormal behavior detection (in this case: abandoning a luggage). The resolution of the videos are of 720 × 576 pixels for 25 Hz.
- **MOTChallenge Dataset:** The 2016 MOTChallenge Dataset[13] is an extension of the precedent version of 2015 proposed by Leal-Taixé et al. (2015) (Milan et al., 2016). It consists of 14 sequences, containing crowded scenarios such as the famous PETS09-S2L1. The dataset presents some challenging conditions like a dynamic camera, illumination changes, and various viewpoints. The annotated objects are pedestrians, sitting persons, vehicles and occluding objects, etc. This dataset encourages the use of the CLEAR metrics (Kasturi et al., 2009), in addition to other metrics such as those evoked in Huang, Nevatia, and Li (2009), to evaluate methods used for object detection and tracking.
- **KITTI:** The videos of the KITTI Dataset[14] were obtained from a moving VM station wagon that was recording for 6 h at 10–100 Hz, in the city of Karlsruhe, Germany (Geiger et al., 2013). Several sensors were involved for the recording such as a color and a grayscale stereo camera, and a Velodyne 3D laser scanner. Color and grayscale images are stored under the 8-bit PNG files format. The dataset is used, inter alia, for object detection and tracking. Eight classes of objects can be tracked and detected in challenging conditions, because the objects may be static or dynamic, and the camera is constantly moving. These eight classes are: 'Car', 'Van', 'Truck', 'Pedestrian', 'Person (sitting)', 'Cyclist', 'Tram' and 'Misc'. The provided annotations are in the form of 3D bounding boxes tracklets. The total size of the provided data is 180 GB.
- **Toulouse Campus Surveillance:** The Toulouse Campus Surveillance Dataset[15] can be used, among other things, for object detection/tracking and audio event recognition (Malon et al., 2018). The dataset results from a multi-camera setup, and thereby, can be useful for assessing Multi-source fusion methods. Some cameras are fixed, others are mobile. The setup gave birth to 50 videos fairly broken down into two scenarios. Each video is depicted by three resolution qualities: 1920 × 1080, 960 × 540, 640 × 360. The Ground Truth annotations for detection and tracking are only provided for the first scenario.
- **The PathTrack MOT:** Assisted by the crowd-sourcing permitted by Amazon Mechanical Turk, where approximately 80 AMT workers were involved, Manen et al. (2017) propose the large-scale PathTrack MOT dataset.[16] This dataset was released following the creation of the PathTrack annotator. It contains the annotation of 720 sequences depicting pedestrian movements from video-surveillance scenes that were captured by either static or moving cameras. The sequences include 16.287 annotated trajectories.

[7] For the sake of completeness; we invite you to explore these three websites:
- http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm.
- http://riemenschneider.hayko.at/vision/dataset/index.php.
- https://www.di.ens.fr/~miech/datasetviz/.

[8] WWW Crowd Dataset: https://computervisiononline.com/dataset/1105138602.
[9] UCF Crowd 50: http://crcv.ucf.edu/data/ucf-cc-50/.
[10] FLICKR: https://www.flickr.com/.
[11] Script hook v. http://www.dev-c.com/gtav/scripthookv/.
[12] i-Lids Dataset: http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html.
[13] MOT Challenge datasets: https://motchallenge.net/.
[14] The KITTI Dataset: http://www.cvlibs.net/datasets/kitti/index.php.
[15] Toulouse Campus Surveillance: http://ubee.enseeiht.fr/dokuwiki/doku.php?id=public:tocada.
[16] The PathTrack MOT dataset: https://data.vision.ee.ethz.ch/daid/MOT/pathtrack_release_v1.0.zip.

- **Optical Flow Dataset:** Also named TUB CrowdFlow Dataset, the Optical Flow Dataset[17] was created by Schröder et al. (2018). The authors graphically generated the dataset by using Unreal Engine. They simulate crowd motion in five different situations. Moreover, the crowd is captured by a static and a dynamic camera. The dataset contains 10 sequences whose lengths range from 300 to 450 frames. The sequences are characterized by a frame-rate of 25 Hz and a HD resolution. The scenes contain between 371 and 1451 individuals. The authors also verified that the results that state-of-the-art methods obtain on this dataset can be transferred to real-world generated datasets such as UCF crowd tracking (Ali & Shah, 2008). Within the dataset, the ground truth data provided is for optical flow, and dense and sparse trajectories of individuals.
- **UCF crowd tracking:** The UCF crowd tracking dataset[18] was created by Ali and Shah (2008). The dataset is composed of three sequences called Marathon-1, Marathon-2, and Marathon-3. The length of these sequences ranges from 333 to 492 frames. To test their methods, the authors annotated the trajectories of respectively 199 individuals, 120 individuals, and 50 individuals on the three sequences.

5.2.3. Pedestrian detection

- **INRIA Person:** The Inria Person Dataset[19] is used for pedestrian detection. The dataset contains bounding-boxes annotations of original images, resized positive images of pedestrians in $64 \times 128 pixels$ format with original negative images. The dataset contains 1805 images of people in challenging situations: different poses and orientations, and within a wide range of backgrounds (Dalal & Triggs, 2005).
- **Caltech Pedestrian:** The Caltech Dataset[20] is commonly used for Pedestrian Detection (Dollar et al., 2012). It can also be used for occlusion handling. The dataset contains 10 h of 640 × 480 pixels video recorded from a mono-camera mounted on a vehicle moving in the street. These 10 h video are also depicted by 250.000 frames. The provided Ground Truth is for 2300 different pedestrians, in a form of 350.000 bounding boxes.
- **COCO Common Object in Context:** Lin et al. (2014) propose the COCO dataset.[21] It is mainly used for object and keypoint detection, and image segmentation. In addition to pedestrians, 79 other objects can be detected. The dataset contains 330 K images and more than 200 K are labeled.

5.2.4. Action recognition in individual scenes

- **UCF-101:** Soomro et al. (2012) propose UCF-101,[22] an extension of UCF-50. The dataset consists of 13320 web clips depicting 27 h of video, obtained from Youtube.[23] The resolution of each clip is of 320 × 240 at a frame rate of 25 FPS. The clips are categorized into 101 action classes. The videos are subjected to background clutter, camera motion, lighting changes, partial occlusion and low quality frames. Was assessed, on this dataset, a bag-of-words (BoW) approach used for action recognition that resulted in an overall accuracy of 43.9%.
- **HMDB-51:** Kuehne et al. (2011) propose HMDB-51.[24] The dataset consists of 6766 annotated video clips gleaned from various sources such as Youtube or Movies. In terms of resolution, the videos share all the same 240 pixels height. However, the width of each video clip is re-sized so as to maintain its proper aspect ratio. The frame-rate of all the clips is of 30 FPS. The videos are categorized into 51 action classes. Meta-information within the dataset details for each video clip a range of information such as camera viewpoint, occlusion, occurrence of camera motion (that concerns 2/3's of the database), video quality (which is categorized into high, medium, or low), the number of individuals appearing in the video. Two methods were assessed on this dataset: Jhuang, Serre, Wolf, and Poggio (2007)'s and Laptev (2005), Laptev et al. (2008), Wang, Ullah, Klaser, Laptev, and Schmid (2009)'s, and achieved at around 23% accuracy.
- **Kinetics:** Kay et al. (2017) propose the Kinetics dataset,[25] a dataset that is similar to UCF-101 and HMDB-51, and that is mainly used for video classification. It consists of 400 classes of different actions performed by a wide spectrum of different persons. There are from 400 to 1150 video clips for each class. Each video clip is taken from Youtube and lasts for approximately 10 s. These actions cover human–human and human-object interactions. The challenges presented by this dataset are illumination changes, background clutter, camera motion and vibrations, shadows, etc. The annotation process relied on the Amazon Mechanical Turk (MTurk).[26] The clips are not exhaustively annotated. Some of them may incorporate more than one action, but the authors made sure that each clip is labeled with the name of, at least, one of occurring actions.
- **Sports-1m:** Karpathy et al. (2014) propose the Sports-1 m dataset.[27] This dataset contains 1 million Youtube videos that lasts for 5 min and 36 s on average. As the name of the dataset suggests it, the videos are solely about sports-related activities. Each activity is a class-name, and within the dataset, there are 487 different activities. There is, at around, 1000–3000 videos per class, and about 5% of the videos possess more than one label.

5.2.5. Anomaly detection

- **Aerial Violent Individual (AVI):** Proposed by Singh et al. (2018) to train their ScatterNet Hybrid Deep Learning (SHDL) network for pose estimation, this dataset contains 2000 images. The recorded scenes are mildly dense and contain at around 2 to 10 individuals. The interesting part of the dataset is that each individual is annotated with 14 key-points allowing a detailed estimation of its pose. The images were recorded by a drone from four different heights. Unfortunately, this dataset is not publicly available.
- **UMN SocialForce and Web datasets:** UMN SocialForce and Web datasets[28] are two datasets used by Mehran et al. (2009) in their work. UMN SocialForce consists of 11 videos all illustrating a normal-starting situation and an abnormal ending. Web Dataset consists of 20 videos, 8 videos containing abnormal events such as panic, clashes, fights, and 12 video clips of a normal situation (pedestrians walking).
- **UCSD Anomaly Detection:** UCSD Anomaly Dataset[29] is commonly used for anomaly detection and it consists of at around 100 video clips. These videos are divided into two sub-datasets: Peds1 and Peds2 (Chan et al., 2008). The anomalies are linked

[17] Optical Flow Dataset: https://github.com/tsenst/CrowdFlow.
[18] UCF crowd tracking: https://www.crcv.ucf.edu/data/tracking.php.
[19] The Inria Person Dataset: http://pascal.inrialpes.fr/data/human/.
[20] The Caltech Dataset: http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/.
[21] The Coco dataset: http://cocodataset.org/#detection-2018.
[22] UCF-101 dataset: http://crcv.ucf.edu/data/UCF101.php.
[23] Youtube: https://www.youtube.com/.
[24] HMDB-51 dataset: http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#dataset.
[25] The Kinetics dataset: https://deepmind.com/research/open-source/open-source-datasets/kinetics/.
[26] Amazon Mechanical Turk: https://www.mturk.com/.
[27] The Sports-1 m dataset: https://cs.stanford.edu/people/karpathy/deepvideo/.
[28] UMN SocialForce and Web Datasets: http://crcv.ucf.edu/projects/Abnormal_Crowd/.
[29] UCSD Anomaly Detection Dataset: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm.

to abnormal elements appearing in the video clip such as the circulation of non-pedestrians. However, anomalies pertain also to pedestrians adopting abnormal motion patterns. The ground truth indicate the occurrence of an anomaly on frame via a binary flag. The bounding-box localization of the abnormal element is also provided.

- **Agoraset:** Created by Allain et al. (2012), this dataset can be used for pedestrian tracking, abnormal event analysis and density estimation. The simulations have been generated using a model based on the Lagrangian forces proposed by Helbing, Farkas, and Vicsek (2000). The dataset consists of 7 scenarios that can be broken down into several video clips by changing the crowd state (soft motion or panic) and/or the illumination rendering (shading or no shading).
- **Violent Flows (ViF):** Following the observation about the scarcity of even Action Recognition-related datasets, and the almost non-existence of those pertaining to Violence Behavior, Hassner et al. propose The Violent Flows Dataset[30] (Hassner et al., 2012). The dataset, that consists of 246 videos, is mainly collected from Youtube. The video frames were resized to 320 × 240 pixels. The average video duration is of 3.60 s. To assess their dataset, they compared their technique to two other existing ones: an interest-point driven method (Laptev, 2005), and a frame-based descriptor, the LTP (Local Trinary Patterns) (Yeffet & Wolf, 2009). Their technique is based on the use of Optical Flow and Support Vector Machines. They also developed for the need of their method the ViF (Violent Flows) descriptor. The code to compute the violence descriptors is publicly available.[31]
- **Caviar:** The Caviar Dataset[32] can be used for pedestrian and group behavior analysis in a mildly crowded scene, but mainly for anomaly detection. All the scenes were scripted (ED, 2003). The sequences are divided into two scenarios depending on where they come from: an indoor scene from the entrance lobby of INRIA Labs at Grenoble, and a shopping center from Portugal. The first set consists of 28 clips stemming from a mono-camera. The second set consists of 26 video clips originating from two cameras yielding two views for each scenario.
- **Crowd-11:** The Crowd-11 dataset is a totally annotated dataset that was created by Dupont et al. (2017) for crowded scene classification. The dataset consists of 6272 video sequences captured from a mono-camera. Each sequence is made up of 100 frames each. Each video sequence lasts for 5 s. The resolution depends on each video. The dataset is mainly intended for behavior classification and violence detection. The sequences can be classified into 11 categories, namely: No Crowd, Laminar flow, Turbulent flow, Crossing flows, Merging flow, Diverging Flow, Gas free, Gas jammed, Static agitated, Interacting crowd. This very large dataset originates from video-sharing websites such as Youtube, Pond5 and Gettyimages, and from other datasets, namely Agoraset, UMN, Violent Flows, CUHK Crowd, WWW Crowd Attributes, Shanghai WorldExpo'10 Crowd, Hockey fights and movies, PETS-2009.
- **Motion Emotion:** The Motion Emotion Dataset[33] was created by Rabiee, Haddadnia, Mousavi, Kalantarzadeh et al. (2016) and Rabiee, Haddadnia, Mousavi, Nabi et al. (2016), and is used for anomaly detection in human motions and emotions. The dataset is made up of approximately 44.000 frames that are divided in 31 video clips. The dataset depicts 5 types of motions, which are: panic, fight, dealing with an obstacle, congestion, and neutral behaviors; and 6 types of emotions, which are: angry, happy, excited, scared, sad, and neutral. The dataset is exhaustively annotated. The scenes are recorded from a camera fixed at a height oriented to the ground to capture people walking. The video resolution is of 554 × 235 pixels and consists of 30 frames per second (FPS). The density of the scenes varies from an intermediate to a high concentration of individuals.
- **UCF Crime:** The UCF Crime dataset[34] was created by Sultani et al. (2018). The dataset consists of 1900 long and untrimmed surveillance videos, that totalized 128 h of recording, and that were acquired from Youtube and Liveleak[35] via text queries expressed in different languages. UCF Crime can be used for two tasks: General anomaly detection as we can find normal and abnormal behaviors; and anomalous activities classification, because the abnormal behaviors can be classified into 13 anomalous activities which are: Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. When the authors evaluated methods on this dataset, they could fix the frame rate to 30 FPS and the resolution to 240 × 320 pixels.
- **CCTV-Fights:** The CCTV-Fights dataset[36] was created by Perez et al. (2019). The dataset consists of 1000 temporally annotated videos of real-world fights that required more than 17 h of camera recording. These videos are made up of 280 CCTV videos, whose duration varies between 5 s to 12 min, in addition to 720 videos that come from Non-CCTV sources, whose duration varies between 3 s to 7 min. The resolutions of the videos are diverse.

5.2.6. Group detection and behavior analysis

- **The Friends Meet (FM):** Bazzani et al. (2012) propose The Friends Meet Dataset[37] for their Decentralized Particle Filter technique used for joint pedestrian-group tracking. The Friends Meet is suitable for the development of bottom-up group tracking and detection approaches. The dataset contains 53 sequences broken down into 16286 frames, involving 3 to 16 individuals per frame. The sequences are divided into two groups: a synthetic set, of 28 sequences and 200 frames each, and a real dataset. The synthetic set includes simple and challenging events. The real dataset concerns only outdoor scenarios. The events described within the dataset are: groups appearing, disappearing, going through split/merge and queue events.
- **CUHK Crowd Dataset:** Created by Shao et al. (2014, 2017), the CUHK Crowd dataset[38] contains 474 video clips of different lengths depicting 215 crowded scenes taken from various environments. The dataset is mainly used for mesoscopic group detection: clustering pedestrians into groups and crowd segmentation, eventually group behavior analysis and anomaly detection. It can also be used for crowd statistics. The annotations for group detection and state analysis, and crowd video classification are provided in the dataset.
- **MuseumVisitors** The MuseumVisitors Dataset,[39] created by Bartoli, Lisanti et al. (2015), was recorded by a multi-camera setup of three IP cameras at a resolution of 1280 × 800 pixels at a frame rate of 5 FPS, inside the National Museum of Bargello in Florence, Italy. This challenging dataset, in terms of occlusion, lighting and scale changes, is intended for pedestrian and group detection under occlusion, gaze estimation, behavior analysis,

[30] Violent Flows (ViF): https://www.openu.ac.il/home/hassner/data/violentflows/.
[31] https://talhassner.github.io/home/projects/violentflows/index.html.
[32] Caviar Dataset: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/.
[33] Motion Emotion Dataset (MED): https://github.com/hosseinm/med.
[34] UCF Crime: https://webpages.uncc.edu/cchen62/dataset.html.
[35] https://www.liveleak.com/.
[36] CCTV-Fights: http://rose1.ntu.edu.sg/Datasets/cctvFights.asp.
[37] The Friends Meet Dataset: https://www.iit.it/research/lines/pattern-analysis-and-computer-vision/pavis-datasets/533-friends-meet-dataset.
[38] CUHK Crowd dataset: http://www.ee.cuhk.edu.hk/~jshao/CUHKcrowd_files/cuhk_crowd_dataset.htm.
[39] The MuseumVisitors Dataset: https://www.micc.unifi.it/resources/datasets/museumvisitors/.

person re-identification. The dataset contains two scenarios, the first one depicts individual visitors watching artworks, and the second one depicts groups of visitors watching artworks. Provided annotations are in the form of Bounding Boxes containing the visitors. If a person is occluded, its visible part is surrounded by a second Bounding Box. The Ground Truth furnishes also a unique identifier for each Group and each Pedestrian in all the frames. Seven detectors, pre-trained either on the Caltech pedestrian or the INRIA pedestrian datasets, were assessed on the dataset and improved their miss rate at 10^{-1} false positive per image (FPPI) for Pedestrian and Group detection scenarios.
- **Behave:** Blunsden and Fisher (2010) observed that most of the available ground truthed datasets are used for target detection, tracking, and individual behavior analysis. Consequently, they proposed the Behave Dataset.[40] The dataset is used for pedestrian detection, bottom-up group behavior recognition, and abnormal behavior analysis. It consists of 4 WMV video clips, depicted by 76800 frames. Qualitatively, the resolution is at 640 × 480 at a frame-rate of 25 FPS. There are 125 individuals involved in the actions within the scenes. The detected pedestrians are surrounded by rectangular bounding boxes, which results in 83545 bounding boxes. The annotations are yielded using the Viper-GT[41] tool. The group behavioral interactions are categorized in 10 classes. The annotation policy in this dataset considers the unique pedestrian as the smallest group unit. A group may include many pedestrians that are involved in the same activities and showing a certain proximity. A group activity is labeled by an activity name that happens between two persons. The challenges presented by this dataset are light changes and the recurrent occlusions. A Hidden Markov Model (HMM) classifier had been tested on the dataset. The authors varied several times the size of the window centered on a current frame. Each frame is represented by a feature vector. For a window size of 100, the reached performance is at around 93.67%.
- **SALSA:** Alameda-Pineda et al. (2016) propose the SALSA dataset[42] a dataset dedicated to group detection and behavior analysis, and more precisely for the study of free-standing conversational groups (FCGs). Within the dataset, two scenarios are depicted in an indoor setup: a poster presentation and a cocktail party. In each of them, 18 participants are involved and their behaviors were not scripted. The recorded scenes last for 60 min. The challenges presented by this dataset are the low resolution, the lighting variations, and occlusions. The provided annotations are about each individual's personality as well as their position, head and body orientation, the group F-formation information. The individual's personality annotation are scores related to Extraversion, Agreeableness, Emotional Stability, and Creativity. These scores were obtained following the Big Five personality questionnaire (John & Srivastava, 1999), that each participant filled beforehand.

5.3. Conclusion on the existing sources of data

As you can observe from the range of datasets presented in this section, most of the publicly available datasets can be used for individual-scene action detection/recognition, pedestrian detection/tracking and sometimes group detection/tracking, crowd counting and density estimation, crowded scene classification, few cases of anomaly detection, and very few cases of group activity recognition. However, there is a lack of datasets reproducing crowd motions, in roads or avenues of populated cities. Besides, to the best of our knowledge, there is not any dataset that can be used for massive upstream crowd behavior analysis and motion forecasting that can be utilized to train and test a deep Neural Network.

The access to the video data, provided by live video-surveillance presented in Section 5.1, can widen our possibilities as the crowded situations happens several times with even a high level of service (LoS) in Mecca, Saudi Arabia, often in the Wailing Wall, Jerusalem, Israel (Earthcam), medium to low LoS in St. Pietro Square, Vatican, medium LoS in Times Square, Manhattan, New York, United States (EarthCam), etc. However, publishing results or snapshots of those places may require a prior authorization from the providers of those feeds (see Fig. 12).

6. Annotators

As you may have observed from the precedent section, data scarcity, and more precisely the lack of relevant labeled data, is one of the major problems of crowd analysis. One solution to this problem is the use of annotators assisted with massive crowd-sourcing. In this section, we provide a small list of annotators that can be used for crowd analysis-related tasks. Table 6 summarizes their characteristics.

6.1. Image annotators

Russell, Torralba, Murphy, and Freeman (2008) propose LabelME, a publicly available web-based tool intended for image annotation. Its purpose is to create ground truthed image datasets intended for object detection and recognition tasks.

In the same line, Dutta and Zisserman (2019) propose the VGG Image Annotator (VIA), a web-based tool used to define and describe regions in an image using six different shapes such as: a rectangle (or a bounding box), a point and polylines which can be used for mask annotation. The tool allows a preliminary annotation of images by applying object detectors before altering the annotations manually, which alleviates the burden of manual image annotation. The code source of the VIA tool is publicly available. Although the annotator can be used for video annotation like face tracking, the authors mention that this feature will become officially a part of the software in its upcoming updates.

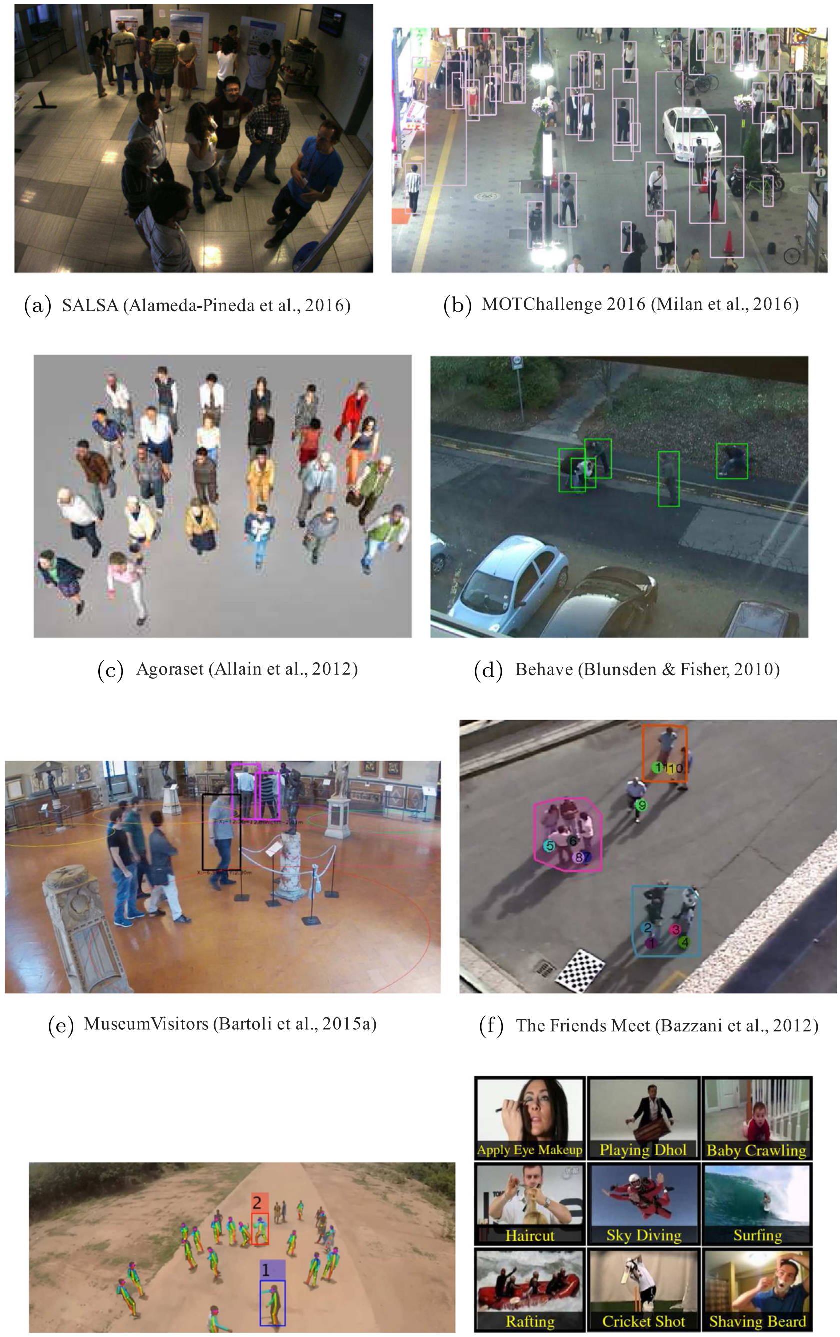
Andriluka, Uijlings, and Ferrari (2018) propose Fluid Annotation. An image annotator that relies on three principles: Machine-learning aid which means that a pre-annotation is performed by a machine learning model; a full image annotation in a single pass, which means that many tasks are performed during image annotation such as drawing bounding boxes and the segmentation of the image, which appears to be more handy than the previously presented VIA tool; empower the annotator which means that the annotator tool sees itself what it is suitable to annotate allowing the human expert to intervene only on the errors made by the tool. This reduces the workload of the human annotator. For instance, the authors compared their use of both of Fluid Annotation and LabelMe, and concluded that annotation time is reduced by a factor of 3× when using their own tool.

Despite being three times quicker than LabelMe, and more handy than VIA, Fluid Annotation does not seem to perform video annotation contrary to VIA.

6.2. Group and crowd behavior annotators

Bartoli et al. (2017) propose PACE an open-source collaborative tool used for the annotation of crowded scenes, that is an improvement of WATSS (Bartoli, Seidenari et al., 2015), and is targeted to indoor multi-camera setups and group behavior understanding. Developed using HTML5 and Javascript, it has two back-ends: a PHP-based that can mainly be summarized to a relational database, and a Python-based REST server that is used for computer vision tasks. The tool allows to determine person location and identity (through respectively an

[40] Behave Dataset: http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/.
[41] Viper-GT: http://viper-toolkit.sourceforge.net/.
[42] SALSA dataset: http://tev.fbk.eu/salsa.



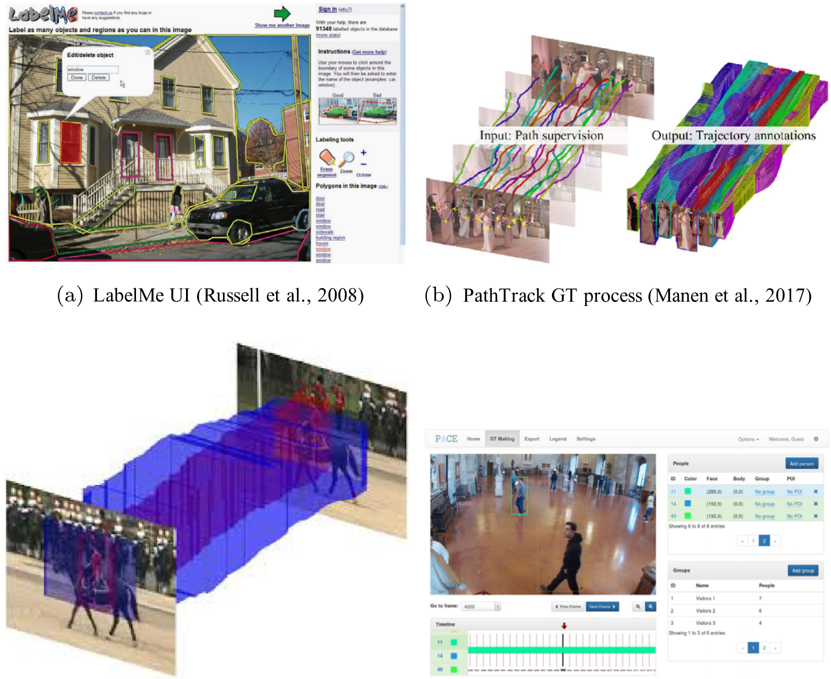
(a) SALSA (Alameda-Pineda et al., 2016)
(b) MOTChallenge 2016 (Milan et al., 2016)
(c) Agoraset (Allain et al., 2012)
(d) Behave (Blunsden & Fisher, 2010)
(e) MuseumVisitors (Bartoli et al., 2015a)
(f) The Friends Meet (Bazzani et al., 2012)
(g) Aerial Violent Individual (AVI) (Singh et al., 2018)
(h) UCF-101 (Soomro et al., 2012)

Fig. 12. Snapshots of some of the aforementioned datasets.

Table 6
Summarized presentation of annotators.

Annotator	Release date	Used for	Availability	Reference
LabelMe	2008	Object detection, image segmentation	Public	Russell et al. (2008)
VIA	2017	Object detection, image segmentation	Public	Dutta and Zisserman (2019)
Fluid Annotation	2018	Object detection, image segmentation	Public	Andriluka et al. (2018)
PathTrack	2017	Pedestrian detection/tracking	Private	Manen et al. (2017)
Pace	2015	Group/pedestrian detection, gaze and body orientation	Public	Bartoli et al. (2017)
SpotOn	2016	Action detection/recognition	Private	Mettes, van Gemert, and Snoek (2016)
WATSS	2015	Group/pedestrian detection	Public	Bartoli, Seidenari, Lisanti, Karaman, and Del Bimbo (2015)

embedding bounding box and bounding box inside), occluded parts, group membership, gaze and body orientation. Although, it is open for public use to encourage collaborative work, the tool is also supported with predictive algorithms. The predictive annotation works as follows:



(a) LabelMe UI (Russell et al., 2008)

(b) PathTrack GT process (Manen et al., 2017)

(c) SpotOn GT process (Mettes et al., 2016)

(d) Pace UI (Bartoli et al., 2017)

Fig. 13. Examples of annotators' User Interfaces (UI) or Ground Truthing (GT) process.

when a person is annotated on frames $t-k$ to t, its label is inferred on the frames $t+1$ to $t+m$ by using tracking based on the Kalman Filtering (Kalman, 1960). However, motion detection and person detection are respectively based on Mixture of Gaussians (MoG) (Godbehere, Matsukawa, & Goldberg, 2012) and Histogram of Gradients (HoG) (Dalal & Triggs, 2005), methods that are quite outdated.

6.3. Tracking annotators

Manen et al. (2017) propose PathTrack, a realtime trajectory annotator, and aspire to enrich MOTChallenge-like datasets. The annotations are produced by users watching a video-surveillance scene and following targets with a cursor. The annotations are then turned into frame-by-frame bounding boxes. The use of PathTrack halved the misclassification rate on a person matching method trained on Multiple Object Tracking Challenge 2015 (Leal-Taixé et al., 2015), and improved NOMT performance reducing the identity switches by 18% and the fragments by 5%. As aforementioned, the crowd-sourced use of Pathrack allowed the creation of a public dataset, named the PathTrack MOT dataset. This annotator does not seem to be publicly available.

6.4. Action recognition annotators

Mettes et al. (2016) propose Spot On, a spatio-temporal action annotator in video clips. Instead of annotating accurate action boxes, Spot On relies on action proposals inferred by point annotations on a subset of frames. This annotator does not seem to be publicly available.

6.5. Conclusion on the existing annotators

We saw in this section some popular annotators used for important tasks in crowd analysis such as pedestrian detection, tracking, action and behavior recognition (see Fig. 13).

Currently, a lot of annotation tools intended for object detection/tracking and image segmentation are being developed (Andriluka et al., 2018; Dutta & Zisserman, 2019). However, except Spot On and PACE, few annotators are used for action recognition, crowd behavior analysis and other more precised tasks such as group detection and behavior recognition. Besides, we did not found any public or private annotator on other problems intended for less studied, but still important, topics in crowd analysis such as massive crowd motion analysis and behavior recognition.

7. Conclusion and discussion

The deployment of intelligent surveillance systems is linked with the development of smart cities. The use of these systems requires the development of a framework capable of scanning adequately video-surveillance scenes. As video-surveillance occurs most of the time in public areas (Krausz & Bauckhage, 2012), crowd analysis-related methods are becoming highly demanded.

The purpose of a review paper is to provide a panoramic view of a precised field of research through particular lenses. A review paper captures this overview by taking into account recent trends in the field itself and by taking into consideration parent fields. This review paper being dedicated to crowd analysis, we talked about recent trends in this field. Throughout this paper, we explored previous reviews on crowd analysis. We saw recent studies pertaining to pedestrian and group detection, as well as on the branches and several sub-branches of crowd analysis. We enumerated the sources of video/image data we came across, and due to the paucity of datasets, we found it relevant to talk about annotators. These latter are somehow neglected by the research community in some subtopics of crowd analysis which are: crowd statistics, action recognition, and crowd behavior analysis. This review allowed us to find out that group analysis-related tasks are not widely explored using Deep Learning methods, despite their widespread use in crowd analysis. Moreover, upstream massive crowd analysis for motion tracking and/or anomaly detection is not widely explored by the Deep Learning literature, due to the non-existence of relevant datasets.

Future research should focus on creating annotators dedicated to massively ground truthing datasets that can be used for crowded scenes analyses tasks such as crowd behavior recognition, crowd tracking and motion prediction. Moreover, it should focus on the creation of annotated datasets depicting several group activities, and other annotated datasets for crowded scenes classification such as Crowd-11 (Dupont et al., 2017).

CRediT authorship contribution statement

Mounir Bendali-Braham: Conceptualization, Methodology (lead), Investigation, Writing - original draft. **Jonathan Weber:** Methodology (supporting), Supervision, Writing - review & editing (equal), Resources (equal). **Germain Forestier:** Methodology (supporting), Supervision (equal), Writing - review & editing (equal), Resources (equal). **Lhassane Idoumghar:** Supervision (equal), Project administration (equal), Funding acquisition (equal), Resources (equal). **Pierre-Alain Muller:** Supervision (equal), Project administration (equal), Funding acquisition (equal), Resources (equal).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the ANR OPMoPS project (grant ANR-16-SEBM-0004) funded by the French National Research Agency.

References

Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(3), 555–560.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Josa a*, *2*(2), 284–299.

Aggarwal, C. C. (2004). A human-computer interactive method for projected clustering. *IEEE Transactions on Knowledge and Data Engineering*, *16*(4), 448–460.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 961–971).

Alahi, A., Ramanathan, V., & Fei-Fei, L. (2017). Tracking millions of humans in crowded spaces. In *Group and crowd behavior for computer vision* (pp. 115–135). Elsevier.

Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., et al. (2016). Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(8), 1707–1720.

Ali, S., & Shah, M. (2007). A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *2007 IEEE conference on computer vision and pattern recognition* (pp. 1–6). IEEE.

Ali, S., & Shah, M. (2008). Floor fields for tracking in high density crowd scenes. In *European conference on computer vision* (pp. 1–14).

Allain, P., Courty, N., & Corpetti, T. (2012). AGORASET: a dataset for crowd video analysis. In *1st ICPR international workshop on pattern recognition and crowd analysis* (pp. 1–6).

Andriluka, M., Uijlings, J. R., & Ferrari, V. (2018). Fluid Annotation: a human-machine collaboration interface for full image annotation. In *ACM multimedia conference on multimedia conference* (pp. 1957–1966).

Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A. S., & Ferguson, D. (2015). Real-time Pedestrian detection with deep network cascades. In *BMVC, vol. 2* p. 4.

AVSS (2007). 2007 IEEE international conference on advanced video and signal based surveillance (AVSS 2007).

Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., Garcia-Rodriguez, J., Cazorla, M., & Signes-Pont, M. T. (2016). Group activity description and recognition based on trajectory analysis and neural networks. In *2016 international joint conference on neural networks (IJCNN)* (pp. 1585–1592).

Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., Garcia-Rodriguez, J., & Orts-Escolano, S. (2015). Self-organizing activity description map to represent and classify human behaviour. In *2015 international joint conference on neural networks (IJCNN)* (pp. 1–7).

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011). Sequential deep learning for human action recognition. In *International workshop on human behavior understanding* (pp. 29–39). Springer.

Bartoli, F., Lisanti, G., Ballan, L., & Del Bimbo, A. (2018). Context-aware trajectory prediction. In *2018 24th international conference on pattern recognition (ICPR)* (pp. 1941–1946). IEEE.

Bartoli, F., Lisanti, G., Seidenari, L., & Del Bimbo, A. (2017) PACE: Prediction-based annotation for crowded environments. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval* (pp. 121–124).

Bartoli, F., Lisanti, G., Seidenari, L., Karaman, S., & Del Bimbo, A. (2015). Museumvisitors: a dataset for pedestrian and group detection, gaze estimation and behavior understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 19–27).

Bartoli, F., Seidenari, L., Lisanti, G., Karaman, S., & Del Bimbo, A. (2015). Watts: a web annotation tool for surveillance scenarios. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 701–704).

Bazzani, L., Cristani, M., & Murino, V. (2012). Decentralized particle filter for joint individual-group tracking. In *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on* (pp. 1886–1893).

Benenson, R., Mathias, M., Timofte, R., & Van Gool, L. (2012). Pedestrian detection at 100 frames per second. In *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on* (pp. 2903–2910). IEEE.

Benfold, B., & Reid, I. (2011). Stable multi-target tracking in real-time surveillance video. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on* (pp. 3457–3464).

Bera, A., Kim, S., & Manocha, D. (2018). Modeling trajectory-level behaviors using time varying pedestrian movement dynamics. *Collective Dynamics*, *3*, 1–23.

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. In *Proceedings - International conference on image processing, ICIP, 2016-Augus* (pp. 3464–3468).

Bisagno, N., Zhang, B., & Conci, N. (2018). Group lstm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European conference on computer vision (ECCV)*.

Blunsden, S., & Fisher, R. (2010). The BEHAVE video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, *4*(1–12), 4.

Borja-Borja, L. F., Saval-Calvo, M., & Azorin-Lopez, J. (2017). Machine learning methods from group to crowd behaviour analysis. In I. Rojas, G. Joya, & A. Catala (Eds.), *Advances in computational intelligence* (pp. 294–305).

Branch, H. O. S. D. (2006). Imagery library for intelligent detection systems (i-lids). In *2006 IET conference on crime and security* (pp. 445–448). IET.

Brostow, G. J., & Cipolla, R. (2006). Unsupervised bayesian detection of independent motion in crowds. In *2006 IEEE computer society conference on computer vision and pattern recognition, vol. 1* (pp. 594–601). IEEE.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), 679–698.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer vision and pattern recognition (CVPR), 2017 IEEE conference on* (pp. 4724–4733). IEEE.

Chaaraoui, A. A., Climent-Pérez, P., & Flórez-Revuelta, F. (2012). A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, *39*(12), 10873–10888.

Chan, A. B., Liang, Z.-S. J., & Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1–7). IEEE.

Chan, A. B., & Vasconcelos, N. (2008). Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(5), 909–926.

Chatzis, S. P., & Kosmopoulos, D. (2015). A nonparametric bayesian approach toward stacked convolutional independent component analysis. In *Proceedings of the IEEE international conference on computer vision* (pp. 2803–2811).

Chau, D., Bremond, F., & Thonnat, M. (2009). Online evaluation of tracking algorithm performance. In *3rd international conference on imaging for crime detection and prevention* (pp. 1–6).

Chen, Y., Nguyen, T. V., Kankanhalli, M., Yuan, J., Yan, S., & Wang, M. (2014). Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(11), 1992–2003.

Chen, M., Wang, Q., & Li, X. (2017a). Anchor-based group detection in crowd scenes. In *Acoustics, speech and signal processing (ICASSP), 2017 IEEE international conference on* (pp. 1378–1382).

Chen, M., Wang, Q., & Li, X. (2017b). Patch-based topic model for group detection. *Science China. Information Sciences*, *60*(11), Article 113101.

Choi, W., Shahid, K., & Savarese, S. (2009). What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Computer vision workshops (ICCV workshops), 2009 IEEE 12th international conference on* (pp. 1282–1289).

Chong, Y., & Tay, Y. (2015). Modeling representation of videos for anomaly detection using deep learning: A review. arXiv:1505.00523.

Coscia, P., Castaldo, F., Palmieri, F. A., Alahi, A., Savarese, S., & Ballan, L. (2018). Long-term path prediction in urban scenarios using circular distributions. *Image and Vision Computing*, *69*, 81–91.

Cui, J., Zha, H., Zhao, H., & Shibasaki, R. (2008). Multi-modal tracking of people using laser scanners and video camera. *Image and vision Computing*, *26*(2), 240–252.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on, vol. 1* (pp. 886–893).

Davies, A. C., Yin, J. H., & Velastin, S. A. (1995). Crowd monitoring using image processing. *Electronics & Communication Engineering Journal*, *7*(1), 37–47.

Dehghan, A., Modiri Assari, S., & Shah, M. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4091–4099).

Dollár, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(8), 1532–1545.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761.

Doretto, G., Chiuso, A., Wu, Y. N., & Soatto, S. (2003). Dynamic textures. *International Journal of Computer Vision*, *51*(2), 91–109.

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, *12*(656–704), 3.

Dupont, C., Tobías, L., & Luvison, B. (2017). Crowd-11: A dataset for fine grained crowd behaviour analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 9–16).

Dutta, A., & Zisserman, A. (2019). The VGG image annotator (VIA). arXiv:1904.10699.

ED (2003). CAVIAR. URL http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/.

Emonet, R., Varadarajan, J., & Odobez, J.-M. (2011). Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model. In *CVPR 2011* (pp. 3233–3240). IEEE.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on image analysis* (pp. 363–370). Springer.

Favaretto, R. M., Dihl, L. L., & Musse, S. R. (2016). Detecting crowd features in video sequences. In *2016 29th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)* (pp. 201–208). IEEE.

Ferryman, J., & Shahrokni, A. (2009). Pets2009: Dataset and challenge. In *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance* (pp. 1–6). IEEE.

FHWA (2004). Traffic analysis tools primer,traffic analysis toolbox. URL http://ops.fhwa.dot.gov/trafficanalysistools/tat-vol1/index.

Gao, Z., Chen, M.-Y., Hauptmann, A. G., & Cai, A. (2010). Comparing evaluation protocols on the KTH dataset. In *International workshop on human behavior understanding* (pp. 88–100). Springer.

Ge, W., Collins, R. T., & Ruback, B. (2009). Automatically detecting the small group structure of a crowd. In *2009 workshop on applications of computer vision (WACV)* (pp. 1–8). IEEE.

Ge, W., Collins, R. T., & Ruback, R. B. (2012). Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(5), 1003–1016.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, *32*(11), 1231–1237.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, *3*(Aug), 115–143.

Godbehere, A. B., Matsukawa, A., & Goldberg, K. (2012). Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In *American control conference (ACC), 2012* (pp. 4305–4312).

Grant, J. M., & Flynn, P. J. (2017). Crowd scene understanding from video: a survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *13*(2), 19.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2255–2264).

Hao, Y., Xu, Z.-J., Liu, Y., Wang, J., & Fan, J.-L. (2019). Effective crowd anomaly detection through spatio-temporal texture analysis. *International Journal of Automation and Computing*, *16*(1), 27–39.

Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), 610–621.

Hassner, T., Itcher, Y., & Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on* (pp. 1–6).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Helbing, D., Farkas, I., & Vicsek, T. (2000). Simulating dynamical features of escape panic. *Nature*, *407*(6803), 487.

Helbing, D., & Molnar, P. (1995). Social force model for pedestrian dynamics. *Physical Review E*, *51*(5), 4282.

Hu, X., Huang, Y., Gao, X., Luo, L., & Duan, Q. (2018). Squirrel-cage local binary pattern and its application in video anomaly detection. *IEEE Transactions on Information Forensics and Security*, *14*(4), 1007–1022.

Huang, C., Nevatia, R., & Li, Y. (2009). Learning to associate: Hybridboosted multi-target tracker for crowded scene. *00*, In *2009 IEEE conference on computer vision and pattern recognition(CVPR)* (pp. 2953–2960).

Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., & Mori, G. (2016). A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1971–1980).

Idrees, H., Saleemi, I., Seibert, C., & Shah, M. (2013). Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2547–2554).

Jhuang, H., Serre, T., Wolf, L., & Poggio, T. (2007). A biologically inspired system for action recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th international conference on* (pp. 1–8).

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, *2*(1999), 102–138.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45.

Kang, D., & Chan, A. (2018). Crowd counting by adaptively fusing predictions from an image pyramid. arXiv:1805.06115.

Karamouzas, I., Heil, P., Van Beek, P., & Overmars, M. H. (2009). A predictive collision avoidance model for pedestrian simulation. In *International workshop on motion in games* (pp. 41–52).

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1725–1732).

Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., et al. (2009). Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 319–336.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., et al. (2017). The kinetics human action video dataset. arXiv:1705.06950.

Ke, Y., Sukthankar, R., & Hebert, M. (2007). Event detection in crowded videos. In *2007 IEEE 11th international conference on computer vision* (pp. 1–8). IEEE.

Kim, J., & Grauman, K. (2009). Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 2921–2928). IEEE.

Kiran, B. R., Thomas, D. M., & Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, *4*(2), 36.

Kong, D., Gray, D., & Tao, H. (2005). Counting pedestrians in crowds using viewpoint invariant training. In *BMVC, vol. 1* (p. 2). Citeseer.

Krausz, B., & Bauckhage, C. (2012). Loveparade 2010: Automatic video analysis of a crowd disaster. *Computer Vision and Image Understanding*, *116*(3), 307–319.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: a large video database for human motion recognition. In *Computer vision (ICCV), 2011 IEEE international conference on* (pp. 2556–2563).

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1–2), 83–97.

Lamba, S., & Nain, N. (2017). Crowd monitoring and classification: a survey. In *Advances in computer and computational sciences* (pp. 21–31). Springer.

Lamba, S., & Nain, N. (2019). Segmentation of crowd flow by trajectory clustering in active contours. *The Visual Computer*, 1–12.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, *64*(2–3), 107–123.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE conference on* (pp. 1–8).

Le, Q., Zou, W. Y., Yeung, S. Y., & Ng, A. Y. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR 2011* (pp. 3361–3368).

Leal-Taixé, L., Milan, A., Reid, I., Roth, S., & Schindler, K. (2015). MOTChallenge 2015: Towards a benchmark for multi-target tracking. arXiv:1504.01942.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lerner, A., Chrysanthou, Y., & Lischinski, D. (2007). Crowds by example. *Computer Graphics Forum*, *26*(3), 655–664.

Leyva, R., Sanchez, V., & Li, C. (2017). The LV dataset: A realistic surveillance video dataset for abnormal event detection. In *2017 5th international workshop on biometrics and forensics (IWBF)* (pp. 1–6).

Li, T., Chang, H., Wang, M., Ni, B., & Hong, R. (2015). Crowded scene analysis : A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, *25*(3), 367–386.

Li, X., Chen, M., Nie, F., & Wang, Q. (2017). A multiview-based parameter free framework for group detection. In *AAAI* (pp. 4147–4153).

Li, B., Liang, X., & Jin, L. (2016). Video classification via spatial-temporal subspace learning. In *2016 6th international conference on digital home (ICDH)* (pp. 38–43). IEEE.

Li, Y., Liu, H., Zheng, X., Han, Y., & Li, L. (2019). A top–bottom clustering algorithm based on crowd trajectories for small group classification. *IEEE Access*, *7*, 29679–29698.

Li, W., Mahadevan, V., & Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(1), 18–32.

Li, H., Wu, Z., & Zhang, J. (2016). Pedestrian detection based on deep learning model. In *Image and signal processing, biomedical engineering and informatics (CISP-BMEI), international congress on* (pp. 796–800).

Li, Y., Zhang, X., & Chen, D. (2018). Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1091–1100).
Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings - 30th IEEE conference on computer vision and pattern recognition* (pp. 936–944).
Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).
Lin, S., Yang, H., Tang, X., Shi, T., & Chen, L. (2019). Social MIL: Interaction-aware for crowd anomaly detection. In *2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1–8). IEEE.
Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21–37).
Liu, W., Salzmann, M., & Fua, P. (2019). Context-aware crowd counting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5099–5108).
Lucas, B. D., Kanade, T., et al. (1981). *An iterative image registration technique with an application to stereo vision*. Vancouver, British Columbia.
Luo, P., Tian, Y., Wang, X., & Tang, X. (2014). Switchable deep network for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 899–906).
Lyu, S., Chang, M.-C., Du, D., Wen, L., Qi, H., Li, Y., et al. (2017). UA-DETRAC 2017: Report of AVSS2017 & IWT4S challenge on advanced traffic monitoring. In *Advanced video and signal based surveillance (AVSS), 2017 14th IEEE international conference on* (pp. 1–7).
Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 1975–1981). IEEE.
Malon, T., Roman-Jimenez, G., Guyot, P., Chambon, S., Charvillat, V., Crouzil, A., et al. (2018). Toulouse campus surveillance dataset: scenarios, soundtracks, synchronized videos with overlapping and disjoint views. In *Proceedings of the 9th ACM multimedia systems conference* (pp. 393–398).
Manen, S., Gygli, M., Dai, D., & Van Gool, L. (2017). Pathtrack: Fast trajectory annotation with path supervision. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 290–299).
Marsden, M., McGuinness, K., Little, S., & O'Connor, N. E. (2016a). Fully convolutional crowd counting on highly congested scenes. arXiv:1612.00220.
Marsden, M., McGuinness, K., Little, S., & O'Connor, N. E. (2016b). Holistic features for real-time crowd behaviour anomaly detection. In *Image processing (ICIP), 2016 IEEE international conference on*, (pp. 918–922).
Marsden, M., McGuinness, K., Little, S., & O'Connor, N. E. (2017). ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In *IEEE international conference on advanced video and signal based surveillance*.
Mehran, R., Oyama, A., & Shah, M. (2009). Abnormal crowd behavior detection using social force model. In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on* (pp. 935–942).
Mettes, P., van Gemert, J. C., & Snoek, C. G. (2016). Spot on: Action localization from pointly-supervised proposals. In *European conference on computer vision* (pp. 437–453).
Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. arXiv:1603.00831.
Milan, A., Roth, S., & Schindler, K. (2014). Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(1), 58–72.
Mousavi, H., Nabi, M., Kiani, H., Perina, A., & Murino, V. (2015). Crowd motion monitoring using tracklet-based commotion measure. In *Image processing (ICIP), 2015 IEEE international conference on* (pp. 2354–2358).
Onoro-Rubio, D., & López-Sastre, R. J. (2016). Towards perspective-free object counting with deep learning. In *European conference on computer vision* (pp. 615–629).
Pellegrini, S., Ess, A., Schindler, K., & Van Gool, L. (2009). You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International conference on computer vision* (pp. 261–268). IEEE.
Penmetsa, S., Minhuj, F., Singh, A., & Omkar, S. N. (2014). Autonomous UAV for suspicious action detection using pictorial human pose estimation and classification. *Electronic Letters on Computer Vision and Image Analysis*, *13*(1), 18–32.
Perez, M., Kot, A. C., & Rocha, A. (2019). Detection of real-world fights in surveillance videos. In *IEEE international conference on acoustics, speech and signal processing* (pp. 2662–2666).
Peter, T., Richard, J. M., Murray, M., & Krause, A. (2013). Robot navigation in dense human crowds: the case for cooperation. In *Robotics and automation (ICRA), 2013 IEEE international conference on* (pp. 2153–2160).
Porikli, F., Bremond, F., Dockstader, S. L., Ferryman, J., Hoogs, A., Lovell, B. C., et al. (2013). Video surveillance: past, present, and now the future [DSP Forum]. *IEEE Signal Processing Magazine*, *30*(3), 190–198.
Qasim, T., & Bhatti, N. (2019). A low dimensional descriptor for detection of anomalies in crowd videos. *Mathematics and Computers in Simulation*, *166*, 245–252.
Rabaud, V., & Belongie, S. (2006). Counting crowded moving objects. *1*, In *2006 IEEE Computer society conference on computer vision and pattern recognition* (pp. 705–711). IEEE.
Rabiee, H., Haddadnia, J., Mousavi, H., Kalantarzadeh, M., Nabi, M., & Murino, V. (2016). Novel dataset for fine-grained abnormal behavior understanding in crowd. In *2016 13th IEEE international conference on advanced video and signal based surveillance* (pp. 95–101).
Rabiee, H., Haddadnia, J., Mousavi, H., Nabi, M., Murino, V., & Sebe, N. (2016). Emotion-based crowd representation for abnormality detection. arXiv:1607.07646.
Ramos, J., Nedjah, N., de Macedo Mourelle, L., & Gupta, B. B. (2017). Visual data mining for crowd anomaly detection using artificial bacteria colony. *Multimedia Tools and Applications*, 1–23.
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.
Ranjan, R. K., & Agrawal, A. (2016). Video summary based on F-sift, tamura textural and middle level semantic feature. *Procedia Computer Science*, *89*, 870–876.
Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E., & Sebe, N. (2016). Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. arXiv:1610.00307.
Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., & Sebe, N. (2017). Abnormal event detection in videos using generative adversarial nets. In *Image processing (ICIP), 2017 IEEE international conference on* (pp. 1577–1581).
Ravanbakhsh, M., Sangineto, E., Nabi, M., & Sebe, N. (2017). Training adversarial discriminators for cross-channel abnormal event detection in crowds. arXiv:1706.07680.
Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).
Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1–3), 157–173.
Sadeghian, A., Alahi, A., & Savarese, S. (2017). Tracking the untrackable: Learning to track multiple cues with long-term dependencies. *4*, (5), (p. 6). arXiv:1701.01909.
Sam, D. B., Peri, S. V., Sundararaman, M. N., Kamath, A., & Babu, R. V. (2019). Locate, size and count: Accurately resolving people in dense crowds via detection. arXiv:1906.07538.
Schröder, G., Senst, T., Bochinski, E., & Sikora, T. (2018). Optical flow dataset and benchmark for visual crowd analysis. In *IEEE international conference on advanced video and signal based surveillance* (pp. 1–6).
Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local SVM approach. In *Proceedings of the 17th international conference on pattern recognition, 2004, vol. 3* ICPR 2004, (pp. 32–36). IEEE.
Shao, J., Change Loy, C., & Wang, X. (2014). Scene-independent group profiling in crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2219–2226).
Shao, J., Dong, N., & Zhao, Q. (2018). A real-time algorithm for small group detection in medium density crowds. *Pattern Recognition and Image Analysis*, *28*(2), 282–287.
Shao, J., Kang, K., Change Loy, C., & Wang, X. (2015). Deeply learned attributes for crowded scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4657–4666).
Shao, J., Loy, C. C., & Wang, X. (2017). Learning scene-independent group descriptors for crowd understanding. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*(6), 1290–1303.
Shi, J., & Malik, J. (1998). Motion segmentation and tracking using normalized cuts. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)* (pp. 1154–1160). IEEE.
Shu, T., Todorovic, S., & Zhu, S.-C. (2017). CERN: Confidence-energy recurrent network for group activity recognition. In *2017 IEEE conference on computer vision and pattern recognition*.
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
Sindagi, V. A., & Patel, V. M. (2017). Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *Advanced video and signal based surveillance (AVSS), 2017 14th IEEE international conference on* (pp. 1–6).
Sindagi, V. A., & Patel, V. M. (2018). A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, *107*, 3–16.
Singh, A., Hazarika, D., & Bhattacharya, A. (2017). Texture and structure incorporated scatternet hybrid deep learning network (ts-shdl) for brain matter segmentation. In *International conference on computer vision workshop* (pp. 1181–1188).
Singh, A., & Kingsbury, N. (2017a). Dual-tree wavelet scattering network with parametric log transformation for object classification. In *Acoustics, speech and signal processing (ICASSP), 2017 IEEE international conference on* (pp. 2622–2626).
Singh, A., & Kingsbury, N. (2017b). Efficient convolutional network learning using parametric log based dual-tree wavelet scatternet. In *Computer vision workshop (ICCVW), 2017 IEEE international conference on* (pp. 1140–1147).
Singh, A., & Kingsbury, N. (2017c). Scatternet hybrid deep learning (shdl) network for object classification. In *Machine learning for signal processing (MLSP), 2017 IEEE 27th international workshop on* (pp. 1–6).
Singh, A., & Kingsbury, N. (2018). Generative scatternet hybrid deep learning (g-shdl) network with structural priors for semantic image segmentation. arXiv:1802.03374.
Singh, A., Patil, D., & Omkar, S. (2018). Eye in the sky: Real-time drone surveillance system (DSS) for violent individuals identification using scatternet hybrid deep learning network. arXiv:1806.00746.

Singh, K., Rajora, S., Vishwakarma, D. K., Tripathi, G., Kumar, S., & Walia, G. S. (2020). Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets. *Neurocomputing*, *371*, 188–198.

Siva, P., & Xiang, T. (2010). Action detection in crowd. In *BMVC* (pp. 1–11).

Solera, F., Calderara, S., & Cucchiara, R. (2016). Socially constrained structural learning for groups detection in crowd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(5), 995–1008.

Solmaz, B., Moore, B. E., & Shah, M. (2012). Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(10), 2064–2070.

Song, X., Zhao, H., Cui, J., Shao, X., Shibasaki, R., & Zha, H. (2013). An online system for multiple interacting targets tracking: Fusion of laser and vision, tracking and learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(1), 18.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402.

Spampinato, C., Palazzo, S., & Giordano, D. (2012). Evaluation of tracking algorithm performance without ground-truth data. In *2012 19th IEEE international conference on image processing* (pp. 1345–1348).

Sugimura, D., Kitani, K. M., Okabe, T., Sato, Y., & Sugimoto, A. (2009). Using individuality to track individuals: Clustering individual trajectories in crowds using local appearance and frequency trait. In *2009 IEEE 12th international conference on computer vision* (pp. 1467–1474). IEEE.

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6479–6488).

Sun, L., Jia, K., Chan, T., Fang, Y., Wang, G., & Yan, S. (2014). DL-SFA: Deeply-learned slow feature analysis for action recognition. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 2625–2632).

Tang, Y., Ma, L., Liu, W., & Zheng, W. (2018). Long-term human motion prediction by modeling motion context and enhancing motion dynamic. arXiv:1805.02513.

Taylor, G. W., Fergus, R., LeCun, Y., & Bregler, C. (2010). Convolutional learning of spatio-temporal features. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), *Computer vision – ECCV 2010* (pp. 140–153).

Thida, M., Yong, Y. L., Climent-Pérez, P., Eng, H.-l., & Remagnino, P. (2013). A literature review on video analytics of crowded scenes. In *Intelligent multimedia surveillance* (pp. 17–36). Springer.

Tian, Y., Luo, P., Wang, X., & Tang, X. (2015). Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 1904–1912).

Tomasi, C., & Kanade, T. (1991). Detection and tracking of point features. *Technical report, International Journal of Computer Vision*.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6450–6459).

TRB (2000). Highway capacity manual. Transportation Research Board, National Research Council, Washington, DC.

Tripathi, G., Singh, K., & Vishwakarma, D. (2018). Convolutional neural networks for crowd behaviour analysis: a survey. *The Visual Computer*, 1–24.

UCSD (2013). UCSD Anomaly dataset. URL http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm.

Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*(2), 154–171.

Ullah, H., & Conci, N. (2012). Crowd motion segmentation and anomaly detection via multi-label optimization. In *ICPR workshop on pattern recognition and crowd analysis vol. 75*.

Ullah, H., Khan, S. D., Ullah, M., Cheikh, F. A., & Uzair, M. (2019). Two stream model for crowd video classification. In *2019 8th european workshop on visual information processing (EUVIP)* (pp. 93–98). IEEE.

Ullah, M., Ullah, H., Conci, N., & De Natale, F. G. (2016). Crowd behavior identification. In *2016 IEEE international conference on image processing (ICIP)* (pp. 1195–1199). IEEE.

Vahora, S., & Chauhan, N. (2018). Deep neural network model for group activity recognition using contextual relationship. *Engineering Science and Technology, An International Journal*.

Vascon, S., & Bazzani, L. (2017). Group detection and tracking using sociological features. In *Group and crowd behavior for computer vision* (pp. 29–66). Elsevier.

Vishwakarma, S., & Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, *29*(10), 983–1009.

Voon, W. P., Mustapha, N., Affendey, L. S., & Khalid, F. (2019). Collective interaction filtering approach for detection of group in diverse crowded scenes. *TIIS*, *13*(2), 912–928.

Walach, E., & Wolf, L. (2016). Learning to count with CNN boosting. In *European conference on computer vision* (pp. 660–676).

Walia, G. S., & Kapoor, R. (2016). Recent advances on multicue object tracking: a survey. *Artificial Intelligence Review*, *46*(1), 1–39.

Wan, J., & Chan, A. (2019). Adaptive density map generation for crowd counting. In *Proceedings of the IEEE international conference on computer vision* (pp. 1130–1139).

Wan, J., Kumar, N. S., & Chan, A. B. (2020). Fine-grained crowd counting. arXiv:2007.06146.

Wang, Z., Cheng, C., & Wang, X. (2018). A fast crowd segmentation method. In *2018 international conference on audio, language and image processing (ICALIP)* (pp. 242–245).

Wang, Q., Gao, J., Lin, W., & Yuan, Y. (2019). Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8198–8207).

Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2011). Action recognition by dense trajectories. In *CVPR 2011* (pp. 3169–3176). IEEE.

Wang, W., Lin, W., Chen, Y., Wu, J., Wang, J., & Sheng, B. (2014). Finding coherent motions and semantic regions in crowd scenes: A diffusion and clustering approach. In *European conference on computer vision* (pp. 756–771). Springer.

Wang, X., Ma, K. T., Ng, G.-W., & Grimson, W. E. L. (2011). Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *International Journal of Computer Vision*, *95*(3), 287–312.

Wang, H., & O'Sullivan, C. (2016). Globally continuous and non-Markovian crowd activity analysis from videos. In *European conference on computer vision* (pp. 527–544). Springer.

Wang, L., Shi, J., Song, G., & Shen, I.-f. (2007). Object detection combining recognition and segmentation. In *Asian conference on computer vision* (pp. 189–199).

Wang, X., Tieu, K., & Grimson, E. (2006). Learning semantic scene models by trajectory analysis. In *European conference on computer vision* (pp. 110–123). Springer.

Wang, H., Ullah, M. M., Klaser, A., Laptev, I., & Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British machine vision conference* 124–1.

Wei, X., Du, J., Xue, Z., Liang, M., Geng, Y., Xu, X., et al. (2020). A very deep two-stream network for crowd type recognition. *Neurocomputing*, *396*, 522–533.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *Image processing (ICIP), 2017 IEEE international conference on* (pp. 3645–3649). IEEE.

Wu, H., Sankaranarayanan, A. C., & Chellappa, R. (2010). Online empirical evaluation of tracking algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(8), 1443–1458.

Wu, M., & Schölkopf, B. (2007). A local learning approach for clustering. In *Advances in neural information processing systems* (pp. 1529–1536).

Wu, S., Yang, H., Zheng, S., Su, H., Fan, Y., & Yang, M.-H. (2017). Crowd behavior analysis via curl and divergence of motion trajectories. *International Journal of Computer Vision*, *123*(3), 499–519.

Xie, S., Zhang, X., & Cai, J. (2019). Video crowd detection and abnormal behavior model detection based on machine learning method. *Neural Computing and Applications*, *31*(1), 175–184.

Yamaguchi, K., Berg, A. C., Ortiz, L. E., & Berg, T. L. (2011). Who are you with and where are you going? In *CVPR 2011* (pp. 1345–1352). IEEE.

Yan, L., Zhu, M., & Yu, C. (2019). Crowd video captioning. arXiv:1911.05449.

Yeffet, L., & Wolf, L. (2009). Local trinary patterns for human action recognition. In *Computer vision, 2009 IEEE 12th international conference on* (pp. 492–497).

You, Q., & Jiang, H. (2018). Action4D: Real-time action recognition in the crowd and clutter. arXiv:1806.02424.

Yuan, Y., Lu, Y., & Wang, Q. (2017). Tracking as a whole: Multi-target tracking by modeling group behavior with sequential detection. *IEEE Transactions on Intelligent Transportation Systems*, *18*(12), 3339–3349.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).

Zhan, B., Monekosso, D. N., Remagnino, P., Velastin, S. A., & Xu, L. Q. (2008). Crowd analysis: A survey. *Machine Vision and Applications*, *19*(5–6), 345–357.

Zhang, L., Lin, L., Liang, X., & He, K. (2016). Is faster R-CNN doing well for pedestrian detection? In *European conference on computer vision* (pp. 443–457).

Zhang, J., Liu, Y., & Jiang, J. (2018). Tensor learning and automated rank selection for regression-based video classification. *Multimedia Tools and Applications*, *77*(22), 29213–29230.

Zhang, X., Zhang, Q., Hu, S., Guo, C., & Yu, H. (2018). Energy level-based abnormal crowd behavior detection. *Sensors*, *18*(2), 423.

Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 589–597).

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., et al. (2016). Mars: A video benchmark for large-scale person re-identification. In *European conference on computer vision* (pp. 868–884).

Zhou, B., Tang, X., & Wang, X. (2012). Coherent filtering: Detecting coherent motions from crowd clutters. In *European conference on computer vision* (pp. 857–871). Springer.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *European conference on computer vision* (pp. 391–405).

Zitouni, M. S., Bhaskar, H., & Al-Mualla, M. E. (2016). Robust background modeling and foreground detection using dynamic textures. In *VISIGRAPP (4: VISAPP)* (pp. 403–410).

Zitouni, M. S., Sluzek, A., & Bhaskar, H. (2020). Towards understanding socio-cognitive behaviors of crowds from visual surveillance data. *Multimedia Tools and Applications*, *79*(3), 1781–1799.