A Comprehensive Review on Vision-Based Violence Detection in Surveillance Videos

FATH U MIN ULLAH, Sejong University, Seoul, South Korea
MOHAMMAD S. OBAIDAT, Life Fellow of IEEE, Chair & Professor, Computer Science Department, and Director of Cybersecurity Center, University of Texas-Permian Basin, Odessa, TX, with King Abdullah II School of Information Technology, University of Jordan, Amman, Jordan, School of Computer and Communication Engineering, with University of Science and Technology Beijing, China and with The Amity University, Noida, UP, India
AMIN ULLAH, Collaborative Robotics and Intelligent Systems (CoRIS) Institute, Oregon State University, Corvallis, Oregon
KHAN MUHAMMAD, Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul, South Korea
MOHAMMAD HIJJI, Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia
SUNG WOOK BAIK, Sejong University, Seoul, South Korea

Recent advancements in intelligent surveillance systems for video analysis have been a topic of great interest in the research community due to the vast number of applications to monitor humans' activities. The growing demand for these systems aims towards automatic violence detection (VD) systems enhancing and comforting human lives through artificial neural networks (ANN) and machine intelligence. Extremely overcrowded

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2019R1A2B5B01070067) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00062).
Authors' addresses: F. U Min Ullah and S. W. Baik (corresponding author), Sejong University, Gwangjin-gu, Seoul 143-747, 05006, South Korea; email: sbaik@sejong.ac.kr; M. S. Obaidat, Chair & Professor, Computer Science Department, and Director of Cybersecurity Center, University of Texas-Permian Basin, 4901 E. University Blvd., Odessa, TX 79762, 79762, USA, with the King Abdullah II School of Information Technology, University of Jordan, Amman 11942, 11942, Jordan and with School of Computer and Communication Engineering, University of Science and Technology Beijing, 100083, China, Honorary Distinguished Professor, Amity University, Noida, Uttar Pradesh 201301, 201303, India; A. Ullah, Collaborative Robotics and Intelligent Systems (CoRIS) Institute, Oregon State University, Corvallis, OR, 97331, USA; K. Muhammad (corresponding author), Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Jongno-gu, Seoul 03063, 16419, South Korea; email: khan.muhammad@ieee.org; M. Hijji, Faculty of Computers and Information Technology, University of Tabuk, Tabuk 47711, 71491, Saudi Arabia.
Updated author affiliation: Mohammad S. Obaidat, Chair & Professor, Computer Science Department, and Director of Cybersecurity Center, University of Texas-Permian Basin, 4901 E. University Blvd., Odessa, TX 79762, USA, with the King Abdullah II School of Information Technology, University of Jordan, Amman 11942, Jordan and with School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China, Honorary Distinguished Professor, Amity University, Noida, UP 201301, India.

0360-0300/2022/10-ART200 $15.00
https://doi.org/10.1145/3561971

regions such as subways, public streets, banks, and the industries need such automatic VD system to ensure safety and security in the smart city. For this purpose, researchers have published extensive VD literature in the form of surveys, proposals, and extensive reviews. Existing VD surveys are limited to a single domain of study, i.e., coverage of VD for non-surveillance or for person-to-person data only. To deeply examine and contribute to the VD arena, we survey and analyze the VD literature into a single platform that highlights the working flow of VD in terms of machine learning strategies, neural networks (NNs)-based patterns analysis, limitations in existing VD articles, and their source details. Further, we investigate VD in terms of surveillance datasets and VD applications and debate on the challenges faced by researchers using these datasets. We comprehensively discuss the evaluation strategies and metrics for VD methods. Finally, we emphasize the recommendations in future research guidelines of VD that aid this arena with respect to trending research endeavors.

CCS Concepts: • **Computing methodologies** → *Artificial intelligence*; *Machine learning*; • **Applied computing** → *Surveillance mechanisms*; • **Human-centered computing** → *Human computer interaction (HCI)*;

Additional Key Words and Phrases: Artificial Intelligence, machine learning, smart surveillance, neural networks, deep learning, violence detection, big data, video data, activity recognition

ACM Reference format:
Fath U Min Ullah, Mohammad S. Obaidat, Amin Ullah, Khan Muhammad, Mohammad Hijji, and Sung Wook Baik. 2022. A Comprehensive Review on Vision-Based Violence Detection in Surveillance Videos. *ACM Comput. Surv.* 55, 10, Article 200 (October 2022), 44 pages.
https://doi.org/10.1145/3561971

1 INTRODUCTION

The advancements in technology and increase in the number of installed surveillance cameras across the globe have made human lives comfortable through adapting machine intelligence and **neural networks (NNs)-based** learning. The purpose of these cameras is the monitoring of human activities [1, 2], object detection [3, 4], protection of human assets, and finding the state of certain actions via CCTV footage. However, as the number of these cameras has increased, involving humans in their monitoring becomes costly and problematic to manage intelligently [5]. An automatic system for monitoring these activities will ease the detection and recognition of ongoing events. The main objective of detecting these events is to reduce crime rates and create a more secure and safe environment. These events involve different abnormalities such as VD [6], crowded scenes [7], other anomalies [8–11], and the like. Consequently, violence is an abnormal action which involves physical force that harms human beings or damages the state of something. According to a report [12], 48% of people in South Korea killed by interpersonal violence in 2015 were killed with sharp objects, such as knives, whereas the deaths from sharp objects were around 25%. Similarly, Yoon et al. [13] stated that arrests happened due to violent events in South Korea were 41 thousand, remained more than half of the other violent offenses and the arrests stood at 67 thousand. Another report [14] states that 1.2 million violent crimes happened in the United States in 2019. These crimes encompassed mass shootings, fighting, and other aggressive actions, while in China [15], this rate was 4.86 million in 2019, which was less than the previous year. Such activities need to be detected before the occurrence of any catastrophic situation. Several techniques with realistic results came into being to ensure the safety via secure surveillance. VD is closely related to public and commercial security, and its localization has gained widespread attention among academia, security headquarters, law-enforcement, and industry. In surveillance scenarios, certain activities occur that are widely normal or abnormal activities. The normal activities are usually related to daily human actions and movement without making any interruption/disruption in the environment such as walking, eating, running, and the like. Though, there are certain abnormal events that are further categorized into several events such as Robbery, Vandalism,

Table 1. The Representation and Taxonomy of NNs Used throughout the VD Works

Category	Publications
RNN	[16–30]
CNN/ConvNets	[17, 20, 24, 31–63]
CNN-LSTM/CNN-BiLSTM	[26, 27, 29, 64–67]

Accidents, Fighting, Snatching, and so on. All these categories belong to the class "Anomaly" and their identification in surveillance videos is called "anomaly recognition". However, there are higher chances of Fighting in real-world surveillance that happens in higher number. Therefore, the detection of fighting is termed as "violence detection" in the computer vision literature.

Recently, an increasing interest in NNs-based learning approaches for video analytics has been observed in many domains of computer vision. Consequently, motivations have been obtained from several aspects of these NNs such as CNN [68], RNN [69], deep autoencoder [70], GNN [71], GAN [72], and so on. The important operations and learning-network engineering have made these NN models able to handle the complex tasks of video analytics. The number of reviews to briefly cover the NNs for VD are limited. Using the term "deep learning", Jain et al. [73] introduced the overview of ConvNets in the domain of VD in the first survey that concisely envelops the existing NNs and their categorization for VD. We categorized the VD literature based on NNs categorization that is given in Table 1. Similarly, in the VD literature, the majority of methods follow machine- and deep learning strategies to effectively learn the data patterns that are based on different invariants. The common techniques in the literature that are used for VD are clustering, sequential learning, ANN, optical flow, and CNN, along with different types of classifiers, such as SVM, decision trees, and so on. From this perspective, we have covered the most recent literature from 2012 onwards, encompassing the overall learning systems from both the journals and conferences with a focus on NNs. Surveying the literature, we examined that most of the work is covered in conferences of IEEE and ACM. Going deeper, we used the most popular search engines to gather the VD literature, including Google, Google Scholar, ACM Digital Library, Hindawi, ScienceDirect, and IEEE Explore.

In collecting the literature, a year-wise search procedure was undertaken that made the searching strategy more convenient. The initial search for each paper was primarily performed in Google and Google Scholar, followed by ScienceDirect, then IEEE Explore, and other search engines. We noted from the search that the number of papers on VD increased with the passage of time, i.e., year-wise. For instance, initial few years obtained very few researchers' attention for VD methods and got popularity in the later years. The number of articles published each year is visually presented in Figure 1(a), which clearly highlights the increase in research interest in VD work from past to future.

Nowadays, CCTV cameras are widely installed everywhere for security, safety, and asset protection. According to the statistics [74], both the national and local South Korean governments are installing more than ten thousand CCTV cameras every year in public. In 2018, about one million CCTV cameras were in operation, which is a 200% increase in comparison to the previous five years. Of this number, half of them are installed for facility safety and fire prevention. About 45% and 4% are installed for crime prevention and traffic law enforcement, respectively. Similarly, this number for China was about 626 million by 2020, encompassing both privately and publicly installed cameras [75]. However, in the United States, around 50 million cameras are installed in total compared to China, which shows roughly 200 million according to this report [76]. These surveillance systems are proving to be very beneficial particularly in law enforcement services in helping them with the records to solve the majority of serious crimes. So far, when surveillance systems

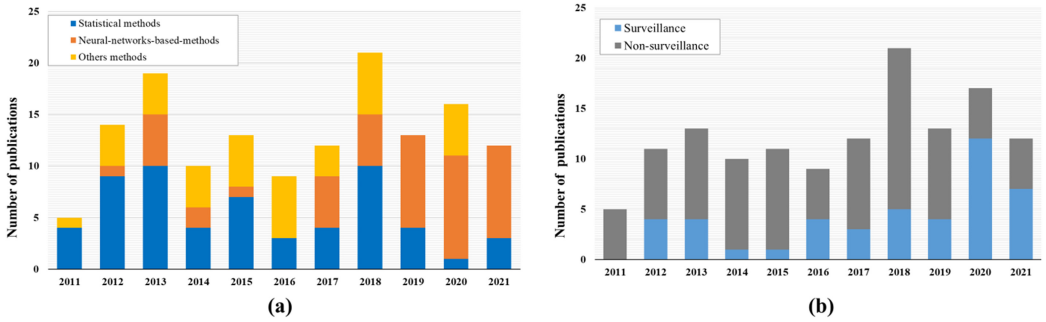

Fig. 1. Overview of the VD statistics. **(a)** Advancements in the field of VD and its research trends from 2011 to 2020. Similarly, the year-wise categorization of publications is visualized based on the used methods. From this, it is cleared that there is huge number of VD works published in 2013 and 2018, but we also noted that there is a huge increase in neural networks-based methods. **(b)** Year-wise distribution and number of papers of VD methods based on surveillance and non-surveillance. It also shows that a great increase in surveillance-based methods occurred.

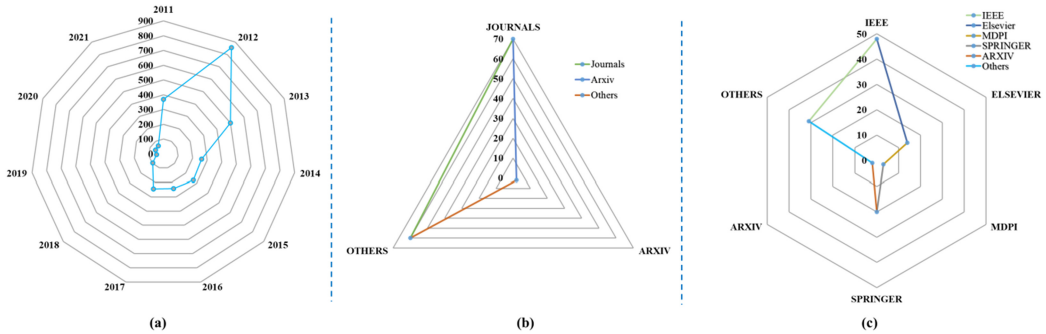

Fig. 2. Statistics for the citations score, proceedings-wise, and publisher-wise distribution of VD methods. **(a)** The citations achieved by VD research for each year, where 2012 is the most cited year as the VD methods of this year have been well explored in recent research; **(b)** Distribution among different journals and other sources, covering 50% of the articles from the journals; **(c)** Division of publications in each portal in which IEEE shows a large number of published articles.

are operating 24 hours a day, that makes it tough to monitor them manually for suspect detection. On account of this, an intelligent surveillance system is needed to overcome this problem. Inspired by these statistics, we have a keen interest in the VD literature about both surveillance and non-surveillance-based methods. This distribution is well illustrated in Figure 1(b).

The methods that use data obtained through CCTV footage are considered in the surveillance-based category. The majority of initial methods are based on non-surveillance systems that cover violent scenes in movies, videos taken by phones, dashboard cameras in cars, or scenes from normal web cameras. Furthermore, we noted from the year-wise distribution that the earlier methods are less explored for surveillance operations. However, as a growing popularity along with the installation of surveillance systems in both the public and private organizations occur, they are included in the VD methods. So far, we have divided the VD methods based on the datasets and mechanisms they used.

Hallmarks of good research require deep investigation with the ability to acquire quality research. Citing articles increases the impact of work and journals. The highest citation score indicates that certain articles have had a positive influence on the research community. In view

Table 2. Keywords used as Query for Searching VD Papers in Different Portals

#	Terms	Portal			Total retrieved	Selected	Rejected
		IEEE Xplore	Google Scholar	Other			
1	Violence detection	✓	✗	✗	60	48	12
		✗	✓	✗	1350	35	1315
		✗	✗	✓	28250	35	28215
2	Violence detection in videos/images	✓	✗	✗	3	3	0
		✗	✓	✗	89	9	80
		✗	✗	✓	5910	9	5901
3	Violence detection in surveillance videos/images	✓	✗	✗	1	1	0
		✗	✓	✗	14	3	11
		✗	✗	✓	N	4	N-4
4	Violent scene in videos/surveillance videos/images	✓	✗	✗	0	0	0
		✗	✓	✗	1	0	1
		✗	✗	✓	✗	✗	✗
5	Current challenges in violence detection	✓	✗	✗	0	0	0
		✗	✓	✗	0	0	0
		–	✗	✓	N	5	N-5
6	Violence detection reviews/surveys/a review	✓	✗	✗	0	0	0
		✗	✓	✗	0	0	0
		✗	✗	✓	N	4	N-5

The search was carried out between 2011 to 2020 in specific portals where the total retrieved is labeled as *n*, while *n*-selected shows the rejected papers. Total portal consists of three parts, namely IEEE XPLORE, Google Scholar, and other. The other contains Google and MDPI.

of this, we distributed all the articles based on the citations of each year that are presented in Figure 2(a). Similarly, Figure 2(b) shows the overall article coverage included in journals, arXiv, and other sources. These sources include 48% of papers from conferences and book chapters.

Similarly, there are different portals in the research community that publish papers in the field of VD. We categorized the papers on the basis of these portals, such as Elsevier, Springer, MDPI, and IEEE. This distribution is shown in Figure 2(c). We applied various key terms for querying these methods. These terms are given in Table 2 where we have shown the total number of retrieved and selected publications from the retrieved results for VD in 2012 using Google Scholar. This representation provides an excellent view of other queries. Furthermore, there are some abbreviations used throughout the survey for representation. The nomenclature details of these abbreviations are given in Table 3.

There has been a race in the field for automatic VD in surveillance systems using deep learning. Several problems and issues are faced by the researchers due to the VD subjective nature. The main aim of this survey is to form a compact representation of VD methods in a solo platform with a diverse range of paper coverage focusing on NNs and their learning strategies in surveillance and non-surveillance systems. Existing surveys lack coverage of important articles, surveillance datasets, and VD applications in real-world scenarios, and they only emphasized VD for person-to-person data. There is no proper explanation for the detailed NNs and their learning for VD, VD applications, challenges, along with problems in VD and their solutions. To overcome these challenges and limitations, we survey the VD that mainly focuses on both surveillance and non-surveillance setups.

The notable contributions of this survey are summarized as follows:

Our contributions:

Table 3. Nomenclature of Various Terms and Parameters Used throughout This Study

Symbols	Descriptions	Symbols	Descriptions
ASLAN	The Action Similarity Labeling	KDE	Kernel Density Estimation
AUC	Area Under the Curve	LSTM	Long/Short-Term Memory
AI	Artificial Intelligence	LHOG	Local Histogram of Oriented Gradient
B-LSTM	Bidirectional LSTM	LBP	Local Binary Pattern
BRISK	Binary Robust Invariant Scalable Key points	MAP	Mean Average Precision
BOW	Bag of Words	MoSIFT	Motion SIFT
CNN	Convolutional Neural Network	ORB	Oriented-FAST and Rotated-BRIEF
ConvLSTM	Convolutional LSTM	ROC	Receiver Operating Characteristic Curve
EER	Equal Error Rate	RD	Rate of Detection
FAST	Features from Accelerated Segment Test	RIMOC	Rotation Invariant Feature Modeling Motion Coherence
GAN	Generative Adversarial Neural Network	RMV	Region Motion Vector
GMM	Gaussian Mixed Model	RNN	Recurrent Neural Network
GNN	Graph Neural Network	RD	Rate of Detection
GLCM	Gray-Level Co-Occurrence Matrix	SIFT	Scale-Invariant Feature Transform
HoG	Histogram of Oriented Gradient	SVM	Support Vector Machine
HoF	Histogram of Oriented Optical Flow	UCF	University of Central Florida
HOT	Histogram of Oriented Tracklets	UCSD	University of California, San Diego
HOMO	Histogram of Optical Flow Magnitude and Orientation	UAR	Unweighted Average Recall
IWLD	Improved Weber Local Descriptor	VSD	Violent Scenes Detection
IFV	Improved Fisher Vectors	ViF	Violent Flow

(1) *New Taxonomy.* To the best of our knowledge, this is the first survey that delivers a new taxonomy for the most recent [2011 ~ onward] VD literature in terms of NNs that analyzes the nature of video patterns for VD with both deep learning and traditional methods, making a compact presentation.
(2) *Comprehensive Review.* We provide comprehensive reviews of each VD method, publishers/portals, journal/conference details, citation coverage, and VD applications in both surveillance and non-surveillance domains. Next, we comparatively surpass the existing reviews/surveys by overcoming their limitations.
(3) *Abundant Coverage.* Existing surveys cover either person-to-person violence or crowd violence-related articles, which is a tedious way for readers to handle the distinct literature on different platforms. For this purpose, we incorporate both the VD articles based on their baseline strategy for person-to-person and crowded violence articles including surveillance and non-surveillance, helping the readers to avoid several sources. Next, we visually present the summarized working flow of VD from the input data module to the final output with the intermediate steps.
(4) *Guidelines and Future Directions.* Unlike existing VD survey literature, we comprehensively discuss the VD datasets used in surveillance and non-surveillance domains and the challenges faced by researchers using these datasets. Similarly, we cover the applied techniques on these datasets with deep analysis and their evaluation strategies/metrics for VD. Furthermore, we present the current challenges, their solutions with modern machine intelligence, and future research guidelines.

The rest of this article is organized as follows: Section 2 covers the existing surveys and their limitations. Section 3 discusses the working flow of VD methods, while Section 4 explores the datasets being used in VD. Section 5 explains VD applications. Subsequently, we include current

Table 4. Coverage of the Most Recent Surveys on VD with Their Deep Learning (DL) Usage, Number of VD Papers Discussed, Their Year Coverage, Key Contributions, and Limitations Compared to Our Survey

Ref.	Year	DL	Number of papers	Years Coverage	Key Contributions	Remarks
[80]	2018	✗	30	2016~2018	Surveyed VD techniques along with their advantages and disadvantages and presented their comparison based on performance. Focused on the techniques, including statistical hypothesis detector, sparse reconstruction, and GMM.	The working flow of the techniques is limited to being explained as existing surveys do. Coverage of very few articles is found. No paragraph explains the methods and background.
[79]	2018	✗	––	2012~2017	Explained the steps that are mainly composed of surveillance systems for behavior modeling and its representation. Discussion of features extraction and behavior representation. Classification methods and the frameworks to model the behavior.	Discussion on approaches and systems for crowded scenes only. Coverage of non-surveillance datasets only, which are fewer in numbers. No challenges in VD are found.
[78]	2019	✗	20	Detail coverage is missing.	Categorized VD methods, and their challenges are highlighted. Datasets relevant to their published articles are given.	Coverage of a limited number of articles related to VD obtained from the old literature. No details about the evaluation of the approaches. Visual representation to cover the approaches is missing.
[77]	2019	✗	28	Major coverage is given to 2016~ 2018	Division of the VD methods into three categories, such as traditional learning, deep learning, and methods based on SVM. The features extraction procedure for each method is presented. Brief explanation about VD datasets is given.	Coverage of majority of traditional machine learning approaches. (*Less focus on deep learning models*). No discussion about future directions or guidelines and challenges in VD.
[81]	2020	✓	57	––	Categorization of VD methods in terms of ConvNets.	VD methods other than ConvNets are missing. No discussion about the future directions or research guidelines.
Ours	2022	✓	125	2011~2021	Thorough analysis of VD literature in terms of both surveillance and non-surveillance, NNs categorization for VD, covering the current reviews, criticism of them, index of importance through citation score, publishers, and journal details, and applications. Incorporation of person-to-person VD and crowd VD, surveillance and non-surveillance datasets, challenges faced by researchers, evaluation metrics for VD, and future research directions.	Discussion of both deep learning and traditional methods for VD with working flow, visual presentation, and challenges. Index of impact analysis for VD and future research direction with recommendation is provided.

challenges and future directions with recommendations in Section 6. Finally, Section 7 concludes the overall survey.

2 OUTLINES AND COVERAGE OF EXISTING SURVEYS AND THEIR LIMITATIONS

This section discusses the existing surveys published in the VD domain. We apply several search queries and exploit VD-related keywords to obtain the most significant papers from large repositories. The origin of different papers is explored, such as conferences, arXives, journals, and the like. The majority of VD literature is published in Elsevier, followed by IEEE.

There are several contributions to VD in the form of reviews and surveys. The review articles are presented in Table 4, which thoroughly shows a detailed comparison in terms of key contributions, their remarks, and the NN-based categorization. We focus on the most recent and relevant VD surveys. For instance, recently, a review article [77] published in *IEEE Access* covered the VD

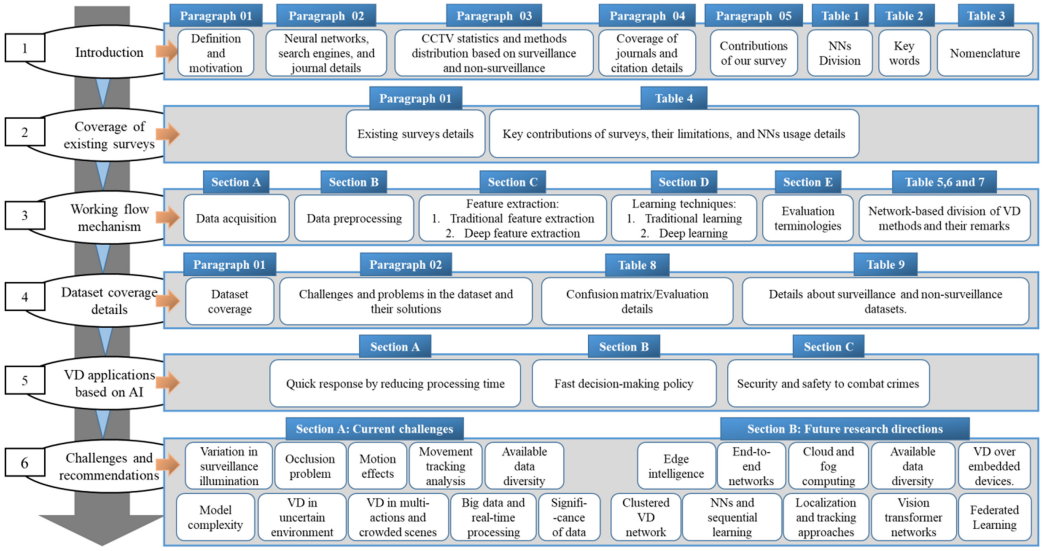

Fig. 3. Structure and contents flow of the overall survey comprised of the compact and summarized details for each section. Information about certain VD directions can be easily followed in all the sections and subsections.

techniques proposed before 2019. This survey explained each method/paper in a separate section, with a focus on features extraction. We determined that most of the old VD literature is covered in their survey, and there is no discussion about NNs or their categorization. [77] Further, another survey [78] reviewed the current challenges and problems in the field of VD. They briefly explained VD methods and covered very few papers from the VD literature. They also skipped the most recent techniques, NNs, and algorithms used for VD. Furthermore, the review presented in [79] focused on abnormal behavior and its modeling for surveillance systems. Secondly, they considered the features extraction and classification methods. They also discussed surveillance systems that are implemented in real world. In reviewing the VD survey, we found that most of the surveys have covered either the crowd scenes or a violent scene for person-to-person data only. Therefore, in our survey, we cover the most important and significant works related to violence both in crowded and non-crowded scenes with NN-based distribution.

3 WORKING FLOW OF VD METHODS

In this section, we discuss all the basic steps performed in VD. The basic steps include data acquisition, its preprocessing, features extraction, and classification of video based on the features. The final action is taken according to the results obtained from the classification. We visually illustrate the overall survey structure and its flow in Figure 3, while the compact representation of the statistical, NN/Deep learning, and other VD methods, is described in Tables 5–7, respectively. Similarly, the generalized VD framework is presented in Figure 4.

3.1 Data Acquisition

In this section, we discuss the data collection for VD, which are obtained through different sources. These sources contain both indoor and outdoor surveillance, videos from movies, and videos recorded through smartphones or simple cameras. Over the last few years, infrastructure growth has been found for security- and crime-related issues related to public areas, streets, marts, and inside buildings. With increased demand for security and safety, surveillance-based video analysis

Table 5. The Working Flow of the VD Methods Based on Statistical Learning Systems Employed along with Evaluation Metrics, Datasets, and Applicability for Surveillance (SU)

SU?	Ref.	System Flow	Evaluation Metrics	Datasets
Non-Surveillance Methods	[110]	• BoW Concept is used to recognize the action specific to fight detection with two major descriptors such as STIP and MoSIFT.	▪ Accuracy	• Hockey fight • Movies
	[111]	• Mainly focused on semantic-complete structure of a scene in a video. • Features extracted from the segmented scenes are fed into SVM for classification.	▪ Accuracy ▪ Precision ▪ Recall	• Hollywood movies
	[112]	• A classification scheme to detect violence using a set of features that are formed from video, audio, and subtitles of movies is proposed.	▪ Precision ▪ Recall	• Hollywood movies
	[113]	• An optical flow context histogram-based method is proposed that detects the abnormal event, especially the fighting scenes.	N/A	• Videos from Internet.
	[114]	• A multimodel technique is developed that uses both the visual and audio features, where the visual features are extracted through SIFT, and a BoW approach is applied to learn by SVM.	N/A	• Media-Eval 2011 dataset
	[115]	• Database with violent and non-violent classes is created and uses the BoW model in the training process.	▪ Average Classification Accuracy	• VID
	[116]	• System for person-to-person VD is proposed using acoustic and visual signals to detect violence indoors. • Based on the detection, final fusion is done from each acoustic and visual detector.	▪ Accuracy	• Created dataset
	[117]	• Discriminative slow feature analysis is presented, where the extracted features are learned through the slow features function. • Next, accumulated squared derivative features are used to represent the video. SVM is trained for final classification.	N/A	• Violence video dataset
	[118]	• The mid-level visual and auditory features are used to predict the set of mid-level concepts, and multi-layer perceptron is used for classification purposes.	▪ Precision ▪ Recall ▪ F1-score	• Hollywood movies
	[119]	• The dynamic pedestrian agent is used to learn the behavior patterns of the crowd.	▪ Mean deviation	• Dataset from New York Grand Central
	[120]	• The low-level features are used to infer the concept of prediction. • A classifier comprised of multilayer perceptron is employed.	▪ Precision ▪ Recall ▪ F1-score	• Media-Eval 2012
	[121]	• Segmental features are combined with audio-visual to detect violence.	▪ MAP ▪ UAR	• Media-Eval 2012
	[122]	• Mid-level features are employed and compared to low-level audio and visual features, where the mid-level audio cues are fused with low-level visual ones. • Next, SVM is trained to detect violent scenes.	N/A	N/A
	[123]	• This method mostly focused on fight detection and is inspired by the kinematic features. • It used extreme acceleration patterns, which are estimated through radon transform, as the main features.	▪ Precision ▪ Recall ▪ Accuracy ▪ MAP	• Media-Eval
	[124]	• Conditional random fields used to refine the movie shots detection holistically.	▪ MAP	• Media-Eval

(Continued)

Table 5. Continued

SU?	Ref.	System Flow	Evaluation Metrics	Datasets
	[125]	• Mid-level audio features are employed based on BoW with mel-frequency cepstral coefficient quantization-based (VQ-based) method, • Binary SVM is used for classification.	▪ AP ▪ Precision	• Media-Eval VSD
	[126]	• Human interaction is detected and localized in videos by focusing on two major facts, i.e., interaction of single subject and interaction of subject with other subjects.	▪ AUC ▪ ROC	• UT-dataset
	[127]	• Audio and visual features are fused to detect violent content, mid- and low-level features known to be audio-visual features, are used • Next, the Mel-frequency cepstal coefficient, ViF, and HoF are extracted. • Color-related descriptors are used and partitioning in the feature space of violence samples through k-means clustering, is performed.	▪ Precision ▪ Recall ▪ MAP	• Hollywood movies • User-generated videos • Web videos
	[128]	• BoW is utilized for fight detection and spatiotemporal features are extracted from videos to classify the violent scenes. • Similarly, features like motion blobs from the video sequence are used to discriminate between the fight and non-fight scenes.	▪ ROC	• Violence in movies • Hockey fight
	[129]	• Low-level visual and audio features are used to detect violent scenes.	▪ MAP ▪ AP	• Media-Eval
	[130]	• MoSIFT algorithm is used as low-level description of a query video, and the KDE is exploited for feature selection on MoSIFT descriptor.	▪ Accuracy	• Violent flow • Hockey fight
	[131]	• Statistic method based on optical flow is proposed to detect violent scenes in a crowded environment. • Statistical characteristic of optical flow to represent the sequence of video frames are used, which are classified by SVM.	▪ Accuracy	• Hockey fight • Crowd dataset
	[132]	• Features that provide the strong visual and audio cues revealing multimodel patterns are applied to detect the violent scenes.	▪ Average ▪ Precision	• Media-Eval 2013 multimedia benchmark
	[133]	• Fight detection method based on spatiotemporal interest points is proposed and combines spatiotemporal features with motion energy to discuss video information.	N/A	N/A
	[134]	• Local features based on the SIFT algorithm are incorporated to introduce the appearance and Lagrangian motion-based models.	▪ Accuracy ▪ AUC	• Crowd violence • Hockey fight
	[135]	• Dense trajectories are refined to select the most discriminative trajectories. • Motion boundary histogram and oriented gradient are computed. Next, the extracted features descriptor is encoded via the super descriptor vector method where they are arranged as tensors to retain the spatiotemporal structure. • The final features are fed into SVM for classification.	▪ Accuracy	• Media-Eval VSD
	[136]	• Enhanced histogram of oriented tracklets-based method is improved, and the naïve count-based histogram is replaced by richer statistics of crowd movements.	▪ EER	• UCSD • BEHAVE • UMN

(Continued)

Table 5. Continued

SU?	Ref.	System Flow	Evaluation Metrics	Datasets
	[137]	• ViF with Horn-Schunck is proposed instead of iterative reweighted least squares, and SVM is used for classification.	▪ Accuracy	• Hockey fight • Violence in movies • Crowded dataset
	[108]	• Input video stream is converted into frames, and the optical flow features are extracted from the consecutive frames. • Next, the acceleration field is extracted in accordance with the optical flow field.	▪ Accuracy	• VID
	[138]	• Spatiotemporal features are used for VD, and a descriptor known as distribution of magnitude and orientation of local interest frame is utilized to train the binary class SVM.	▪ Accuracy ▪ AUC ▪ ROC	• Violent Flows • Hockey Fight
	[139]	• SIFT descriptor is used to extract the features from the images, and a set of descriptors is defined as a means to provide fast and accurate comparisons between violent and non-violent.	▪ Accuracy	N/A
	[140]	• Mainly two steps, such as object tracking and behavior understanding, are used to detect violence. • Features such as speed, direction, dimensions, and centroid are identified. Similarly, two features vectors, i.e., violent flow and LBP, that are fed into SVM for classification are used.	▪ Accuracy	• Hockey fight • Violent flow
	[43]	• Hybrid approach of handcrafted features is learned through a framework for VD.	▪ Accuracy	• Hockey fight • Violence in movies
	[141]	• The candidate features are extracted, and the redundant information is eliminated. • Similarly, HOG-LBP and HOF are calculated, which described the appearance and motion of regions. • Finally, the violent behavior is detected using a one-class SVM model.	▪ Accuracy ▪ AUC ▪ ROC	• UCSD
	[142]	• A features descriptor known as HOMO is introduced. The input frames are converted into grayscale, and the optical flow is computed between the frames. • Finally, the SVM classifier is trained through HOMO descriptor.	▪ Accuracy	• Hockey fight • Violent flows
	[143]	• Moving filtering algorithm is used to check the surveillance videos based on a temporal derivative, where the filter avoided the non-violent actions, and only filtered frames are entered into features extraction. • Next, SIFT and histogram of optical flow feature (HOFF) with motion boundary are combined to form a MoBSIFT descriptor.	▪ Accuracy	• Hockey fight • Violence in movies
	[144]	• Constructing approach based on HoG3D and BoW model is proposed, where the HoG3D features are extracted at block level for video. Next, the K-means clustering is used to generate visual words. • BoW is used to quantize the features, and the feature-pooling technique is used to generate the feature vector.	▪ Accuracy	• Hockey fight • Violence in movies
	[145]	• Descriptor based on statistical features is proposed to detect violent activities in surveillance, and spatiotemporal features are extracted from videos that define the motion cues. • Next, a discriminative SVM classifier is used to classify violent and non-violent scenes.	▪ Accuracy	• Hockey fight

(Continued)

Table 5. Continued

SU?	Ref.	System Flow	Evaluation Metrics	Datasets
	[146]	• Frames and differential images are forked to obtain the appearance and motion features. • Next, the linear SVM is used to classify the features, while label fusion is used to improve the detection performance.	▪ Accuracy	• Hockey fight • Crowd violence
	[147]	• Several methods are explored to fuse the visual and audio information. A network is proposed that consists of three modules such as attention module, fusion module, and mutual learning.	N/A	• XD-Violence
	[148]	• Three deep learning models are proposed to test the AIRTLab data.	▪ Accuracy	• AIRTLab • Hockey fight • Crowd violence
Surveillance-based Methods	[149]	• Dynamic system defined by optical flow is initialized where a scene is overlaid via a particles grid, providing trajectories to represent motion and find the region of interest in the scene. • Jacobian matrix for behavior classification is used.	▪ Accuracy ▪ AUC	• (PETS) 2009
	[150]	• Context-based system is proposed to detect suspicious behavior, distinguishing between the contexts. • Data stream clustering is applied to update and retrieve knowledge. • Inference algorithm is used to make context-sensitive decisions.	▪ AUC ▪ P-value	• CAVIAR • Z-block
	[151]	• The flow-vector magnitude and its statistics are collected for frames sequence via an optical flow descriptor, which is used to classify violent and non-violent scenes.	▪ Accuracy	• Violent flow • Hockey fight • Aslan
	[152]	• Features from objects with interobject motion are extracted to detect semantic behaviors.	▪ Accuracy	• BEHAVE • CAVIAR • PETS
	[153]	• The sparse representation is used to detect abnormal events using an LSD model, which has low constraint rank. • Similarly, they designed the sparse reconstruction cost to measure the outlier for abnormal event detection.	▪ EER ▪ RD ▪ AUC	• UCSD
	[154]	• System for abnormal activity detection is developed where the spatiotemporal segmentation is performed from the video to describe the appearance information and motion of the spatiotemporal segment. • The abnormal event is considered to be a matching problem through searching the best match, and the compact random projections are adopted to speed up the search process.	▪ EER ▪ RD ▪ AUC	• UCSD
	[155]	• GMM is used in optical flow domain to detect the violence region in surveillance. • The histogram of optical flow orientation is used to measure the spatiotemporal features in each region.	N/A	• BEHAVE • CAVIAR
	[107]	• Method based on a motion weber local descriptor for motion analysis is proposed and extended through adding a temporal component that explicitly captures motion information with low-level image appearance information. • They eliminated the redundant features by proposing KDE and later on used sparse coding with max pooling to extract discriminative features.	▪ Accuracy	• Hockey fight • BEHAVE • Violent flows

(Continued)

Table 5. Continued

SU?	Ref.	System Flow	Evaluation Metrics	Datasets
	[156]	• Features using the Lagrangian direction field are presented that are based on a spatiotemporal model and utilized appearance and background compensation. • Next, they applied extended BoW in late fusion way as a classification.	▪ Accuracy ▪ AUC ▪ ROC	• Violence in movies • Hockey fight • Violent crowd • London riots
	[157]	• Motion region that depends on the distribution of optical flow fields is segmented. Features such as LHOG and local histogram of optical flow are extracted from the RGB images and optical flow images, respectively. • Next, BoW is used to code the extracted features to eliminate redundancy and the video-level vectors are classified using SVM.	▪ Accuracy ▪ AUC ▪ ROC	• Crowd violence dataset • Hockey fight • BEHAVE
	[158]	• Semi-supervised learning method is developed for VD with dictionary learning from labeled samples to discriminate.	▪ Accuracy	• Hockey fight • BEHAVE • Crowd violence
	[159]	• The spatiotemporal features and cubic trajectories are proposed to detect a fight. • This method is based on blob movement that creates trajectories and captures the motion specific to a fight.	▪ Accuracy	• UT-interaction • Violence in movies • Hockey fight
	[160]	• SVM classifier is trained with HoG features extracted from video frames.	▪ Accuracy	• UT-interaction
	[88]	• This method built two types of perception layers which are based on situation graph trees and SVM. • Next, they are fused through late fusion into an activity score.	▪ Sensitivity ▪ Specificity ▪ Precision ▪ Accuracy ▪ F1-Score	• BEHAVE • NUS-HGA • Videos from YouTube
	[161]	• Gaussian model of optical flow is developed to collect the candidate violence region where the features are modeled to be a deviation from normal crowd behavior. • Next, an orientation histogram of optical flow is proposed which is fed into SVM.	▪ Accuracy ▪ ROC ▪ AUC	• BEHAVE • CAVIAR • Crowd violence
	[162]	• Machine is made enable to understand high-level violence by breaking into smaller and more objectives including explosions, fights, blood, etc.	▪ Accuracy	• BEHAVE • Hockey fight • Violent flow • Violence in movies

has become an important arena in the research community [82]. In view of this, numerous indoor and outdoor surveillance cameras have been installed to provide safety and prevent crimes. The scope, such as prevention, intervention, and detection of violence, has led to the development of consistent and real-world surveillance systems capable of intelligent video processing in both indoor and outdoor surveillance [83]. The main purpose of data collection from both indoor and outdoor environments is to analyze the video scenes for abnormal events and to prevent any catastrophic situations. These data face different challenges, such as resolution, lighting, or blurring, that affect the detection process. For certain noises or outliers, the data enter preprocessing step to refine them for further investigation.

Content-based analysis of multimedia to search the violence in videos has several applications related to children protection and the care of elderly people. Several datasets are available where the video data are taken from Hollywood movies, mobile phones, or other action movies that contain the violent scenes. The source of data collection is based on applications of the VD system that need to be installed. Sometimes, the videos are recorded through smartphones when fight scenes are caught on the street. After acquiring this data, the violent scenes are detected in the

Table 6. The Working Flow of the VD Methods based on Neural Network/Deep Learning Employed along with Evaluation Metrics, Datasets, and Applicability for Surveillance (SU)

SU?	Ref	Main Contributions	Evaluation Metrics	Datasets
Non-Surveillance Methods	[44]	• 3D ConvNets model is developed to detect violent scenes in videos, without any prior knowledge.	▪ AUC ▪ ROC	• Hockey fight
	[32]	• An algorithm based on local spatiotemporal and optical flow features is introduced to detect violence behavior, and the Harris 3D spatiotemporal point detector is used, which is combined with optical flow.	N/A	N/A
	[33]	• Optical flows from the input videos are computed using Lucas-Kanade, and few templates are applied that overlap with optical flow magnitude. • Similarly, a pre-trained CNN model received those templates as input, and the deep features are extracted from the templates.	▪ Accuracy	• Violence in movies • Hockey fight • Violent crowd
	[20]	• The deep CNN is proposed to extract frame-level features which are then aggregated via variants of LSTM. • Next, the CNN with LSTM is enabled, which is capable of extracting spatiotemporal features to analyze the motion in video.	▪ Accuracy	• Violence in movies • Hockey fight • Violent crowd
	[34]	• This method integrated the trajectories and deep CNNs that use handcrafted features and deep learned features.	▪ Accuracy	• Hockey fight • Crowd violence dataset
	[35]	• The blur and radon transform with CNNs is developed to detect fight scenes in videos. • Next, the local motion is extracted from the blur information, and random transform is applied on local motion. • The video frames are then identified via transfer learning with the use of a pre-trained deep learning model (VGG-Net)	▪ Accuracy	• Hockey fight
	[36]	• Hundred image dataset is constructed as multitask crowd to detect violent behavior and calculate the density-level classification. • Next, a pre-trained ResNet Crowd model is used.	▪ AUC	• Multitask crowd dataset
	[21]	• Spatiotemporal encoder is introduced built on bi-convolutional LSTM architecture.	▪ Accuracy	• Hockey fight • Violence in movies • Violent flows
	[37]	• Framework based on deep learning is proposed to detect early violence, which is obtained through combination of the dynamic image and VGG16 network. • The former used the simple image which represents a video, and the latter scored the dynamic image from the partial event through monitoring the event completion degree.	▪ Accuracy ▪ Sensitivity ▪ Specificity	• Hockey fight • Crowd violence
	[22]	• Deep features are extracted from the image-based convolution that describes spatial information. • Next, multiscale convolutional features are described to handle video data variations, and bi-directional LSTM is used to learn the features.	▪ Accuracy	• Hockey fight • Movies • Real violence
	[19]	• A mini-drone video dataset recorded via drone is used, and the anomaly is detected via deep neural network which is composed of CNN and RNN.	▪ AUC	• UMN
	[23]	• The input frames are processed through the Spark framework where the features from each frame are extracted through the HoG function, which are then fed into the BDLSTM network for training.	▪ Precision ▪ Recall	• VID

(Continued)

Table 6. Continued

SU?	Ref	Main Contributions	Evaluation Metrics	Datasets
	[38]	• This method first extracted the salient features with spatiotemporal interest points. • Also, it extracted the descriptor in those regions with motion effect that lead to detecting violence in videos.	N/A	N/A
	[39]	• A triple-staged framework based on deep learning is proposed. • Persons in surveillance are detected via CNN model and spatiotemporal features are extracted from 16 frames through 3D CNN to detect violence.	▪ Accuracy ▪ Precision ▪ Recall ▪ AUC ▪ ROC	• Hockey fight • Violent flows • Violence in movies
	[47]	• This method relied on two DNN frameworks which learn the spatiotemporal features from videos with conceptual and subject-based way.	▪ Accuracy	• Media-Eval 2013 • VSD
	[40]	• This method focused on the VGG-16 network and the extraction of conceivable features descriptor from video spatial, rhythmic, depth information, and temporal information is performed.	▪ Accuracy ▪ Specificity ▪ Sensitivity	• Hockey fight
	[64]	• Features are extracted through ResNet-50 from video frames and are fed into ConvLSTM block.	▪ Accuracy	• KTH • Hockey fight • Violent flows
	[41]	• Multiplayer VD method that is based on 3D CNN, extracts the spatiotemporal features.	▪ Accuracy	• Multiplayer violence
	[183]	• Features are extracted via C3D network.	▪ Accuracy ▪ Precision ▪ Recall ▪ F1-score	N/A
	[184]	• A VD framework based on features derived from the handcrafted methods, is proposed. • The features include appearance, movement speed, and representative image that are fed into CNN.	▪ Accuracy ▪ Precision ▪ RecallF1-score	• Hockey fight • Violence in movies • Violent flows
	[185]	• An approach that extract the temporal features while exploiting 1D-CNN, is proposed.	▪ Accuracy	• Hockey fight • Violent flow
Surveillance-based Methods	[18]	• This method used fused spatial feature maps extracted through CNN, and multilevel fusion method is used to combine the spatial features map. • These maps are based on CNN and LSTM units. • They added additional residual layer blocks to boost accuracy, and the overall features are combined and fed into LSTM units for learning global temporal information.	▪ Accuracy	• Hockey fight • Violent flows • BEHAVE
	[17]	• Different features extractors are used to detect fights in surveillance videos, descriptors, such as CNN, 3D CNN, local interest points with different classifiers, i.e., end-to-end CNN, SVM, and LSTM are used.	▪ mAP ▪ F-Measure	• Hockey fight • Violence in movies • CCTV fights
	[16]	• LSTM is explored where attention layer is utilized for the VD problem in surveillance.	▪ Accuracy	• Hockey fight • Violence in movies • Surveillance dataset
	[31]	• Localized framework for fights in surveillance videos is presented, and optical flow maps are exploited to extract the motion information that indicates the location of the moving region. • Next, this method used two-stream 3D convolution network with novel motion acceleration on a temporal stream.	▪ Accuracy	• FADS • Hockey fight • Violence in movies • USF101

(Continued)

Table 6. Continued

SU?	Ref	Main Contributions	Evaluation Metrics	Datasets
	[58]	• CNN is used to extract the spatiotemporal features from the video frames that are combined with the trajectory features to detect violence.	▪ Accuracy	• Hockey fight • Crowd dataset
	[83]	• The features from the preprocessed frames are extracted via ConvLSTM and the generated latter is propagated into GRU for learning.	▪ Accuracy ▪ Precision ▪ Recall ▪ F1-score	• Surveillance fight • Hockey fight • RWF-2000
	[186]	• Existing VD video datasets are summarized and a dataset known as RWF-2000 is proposed that is captured via surveillance cameras	▪ Accuracy	• RWF-2000
	[82]	• A computationally intelligent approach is proposed for VD in surveillance where the frames containing objects are processed via CNN.	▪ Accuracy ▪ Precision ▪ Recall ▪ F1-score	• Hockey fight • Violent flow • Surveillance fight
	[187]	• A two-stream deep architecture that leverage the Separable LSTM and MobileNet, is proposed.	▪ Accuracy	• RWF-2000 • Hockey fight • Violence in movies
	[188]	• A VD pipeline that applied conventional 2D-CNNs is proposed. • The frames grouping procedure is proposed to make the network capable of learning the spatiotemporal representation.	▪ Accuracy	• Hockey fight • Violence in movies • Surveillance fight • RWF-2000

videos and are removed from them. After the removal of those violent scenes, children and old people can watch violence-free videos. Furthermore, the data collected through smartphones or Hollywood movies also pass through a preprocessing step for noise removal.

3.2 Data Preprocessing

This section discusses the preprocessing of video data prior to the actual processing or features extraction. At first, the videos obtained from the data acquisition stage are converted into frames or sequences of frames. Afterwards, the data are refined through different cleaning filters to enhance the quality of the images that help in effective processing. These enhancement filters contain contrast adjustment and smoothing filters that work on the image edges and corners. Some ideal low-pass and Gaussians filters are also applicable to refine the data. Once the data are purified, the data are ready for processing and training. Before training, the data are converted into phases for validation purposes. There are several ways of cross validation [84] in machine learning, but mainly two cross-validation methods are used in VD, i.e., the holdout method and k-fold method [85]. The holdout method obtains the data and makes their training, testing, and validation sets prior to training. Similarly, in some cases, the k-fold cross validation is used, in which the value of k is selected, such as $k = 10$ or $k = 5$. For instance, selecting $k = 10$ divides the whole data into 10 sets and the k-1 sets are used for training, while the remaining k is used for testing purposes.

3.3 Features Extraction

This section discusses various types of features extraction methods worthy of being included for VD. Features extraction is one of the core steps to obtain meaningful information from large-scale data for processing, which are used to solve several computer vision tasks, such as fire detection, summarizing videos, and activity analysis. Altogether, the learning models work on the extraction of useful characteristics from the data. It is a primary step for multimedia processing. Furthermore, various types of features that are significant, such as edges and corners with motion blobs [86],

Table 7. The Working Flow of the VD methods other than Statistical and Deep Learning Employed along with Evaluation Metrics, Datasets, and Applicability for Surveillance (SU)

SU?	Ref	Main Contributions	Evaluation Metrics	Datasets
Non-Surveillance Methods	[194]	• Different kinds of Bayesian network structures are used, and the system is trained using Media-Eval 2011 Affect.	▪ Media-Eval Cost	• Media-Eval 2011
	[195]	• The proposed method consists of some portion of movies which have violent scenes, and the dataset consisting of 15 Hollywood movies is annotated and used for the method.	▪ MAP	• 2011 benchmark
	[196]	• Detailed description about the Media-Eval 2013 task that is performed for violent scene detection, is provided.	N/A	• Media-Eval 2013
	[197]	• System to differentiate between the real violence and martial arts videos is presented, and the Gestalt law is applied to detect jitter and local interest point.	▪ Precision ▪ Recall ▪ F1-score	• Videos from YouTube and from websites
	[198]	• Violent scenes are detected using Media-Eval 2014 evaluation.	▪ MAP	• Hollywood Movies • Web Videos
	[199]	• The spatiotemporal features are extracted from violent scene sequences and are used for classification. • Next, the extreme acceleration patterns are utilized as main features which are estimated by radon transform.	▪ Accuracy	• Violence in movies • Hockey fight
	[200]	• The 2013 dataset is explained, and the task's rules are defined. • Next, the system is analyzed, and several metrics are used to draw conclusion.	▪ MAP	• 2013 movie
	[201]	• Threats of violence are detected via machine learning.	N/A	• Eight videos from YouTube
	[202]	• Four types of features from acceleration and gyro data are extracted. • Next, an instance-based classifier is applied on those features.	N/A	N/A
	[203]	• An approach based on a natural and robust technique is used to detect fight scenes in surveillance. • These techniques are based on motion analysis.	▪ Precision ▪ Recall ▪ F-measure ▪ Accuracy	• Hockey fight • Action movies
	[180]	• This method extended the IFV that represents a video via both local and spatiotemporal features and used sliding window approach for VD. • Next, the IFV is reformulated where they used area table to speed up.	▪ Accuracy ▪ ROC	• Violent flows • Hockey fight • Movies
	[177]	• The oriented violent flow is proposed to detect violence in videos, and the features fusion and multiclassifier strategies are used.	▪ Accuracy ▪ ROC ▪ AUC	• Hockey fight • Violent flow
	[204]	• Problem of model selection is addressed for the case when there are limited resources, such as central processing unit and graphics processing unit. • Different features are combined considering different specifications.	▪ Accuracy ▪ Time	• Media-Eval VSD 2014
	[205]	• The ViF descriptor is proposed with Horn-Schunk for VD and applied super image resolution after detection to enhance the image quality for the detection of faces for individuals involved in the violence.	▪ Accuracy	• Boss dataset • Violence dataset
	[206]	• CENTRIST-based features are used to identify the violent content from video sequences.	N/A	• Violent flows • Hockey fight

(Continued)

Table 7. Continued

SU?	Ref	Main Contributions	Evaluation Metrics	Datasets
	[207]	• This method developed the real time descriptor to measure the crowd dynamics through crowd texture and temporal summaries of co-occurrence matrix features. • Next, this method introduced a measure of interframe uniformity to demonstrate the appearance of the violent activity.	▪ Accuracy ▪ AUC ▪ ROC	• Violent flows • UCF web abnormality • UMN abnormal crowd
	[208]	• The spatiotemporal features and the motion trajectory are combined to detect violent scenes in videos, and dense sampling is carried out on spatiotemporal volume to extract a differential histogram of optical flow features and deviation of motion features. • The features are fed into SVM to classify the videos.	▪ Accuracy	• CAVIAR
	[209]	• This method collects the videos and generates the violence features through applying motion tracking that slices video frames based on object movement.	▪ Accuracy ▪ Precision ▪ Recall ▪ F-Measure ▪ AUC	• Violence in movies • Violent crowd
	[210]	• This method first collected the statistical models for each person to track the time of different people via GMM and is further improved to detect the crowd. They developed the distribution of magnitude of optical flow to detect the violent scene.	▪ Accuracy	• UMN • PETS
	[211]	• The pedestrian key nodes are added to determine the individual behavior with optical detection method. Next, the response graph is obtained in the image convolution structure, and the pedestrian limbs are represented through key nodes connection.	▪ Accuracy	• COCO
	[212]	• The streakline model is combined with a variational model to obtain features that are used in discriminating between fights and non-fights.	▪ Accuracy	• Violent crowd
	[213]	• A descriptor based on the histogram of oriented tracklets to detect abnormal events in the crowded scene is presented. • A descriptor over long-range motion trajectories, known as tracklets, is developed.	▪ EER	• UCSD • Violence in crowd • UMN
	[214]	• This method used an optical flow algorithm to detect violent scenes in surveillance videos.	N/A	• SDHA 2000
Surveillance-based Methods	[215]	• The violent activity is detected via categorizing the human behavior into individual and group interaction. • The individual behavior is analyzed by the native motion, such as speed and motion direction, while the group behavior is defined by the information about the interaction motion among the neighbors.	▪ Sensitivity	• UMN • BEHAVE • PETS 2009 • UCSD
	[216]	• Descriptor based on substantial derivative is developed.	▪ Average ▪ Accuracy	• Violence in movies • Violence in crowd • BEHAVE • Two surveillance datasets

(Continued)

Table 7. Continued

SU?	Ref	Main Contributions	Evaluation Metrics	Datasets
	[217]	• An IWLD is proposed to depict the information of low-level image appearance and extended the spatial descriptor IWLD by adding temporal component. • Next, the sparse representation based on the classification model is developed to reconstruct the coefficient of coding error and minimize the error of classification.	▪ Accuracy ▪ AUC	• Hockey fight • BEHAVE • Crowd violence
	[218]	• Computer vision techniques are utilized to detect the violent scene in a crowded area in the city center and measured the visual texture that has been proven to be effective in encoding the crowd appearance.	▪ Accuracy	• City center environment • Violent flows • UCF web crowd abnormality • UMN crowd abnormality
	[219]	• The corner joints in an image are detected via Shi-Tomasi corner detection algorithm, and the optical flow is calculated from the moving objects via an improved Lucas-Kanade pyramid. • Next, the violence is detected via the histogram of computed optical flow values.	▪ Accuracy	• CAVIAR • Online movie video clips • Surveillance videos
	[220]	• Motion co-occurrence feature is used as a key motion feature. • Next, the motion vectors are extracted through a block-matching algorithm where the direction and magnitude values are quantized.	▪ Accuracy	• BEHAVE • CAVIAR
	[221]	• A technique called DEARESt is employed that extracts the appearance and motion flow features, which are concatenated into a single features vector.	▪ Accuracy	• UCF-crime

motion speed and direction [87], movement direction and speed [88], spatiotemporal and motion streams [77], motion space and time [79], speed direction and centroid [89], optical flow and motion region [90], motion vector [91], acceleration and movement [92], speed direction and density [93], motion and postures [94], optical flow, and appearance with motion [95]. To extract features, several methods are used for VD. There are two basic types such as traditional features collection and NN-based (deep) features. Each type is explained in the following subsections.

3.3.1 Traditional Features Extraction. Several features descriptors are developed to understand the nature of video frames containing low-level features, such as shape features characterizing certain image information. Low-level features encompass a wide range of techniques, of which key points detection, which finds the points in an image distinct of its surroundings, is the most popular. An ideal key points detector considers two segments, such as robustness and computational speed. Robustness refers to the detection of the same key points in any circumstances or at low video quality. When detecting features, consideration of computational speed is an essential part for optimal performance of a certain task. Some state-of-the-art features extraction techniques include SIFT [96, 97], SURF [98], BRISK [99], ORB [100], FAST [101], BoW [102], and some spatiotemporal features, which can be used for VD methods [103–105]. The SIFT features extracted from video frames do not suffer from certain conditions, such as rotation or changes in scale. SIFT converts the input VD frames into a vector of a large collection of local features, where each vector is invariant to any scale or changes, while SURF is based on principles of the SIFT descriptor. The SURF descriptor [98] is more suitable for processing video frames for VD [106] due to its efficiency. Another feature descriptor known as ORB is an alternative to the SURF and SIFT algorithms. It

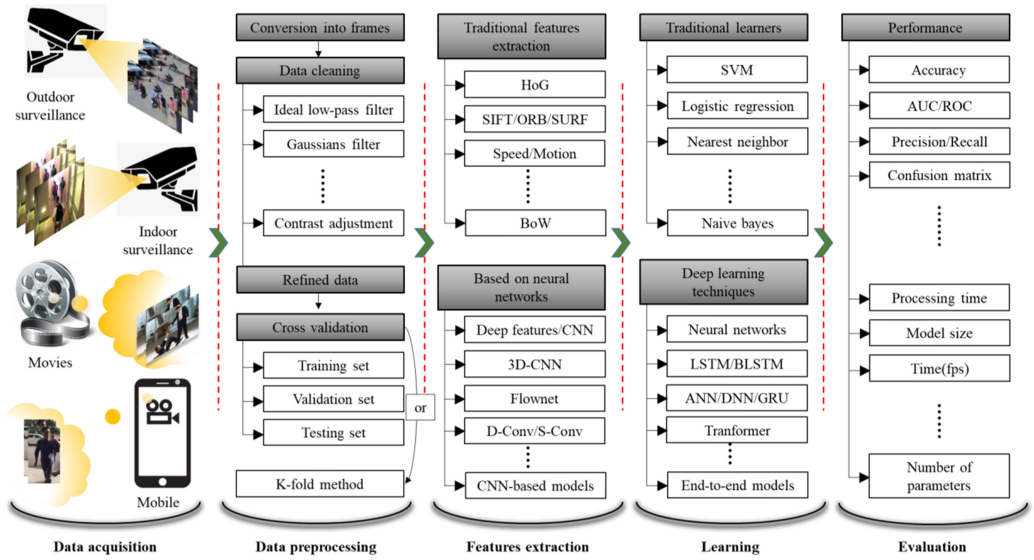

Fig. 4. Basic steps of VD methods. Data acquisition collects data from distinct sources, such as indoor or outdoor surveillance, mobile phones, and non-surveillance cameras. After this step, the data are refined and prepared in the data preprocessing step, which are then fed into the features' extraction step for detailed information collection. D-Conv represents dilated convolution while S-Conv shows sparse convolution. These features are learned through different classifiers or NN-based models to detect violence. After features learning, their performance is computed in the evaluation step using different metrics.

is the combination of FAST [101] and BRISK [99] and has been proven to be invariant in rotation and scale, robust to affine transformation and noise. Furthermore, the BoW [107] represents the VD frames as a single vector that reduces the sensitivity and variability of noise, thus enabling the classifier for training. It builds the vocabulary of the visual words obtained from the frames, i.e., chooses the most representative features for classification purposes. Estimating optical flow calculates the velocities of objects and enables the classifier for training the video sequences. The key point descriptors only determine the area of interest, while optical flow uses multiple video frames and finds the movement difference among them.

$$I_x h + I_y v + It = 0 \tag{1}$$

Equation (1) describes the steps for optical flow calculation where the v and h represent the vertical and horizontal optical flow, respectively, forming the velocity vectors.

To conclude, most traditional features-based methods for VD or their sub-steps, are very complex and are always limited to certain applications. Numerous methods and techniques have been developed that detect the harmful or brutally events patterns [108, 109]. In all these methods, different tactics have been forwarded which work with distinct parameters settings. These parameters are mainly the features or attributes of videos such as the flow, acceleration, appearance, and time, etc. Usually, they divide the whole video into minor segments/frames [89]. In next step, they detect the objects in the frames which vary depending on the detection method. Several researchers developed different methods to boost the VD efficiency, performance, and accuracy. Furthermore, these traditional features extraction methods widely struggle for accurate VD by applying hand-carried engineering tools to achieve their desire goals. In the violent scenes, the motion blobs comprise of particular position and shape. So, these methods first compute the difference between the consecutive frames where the resulted image is binarized that lead into motion blobs number, marking

the largest one at violent and non-violent sequence. Further, it is hard to define the aggression because of no consistency. It also requires high level of interpretation. Detecting the violent scene in low cost and usual manner, mostly these approaches which are based on motion are focused [95]. Some of the methods applied skeleton-based information given by the Kinect camera to recognize the abnormal postures and track the bone joints and their points [94]. For surveillance-based VD, they require to detect the moving object, track it, and understand the behavior of activity.

3.3.2 Deep Learning based Methods. Deep learning based techniques play a vital role in several computer vision tasks, such as activity recognition, disaster management, time series data analysis [163, 164], and multimedia summarization. These techniques have the most popular subfields, owing to a wide range of distributed data representation. Diverse ranges of these algorithms are employed to solve conventional AI tasks, including VD. Similarly, the CNNs, which are trained through large-scale VD datasets and are considered to be momentous approaches [165], are known to play a significant role in VD. The CNNs mainly consist of three kinds of layers, namely convolutional layers, pooling layers, and fully connected layers. The major advantage of convolutions is the extraction of deep features information from frames, the correlation between the pixels, and the object locations. Pooling layers have similar convolutions, but they reduce the feature maps and the parameters of the network [166], where max and average pooling are generally used. Next, 2D convolutional operation is performed in 2D CNN at convolutional layers that extract features with the generation of feature maps. After this, the additive biases are applied, and the results are fed via the activation function described in (2). Usually, the value obtained at a certain position (a, b) in the features map, *j*th with layer *i*th is denoted by U_{ij}^{xy}.

$$U_{ij}^{xy} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{p_i-1}\sum_{q=0}^{q_i-1} w_{ijm}^{pq} u_{(i-1)m}^{(x+p)(y+q)}\right) \tag{2}$$

The 3D convolution is the processing of convolving 3D kernels to compute features from both spatial and temporal dimensions where multiple frames are stacked together. Next, constructing it thus, the obtained feature maps are connected to multiple contiguous frames capturing violent motion information. In this way, the value of position (*x, y, z*) at *i*th layer with *j*th feature map is given in (3) as:

$$U_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} u_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{3}$$

Here R_i shows the 3D kernel size with a temporal dimension, while w_{ijm}^{pqr} indicates the (*p, q, r*)th value of the kernel that is connected to *m*th feature map.

Furthermore, different CNN-based networks are used to train the violence activity sequences such as 3D ResNet [167], FlowNet [168], and ConvLSTM [20]. A typical method to detect violence in videos depends on the domain knowledge that constructs the handcrafted features, while on the other hand, the deep learning models directly act to extract the features. For instance, a method presented in [169] used 3D ConvNets to detect violence with prior knowledge. This network helped in computing the convolution on the video's frame sets, thus the motion information was obtained. They trained the network model in a supervised manner and computed the gradients through a back-propagation method.

Despite this, several other methods have combined the CNN with other deep models such as LSTM to detect the violent scenes in videos. A method presented in [42] utilized acoustic information to detect the violent scenes. Based on that information, the CNN is investigated, which is applied in two ways: as a classifier and as a feature extractor. In another method [20], the

CNN is employed to extract deep features from video frames that are accumulated through several LSTM variants that use the convolutional gates. They combined the CNN with ConvLSTM to make a hybrid network connection that takes the local spatiotemporal features to detect violent frames.

Apart of these networks, dilated convolutions can be the way ahead for VD. These are techniques where the kernel is expanded through the insertion of holes between the consecutive elements. Dilated convolution is same as simple convolution, however, it applies pixel skipping procedure to cover larger input area [170]. An extra parameter such as dilated factor describes how much the input area is expanded. Based on this parameter, the dilated factor minus one, pixels are ignored in the kernel. It helps to expand the input frame area covered without pooling. The main objective is to gain a detailed information from the output at every convolutional operation, i.e., provide a wider view with the same computation [171]. Dilated convolution ensures the aggregation of multi-scale information with no effects on resolution. In past, the dilated convolution used to refer as convolution with dilated filter [172]. Investigation of such network helps to cover a wider range of violent action inside a frame that held at the edge of the frame. Similarly, sparse convolutions are also counselled for their consideration in VD tasks. These convolutions are introduced in [173]. In sparse convolution model, the sparse convolutional layer performs its operation with few kernels followed by sparse multiplication. It is assumed that sparse matrix leads to highly computation, though, the computing sparse multiplication includes severe overhead, making it difficult to attain the acceleration. Furthermore, deformable convolutions are gaining popularity and are applied in different vision tasks [174], showing that their inspiring architecture will boost performance. CNNs are exciting networks for recognition purposes, however, they are limited to model geometric transformation or variations in pose, viewpoint, scale, and part deformation. Tackling such problems, deformable convolutions are introduced [175]. To have different receptive fields and to factor in scale of distinct objects, 2D offset is added to regular grid location. The constant receptive field that are preceding the activation unit are deformed. The added offset is learnable from preceding features via additional convolutional layer [176]. The deformation depends on input features in dense, adaptive, and local manner. The deformable layers add up little number of parameters and computation.

3.4 Learning

Numerous techniques have been used to learn the features extracted from the violent activity sequences. These features are either collected from the still images or frames sequence. The learning procedure consists of two types, such as traditional machine learning and deep learning-based methods. The details of each method are given in the following subsections.

3.4.1 Traditional Learners. In this section, we discuss the traditional machine learning techniques for VD. Once features are extracted, they are fed into classifier for training and learning process. Several classifiers are used, such as SVM [177, 178], KNN [179], etc. The SVM algorithm is used to solve the problem of classifications with a supervised learning process. The SVM plots the data (features) dimension space to differentiate between classes, i.e., violent and non-violent. This classifier is based on a kernel that acts as a function which converts the input into high dimensional space to solve the problem. A method presented in [180] has firstly used IFV as an extension, whereby the local features and positions of spatiotemporal information are used to represent complete video. They applied the sliding window technique to detect violence. Furthermore, the other classifiers have distinct classification procedures that are outside the scope of this survey. Learning techniques include various classifiers, such as SVM [144], Naïve Bayes [120], KNN [181] and decision trees [115].

There are several classification algorithms, but they depend on the data and the applications. For instance, if classes are separable linearly, then a classifier like logistic regression or Fisher's linear will outperform the sophisticated models. Next, SVM is the most popular classifier that is abundantly used, i.e., for violent scenes, the binary SVM is used where the data has two classes, such as violent and non-violent. This classifier finds the best hyperplane which separates the data points of violent class from non-violent data points. The best hyperplane means the one with the largest margin between two classes. Margins are the maximum parallel slab width to hyperplane. Next, the Naïve Bayes [120] is used to distinguish between the violent and non-violent frames. It is a probabilistic classifier which is inspired from the Bayes theorem. Detecting motion and tracking is a complex method for VD. For instance, a method based on motion vectors [182] extracted the motion vectors from the video sequence, where the motion vectors of each frame are analyzed and the region motion vector (RMV) is attained. The authors in [182] used a radial basis to classify RMV using SVM.

Similarly, a technique [89] presented three main stages, namely detection of a moving object, its tracking, and understanding its behavior to recognize the activity. This technique extracted several types of features that include object direction, speed, dimension, and centroid, which help its tracking in video frames. In the final stage, this technique applied a rule-based classification method to categorize the output video into violent or non-violent. It is important to note that in computer vision the Lagrangian theory provides rich tools to analyze the long-term and non-local motion information. Based on this theory, a technique presented in [156] used spatiotemporal features that are based on Lagrangian-directed fields. They used the background motion information, appearance, and long-term motion. BoW was applied in a late fusion manner, ensuring the spatial and temporal scale. They demonstrated that the temporal scale obtained through Lagrangian is crucial for VD and showed how they correlated to the spatial scale of event characteristics.

3.4.2 Deep Learning Techniques. Numerous VD methods have obtained reasonable results using deep learning-based methods, including NNs, LSTM [189], DNN [190], or other end-to-end learning models. The NNs consist of arranged layers with small units called neurons where each unit obtains the input and applies the function and passes the output to next layer. Generally, the networks are defined to be feed-forward, whereby a unit feeds the output into the next layer with no feedback to the previous layer. The weights are applied to pass signals from one unit into another; the same weights are trained to adapt a NN. RNN is introduced to analyze the hidden violence patterns in both spatial and temporal sequential data [191]. As violence data are also video data where the movement in the visual content is represented in several frames, this helps in understanding the context of certain activities. In view of this, RNNs interpret these types of sequences easily, but forget long-term sequential information. This problem is known as vanishing gradient, which can be solved through a special type of RNN known as LSTM [189]. LSTM networks are the RNN extended form that have the ability to find the spatiotemporal patterns in the violence activity data. In this way, they show the loss of the initial dependencies present in the sequence. The LSTM includes mainly three kinds of gates, namely the input gate, output gate, and forget gate. These gates assist in learning the violence sequences. These gates are adjusted by a sigmoid function that learns the opening and closing in training.

Training of large-scale data is challenging due to the presence of complex sequential patterns of violence that cannot be easily identified and learned by a single LSTM. For this purpose, VD researchers used multilayer LSTM networks that are made by stacking multiple LSTMs. In addition, DNN boosts their performance through adding multiple layers to the network. Adding additional layers allows the RNN to capture high-level sequence information [192]. In simple RNN, the data is passed into a single layer and is activated, but in sequential data, the violence sequences are

processed by several layers. Consequently, LSTM preserves only the core information that has been previously processed through a hidden state. For this purpose, bi-directional LSTM can be used to preserve the violence sequence information from both the previous and upcoming hidden states that are more precise for learning the sequential patterns. Therefore, several VD methods exploit the B-LSTM that runs the input in two directions, i.e., from the past into future and from the future into past. In B-LSTM, the output at a certain time is not only dependent on previous frames in the sequence, but also on the upcoming frames [193]. B-LSTM comprises two stacked RNNs, i.e., one goes in a forward direction, while another goes in a backward direction. Their combined output is computed through the hidden state of both RNNs.

3.5 Evaluation Terminologies

In multimedia analysis, the performance evaluation of certain methods is required to boost and improve their accuracy. Several evaluation metrics are used in computer vision that are related to the problems solved by these methods, such as classification, regression, detection, and summarization. For instance, the performance of a regression model can be evaluated through mean square error or root mean square. Similarly, various metrics are used to evaluate the VD methods such as accuracy, precision, recall, AUC values, ROC, the confusion matrix, specificity, and sensitivity. In these metrics, the basic evaluation metric studies are the confusion metric that consists of certain parameters such as TP, TN, FP, and FN. By computing these values, we obtain another metrics information. These metrics are formulized in (4) to (9) where accuracy B is for balance data. Furthermore, the general structure of confusion metrics is given in Table 8. In classification problems, the precision means the number of TP (number of correctly classified) divided by the total elements belonging to the positive class. In other words, the terms TP, TN, FP, and FN compare the classifier results under examination. Furthermore, the AUC-ROC is the performance measurement at several threshold settings. Here, the Receiver Operating Characteristics (ROC) is the curve, while AUC indicates the degree of separability. Some methods [57, 194] compute the false alarm rate known to be False Positive Rate (FPR) that is the number of false alarms per total warnings/alarms in a situation. This metric is formulated in (10).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$AccuracyB = \frac{TPR + TNR}{2} \tag{7}$$

$$Specificity = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{10}$$

Table 8. Confusion Matrix Structure for the Evaluation of VD Method

Conditions	Violent	Non-violent
Violent	True Positive	False Positive
Non-violent	False Negative	True Negative

The metric tells how much the model distinguishes between the violent and non-violent class. A high value for AUC indicates a better distinguishing performance between violent and non-violent. In this case, the ROC is plotted against FPR with TPR, which is on the Y-axis and FPR is on the X-axis. In term of computation cost, several facts are preserved, such as model size, number of parameters, FPs, and training time. Model size has a great impact on certain methods. A model with a small size is preferable, as that is easily deployable on resource-constrained devices to overcome the power consumption and cost.

4 VD DATASETS AND THEIR DISTRIBUTION FOR SURVEILLANCE AND NON-SURVEILLANCE DOMAINS

This section discusses the VD datasets that are widely used in both surveillance and non-surveillance settings. The datasets used for VD are covered from different sources, such as YouTube, real-time CCTV footage, movies, or those recorded with mobile phone. Each setup has different resolutions and video quality that make them more challenging. Covering the datasets literature, we determined that most of the videos are obtained from movies or recorded with mobiles. From our deep analysis, we noted that the most explored datasets in the field of VD are hockey fights [110], violence in movies [110], and violent crowd datasets [151] due to their challenging nature, which can be verified from Tables 5–7. Further, we demonstrate all the VD datasets in Table 9, where details, videos source, and resolution are given. We considered the datasets as surveillance by viewing each video in each dataset. We also added the state-of-the-art accuracy on the datasets along with the corresponding methods. After examination, we found that the RWF-200 dataset needs more improvement in terms of accuracy. Some of the methods are represented with NA, showing that they applied different metrics such as AUC, ROC, Recall, and so on and their inclusion is not applicable due to not using accuracy for evaluation.

There exist numerous challenges and problems with the VD datasets, such as small amounts of data, the video quality, and size [82]. These parameters greatly impact the detection of violent scenes in both surveillance and non-surveillance domains. There are several solutions to these problems that are applied to overcome them. Sometimes the videos are of very low quality, i.e., the resolution is very low. In this case, different smoothing and cleaning filters are applied that enhance the frames' quality and increase the accuracy of recognition. In some cases, the datasets consist of a small amount of video data, but as the deep learning models require a large amount of data for training, several data augmentation techniques [222] are utilized to increase data diversity and assist in training the network models. These techniques include different geometrical operations, such as vertical and horizontal flipping, color variations, and so on. In the surveillance domain, indoor and outdoor surveillance systems exist. As the outdoor surveillance cameras are subject to weather and light conditions, such as smoke, fog, and illumination problems, these parameters have an influence on the detection of violent scenes, as they produce noise and abnormalities in the data. For this purpose, defogging techniques are applied for the detection of violent scenes in uncertain environments. We illustrated sample frames from certain datasets in Figure 5.

Table 9. Details about the Surveillance and Non-Surveillance Datasets including Resolution, Overall Accuracy, and Frame per Second (FPS)

Reference	Total Video Samples	Domain		Frames per Second	Frame Resolution (pixels)	Overall Accuracy (%)	State-of-the-art Accuracy (%)	
		Surveillance	Non-surveillance				Accuracy (%)	Methods
Hockey Fight [110]	1,000	✗	––	25	360 × 288	91.7	99.9	[223]
Violence in Movies [110]	200	✗	––	25	360 × 250	89.5	100	[223]
Behave [224]	4	✓	––	25	640 × 480	93.67	99.3	[225]
Violent Flows[151]	246	––	✓	25	320 × 240	81.30	96	[226]
CAVIAR [227]	––	––	––	25	384 × 288	––	98.86	[228]
VSD [229]	––	––	––	30	––	––	NA	NA
Web Abnormality [230]	20	✓	✓	––	––	––	NA	NA
UMN [230]	––	––	––	––	320 × 240	––	NA	NA
RWF2000 [231]	2,000	✓	––	––	Diverse	86.75	89.25	[187]
Surveillance Camera Fight [232]	300	✓	––	25	480 × 360	72	95.62	[233]
Real-world Fight [17]	1,000	✓	✓	–	Diverse	––	NA	NA
SDHA 2000 [214]	––	✓	––	30	720×480	––	NA	NA
Industrial Surveillance [234]	300	✓	––	––	Diverse	80	80	[234]

Fig. 5. Sample representation of each dataset having two classes: violent/abnormal and non-violent/normal.

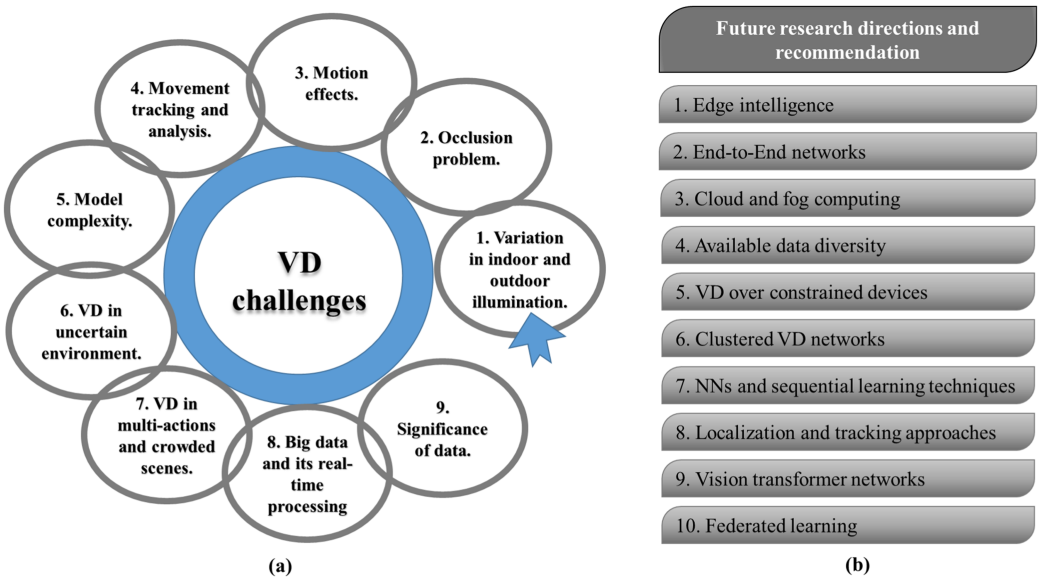

Fig. 6. There exist several challenges and problems faced by researchers in VD that can be solved by future research. Incorporating these recommendations will not only assist VD methods, but will also strengthen the VD arena.

5 COVERAGE OF VD APPLICATION WITH AI-BASED TECHNOLOGIES

VD has a wide range of applications for both surveillance and non-surveillance areas based on AI. We give a thorough overview of its applications [235] in movies, surveillance videos, images, and the rest of non-surveillance setups.

5.1 Quick Response by Reducing Processing Time

In traditional monitoring systems, cameras are continuously operating for 24 hours to check for suspects or abnormal events in surroundings. Using this method, the processing becomes complicated, and much time is spent in monitoring. For this purpose, an automatic and intelligent deep learning-based VD model can reduce the extensive processing time by its deployment in surveillance setups. These VD models can help in automatic detection of violent scenes and reduce the false alarm rates [57, 194]. Next, several object detection techniques are applied prior to the actual VD step. These techniques help to collect the most significant information, such as important object collection including humans and vehicles, as these objects are more subject to violence. Once object detection is performed, only the important frames having the objects details are proceeded and extra frames are avoided from its processing. This procedure helps to reduce the computational cost and time.

5.2 Fast Decision-Making Policy

An automatic VD method has a significant role for decision concerns in departmental and organizational aspects, where the chances of abnormal events are high. The implementation and deployment of an intelligent VD system will assist in improving the decision-making policy before any further catastrophes occur. For this purpose, several VD methods [161, 236] have been developed for surveillance scenes analysis to detect and report violence to the department con-

cerned. Next, movies and animated cartoon contain a high risk of violence with the blood, injuries, guns and horror scenes they contain. These types of violence have a great influence on users, such as underage people or children and even old people. This type of risk needs to be prevented to reduce stress and depression. For this purpose, the detection of violence plays a significant role in this case.

5.3 Security and Safety for Combating Crime

Nowadays, surveillance cameras are installed at every corner in areas like shopping malls, prisons, schools, border points, and entertainment venues [235], which are comprised of both indoor and outdoor access areas. The main aim of these cameras is to assist law enforcement services to prevent crimes and secure the areas through smart and intelligent VD methods. For this purpose, researchers are discovering machine-intelligence-based technologies for VD. Their designed networks are implemented to distinguish violent scenes from non-violent ones. They deploy an alarm-based tool; upon the detection of violent scenes, the alarm is sounded to report the violent activity to the nearest headquarters [237]. Furthermore, several AI-based VD methods [238] have been proposed, which are comprised of gunfire detection and crime spot detection. Law enforcement can reach the gunfire location without calling them or without an officer witnessing the gunfire. In this way, the sensors are installed as a cloud-based application that accurately detects and locates the gunfire and where every sensor captures the gunfire sound and time.

6 CURRENT CHALLENGES AND FUTURE RESEARCH DIRECTIONS

In this section, we discuss the challenges faced by researchers and their possible solutions through future directions. There are numerous challenges and problems in the VD domain that are explained briefly here. These challenges range from different perspectives, such as illumination variations, non-stationary background, occlusion, resolution, motion blur, and the like.

6.1 Current Challenges

The challenges faced by research in the VD arena are given below:

6.1.1 Variations in Indoor and Outdoor Illumination. Detecting violence in videos is highly affected by dynamic illumination, which makes tracking objects and analyzing scenes more difficult. These illuminations result from naturalistic change or lightening effects outdoors and indoors, respectively. Mostly, outdoor surveillance is subject to natural illumination changes, which result in poor contrast while recording at night and make the content description difficult. Furthermore, in an indoor environment, the lighting from the source is dispersed at each edge or gathered on a single point, making the scene more complex. Similarly, light glowing from each corner makes it difficult to detect the event in a live stream. This problem is rarely taken into consideration, although several solutions have been presented for it. For instance, Zhou et al. [157] proposed better tolerance to variation in illumination. They introduced the local histogram of oriented gradient (LHOG). Further, the background subtraction algorithm presented in [239] deals with the light changes and long-term changes in a scene.

6.1.2 Occlusion Problem. Occlusion is another major challenge that occurs when several objects in a scene come closer and merge or combine. This is the most basic problem that cannot be avoided in detecting a violent scene. Several algorithms suffer due occluded objects present in the scene. Various methods have provided a solution to this problem, such as the review presented in [240] that reported an algorithm for occlusion problem and reviewed numerous occlusion algorithms. In some methods, the camera is defined to ensure which part is occluded. In [241], several techniques are proposed, for instance, a minimization procedure, temporal section, sum of squared

distance, and graph cut methods. They opined about the fusion of several cameras that find the depth and estimate the hidden part where the depth is most challenging. They stated that the polynocular stereo algorithm assists in overcoming the occlusion. Sometimes, the multi-view camera setup can overcome this problem if the cameras jointly work and share the outcomes.

6.1.3 Motion Effects. During the scene recognition, the videos are of low resolution and contain background motion that is caused by camera motion, light changes, or rendering noise. The resultant frames obtained during this are blurry and contain noise, which causes poor results for the model. Regarding this problem, an approach [203] using optical flow is adopted to analyze the motion. The method in [194] considered the optical flow as a strong vector and cue to measure the amount of motion and flow direction. Furthermore, a method presented in [206] developed several procedures that include 3×3 Gaussian filters, which reduce noise, and the histogram equalization is applied to distribute pixel intensities into a large range of contrast. The same method further applied a mixture of Gaussians (MoG) that avoids the objects, which are not related to action. The current challenges and future recommended directions are given in Figures 6(a) and 6(b), respectively.

6.1.4 Movement Tracking and Analysis. Surveillance cameras are installed everywhere in both indoor and outdoor environments to keep track of every object, such as humans [242], vehicles [243], and the like. Sometimes in clustered zones or overpopulated areas, the human density is too high, and the VD algorithm failed to ascertain the violent scenes. In such areas, human exchange goods or other stuff to each other, or they shake hands. During this, the tracking of hand movements and body gesture analysis for violence is a challenging task that needs to be intelligently tracked [244]. Existing methods are trained for person-to-person violence datasets, which fail to keep track of each individual movement. For this purpose, such algorithms and methods are suggested to keep track of individual hand gestures of humans and to identify them for occurrences of violence. Algorithms with consistency in tracking are more significant for following the object position in each frame sequence in a complex environment.

6.1.5 Model Complexity. Model complexity is characterized by many factors and is subject to several parameters. It refers to a number of features or the consideration of different terms, such as whether the model is linear or non-linear and so on. Similarly, the complex models are less interpreted and at a greater risk of overfitting, they become computationally expensive. Further, due to the large number of parameters, their size increases and becomes tedious to load over resource-constrained devices, such as mobiles, and Raspberry Pi boards [245]. This problem arises when the complex architectures with large numbers of layers are used. These layers are added to boost the performance of the model, but on the other hand, they increase the model size and make it computationally complex. Therefore, models with small number of parameters are required to made them able for their real-time deployment.

6.1.6 VD in an Uncertain Environment. To keep the assets safer under surveillance and secure the location from the occurrence of different abnormalities, it is important to detect the violence everywhere (certain and uncertain environments), prior to any risk. So far, VD in a certain environment can be performed easily enough where the data from certain environment can be easily processed, but the models are not reliable in uncertain environments, as sometimes the situation in outdoor surveillance is unclear due to different uncertainties, such as rain, smoke, storms, among others [246]. These uncertainties reduce the correct detection in the scene. In some cases, researchers apply the uncertainty to the data for its prediction in uncertain environment. Thus, making such a system functional in all uncertainties is a big challenge in terms of accuracy and computation cost.

6.1.7 VD in Multi-Actions and Crowded Scenes. Surveillance cameras are now installed in every place, such as streets, malls, playgrounds, schools, and airports, capturing multiple scenes for 24 hours/7 Days a week. So far, these scenes cover several activities and actions in the same footage, such as moving of different objects, their interaction, i.e., handshaking, and hugging of the individuals, thus making the system more complex. Moreover, distinct activities by these individuals are included in the same surveillance. Therefore, the detection of violent activities in a complex environment is a big challenge. Developing a method that is able to distinguish among several activities is another challenge.

6.1.8 Big Data and its Real-Time Processing. For continuous monitoring to ensure security and safety, the installed vision sensors capture large amounts of data that result in generating big data. Similarly, for efficient processing, high-quality data (higher resolution) are considered prior to VD via IoT. Thus, processing this huge amount of multimedia data raises another big challenge for its real-time processing in term of saving the dissipated power over both the PC and resource constrained devices.

6.1.9 Significance of Data. As previously discussed, continuous capturing generates big data containing a variety of useful and non-useful information. Processing overall data is difficult and computationally costly. A data mechanism is needed to scrutinize the important content from the data for efficient and smooth processing through discarding the unnecessary data. For instance, this scrutiny can be performed for important object collections, such as humans, vehicles, and so on, as the violence is basically concerned with these objects. Devising a mechanism to deal with this crucial matter, like the preprocessing step performed in [39], which significantly improved the performance of the system, should be a priority.

6.2 Future Research Directions

The detailed investigation and thorough study of all existing VD methods since 2012 has brought to light numerous limitations and drawbacks that need to be tackled and solved in future research, along with some recommendations. We discuss these limitations in the following paragraphs and elaborate upon them in the subsequent sections.

Modern techniques and advancements of deep learning, machine learning and AI tools have been applied in several domains of data science for various applications that are based on big data as input and that generate intelligent outputs. Data collected from various sources are passed into numerous stages of deep learning networks that process the data for certain output results. Such types of training networks are missing in VD literature. Next, several VD methods process the data without preprocessing mechanisms to refine them and make them fruitful to produce reasonable results. Today, the greatly significant role of IoT in different domains has made it possible to reduce the cost and computational resources. Such IoT setups will further enhance intelligent VD technology through the usage of resource-constrained devices such as Raspberry Pi, smart phones, tablets, etc. Consequently, scarcity was found in VD methods that can process the data at the generation stage, i.e., edge computing intelligence and cloud-based intelligent terminologies. IoT is widely in progress in several real-world applications that are connecting plenty of agile devices that share the sensors' data for ease and prompt information collection. Employing such techniques will strengthen the VD technology through machine-intelligent aspects. Surveillance systems have applications for VD in both indoor and outdoor scenes where both scenes are less focused in VD. The most challenging and positive points for researchers are reducing energy bandwidth, processing time, number of parameters, and model size. The recommendation and future research directions are given in Figure 6(b).

6.2.1 Edge Intelligence. Edge intelligence plays a vital role in several AI-based applications. These applications include activity recognition [247] and disaster management [248]. Combining edge computing with machine learning enhances data processing and ensures real-time processing for quick decisions. VD over edge intelligence improves connectivity and security. In this way, the data are processed over the edge where they are acquainted. Using edge intelligence [234], the devices over the network control the data and communication management improves the quality and reduces time delays. For the VD domain, it is a significant concept to detect violent scenes at the edge and make decisions before further catastrophes occur. Several advantages to edge intelligence are to make instantaneous decisions during data collection. Implementing VD over edge intelligence will further enhance the detection of violent scenes. In this view, several devices can be clustered together, forming an IoT-based network comprised of a connected CCTV grid. During this, all the devices are connected to each other and share their information through smart sensors.

6.2.2 End-to-End Friendly Networks. In the field of computer vision, end-to-end network models are widely used for different tasks, such as speech recognition [249] and object detection [41]. An end-to-end network saves time in writing a multiple optimization code for each component in machine learning. Several VD methods pass the input VD data from several networks' layers through convolutions. Other complex networks require huge processing at the initial step as a prerequisite that delays the process and needs further additional processing. Consequently, the same models with post-processing consumed much time for the VD task. In contrast, the end-to-end models take the input data and further propagate it into the hidden layers of the same network and produce an output at the end layer.

6.2.3 Cloud and Fog Computing. Several video analytic systems depend and work on shallow networks that are unable to harness the power of processing in training and inference. Therefore, cloud computing, which is the most effective and fastest approach for VD in which the input data is analyzed and distributed efficiently for fast processing, will overcome this issue. Several existing computer vision tasks, such as summarization and cloud approach [250] over IoT applications [251], are using it. Furthermore, this concept can be implemented to secure the process of VD where the response from the cloud server is obtained promptly. Thus, in reviewing the applications of cloud computing, it can facilitate the process of VD in terms of different aspects such as fast processing, accuracy, and consumption of lower power. Similarly, it is usable in the near future for instant responses. Apart from this, sensor-based VD methods have been addressed in current literature or the proposed solutions utilize a single device to accomplish the data acquisition and recognize the activity [252, 253]. Still, to achieve more demanding tasks like classification of complex activities, the resource constrained devices need a support via solid infrastructure for capturing, processing, managing, and storing the data acquiring from heterogeneous sensors. In such case, cloud computing offers a reasonable solution to transfer the huge computation into cloud storage [254] for its execution [255] while using mobile as a sensing platform. Yet, the advantages obtained through this approach can be negligible in real-time processing when the data is transferred to/from the cloud [256]. The paradigm of fog computing was developed (as extension) of cloud computing on the edge. Nowadays, Fog computing is widely accepted as good alternative to cloud [256] while dealing with largescale data to be locally and timely processed.

6.2.4 Available Data Diversity. There exist numerous VD datasets published by researchers that consist of private data and are not accessible for use by researchers. Such datasets can be made public to further ease the tasks of VD. In an overview of VD literature, the majority of non-surveillance datasets are utilized, whereas the surveillance datasets, such as [17], [231], and [232], are missing. In this way, VD over surveillance setups, which need such benchmarks in order to make them

talented for machine-based intelligent VD over CCTV, remains untouched. Furthermore, the complex datasets with real-world violence are missing for VD. An examination of the literature reveals that the most frequently used datasets are taken from Hollywood and Bollywood videos that are not enough to detect real-world violence. Next, in VD literature, multi-view violence datasets are missing that will further improve the detection process in surveillance. During multi-view VD, the video is captured from different views and each view is analyzed for violence and the detection of suspicious objects. The object is located in each view, and the violence across the views is localized.

6.2.5 VD over Embedded Devices. Nowadays, resource-constrained devices are used in several fields of computer vision, such as summarization and image processing. To execute the VD process, setups such as Raspberry Pi, Arduino, and other programmable constrained devices are used. Further, such methods exist that cluster the processing of these devices to compute and solve their tasks using parallel processing. Moreover, the constrained devices can be set up with the cloud computing to train the network model at the cloud server and obtain the decision of VD at the resource devices. Adding these capabilities to the field of VD will further enhance the VD method. The purpose of these devices is the lower computational cost and processing in real time. These devices are clustered to make a connected network that shares the information, thus making it computationally intelligent and smart. Methods such as the ones reported in [257] evaluate their performance over constrained devices.

6.2.6 Clustered VD Network. Surveillance of every point is streaming continuously for monitoring purposes; however, the present approaches are unable to cover and deploy their models over the multiple cameras and shared network for their smooth operation. Therefore, a clustered network of multiple cameras captures the whole scene that requires an energy-friendly mechanism workable on a shared VD network [258]. This clustered network is the collection of multiple sensors sharing information about a single activity [259] or group activities [260]. The information about the detection of violence at one camera is shared with another in the case when an individual is moving. This type of system is the missing piece in the state of the art. Consequently, the deployment of such a system will automate the execution of the overall process and the on-the-spot response to the event.

6.2.7 NNs and Sequential Learning Techniques. Sequential learning techniques such as RNN, its several variants, such as LSTM, GRU, and so on, have shown promising results in several computer vision and other machine learning tasks, such as activity analysis, video summarization, and energy clouds. However, their usage for VD is very limited in both surveillance and non-surveillance video pattern analysis [261]. Also, their hybrid approaches have given promising results and adding them will further strengthen the VD process. The sequential learning methods aid in learning long-term dependencies in the video sequence. These networks consist of different gates with hidden neurons. These neurons store and memorize the information obtained from the sequence.

6.2.8 Localization and Tracking Approaches. In VD, the main concerns are the humans and the objects, which can lead the activity into violence. As these objects are subject to violence, they continue to keep changing their location in the scenes. However, to keep track of the activity in the scene and its detailed information, such as object location, detection [262, 263], positions [264], and understanding the behavior, we need to localize the activity in the scene and track the objects' movement with respect to their surroundings. Locating and tracking the object will not only provide the object information, but will also provide the event information. The tracking object can be a human or vehicle which is involved in the event [265]. Once the objects are tracked

and located, their detailed information are recorded for the future analysis to prevent the violence in such cases.

6.2.9 Vision Transformer Networks. There emerged the most popular networks such as Transformers that attained the outstanding results compared to mainstream competitive convolutional networks and require extensively fewer resources of computation for training [266]. Instead of prior works, they do not utilize the inductive biases in the architecture, but interpret the image as patches sequence which is processed through Transformer encoder. They are built on inspired self-attention mechanism and they demonstrated notable performance in various vision tasks such as image recognition [267], action recognition [268], and so on. Inclusion of such networks will further assist and boost the VD performance by improving accuracy and reducing the computation. These networks exceed or match state-of-the-art on several image classification data sets and are relatively cheaper when training them [267].

6.2.10 Federated Learning. The detection of violence in computer vision is the recognition of the activity which is the detection of fighting scene. So, the significant data-driven approaches applied for recognition of such activities are widely based on supervised learning. Though, such methods perform better on limited set of activities for which there is labelled dataset. Still it is a challenging task to cope with the inter- and intra-variability of event held among different subjects [269]. In this regard, semi-supervised approaches are proposed to face the challenges of acquiring huge amount of labeled data mandatory for realistic settings. Federated Learning (FL) is the most significant paradigm to counter these problems. In FL, the learning is accomplished by the federation of the devices that participated, Further, it produces training data, instead to centralize the training data on server or cloud [270].

7 CONCLUSION

Nowadays, surveillance cameras are installed everywhere to continuously monitor human's activities. These activities include running, jogging, jumping, clapping, and the like. However, the detection and monitoring of abnormal events, such as violence, aggressive behavior, and the like, is comparatively less focused. In this way, the installation of surveillance cameras and automating them through computer vision and AI techniques will further play a vital role in society. After deep analysis, several guidelines in the form of reviews and surveys have been found to guide the VD community. For this purpose, we comprehensively surveyed the existing methods from 2011 to date with a focus on NNs.

In this survey, we briefly studied the VD methods from different perspectives that are missing in the existing literature. We covered the concept and motivation for VD along with its applications, followed by its division on the basis of surveillance and non-surveillance setups. We also examined the statistics concerning cameras installed in different countries such as South Korea, China, and the United States. Next, we covered their sources of origin, from which we sought the VD methods in distinct type of portals. Afterwards, we thoroughly explained the challenges and problems faced by researchers during VD. Also, we mentioned solutions to tackle and overcome these challenges. Consequently, we highlighted the main contributions of the survey that reflect its actual theme and help readers grasp the actual points of the scenarios covered. We have provided an extensive categorization of NNs for VD. Next, the existing surveys are covered in Section 2, where various related contributions are explained. In addition, we presented the limitations, which can be overcome in the proposed survey. The most significant aspect for VD is the working flow. Section 3 covered the working flow of the VD methods along with the minor and major steps taken from the data collection phase to performance evaluation. The same section mainly consists of data

acquisition, its preprocessing, feature extractions, and learning on the basis of these features. Once training is completed, the methods are evaluated via different metrics, which are explained in the same section. Further, Section 4 comprehensively studied the datasets utilized for VD tasks related to surveillance and non-surveillance, which are summarized in Table 9. Furthermore, after reviewing the VD literature, we found that the existing surveys and other articles have not revealed the applications of VD. Adding the applications will convince more researchers to contribute to VD literature. For this purpose, we added Section 5, which briefly explains VD applications in various domains. At the end, Section 6 incorporated the future directions and recommendations for VD methods, i.e., how to enhance and strengthen them in various domains, ensuring efficient processing in real-time scenarios.

VD is the most rapidly emerging area of computer vision due to its wide range of applications, such as security and safety for public assets in a smart city and on the industrial level through the in-depth study of ANN. Literature on ANN, deep learning, and statistical learning for VD methods and their NN-based categorization delivered in this review have the potential to lead the research community forward. Similarly, the unprecedented platform given in this overview for future research recommendations supported by AI-based smart techniques will enable the research methods. We believe that this survey will be the role model that will gain support among the research community. It will be a vital source for researchers in VD domain.

REFERENCES

[1] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu. 2021. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)* 54 (2021), 1–40.

[2] S. B. Atitallah, M. Driss, W. Boulila, and H. B. Ghézala. 2020. Leveraging deep learning and IoT big data analytics to support the smart cities development: Review and future directions. *Computer Science Review* 38 (2020), 100303.

[3] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel. 2021. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)* 54 (2021), 1–38.

[4] V. Sharma and R. N. Mir. 2020. A comprehensive and systematic look up into deep learning based object detection techniques: A review. *Computer Science Review.* 38 (2020), 100301.

[5] D. Singh and C. K. Mohan. 2017. Graph formulation of video activities for abnormal activity recognition. *Pattern Recognition* 65 (2017), 265–272.

[6] P. Wu, J. Liu, and F. Shen. 2019. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE Transactions on Neural Networks and Learning Systems.*

[7] X. Jiang, L. Zhang, P. Lv, Y. Guo, R. Zhu, Y. Li, Y Pang, X Li, B Zhou, and M Xu. 2019. Learning multi-level density maps for crowd counting. *IEEE Transactions on Neural Networks and Learning Systems* 31, 8 (2019), 2705–15.

[8] T. Ergen and S. S. Kozat. 2019. Unsupervised anomaly detection with lstm neural networks. *IEEE Transactions on Neural Networks and Learning Systems.*

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

[10] E. Epaillard and N. Bouguila. 2018. Variational Bayesian learning of generalized Dirichlet-based hidden Markov models applied to unusual events detection. *IEEE Transactions on Neural Networks and Learning Systems* 30 (2018), 1034–1047.

[11] O. Isupova, D. Kuzin, and L. Mihaylova. 2018. Learning methods for dynamic topic modeling in automated behavior analysis. *IEEE Transactions on Neural Networks and Learning Systems* 29 (2018), 3980–3993.

[12] W. So. 2018. Perceived and actual leading causes of death through interpersonal violence in South Korea as of 2018. https://www.statista.com/statistics/953168/south-korea-perceived-and-actual-leading-causes-of-violent-death/.

[13] L. Yoon. 2022. Number of violent crime arrests in Seoul South Korea 2020, https://www.statista.com/statistics/1290949/south-korea-number-of-violent-crime-arrests-in-seoul-by-type/.

[14] H. Ward. 2020. Violent crime statistics in the U.S. https://www.statista.com/topics/1750/violent-crime-in-the-us/.

[15] C. Textor. 2020. Number of crimes committed in China between 2009 and 2019. https://www.statista.com/statistics/224778/number-of-crimes-in-china/.

[16] sayibet. 2019. Vision-based fight detection from surveillance cameras. https://github.com/sayibet/fight-detection-surv-dataset.

[17] M. Perez, A. C. Kot, and A. Rocha. 2019. Detection of real-world fights in surveillance videos. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*, 2662–2666.

[18] M. Asad, Z. Yang, Z. Khan, J. Yang, and X. He. 2019. Feature fusion based deep spatiotemporal model for violence detection in videos. In *International Conference on Neural Information Processing*, 405–417.

[19] J. Henrio and T. Nakashima. 2018. Anomaly detection in videos recorded by drones in a surveillance context. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2503–2508.

[20] S. Sudhakaran and O. Lanz. 2017. 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)*, 1–6.

[21] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis. 2018. Bidirectional convolutional LSTM for the detection of violence in videos. In *European Conference on Computer Vision (ECCV)*, 0–0.

[22] E. Ditsanthia, L. Pipanmaekaporn, and S. Kamonsantiroj. 2018. Video representation learning for CCTV-Based violence detection. In *2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, (2018), 1–5.

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

[24] S. A. Sumon, R. Goni, N. B. Hashem, T. Shahria, and R. M. Rahman. 2020. Violence detection by pretrained modules with different deep learning approaches. *Vietnam Journal of Computer Science* 7 (2020), 19–40.

[25] Z. Dong, J. Qin, and Y. Wang. 2016. Multi-stream deep networks for person to person violence detection in videos. In *Chinese Conference on Pattern Recognition*, (2016), 517–531.

[26] S. Vosta and K.-C. Yow. 2022. A CNN-RNN combined structure for real-world violence detection in surveillance cameras. *Applied Sciences* 12 (2022), 1021.

[27] M. Haque, S. Afsha, and H. Nyeem. 2022. Developing BrutNet: A new deep CNN model with GRU for realtime violence detection. In *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, (2022), 390–395.

[28] F. Eyben, F. Weninger, S. Squartini, and B. Schuller. Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, (2013), 483–487.

[29] A. Traoré and M. A. Akhloufi. 2020. Violence detection in videos using deep recurrent and convolutional neural networks. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, (2020), 154–159.

[30] R. Choudhary and A. Solanki. 2022. Violence detection in videos using transfer learning and LSTM. In *Advances in Data Computing, Communication and Security*, (ed.). Springer, (2022), 51–62.

[31] Q. Xu, J. See, and W. Lin. 2019. Localization guided fight action detection in surveillance videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, (2019), 568–573.

[32] Y. Lyu and Y. Yang. 2015. Violence detection algorithm based on local spatio-temporal features and optical flow. In *2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration*, (2015), 307–311.

[33] A. Keçeli and A. Kaya. 2017. Violent activity detection with transfer learning method. *Electronics Letters* 53 (2017), 1047–1048.

[34] Z. Meng, J. Yuan, and Z. Li. 2017. Trajectory-pooled deep convolutional networks for violence detection in videos. In *International Conference on Computer Vision Systems*, (2017), 437–447.

[35] S. Mukherjee, R. Saini, P. Kumar, P. P. Roy, D. P. Dogra, and B.-G. Kim. 2017. Fight detection in hockey videos using deep network. *Journal of Multimedia Information System* 4 (2017), 225–232.

[36] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor. 2017. 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), 1–7.

[37] Y. Fan, G. Wen, D. Li, S. Qiu, and M. D. Levine. 2018. Early event detection based on dynamic images of surveillance videos. *Journal of Visual Communication and Image Representation* 51 (2018), 70–75.

[38] G. Singh, A. Khosla, and R. Kapoor. 2019. Salient region guided deep network for violence detection in surveillance systems. *Journal of Computer Technology & Applications* 10 (2019), 19–28.

[39] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik. 2019. Violence detection using spatiotemporal features with 3D convolutional neural network. *Sensors* 19 (2019), 2472.

[40] S. A. Carneiro, G. P. da Silva, S. J. F. Guimaraes, and H. Pedrini. 2019. Fight detection in video sequences based on multi-stream convolutional neural networks. In *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, (2019), 8–15.

[41] C. Li, L. Zhu, D. Zhu, J. Chen, Z. Pan, X. Li, and B Wang. 2018. End-to-end multiplayer violence detection based on deep 3D CNN. In *2018 VII International Conference on Network, Communication and Computing*. 227–230.

[42] G. Mu, H. Cao, and Q. Jin. 2016. Violent scene detection using convolutional neural networks and deep audio features. In *Chinese Conference on Pattern Recognition*, (2016), pp. 451–463.

[43] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno. 2018. Fight recognition in video using hough forests and 2D convolutional neural network. *IEEE Transactions on Image Processing* 27 (2018), 4787–4797.
[44] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia. 2014. Violence detection in video by using 3D convolutional neural networks. In *International Symposium on Visual Computing*, (2014), 551–558.
[45] B. Jiang, F. Xu, W. Tu, and C. Yang. 2019. Channel-wise attention in 3D convolutional networks for violence detection. In *2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA)*, 59–64.
[46] A. Jain and D. K. Vishwakarma. 2020. Deep NeuralNet for violence detection using motion features from dynamic images. In *2020 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT)*, (2020), 826–831.
[47] B. Peixoto, B. Lavi, J. P. P. Martin, S. Avila, Z. Dias, and A. Rocha. 2019. Toward subjective violence detection in videos. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*, 8276–8280.
[48] B. Peixoto, B. Lavi, P. Bestagini, Z. Dias, and A. Rocha. 2020. Multimodal violence detection in videos. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'20)*, 2957–2961.
[49] J. Li, X. Jiang, T. Sun, and K. Xu. 2019. Efficient violence detection using 3D convolutional neural networks. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, (2019), 1–8.
[50] D. G. C. Roman and G. C. Chávez. 2020. Violence detection and localization in surveillance video. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, (2020), 248–255.
[51] S. Mondal, S. Pal, S. K. Saha, and B. Chanda. 2017. Violent/non-violent video classification based on deep neural network. In *2017 9th International Conference on Advances in Pattern Recognition (ICAPR)*, (2017), 1–6.
[52] M. Mahmood, A. Jalal, and M. Sidduqi. 2018. Robust spatio-temporal features for human interaction recognition via artificial neural network. In *2018 International Conference on Frontiers of Information Technology (FIT)*, (2018), 218–223.
[53] A. Jalal, M. Mahmood, and A. S. Hasan. 2019. Multi-features descriptors for human activity tracking and recognition in indoor-outdoor environments. In *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, (2019), 371–376.
[54] Z. Zhou, M. Zhu, and K. Yahya. 2017. Violence behavior detection based on 3D-CNN. *Computer Systems & Applications* 12 (2017), 034.
[55] K. M. Yew. 2019. Violent scene detection in videos. Universiti Tunku Abdul Rahman (2019).
[56] Y. Zhao, W. W. Fok, and C. Chan. 2019. Video-based violence detection by human action analysis with neural network. In *2019 International Conference on Image and Video Processing, and Artificial Intelligence*, (2019), 113212N.
[57] M. Baba, V. Gui, C. Cernazanu, and D. Pescaru. 2019. A sensor network approach for violence detection in smart cities using deep learning. *Sensors* 19 (2019), 1676.
[58] P. Wang, P. Wang, and E. Fan. 2021. Violence detection and face recognition based on deep learning. *Pattern Recognition Letters* 142 (2021), 20–24.
[59] T. Hussain, A. Iqbal, B. Yang, and A. Hussain. 2022. Real time violence detection in surveillance videos using convolutional neural networks. *Multimedia Tools and Applications,* 1–23.
[60] S. Abdul-Rahman, Y. Mahmud, and M. Nasrullah. 2022. Violence recognition using convolutional neural networks. In *Computational Intelligence in Machine Learning*, (ed.). Springer, 81–94.
[61] J. Mahmoodi, H. Nezamabadi-pour, and D. Abbasi-Moghadam. 2022. Violence detection in videos using interest frame extraction and 3D convolutional neural network. *Multimedia Tools and Applications,* 1–17.
[62] S. M. Mohtavipour, M. Saeidi, and A. Arabsorkhi. 2022. A multi-stream CNN for deep violence detection in video sequences using handcrafted features. *The Visual Computer* 38 (2022), 2057–2072.
[63] J. Selvaraj and J. Anuradha. 2022. Violence detection in video footages using I3D ConvNet. In *Innovations in Computational Intelligence and Computer Vision*, (ed.). Springer, 63–75.
[64] M. Sharma and R. Baghel. 2020. Video surveillance for violence detection using deep learning. In *Advances in Data Science and Management*, (ed.). Springer, 411–420.
[65] A.-M. R. Abdali and R. F. Al-Tuma. 2019. Robust real-time violence detection in video using CNN and LSTM. In *2019 2nd Scientific Conference of Computer Sciences (SCCS)*, (2019), 104–108.
[66] N. Convertini, V. Dentamaro, D. Impedovo, G. Pirlo, and L. Sarcinella. 2020. A controlled benchmark of video violence detection techniques. *Information* 11 (2020), 321.
[67] R. Halder and R. Chatterjee. 2020. CNN-BiLSTM model for violence detection in smart surveillance. *SN Computer Science* 1 (2020), 1–9.
[68] S. Albawi, T. A. Mohammed, and S. Al-Zawi. 2017. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, 1–6.
[69] W. Zaremba, I. Sutskever, and O. Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329,* 2014.
[70] S. Lange and M. Riedmiller. 2010. Deep auto-encoder neural networks in reinforcement learning. In *2010 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

[71] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20 (2008), 61–80.
[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al. 2020. Generative adversarial networks. *Communications of the ACM* 63 (2020), 139–144.
[73] A. Jain and D. K. Vishwakarma. 2020. State-of-the-art Violence Detection using ConvNets, in *2020 International Conference on Communication and Signal Processing (ICCSP)*, (2020), 0813–0817.
[74] Statista. 2020. Number of installed closed-circuit television (CCTV) cameras in public places in South Korea from 2013 to 2019. https://www.statista.com/statistics/651509/south-korea-cctv-cameras/.
[75] 2020. Number of surveillance cameras installed in public and private areas of China in 2017 with a projection for 2020, https://www.statista.com/statistics/879198/china-number-of-installed-surveillance-cameras/.
[76] 2019. The U.S. Has More Surveillance Cameras per Person than China, New Study Shows, https://www.inverse.com/article/61552-united-states-china-surveillance-cameras@:~:text=The%20United%20States%20has%20roughly,to%20China's%20roughly%20200%20million.
[77] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, and M Ilyas. 2019. A review on state-of-the-art violence detection techniques. *IEEE Access* 7 (2019), 107560–107575.
[78] W. Lejmi, A. B. Khalifa, and M. A. Mahjoub. 2019. Challenges and methods of violence detection in surveillance video: A survey. In *International Conference on Computer Analysis of Images and Patterns*, 62–73.
[79] A. B. Mabrouk and E. Zagrouba. 2018. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications* 91 (2018), 480–491.
[80] A. Ahir, P. K. Pateriya, D. Kaur, and M. K. Rai. 2018. A review on abnormal activity detection methods. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 521–526.
[81] G. Tripathi, K. Singh, and D. K. Vishwakarma. 2020. Violence recognition using convolutional neural network: A survey. *Journal of Intelligent & Fuzzy Systems* 39 (2020), 7931–7952.
[82] F. U. M. Ullah, M. S. Obaidat, K. Muhammad, A. Ullah, S. W. Baik, F. Cuzzolin, J. J. P. C. Rodrigues, and V. H. C. D. Albuquerque, 2021. An intelligent system for complex violence pattern analysis and detection. *International Journal of Intelligent Systems*. July 2021.
[83] F. U. M. Ullah, K. Muhammad, I. U. Haq, N. Khan, A. A. Heidari, S. W. Baik, and V. H. C. d Albuquerque. 2021. AI assisted edge vision for violence detection in IoT based industrial surveillance networks. *IEEE Transactions on Industrial Informatics* 18, 8 (2021), 5359–5370.
[84] P. Refaeilzadeh, L. Tang, and H. Liu. 2009. Cross-validation. *Encyclopedia of Database Systems* 5 (2009), 532–538.
[85] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella. 2012. The 'K' in K-fold cross validation. In *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 441–446.
[86] W. Lejmi, A. B. Khalifa, and M. A. Mahjoub. 2017. Fusion strategies for recognition of violence actions. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, (2017), 178–183.
[87] P. C. Ribeiro, R. Audigier, and Q. C. Pham. 2016. RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance. *Computer Vision and Image Understanding* 144 (2016), 121–143.
[88] D. Song, C. Kim, and S.-K. Park. 2018. A multi-temporal framework for high-level activity analysis: Violent event detection in visual surveillance. *Information Sciences* 447 (2018), 83–103.
[89] S. Chaudhary, M. A. Khan, and C. Bhatnagar. 2018. Multiple anomalous activity detection in videos. *Procedia Computer Science* 125 (2018), 336–345.
[90] M. Al-Nawashi, O. M. Al-Hazaimeh, and M. Saraee. 2017. A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments. *Neural Computing and Applications* 28 (2017), 565–572.
[91] D. Maniry, E. Acar, F. Hopfgartner, and S. Albayrak. 2014. A visualization tool for violent scenes detection. In *International Conference on Multimedia Retrieval*, 522–523.
[92] E. Y. Fu, M. X. Huang, H. V. Leong, and G. Ngai. 2018. Cross-species learning: A low-cost approach to learning human fight from animal fight. In *26th ACM International Conference on Multimedia*, 320–327.
[93] M. Alvar, A. Torsello, A. Sanchez-Miralles, and J. M. Armingol. 2014. Abnormal behavior detection using dominant sets. *Machine Vision and Applications* 25 (2014), 1351–1368.
[94] R. Nar, A. Singal, and P. Kumar. 2016. Abnormal activity detection for bank ATM surveillance. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2042–2046.
[95] E. Y. Fu, H. V. Leong, G. Ngai, and S. C. Chan. 2017. Automatic fight detection in surveillance videos. *International Journal of Pervasive Computing and Communications*.
[96] H. Zhou, Y. Yuan, and C. Shi. 2009. Object tracking using SIFT features and mean shift. *Computer Vision and Image Understanding* 113 (2009), 345–352.
[97] D. G. Lowe. 1999. Object recognition from local scale-invariant features. In *7th IEEE International Conference on Computer Vision*, 1150–1157.

[98] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding* 110 (2008), 346–359.
[99] S. Leutenegger, M. Chli, and R. Y. Siegwart. 2011. BRISK: Binary robust invariant scalable keypoints. In *2011 International Conference on Computer Vision*, 2548–2555.
[100] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, 2564–2571.
[101] E. Rosten and T. Drummond. 2006. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, 430–443.
[102] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. In *International Workshop on Multimedia Information Retrieval*, 197–206.
[103] I. Laptev. 2005. On space-time interest points. *International Journal of Computer Vision* 64 (2005), 107–123.
[104] F. D. De Souza, G. C. Chavez, E. A. do Valle Jr, and A. d. A. Araújo. 2010. Violence detection in video using spatio-temporal features. In *2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, 224–230.
[105] M.-Y. Chen and A. Hauptmann. 2009. Mosift: Recognizing human actions in surveillance videos, Technical Report. Carnegie Mellon University, Pittsburgh, PA, USA, 2009.
[106] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, et al. 2017. Temporal robust features for violence detection. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 391–399.
[107] T. Zhang, W. Jia, B. Yang, J. Yang, X. He, and Z. Zheng. 2017. MoWLD: A robust motion image descriptor for violence detection. *Multimedia Tools and Applications* 76 (2017), 1419–1438.
[108] P. Zhou, Q. Ding, H. Luo, and X. Hou. 2017. Violent interaction detection in video based on deep learning. In *Journal of Physics: Conference Series*, 012044.
[109] C. Dhiman and D. K. Vishwakarma. 2019. A review of state-of-the-art techniques for abnormal human activity recognition. *Engineering Applications of Artificial Intelligence* 77 (2019), 21–45.
[110] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar. 2011. Violence detection in video using computer vision techniques. In *International Conference on Computer Analysis of Images and Patterns*, 332–339.
[111] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su. 2011. Violence detection in movies. In *2011 8th International Conference Computer Graphics, Imaging and Visualization*, 119–124.
[112] G. Gninkoun and M. Soleymani. 2011. Automatic violence scenes detection: A multi-modal approach. In *Working Notes Proceedings of the MediaEval 2011 Workshop*.
[113] Y. Chen, L. Zhang, B. Lin, Y. Xu, and X. Ren. 2011. Fighting detection based on optical flow context histogram. In *2011 2nd International Conference on Innovations in Bio-inspired Computing and Applications*, 95–98.
[114] E. Acar, S. Spiegel, S. Albayrak, and D. Labor. 2011. MediaEval 2011 affect Task: Violent scene detection combining audio and visual features with SVM. In *MediaEval*.
[115] D. Wang, Z. Zhang, W. Wang, L. Wang, and T. Tan. 2012. Baseline results for violence detection in still images. In *2012 IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance*, 54–57.
[116] Y. Lee, K. Kim, D. K. Han, and H. Ko. 2012. Acoustic and visual signal based violence detection system for indoor security application. In *2012 IEEE International Conference on Consumer Electronics (ICCE)*, 737–738.
[117] K. Wang, Z. Zhang, and L. Wang. 2012. Violence video detection by discriminative slow feature analysis. In *Chinese Conference on Pattern Recognition*, 137–144.
[118] J. Schlüter, B. Ionescu, I. Mironica, and M. Schedl. 2012. ARF@ MediaEval 2012: An uninformed approach to violence detection in Hollywood movies. In *MediaEval*.
[119] B. Zhou, X. Wang, and X. Tang. 2012. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2871–2878.
[120] B. Ionescu, J. Schlüter, I. Mironica, and M. Schedl. 2013. A naive mid-level concept-based fusion approach to violence detection in Hollywood movies. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, 215–222.
[121] F. Eyben, F. Weninger, N. Lehment, B. Schuller, and G. Rigoll. 2013. Affective video retrieval: Violence detection in Hollywood movies by large-scale segmental feature extraction. *PloS one* 8 (2013), e78506.
[122] E. Acar, F. Hopfgartner, and S. Albayrak. 2013. Violence detection in Hollywood movies by the fusion of visual and mid-level audio cues. In *21st ACM International Conference on Multimedia*, 717–720.
[123] I. Serrano, O. Déniz, and G. B. García. 2013. VISILAB at MediaEval 2013: Fight Detection. In *MediaEval*.
[124] C. C. Tan and C.-W. Ngo. 2013. The Vireo Team at MediaEval 2013: Violent Scenes Detection by Mid-level Concepts Learnt from Youtube. In *MediaEval*.
[125] E. Acar, F. Hopfgartner, and S. Albayrak. 2013. Detecting violent content in Hollywood movies by mid-level audio representations. In *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 73–78.
[126] P. Rota, N. Conci, N. Sebe, and J. M. Rehg. 2015. Real-life violent social interaction detection. In *2015 IEEE International Conference on Image Processing (ICIP)*, 3456–3460.

[127] E. Acar, M. Irrgang, D. Maniry, and F. Hopfgartner. 2015. Detecting violent content in hollywood movies and user-generated videos. In *Smart Information Systems*, (ed.). Springer, (2015), 291–314.
[128] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T.-K. Kim. 2015. Fast fight detection. *PloS one* 10 (2015).
[129] V. Lam, D.-D. Le, S. Phan, S. i. Satoh, D. A. Duong, and T. D. Ngo. 2013. Evaluation of low-level features for detecting violent scenes in videos. In *2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, 213–218.
[130] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao. 2014. Violent video detection based on MoSIFT feature and sparse coding. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3538–3542.
[131] J.-F. Huang and S.-L. Chen. 2014. Detection of violent crowd behavior based on statistical characteristics of the optical flow. In *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 565–569.
[132] N. Derbas and G. Quénot. 2014. Joint audio-visual words for violent scenes detection in movies. In *International Conference on Multimedia Retrieval*, 483–486.
[133] J. Hu, X. Qi, and J. F. Chen. 2014. Fights behavior detection based on space-time interest points. In *Applied Mechanics and Materials*, 659–663.
[134] T. Senst, V. Eiselein, and T. Sikora. 2015. A local feature based on Lagrangian measures for violent video classification. *Proc. 6th IET Int. Conf. Imag.* Crime Detection Prevention, 1–6.
[135] M. R. Khokher, A. Bouzerdoum, and S. L. Phung. 2015. Violent scene detection using a super descriptor tensor decomposition. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–8.
[136] H. Mousavi, M. Nabi, H. K. Galoogahi, A. Perina, and V. Murino. 2015. Abnormality detection with improved histogram of oriented tracklets. In *International Conference on Image Analysis and Processing*, 722–732.
[137] V. M. Arceda, K. F. Fabián, and J. C. Gutíerrez. 2016. Real time violence detection in video. In *IET Conference Proceedings, no. 1, Talca, Chile.* 6–7.
[138] A. B. Mabrouk and E. Zagrouba. 2017. Spatio-temporal feature using optical flow based distribution for violence detection. *Pattern Recognition Letters* 92 (2017), 62–67.
[139] M. Khan, M. A. Tahir, and Z. Ahmed. 2018. Detection of violent content in cartoon videos using multimedia content detection techniques. In *2018 IEEE 21st International Multi-Topic Conference (INMIC)*, 1–5.
[140] P. Vashistha, C. Bhatnagar, and M. A. Khan. 2018. An architecture to identify violence in video surveillance system using ViF and LBP. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)*, 1–6.
[141] S. Amraee, A. Vafaei, K. Jamshidi, and P. Adibi. 2018. Abnormal event detection in crowded scenes using one-class SVM. *Signal, Image and Video Processing* 12 (2018), 1115–1123.
[142] J. Mahmoodi and A. Salajeghe. 2019. A classification method based on optical flow for violence detection. *Expert Systems with Applications* 127 (2019), 121–127.
[143] I. Febin, K. Jayasree, and P. T. Joy. 2019. Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. *Pattern Analysis and Applications,* 1–13.
[144] J. Yu, W. Song, G. Zhou, and J.-J. Hou. 2019. Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation. *Multimedia Tools and Applications* 78 (2019), 8497–8512.
[145] K. Deepak, L. Vignesh, G. Srivathsan, S. Roshan, and S. Chandrakala. 2020. Statistical features-based violence detection in surveillance videos. In *Cognitive Informatics and Soft Computing*, (ed.). Springer, 197–203.
[146] Q. Xia, P. Zhang, J. Wang, M. Tian, and C. Fei. 2018. Real time violence detection based on deep spatio-temporal features. In *Chinese Conference on Biometric Recognition*, 157–165.
[147] W.-F. Pang, Q.-H. He, Y.-J. Hu, and Y.-X. Li. 2021. Violence detection in videos based on fusing visual and audio information. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'21)*, 2260–2264.
[148] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni. 2021. Deep learning for automatic violence detection: Tests on the AIRTLab dataset. *IEEE Access* 9 (2021), 160580–160595.
[149] B. Solmaz, B. E. Moore, and M. Shah. 2012. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (2012), 2064–2070.
[150] A. Wiliem, V. Madasu, W. Boles, and P. Yarlagadda. 2012. A suspicious behaviour detection using a context space model for smart surveillance systems. *Computer Vision and Image Understanding* 116 (2012), 194–209.
[151] T. Hassner, Y. Itcher, and O. Kliper-Gross. 2012. Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1–6.
[152] M. Elhamod and M. D. Levine. 2012. Real-time semantics-based detection of suspicious activities in public spaces. In *2012 9th Conference on Computer and Robot Vision*, 268–275.
[153] Y. Cong, J. Yuan, and J. Liu. 2013. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition* 46 (2013), 1851–1864.

[154] Y. Cong, J. Yuan, and Y. Tang. 2013. Video anomaly search in crowded scenes via spatio-temporal motion context. *IEEE Transactions on Information Forensics and Security* 8 (2013), 1590–1599.

[155] Z. Yang, T. Zhang, J. Yang, Q. Wu, L. Bai, and L. Yao. 2013. Violence detection based on histogram of optical flow orientation. In *6th International Conference on Machine Vision (ICMV'13)*, 906718.

[156] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora. 2017. Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation. *IEEE Transactions on Information Forensics and Security* 12 (2017), 2945–2956.

[157] P. Zhou, Q. Ding, H. Luo, and X. Hou. 2018. Violence detection in surveillance video using low-level features. *PLoS one* 13 (2018).

[158] T. Zhang, W. Jia, C. Gong, J. Sun, and X. Song. 2018. Semi-supervised dictionary learning via local sparse constraints for violence detection. *Pattern Recognition Letters* 107 (2018), 98–104.

[159] I. Serrano, O. Deniz, G. Bueno, G. Garcia-Hernando, and T.-K. Kim. 2018. Spatio-temporal elastic cuboid trajectories for efficient fight recognition using Hough forests. *Machine Vision and Applications* 29 (2018), 207–217.

[160] P. K. Roy and H. Om. 2018. Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos. In *Advances in Soft Computing and Machine Learning in Image Processing*, (ed.). Springer, 277–294.

[161] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He. 2016. A new method for violence detection in surveillance scenes. *Multimedia Tools and Applications* 75 (2016), 7327–7349.

[162] B. M. Peixoto, B. Lavi, Z. Dias, and A. Rocha. 2021. Harnessing high-level concepts, visual, and auditory features for violence detection in videos. *Journal of Visual Communication and Image Representation* 78 (2021), 103174.

[163] F. U. M. Ullah, N. Khan, T. Hussain, M. Y. Lee, and S. W. Baik. 2021. Diving deep into short-term electricity load forecasting: Comparative analysis and a novel framework. *Mathematics* 9 (2021), 611.

[164] N. Kumar, A. V. Vasilakos, and J. J. Rodrigues. 2017. A multi-tenant cloud-based DC nano grid for self-sustained smart buildings in smart cities. *IEEE Communications Magazine* 55 (2017), 14–21.

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

[166] R. K. Sinha, R. Pandey, and R. Pattnaik. 2018. Deep learning for computer vision tasks: A review. *arXiv preprint arXiv:1804.03928,* (2018).

[167] S. Dubey, A. Boragule, and M. Jeon. 2020. 3D ResNet with Ranking Loss Function for Abnormal Activity Detection in Videos. *arXiv preprint arXiv:2002.01132,* (2020).

[168] A. S. Saif, M. A. S. Khan, A. M. Hadi, R. P. Karmoker, and J. J. Gomes. 2019. Aggressive action estimation: A comprehensive review on neural network based human segmentation and action recognition. *International Journal of Education and Management Engineering* 9 (2019), 9.

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

[170] C. Schmidt, A. Athar, S. Mahadevan, and B. Leibe.2022. D2Conv3D: Dynamic dilated convolutions for object segmentation in videos. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 1200–1209.

[171] Y. Li, X. Zhang, and D. Chen. 2018. CSRNET: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1091–1100.

[172] F. Yu and V. Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122,* (2015).

[173] B. Graham. 2014. Spatially-sparse convolutional neural networks. *arXiv preprint arXiv:1409.6070,* (2014).

[174] J. Li, X. Liu, M. Zhang, and D. Wang. 2020. Spatio-temporal deformable 3D convnets with attention for action recognition. *Pattern Recognition* 98 (2020), 107037.

[175] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, et al. 2017. Deformable convolutional networks. In *IEEE International Conference on Computer Vision*, 764–773.

[176] X. Zhu, H. Hu, S. Lin, and J. Dai. 2019. Deformable convnets v2: More deformable, better results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9308–9316.

[177] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu. 2016. Violence detection using oriented violent flows. *Image and Vision Computing* 48 (2016), 37–41.

[178] C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (1995), 273–297.

[179] J. M. Keller, M. R. Gray, and J. A. Givens. 1985. A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics,* 580–585.

[180] P. Bilinski and F. Bremond. 2016. Human violence recognition and detection in surveillance videos. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 30–36.

[181] M. Alhammami, C. P. Ooi, and W.-H. Tan. 2015. Violence recognition using harmonic mean of distances and relational velocity with *K*-Nearest neighbour classifier. In *International Visual Informatics Conference*, 132–139.
[182] J. Xie, W. Yan, C. Mu, T. Liu, P. Li, and S. Yan. 2016. Recognizing violent activity without decoding video streams. *Optik* 127 (2016), 795–801.
[183] L. Ye, T. Liu, T. Han, H. Ferdinando, T. Seppänen, and E. Alasaarela. 2021. Campus violence detection based on artificial intelligent interpretation of surveillance video sequences. *Remote Sensing* 13 (2021), 628.
[184] S. M. Mohtavipour, M. Saeidi, and A. Arabsorkhi. 2021. A multi-stream CNN for deep violence detection in video sequences using handcrafted features. *The Visual Computer,* 1–16.
[185] N. Honarjoo, A. Abdari, and A. Mansouri. 2021. Violence detection using one-dimensional convolutional networks. In *2021 12th International Conference on Information and Knowledge Technology (IKT)*, 188–191.
[186] M. Cheng, K. Cai, and M. Li. 2021. RWF-2000: An open large scale video database for violence detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 4183–4190.
[187] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, and M. Farazi. 2021. Efficient two-stream network for violence detection using separable convolutional LSTM In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8.
[188] M.-S. Kang, R.-H. Park, and H.-M. Park. 2021. Efficient spatio-temporal modeling methods for real-time violence recognition. *IEEE Access* 9 (2021), 76270–76285.
[189] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9 (1997), 1735–1780.
[190] A. Ali and N. Senan. 2018. Violence video classification performance using deep neural networks. In *International Conference on Soft Computing and Data Mining*, 225–233.
[191] K.-I. Funahashi and Y. Nakamura. 1993. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks* 6 (1993), 801–806.
[192] H. Sak, A. W. Senior, and F. Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Proceedings INTERSPEECH-2014.* 338–342.
[193] A. Ogawa and T. Hori. 2017. Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. *Speech Communication* 89 (2017), 70–83.
[194] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. 2012. Multimodal information fusion and temporal integration for violence detection in movies. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2393–2396.
[195] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. 2012. A benchmarking campaign for the multimodal detection of violent scenes in movies. In *European Conference on Computer Vision*, 416–425.
[196] C.-H. Demarty, C. Penet, M. Schedl, I. Bogdan, V. L. Quang, and Y.-G. Jiang. 2013. The MediaEval 2013 Affect Task: Violent Scenes Detection. In *Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 17–19 October 2013.* 383–395, BioMedical Engineering and Informatics (CISP-BMEI), 1–5.
[197] M. Hörhan and H. Eidenberger. 2013. New content-based features for the distinction of violent videos and martial arts. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1836–1839.
[198] M. Sjöberg, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, and C.-H. Demarty. The MediaEval 2014 Affect Task: Violent Scenes Detection. In *MediaEval.*
[199] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim. 2014. Fast violence detection in video. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, 478–485.
[200] C.-H. Demarty, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, and C. Penet. 2014. Benchmarking violent scenes detection in movies. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 1–6.
[201] H. L. Hammer. 2014. Detecting threats of violence in online discussions using bigrams of important words. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, 319–319.
[202] L. Ye, H. Ferdinando, T. Seppänen, T. Huuki, and E. Alasaarela. 2015. An instance-based physical violence detection algorithm for school bullying prevention. In *2015 International Wireless Communications and Mobile Computing Conference (IWCMC)*, (2015), 1384–1388.
[203] E. Y. Fu, H. V. Leong, G. Ngai, and S. Chan. 2015. Automatic fight detection based on motion analysis,. In *2015 IEEE International Symposium on Multimedia (ISM)*, 57–60.
[204] V. Lam, S.-P. Le, T. Do, T. D. Ngo, D.-D. Le, and D. A. Duong. 2016. Computational optimization for violent scenes detection. In *2016 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, 141–146.
[205] V. M. Arceda, K. F. Fabián, P. L. Laura, J. R. Tito, and J. C. Gutiérrez-Cáceres. 2016. Fast face detection in violent video scenes. *Electr. Notes Theor. Comput. Sci.* 329 (2016), 5–26.
[206] F. De Souza and H. Pedrini. 2017. Detection of violent events in video sequences based on census transform histogram. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 323–329.
[207] K. Lloyd, P. L. Rosin, D. Marshall, and S. C. Moore. 2017. Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures. *Machine Vision and Applications* 28 (2017), 361–371.

[208] T. Z. Ehsan and M. Nahvi. 2018. Violence detection in indoor surveillance cameras using motion trajectory and differential histogram of optical flow. In *2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)*, 153–158.

[209] K. Singh, K. Y. Preethi, K. V. Sai, and C. N. Modi. 2018. Designing an efficient framework for violence detection in sensitive areas using computer vision and machine learning techniques. In *2018 10th International Conference on Advanced Computing (ICoAC)*, 74–79.

[210] M. Gnouma, R. Ejbali, and M. Zaied. 2018. Abnormal events' detection in crowded scenes. *Multimedia Tools and Applications* 77 (2018), 24843–24864.

[211] H. Pan, J. Yin, H. Ku, C. Liu, F. Feng, J. Zheng, and S Luo. 2018. Fighting detection based on pedestrian pose estimation. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. 1–5.

[212] X. Wang, L. Yang, J. Hu, and H. Dai. 2018. A violent behavior detection algorithm combining streakline model with variational model. In *International Conference on Frontiers in Cyber Security*, 216–224.

[213] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino. 2015. Analyzing tracklets for the detection of abnormal crowd behavior. In *2015 IEEE Winter Conference on Applications of Computer Vision*, 148–155.

[214] P. D. Garje, M. Nagmode, and K. C. Davakhar. Optical flow based violence detection in video surveillance. In *2018 International Conference on Advances in Communication and Computing Technology (ICACCT)*, 208–212.

[215] S.-H. Cho and H.-B. Kang. 2014. Abnormal behavior detection using hybrid agents in crowded scenes. *Pattern Recognition Letters* 44 (2014), 64–70.

[216] S. Mohammadi, H. Kiani, A. Perina, and V. Murino. 2015. Violence detection in crowded scenes using substantial derivative. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6.

[217] T. Zhang, W. Jia, X. He, and J. Yang. 2016. Discriminative dictionary learning with motion weber local descriptor for violence detection. *IEEE Transactions on Circuits and Systems for Video Technology* 27 (2016), 696–709.

[218] K. Lloyd, D. Marshall, S. C. Moore, and P. L. Rosin. 2016. Detecting violent crowds using temporal analysis of GLCM texture. *arXiv preprint arXiv:1605.05106,* (2016).

[219] Z. Guo, F. Wu, H. Chen, J. Yuan, and C. Cai. 2017. Pedestrian violence detection based on optical flow energy characteristics. In *2017 4th International Conference on Systems and Informatics (ICSAI)*, 1261–1265.

[220] E. Esen, M. A. Arabaci, and M. Soysal. 2013. Fight detection in surveillance videos. In *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, (2013), pp. 131–135.

[221] K. Biradar, S. Dube, and S. K. Vipparthi. 2018. DEARESt: Deep convolutional aberrant behavior detection in real-world scenarios. In *2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS)*, 163–167.

[222] D. A. Van Dyk and X.-L. Meng. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics* 10 (2001), 1–50.

[223] X. Hu, Z. Fan, L. Jiang, J. Xu, G. Li, W. Chen, X Zeng, G. Yan, and D. Zhang. 2022. TOP-ALCM: A novel video analysis method for violence detection in crowded scenes. *Information Sciences* 606 (2022), 313–327.

[224] S. Blunsden and R. Fisher. 2010. The BEHAVE video dataset: Ground truthed video for multi-person behavior classification. *Annals of the BMVA* 4 (2010), 4.

[225] I. Mugunga, J. Dong, E. Rigall, S. Guo, A. H. Madessa, and H. S. Nawaz. 2021. A frame-based feature model for violence detection from surveillance cameras using ConvLSTM network. In *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, 55–60.

[226] H. M. B. Jahlan and L. A. Elrefaei. 2022. Detecting Violence in Video Based on Deep Features Fusion Technique. *arXiv preprint arXiv:2204.07443,* (2022).

[227] J. S.-V. Robert Fisher and James Crowley. 2004. CAVIAR: Context Aware Vision using Image-based Active Recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[228] M. Q. Gandapur. 2022. E2E-VSDL: End-to-end video surveillance-based deep learning model to detect and prevent criminal activities. *Image and Vision Computing* 123 (2022), 104467.

[229] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier. 2015. VSD, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation. *Multimedia Tools and Applications* 74 (2015), 7379–7404.

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

[231] M. Cheng, K. Cai, and M. Li. 2019. RWF-2000: An Open Large Scale Video Database for Violence Detection. *arXiv preprint arXiv:1911.05913,* (2019).

[232] Ş. Aktı, G. A. Tataroğlu, and H. K. Ekenel. 2019. Vision-based fight detection from surveillance cameras. In *2019 9th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6.

[233] W. Tan and J. Liu. 2022. Detection of Fights in Videos: A Comparison Study of Anomaly Detection and Action Recognition. *arXiv preprint arXiv:2205.11394,* (2022).

[234] F. U. M. Ullah, K. Muhammad, I. U. Haq, N. Khan, A. A. Heidari, S. W. Baik, et al. 2021. AI-Assisted edge vision for violence detection in IoT-Based industrial surveillance networks. *IEEE Transactions on Industrial Informatics* 18 (2021), 5359–5370.

[235] A. Software. Violence Detection for Smart Surveillance Systems. https://www.abtosoftware.com/blog/violence-detection.

[236] A. Mumtaz, A. B. Sargano, and Z. Habib. 2018. Violence detection in surveillance videos with deep network using transfer learning. In *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*, 558–563.

[237] 파튜, 유민, 울라, 아민, 울라, 이미영, et al. 2018. 스마트 감시 애플리케이션을 위해 Deep CNN 을 이용한 폭력 인식. *한국차세대컴퓨팅학회 논문지* 14 (2018), 53-59.

[238] Allerin. 2019. The rise of AI in crime prevention and detection. https://www.allerin.com/blog/the-rise-of-ai-in-crime-prevention-and-detection.

[239] A. Datta, M. Shah, and N. D. V. Lobo. 2002. Person-on-person violence detection in video data. In *Object Recognition Supported by User Interaction for Service Robots*, 433–438.

[240] H. Chandel and S. Vatta. 2015. Occlusion detection and handling: A review. *International Journal of Computer Applications* 120, (2015).

[241] S. B. Kang, R. Szeliski, and J. Chai. 2001. Handling occlusions in dense multi-view stereo. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, (2001), pp. I–I.

[242] S. Liu, C. Guo, F. Al-Turjman, K. Muhammad, and V. H. C. de Albuquerque. 2020. Reliability of response region: A novel mechanism in visual tracking by edge computing for IIoT environments. *Mechanical Systems and Signal Processing* 138 (2020), 106537.

[243] N. Kumar, S. Misra, J. J. Rodrigues, and M. S. Obaidat. 2015. Coalition games for spatio-temporal big data in internet of vehicles environment: A comparative analysis. *IEEE Internet of Things Journal* 2 (2015), 310–320.

[244] T. Wang, K. Chen, W. Lin, J. See, Z. Zhang, Q. Xu, and X Jia. 2020. Spatio-temporal point process for multiple object tracking. *IEEE Transactions on Neural Networks and Learning Systems*.

[245] K. Muhammad, S. Khan, J. Del Ser, and V. H. C. De Albuquerque. 2020. Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey. *IEEE Transactions on Neural Networks and Learning Systems* 32 (2020), 507–522.

[246] L. Joshila Grace, P. Asha, J. Refonaa, S. Jany Shabu, and A. Viji Amutha Mary. 2022. Detect fire in uncertain environment using convolutional neural network. In *Advances in Intelligent Computing and Communication*, (ed.). Springer, 399–404.

[247] Z. Huang, K.-J. Lin, B.-L. Tsai, S. Yan, and C.-S. Shih. 2018. Building edge intelligence for online activity recognition in service-oriented IoT systems. *Future Generation Computer Systems* 87 (2018), 557–567.

[248] K. Muhammad, S. Khan, V. Palade, I. Mehmood, and V. H. C. De Albuquerque. 2019. Edge intelligence-assisted smoke detection in foggy surveillance environments. *IEEE Transactions on Industrial Informatics*.

[249] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. 2018. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6548–6552.

[250] Y. Dai, D. Xu, S. Maharjan, G. Qiao, and Y. Zhang. 2019. Artificial intelligence empowered edge computing and caching for internet of vehicles. *IEEE Wireless Communications* 26 (2019), 12–18.

[251] H. R. Boveiri, R. Khayami, M. Elhoseny, and M. Gunasekaran. 2019. An efficient Swarm-Intelligence approach for task scheduling in cloud-based internet of things applications. *Journal of Ambient Intelligence and Humanized Computing* 10 (2019), 3469–3479.

[252] Z. Li, Z. Wei, Y. Yue, H. Wang, W. Jia, L. E. Burke, T. Baranowski, and M. Sun. 2015. An adaptive hidden Markov model for activity recognition based on a wearable multi-sensor device. *Journal of Medical Systems* 39 (2015), 1–10.

[253] A. Mannini, M. Rosenberger, W. L. Haskell, A. M. Sabatini, and S. S. Intille. 2017. Activity recognition in youth using single accelerometer placed at wrist or ankle. *Medicine and Science in Sports and Exercise* 49 (2017), 801.

[254] E. Bacis, S. D. C. di Vimercati, S. Foresti, S. Paraboschi, M. Rosa, and P. Samarati. 2019. Dynamic allocation for resource protection in decentralized cloud storage. In *2019 IEEE Global Communications Conference (GLOBECOM)*, 1–6.

[255] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, G. Livraga, S. Paraboschi, and P. Samarati. 2022. An authorization model for query execution in the cloud. *The VLDB Journal* 31 (2022), 555–579.

[256] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. 2012. Fog computing and its role in the internet of things. In *1st Edition of the MCC Workshop on Mobile Cloud Computing*, 13–16.

[257] N. Kumar, S. Zeadally, and J. J. Rodrigues. 2016. Vehicular delay-tolerant networks for smart grid data management using mobile edge computing. *IEEE Communications Magazine* 54 (2016), 60–66.

[258] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie. 2019. A semisupervised recurrent convolutional attention model for human activity recognition. *IEEE Transactions on Neural Networks and Learning Systems* 31 (2019), 1747–1756.

[259] X.-Y. Zhang, C. Li, H. Shi, X. Zhu, P. Li, and J. Dong. 2020. AdapNet: Adaptability decomposing encoder-decoder network for weakly supervised action recognition and localization. *IEEE Transactions on Neural Networks and Learning Systems*.

[260] X. Shu, L. Zhang, Y. Sun, and J. Tang. 2020. Host-Parasite: Graph LSTM-in-LSTM for group activity recognition. *IEEE Transactions on Neural Networks and Learning Systems*.

[261] N. Kumar, J.-H. Lee, and J. J. Rodrigues. 2014. Intelligent mobile video surveillance system as a Bayesian coalition game in vehicular sensor networks: Learning automata approach. *IEEE Transactions on Intelligent Transportation Systems* 16 (2014), 1148–1161.

[262] Y. Shen, R. Ji, C. Wang, X. Li, and X. Li. 2018. Weakly supervised object detection via object-specific pixel gradient. *IEEE Transactions on Neural Networks and Learning Systems* 29 (2018), 5960–5970.

[263] G. Li and Y. Yu. 2018. Contrast-oriented deep neural networks for salient object detection. *IEEE Transactions on Neural Networks and Learning Systems* 29 (2018), 6038–6051.

[264] J. Yang, J. Man, M. Xi, X. Gao, W. Lu, and Q. Meng. 2019. Precise measurement of position and attitude based on convolutional neural network and visual correspondence relationship. *IEEE Transactions on Neural Networks and Learning Systems*.

[265] A. Ullah, K. Muhammad, W. Ding, V. Palade, I. U. Haq, and S. W. Baik. 2021. Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications. *Applied Soft Computing* 103 (2021), 107102.

[266] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Geigold, S. Gelly, J. Uszkoreit, and N. Houlsby. 2020. An image is 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, (2020).

[267] L. Meng, H. Li, B.-C. Chen, S. Lan, Z. Wu, and Y.-G. Jiang, 2022. AdaViT: Adaptive vision transformers for efficient image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12309–12318.

[268] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge. 2022. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition* 124 (2022), 108487.

[269] C. Bettini, G. Civitarese, and R. Presotto. 2021. Personalized semi-supervised federated learning for human activity recognition. *arXiv preprint arXiv:2104.08094,* 2021.

[270] K. Sozinov, V. Vlassov, and S. Girdzijauskas. 2018. Human activity recognition using federated learning. In *2018 IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 1103–1111.

Received 6 March 2022; revised 10 July 2022; accepted 21 August 2022