

Multimedia Datasets for Anomaly Detection: A Review

Pratibha Kumari, Anterpreet Kaur Bedi and Mukesh Saini, *Indian Institute of Technology Ropar, India*

Abstract— Multimedia anomaly datasets play a crucial role in automated surveillance. They have a wide range of applications expanding from outlier objects/ situation detection to the detection of life-threatening events. For more than 1.5 decades, this field has attracted a lot of research attention, and as a result, more and more datasets dedicated to anomalous actions and object detection have been developed. Tapping these public anomaly datasets enable researchers to generate and compare various anomaly detection frameworks with the same input data. This paper presents a comprehensive survey on a variety of video, audio, as well as audio-visual datasets based on the application of anomaly detection. This survey aims to address the lack of a comprehensive comparison and analysis of multimedia public datasets based on anomaly detection. Also, it can assist researchers in selecting the best available dataset for benchmarking frameworks. Additionally, we discuss gaps in the existing dataset and insights for future direction towards developing multimodal anomaly detection datasets.

Index Terms— Anomaly datasets survey, Concept drift, Multimodal anomaly detection, Long term surveillance

I. INTRODUCTION

Anomaly detection corresponds to identifying unexpected or rare events in a dataset [1]. Detection of anomalous events from a scenario can offer important insights into a large number of monitoring and safety-critical real-world applications such as disaster forecast, detection of extreme climate event, mechanical fault, disease outbreak, fire/ blast, fraud, etc. Recently, there has been a significant rise in the use of audio and video sensors for monitoring situations [2]–[4]. Monitoring of scenes by humans is unscalable and error-prone due to inherent human limitations [5], [6]. Therefore, we need to automatically analyse audio/video data to detect anomalous events and objects. Consequently, many frameworks for anomaly detection have been proposed. In order to evaluate and validate a framework, researchers require a comprehensive dataset. Towards this end, numerous datasets have been created in literature for anomaly detection. These are either based on a specific type of anomaly, or for generic anomaly detection.

In this survey paper, we present a comprehensive review on video, audio, and audiovisual datasets available for evaluating anomaly detection frameworks. We compare these datasets based on various perspectives, which will help researchers to choose suitable datasets for an application. To the best of our knowledge, there exists only one short survey paper [7] on video datasets published in 2016. We could not find any survey on audio or audio-visual anomaly datasets. We believe that a consolidated review of audio, video, and audio-visual datasets will promote use of multimedia for anomaly detection. Additionally, there are more than fifteen new datasets contributed to the video anomaly detection domain after the year 2016, which need to be reviewed.

A comprehensive taxonomy of anomaly datasets is given in Figure 1. Based on the recording media, anomaly datasets can be categorized into video, audio, and audio-visual datasets.

Further, we categorize them on the basis of four attributes, viz., area of application, scene-type, anomaly induction mode, and labeling type. The characterization is expanded to include additional attributes in multiple figures and tables across Sections II, III, and IV. This enables a quick comparison of existing anomaly datasets, and thus helpful in selecting most suitable dataset to evaluate the given work. Further, to facilitate an easy access of datasets, we assemble them together and provide brief description including common information such as how and where the data was collected, what are the anomalies, merit/ demerit identified by researchers, etc., along with link to their website via supplementary file attached with this manuscript.

Rest of the paper is organized as follows. Public video datasets for anomaly detection are discussed in the Section II. A discussion about the audio datasets has been provided in Section III. Further, a comparison of public audio-visual datasets is given in Section IV. Section V discusses on future directions. Finally, we conclude the paper in Section VI.

II. VIDEO ANOMALY DATASETS

There are more than 30 publicly available video datasets that are currently being used for the purpose of anomaly detection. Initial datasets are comprised of simple events and scenarios with a very constrained amount of anomalies. The anomalies are performed by a group of actors and therefore lack the natural flow of events. They are of very short duration as well. The presence of a few rare objects or events in unimodal background events was regarded as anomaly in these datasets. Some examples of such datasets are Canoe [8] (a boat occurring once in the scene is regarded anomaly), UMN [9] (few people acting for sudden evacuation is regarded anomaly here), Web [9] (panic-escape and crowd fighting regarded as anomaly), Subway entrance/exit [10] (movement

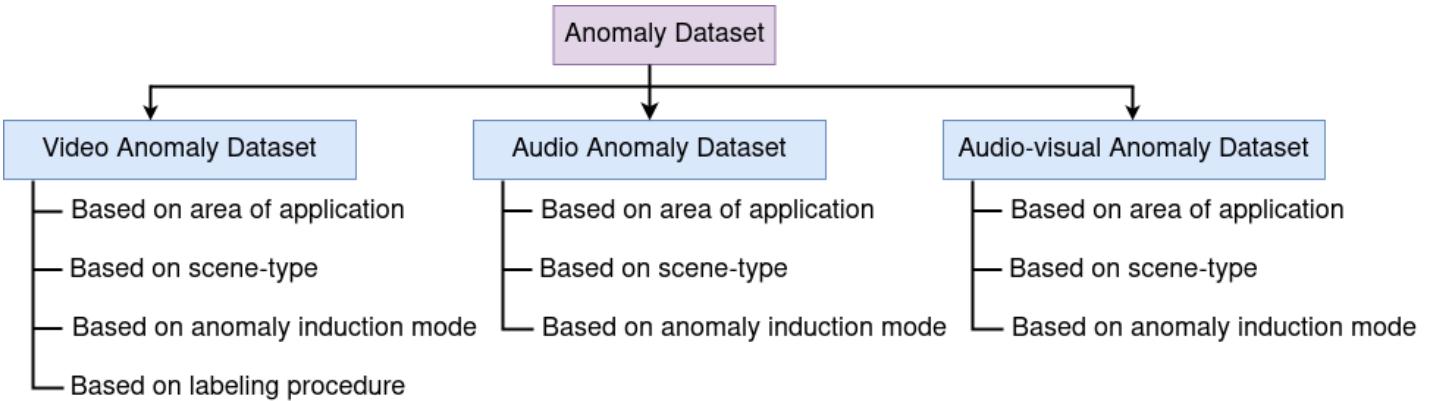


Fig. 1. Taxonomy for multimedia anomaly datasets

in wrong direction regarded as anomaly), various sub-clips in abnormal behavior dataset (appearance of only one type of rare object/event), etc.

Later datasets consist of more types of anomalies and scene variability such as UCSD [11], AVENUE [12], ARENA [13], ShanghaiTech [14]. In contrast, some recently developed datasets are gigantic in terms of duration and variability of scene and events such as UCF-Crime [15], VIRAT [16], LV [17], ADOC [18], HTA [19], Rodriguez's [20], etc. They leverage the online sources of CCTV footage and public videos and curate large databases of realistic video clips from our daily life. They are equipped with a variety of events and actions in with complex scene settings. Some datasets are suitable for generic scene monitoring, such as UCSD [11], ADOC [18], AVENUE [12], ShanghaiTech [14], etc., while others are suitable for specific classes of anomaly detection such as Subway entrance/exit [10], UMN [9], Web [9], i-Lids [21], etc. Further, few datasets such as UCF-Crime, LV, and NVIDIA AI CITY [22] contain specific anomaly classes that are important for usage in frameworks for detecting life-threatening events.

Table I presents a comparison of video anomaly datasets on various attributes of recorded videos such as continuity, duration, number of frames, number of video clips, resolution, frame rate, and camera motion. The datasets are listed in chronological order of their release year. Continuity (yes if untrimmed) information helps to see the applicability of a dataset for the task at hand. If the aim is to examine the context adaptation of anomaly detection framework, then the footage should be untrimmed as well as recorded at a single location [4], [6]. The adaptive models need sufficient data to learn the context on their own and adapt over time; hence a dataset with too short clips from different spatial and temporal aspects is not useful for evaluation. The approximate amount of duration is also specified in the table. Some of the dataset papers do not specify the exact duration and some provide just the number of frames; hence that entry is vacant in the table.

Generally, authors have used clips comprising of only normal frames in training and those with normal and abnormal frames in the testing phase. If the exact number of normal (N), abnormal (A), normal-abnormal (AN) clips are specified by the authors, then we have added this information in the

table. Otherwise, the total number of clips is specified there. Some datasets, mainly scraped from online sources (YouTube, Pond5, Getty Images), have multiple resolutions; therefore, they need to be scaled on the same scale before processing. Also, for the same reason, these datasets may have multiple FPS. Some datasets have camera motion present too. For some of the datasets, motion is intentionally added to make the dataset more challenging and realistic (surveillance in moving trains such as Train [23], etc., surveillance with head-mounted cameras such as HTA [19], etc.). In the following subsections, we present a systematic comparison and remark on publicly available video datasets based on (1) area of application, (2) scene type, (3) anomaly induction mode, and (4) labelling procedure. We describe and compare video datasets through Figures 2 to 5, following the categorization basis mentioned in Fig. 1. At the end of this section, we provide a more detailed overview of each dataset in Section II-E with Tables II to IV.

A. Area of application

Based on the area of application, the video datasets can be broadly categorized into traffic monitoring and public monitoring applications. Some datasets have a mix of utility for both traffic as well as public monitoring; we keep them in the miscellaneous application category. This categorization is shown in Fig. 2. The datasets built for traffic monitoring application have only traffic-related anomalies and hence owns exclusive application only to traffic surveillance. This category has seven datasets, and all of them possess mainly safety-critical anomalies, e.g., accident, tire skidding, wrong turn, etc. On the other hand, the public monitoring application has diversity in the dataset recording location and can be used across indoor to outdoor public place monitoring, e.g., mall, railway station, park, subway, street, etc. Most of the available video anomaly datasets fall into this category. Some of these datasets only possess safety-critical anomalies such as blast, fight, robbery, gunshot, etc., while others possess non-safety-critical anomalies such as walking with a cart in a pedestrian area, throwing papers in the air, walking on grass, etc. Further, some datasets have both types of anomalies. Thus, the public monitoring category is further categorized into ‘security-threat’, ‘non-security-threat’, and ‘both’ sub-categories, based on the criticalness of present anomalies.

TABLE I
SPECIFICATIONS OF VIDEO ANOMALY DATASETS

Dataset	Continuity	Total Duration	Total No. of Frames	Total No. of Videos	Resolution	FPS	Camera motion
i-Lids (2007)	no	≈24 min	35000	7	720x576	25	none
QMUL (2008)	no	22 min	34000	104(N)+8(A)	360x288	25	none
Canoe (2008)	yes	34 s	1050	1	320 × 240	30	none
Subway Entrance (2008)	yes	96 m 9 s	144225	1	512x384	25	none
Subway Exit (2008)	yes	43 m 16 s	64900	1	512x384	25	none
UMN (2009)	no	4 min 17 sec	7710	11 (available as 1) 12(N)+8(A)	320×240	30	none
Web (2009)	no	7 min 35 sec	11,962		multiple	multiple	slight jerks
PETS2009 (2009)	no	≈ 1-2 hrs	42182	59	768x576	7	none
U-Turn (2009)	no	≈20 min	≈25182	8	360 × 240	multiple	none
Idiap (2009)	yes	44.13 min	66324	1	288x360	25	none
USCD Ped1 (2010)	no	≈ 5-7 min	14000	34(N)+36(AN)	238x158	-	none
UCSD Ped2 (2010)	no	≈ 2-3 min	4560	16(N)+14(AN)	360×240	-	none
Train (2010)	yes	12 min	19218	1	288x386	25	moving
Belleview (2010)	no	4 min 51 sec	2918	1	320 × 240	10	jitter
Boat-Sea (2010)	yes	1 min 56 sec	450	1	720x576	19	none
Boat-River (2010)	yes	1 min 8 sec	250	1	704x576	5	none
Caouflage (2010)	yes	54 sec	1629	1	320×240	29.97	none
Rodriguez's	no	10 hrs 24 min	-	520	720x480	-	none
Hockey (2011)	no	27 min	50000	1000	720x576	-	none
Movie (2011)	no	6 min	-	200	multiple	multiple	in some
UCF Crowd (2012)	no	≈11 min	≈16320	38	multiple	multiple	none
Violent-Flows (2012)	no	14 min 6 sec	22,156	246	320×240	multiple	significant
Grand Central Station (2012)	yes	33 min 20 sec	50010	1	480x720	25	none
AGORASET (2012)	no	≈20 min	>33641	23	640x480	30	none
Meta-tracking (2013)	no	-	>4000	12	multiple	multiple	none
Avenue (2013)	no	20 min 26 sec	30652	16(N)+21(AN)	640x360	25	slight in few
ARENA (2014)	no	-	-	22	1280 x 960	30	none
PWPD (2015)	yes	1 hour	5000	1	1920×1080	1.25	none
RE-DID (2015)	no	<2 hrs	-	30	1280X720	multiple	in some
MED (2016)	no	≈24-25 min	43,626	31	554x235	31	none
ShanghaiTech (2017)	no	≈ 3-3.5 hrs	317398	330(N)+107(AN)	856x480	24	none
LV (2017)	no	3.93 hrs	-	30	multiple	multiple	in some
UCF-Crime (2018)	no	≈ 128 hrs	≈ 13 million	950(N)+950(A)	320×240	30	slight in few
IITH Accidents (2018)	no	-	128001	-	-	30	none
CCTV-Fights (2019)	no	≈ 10 hrs	-	1000	multiple	multiple	in some
Street Scene (2020)	no	≈ 3-4 h	203257	46(N)+35(AN)	1280×720	15	none
ADOC (2020)	yes	24 hrs	259127	4	2048 × 1536	3	none
HTA (2020)	no	≈ 4 hrs	≈ 0.4 million	286(N)+107(AN)	1280×720	30	moving

Some datasets can be used for both traffic as well as public monitoring as they possess anomalies for both categories. We list such datasets under the miscellaneous category. The datasets in each category are listed in order with the count of distinct safety-critical anomalies present in them. A dataset that offers a larger number of distinct safety-critical anomalies is mentioned first.

B. Scene-type

Some of the video anomaly datasets are recorded by placing the recording device at one place, while others collect video data from multiple places. Therefore, based on the dataset recording scene, there can be two categories of datasets: single-scene and multi-scene. They can be further categorized based on whether the scene is indoor or outdoor. A dataset recorded indoor or outdoor may help to evaluate the robustness or claim that the framework works in indoor and/or outdoor scenarios. We show the taxonomy for scene-based video datasets categorization via Fig 3. For datasets falling in the multi-scene category, we mention the exact number of locations in parenthesis along with the dataset if provided by

the authors of the datasets. Also, the datasets in each sub-category are listed in order of crowd density present in the dataset. Some frameworks are specifically designed to handle crowded scenes, whereas others work only for low density of people; thus, the information of density in a dataset may help to choose suitable datasets for evaluations.

In the multi-scene category, the datasets mainly possess non-contextual anomalies, which means if an event is anomalous in one scene, it will be anomalous in other scenes as well. For example, fighting with hands is an anomaly for a boxing game scene as well as in a park scene. Therefore, datasets falling in the multi-scene category largely ignore the context. On the other hand, in the case of single-scene datasets, events are regarded as anomalies based on their frequency. An event that is rare for the specific scene is regarded anomaly for that scene only. If the aim is to develop a generic scene monitoring, then single-scene datasets are useful, whereas, for detecting specific anomaly events, multi-scene datasets are suitable.

C. Anomaly induction mode

Based on the induction mode of anomalies, the datasets can be categorized into four groups, viz., acted, natural, acted

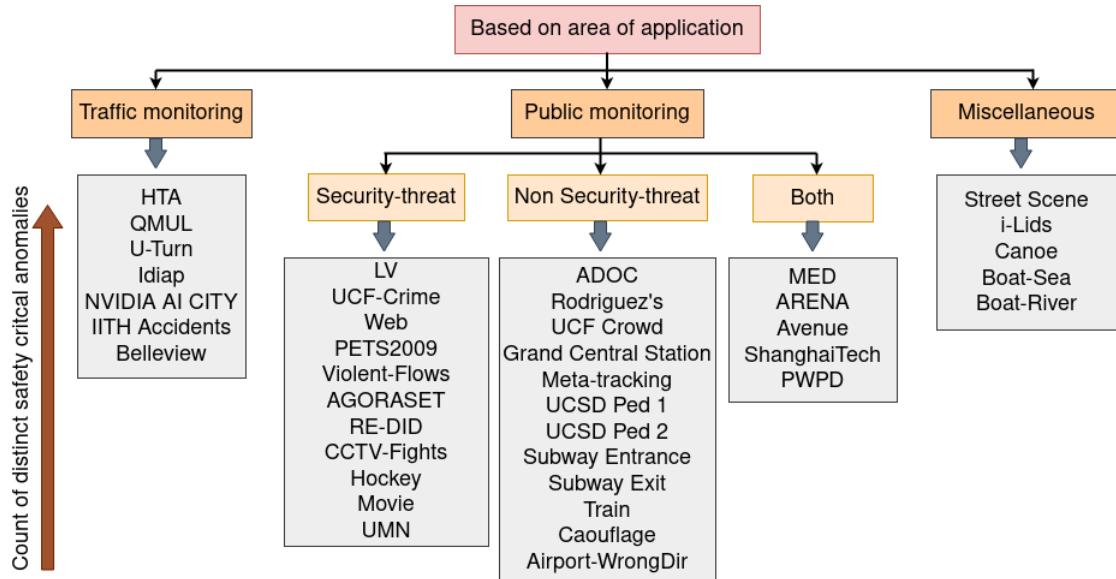


Fig. 2. Categorization of video anomaly datasets based on area of application

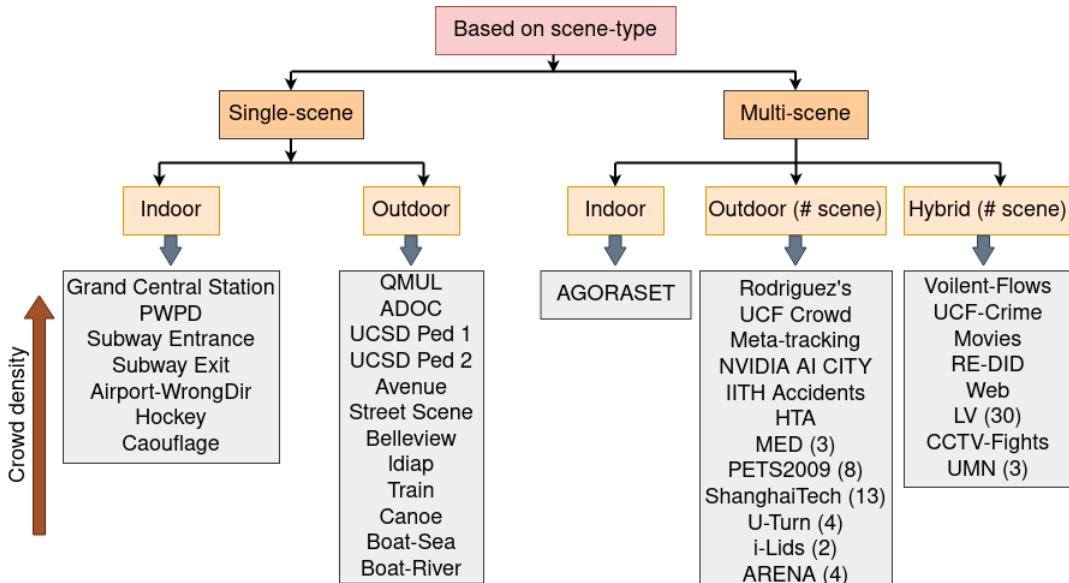


Fig. 3 Categorization of video anomaly datasets based on scene-type

as well as natural, and artificial, as shown in Fig. 4. Since anomaly is a rare phenomenon, it will require a longer waiting time to get the anomalous samples/ events; consequently, researchers have also tried inducing anomalies manually with the help of actors or artificially with the help of computer software. However, the acted events add unrealistic flavor, and hence less complex anomaly events are generated. In Fig. 4, we also mention the distinct number of anomalies present in each dataset. Further, if the dataset has acted anomalies too, the number of actors involved in inducing anomalies are also reported in parenthesis.

Generally, for traffic surveillance-related datasets such as HTA [19], QMUL [24], Idiap [25], IITH Accidents [26], etc., getting natural anomaly samples is easier as the anomalies in these datasets, which include accidents, illegal turns, or other

traffic rule violations, occur a bit frequently. Also, safety-critical events in public monitoring such as fights, blasts, robbery, gunshots are easier to collect from CCTV footage, movies, or YouTube. Hence, datasets such as UCF-Crime [15], LV [17], CCTV-Fights [27], Movie [28], etc., have natural anomalies. On the other hand, datasets having fewer safety-critical anomalies are built by inducing acted events too. Some of these include ARENA [13], Avenue [12], MED [29], ShanghaiTech [14], etc.

D. Labelling procedure

The existing video anomaly datasets have diversity in terms of the type of labeling too. We categorize them based on the available annotation type, as shown in Fig. 5. Some datasets give only clip-level annotation; if a clip is annotated anomalous

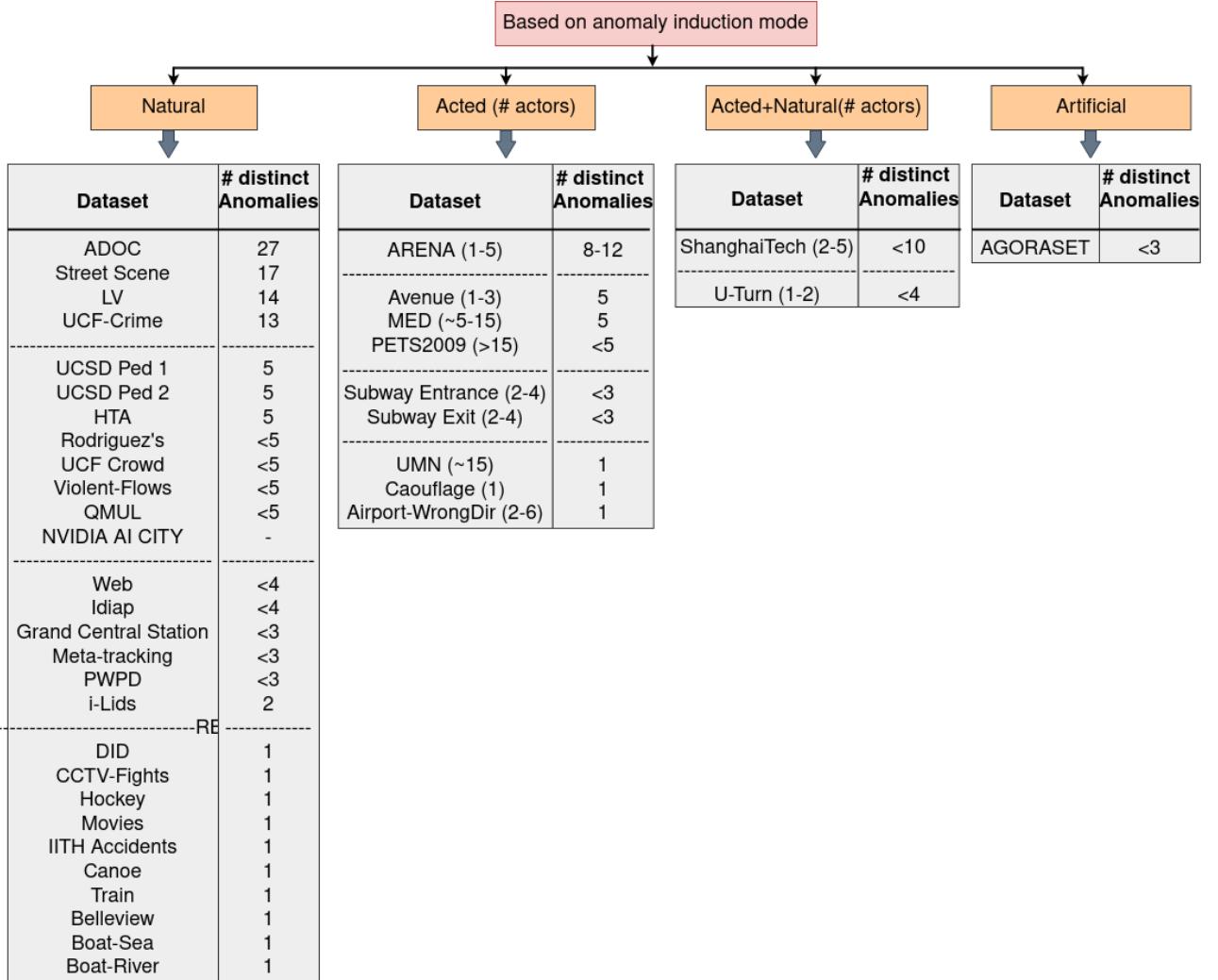


Fig. 4. Categorization of video anomaly datasets based on anomaly induction mode

ious, then it is assumed that anomaly is present or spanned to all the frames in the clip. The assumptions hold true in most cases, especially for small databases, but it is not scalable to validate this assumption for larger datasets. Hence, some frames may be wrongly annotated. Frame-level annotations are more precise and accurate [30]. For some datasets, pixel-level annotations are also provided, which are helpful for models which aim to precisely localize the occurrence of an anomaly in the spatial domain along with the temporal domain. Pixel-wise annotations are too precise and hence difficult to do manually; therefore, some datasets such as Train [23], Bellevue [23], Boat-Sea [23], BoatRiver [23], Airport-WrongDir [23], etc. first divide the frame into grids (a grid can have multiple pixels) and then annotate the grids that can be served for spatial anomaly localization. Some also provide the rectangle bounding boxes around the area of interest, i.e., anomalous object/person. Some datasets, especially for people tracking crowd datasets, e.g., PWPD [31], meta-tracking [32], etc., provide the trajectories and anomaly information on these.

E. Dataset Overview

A brief introduction of all the publicly available video anomaly datasets is given through Tables II to IV. The datasets are listed in chronological order of their release year. Some examples of anomalies present and the recording location of the dataset are also mentioned. Information about dataset collection scene, i.e., whether it is a mall or an airport or a traffic junction, etc., may help to choose scene-specific dataset for the evaluation of scene-specific surveillance frameworks. The table also mentions the area of application which gives an idea of different applications where the dataset has been used for bench-marking. Visuals from the dataset, i.e., a sample normal and an abnormal frame from the datasets, are also shown.

There are some other datasets that are not primarily developed for anomaly detection but other sub-tasks of scene monitoring such as detecting group formation, tracking people, counting people, human interaction/action classification, etc. We discuss them under the additional video dataset category in supplementary material attached with this survey, as they can also be utilized for generic scene surveillance or can be

TABLE II
VIDEO ANOMALY DATASETS: PRIMARY INFORMATION

Dataset	Anomalies: collection scenario	Applications	Normal image example	Anomaly image example
i-Lids [21] (2007)	abandon bag, illegally parked vehicle: at railway station, road	abandon baggage detection [33], traffic surveillance [34]		
QMUL [24] (2008)	unusual traffic trajectory, rare behaviour of vehicles: at road (traffic junction)	vehicle tracking [35], trajectory classification [36], anomaly detection [37]–[39]		
Canoe [8] (2008)	canoe: in river	anomaly detection [23], [40]		
Subway Entrance [10](2008)	wrong direction: at subway entrance	Anomaly detection [41], [42], abnormal behavior modeling [43], [44]		
Subway Exit [10](2008)	wrong direction: at subway exit	Anomaly detection [41], [42], abnormal behavior modeling [43], [44]		
UMN [9] (2009)	run: in field, courtyard, hallway	anomaly detection [11], [41], abnormal crowd behaviour [45], [46], crowd aggregation detection [47], crowd escape detection [48]		
Web [9] (2009)	panic-escape, clashing/fighting: at multiple location	anomaly detection [49]		
PETS2009 [50] (2009)	run, panic: in university	tracking [51], [52], crowd profiling/counting [53], [54], crowd analysis [46], human detection [55], [56], person re-identification [57], crowd escape behaviour detection [48]		
U-Turn [58] (2009)	illegal u-turns, running, abandon baggage: at road (intersection), university	traffic based anomaly detection [59]–[61]		
Idiap [25] (2009)	wrong road crossing, wrong vehicle parking, etc.,: at road	anomaly detection [38], [39], recurrent activity mining [62]		
USCD Ped1 [11] (2010)	bikers, small carts, walking across walkways: at campus (walkway at UCSD)	crowd profiling/counting [53], [63], anomaly detection [64], [65], crowd density estimation [66]		
USCD Ped2 [11] (2010)	bikers, small carts, walking across walkways: at campus (walkway at UCSD)	anomaly detection [67], [68], crowd profiling/counting [53], action classification [69]		

TABLE III
CONTINUED FROM TABLE II: VIDEO ANOMALY DATASETS: PRIMARY INFORMATION

Dataset	Anomalies: collection scenario	Applications	Normal image example	Anomaly image example
Train [23] (2010)	people movement: inside train	anomaly detection [40], [68], [70]		
Bellevue [23] (2010)	illegal turns: at road (intersection)	anomaly detection [40], [70]		
Boat-Sea [23] (2010)	boat: in sea	anomaly detection [40], [71]		
Boat-River [23] (2010)	boat: in river	anomaly detection [40], [44]		
Caouflage [23] (2010)	person in Caouflage: in room	anomaly detection [44]		
Airport-WrongDir [23] (2010)	movement in wrong direction: at security check-point	anomaly detection [44], [72]		
Rodriguez's [20] (2011)	anomalous trajectory: from multiple scenarios	crowd profiling/counting [54], crowd saliency detection [73], tracking [52], crowd analysis [74], crowd segmentation [75]		
Hockey [28] (2011)	fight: at ice hockey game	violence detection [76]–[78]	available upon request	
Movie [28] (2011)	fight: in movie clips	violence detection [77]–[80]	available upon request	
UCF Crowd [81] (2012)	anomalous trajectory: from multiple scenarios	abnormal crowd detection [82], crowd profiling/counting [54], crowd saliency detection [73], crowd segmentation [83]		
Violent-Flows [84] (2012)	crowd violence: at multiple scene	abnormal crowd behaviour detection [45], violence detection [77], [79]		
Grand Central Station [85] (2012)	rare walking pattern: at terminal station	pedestrian trajectory prediction [86], tracking [35], [87] pedestrian behavior modeling [88], crowd counting [63], [89] crowd behavior analysis [90], [91], human re-identification [92]		trajectory based
AGORASET [93] (2012)	evacuation, dispersion: at multiple scene	panic behaviour detection [94], crowd flow tracking [95], crowd motion classification [96], crowd behaviour analysis [97]		

TABLE IV
CONTINUED FROM TABLE III: VIDEO ANOMALY DATASETS: PRIMARY INFORMATION

Dataset	Anomalies: collection scenario	Applications	Normal image example	Anomaly image example
Meta-tracking [32] (2013)	anomalous trajectory: at multiple location	tracking [98], [99], anomalous walking pattern detection [98] measuring collectiveness [99], crowd segmentation [100]		trajectory based
Avenue [12] (2013)	loitering, running, throwing objects, new object: in campus (avenue)	anomaly detection [101], [102]		
ARENA [13] (2014)	abnormal behaviour, threats: in campus (university of Reading)	abnormal activity/behaviour detection [103]–[105], group walking event detection [106], Human Full-Body/ Body-Parts detection and tracking [107]		
PWPD [31] (2015)	anomalous trajectory, abnormal crowd: at terminal station	trajectory prediction [108], anomaly detection [108], pedestrian speed detection [109], crowd behaviour analysis [110]		
RE-DID [111] (2015)	fight: from multiple vehicle dash-cams	fight detection [111]		
MED [29] (2016)	Panic, fight, congestion, obstacle, neutral: at campus (walkway)	anomaly detection [112], panic detection [113]		
ShanghaiTech [14] (2017)	bicycle, small carts, fight, etc.: at campus (ShanghaiTech)	anomaly detection [101], [114], action classification [69], crowd counting [115]		
LV [17] (2017)	realistic security threats: at multiple location	anomaly detection [116]–[118], panic detection [119]		
UCF-Crime [15] (2018)	abuse, arrest, assault, accident, burglary, etc.: at multiple location	anomaly detection [120], [121], action classification [69]		
IITH Accidents [26](2018)	accidents: at road (intersection, junction)	road accident detection [26]		
CCTV-Fights [27](2019)	fight: at multiple location	fight detection [122]		
Street Scene [30](2020)	jaywalking across road, pedestrians loitering, u-turns: on road (two- lane street)	anomaly detection [123], road surveillance [124]		
ADOC [18] (2020)	walking with balloons/dog, person on vehicle, crowd gathering, etc.: at campus (walkway)	anomaly detection [123]		
HTA [19] (2020)	accident, speeding vehicle, close merge: at multiple location	anomalous motion detection [19]		

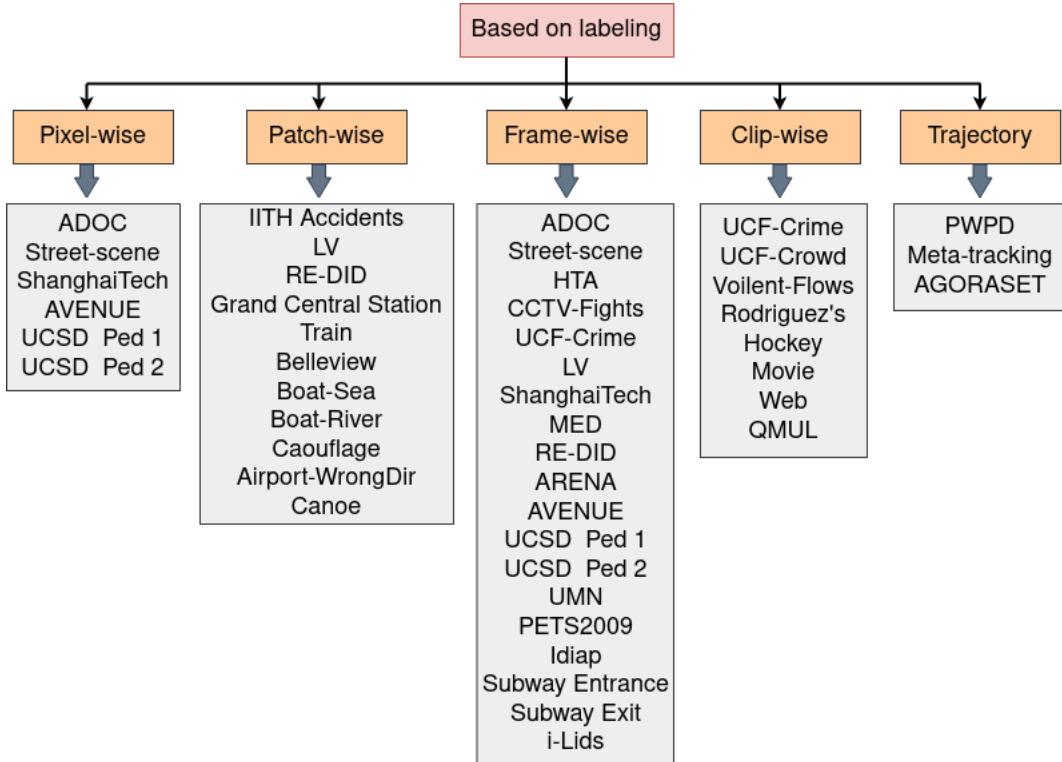


Fig. 5. Categorization of video anomaly datasets based on labeling procedure

useful in creating some giant anomaly datasets.

III. AUDIO ANOMALY DATASETS

Audio anomaly detection is gaining a lot of attention in many applications in various fields, such as traffic surveillance, industries, music, medicines, etc. Unlike for video, not many audio anomaly datasets are available for study. A few datasets have been created based on applications and further provided publicly. In this section, we describe and discuss the features of publicly available audio datasets. Table V gives the specifications of the audio samples collected for different datasets. The specifications include time duration of each dataset, number of samples collected, and sampling rate for audio-digitization. It can be observed from the table that the duration of audio for different datasets varies from a few seconds to a number of hours. AudioSet soundscape dataset has the longest duration of audios, followed by ToyADMOS2 [125] and ToyADMOS [126] datasets. A very small recording of 550 seconds has been collected for the LIFE DYNAMAP project [127].

Further, from the table, it can be seen that the sampling rate for different audio datasets varies between the range of 4kHz to 48 kHz. The sampling rate of 48 kHz has been mostly considered for audio digitization in various datasets. Also, for ICBHI 2017 [128], there is no fixed value of sampling rate for the data; rather, it varies between the range of 4 to 44.1 kHz. Thus, different sampling rates have been used as per the requirement of the tasks.

Dataset classification based on various attribute as mentioned in Fig. 1 is discussed in this section. Unlike in video

datasets, since only segment level labelling is followed for the audio datasets, hence, a separate classification of audio datasets based on labelling has not been provided.

A. Area of application

Audio anomaly datasets cover a variety of applications such as monitoring of traffic, machine, public, patient, etc. The area of application based classification is shown in Fig. 6. Majority of the audio anomaly datasets have been developed for machine monitoring, whereas a few datasets are available for monitoring traffic and public. Health supervision is also made possible using ICBHI [128] dataset by monitoring respiratory patterns in the patients.

B. Scene-type

Scene based categorization for audio datasets is shown via Fig. 7. Owing to different recording locations for each dataset, the number of multi-scene datasets is much higher and more easily available as compared to those collected from one location, such as ICBHI [128] and SMD [129]. The single-scene datasets have been collected in indoor scenarios, whereas multi-scene datasets have been created at various indoor and outdoor locations such as industry, road, forests, etc.

C. Anomaly induction mode

Fig. 8 shows anomaly induction mode based categorization of datasets. The count of distinct anomalies present in the dataset is also listed along with. As it can be seen from the

TABLE V
SPECIFICATIONS OF AUDIO ANOMALY DATASETS

Dataset	Continuity	Total duration	No. of samples	Sampling rate
MIVIA (2014)	no	3580 seconds	-	32 kHz
AudioSet Soundscape: Ithaca, New York	no	#1: 797 hrs #2: 638 hrs	-	48 kHz 48 kHz
AudioSet Soundscape: Sabah, Malaysia	no	Tuscam: 27 hrs 40 min Audiomoth: 784 hrs	-	44.1 kHz 16 kHz
AudioSet Soundscape: New Zealand (2015)	no	240 hrs	-	32 kHz
AudioSet Soundscape: Sulawesi	no	64 hrs	-	48 kHz
AudioSet Soundscape: Republic of Congo	no	238 hrs 20 min	-	8 kHz
DCASE 2017 (2017)	no	39 min	1170	44.1 kHz
ICBHI 2017 (2017)	no	> 5.5 hrs	-	4-44.1 kHz
SMD (2018)	no	≈ 1 hour	2048	16.3 kHz
MIMII DUE (2019)	no	>420000 sec	32157	16 kHz
ToyADMOS (2019)	no	≈ 540 hrs	> 12000	48 kHz
LIFE DYNAMAP (2020)	no	550 sec	-	48 kHz
ToyADMOS2 (2021)	no	604 hrs	≈264k	48 kHz
DCASE 2021 (2021)	no	≈82 hours	29463	16 kHz

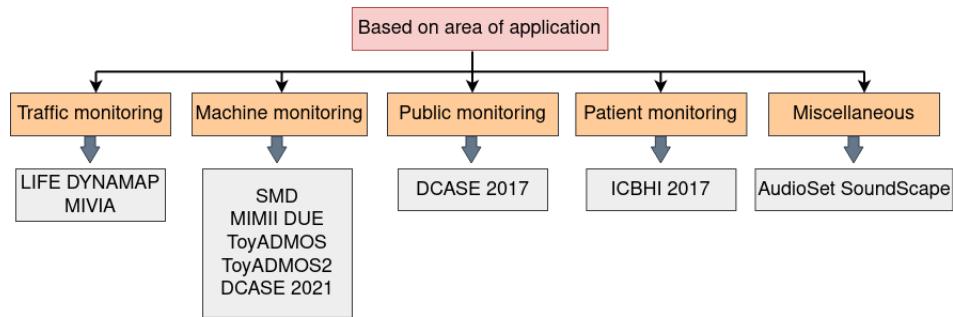


Fig. 6. Categorization of audio anomaly datasets based on area of application

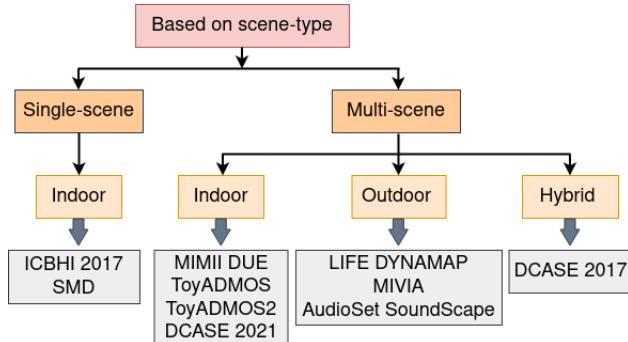


Fig. 7. Categorization of audio anomaly datasets based on scene-type

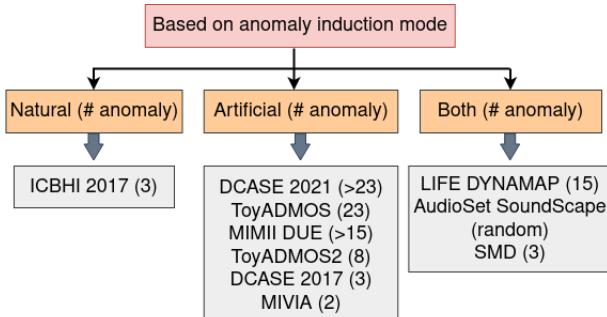


Fig. 8. Categorization of audio anomaly datasets based on anomaly induction mode

figure, most of the audio datasets have been created artificially, except for ICBHI 2017 [128] dataset, which records the natural respiratory sounds from a human body. Since it is difficult to generate natural audio datasets consisting of anomalies, hence anomalies have been generated deliberately to simulate the system for its detection and for future anomaly detection applications.

D. Dataset Overview

Table VI list all the publicly available audio anomaly datasets in chronological order of their release year. Some of the anomalies that have been generated are mentioned for every dataset in the table. The datasets have been used for applications in different fields, such as in industries for machine monitoring and inspection, road surveillance, etc. A dataset for anomaly detection in respiratory patterns in human bodies has also been created. Most of the datasets have been artificially generated by deliberately adding anomalies and made available to the public recently.

IV. AUDIO-VISUAL DATASETS

So far, we have discussed a number of publicly available datasets for video and audio anomaly detection applications individually. Since a course of actions consists of both audio

TABLE VI
AUDIO ANOMALY DATASETS: PRIMARY INFORMATION

Dataset	Anomalies: collection scenario	Application
MIVIA [130] (2014)	crashes and tire skidding: on road	road surveillance [131]–[133]
AudioSet Soundscape [134] (2015)	music, human speech, machine noise, etc.: from nature	anomaly detection from eco-acoustic data [134]
DCASE 2017 [135] (2017)	baby cry, glass break, gun shot: city	detection of rare sounds [136], [137]
ICBHI 2017 [128] (2017)	crackle, wheeze, crackle and wheeze: in human body	anomalous respiratory pattern detection in humans [138]–[140]
SMD [129] (2018)	non-greased line, B-line, C-line to B-line: from industry	machine monitoring in industries [129]
MIMII DUE [141] (2019)	wing damage, clogging, gear, contamination etc.: from industry	investigation and inspection of industrial machine
ToyADMOS [126] (2019)	deformed gears, over/under voltage, pulley, chipped wheel, axle, excessive tension, etc.: from industry	anomalous sound detection in miniature machines [142], [143]
LIFE DYNAMAP [127] (2020)	horns, church bells, birds, thunder etc.: on road	detection and removal of anomalous events for road traffic mapping [144], [145]
ToyADMOS2 [125] (2021)	bent shaft, melted gears, flat tire etc.: from industry	anomalous sound detection in miniature machines
DCASE 2021 [146] (2021)	wing damage, clogging, chipped wheel, axle etc.: from industry	detection of anomalous sounds during machine monitoring [147]–[149]

TABLE VII
SPECIFICATIONS OF AUDIO-VISUAL ANOMALY DATASETS

Dataset	Continuity	Total Duration	Total No. of Frames	Total No. of Videos	Video resolution	FPS (video)	Camera motion
Human-human interaction (2014)	no	32.24 min	-	8	-	-	none
VSD (2015)	no	35 hrs 18 min	-	25	multiple	multiple	in some
EMOLY (2018)	no	-	-	123	-	-	none
XD-Violence (2020)	no	217 hrs	-	4754	multiple	multiple	in some
BAREM (2021)	no	≈ 6 hrs	≈ 5,40,000	72	-	25	none

and video components dependent on each other, hence, audio-visual analysis can result in more accurate results compared to audio and video recordings being used individually [4], [150]. There are very few audio-visual datasets for anomaly detection. They are mainly developed for applications such as detection of anomalous expression, violence, stress, etc.

Table VII provides details about the features and specifications of the audio-visual clips. The table enables a quick comparison of datasets based on their total frame count, resolution, duration, etc. Mainly, information about the total number of videos and total duration is made available by the authors. The datasets for violence detection records the largest length of 217 hours compared to others. This is because violence event is comparatively more often to occur and hence easily available in CCTV footage. The total number of videos available for XD-Violence [151] is maximum, i.e., 4754, followed by the EMOLY [152] dataset consisting of 123 videos. Least number of videos have been provided in the Human-human interaction dataset, with a total duration of 32.24 minutes. Clips have been recorded at multiple resolutions and at multiple frame rates. However, the sampling rate of audios recorded in the

databases has not been made available by the authors. Few clips from VSD [153] as well as XD-Violence [151] datasets consist of camera motion, thus making them challenging for further analysis.

Also, it can be observed from the table that none of the datasets is collected in continuation and hence are not suitable for evaluation of adaptive anomaly detection frameworks. On the same categorization basis as for audio datasets, we present and discuss the categorization of audio-visual datasets through Fig. 9 to Fig. 11. Unlike in video datasets, since only frame level labelling is followed for the audio-visual datasets, hence, a separate class of audio-visual datasets based on labelling has not been provided. Since all the datasets possess frame-level annotations, hence, anomaly localization is not yet explored in audio-visual datasets.

A. Area of application

The audio-visual datasets have been designed and made available to the users recently. They all fall into public monitoring applications. The categorization is shown in Fig. 9. Different sub-applications of public monitoring, such as anom-

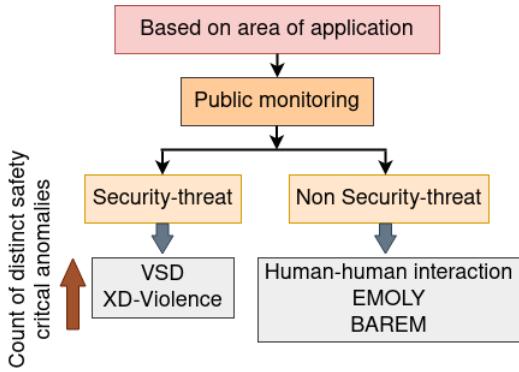


Fig. 9. Categorization of audio-visual anomaly datasets based on area of application

lous expression detection, violence detection, detection of stress in a particular situation, etc., have been covered for these datasets. Based on whether the anomalies posses security threats or not, they are further categorized into two groups viz., security-threat and non security-threat. VSD [153] and XD-Violence [151] datasets have security threat-based anomalies, and the rest datasets do not possess any security threat-based anomalies.

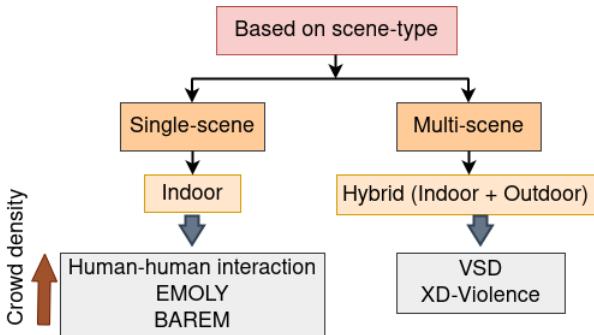


Fig. 10. Categorization of audio-visual anomaly datasets based on scene-type

B. Scene-type

Scene-type-based dataset categorization is shown via Fig. 10. The audio-visual dataset collection scenarios are mostly limited to indoor. Human-human interaction [154], EMOLY [152], and BAREM [155] are single-scene datasets, all collected in indoor environments. Rest two datasets, viz., VSD [153] and XD-Violence [151], are multiscene datasets comprising of both indoor as well as the outdoor scenes. The datasets for violence detection, i.e., VSD [153] and XD-Violence [151], are mainly collected from real CCTV footage and hence have high crowd density; whereas the other datasets mainly contain actors and thus have low density. Among the three single-scene datasets, BAREM [155] has the least crowd density (single person at a time), followed by EMOLY [152] and then Human-human interaction [154] dataset.

C. Anomaly induction mode

Categorization based on anomaly induction mode, i.e., natural or acted, is shown via Fig. 11. All the datasets contain some

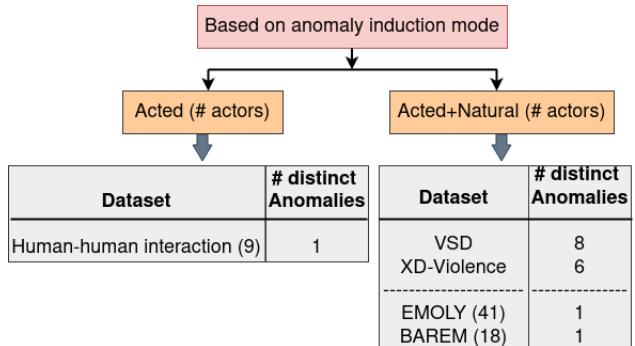


Fig. 11. Categorization of audio-visual anomaly datasets based on anomaly induction mode

acted anomalies too. The number of actors in the datasets is also mentioned in parenthesis. Each dataset has only one type of distinct anomaly except for the violence detection datasets, i.e., VSD [153] and XD-Violence [151]. VSD [153] consists of 8 distinct anomalies, while XD-Violence [151] contains 6 distinct anomalies. Some of the example anomalies in audio-visual datasets include fights, stress, explosions, abuse, car accident, shooting, frustration, etc.

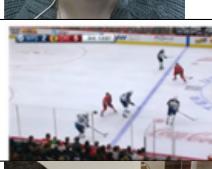
D. Dataset Overview

For a quick analysis of all the publicly available audio-visual anomaly datasets, we list them in increasing order of their release year via Table VIII. The table has other information, viz., where it was collected, what anomalies are in them, applications where it has been used for benchmarking, a normal sample image, and an abnormal sample image. These datasets mainly focus on the specific type of anomaly. VSD [153] dataset, XD-Violence [151] dataset, etc., consist of only fight events, whereas EMOLY [152] dataset has an anomalous mood of a person as an anomaly. Human-human interaction dataset has agitation behavior as an anomaly. Apart from having one class of anomaly, the datasets are mainly recorded in a controlled environment (except for fight anomaly, which is easy to collect) by some human actors. The existing audio-visual dataset is good for application-specific anomaly detection; however, there is a lack of audio-visual datasets for generic scene surveillance. Some researchers have tried collecting such datasets, but they are not released publicly due to privacy or legal issues.

V. DISCUSSION: TOWARDS THE FUTURE

During the last decade, there has been a drastic shift from datasets with less number of anomaly samples and total duration to those with diverse range of anomalies and gigantic volume. It may be observed that the availability of datasets with crime-specific anomalies has facilitated the development of automated surveillance frameworks for crime detection. For videos, the category spans detection of the explosion, abuse, panic escape, assault, accident, violence, fight, etc. In the case of audio, they span as detection of the gunshot, shout, etc. Further, for the audio-visual dataset, the categories span as detection of riots, fight, and violence.

TABLE VIII
AUDIO-VISUAL ANOMALY DATASETS: PRIMARY INFORMATION

Dataset	Anomalies: collection scenario	Application	Anomaly image example	Normal image example
Human-human interaction [154] (2014)	stress: at help-desk	Detection of stressful situations at a help-desk [156]		
VSD [153] (2015)	fights, fire, gunshot, cold weapons, car chases, gory, explosion, screams: at multiple location	violence detection [80], [157], [158]		
EMOLY [152] (2018)	anomalous expression: in lab	abnormal expression detection [152]		
XD-Violence [151] (2020)	abuse, car accident explosion, fight, riot, shoot: at multiple location	violence detection [159], anomaly detection [160]		
BAREM [155] (2021)	frustration: on e-service platform	Behaviour Analysis for Reverse Efficient Modeling [155]		Available upon request

By analyzing the applications and anomalies present in the datasets discussed across Sections II to IV, we can see that they have mainly specific types of anomalies. There are a very few datasets of heterogeneous nature. This is almost zero in the audio-visual category. Thus, existing datasets are useful for specific anomaly detection, particularly crime-oriented anomaly detection, traffic rule violation, etc. However, for future applications like smart city surveillance we need generic scene monitoring where we have to raise an alarm for interesting events, which may or may not be crime-oriented. This shows a strong need to develop datasets having diverse ranges of anomalous samples. Also, the datasets in the audio-visual category mainly contain acted anomalies, which limits its usefulness. Even if some acted situation/events need to be added, it should be in such a manner that it appears natural and contains the necessary amount of variations mimicking real life.

Apart from this, all the existing multimedia datasets lack anomalies with concept drift [6]. If an event/ object is regarded anomaly in a dataset, it is always regarded as an anomaly for that scene, no matter how frequently it may occur in the distant future. This is due to the fact that datasets are too short to contain this effect. We should pay attention that the datasets which are more than a duration of 5-10 hours are not suitable here because the samples are collected from different time-stamp, location, and have only specific anomalies in rare amounts. Thus, interclass shift, i.e., abnormal to normal

class and vice-versa, is not observed. There are attempts to record long untrimmed footage at one place, e.g., QMUL [24], ADOC [18], etc. However, the authors do not attempt to provide annotations in accordance with concept drift.

VI. CONCLUSION

This paper presents a survey of multimedia datasets for anomaly detection to researchers working towards automated surveillance. The structured comparison of datasets on various attributes also helps to understand datasets better. There are a large number of short-length and giant video datasets available. Some are developed for generic scene surveillance, whereas other are specific anomaly datasets. Datasets for heterogeneous anomaly are far less compared to specific anomaly datasets. In case of audio, datasets are mainly developed for machine surveillance such as defect detection, fault detection, etc. Generally, surveillance using audio alone in outdoor scenarios is not efficient due to the presence of multiple auditory signals superimposed together. However, when analyzed together with video, they can offer crucial and complementary information about the target scene. However, the datasets developed towards audio-visual surveillance are far less compared to that for audio or video. There are a few audio-visual datasets for a specific action, such as fight and agitation detection. A recently released dataset, viz., the EMOLY dataset, has used only the upper body of individuals and their speech information to facilitate abnormal behavior detection. However, there is a

strong need to develop more audio-visual datasets for generic scene surveillance. We believe the survey presented in this article will help the prospective researchers who intend to contribute datasets or research in this field.

REFERENCES

- [1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys*, vol. 41, no. 3, pp. 1–58, 2009.
- [2] N. Jaafar and Z. Lachiri, "Audio-visual fusion for aggression detection using deep neural networks," in *2019 International Conference on Control, Automation and Diagnosis (ICCAD)*. IEEE, 2019, pp. 1–5.
- [3] A.-U. Rehman, H. S. Ullah, H. Farooq, M. S. Khan, T. Mahmood, and H. O. A. Khan, "Multi-modal anomaly detection by using audio and visual cues," *IEEE Access*, vol. 9, pp. 30 587–30 603, 2021.
- [4] P. Kumari, "Situational anomaly detection in multimedia data under concept drift," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2969–2973.
- [5] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," *Journal of Visual Communication and Image Representation*, vol. 77, p. 103116, 2021.
- [6] P. Kumari and M. Saini, "Multivariate adaptive gaussian mixture for scene level anomaly modeling," in *BigMM*. New Delhi, India: IEEE, 2020, pp. 54–62.
- [7] N. Patil and P. K. Biswas, "A survey of video datasets for anomaly detection in automated surveillance," in *ISED*. IEEE, 2016, pp. 43–48.
- [8] P.-M. Jodoin, J. Konrad, and V. Saligrama, "Modeling background activity for behavior subtraction," in *ICDSC*. IEEE, 2008, pp. 1–10.
- [9] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *CVPR*. IEEE, 2009.
- [10] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *Transactions on pattern analysis and machine intelligence*, vol. 30, 2008.
- [11] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*. IEEE, 2010, pp. 1975–1981.
- [12] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *ICCV*. IEEE, 2013.
- [13] ARENA, "Dataset, pets," 2014. [Online]. Available: <http://www.cvg.reading.ac.uk/PETS2014/a.html>
- [14] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *ICCV*, 2017, pp. 341–349.
- [15] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *CVPR*. IEEE, 2018.
- [16] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR*. IEEE, 2011, pp. 3153–3160.
- [17] R. Leyva, V. Sanchez, and C.-T. Li, "The lv dataset: A realistic surveillance video dataset for abnormal event detection," in *IWBF*. IEEE, 2017, pp. 1–6.
- [18] M. Pranav, L. Zhenggang *et al.*, "A day on campus-an anomaly detection dataset for events in a single camera," in *ACCV*, 2020.
- [19] H. Singh, E. M. Hand, and K. Alexis, "Anomalous motion detection on highway using deep learning," in *ICIP*. IEEE, 2020, pp. 1901–1905.
- [20] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *ICCV*. IEEE, 2011, pp. 1235–1242.
- [21] i Lids, "dataset for avss," 2007. [Online]. Available: <http://www.eecs.qmul.ac.uk/andrea/avss2007.d.html>
- [22] NVIDIA, "Ai city," 2021. [Online]. Available: <https://www.aicitychallenge.org/>
- [23] A. Zaharescu and R. Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *ECCV*. Springer, 2010, pp. 563–576.
- [24] C. C. Loy, T. Xiang, and S. Gong, "From local temporal correlation to global anomaly detection," in *MLVMA*, 2008.
- [25] J. Varadarajan and J.-M. Odobez, "Topic models for scene analysis and abnormality detection," in *ICCV*. IEEE, 2009, pp. 1338–1345.
- [26] D. Singh and C. K. Mohan, "Deep spatio-temporal representation for detection of road accidents using stacked autoencoder," *Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 879–887, 2018.
- [27] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveillance videos," in *ICASSP*. IEEE, 2019, pp. 2662–2666.
- [28] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *CAIP*. Springer, 2011, pp. 332–339.
- [29] H. Rabiee, J. Haddadnia, H. Mousavi, M. Kalantarzadeh, M. Nabi, and V. Murino, "Novel dataset for fine-grained abnormal behavior understanding in crowd," in *AVSS*. Colorado Springs, CO, USA: IEEE, 2016.
- [30] B. Ramachandra and M. Jones, "Street scene: A new dataset and evaluation protocol for video anomaly detection," in *WACA*, 2020, pp. 2569–2578.
- [31] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *CVPR*, 2015, pp. 3488–3496.
- [32] P.-M. Jodoin, Y. Benetech, and Y. Wang, "Meta-tracking for video scene understanding," in *AVSS*. IEEE, 2013, pp. 1–6.
- [33] Q. Li, Y. Mao, Z. Wang, and W. Xiang, "Robust real-time detection of abandoned and removed objects," in *ICIG*. IEEE, 2009, pp. 156–161.
- [34] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *CVPR*. IEEE, 2011, pp. 3161–3167.
- [35] K. K. Santhosh, D. P. Dogra, P. P. Roy, and B. B. Chaudhuri, "Trajectory-based scene understanding using dirichlet process mixture model," *Transactions on cybernetics*, 2019.
- [36] K. K. Santhosh, D. P. Dogra, P. P. Roy, and A. Mitra, "Vehicular trajectory classification and traffic anomaly detection in videos using a hybrid cnn-vae architecture," *Transactions on Intelligent Transportation Systems*, 2021.
- [37] C. C. Loy, T. Xiang, and S. Gong, "Detecting and discriminating behavioural anomalies," *Pattern Recognition*, vol. 44, no. 1, pp. 117–132, 2011.
- [38] J. Varadarajan, R. Subramanian, N. Ahuja, P. Moulin, and J.-M. Odobez, "Active online anomaly detection using dirichlet process mixture model and gaussian process classification," in *WACV*. IEEE, 2017, pp. 615–623.
- [39] V. Kaltsa, A. Briassoulis, I. Kompatsiaris, and M. G. Strintzis, "Multiple hierarchical dirichlet processes for anomaly detection in traffic," *Computer Vision and Image Understanding*, vol. 169, pp. 28–39, 2018.
- [40] F. P. dos Santos, L. S. Ribeiro, and M. A. Ponti, "Generalization of feature embeddings transferred from different video anomaly detection domains," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 407–416, 2019.
- [41] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR*. IEEE, 2011, pp. 3449–3456.
- [42] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *CVPR*. IEEE, 2009, pp. 1446–1453.
- [43] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—a review," *Transactions on Systems, Man, and Cybernetics*, vol. 42, no. 6, pp. 865–878, 2012.
- [44] M. Javan Roshtkhari and M. D. Levine, "Online dominant and anomalous behavior detection in videos," in *CVPR*, 2013, pp. 2611–2618.
- [45] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino, "Analyzing tracklets for the detection of abnormal crowd behavior," in *WACA*. IEEE, 2015, pp. 148–155.
- [46] H. Su, H. Yang, S. Zheng, Y. Fan, and S. Wei, "The large-scale crowd behavior perception based on spatio-temporal viscous fluid field," *Transactions on Information Forensics and security*, vol. 8, no. 10, pp. 1575–1589, 2013.
- [47] M. Xu, C. Li, P. Lv, N. Lin, R. Hou, and B. Zhou, "An efficient method of crowd aggregation computation in public areas," *Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2814–2825, 2017.
- [48] S. Wu, H.-S. Wong, and Z. Yu, "A bayesian model for crowd escape behavior detection," *Transactions on circuits and systems for video technology*, vol. 24, no. 1, pp. 85–98, 2013.
- [49] X. Zhu, J. Liu, J. Wang, C. Li, and H. Lu, "Sparse representation for robust abnormality detection in crowded scenes," *Pattern Recognition*, vol. 47, no. 5, pp. 1791–1799, 2014.
- [50] J. Ferryman and A. Shahrokn, "An overview of the pets 2009 challenge," 2009.
- [51] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker," in *ICCV*. IEEE, 2011, pp. 120–127.
- [52] C.-T. Chu, J.-N. Hwang, H.-I. Pai, and K.-M. Lan, "Tracking human under occlusion based on adaptive multiple kernels with projected gradients," *Transactions on Multimedia*, vol. 15, no. 7, pp. 1602–1615, 2013.

- [53] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd counting and profiling: Methodology and evaluation," in *Modeling, simulation and visual analysis of crowds*, 2013, pp. 347–382.
- [54] H. Fradi and J.-L. Dugelay, "Towards crowd density-aware video surveillance applications," *Information Fusion*, vol. 24, pp. 3–15, 2015.
- [55] D. S. Bolme, Y. M. Lui, B. A. Draper, and J. R. Beveridge, "Simple real-time human detection using a single correlation filter," in *PETS*. IEEE, 2009, pp. 1–8.
- [56] C. Conde, D. Moctezuma, I. M. De Diego, and E. Cabello, "Hogg: Gabor and hog-based human detection for surveillance in non-controlled environments," *Neurocomputing*, vol. 100, pp. 19–30, 2013.
- [57] J. Yang, Z. Shi, and P. A. Vela, "Person reidentification by kernel pca based appearance learning," in *CRV*. IEEE, 2011, pp. 227–233.
- [58] Y. Beneteth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurrences," in *CVPR*. IEEE, 2009, pp. 2458–2465.
- [59] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Processing: Image Communication*, vol. 47, pp. 358–368, 2016.
- [60] Y. Beneteth, P.-M. Jodoin, and V. Saligrama, "Abnormality detection using low-level co-occurring events," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 423–431, 2011.
- [61] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *Transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.
- [62] J. Varadarajan, R. Emonet, and J.-M. Odobez, "A sequential topic model for mining recurrent activities from long term video logs," *International journal of computer vision*, vol. 103, no. 1, pp. 100–126, 2013.
- [63] T. Xu, X. Chen, G. Wei, and W. Wang, "Crowd counting using accumulated hog," in *ICNC-FSKD*. IEEE, 2016, pp. 1877–1881.
- [64] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection," in *WACV*. IEEE, 2018, pp. 1689–1698.
- [65] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognition*, vol. 46, no. 7, pp. 1851–1864, 2013.
- [66] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in *ICCV*. IEEE, 2015, pp. 3253–3261.
- [67] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, vol. 172, pp. 88–97, 2018.
- [68] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, 2017.
- [69] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *CVPR*, 2019, pp. 1237–1246.
- [70] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *Signal Processing Letters*, vol. 22, no. 9, pp. 1477–1481, 2015.
- [71] M. J. Roshtkhari and M. D. Levine, "An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions," *Computer vision and image understanding*, vol. 117, no. 10, pp. 1436–1452, 2013.
- [72] K. V. Joshi and N. M. Patel, "A cnn based approach for crowd anomaly detection," *International Journal of Next-Generation Computing*, vol. 12, no. 1, 2021.
- [73] M. K. Lim, V. J. Kok, C. C. Loy, and C. S. Chan, "Crowd saliency detection via global similarity structure," in *ICPR*. IEEE, 2014, pp. 3957–3962.
- [74] A. Bera, S. Kim, and D. Manocha, "Realtime anomaly detection using trajectory-level crowd behavior learning," in *CVPR*, 2016, pp. 50–57.
- [75] V. J. Kok and C. S. Chan, "Gres: Granular computing-based crowd segmentation," *Transactions on cybernetics*, vol. 47, no. 5, pp. 1157–1168, 2016.
- [76] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image and vision computing*, vol. 48, pp. 37–41, 2016.
- [77] X. Xu, X. Wu, G. Wang, and H. Wang, "Violent video classification based on spatial-temporal cues using deep learning," in *ISCID*, vol. 1. IEEE, 2018, pp. 319–322.
- [78] M. Cheng, K. Cai, and M. Li, "Rwf-2000: An open large scale video database for violence detection," in *ICPR*. IEEE, 2021, pp. 4183–4190.
- [79] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3d convolutional neural networks," *IEEE Access*, vol. 7, pp. 39172–39179, 2019.
- [80] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: A novel deep-learning-based approach towards violence-detection in movies," *Applied Sciences*, vol. 9, no. 22, p. 4963, 2019.
- [81] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *Transactions on pattern analysis and machine intelligence*, vol. 34, no. 10, pp. 2064–2070, 2012.
- [82] Y. Zhang, L. Qin, R. Ji, H. Yao, and Q. Huang, "Social attribute-aware force model: exploiting richness of interaction for abnormal crowd detection," *Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 7, pp. 1231–1245, 2014.
- [83] J. Li, H. Yang, and S. Wu, "Crowd semantic segmentation based on spatial-temporal dynamics," in *AVSS*. IEEE, 2016, pp. 102–108.
- [84] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *CVPR*. IEEE, 2012, pp. 1–6.
- [85] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *CVPR*. IEEE, 2012, pp. 2871–2878.
- [86] Y. Xu, Z. Piao, and S. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in *CVPR*. Salt Lake City, UT, USA: IEEE, 2018, pp. 5275–5284.
- [87] A. Maksai, X. Wang, F. Fleuret, and P. Fua, "Non-markovian globally consistent multi-object tracking," in *ICCV*. Venice, Italy: IEEE, 2017, pp. 2544–2554.
- [88] S. Yi, H. Li, and X. Wang, "Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance," *Transactions on image processing*, vol. 25, no. 9, pp. 4354–4368, 2016.
- [89] M. A. K. Sağın and B. Bolat, "A novel approach for people counting and tracking from crowd video," in *INISTA*. IEEE, 2017, pp. 277–281.
- [90] J. Zhong, W. Cai, L. Luo, and H. Yin, "Learning behavior patterns from video: A data-driven framework for agent-based crowd modeling," in *AAMAS*, 2015, pp. 801–809.
- [91] B. Zhou, X. Tang, and X. Wang, "Learning collective crowd behaviors with dynamic pedestrian-agents," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 50–68, 2015.
- [92] S. M. Assari, H. Idrees, and M. Shah, "Human re-identification in crowd videos using personal, social and environmental constraints," in *ECCV*. Springer, 2016, pp. 119–136.
- [93] P. Allain, N. Courty, and T. Corpetti, "Agoraset: a dataset for crowd video analysis," in *ICPR*. IAPR, 2012, pp. 1–6.
- [94] D. Shehab and H. Ammar, "Statistical detection of a panic behavior in crowded scenes," *Machine Vision and Applications*, vol. 30, no. 5, pp. 919–931, 2019.
- [95] A. Fagette, P. Jamet, D. Racoceanu, and J.-Y. Dufour, "Particle video for crowd flow tracking," 2013.
- [96] A. Basset, P. Bouthemy, and C. Kerivenn, "Frame-by-frame crowd motion classification from affine motion models," in *AVSS*. IEEE, 2013, pp. 282–287.
- [97] A. Pennisi, D. D. Bloisi, and L. Iocchi, "Online real-time crowd behavior detection in video sequences," *Computer Vision and Image Understanding*, vol. 144, pp. 166–176, 2016.
- [98] S. Kim, A. Bera, and D. Manocha, "Interactive crowd content generation and analysis using trajectory-level behavior learning," in *ISM*. IEEE, 2015, pp. 21–26.
- [99] X. Li, M. Chen, and Q. Wang, "Measuring collectiveness via refined topological similarity," *Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 2, pp. 1–22, 2016.
- [100] Z. Fan, J. Jiang, S. Weng, Z. He, and Z. Liu, "Adaptive crowd segmentation based on coherent motion detection," *Journal of Signal Processing Systems*, vol. 90, no. 12, pp. 1651–1666, 2018.
- [101] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *ICCV*, 2019, pp. 1705–1714.
- [102] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *CVPR*, 2016, pp. 733–742.
- [103] L. Patino and J. Ferryman, "Detecting threat behaviours," in *AVSS*. IEEE, 2016, pp. 88–94.

- [104] G. J. Burghouts, P. van Slagerland, R. ten Hove, R. J. den Hollander, and K. Schutte, "Complex threat detection: Learning vs. rules, using a hierarchy of features," in *AVSS*. IEEE, 2014, pp. 375–380.
- [105] L. Patino and J. Ferryman, "Multiresolution semantic activity characterisation and abnormality discovery in videos," *Applied Soft Computing*, vol. 25, pp. 485–495, 2014.
- [106] V. Bastani, D. Campo, L. Marcenaro, and C. Regazzoni, "Online pedestrian group walking event detection using spectral analysis of motion similarity graph," in *AVSS*. IEEE, 2015, pp. 1–5.
- [107] H.-W. Chen and M. McGurr, "Improved color and intensity patch segmentation for human full-body and body-parts detection and tracking," in *AVSS*. IEEE, 2014, pp. 361–368.
- [108] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural networks*, vol. 108, pp. 466–478, 2018.
- [109] S. Yi, H. Li, and X. Wang, "Pedestrian travel time estimation in crowded scenes," in *ICCV*. Santiago, Chile: IEEE, 2015, pp. 3137–3145.
- [110] Y. Li, "A deep spatiotemporal perspective for understanding crowd behavior," *Transactions on multimedia*, vol. 20, no. 12, pp. 3289–3297, 2018.
- [111] P. Rota, N. Conci, N. Sebe, and J. M. Rehg, "Real-life violent social interaction detection," in *ICIP*. IEEE, 2015, pp. 3456–3460.
- [112] L. Lazaridis, A. Dimou, and P. Daras, "Abnormal behavior detection in crowded scenes using density heatmaps and optical flow," in *EUSIPCO*. IEEE, 2018, pp. 2060–2064.
- [113] H. Ammar and A. Cherif, "Deeprod: A deep learning approach for real-time and online detection of a panic behavior in human crowds," *Machine Vision and Applications*, vol. 32, no. 3, pp. 1–15, 2021.
- [114] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—a new baseline," in *CVPR*, 2018, pp. 6536–6545.
- [115] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *AVSS*. IEEE, 2017, pp. 1–6.
- [116] M. U. K. Khan, H.-S. Park, and C.-M. Kyung, "Rejecting motion outliers for efficient crowd anomaly detection," *Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 541–556, 2018.
- [117] R. Leyva, V. Sanchez, and C.-T. Li, "Abnormal event detection in videos using binary features," in *TSP*. IEEE, 2017, pp. 621–625.
- [118] K. Deepak, S. Chandrakala, and C. K. Mohan, "Residual spatiotemporal autoencoder for unsupervised video anomaly detection," *Signal, Image and Video Processing*, vol. 15, no. 1, pp. 215–222, 2021.
- [119] M. George, C. Bijitha, and B. R. Jose, "Crowd panic detection using autoencoder with non-uniform feature extraction," in *ISED*. IEEE, 2018, pp. 11–15.
- [120] S. Majhi, R. Dash, and P. K. Sa, "Temporal pooling in inflated 3dcnn for weakly-supervised video anomaly detection," in *ICCCNT*. IEEE, 2020, pp. 1–6.
- [121] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "Cnn features with bi-directional lstm for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16979–16995, 2021.
- [122] S. Akti, G. A. Tataroğlu, and H. K. Ekenel, "Vision-based fight detection from surveillance cameras," in *PTA*. IEEE, 2019, pp. 1–6.
- [123] M. Pourreza, M. Salehi, and M. Sabokrou, "Ano-graph: Learning normal scene contextual graphs to detect video anomalies," *arXiv preprint arXiv:2103.10502*, 2021.
- [124] S. Kapoor and R. Bhatia, *IntelliSys, Volume 1*, 2020, vol. 1250.
- [125] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "Toyadmos2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," *arXiv preprint arXiv:2106.02369*, 2021.
- [126] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection," in *WASPAA*. IEEE, 2019, pp. 313–317.
- [127] J. C. Socoró, G. Ribera, X. Sevillano, and F. Alías, "Development of an anomalous noise event detection algorithm for dynamic road traffic noise mapping," in *ICSV*, Florence, Italy, 2015, pp. 12–16.
- [128] L. Pham, I. McLoughlin, H. Phan, M. Tran, T. Nguyen, and R. Palaniappan, "Robust deep learning framework for predicting respiratory anomalies and diseases," in *EMBC*. IEEE, 2020, pp. 164–167.
- [129] D. Y. Oh and I. D. Yun, "Residual error based anomaly detection using auto-encoder in smd machine sound," *Sensors*, vol. 18, no. 5, p. 1308, 2018.
- [130] P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento, "Cascade classifiers trained on gammatonegrams for reliably detecting audio events," in *ICASSP*. IEEE, 2014, pp. 50–55.
- [131] N. Strisciuglio, M. Vento, and N. Petkov, "Learning representations of sound using trainable cope feature extractors," *Pattern recognition*, vol. 92, pp. 25–36, 2019.
- [132] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "Automatic detection and classification of audio events for road surveillance applications," *Sensors*, vol. 18, no. 6, p. 1858, 2018.
- [133] A. Greco, N. Petkov, A. Saggese, and M. Vento, "Aren: A deep learning approach for sound event recognition using a brain inspired representation," *Transactions on Information Forensics and Security*, vol. 15, pp. 3610–3624, 2020.
- [134] S. S. Sethi, N. S. Jones, B. D. Fulcher, L. Picinali, D. J. Clink, H. Klinck, C. D. L. Orme, P. H. Wrege, and R. M. Ewers, "Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set," *Proceedings of the National Academy of Sciences*, vol. 117, no. 29, pp. 17049–17055, 2020.
- [135] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE*, 2017.
- [136] O. I. Provotor, Y. M. Linder, and M. M. Veres, "Unsupervised anomaly detection in time series using lstm-based autoencoders," in *ATIT*. IEEE, 2019, pp. 513–517.
- [137] E. Rushe and B. Mac Namee, "Anomaly detection in raw audio using deep autoregressive networks," in *ICASSP*. IEEE, 2019, pp. 3597–3601.
- [138] H. Chen, X. Yuan, Z. Pei, M. Li, and J. Li, "Triple-classification of respiratory sounds using optimized s-transform and deep residual networks," *IEEE Access*, vol. 7, pp. 32845–32852, 2019.
- [139] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *Transactions on biomedical circuits and systems*, vol. 14, no. 3, pp. 535–544, 2020.
- [140] F. Demir, A. Sengur, and V. Bajaj, "Convolutional neural networks based efficient approach for classification of lung diseases," *Health information science and systems*, vol. 8, no. 1, pp. 1–8, 2020.
- [141] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "Mimii due: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," *arXiv preprint arXiv:2105.02702*, 2021.
- [142] P. Primus, V. Haunschmid, P. Praher, and G. Widmer, "Anomalous sound detection as a simple binary classification problem with careful selection of proxy outlier examples," *arXiv preprint arXiv:2011.02949*, 2020.
- [143] Y. Koizumi, M. Yasuda, S. Murata, S. Saito, H. Uematsu, and N. Harada, "Spidernet: Attention network for one-shot anomaly detection in sounds," in *ICASSP*. IEEE, 2020, pp. 281–285.
- [144] J. C. Socoró, F. Alías, and R. M. Alsina-Pagès, "An anomalous noise events detector for dynamic road traffic noise mapping in real-life urban and suburban environments," *Sensors*, vol. 17, no. 10, p. 2323, 2017.
- [145] F. Alías and J. C. Socoró, "Description of anomalous noise events for reliable dynamic traffic noise mapping in real-life urban and suburban soundscapes," *Applied Sciences*, vol. 7, no. 2, p. 146, 2017.
- [146] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on dcse 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *arXiv preprint arXiv:2106.04492*, 2021.
- [147] K. Morita, T. Yano, and K. Tran, "Anomalous sound detection using cnn-based features by self supervised learning," *DCASE2021 Challenge, Tech. Rep.*, Tech. Rep., 2021.
- [148] K. Wilkinghoff, "Utilizing sub-cluster adacos for anomalous sound detection under domain shifted conditions," *DCASE2021 Challenge, Tech. Rep.*, Tech. Rep., 2021.
- [149] I. Kuroyanagi, T. Hayashi, Y. Adachi, T. Yoshimura, K. Takeda, and T. Toda, "Anomalous sound detection with ensemble of autoencoder and binary classification approaches," *DCASE2021 Challenge, Tech. Rep.*, Tech. Rep., 2021.
- [150] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.
- [151] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *ECCV*. Springer, 2020, pp. 322–339.

- [152] C. Fayet, A. Delhay, D. Loline, and P.-F. Marteau, “Emo&ly (emotion and anomaly): A new corpus for anomaly detection in an audiovisual stream with emotional context,” in *LREC*, 2018.
- [153] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier, “Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation,” *Multimedia Tools and Applications*, vol. 74, no. 17, pp. 7379–7404, 2015.
- [154] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, “An audio-visual dataset of human–human interactions in stressful situations,” *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 29–41, 2014.
- [155] R. Belmonte, A. Aissaoui, S. Mihoubi, B. Allaert, J. Mennesson, I. M. Bilasco, and L. Goncalves, “Barem: A multimodal dataset of individuals interacting with an e-service platform,” in *CBMI*, 2021.
- [156] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, “Recognizing stress using semantics and modulation of speech and gestures,” *Transactions on Affective Computing*, vol. 7, no. 2, pp. 162–175, 2015.
- [157] B. M. Peixoto, B. Lavi, Z. Dias, and A. Rocha, “Harnessing high-level concepts, visual, and auditory features for violence detection in videos,” *Journal of Visual Communication and Image Representation*, p. 103174, 2021.
- [158] X. Li, Y. Huo, Q. Jin, and J. Xu, “Detecting violence in video using subclasses,” in *ACM MM*. ACM, 2016, pp. 586–590.
- [159] W.-F. Pang, Q.-H. He, Y.-j. Hu, and Y.-X. Li, “Violence detection in videos based on fusing visual and audio information,” in *ICASSP*. IEEE, 2021, pp. 2260–2264.
- [160] P. Wu and J. Liu, “Learning causal temporal relation and feature discrimination for anomaly detection,” *Transactions on Image Processing*, vol. 30, pp. 3513–3527, 2021.