Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Information Fusion 64 (2020) 318–335

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects

Francisco Luque Sánchez [a,*], Isabelle Hupont [b], Siham Tabik [a], Francisco Herrera [a]

[a] Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, Granada, Spain
[b] Herta Security, Barcelona, Spain

ARTICLE INFO

Keywords:
Crowd behaviour analysis
Crowd anomaly detection
Crowd emotions
Review
Deep learning
Models fusion

ABSTRACT

Crowd behaviour analysis is an emerging research area. Due to its novelty, a proper taxonomy to organise its different sub-tasks is still missing. This paper proposes a taxonomic organisation of existing works following a pipeline, where sub-problems in last stages benefit from the results in previous ones. Models that employ Deep Learning to solve crowd anomaly detection, one of the proposed stages, are reviewed in depth, and the few works that address emotional aspects of crowds are outlined. The importance of bringing emotional aspects into the study of crowd behaviour is remarked, together with the necessity of producing real-world, challenging datasets in order to improve the current solutions. Opportunities for fusing these models into already functioning video analytics systems are proposed.

1. Introduction

Last years have known a significant increase of crime and terrorism. Video-surveillance has become a crucial tool for preventing violence and crimes, especially in crowded places and events. The number of video cameras installed in both, public and private places, multiplied in the last years. On the other hand, technology advances allowed an unprecedented improvement in video quality but at the cost of higher computational requirements. Manually analysing such volumes of data is impossible, and automatic processing turns out to be essential. In this context, automatic video-surveillance [1] emerges as a research field, attracting the attention of a large community. Areas such as dangerous object detection [2] or face recognition and identification [3] have been broadly studied.

Automatic video-surveillance, including crowd behaviour analysis, is increasingly attracting much attention [4]. This area aims to understand how individuals behave when they are part of a large group, and extract meaningful information from videos in which crowds of people are present. For example, the automatic analysis of the motion flow of pedestrians when accessing a crowded pilgrimage site, or monitoring the behaviour of large amounts of fans in a sport stadium is crucial in order to detect dangerous situations previous to a catastrophe.

State-of-the-art Deep Neural Networks (DNNs), also known as Deep Learning models [5], have demonstrated impressive results in different computer vision tasks and time series analysis. These models are suitable for automatically analysing video sources, and thereby the use of Deep Learning for crowd behaviour analysis is an emerging trend.

A large number of publications have addressed crowd behaviour analysis using Deep Learning techniques in their pipelines [6–8]. Nevertheless, most of these works are sparse and difficult to compare. This is particularly critical when developing new solutions, since it is difficult to gather the previously developed knowledge. This dispersion is due to three different factors: (1) There is a lack of consensus on what crowd behaviour analysis is; (2) the sub-tasks that constitute crowd behaviour analysis are not clearly determined; and (3) there is not an available taxonomy on how these sub-tasks should be organised and addressed.

The three aforementioned problems are caused by the uncertainty that surrounds this research area. The definition of the task itself and most of the related concepts to the topic are ambiguous. In fact, it is not even clear what should be considered as a crowd. For example, a group of twenty persons can be considered as crowd in an underground station, but it will be hardly interpreted as so in other environments, such as a pilgrimage site like the Mecca. For this reason, designing a general purpose model for crowd behaviour analysis is very difficult. This is also problematic in order to perform a thorough review of the published works, since most studies are not comparable to others, because the specific problem they solve is completely different. Similarly, there are very few standard datasets for benchmarking crowd

* Corresponding author.
E-mail addresses: fluque@decsai.ugr.es (F. Luque Sánchez), isabelle.hupont@hertasecurity.com (I. Hupont), siham@ugr.es (S. Tabik), herrera@decsai.ugr.es (F. Herrera).

https://doi.org/10.1016/j.inffus.2020.07.008
Received 8 June 2020; Received in revised form 20 July 2020; Accepted 21 July 2020
Available online 29 July 2020

behaviour analysis algorithms; most of the works make use of their own datasets, conceived for the specific problem they want to solve.

With this paper, our aim is to establish a common taxonomy to properly categorise the sub-tasks of the topic of crowd behaviour analysis. Afterwards, a comprehensive analysis of related works based on this taxonomy will be performed, specially focusing on crowd anomaly detection and emotional aspects of crowds, two key issues that are still under-explored. To sum up, our contributions with this paper are the following:

1. We review the previous categorisations of crowd behaviour analysis, and give clear definitions of the concepts of group and crowd.
2. We propose a novel taxonomy that organises the existing approaches as a pipeline, using the hierarchical relation among the different involved sub-tasks. This taxonomy comes to fix the lack of structure in previous organisations, which were mere enumerations of tasks without explicit relationships among them.
3. We thoroughly analyse the advantages and limitations of the existing public datasets.
4. We perform a comprehensive review of the works that employ Deep Learning to solve the task of crowd anomaly detection, organising them according to the new taxonomy.
5. We expose the importance of taking into account emotional aspects of the crowd in the analysis of its behaviour.
6. We outline the need of overcoming the current limitations of available datasets, going beyond simulated scenarios towards real-world environments.
7. We point out some possible future directions for the fusion of these models into existing video analytics solutions, in order to deploy these models in real-world video-surveillance contexts.

The rest of the paper is organised as follows, and as described in Fig. 1. In Section 2, some preliminary definitions related to crowd behaviour analysis are given, and previous efforts to categorise the works in the area are described. In Section 3, a comprehensive taxonomy based on the hierarchical relation between sub-tasks is proposed. Section 4 provides a review of the available datasets on the topic. Section 5 describes the common metrics employed to evaluate crowd behaviour analysis models. In Section 6, a thorough review of the works using Deep Learning for crowd anomaly detection is conducted, with a numerical comparison between them on the UCSD Pedestrians dataset (the most widely used in the topic). The importance of bringing emotional aspects into crowd anomaly detection is outlined in Section 7. Section 8 discusses main limitations in the field, and Section 9 highlights future directions in which crowd behaviour analysis solutions could be fused with other video analytics in order to build richer systems. Finally, Section 10 exposes conclusions.

2. Crowd behaviour analysis: Concepts and previous work

There is a need for establishing a common ground for the analysis and characterisation of crowd behaviour. The first definition to be remarked is about what should be considered a crowd. A proper way to define the concept of crowd is through its differentiation from the concept of group:

- Group: It consists on a collection of people that can range from a size of two persons to hundreds [9], in mutual presence at a given moment, who are having some form of social interaction [10]. Its members are close to each other, with a similar speed and with a similar direction of motion [11].
- Crowd (or mass): A crowd is a unique large group of individuals sharing a common physical location [4]. It is usually formed when people with the same goal become one single entity, losing their individuality and adopting the behaviour of the crowd entity [12]. Complex crowd behaviours may result from repeated simple interactions among its constituent individuals, i.e., individuals locally coordinate their behaviours with their neighbours, and then the crowd is self-organised into collective motions without external control [13].

Therefore, the definition of these two terms can be clarified by using two different features: the density of individuals, which is higher in crowds; and the relationships and interactions between the individuals, which tend to be higher in groups. A group is usually formed by less people, with a stronger relation and cohesion. A crowd typically refers to a much larger collection of subjects, whose relationships are less stronger, and with an organisation that emerges from the individual interaction between agents.

Once the concept of crowd is properly defined, delimiting the scope of the area becomes the next challenge. There are many different suitable interpretations for the task of crowd behaviour analysis. In general terms, the main goal in the area is to be able to understand how a concentration of individuals behaves, using information retrieved mainly from video sources. However, there are a lot of different aspects in the behaviour of a crowd we can be interested in. For example, the number of subjects on a certain location and how this number varies in a period of time can be useful to prevent dangerous stampedes of pedestrians. Also, the understanding of the predominant motion directions in a moving crowd (e.g. when accessing a sport stadium) may be useful to detect the persons whose movement differs from the main flows and identify their reasons.

Few works have tried to organise the advances in the field by categorising the problems/tasks that it constitutes, even though such categorisations could be very useful for comparing related works that address the same task. To the best of our knowledge, the most popular categorisation was first proposed in [14]. After the original definition, it has been used in subsequent works [4]. According to this categorisation, crowd behaviour analysis is divided into four main problems:

- Crowd behaviour classification. This task involves the identification and classification of behaviours usually known a priori.
- Crowd counting. The works whose aim is to estimate the number of individuals present in a video are included in this category.
- People detection and tracking. It covers the works whose task is to follow the trajectory of pedestrians in a video. Multiple Object Tracking (MOT), when focused on pedestrians, belongs to this subarea. It also covers the tracking of crowds, i.e., the estimation of the flow of a large group of people moving.
- Crowd anomaly detection. In this case, the task is to identify abnormal behaviours in a crowd, not known a priori. It is an anomaly detection approach to the problem.

This categorisation includes the main sub-tasks in the topic, but places all of them at the same level. However, it is clear that there is a hierarchical relation among the four mentioned tasks, and some of them can be performed as consecutive steps of a pipeline. For instance, the crowd counting stage is often performed after the detection and tracking of individuals. The output of crowd counting may also be used as an input feature for detecting abnormal behaviour, such as congestion. In the next section, we propose a categorisation that overcomes the absence of hierarchy of the previous works, while keeping their contributions as different parts of a pipeline.

3. A taxonomy for crowd behaviour analysis

Previous categorisations in the topic of crowd behaviour analysis are mere enumerations of different sub-tasks, without establishing relations between them. With this new taxonomy, our aim is to link related sub-areas by providing a hierarchical relationship between them.

The proposed taxonomy is based on two complementary aspects.

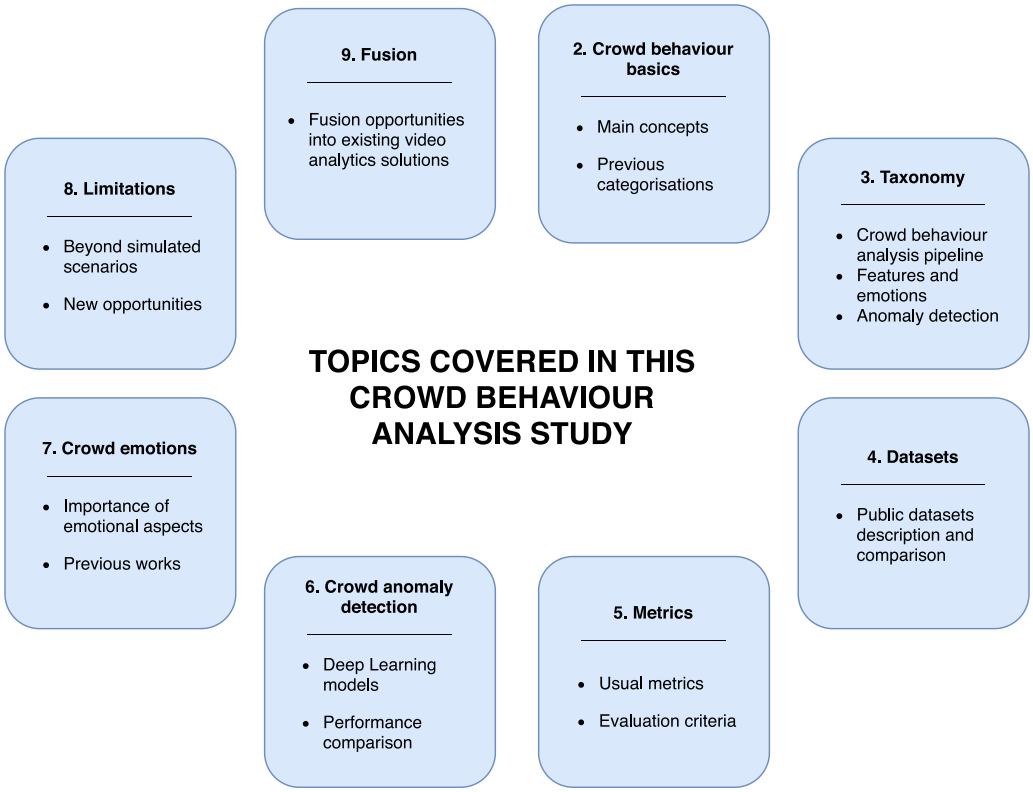


Fig. 1. Organisation and contributions of this study. Numbers indicate corresponding sections in the paper.

- On the one hand, there are two main ways to face the problem of crowd behaviour analysis, depending on the connection between the individuals and the crowd. This is the distinction between macroscopic and microscopic approaches, which will be further analysed.
- On the other hand, independently of the chosen approach, the different related sub-tasks are organised following a pipeline. This pipeline has four main stages, as shown in Fig. 2, in which subsequent stages strongly depend on the previous ones.

As it can be seen in the figure, these two aspects are presented in two different axes, in order to show the interdependence between them. Every stage of the pipeline could be tackled using one of the two approaches. However, there will be subtle differences on the solution depending on the approach chosen.

3.1. Macroscopic vs. microscopic approaches

This first distinction distributes works depending on how individuals are considered in relation to the crowd they belong to. As we have stated, this categorisation is not part of the pipeline, but a complementary classification that heavily influences the different stages in it. Two main approaches can be found:

- Microscopic (or bottom-up) approaches. In these works, the crowd is treated as a collection of individuals. Persons in the video are studied individually, and afterwards the knowledge about these individuals is used to infer information at crowd level [15,16].
- Macroscopic (or top-down) approaches. These are holistic approaches, where the crowd is treated as a whole single entity, without the need of individually segmenting and tracking each individual [17,18].

Usually, microscopic approaches tend to perform better in situations where individuals can be tracked properly. That is, when pedestrians are clearly visible, occlusions are not severe, and density is low. However, when the density of individuals increases, tracking quality degrades significantly. In this circumstance, macroscopic approaches are more suitable, since specific individuals are not the main target of interest and crowds are rather studied globally.

3.2. Crowd behaviour analysis pipeline

Independently of whether a microscopic or macroscopic approach is followed, the pipeline for crowd behaviour analysis includes the four stages depicted in Fig. 2:

1. Detection stage. Its objective is to localise the position of individuals (microscopic settings) and crowds (macroscopic approaches) in each frame. It is a broadly studied sub-task, and several detection models with great performance and accuracy are already available [19].
2. Tracking stage. It aims at uniquely identifying the specific persons and crowd trajectories across a sequence of consecutive frames. Frequently, the dominant flows of movement in the crowd are also determined [20]. Many existing works have successfully tackled this research sub-area [21].
3. Feature extraction stage. It computes a set of metrics that describe the dynamics, topological structure and affective state of the crowd. These metrics can be monitored over time, and computed both at the individual level, when the different subjects are studied independently (microscopic approach), or at crowd level, when a mass of pedestrians is considered as a unique entity (macroscopic approach). Examples include crowd density, velocity and arousal monitoring.
4. Crowd behaviour classification and anomaly detection stage. On the basis of extracted features, this last stage aims at recognising particular behaviours and/or abnormal events in video sequences. There are two main approaches for this stage, depending on the type of learning paradigm employed, supervised

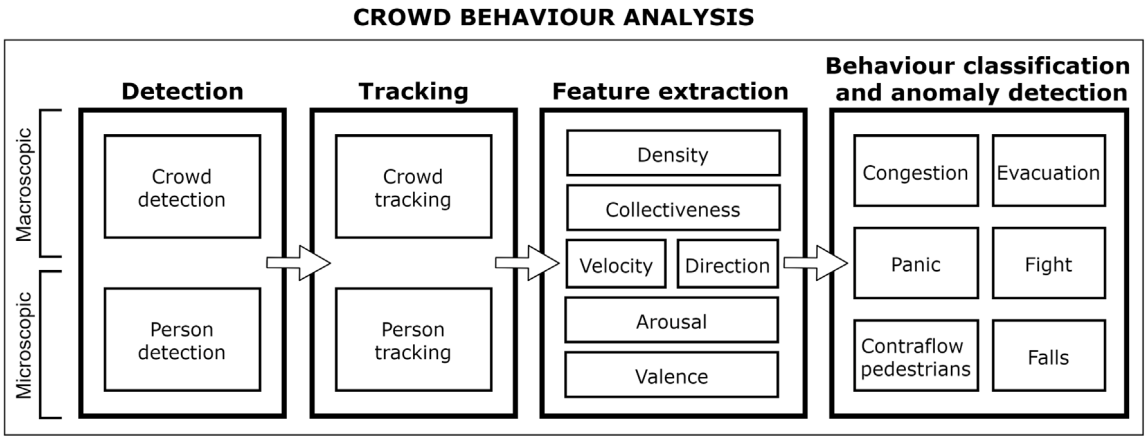


Fig. 2. Proposed taxonomy: The four main stages of the crowd behaviour analysis pipeline.

or unsupervised. Behaviour classification encloses the works that confront the task in a supervised manner. These works previously define a set of behaviours (e.g. people talking, walking together, greeting each other, fighting, snatching, etc.), and train classification models over them. On the other hand, anomalous behaviour detection tries to identify abnormal patterns in the crowd, a priori unknown.

The last two stages of the pipeline will be further described in the following sections. We will not go any deeper into the first two stages of the pipeline, namely detection and tracking, because these two subtasks have been widely developed and studied previously, and thus they are out of the scope of this paper. For example, the reader can refer to [19] for a thorough survey on pedestrian detection, and to [21] for a review on multiple object tracking, with special emphasis on pedestrian tracking.

3.3. Crowd features and related emotional aspects

After the detection and tracking of individuals (or crowds) present in the scene, a feature extraction stage is usually performed. Despite the vast diversity of information that can be extracted from a video sequence, we have identified the following features as very relevant for the understanding of the crowd behaviour:

1. Velocity. It measures the average speed at which individuals (when the approach is bottom-up) or crowds (in top-down approaches) are moving [22].
2. Direction. At the macroscopic level, it determines the number of main directions of movement followed by the crowd [23]. The direction followed by each individual may be extracted as well in microscopic approaches.
3. Density. It quantifies the proximity of individuals in the crowd, determining how dense the crowd is. At the macroscopic level, the objective is to perform density estimation rather than precise people counting, as clutter and severe occlusions make the individual counting problem difficult in very dense crowds.
4. Collectiveness (or cohesion). This feature measures the degree of individuals acting as a union in collective motions [15,24]. When being part of a crowd, instead of behaving independently, individuals tend to follow the behaviours of others and move along the same direction as their neighbours [25]. Then, some spatially coherent structures emerge from the movements of individuals. Collectiveness aims at quantifying the stability of local geometric and topological structures of the crowd. It depends on multiple factors, such as crowd density, velocity, direction and scene structures.
5. Valence. It aims at measuring the positive and negative affect of the crowd. According to the literature on Psychology, it is usually presented as a $[-1;1]$ continuous scale, ranging from unpleasantness to pleasantness.
6. Arousal. It aims at monitoring how calmed or excited the crowd is. It is also presented in a $[-1;1]$ continuous scale, ranging from passive to active.

Note that the latter two features, namely valence and arousal, are related to emotional aspects [26]. Crowd emotions occur when the same event yields a similar appraisal and elicits a common emotion among the members of a crowd [27]. Crowds suffer from strong emotion regulation and contagion processes [28], which are critical to monitor. For example, in crisis situations, negative emotions such as panic, anxiety and fear may spread among the crowd and exert a considerable adverse impact on human decision-making [29]. As a result, stampedes or other catastrophes may occur.

3.4. Crowd behaviour classification and anomaly detection

Extracted features have to be summarised in order to obtain meaningful information about the behaviour of the crowd. As previously pointed out, there are two main approaches in this stage: crowd behaviour classification, where models are trained in a supervised manner over the extracted features; and crowd anomaly detection, when learning is performed in an unsupervised way. Monitoring crowd features over time opens the door to the detection of abnormal behaviours in crowds, since sudden changes in these features are indicative of strange patterns. For example, a sudden drift in crowd speed values is usually an indicator of alert; unwanted congestion can be characterised in terms of lower speed and higher density; and extreme values of valence and arousal can lead to violent situations between groups of people.

When solving the problem from an anomaly detection perspective, another possible organisation emerges from the source of the anomaly itself. As we stated before, the nature of the anomaly may be diverse, and thus the approach to solve the problem may differ slightly. In this work, we have identified five types of anomalies. Four of them were identified when reviewing the specialised literature, and the fifth one is proposed by us, due to its importance despite the lack of works about it:

- Anomalous position. The source of this anomaly comes from an atypical position of an object in the scene. This kind of anomaly occurs, e.g., when a non-authorised individual enters a restricted area, or a pedestrian is detected on a dangerous zone. It is considered the easiest kind of anomaly to detect, since it usually involves just a pedestrian detection stage, combined with bounding box overlapping computation.

Table 1
Public datasets for crowd anomaly detection.

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

[a]Only first video labelled.
[b]Labelled for behaviour classification.
[c]13 different types of uncivil behaviours reported together with normal videos.
[d]Short videos of fights and no-fights.

- Anomalous movement. In this case, the anomalous pattern is produced by an unexpected trajectory of one individual or group in the scene. Two different sources of irregularity can be found: speed, when someone moves faster or slower than his/her surroundings; and direction, when predominant flows exist and the movement of an individual deviates from these trends.
- Anomalous appearance. This abnormality occurs when a non-recognised object enters the scene. A typical example of this anomaly is the presence of a vehicle in a pedestrian path.
- Anomalous action. It is the most difficult anomaly to be identified. It involves the understanding of the usual behavioural patterns of the individuals in the scene, and the detection of non-common ones.
- Anomalous affect. This anomaly is produced by the presence of abnormal or extreme emotions in the crowd. It is the most under-studied topic, due to the lack of properly annotated datasets, but constitutes a promising direction for research, since emotional aspects often arise prior to anomalous situations such as violence.

In practice, the detection of these types of anomalies is often combined. For example, anomalous movement and anomalous appearance are usually tackled together. A clear example of this combination is present in the UCSD Pedestrian dataset [30], which is the most employed one in the literature. In this dataset, two types of anomalies are present: people walking in strange directions (motion) and presence of unauthorised vehicles (appearance). Some examples of different types of anomalies are illustrated in Fig. 3.

3.5. Focus of this study: Why?

As discussed in Section 2, previous surveys have mostly focused on detection, tracking and crowd counting tasks, treating them as independent subareas of study, regardless of their role in the crowd behaviour analysis pipeline. Detection and tracking are undoubtedly key stages of the process, since inaccurate results in these tasks will lead to poorly performing subsequent stages. Similarly, crowd counting is considered to be one of the most critical metrics to extract, since an abnormal high density of people may produce catastrophes such as stampedes and floods, and thus it is one of the most studied features. However, these stages are incomplete without a last stage of crowd behaviour understanding, in which raw features are converted into proper knowledge of the situation. This study will then focus on the two most under-studied aspects of the crowd behaviour analysis pipeline: crowd emotions and crowd anomaly detection.

On the one hand, emotional aspects have been ignored in previous literature on crowds. While valence and arousal have been widely studied in the fields of Affective Computing, Human–Human, Human–Machine and Human–Robot Interaction, they have been mostly limited to the analysis of individuals [31] or at most of small groups of persons [28,32]. There is a need to bring these emotional concepts to crowd analysis.

On the other hand, crowd anomaly detection is one of the most difficult tasks among all the previously discussed, mainly due to the lack of specificity that usually involves anomaly detection approaches. Despite their difficulty, these approaches have an important advantage when compared to classification methods. The problem of supervised learning in behaviour understanding is that human activities and interactions are really diverse, and it is very difficult to properly represent these behaviours in a database. As a consequence, systems will not perform adequately when dealing with behaviours that are not present in the training database. On the contrary, anomaly detection approaches will deal with these situations smoothly, since they will mark the unknown activities as anomalies, triggering and alarm that can be further analysed afterwards. There are scarce reviews in the literature focusing on anomaly detection, and the few that do present traditional Machine Learning techniques [33] instead of state-of-the-art approaches based on Deep Learning.

The rest of the paper will provide a comprehensive review on these two under-explored aspects of crowd behaviour analysis.

4. Datasets for crowd anomaly detection

Due to the complex nature of the crowd behaviour anomaly detection problem, many different datasets that focus on solving diverse tasks are publicly available. In this section, we will categorise these datasets depending on the main task tackled by each one. In Section 4.1, datasets whose main task is motion anomaly detection will be described. Section 4.2 will focus on datasets for action anomaly detection. Note that, as there are not specific datasets for appearance, position and emotion anomaly detection, we will not devote sections to them. Table 1 briefly summarises the main features of the datasets covered in this section.

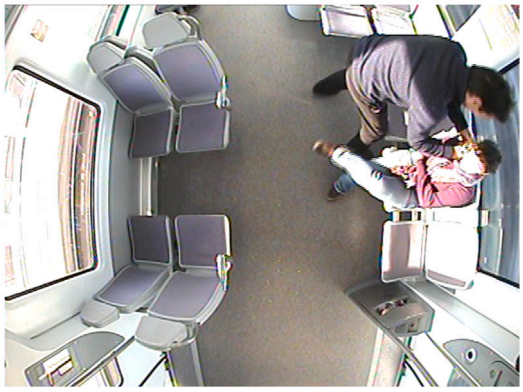
4.1. Datasets for motion anomaly detection

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

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

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

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



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

Fig. 3. Example of different types of crowd anomalies present in publicly available datasets. (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

UCSD Pedestrian dataset. UCSD anomaly detection datasets[1] [30] are the most popular in the literature. There are two different sets of videos, called Peds1 and Peds2. Peds1 contains 34 training and 36 testing video sequences, and Peds2, 16 training and 12 testing sequences. Each clip is approximately 200 frames (20 s) long, with a 158×238 resolution. The main difference between Peds1 and Peds2 is the direction of the moving pedestrians. In Peds1, people walk towards and away from the camera, while in Peds2 individuals move in parallel to the camera plane. No anomalies are present in training videos, which are intended to show what is considered to be a normal behaviour. In testing videos, various abnormal events occur. The frame is considered to be anomalous if there is a non-pedestrian element in the scene (e.g. bikers, skaters, etc.) or if a pedestrian shows an abnormal motion pattern (e.g. somebody running, changing its direction abruptly, and so on). In total, approximately 3400 frames contain anomalies, and 5500 frames are normal. There are two different ways to indicate that an anomaly is present in the frame:

- For each clip, the ground truth includes a binary flag per frame, indicating whether an anomaly is present or not (*frame-level ground truth*).
- In 10 of the videos, a pixel-level mask per frame is provided, which locates the position of the anomaly in the frame (*pixel-level ground truth*).

Despite being the most employed dataset for crowd anomaly detection, it lacks an online leaderboard for algorithm comparison. An example of anomaly in this dataset can be found in Fig. 3, top left side. Most of the works using this dataset solve both motion and appearance anomaly detection in parallel, since the provided annotations do not distinguish between the two different anomalies, and thus there is no way to score both tasks separately.

CUHK Avenue dataset: The CUHK Avenue dataset[2] [34] contains 16 training and 21 testing videos, with 15328 frames for training and 15324 testing. Again, normal samples are formed by people walking in parallel to the camera plane; people moving in other directions, with strange motion patterns or moving vehicles, are considered to be anomalous. In this case, the ground truth for anomalous objects is marked with a bounding box, and the evaluation criteria is the Intersection over Union (IoU) between detection and ground truth [35].

[1] UCSD dataset available at: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm.

[2] The CUHK Avenue dataset can be downloaded at http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html.

UMN dataset: The UMN dataset[3] is a synthetic dataset composed by three different scenes, with a total length of 4 min and 17 s (7725 frames). In each video, an unstructured crowd is walking in the scene, and suddenly everyone starts running, moment that is marked as an anomaly. The objective on this dataset is to accurately detect the change in the movement of the crowd. It can be seen both as motion and behaviour anomaly detection, since the source of anomaly is produced by a sudden change in the speed and direction of people in the scene, but the motion pattern is completely unstructured, in contrast with the structured nature of movement in the previous datasets. Due to this lack of structure, it is not possible to learn the main trends and directions of movement, and thus to use the usual approach for motion anomaly detection. The Web Dataset [36] was proposed as a harder version of UMN dataset, with denser crowds.

ShanghaiTech Campus dataset: The ShanghaiTech Campus dataset[4] [37] is divided into 330 videos for training and 107 videos for testing, taken in 13 different scenarios across the campus. Anomalous events are produced by strange objects in the scene, pedestrians moving at anomalous speed (running or loitering), and moving in unexpected directions.

4.2. Datasets for action anomaly detection

The main task of the datasets presented in this subsection is to identify when a person in the scene presents an abnormal behaviour. Usually, behaviours considered as abnormal are uncivil behaviours such as stealing, fighting, snatching, etc. The most relevant datasets for behaviour anomaly detection are the following:

BEHAVE dataset: The BEHAVE Interactions dataset[5] [38] contains 4 video sequences, of a total length of 2 h. Anomalies are mainly produced by fighting. Only the first sequence is fully annotated. It is divided into 8 fragments, and each fragment is groundtruthed at the frame-level.

CAVIAR dataset: The CAVIAR Test Case Scenarios dataset[6] is a set of videos taken from two different scenes: the entrance hall of a lab building and a hallway in a shopping centre. There are several video sequences for each scenario. In each recording, a person or group of people performs a different action. Most of the anomalies in this dataset are provoked by fighting between pedestrians.

BOSS dataset: The BOSS dataset[7] [39] is a collection of 19 scenes taken inside a moving train, in which groups of people, ranging from single individuals to crowds of more than 10 pedestrians, interact in different manners, both normally and abnormally. For every scene, the action is recorded from different perspectives, using several cameras. Fights, people falling and group panic are examples of anomalies in this dataset.

UT Interactions dataset: The UT Interactions dataset[8] [40] is a collection of 20 videos around 1 min each, presenting six different classes of Human–Human interactions: shake-hands, point, hug, push, kick and punch. All the videos contain several interactions, along with distractor pedestrians. The aim is to correctly detect and classify the type of interaction between subjects. Ground-truth labels for these interactions are provided, including time intervals and bounding boxes.

UCF-Crime dataset: The UCF-Crime dataset[9] [41] was produced in 2018, and contains 1900 videos, 950 of normal events and 950 of abnormal ones, divided into 13 classes: abuse, arrest, arson, assault, road accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting and vandalism. Two main tasks are proposed: detection and localisation of generic anomalies, at a first stage; and specific anomaly classification, at second stage. This dataset is specially relevant due to its large size (more than 13 million samples) and novelty.

Peliculas dataset and Hockey Fight dataset: The Peliculas dataset [42] contains 100 short clips of fights in films, and other 100 clips of normal behaviours. Similarly, the Hockey Fight dataset [43] contains 500 short clips of each class. These two datasets have been typically employed in violence detection systems, which is considered to be an action anomaly detection task.

5. Common evaluation metrics for crowd anomaly detection

The crowd anomaly detection problem can be seen as a binary estimation task. For each frame in the video, a label indicating whether the frame contains an anomaly or not should be generated. Also, it is a heavily unbalanced problem, since the amount of anomalous examples is far scarcer than the amount of normal ones. In this context, the classic metrics that are commonly applied to unbalanced binary classification tasks can be reported. Additionally, as predictions can be given in terms of frames or in terms of areas inside the frame, a distinction between frame-level and pixel-level metrics must be made. This section reviews how both types of metrics apply to crowd anomaly detection.

5.1. Classic metrics

The following classic metrics have been recurrently reported in the reviewed crowd anomaly detection works:

- Accuracy: It is the most common metric for binary classification problems. It is computed as the number of correct predictions divided by the total number of predictions made. The main drawback of this metric is its inadequacy for unbalanced setups.
- Confusion matrix: Given a classification problem with n different classes, the confusion matrix is a $n \times n$ matrix, whose entry a_{ij} is an integer number that represents the amount of elements of class j that have been predicted to belong to class i. Related to this matrix, several metrics are defined. In crowd anomaly detection, three of them are usually reported:
  - True Positive Rate (TPR): Also known as recall, it is the fraction of positive samples that have been correctly identified.
  - False Positive Rate (FPR): It is the fraction of negative samples that have been incorrectly classified.
  - Precision: It computes the fraction of correctly classified elements that belong to the positive class.
  - F-score: It is defined as the harmonic mean of precision and recall.
- Receiver Operating Characteristic (ROC) curve and Area Under the ROC Curve (AUC): The ROC curve is a graphical plot that illustrates the discrimination capability of the model at various discrimination thresholds. In binary classification, it is considered that given two models, a higher AUC indicates higher overall performance.
- Equal error rate (EER): It was initially defined as a performance metric for biometric systems, but now it is widely employed for anomaly detection. It is defined as the operating point at which the miss and false alarm rates are equal. It can be computed directly from the ROC curve as illustrated in Fig. 4.

[3] The UMN dataset is available at http://mha.cs.umn.edu/proj_events.shtml#crowd.

[4] ShanghaiTech Campus dataset can be downloaded at https://svip-lab.github.io/dataset/campus_dataset.html.

[5] The BEHAVE dataset can be downloaded at http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/.

[6] CAVIAR dataset is available at http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/.

[7] The BOSS dataset is available at http://velastin.dynu.com/videodatasets/BOSSdata/index.html.

[8] UT Interactions can be downloaded from https://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html.

[9] UCF-Crime dataset can be downloaded from https://webpages.uncc.edu/cchen62/dataset.html.

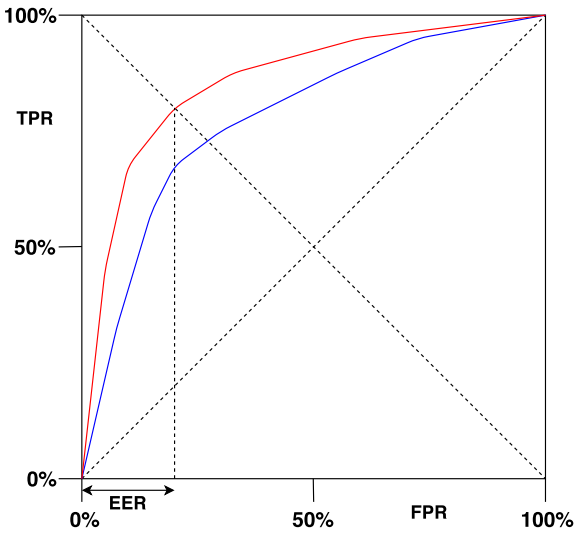


Fig. 4. Example of ROC curves and calculus of the EER for one of them. The model with the red ROC curve is expected to have a better discrimination capability than the blue one, since its AUC is higher. The intersection between the ROC curve and the dotted diagonal marks the point where FPR and TPR are equal, which is defined as the EER. (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

5.2. Frame-level and pixel-level metrics

Since crowd anomaly detection in videos is closely related to image detection, there are different related evaluation criteria that are commonly applied. The two predominant ones are frame-level and pixel-level metrics. In frame-level evaluation, the whole frame is marked as anomalous when an anomaly is present. Detection models are expected to mark the whole frame as abnormal, without stating the exact position at which the anomaly is located. In pixel-level evaluation, instead, the anomaly is marked as a mask over the frame, showing the exact location of the abnormal event. In this case, algorithms are asked to generate a comparable mask, and the anomaly is considered to be correctly identified if the predicted mask is similar to the ground truth one. A typical criterion for this association is the Intersection over Union (IoU). The expected output of the algorithm is a collection of bounding boxes where the anomalies have been detected. Given the ground truth bounding box GT and the predicted bounding box PB, the IoU is defined as:

$$IoU(GT, PB) = \frac{|GT \cap PB|}{|GT \cup PB|}$$

An anomaly is considered to be correctly identified if the IoU between its corresponding GT and predicted bounding box is greater than a predefined threshold, which is usually set to $IoU(GT, PB) \geq 0.5$. Then, results are computed as follows: predicted bounding boxes without a ground truth matching candidate are considered false positives; ground truth boxes marked as anomalous without a matching prediction are considered false negatives; and ground truth boxes marked as normal behaviour without prediction boxes are considered true negatives.

6. Deep learning for crowd anomaly detection: Approaches and numerical analysis

This section presents a thorough review of the works on anomaly detection using Deep Learning. It is structured according to the previous classification of possible anomalies. In Section 6.1, works that perform motion and appearance anomaly detection are shown. We have decided to combine anomalous motion and appearance detection, because these problems are usually tackled together. The most employed dataset for motion anomaly detection, namely UCSD Dataset, marks both types of anomalies in the ground truth indistinctly, and thus the majority of works that use this dataset for evaluation tend to solve both tasks simultaneously. Works that perform action anomaly detection are presented in Section 6.2, while Section 6.3 focuses on position anomaly detection. Finally, in Section 6.4, a numerical comparison of the works that report results on the UCSD Dataset is given.

6.1. Deep learning for motion and appearance anomaly detection

This subsection reviews state-of-the-art deep models for motion and appearance anomaly detection, which have been further grouped depending on the type of approach used to accomplish the task.

6.1.1. Deep feature extraction and one-class SVM classification

A common approach to solve the crowd anomaly detection problem is making use of one-class Support Vector Machines (SVMs) [44] to classify deep features. Deep models are employed as feature extractors, and then a one-class SVM is trained over the extracted features to learn the normal pattern. SVMs learn the smallest region of the feature space that encloses the examples considered to be normal. During inference, new samples located outside the region marked as normal are classified as anomalies.

In [45], three Denoising AutoEncoders (DAE) [46] are trained over original video frames (appearance features), optical flow of the frames (motion features), and concatenation of both inputs (joint appearance and motion features). After the DAEs have been trained over this information, the bottleneck layer (where the most reduced representation of the frames is computed) is used to train three SVMs. Finally, a voting ensemble of the three learned models is used to perform inference. Similarly, in [47], Gutoski et al. trained a Convolutional AutoEncoder (CAE) [48] to reconstruct the input image data using three different channels (original image, edges detected by the Canny Detector and Optical Flow). In a second stage, they calculated the reconstruction error in the three channels separately, and trained the one-class SVM using these three values. In the first work, anomalous regions are marked, which enables pixel-level evaluation. In the second one, whole frames are marked as anomalous.

Authors in [49] proposed an ensemble method of two Stacked DAEs (SDAEs) (see Fig. 5). The first SDAE received the original video as input, while the second received the foreground sequence, calculated using the Kanade–Lucas–Tomasi descriptor [50]. For both inputs, a motion map was calculated and fed into the SDAE. After features extracted, a Deep Belief Network performed a dimensionality reduction and the final classification was given by a one-class SVM for both models. Afterwards, the final anomaly score was computed as a linear combination of both outputs. Authors apply this pipeline both for motion and action anomaly detection, but we have decided to include it in this section since the dataset used to test anomalous action detection is not publicly available.

Fang et al. [51] used two classical features, saliency maps [52] and Multi-scale Histograms of Optical Flow (MHOF) [53], to train a deep network called PCANet [54], whose aim is to perform Cascaded Principal Component Analysis over the input data. After data reduction, a one-class SVM learned the normal pattern. In [55], a pretrained VGG-f network [56] is used as a fast feature extractor, and a SVM is trained over extracted features, resulting in a detection model capable to work at nearly 20 FPS, making it suitable for real-time applications, and achieving a performance close to deeper models. In a similar manner, Sun et al. [57] embedded a one-class SVM layer into a CNN, performing end-to-end training of the whole model.

In [8], an ensemble of several models was proposed. Three different models of CNN were fine tuned as feature extractors (namely VGGNet, AlexNet and GoogLeNet), and their outputs were concatenated to form a feature vector. Afterwards, feature vectors were used to train different SVMs, whose output was combined to perform the final classification.

Huang et al. [58] proposed also an ensemble of features, extracted using Convolutional Restricted Boltzmann Machines. Three different models were trained to extract information from visual patches (regions of original frames), energy patches (feature maps extracted by applying Gaussian filters to input patches), and motion patches (calculated using Optical Flow). After the feature extraction step, all features were fed into a one-class SVM that learned the normal pattern.

Another interesting example can be found in [7]. Authors proposed a model working in two stages. In the first stage, a modified version of Fast R-CNN [59], tuned for multitask learning, was trained in a supervised manner in large scale datasets, so it was able to extract semantic information of the objects in the scene. Specifically, for each object, its class, action and attributes were reported. In the second stage, after the generic semantic extractor was trained, an anomaly detector learned the specific normal pattern for each dataset, and reported an anomaly score for each piece of information in inference time. Different abnormality detectors were tested, getting the best results using one-class SVM. Despite being a general approach, also suitable for anomalous action detection, authors only report results using the motion anomaly dataset, and thus we decided to include it in this section.

6.1.2. Deep feature extraction and Gaussian models

Another approach consists in learning a probabilistic model from features, usually employing Gaussian distributions, and considering as anomalies the samples that deviate from the normal distribution. Again in this case, Deep Learning models are employed as feature extractors.

In [6], two different descriptors were learned. The video was divided into non-overlapping 3D patches (regions of the scene in several subsequent frames), and both local and global descriptors were computed. Local descriptor was a similarity score between the current patch and some adjacent ones, whilst the global descriptor was a sparse representation computed using an autoencoder. In both cases, a Gaussian model was constructed to represent the normal pattern, and new patches were marked as anomalous if both classifiers gave an anomalous response. Same authors refined their work in [60], using the local descriptor as fast rejector of easy patches. When a new patch arrived, the local descriptor classified it. If the answer was normal, the classification was finished. However, if the local descriptor marked the patch as anomalous, it was further processed by the global analyser, marking it as normal or anomalous with more precision.

Additionally, the same authors proposed another cascaded method in [61]. In this case, the first discriminator was computed as the reconstruction error of a shallow autoencoder, used to discard easy patches (specially background patches, whose reconstruction error is low). Afterwards, a deep 3D CNN was trained to detect abnormal situations. Using the CNN extracted features, a Gaussian model was trained, using a Mahalanobis distance between new patches and the Gaussian model as anomaly score.

Feng et al. [62] proposed a method based on 3D gradients. For each video to be classified, 3D gradients were computed (horizontal, vertical and temporal differences of video frames), and a PCANet [54] was trained over these maps. After the PCA features had been computed, a deep Gaussian Mixture Model (Deep GMM) [63] was trained over them. Deep GMMs are an adaptation of Gaussian models to be trained as a Deep Neural Network, performing its optimisation via Gradient Descent. In inference time, if the Deep GMM output is below a threshold, the patch is marked as an anomaly.

6.1.3. Reconstruction-based techniques

The idea behind reconstruction-based techniques relies on training a deep model capable of reconstructing the original image from its compressed representation. After training using normal images, the results of applying these models over abnormal images are irregular, so that an anomaly can be detected from the reconstruction. The measure of irregularity widely employed for this purpose is the reconstruction error.

An example of this approach can be found in Ramchandran et al. [64]. Authors trained a Convolutional AutoEncoder (CAE) with LSTM structure, which is a CAE with a recurrent configuration so that it can deal directly with video fragments instead of individual frames. Particularly, the LSTM-CAE learned to reconstruct normal video fragments from original frames and edges extracted using the Canny detector. When an abnormal video fragment is reconstructed in inference time, the committed error grows noticeably, and thus anomalies are detected if higher than a threshold.

Another example of reconstruction error is presented in [65]. In this model, a temporal CNN with binary output translated an input video into a set of binary feature maps. Using these feature maps, a dictionary of binary codes was computed, and every video block was represented by an histogram of such codes. In inference time, the irregularity of the histogram, measured as the amount of information lost when representing the video using the dictionary, contained information about the degree of abnormality in the video. Moreover, Optical Flow was computed to refine the anomalous region.

6.1.4. End-to-end deep learning approaches

Some authors have trained Deep Learning models whose aim is to output an anomaly score directly, instead of using deep models as feature extractors and then employing another model afterwards. The main advantage of this approach is that models can be trained end-to-end without needing several steps.

Zhou et al. [66] presented a spatio-temporal CNN whose output was the probability of a certain video fragment to contain an anomaly. In their algorithm, each video was divided into small video patches, and the Optical Flow was computed over them. If the patch had a moving object, it was considered to be relevant, and it was further processed by the neural network. If no moving object was detected using Optical Flow, the fragment was directly marked as normal, avoiding useless processing and thus speeding the whole pipeline. This pipeline is also employed for action anomaly detection, but most of the work is focused on motion anomaly, and thus we decided to include it in this section.

In [67], two Generative Adversarial Networks (GANs [68]) were trained. In the first model, the generator had to compute Optical Flow maps from original frames, while in the second one the generator had to perform the inverse task, i.e. computing original frames from Optical Flow. In both cases, GANs were trained only over normal frames. During inference, only the discriminators were used, and since they were not trained over anomalous frames, they tended to mark them as anomalous. The final anomaly score was computed by adding both outputs together.

6.1.5. Other approaches

Some models follow completely different approaches from the previous ones. In [69], a model in cascade was built. In first instance, a shallow CNN discarded easy normal patches (mainly background ones). If the first CNNs could not mark the patch as clearly normal, it was processed by a deeper CNN. The visual features extracted by the CNN were employed to study the motion model of the subject using a flexible Kalman Filter, which is a modification of the classic Karman Filter to measure how much the subject has deviated from the normal motion model. The deviation was considered to be anomalous and it was used as an anomaly indicator.

Hu et al. [70] proposed an interesting solution. In their work, they introduce a new deep model, called D-IncSFA, whose aim is to perform a dimensionality reduction using a technique called Slow Feature analysis (SFA) [71]. This technique performs a dimensionality reduction over a time series, trying to isolate the "most slow varying" feature that defines a time signal. Since the computation of SFA is very time- and resource-consuming, the deep network was trained to compute an approximation. After being trained over normal videos,

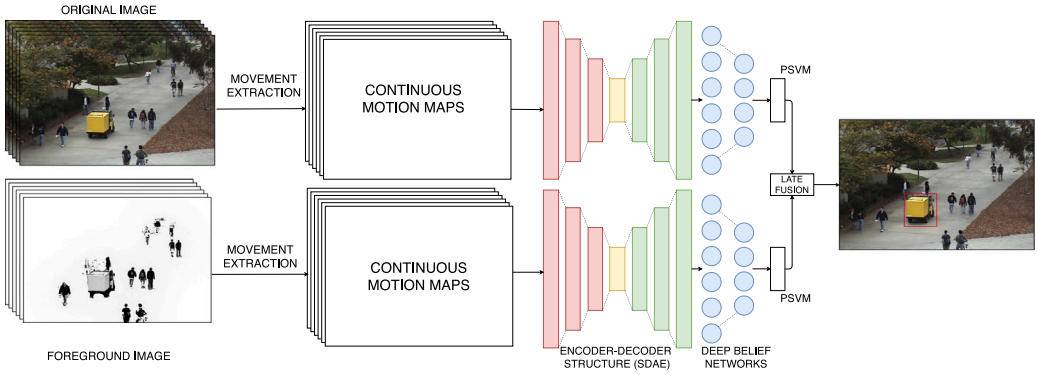


Fig. 5. Example of one-class SVM model for deep feature classification. Two Stacked Denoising Autoencoders receive motion maps from original videos and their corresponding foreground sequences, and try to reconstruct the input map. Afterwards, the one-class SVM model learns the normal pattern, and this information is used in inference time to detect anomalies. Image adapted from method described in [49]. (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

the anomaly score was defined as the square of the derivatives of the output signal. Since the SFA technique tried to reduce this metric when trained over normal videos, when it was applied over anomalous ones it grew noticeably. Two different possible solutions were given using this model. To perform frame-level anomaly detection, the output of the last layer was taken. To perform pixel-level anomaly detection, the output from five different layers was employed, using them as multiscale feature maps.

6.2. Deep learning for action anomaly detection

In this section, models specialised in action anomaly detection are summarised. These works are not further classified since the amount of them is greatly reduced compared to those performing anomalous motion detection.

Following their previous works on motion anomaly detection, [72] proposed a cascaded classifier for anomalous action detection. Using a pretrained fully convolutional neural network (FCN) [73] based on the AlexNet architecture, feature patches at different scales were extracted from the frame. Two different Gaussian classifiers were trained. The first one only used the k shallower layers of the FCN, and a simple discriminator was learned from this information. If the prediction of the classifier was clear (two thresholds marked the clear normal and abnormal behaviours), the classification was considered as completed. If the prediction was unclear, deeper features were computed, and a more specific Gaussian model was employed to refine the decision, at a higher computational cost.

Reconstruction errors have been also employed for anomalous action detection. An example of this kind of models can be found in [74]. In this work, two Stacked Denoising Autoencoders (SDAE) were trained to extract visual and motion features from detected trajectories in a video. After the autoencoders had extracted the mentioned features, a bag of words was computed from them, and the reconstruction error of the original features from the bag of words was considered as an anomaly score.

Authors in [41] proposed an end-to-end Deep learning based solution to the problem (Fig. 6). In their work, the action anomaly detection problem is seen from a multiple instance learning perspective. Since they employ a large scale dataset without per-frame annotations, training videos were split into short sequences, forming a bag of examples, and all were marked as positive examples if there was an anomaly pattern in the complete video (negative if the whole video was normal). Bags were processed by a 3D CNN with a proper loss function to perform multi-instance learning. In inference time, the anomaly score of a whole sequence is computed as the maximum of the predicted values for the video fragments.

In [75], the action anomaly detection problem is solved using a hybrid anomaly detection and behaviour classification approach. A CNN was trained to distinguish among six different classes of anomalous behaviours, given some frames of an anomalous video sequence. Two different experiments were performed in this work. In the first one, the output was a binary label, indicating whether there was an abnormal action present in the video or not. In the second experiment, the actual behaviour category, if present, was also reported.

Violence scenes are widely studied in the context of action anomaly detection. One of the first examples of the use of Deep Learning for violence detection can be found in [76]. In this work, four different traditional feature maps were computed. All of them were derived from Optical Flow representation of input frames. In these feature maps, information about orientation, magnitude and speed of Optical Flow were encoded. This information was fed into a pretrained AlexNet model that extracted visual features. Afterwards, a Relief-F feature selection [77] was employed to reduce data dimensionality, and then two different classifiers were tested, a one-class SVM and a k-NN based model for novelty detection. The same year, Sudhakaran et al. [78] proposed an end-to-end deep model, composed by a FCNN for feature extraction followed by a LSTM for motion model learning. Finally, a group of fully connected layers performed the final classification. The particularity of this model is the resolution of the violence detection as a classification problem, instead of following the anomaly detection approach, but we have decided to include it in this section for the sake of completeness. A visual scheme of this model can be found in Fig. 7.

An interesting solution is provided by Marsden et al. in [79]. Their model was based on a multitask CNN that jointly learned three different tasks: violence detection, crowd counting and density estimation. Outputs from the network were a binary label indicating the presence of violence in the image, a regression neuron that reported the predicted number of subjects in the image, and a heatmap of pedestrians density. As addition, the set of 100 fully annotated images used for training is provided. The main drawback of this model is that it received images (and not videos) as input.

In [80] a 3D CNN based model was proposed. According to the authors, the main contribution of this work was the random frames sampling method. Key frames in the video were detected using a classic clustering technique. A set of 16 random frames was selected between two consecutive key frames, and fed into the 3D CNN to extract spatio-temporal features. These features were then processed by an anomaly detection classifier, though the actual model employed in the work was not reported in the paper. In [81], Dinesh et al. proposed a bidirectional LSTM model, modified in order to work properly in a big data setup. Firstly, the videos were processed inside a Spark engine, and a HoG representation of the frames was computed. This representation was then fed into the LSTM model, that learned the temporal patterns of violent events. Again, the problem was solved as a binary classification. Finally, a benchmark of different models was performed in [82]. The authors tried different models based on CNNs and LSTMs and different

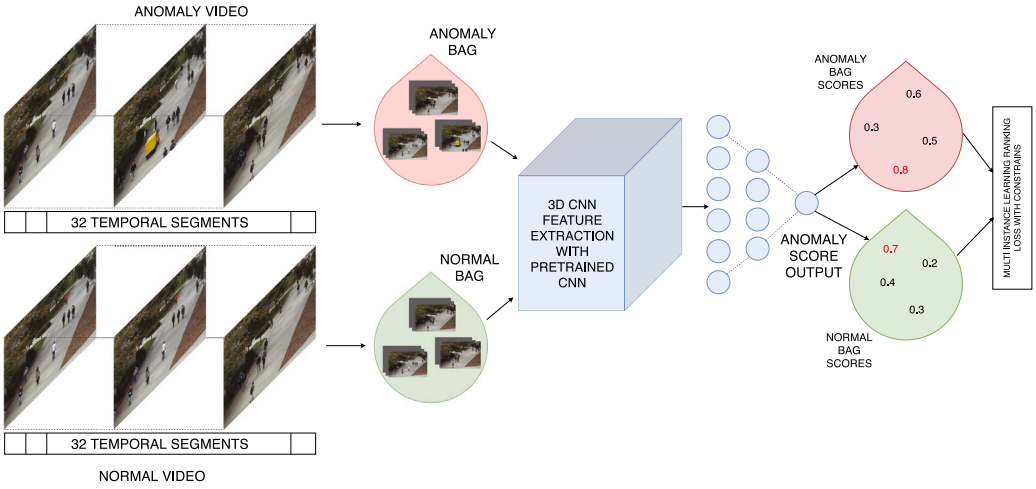


Fig. 6. Example of end-to-end Deep Learning approach for action anomaly detection. In this example, anomaly detection is formulated as a multiple instance learning problem, and the deep model is built to directly output the anomaly score for each video segment. Final anomaly score for a video is the maximum value of the predicted segments that compose the video. Image adapted from method described in [41]. (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

Table 2
Experimental results of crowd anomaly detection algorithms using Deep Learning on UCSD datasets (Ped1 and Ped2).

	UCSD PED1					UCSD PED2				
	FL AUC	FL EER	PL AUC	PL EER	TPR	FL AUC	FL EER	PL AUC	PL EER	TPR
[45]	92.10	16.00	67.20	40.10		90.80	17.00			
[47]			59.00		53.00			61.00		81.00
[7]						90.80	17.10	**87.30**	**19.40**	
[57]	91.40	15.60	69.10	39.30		91.10	16.10			
[8]	93.20					92.10				
[58]	92.60	11.20								
[49]	94.30	10.00	70.30	34.00						
[6]		19.00		**24.00**						
[60]	93.20	8.40			**83.00**	93.90	7.50			**84.00**
[62]	92.50	15.10	69.90	*25.10*						
[64]	**98.40**	**0.75**				**98.50**	**0.92**			
[65]	95.50	8.00	64.50	40.80		88.40	18.00			
[66]	87.00	24.00	**85.00**		*81.30*	88.00	24.40	*86.00*		*81.90*
[67]	*96.80*	*7.00*	*70.80*	34.00		*95.50*	*11.00*			

learning politics, giving a numerical comparison of the results. The dataset was composed by several videos collected from YouTube, where a dense crowd presented some violent episodes. They concluded that the best models, taking into account the lack of data they suffered, were the pretrained ones, specially those for image classification, followed by a fine tuning stage over the training dataset. The models trained from the scratch tended to perform worse, despite being more complex, due to the lack of training samples.

6.3. Deep learning for position anomaly detection

As we have stated before, this type of anomaly is the easiest to detect. An anomaly due to position is produced when an unidentified subject enters a forbidden area. Usually, the pipeline for this problem is simplified to a pedestrian detection step, followed by an overlap calculation between detected bounding boxes and certain regions at the background.

The previous pipeline was employed by Cheng et al. in [83]. In this work, authors considered the anomaly as the trespassing of a forbidden area in a harbour. Videos were recorded by an aerial camera over a harbour, and the constructed model was a combination of a Single-Shot Detector (SSD) fine tuned for pedestrians and an illegal area defined in the background. The system produced an alarm trigger when the area was trespassed by a harbour worker.

6.4. A practical comparison of anomaly detection approaches with the UCSD dataset

Table 2 presents a numerical comparison between existing anomaly detection models. It focuses only on models tested over UCSD datasets, in order to perform a meaningful comparison, since the number of models tested over other datasets is reduced. It is important to note that the reported information has not been double checked, and it is directly extracted from the authors. For each algorithm, the following metrics are reported: frame-level area under ROC curve (FL AUC), frame-level equal error rate (FL EER), pixel-level AUC (PL AUC), pixel-level equal error rate (FL EER) and true positive rate (TPR); both for UCSD Ped1 and UCSD Ped2 sets, when available. Reported metrics are expressed as a percentage. For each column, best result is marked in bold, and second best is underlined.

As we can observe in the table, there are some well-performing models for frame-level anomaly detection. Particularly, [64], a model based on Convolutional AutoEncoders with LSTM structure and using reconstruction error as anomaly score, got top score in metrics AUC and EER for both datasets. Model [67] also obtained high results for both datasets, and it is second best model for frame-level metrics. However, pixel-level anomaly detection results are slightly worse, with the best model getting around 10 points less than the best at frame-level. This means that there is still room from improvement in this aspect.

It is also worth mentioning that no significant differences are found between the different types of algorithms, considering the previous

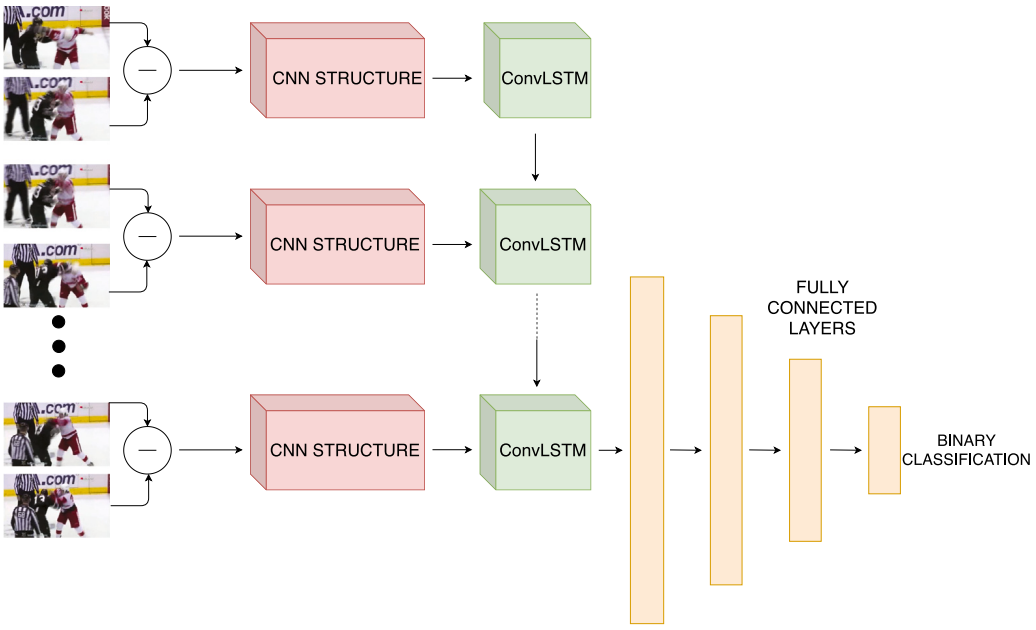


Fig. 7. Example of stacked architecture for violence detection. The model in the image is designated as CNN-LSTM. It consists of a CNN model whose features are fed into an LSTM model that learns the temporal patterns. After temporal relations from visual features have been extracted, a group of fully connected layers produce the final classification. Image adapted from method described in [78]. (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

categorisation. Best performing models are based on reconstruction errors, but the difference with other algorithms is small.

Another important aspect to be remarked is that frame-level metrics are widely employed, but pixel-level metrics are reported in much lesser occasions. This means that it is more common to develop frame-level anomaly detectors than pixel-level ones. However, works that perform pixel-level anomaly detection are far more useful, since the information they provide is more complete: a precise anomaly localisation is crucial to act quickly on a possible threat.

7. Studies of emotions in crowds

The theory of collective emotions in crowds has been widely studied in the field of Psychology [29,84,85], but there is still a need for transferring this knowledge to automatic crowd behaviour analysis. This section presents the challenges that arise when bringing the analysis of crowd affect into practice, and reviews the scarce literature that has been published on the topic.

7.1. From individual to crowd emotions

Affective Computing delves in transferring the theoretical knowledge of emotions and affects into systems capable of recognise, model and express such human aspects [86]. Significant advances in the field have taken place in recent years, but they have mainly focused on single individuals. The most studied human channels for individual emotion recognition are: speech [87], physiological signals [88,89], facial expressions [90,91] and body language [32,92,93]. However, most of these approaches are not transferable to groups and crowds. For example, invasive methods that require people wearing physiological sensors, such as electroencephalography (EEG), galvanic skin response (GSR) or electromyogram (EMG) sensors, are not viable in this context. Approaches grounded on speech (e.g. analysis of prosody or verbal contents) would also be impractical to set up, as it is virtually impossible to capture the sound of the crowd in large outdoor places without a high level of noise and distortion. Video analysis, in the form of facial and bodily expression recognition, turns out as the most appropriate way to tackle the study of affect in crowds, but it is not without limitations either. Existing methods usually require nearly-frontal face/body regions with a resolution above 64 × 64 pixels to accurately identify emotions [94], which is difficult to obtain in a crowded scenario.

Psychological studies on human perception of crowd emotions go in the same direction. When confronted to crowd imagery, we humans mostly focus our attention on facial expressions and body pose of individuals to infer overall crowd affective information. Nevertheless, as facial expressions are not always possible to resolve visually in individuals from afar, body expression is likely to be a more relevant cue [95]. Particularly, findings in [96] demonstrate that the dynamics of body movements play an essential role in the understanding of crowd emotions. Even though all these findings suggest that our brain makes use of a microscopic approach to infer crowd emotion, there is little knowledge on the exact mechanisms that take place to describe and aggregate affective information from individuals to crowd.

Previous studies on the perception of emotion from facial expressions evidence that facial expressions are perceived as categorical, i.e. as belonging to a set of discrete emotional categories such as “joy”, “anger”, “sadness” or “fear” [97,98]. However, it is unknown whether more complex emotional visual stimuli are also perceived in a categorical way. Some works suggest that emotional body language and crowd perception are not [95]. As such, the perception of emotion in groups and crowds may rely on a different description level that has not yet been fully understood by psychologists.

From the Affective Computing perspective, valence and arousal are interesting metrics to explore for the aggregation of emotional information in crowds. The categorical approach is just a discrete list of emotions with no real link between them. It does not represent a dimensional space and has no algebra: each emotion has to be studied and recognised independently. Instead, valence and arousal allow to consider emotions in a continuous affective scale. This continuous approach is attractive because it provides an algebra to aggregate individual contributions [31] and monitoring the intensity of crowd emotion over time.

Overall, there is a need for new inventive ways to address crowd emotion analysis. While both literature on Psychology and Affective Computing seem to point to the microscopic and dynamic analysis of facial and bodily expressions as the most suitable way to tackle the

Fig. 8. Images of groups of people in social events from the HAPPEI database [28]. Each image is manually annotated in terms of group happiness intensity, ranging from 1 (neutral) to 10 (thrilled). The Group Expression Model (GEM) proposed by the authors predicts happiness intensity with close-to-human accuracy. (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

problem, there are still open questions to be answered, such as in which terms crowd affect should be described and measured.

7.2. Emotion analysis in groups

Research in Affective Computing has just started to address the detection of emotions in small groups of persons. The main bottleneck to advance the field is that public datasets showing images/videos involving groups and crowds, such as the ones presented in Section 4.2, do not provide emotional annotations. The only exception is the HAPpy People Images (HAPPEI) dataset [28],[10] which contains 4886 images collect from Flickr and Facebook. Each image is annotated with a 1- to −10 group level mood intensity, corresponding to different stages of happiness: neutral (1), smile (3.5), laugh (6.5) and thrilled (10). HAPPEI was collected for understanding the overall happiness conveyed by a group of people in an image, however it is limited to small groups of at most 15 persons -which is far from crowded situations- and to static images (Fig. 8).

The baseline model proposed by HAPPEI's authors is called Group Expression Model (GEM), and formulates the overall group mood as a weighted average of happiness intensities of all individual faces in the group. Individual faces are analysed using Histogram of Oriented Gradients (HOG) features extracted from the facial region, which are then classified in terms of intensity by an SVM classifier. The contribution of each individual face is weighted depending on several attributes: (i) attributes of the group members such as age, attractiveness and gender; and (ii) context attributes, such as the position of the person inside the group or his/her distance to the camera.

The HAPPEI dataset was then used in the EmotiW 2016 Group-Level emotion recognition challenge [101]. The top performing entry was from Li et al. [102], with a technique based on ensemble of features in Long Short Term Memory (LSTM) and ordinal regression. The first runner-up was the method from Vonikakis et al. [103], which is based on geometric features extracted from faces in an image. Partial least square regression is used to infer the group-level happiness intensity. Sun et al. [104] proposed a LSTM-based approach and fine-tuned the AlexNet model by training it on the HAPPEI database.

The latter work was one of the first to use Deep Learning in its pipeline, but more recent studies also include CNNs in the happiness intensity detection task, either as end-to-end models [99] or as feature extractors [105]. Most popular works utilise two channels of information [99,106]: one channel studying individual faces (microscopic channel) and one channel analysing the whole image (macroscopic channel), as depicted in Fig. 9. Nevertheless, given the small size of the HAPPEI dataset, models often need to be pre-trained on larger datasets (such as ImageNet) and then fine-tuned using HAPPEI [99,104,107].

The emotional analysis of multiple people in terms of valence and arousal still remains an unexplored topic. Mou et al. [100] did some pioneering work on static images of small groups, by using a self-collected dataset of 400 colour images from Flickr and Google (Fig. 10). For each person in an image, they propose to extract a set of facial features (Quantised Local Zernike Moments – QLZM), body features (HOG) and context attributes. These features are then used to feed a k-Nearest Neighbours (k-NN) classifier and detect 3 categories of arousal ("low"–"medium"–"high") and valence ("negative"–"neutral"–"positive"). The final overall group valence and arousal categories are estimated by fusing individual contributions at the decision level. Nevertheless, the fact of discretising valence and arousal into a finite set of categories, instead of using them as a continuous affective scale, makes the approach lose descriptive power.

7.3. Towards emotion analysis in crowds

As we have seen in the previous subsections, most literature on automatic emotion analysis has been limited to single persons or, at most, small groups of persons. The study by Rabiee et al. [108] is the first work tackling the problem of emotion detection in large crowds (Fig. 11). The authors make use of their own private dataset, opening up avenues for both tasks of crowd behaviour classification and emotion recognition, as well as for the analysis of the correlations between these two tasks. Their dataset consists of 31 video clips, each of them annotated with one of 5 crowd behaviour labels ("panic", "fight", "congestion", "neutral" and "obstacle") and one of 6 crowd emotion labels ("angry", "happy", "excited", "scared", "sad" and "neutral"). The

[10] The HAPPEI dataset is available at: https://cs.anu.edu.au/few/Group.htm.

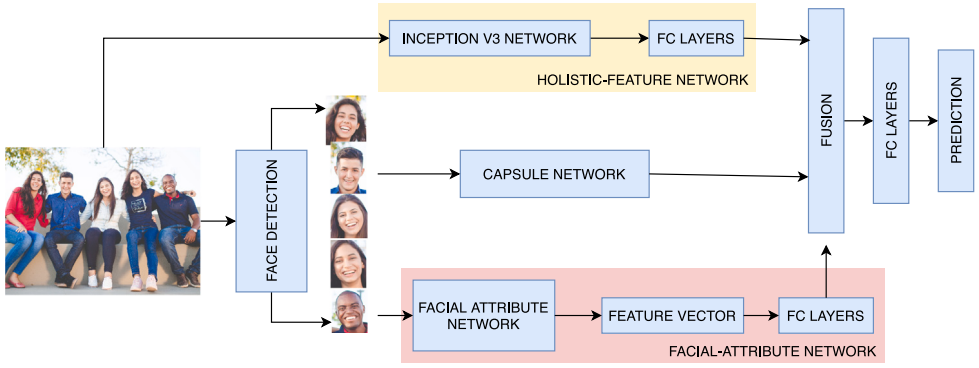


Fig. 9. Example of end-to-end Deep Learning architecture for happiness intensity detection in groups. Top channel analyses the whole scene at the macroscopic level. The bottom channel analyses individual faces, including facial attribute extraction. Image adapted from method described in [99]. (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

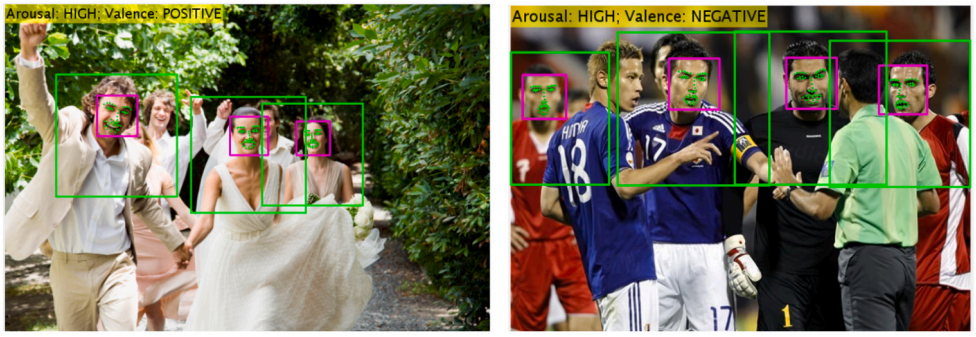


Fig. 10. Images from the private dataset used in [100]. Images are annotated in discrete categories of valence ("negative"–"neutral"–"positive") and arousal ("low"–"medium"–"high"). (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

approach followed for classification is macroscopic, i.e. based on scene features and not on individuals, and takes into account scene dynamics. Their method uses classic Machine Leaning features extracted from the whole scene, including HOG, HOF (Histogram of Optical Flow), MBH (Motion Boundary Histogram) and dense trajectories, which are used to feed a multi-class SVM classifier for the classification of each video. They first detect crowd emotion as mid-level features, and then perform behaviour detection. They prove that by exploiting jointly the complimentary information of these two tasks, all baselines of both tasks significantly outperform. Thus, interestingly, crowd behaviour classification seems to improve when emotional features are taken into account, which is in line with the crowd behaviour analysis pipeline proposed in this paper (c.f. Section 3.3).

To conclude, according to this review, there is a large room for improvement in the field of crowd emotion analysis. Firstly, there is a need for transferring all the psychological knowledge about collective crowd emotions to the field. Secondly, there is a lack of crowd video datasets with emotional annotations, specially in terms of valence and arousal. These datasets could be either built from the scratch or from existing ones (such as UCSD, or any other presented in Table 1), by extending them with affective labels. Finally, Deep Learning methods have not been fully explored yet in the field, and are a very promising line of research.

8. Going beyond simulated scenarios: Limitations of current solutions

Crowd behaviour analysis in video-surveillance sources is a research area experiencing a fast development, gaining increasing attention from the scientific community everyday. However, it is still in its infancy, and there is a large room for improvement on several aspects related to the topic.

8.1. Main limitations

Some of the main problems we have identified while performing this review are the following:

- Lack of definition about the topic
- Lack of realistic datasets available
- Lack of interdisciplinary approaches

As we have remarked in Sections 1 and 3, there is still a lack of consensus about a proper definition of the crowd behaviour analysis problem. This is partly due to the broad interpretation of the terms related to it. This ambiguity makes it easier to enclose various sub-problems within this area of research. These different sub-tasks, despite being closely related, are not exactly identical, and therefore the way to tackle them is not exactly the same. There is a need to organise and categorise all the sub-problems considered in crowd behaviour analysis, in order to ease the research. A precise definition of the problem will simplify work for new researchers, and will facilitate the discovery of new weaknesses and challenges. Our proposed taxonomy is a step towards unifying this branch of knowledge.

Moreover, a clear absence we have detected while gathering information about currently available datasets is that of a standard benchmark with an online table of results or a public maintained leaderboard. UCSD dataset is the standard benchmark for motion anomaly detection, since most of the authors perform experiments over it, but results are not properly gathered and made publicly available. For behaviour anomaly detection, BOSS and BEHAVE datasets are the most employed ones, but comparing results between models also remains a problem for the same reason.

Another problem related to datasets, even more important than the lack of a public standard benchmark, is the fact that available videos are usually recorded in simulated scenarios, and thus they are not

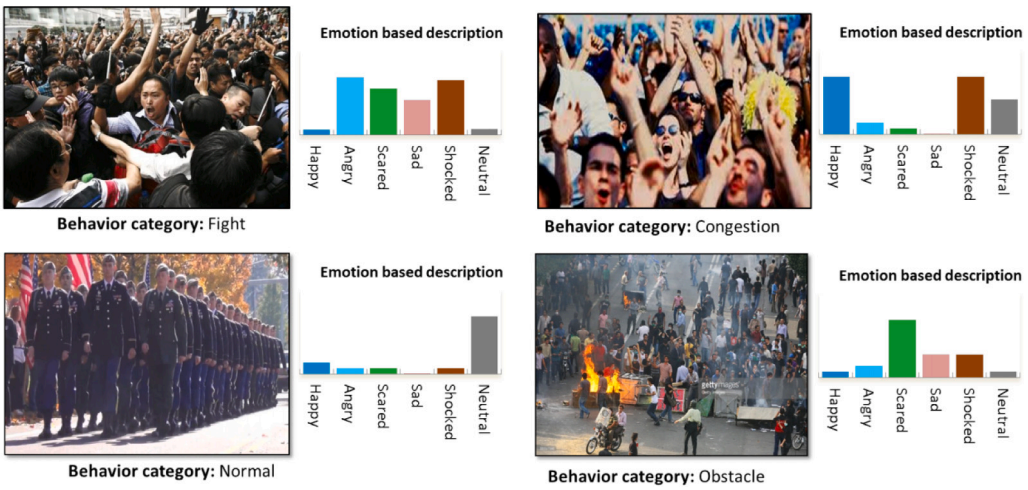


Fig. 11. Emotion-based crowd representation by [108]. (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

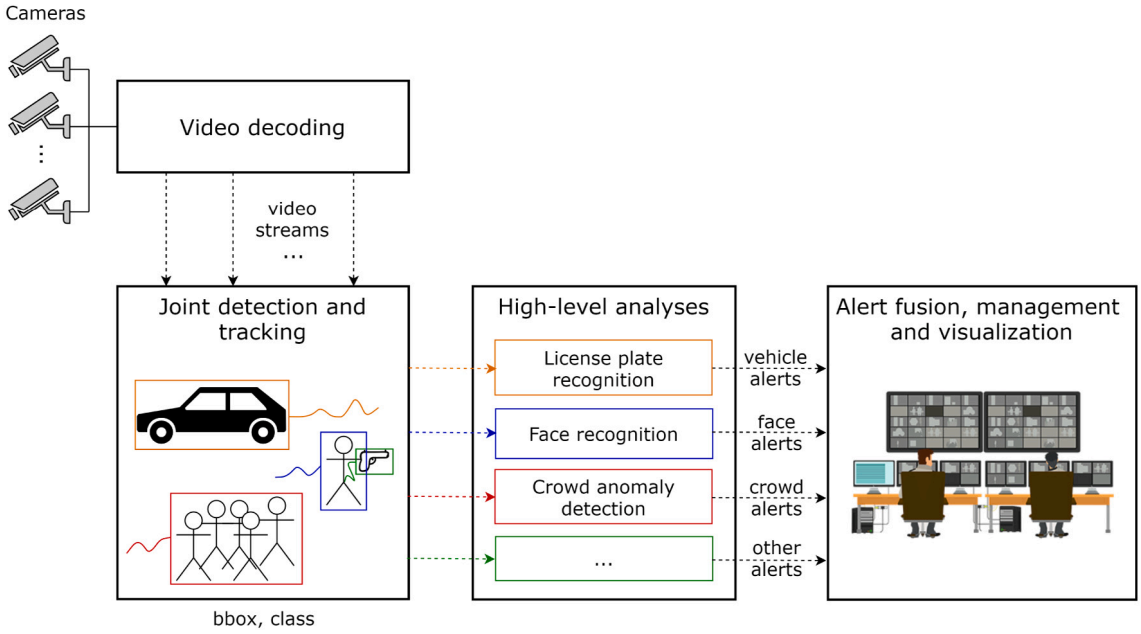


Fig. 12. Efficient fusion of crowd behaviour analysis into a typical video analytics solution. (For a color version of the image, the reader is referred to the web version of this article. Best viewed in color).

realistic, spontaneous and diverse enough. Moreover, most of them are composed by small groups of people rather than actual crowds. The UCF Crime Dataset [41] is a recent effort to provide the scientific community with a dataset for crowd behaviour understanding, but due to its novelty it has not been widely employed in research works for the moment. In addition, to the best of our knowledge, this dataset also lacks a public table of results or leaderboard.

Finally, the lack of interdisciplinary approaches is also a relevant issue. The problem of crowd behaviour analysis is clearly related to several areas of knowledge. There is a lot of research about collective human behaviour and interactions in other fields, like Psychology or Social Studies. However, this knowledge is not widely employed when facing the problem from the perspective of Machine Learning, even further when Deep Learning is employed. Leveraging the advances made in the topic from other approaches would improve the obtained results by a large margin. Furthermore, a close cooperation between researchers and law enforcement bodies (such as Polices) would also bring great benefits to both parts. Researchers would have the opportunity to test and refine their models in realistic scenarios, and social forces would experiment important improvements in the technology they currently employ for the video-surveillance of crowded scenarios.

8.2. New opportunities

Nowadays, despite the visible usefulness of automatic video-surveillance, there is a lack of widely available commercial solutions for crowd behaviour analysis. Research in the topic is not sufficiently advanced in order to exploit it at a large scale. There are some partial solutions, which solve very specific tasks, being tested and deployed, but their aim is usually to analyse individuals rather than complete crowds. Solutions for automatic behaviour understanding in crowds are an increasing need.

Also, ethical concerns are continuously rising about the invasion of surveillance technology in our lives, specially regarding automated and AI-driven tools. Interdisciplinary approaches could help in the adoption of such useful technologies in a fair and ethical manner. It is crucial to ensure that the developments in this area respect the privacy, data protection regulations and integrity of the possible individuals to be

monitored. The collaboration between different agents with diverse backgrounds and perspectives is important to pursue this end. In fact, crowd behaviour analysis from a top-down perspective can be a useful approach to preserve the privacy of the individuals being monitored, since they are not individually focused, while performing a crucial task in catastrophe avoidance.

Finally, the irruption of the Covid-19 pandemic has shown the importance of respecting social distancing politics and certain rules of behaviour in public spaces. This is not always easy in a world as global and crowded as ours. Because of that, monitoring congestion and overcrowding in public spaces has become a top priority around the world. There is an urgent need to incorporate fast solutions to this problem into existing video-surveillance and video analytics solutions. In the next section, different opportunities to fuse novel crowd behaviour analyses and previously employed systems will be addressed.

9. Fusion of crowd behaviour analysis into existing video analytics solutions: Prospects

Automatic video-surveillance is becoming a widely employed tool to ensure public safety around the world. However, crowd behaviour analysis, and more particularly crowd anomaly detection solutions, are still underemployed in terms of commercial applications. There are a lot of opportunities to combine steps from our pipeline into already functioning video analytics systems. By fusing crowd analysis, not only the extracted information from a scene will become richer, but also the total amount of resources and computation needed for the solution will be lower than the required when tackling these problems separately.

As an illustration of the previous idea, Fig. 12 depicts how crowd behaviour analysis could be efficiently fused into a typical video analytics solution. The first two steps of the pipeline, namely detection and tracking of crowds, could be performed jointly with those of individuals, objects, vehicles and other targets. This combination of detection and tracking of different objects avoids processing multiple times the same video stream. Detection and tracking are very computationally expensive Deep Learning processes of the video analysis pipeline, and thus this fusion approach would allow saving a large amount of computational resources. This is particularly critical in video-surveillance settings that require real-time responses from law enforcement bodies.

After detection and tracking, video analytics solutions usually implement separate high-level analyses. For instance, face recognition may be performed over each detected individual; in the case of detected vehicles, license plate and car model recognition may take place. When it comes to detected crowds, this high-level analysis stage would include crowd anomaly detection and crowd behaviour classification.

Each of these high-level analyses generate alerts of different nature. E.g., face recognition may notify the name of an identified person; license plate recognition may notify a vehicle license plate number; and crowd anomaly detection could notify about a potentially dangerous abnormal behaviour of the crowd. There is another opportunity for fusion in this step. Simple alerts from different systems can be fused into advanced information about the scene in a subsequent stage. For example, crowd features monitoring and face recognition can be fused to alert about the presence of a blacklisted person or a weapon in the middle of a high-density crowd, or crowd anomaly detection together with face recognition could be used to automatically tag the people participating in a fight.

Another important advantage of fusing different video analyses is the possibility of cross-validating results where applicable, as it is stated in [109]. The use of different models, that may analyse different targets and obtain information at different timings, enables to fill the gaps of missing or inaccurate data, which is important in order to obtain a robust and complete overview of the situation. This task becomes particularly necessary when attempting to patch coverage holes in a deployment through multiple sources of information.

10. Conclusions

In this work, different aspects related to the crowd behaviour analysis topic have been addressed.

Firstly, a new hierarchical taxonomy taking into account the different stages that conform the crowd behaviour analysis problem has been proposed. Previous categorisations of works placed very different tasks at the same level, presenting a mere enumeration of problems that could be faced from the crowd behaviour analysis perspective. This organisation was incomplete, as it did not take into account the relationship between sub-tasks. With our new organisation, the tasks identified by previous authors are considered as different steps of a global pipeline.

Then, we have focused on one of the last stages of this pipeline, namely crowd anomaly detection, by performing a thorough review of the works that tackle this stage using Deep Learning. We have particularly highlighted the need of considering emotional aspects when studying anomalies in crowds, because sudden changes in crowd emotions are usually a precursor of abnormal situations.

We have also discussed the need of improving the current available material for research, specially in terms of datasets. Currently available data is often created artificially, and thus most of the shown behaviours are unrealistic. Moreover, the density of individuals is quite low compared to real scenarios, which makes research results hardly transferable into real solutions.

Finally, some opportunities for the fusion of crowd behaviour analysis solutions into existing video analytics systems have been outlined. New research areas have been identified, specially related to the pandemic the whole world is suffering nowadays.

As a main conclusion, we would like to remark the relevance of the topic at this moment and how important is its fast and adequate development. Both public and private sectors are demanding accurate solutions to monitor the behaviour of crowds, and there is still a large room for improvement as we have demonstrated in this work.

CRediT authorship contribution statement

Francisco Luque Sánchez: Conceptualization, Investigation, Writing - original draft, Writing - review & editing. **Isabelle Hupont:** Conceptualization, Investigation, Writing - original draft, Writing - review & editing. **Siham Tabik:** Conceptualization, Writing - review & editing, Supervision. **Francisco Herrera:** Conceptualization, Writing - review & editing, Supervision.

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.

Acknowledgements

This work was supported by the project DeepSCOP-Ayudas Fundación BBVA a Equipos de Investigación Científica en Big Data 2018, Spain and the Project AI-MARS, Spain (CIEN CDTI Programme, grant number IDI-20181108). S. Tabik was supported by the Ramon y Cajal Programme, Spain (RYC-2015-18136) and I. Hupont by the Torres Quevedo Programme, Spain (PTQ-16-08735).

References

[1] G. Sreenu, M.S. Durai, Intelligent video surveillance: a review through deep learning techniques for crowd analysis, J. Big Data 6 (1) (2019) 48.

[2] R. Olmos, S. Tabik, F. Herrera, Automatic handgun detection alarm in videos using deep learning, Neurocomputing 275 (2018) 66–72.

[3] S. Bashbaghi, E. Granger, R. Sabourin, M. Parchami, Deep learning architectures for face recognition in video surveillance, in: Deep Learning in Object Detection and Recognition, Springer, 2019, pp. 133–154.

[4] H. Swathi, G. Shivakumar, H. Mohana, Crowd behavior analysis: a survey, in: International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT), 2017, pp. 169–178.

[5] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT press, 2016.

[6] M. Sabokrou, M. Fathy, M. Hoseini, R. Klette, Real-time anomaly detection and localization in crowded scenes, in: IEEE conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 56–62.

[7] R. Hinami, T. Mei, S. Satoh, Joint detection and recounting of abnormal events by learning deep generic knowledge, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3619–3627.

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

[9] H. Hung, D. Gatica-Perez, Estimating cohesion in small groups using audio-visual nonverbal behavior, IEEE Trans. Multimed. 12 (6) (2010) 563–575.

[10] O.A.I. Ramírez, G. Varni, M. Andries, M. Chetouani, R. Chatila, Modeling the dynamics of individual behaviors for group detection in crowds using low-level features, in: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2016, pp. 1104–1111.

[11] L. Bazzani, M. Cristani, V. Murino, Decentralized particle filter for joint individual-group tracking, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1886–1893.

[12] B. Hellmuth, Atlas de la psychologie, 1995, Librairie Générale Française–1995. Dictionnaire de sociologie.

[13] M. Moussaid, S. Garnier, G. Theraulaz, D. Helbing, Collective information processing and pattern formation in swarms, flocks, and crowds, Top. Cogn. Sci. 1 (3) (2009) 469–497.

[14] M.S. Zitouni, H. Bhaskar, J. Dias, M.E. Al-Mualla, Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques, Neurocomputing 186 (2016) 139–159.

[15] B. Zhou, X. Tang, X. Wang, Measuring crowd collectiveness, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3049–3056.

[16] X. Zhang, X. Zhang, Y. Wang, H. Yu, Extended social force model-based mean shift for pedestrian tracking under obstacle avoidance, IET Comput. Vis. 11 (1) (2016) 1–9.

[17] Y. Cong, J. Yuan, J. Liu, Sparse reconstruction cost for abnormal event detection, in: CVPR 2011, pp. 3449–3456.

[18] G. Xiong, X. Wu, Y.-l. Chen, Y. Ou, Abnormal crowd behavior detection based on the energy model, in: 2011 IEEE International Conference on Information and Automation, 2011, pp. 495–500.

[19] D.T. Nguyen, W. Li, P.O. Ogunbona, Human detection from images and videos: A survey, Pattern Recognit. 51 (2016) 148–175.

[20] C. Garate, P. Bilinsky, F. Bremond, Crowd event recognition using hog tracker, in: Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2009, pp. 1–6.

[21] G. Ciaparrone, F.L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, F. Herrera, Deep learning in video multi-object tracking: A survey, Neurocomputing 381 (2020) 61–88.

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

[23] S.R. Musse, C.R. Jung, J.C. Jacques Jr, A. Braun, Using computer vision to simulate the motion of virtual agents, Comput. Anim. Virt. Worlds 18 (2) (2007) 83–93.

[24] C. Zhang, K. Kang, H. Li, X. Wang, R. Xie, X. Yang, Data-driven crowd understanding: A baseline for a large-scale crowd dataset, IEEE Trans. Multimed. 18 (6) (2016) 1048–1061.

[25] D.R. Forsyth, Group Dynamics, Cengage Learning, 2018.

[26] J.A. Russell, A circumplex model of affect, J. Personal. Soc. Psychol. 39 (6) (1980) 1161.

[27] O.J. Urizar, E.I. Barakova, L. Marcenaro, C.S. Regazzoni, M. Rauterberg, Emotion estimation in crowds: a survey, International Conference of Pattern Recognition Systems (2017).

[28] A. Dhall, R. Goecke, T. Gedeon, Automatic group happiness intensity analysis, IEEE Trans. Affect. Comput. 6 (1) (2015) 13–26.

[29] G. Zhang, D. Lu, H. Liu, Strategies to utilize the positive emotional contagion optimally in crowd evacuation, IEEE Trans. Affect. Comput. (2018).

[30] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 1975–1981.

[31] I. Hupont, E. Cerezo, S. Baldassarri, Sensing facial emotions in a continuous 2D affective space, in: 2010 IEEE International Conference on Systems, Man and Cybernetics, 2010, pp. 2045–2051.

[32] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, M. Chetouani, Fully automatic analysis of engagement and its relationship to personality in human-robot interactions, IEEE Access 5 (2016) 705–721.

[33] A. Afiq, M. Zakariya, M. Saad, A. Nurfarzana, M.H.M. Khir, A. Fadzil, A. Jale, W. Gunawan, Z. Izuddin, M. Faizari, A review on classifying abnormal behavior in crowd scene, J. Vis. Commun. Image Represent. 58 (2019) 285–303.

[34] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 fps in matlab, in: IEEE International Conference on Computer Vision, 2013, pp. 2720–2727.

[35] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[36] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 935–942.

[37] W. Liu, W. Luo, D. Lian, S. Gao, Future frame prediction for anomaly detection–a new baseline, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6536–6545.

[38] S. Blunsden, R. Fisher, The BEHAVE video dataset: ground truthed video for multi-person behavior classification, Ann. BMVA 4 (1–12) (2010) 4.

[39] S.A. Velastin, D.A. Gómez-Lira, People detection and pose classification inside a moving train using computer vision, in: International Visual Informatics Conference, Springer, 2017, pp. 319–330.

[40] M.S. Ryoo, J. Aggarwal, UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA), in: IEEE International Conference on Pattern Recognition Workshops, vol. 2, 2010, p. 4.

[41] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479–6488.

[42] E.B. Nievas, O.D. Suarez, G.B. Garcia, R. Sukthankar, Movies fight detection dataset, Comput. Anal. Images Patterns (2011) 332–339.

[43] E.B. Nievas, O.D. Suarez, G.B. Garcia, R. Sukthankar, Hockey fight detection dataset, Comput. Anal. Images Patterns (2011) 332–339.

[44] B. Schölkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor, J.C. Platt, Support vector method for novelty detection, Adv. Neural Inf. Process. Syst. (2000) 582–588.

[45] D. Xu, E. Ricci, Y. Yan, J. Song, N. Sebe, Learning deep representations of appearance and motion for anomalous event detection, 2015, arXiv preprint arXiv:1510.01553.

[46] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: 25th International Conference on Machine Learning, ACM, 2008, pp. 1096–1103.

[47] M. Gutoski, N.M.R. Aquino, M. Ribeiro, E. Lazzaretti, S. Lopes, Detection of video anomalies using convolutional autoencoders and one-class support vector machines, in: XIII Brazilian Congress on Computational Intelligence, 2017.

[48] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, International Conference on Artificial Neural Networks (2011) 52–59.

[49] M. Yang, S. Rajasegarar, S. Erfani, C. Leckie, Deep learning and one-class SVM based anomalous crowd detection, in: 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8.

[50] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: 7th International Joint Conference on Artificial Intelligence, 1981, vol. 81.

[51] Z. Fang, F. Fei, Y. Fang, C. Lee, N. Xiong, L. Shu, S. Chen, Abnormal event detection in crowded scenes based on deep learning, Multimedia Tools Appl. 75 (22) (2016) 14617–14639.

[52] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, C.-W. Lin, Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum, IEEE Trans. Multimed. 14 (1) (2011) 187–198.

[53] Y. Cong, J. Yuan, Y. Tang, Video anomaly search in crowded scenes via spatio-temporal motion context, IEEE Trans. Inf. Forensics Secur. 8 (10) (2013) 1590–1599.

[54] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, PCANet: A simple deep learning baseline for image classification?, IEEE Trans. Image Process. 24 (12) (2015) 5017–5032.

[55] S. Smeureanu, R.T. Ionescu, M. Popescu, B. Alexe, Deep appearance features for abnormal behavior detection in video, in: International Conference on Image Analysis and Processing, Springer, 2017, pp. 779–789.

[56] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, 2014, arXiv preprint arXiv:1405.3531.

[57] J. Sun, J. Shao, C. He, Abnormal event detection for video surveillance using deep one-class learning, Multimedia Tools Appl. 78 (3) (2019) 3633–3647.

[58] S. Huang, D. Huang, X. Zhou, Learning multimodal deep representations for crowd anomaly event detection, Math. Probl. Eng. 2018 (2018).

[59] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[60] M. Sabokrou, M. Fathy, Z. Moayed, R. Klette, Fast and accurate detection and localization of abnormal behavior in crowded scenes, Mach. Vis. Appl. 28 (8) (2017) 965–985.

[61] M. Sabokrou, M. Fayyaz, M. Fathy, R. Klette, Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes, IEEE Trans. Image Process. 26 (4) (2017) 1992–2004.

[62] Y. Feng, Y. Yuan, X. Lu, Learning deep event models for crowd anomaly detection, Neurocomputing 219 (2017) 548–556.

[63] C. Viroli, G.J. McLachlan, Deep gaussian mixture models, Stat. Comput. 29 (1) (2019) 43–51.

[64] A. Ramchandran, A.K. Sangaiah, Unsupervised deep learning system for local anomaly event detection in crowded scenes, Multimedia Tools Appl. (2019) 1–21.

[65] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, N. Sebe, Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1689–1698.

[66] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, Z. Zhang, Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes, Signal Process., Image Commun. 47 (2016) 358–368.

[67] M. Ravanbakhsh, E. Sangineto, M. Nabi, N. Sebe, Training adversarial discriminators for cross-channel abnormal event detection in crowds, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1896–1904.

[68] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014, arXiv:1406.2661.

[69] K. Kumar, A. Kumar, A. Bahuguna, D-CAD: Deep and crowded anomaly detection, in: Proceedings of the 7th International Conference on Computer and Communication Technology, ICCCT-2017, ACM, 2017, pp. 100–105.

[70] X. Hu, S. Hu, Y. Huang, H. Zhang, H. Wu, Video anomaly detection using deep incremental slow feature analysis network, IET Comput. Vis. 10 (4) (2016) 258–265.

[71] L. Wiskott, T.J. Sejnowski, Slow feature analysis: Unsupervised learning of invariances, Neural Comput. 14 (4) (2002) 715–770.

[72] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, R. Klette, Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes, Comput. Vis. Image Underst. 172 (2018) 88–97.

[73] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[74] J. Wang, L. Xia, Abnormal behavior detection in videos using deep learning, Cluster Comput. 22 (4) (2019) 9229–9239.

[75] N. Tay, T. Connie, T.S. Ong, K. Goh, P.S. Teh, A Robust Abnormal Behavior Detection Method Using Convolutional Neural Network: 5th ICCST 2018, Kota Kinabalu, Malaysia, 29-30 August 2018, 2019, pp. 37–47.

[76] A. Keçeli, A. Kaya, Violent activity detection with transfer learning method, Electron. Lett. 53 (15) (2017) 1047–1048.

[77] I. Kononenko, E. Šimec, M. Robnik-Šikonja, Overcoming the myopia of inductive learning algorithms with RELIEFF, Appl. Intell. 7 (1) (1997) 39–55.

[78] S. Sudhakaran, O. Lanz, Learning to detect violent videos using convolutional long short-term memory, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1–6.

[79] M. Marsden, K. McGuinness, S. Little, N.E. O'Connor, Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1–7.

[80] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, A. Wang, A novel violent video detection scheme based on modified 3D convolutional neural networks, IEEE Access 7 (2019) 39172–39179.

[81] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, A. Ahilan, et al., Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM, Comput. Netw. 151 (2019) 191–200.

[82] S.A. Sumon, M.T. Shahria, M.R. Goni, N. Hasan, A. Almarufuzzaman, R.M. Rahman, Violent crowd flow detection using deep learning, in: Asian Conference on Intelligent Information and Database Systems, Springer, 2019, pp. 613–625.

[83] G. Cheng, S. Wang, T. Guo, X. Han, G. Cai, F. Gao, J. Dong, Abnormal behavior detection for harbour operator safety under complex video surveillance scenes, in: 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 2017, pp. 324–328.

[84] J.E. McHugh, R. McDonnell, C. O'Sullivan, F.N. Newell, Perceiving emotion in crowds: the role of dynamic body postures on the perception of emotion in crowded scenes, Exp. Brain Res. 204 (3) (2010) 361–372.

[85] C. Von Scheve, S. Ismer, Towards a theory of collective emotions, Emot. Rev. 5 (4) (2013) 406–413.

[86] R.W. Picard, Affective Computing, MIT press, 2000.

[87] G. Varni, I. Hupont, C. Clavel, M. Chetouani, Computational study of primitive emotional contagion in dyadic interactions, IEEE Trans. Affect. Comput. (2017).

[88] A.-C. Conneau, A. Hajlaoui, M. Chetouani, S. Essid, EMOEEG: A new multimodal dataset for dynamic EEG-based emotion recognition with audiovisual elicitation, in: 2017 25th European Signal Processing Conference (EUSIPCO), 2017, pp. 738–742.

[89] G. Fortino, S. Galzarano, R. Gravina, W. Li, A framework for collaborative computing and multi-sensor data fusion in body sensor networks, Inf. Fusion 22 (2015) 50–70.

[90] I. Hupont, M. Chetouani, Region-based facial representation for real-time action units intensity detection across datasets, Pattern Anal. Appl. 22 (2) (2019) 477–489.

[91] M.Z. Uddin, M.M. Hassan, A. Almogren, M. Zuair, G. Fortino, J. Torresen, A facial expression recognition system using robust face features from depth videos and deep learning, Comput. Electr. Eng. 63 (2017) 114–125.

[92] Q. Li, R. Gravina, G. Fortino, Posture and gesture analysis supporting emotional activity recognition, in: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2018, pp. 2742–2747.

[93] R. Gravina, Q. Li, Emotion-relevant activity recognition based on smart cushion using multi-sensor fusion, Inf. Fusion 48 (2019) 1–10.

[94] B. Cheng, Z. Wang, Z. Zhang, Z. Li, D. Liu, J. Yang, S. Huang, T.S. Huang, Robust emotion recognition from low quality and low bit rate video: A deep learning approach, in: 2017 7th International Conference on Affective Computing and Intelligent Interaction (ACII), 2017, pp. 65–70.

[95] J.E. McHugh, R. McDonnell, C. O'Sullivan, F.N. Newell, Perceiving emotion in crowds: the role of dynamic body postures on the perception of emotion in crowded scenes, Exp. Brain Res. 204 (3) (2010) 361–372.

[96] E.M. Huis in 't Veld, B. de Gelder, From personal fear to mass panic: The neurological basis of crowd perception, Human Brain Mapp. 36 (6) (2015) 2338–2351.

[97] A.W. Young, D. Rowland, A.J. Calder, N.L. Etcoff, A. Seth, D.I. Perrett, Facial expression megamix: Tests of dimensional and category accounts of emotion recognition, Cognition 63 (3) (1997) 271–313.

[98] P. Ekman, Darwin and Facial Expression: A Century of Research in Review, ISHK, 2006.

[99] S. Ghosh, A. Dhall, N. Sebe, Automatic group affect analysis in images via visual attribute and feature networks, in: 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 1967–1971.

[100] W. Mou, O. Celiktutan, H. Gunes, Group-level arousal and valence recognition in static images: Face, body and context, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, vol. 5, pp. 1–6.

[101] A. Dhall, R. Goecke, J. Joshi, J. Hoey, T. Gedeon, Emotiw 2016: Video and group-level emotion recognition challenges, in: 2016 18th ACM International Conference on Multimodal Interaction, 2016, pp. 427–432.

[102] J. Li, S. Roy, J. Feng, T. Sim, Happiness level prediction with sequential inputs via multiple regressions, in: 2016 18th ACM International Conference on Multimodal Interaction, 2016, pp. 487–493.

[103] V. Vonikakis, Y. Yazici, V.D. Nguyen, S. Winkler, Group happiness assessment using geometric features and dataset balancing, in: 2016 18th ACM International Conference on Multimodal Interaction, 2016, pp. 479–486.

[104] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, L. Yu, LSTM for dynamic emotion and group emotion recognition in the wild, in: 2016 18th ACM International Conference on Multimodal Interaction, 2016, pp. 451–457.

[105] X. Huang, A. Dhall, R. Goecke, M. Pietikainen, G. Zhao, A global alignment kernel based approach for group-level happiness intensity estimation, 2018, arXiv preprint arXiv:1809.03313.

[106] A. Cerekovic, A deep look into group happiness prediction from images, in: 2016 18th ACM International Conference on Multimodal Interaction, 2016, pp. 437–444.

[107] X. Guo, B. Zhu, L.F. Polanía, C. Boncelet, K.E. Barner, Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions, in: 2018 20th ACM International Conference on Multimodal Interaction, 2018, pp. 635–639.

[108] H. Rabiee, J. Haddadnia, H. Mousavi, M. Nabi, V. Murino, N. Sebe, Emotion-based crowd representation for abnormality detection, 2016, arXiv preprint arXiv:1607.07646.

[109] A.-E.M. Taha, A. Ali, Monitoring a crowd's affective state: Status quo and future outlook, IEEE Commun. Mag. 57 (4) (2019) 26–32.