IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS. 2017.Doi Number

A Review on state-of-the-art Violence Detection Techniques

Muhammad Ramzan[1,2] , Adnan Abid[1], Hikmat Ullah Khan[3], Shahid Mahmood Awan[1], Amina Ismail[3], Muzamil Ahmed[4], Mahwish Ilyas[2], Ahsan Mahmood[5]

[1]School of Systems and Technology, University of Management and Technology, Lahore 54782, Pakistan
[2]Department of Computer Science and Information Technology, University of Sargodha, Sargodha 40100, Pakistan
[3]Department of Computer Science, COMSATS University, Islamabad at Wah Campus, Wah Cantt 47040, Pakistan
[4]Department of Computer Science, University of Lahore, Sargodha Campus, Pakistan
[5]Department of Computer Science, COMSATS University, Islamabad at Attock Campus, Attock 43600, Pakistan

Corresponding author: Hikmat Ullah Khan (hikmat.ullah@ciitwah.edu.pk)

ABSTRACT: With the rapid growth of surveillance cameras to monitor the human activity demands such system which recognize the violence and suspicious events automatically. Abnormal and violence action detection has become an active research area of computer vision and image processing to attract new researchers. The relevant literature presents different techniques for detection of such activities from the video proposed in the recent years. This research study reviews various state-of-the-art techniques of violence detection. In this paper, the methods of detection are divided into three categories that is based on classification techniques used: violence detection using traditional machine learning, using Support Vector Machine (SVM) and using Deep Learning. The feature extraction techniques and object detection techniques of each single method are also presented. Moreover, datasets and video features that used in the techniques, which play a vital role in recognition process are also discussed. For better understanding, the steps of the research approaches have been presented in an architecture diagram. The overall research findings have been discussed which may be helpful for finding the potential future work in this research domain.

教育机构，医院，
银行，市场，街道

Keywords: Violence detection, violent behavior, Support Vector Machine, deep learning, machine learning, surveillance camera, computer vision

I. Introduction

Violence detection techniques using computer vision, analyze the surveillance camera videos. Over the last few years, these cameras and other surveillance equipment are installed on different places for the public safety e.g. Educational institutions, hospitals, banks, markets, streets etc. to monitor the activates of people [1]. Monitoring include the behaviors analysis of people, whether their activities are suspicious or normal. This detection of suspicious activity during 24/7 or finding such activity within huge data consist of recorded videos is a very difficult task [2][3]. For this purpose, different methods have been developed to recognize the human activities in real life. These methods help to detect the suspicious activities in the surveillance videos.

Violence detection from the surveillance videos [4] is also a kind of activity detection. Several techniques and methods are developed to detect the brutally events and other harmful patterns in videos [5], [6]. In these methods different approaches are proposed that work with different input parameters. The parameters are basically different attributes or features of the video such as acceleration, flow, time, appearance etc. In the violence activity detection process, first step is to divide a whole video into segments and frames [7]. Secondly, detects the object from the video frames. Thirdly, extract the features of the video according to the applied method. Lastly, detects the anomalous activity from the frames. The step varies according to the method which is applied for detection. The basic steps of violence detection techniques are shown in Figure 1**.** Many researchers proposed different techniques to increase the efficiency, accuracy and performance of the detection process. In this paper, different methods of the violence detection from surveillance videos using computer vision are explored and discussed in detail using systematic literature review.

The main goal of this paper is to present an exhaustive systematic literature review of the methods of violence detection. Over the last decade, different methods of violence and anomalous activity detection are proposed. It is important to classify, analyze and summarizes the proposed methods. To conduct a comprehensive research study, we set up the basic search strings for gathering the most relevant study on the violence activity detection available on the digital libraries. For quality assurance and quality assessment also formulize the assessment criteria directed in [8][9]. The research area of violence detection recently attracts the researchers, when it is noticed that the there is a fine difference between violence and abnormal behavior. Such activities that are separated from normal routine are called abnormal activity and activity which contain fighting, beating; stealing etc. is called violent activity[7], [10], [11]. According to our best knowledge, there is still no systematic literature review of the methods of violent detection.

In this survey, our research contribution can be summarized as follows:

- Classification of the existing models into diverse catergories for better discussion.
- Critical review of each model in chronological order sharing its novelty, main features and limitations.
- Exploration and ranking and significance of the video features for violence detection.
- Discussion of the real-world datasets which are widely used.

The rest of the paper is divided into five main sections. Section II discusses the basic concept related to the violence detection. Section III presents the research methodology based on the literature review and selection process. In Section IV, the techniques of violence are discussed in detail. Section V discusses the video features, and, in the end, section VI summarizes the datasets following the conclusion.

II. Basic Concepts

Here some basic concepts related to violence, computer vision, video features and vision-based recognition of activities are discussed. **Table** ***1*** describes their basic descriptions.

Table 1: Basic Concepts

Sr. No	Feature	Description
1.	**Computer Vision**	with the help of camera and computer images, object shapes etc. from the videos are discovered like human vision. What is happening in images or in a sequence of images are understand on the basis of Machine learning algorithms[12].
2.	**Video Feature**	Video feature are used to detect the activity of an object from surveillance videos[13]
3.	**Centroid**	All point average position of an object shape or space dimension of an object is called the centroid.
4.	**Movement**	An action in which object changes their position in videos.
5.	**Speed**	Movement speed of an object from a specific place to another place
6.	**Direction**	Any object lies along a line from the point where the object is directed.
7.	**Dimension**	Property of a space which measure in length, width and object thickness toward a given direction.
8.	**Acceleration of Images**	Change of velocity or speed over the time unit. Field of acceleration consists of two directions x and y.
9.	**Optical flow field**	Patterns related to the motion of an object, surface and edges of objects.
10.	**Spatiotemporal**	This feature related to time and space of the object.
11.	**Violence**	Such activities that are separated from normal routine are called abnormal activity and activity which contain fighting, beating; stealing etc. is called violent activity[7], [10], [11].

Many researchers attract toward the computer Vision field because the wide range of applications concerned with analysis of images and videos. Analysis of images and videos includes the detection of objects and recognition the activity that is performed by the object. Activity recognition is the complete process which starts from taking a picture or making movies. The basic architecture is shown in

Figure 1. Different researchers proposed different methods in which they use such features according to their own method [14],[15],[7],[13].

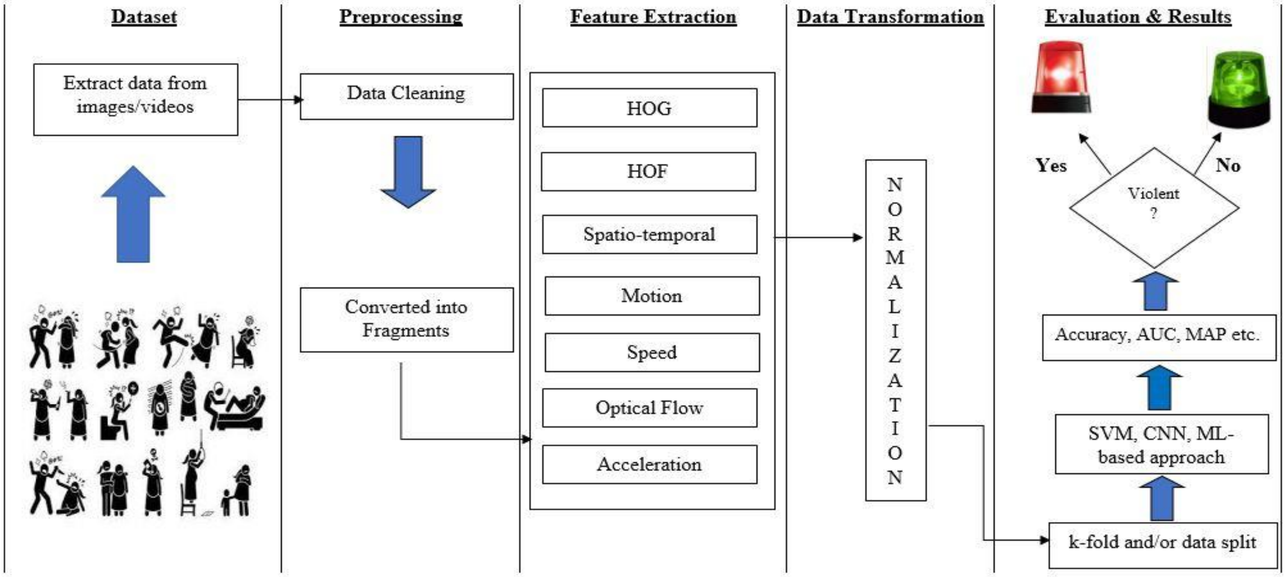

Figure 1: Basic steps for a typical Violence detection technique

III. Research Methodology

The basic goal of this review is to collect categories and recognize the most efficient available methods or techniques that are used in violence and anomalous activity detection from the videos using computer vision. In this domain we analyze the available research with the help of the systematic review process. The available study is evaluated based on predefined criteria. Classify the result of above evaluation process according to relevancy. The population of systematic review contain research papers related to the detection of violence.

A. Data Acquisition and Selection

A well organized and systematic, search is conducted to extract relevant and meaningful information from the heap of data. The basic target of this review process is to sum up the ongoing existing methods in the domain of violence and anomalous activity detection from videos and to discover the most efficient method for the detection of violence and anomalous events. Irrelevant studies are filtered and focus on the research area for taking out the

Table 2 shows the keywords which are selected for the searching literature for this study. In table use wildcard (*) with the term word which mean this word is prefix of different words use in the search string as keyword e.g. "Anom" term is a prefix of anomalous, anomalies, anomaly, etc. Finally, the following is search string which is formulated for execution of automatic research process.

appropriate knowledge for a well-organized systematic literature review. Effective planning for search plays a dynamic role in the review process to find the meaning full available study. There are two types of search: automatic and manual search. Both searches are applied in this process. Starting from the execution of automatic search which is followed by the manual search. In digital data libraries, automatic search is conducted by entering the strings. Gathering more literature, we perform the manual search to explore more information on the violence and anomalous activity detection. Search data from the reference of primary selected papers and gray literature are the types of manual search. Search are limited by following the terms and condition to extract most relevant and exact data related to the violence and anomaly event detection.

The combination of major keywords that return the most relevant information from the massive data is called search Term. The terminologies are analyzed for ensuring the accurate and reliable research in the field of violence and anomaly detection for surveillance videos.

((Violen*) OR (Anomal*) OR (Fight*) AND (activity) OR (event) OR (sequence) AND (Detec*) OR (Recogni*) AND (for) AND (surveillance) OR (observation) AND (Vi*) OR (motion) AND (through) OR (by) OR (via) OR (using) AND (Computer Vision) OR (Deep learning) OR (Machine Learning)).

Table 2: Search word Formulation

Terms	Postfix and Alternative words
Violen*	Violent, Violence, etc.
Anomal*	Anomaly, Anomalous, Anomalies, etc.
Fight*	Fighting, Fight, etc.
Activity	Activity, Event, Sequence, etc.
Detec*	Detection, Detect, Detected, etc.
Recogni*	Recognition, Recognized
Surveillance	Surveillance, Observation, etc.
Vi*	Video, Visual, Visualization, Vision, Motion, etc.
Through	Thorough, by, via, using, etc.

For conducting solid review and finding clear and comprehensive data, different digital data based, science libraries and their search engines are used. To execute the above mention search string to find the different journal publication, conference proceedings and other data related queries, use search engines of following digital databases:

a) Science-Direct Elsevier (https://www.sciencedirect.com)
b) ACM (DL) Digital Library (https://dl.acm.org)
c) IEEE Xplore Access Digital Library (https://ieeexplore.ieee.org)
d) Google Scholar and Google (https://scholar.google.com.pk)
e) Springer-Link Electronic Library (https://www.springer.com)

Initial criteria of the search are to explore the most relevant study to the relevant domain. The following criteria are set to extract the relevant publications.

- Range of years for selecting the publication is 2012 to November 2019.
- Only Full-Length Papers are valid for primary selected study.
- Only Paper written in English is selected, Research papers publish in other languages is excluded.
- Include research papers that are relevant to the domain and search strings.
- Assessment criteria must be addressed in the research.

Quality of research means increasing the relevancy with the domain and decrease the business constraints. Quality assurance criteria is set [16] to extract the study which is valid internally and externally as well.

Initially, we get the 2853 papers from the relevant literature based on the search string which is also called primary search. In the papers initially selected, 120 papers are found to be relevant to the study based on the title of the paper, then based on abstract, 69 papers selected which have the proposed methods and techniques for the detection of violence. Then the whole papers are studied to analyze the proposed techniques of violence detection and assess the quality of paper. Analyze the input on proposed methodology, in this study the features of video use as input. Different methods use different features and feature extraction technique to increase performance and accuracy of detection, 26 papers finalized based on the full text analysis. Then 3 papers are selected from manual searching. In add up of 3 with 26, 29 papers are finalized. The strategy of searching is explained in Figure 2.

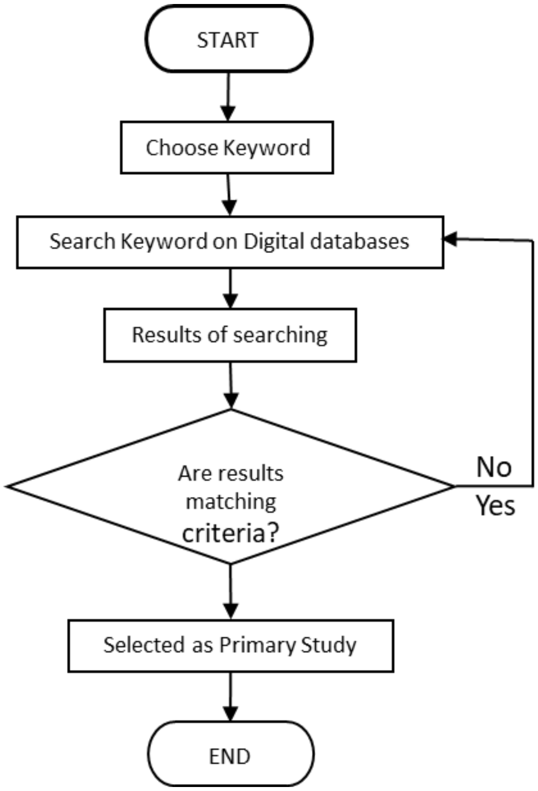
START
Choose Keyword
Search Keyword on Digital databases
Results of searching
Are results matching criteria?
No
Yes
Selected as Primary Study
END

Figure 2: Research Paper Search Strategy

The ratio of selected papers includes 24% from IEEE, 24% of Science Direct, and 17% from ACM (DL), 13% from Springer Link and 20 % from Google Scholar. All the selected study is assessed by the selection process illustrated in Figure 3.

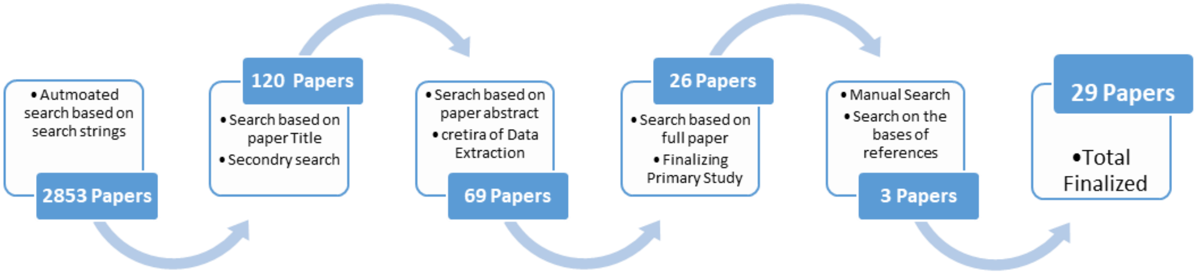
• Autmoated search based on search strings
2853 Papers
120 Papers
• Search based on paper Title
• Secondry search
• Serach based on paper abstract
• cretira of Data Extraction
69 Papers
26 Papers
• Search based on full paper
• Finalizing Primary Study
• Manual Search
• Search on the bases of references
3 Papers
29 Papers
•Total Finalized

Figure 3: **Research Paper Selection Process**

Table 3 represents the detail of finalizing paper, 50 % papers are selected from the IEEE Explore and Science Direct in which mostly we target journal papers 79% and conference papers 13%. Table 4 present the year wise growth rate of selected studies in which column NOP represents number of papers, in cumulative number of papers from pervious and current year, where RGR is the relative growth rate and DT is double time.

$$RGR= \ln \left(\frac{NOP_{current\ year}}{NOP_{previous\ year}}\right)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

$$DT= \left(\frac{\ln 2}{RGR}\right)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (2)$$

Above two equations are used to calculate the value of RGR and DT, such as, RGR of 2018 is calculated by dividing the NOP of 2018 by 2017 than ln of resultant value and DT of 2018 is calculated by dividing ln2 by the value of RGR. Relative Growth show that how NOP growth in current year and DT shows when this growth rate becomes double [17].

Table 3 : List of selection Process with Source information

Literature Source	Automated Search	Step1 Filter	Step2 Filter	Step3 Filter	Manual Search	Finalized
IEEE Explore	120	29	17	7	1	7
Science Direct	578	31	15	6	1	6
ACM DL	589	17	13	5	-	5
Springer	779	23	11	4	-	4
Google Scholar	787	20	13	4	1	4
Total	2853	120	69	26	3	29

Table 4:Year wise Growth Rate

Year	NOP	Cumulative	RGR	DT
2012	1	1	N. A	N. A
2014	2	3	0.69	1.00
2015	1	4	0.6	1.1
2016	11	15	2.39	0.29
2017	5	20	0.79	0.87
2018	8	28	0.47	1.47
2019	1	29	2.07	0.33

IV. Classification of Violence Detection Techniques (VDT)

Violence are the suspicious events or activities in normal life. Recognition of such activities in surveillance videos through computer vision becomes the active topic in the field of action detection [18]. Many researchers proposed different techniques and method for detection of violent or abnormal events as the rapid growth of crime rate for more accurate detection. Different techniques of violence detection are presented that are proposed in the recent years. The techniques of violence detection are classified into three categories based on the classifier used: VDT using machine learning, VDT using SVM and VDT using deep learning. SVM and deep learning are classified separately as these algorithms are used widely in computer vision. The details of each method are summarized in tables. The methods are presented in chronological order. Object detection technique and feature extraction method are also presented.

A. Violence Detection using Machine Learning techniques

Here the methods for violence detection have been discussed that uses different traditional algorithms of machine learning such as KNN, Adaboost, etc. as a classifier.

Table 5 presents the list of detection methodologies that use different classification methods. The classification method states the classification method that uses in proposed technique are also presented.

1. Fast Fight Detection

Within computer vision, the recognition of actions become an active topic. To detect the sequences of violence, a novel method is proposed [2]. It is assumed that in the fight scenes, the motion blobs have a specific shape and position. Firstly, the difference between consecutive frames is computed for absolute images. Then the resulting image is binarized, leading towards the number of motion blobs and marked the largest one on a fight sequence and on a non-fighting scene. K largest blobs of motion are selected only. To categorize the k blobs, different parameters are calculated like centroid, area, perimeter and distance between the blobs as well. Then, blobs are characterized as fight and non-fight. The experiments are performed using the Movies dataset which have 200 clips, Hockey dataset which have 1000 clips and UCF-101 dataset which consists of real videos collected from YouTube. The results of experiment show that the proposed method is outperformed by state-of-the-art methods considered that is BoW (MoSIFT), BoW (SIFT), ViF, LMP, variant v-1 and variant v-2 that used SVM, KNN and Ada boost as a classifier in terms of accuracy and ROC. It has a expressively faster computation time therefore creating it agreeable for the real-time applications.

2. Rotation-Invariant feature modeling MOtion Coherence (RIMOC)

As the events of aggression is hard to define due to the lack of consistency and often needs interpretation of high level, that's why it is decided to categorize that what is present frequently in the videos with violent behavior of humans at a low level that is unstructured and jerky motions. To achieve the purpose, an innovative problem-specific Rotation-Invariant feature modeling MOtion Coherence (RIMOC) is proposed [19]. The presented method has been created on the eigen values attained from the Histogram of Optical Flow (HOF) vectors from the instants of consecutive temporal, densely and locally computed and moreover embedded into a spheric Riemannian manifold. The method is used to learn the models of statistics in a weekly supervised manner. A multi-scale scheme applied on an inference-based method that allow the events with erratic motion to be sensed in time and space, as good applicants of violent events. There is no dataset available for aggressive events specifically. For that purpose, a large dataset is created that consists of sequences from the two different sites that is: from an in-lab fake train and from a real underground railway line, real train and then four datasets are formed: fake train, real train, real train station and real-life contexts. The experiments are performed using these datasets and results of experiment show that the proposed method performs better than all the state-of-the-art methods in terms of ROC per frame and false positive rate.

3. Fast Face Detection

To achieve the goal of detecting faces in violent scenes to help the security controls, the method of fast detection faces is proposed [20]. The Violent Flow (ViF) descriptor is used with Horn-Schunck to detect the violent scenes at early stage. Then to improve the video quality, the algorithm of non-adaptive interpolation super resolution is applied. Lastly, Kanade-Lucas-Tomasi (KLT) face detector is fired. To attain a very time processing, face detector and super resolution algorithm is paralleled with CUDA. CUDA consists of functions that runs at the same time in numerous lightweight threads on the GPU. The experiments are performed to assess the proposed method using Boss dataset and another dataset is built named as violent dataset. Promising results are achieved in detecting faces in terms of Area Under the Curve (AUC) and accuracy.

4. *Violent Activity Recognition without Decoding*

To target the detection of motion and their tracking in most of the methods of activity recognition are complex and limited application of such methods. That's why a fast method of violent activity recognition is proposed which is based on motion vectors [21]. Firstly, the motion vectors are extracted from the compressed video sequences directly. Secondly, the attributes of the motion vectors are analyzed in each frame and between the frames and attain the Region Motion Vector (RMV) descriptor. Lastly, radial basis has been taken using SVM as the kernel function to classify the RMV and learn that whether the activity of violence is present or not. There are many datasets available for activity recognition, but no one is suitable directly as they focus on simple individual activity recognition. That's why to evaluate the proposed method, VVAR10 dataset is built which comprises of 296 positive samples and 277 negative samples. Samples are taken from YouTube, UCF50, UCF sports and HMDB51. Experiments are performed using VVAR10 dataset and results show that it can detect 96.1% of violent actions in the video streams and speed of calculation is fast in terms of accuracy, Miss Alarm Rate (MAR) and False Alarm Rate (FAR) that's why it is suitable for embedded systems.

5. *Automatic Fight Detection*

To detect the fights in a low cost and natural manner, an approach based on motion analysis is proposed [22]. The proposed method is based on the motion analysis. Two approaches are applied to detect the fight. First approach is two-level statistical aggregation that generates feature set. Motion pixel are extracted and then regions of motion from the series of frames by calculating the vectors of optical flow. Then the optical flow images are classified according to the nature of vectors after the elimination of noise. After that, motion statistics are calculated according to classified types to produce the set of features for recognition. The second approach used is Bag of Words (BoW), that is used to produce the visual words set. Then the histogram is used across the set of visual words as a vector to characterize the video for the detection of fight. Experiments are performed using the videos based on real fights and results of experiment show that the proposed method outperforms the existing methods based on MoSIFT descriptors with BoW mechanism and the basic motion signal analysis with BoW approach in terms of recall and precision.

识别姿势的骨架信息，跟踪骨骼关节及其位置

6. *Detection of Abnormal Activity for Bank ATM*

In computer vision, the recognition of posture becomes one of the most interesting field due to its several applications in numerous fields. A technique for an effective intelligent monitoring of real-time ATM, based on skeleton information that is given by Kinect 3D camera for the recognition of postures is proposed [23]. Using Kinect for the tracking of bone joints and their positions, recognition of posture can be achieved. System can be able to detect the abnormal positions by analyzing the information of position. Logistic regression is used as a classifier for training. After the training of machine, system can detect the abnormal and normal behavior and generates alarm in the abnormal situation.

7. *Multiple Anomalous Activity Detection*

Surveillance systems are now installed in the malls, colleges, schools, airports and many other places due to rapid increase in the crimes. As the videos are captured 24/7 so it is tough to monitor them manually and detect the suspicious activities, it increases the demand of intelligent surveillance system. To address the challenge, a method is proposed that detects automatically the various anomalous activities in video clips [7]. It involves three major stages: the detection of moving object, object tracking and understanding of behavior for the recognition of activity. In the first phase of preprocessing, the moving objects are detected, and the removal of noise is done. Then the process of feature extraction is used to identify the key features like direction, speed, dimensions and centroid. The extracted features help to track the objects in the video frames. In the last phase, the method of rule-based classification is utilized to categorize the actions from videos and if some suspicious activity is detected it generates alarm. The experiments are performed on the newly created dataset based on 45 videos that contains three activities which are running, walking and crawling. The results of experiments show that the proposed method can detect the various types of anomalous activities in different scenarios and performs best in terms of accuracy.

8. *Crowd Violence Detection*

In computer vision applications, a rich set of tools is provided by the Lagrangian theory to analyze the long-term, non-local information of motion. On the basis of this theory, a specialized lagrangian technique is proposed [24] for the recognition of violence scenes automatically in the video sequences. Spatio-temporal model based Lagrangian direction fields are used for novel features and used the information of background motion compensation, appearances and long-term motion. An extended approach based on bag-of-words is applied in a late-fusion manner on a per-video basis as a classification scheme to ensure the appropriate temporal and spatial feature scale. The experiments are performed to assess the proposed method using the three-benchmark datasets which is hockey fight, violent crowd and violence in movies. Results show that the addition of Lagrangian theory is valuable sign for the detection of violence and the classification performance increased over the state-of-the-art methods like ViF, HOG + BoW, two stream CNN etc. in terms of AUC and accuracy.

9. Cross-Species Fight Detection

In social signal processing, the detection of behavior from human fights in videos is important, especially in the context of surveillance. In real life, the data collection for the detection of fights generally restricts in machine learning and effects the performance of contemporary data-driven methods. To overcome the challenge, an innovative cross-species learning method along with a set of low-computational cost features of motion for the detection of fight is presented [25]. It avoids the problem of limited human fight data for data-demaining methods effectively. The proposed method exploits the essential commonality among the animal and human fights like the physical acceleration of moving body parts. The proposed method takes input from the videos of animal fight and few human fight videos. A Local Motion Features (LMF) based set is proposed that includes motion statistics, segment correlation following the paradigm of motion analysis. LMF are extracted from each video. Temporal features that is based on human heuristics are extracted and to detect the fight, traditional algorithms of machine learning like SVM is adopted. Ensemble classifiers are proposed to perform cross-species fight detection. The experiments are performed using the samples of video clips, dataset of hockey and movies. Results of experiments show that the proposed method performs better than the state-of-the-art methods like ViF, OViF and motion signals in terms of accuracy.

10. Do violent people smile

The admiration of social media platforms has been determined the advent of abuse and violent behavior that reflect the issues of real life in the digital life like raphy, cyberbullying, internet banging and others. To investigate the users that usually adopt the hate speech and used offensive language in Twitter by examining the profile pictures, a model is proposed [26] that automatically detects offensive language using NLP methods, sentiment analysis and meta information. Dictionaries and lexical resources contain the terms which are relevant to hate speech in precise contents but have positive or neutral meaning in the other context, dictionary approach is combined with machine learning method for that purpose to recognize the offensive tweets. Filters are applied to the gathered tweets based on the profanity vocabulary defined and selected 15% from Twitter A and 17% from Twitter B. After that Machine-Learning based model is applied to detect the offensive language and hate speech. The model is based on logistic regression with L2 regularization. The proposed model differentiates the tweets as hate speech tweets (class 0), offensive tweets (class 1) and neutral tweets (class 2). The tweets have been classified as class 0 declared as violent and class 2 as non-violent. The analysis of profile pictures show that violent users are young and smile less. There is a high percentage of females that used offensive language. Results shows that the model gives 94 % in precision and recall.

Table 5 : Violence Detection Techniques using different classification techniques

Method	**Object Detection Method**	**Feature Extraction Method**	**Classification Method**	**Scene Type**	**Accuracy %**
Motion Blob (AMV) acceleration measure vector method for detection of fast fighting from video [2]	Ellipse detection method	An algorithm to find the acceleration	Spatio- temporal features use for classification	Both crowded and less crowded	Near 90 %
RIMOC method focuses on speed and direction of an object on the base of HOF (Histogram Optical flow) [19]	Covariance Matrix method STV based	Spatio- temporal vector method (STV)	STV uses supervised learning	Both crowded and uncrowded	Results for normal situation 97%. Dataset of train station 82 %
The method includes two	Vif objection	Horn shrunk	Interpolation	Less	Lower frame rate

step detection of violent and faces in video by using VIF descriptor and normalization algorithms [20]	recognition CUDA method and KLT face detector	method for Histogram	classification	crowded	14 % too high rate of 35 fs/s 97 %
SVM method for recognition based on statistical theory without decoding of video frames [21]	Vector normalization method	Macro block technique for features extractions	Region motion and descriptor for video classification	Crowded	96.1 %
Detecting Fights with motion blobs [22]	Binarization of images	Spatio-temporal method to extract blobs	Classify on the base of blob length. Largest consider fighting.	Crowded	Depend upon dataset 70% to 98%.
Kinect framework by analyzing the posture for recognition abnormal activates of ATM 3D cameras [23]	Posture recognition using logistic regression	Joint angle for acquiring posture	Gradient decent Method for classification	Less Crowded	85% to 91 %
Lagrangian fields of direction and begs of word framework to recognize the violence in videos [24]	Global compensation of object motion	Lagrangian Theory and STIP method for extract Motion features	Late fusion for classification	Crowded	91% to 94%
A simple approach to form a video to preprocessing then feature extraction and recognition of normal and abnormal event. [7]	Gaussian Mixture model	Apply different formulas on the consecutive frame to extract required feature	Rule based classification using a default threshold	Less crowded	Up to 90 %
A totally different technique in which learns for animal fighting to detect human fight [25]	Motion region and Optical flow method	Vif, OVif and IfV methods	Transfer Learning approach	Less crowded	90 %
Framework to detect violent set a parameter that violent people smile by analyzing social media [26]	Demographic analysis approach	Ethnicity framework	Machine learning models	Less crowded	95%

B. Violence Detection Techniques using SVM

Here the techniques of violence detection are discussed in detail that utilizes the Support Vector Machine (SVM) as a classifier. **Table** ***6*** present a list of violent event recognition techniques. SVM is an algorithm that is used to solve classification problems using supervised learning. In SVM, we plot data on (number features) dimension space and differentiate within two classes. SVM is widely used method in computer vision as it is robust and considers numeric features. It is used for the tasks related to binary classification. SVM is based on kernel. Kernel is a function that converts input to the high dimensional space where the problem is answered. The major disadvantage of SVM is the lack of transparency of the results [27]. Now the methods of violence detection that uses SVM are elaborated in detail separately.

1. Real-Time Detection of Breaking Violence in Crowded Scenes

While the use of surveillance cameras is commonly used but their effectiveness is questionable. To overcome the challenging task of monitoring the violence in crowded scenes, an innovative approach for real-time violence detection in crowded scenes is proposed [28]. The proposed method considers the statistics of how the flow vectors magnitude changes across the time. the two related but different tasks are considered: violence classification and violence detection. The basic goal is to detect the change from violent to non-violent behavior with the shortest delay from the time that the change occurred. Violent Flows (ViF) descriptor is used to represent the statistics that is collected for short frame sequences. Then the Vif descriptors are categorized as violent or non-violent behavior using the Linear SVM. Experiments are performed using the self-made dataset based on surveillance videos collected from YouTube along with the Hockey dataset. The results of experiments show that the proposed method performs better than the existing methods like HOG, HOF etc. by depending on the magnitude of optical flow fields alone in terms of Area Under the Curve (AUC) and accuracy.

2. *Fast Violence Detection*

In the computer vision, the problem of action recognition becomes an active topic. The well-known framework bag-of-words was used in the recent work which is used for the recognition of fight specifically. In this scheme, spatio-temporal features have been taken out from the video frames and used for the classification and the accuracy rates of 90% achieved for this task. Inspired by the results which advises that kinematic features are alone discriminant for the specific actions, an innovative method that used extreme acceleration method as a main feature proposed [3]. To estimate the extreme features efficiently, Random transform is applied on the consecutive frames of video. And used SVM and Adaboost as a classifier. Experiments using two special datasets of hockey and movies was performed and results show that accuracy improves up to 12% and performs better than the state-of-the-art generic action recognition methods that have features like Scale Invariant Feature Transform (SIFT) and Motion Scale Invariant Feature Transform (MoSIFT) [29] and the proposed method is 15 times faster as well.

3. *Gaussian Model of Optical Flow (GMOF)*

For surveillance systems, violence detection becomes an active topic, but not studied as much as action recognition. Earlier methods primarily focus on the violence detection and made few efforts to decide the violence location. That’s why a robust and fast framework to detect and localize the violence in surveillance scenes proposed [10]. In the framework to extract the candidate regions of violence, Gaussian Model of Optical Flow (GMOF) is presented. The model is adaptively designed as deviation from crowds’ normal behavior detected in the scenes. Then to each video volume which is created by densely sampling the regions of candidate violence, violence detection is performed. Then to differentiate the violent scenes from non-violent scenes, an innovative descriptor named Orientation Histogram of Optical Flow (OHOF) proposed as well. Firstly the training module selects training data and extracts OHOF descriptor then using linear SVM, obtains the feature model. Secondly the detection module, GMOF is presented to separate the regions of candidate violence, then using the technique of multi-scale scanning window in densely sampling regions of candidate violence, OHOF descriptor is extracted. In the end, the descriptor is compared with the trained SVM model of known violent scenes. Experiments are performed using three challenging datasets of Behave, Caviar and Crowd Violence. The results of experiment show that the proposed descriptor performs better than the state-of-the-art descriptors like Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOF), Combination of HOG and HOF (HNF), MoSIFT and SIFT in terms of Area Under the Curve (AUC) and accuracy.

4. *Detecting Violence in Videos using Subclasses*

To address the challenging problem of detecting violence in videos, Li et.al [30] specifically focuses on the combination of multi-modal features by adding and exploiting subclasses visually linked to violence. The dataset of MediaEval 2015 is used to fulfil the purpose. The videos of dataset are labeled manually with respect to the subclasses and results in the 10 subclasses of violence like blood, gun, death and others. The concept of subclasses taken from the existing studies. SVM is used as classifier to train the subclasses and the set is divided in to two randomly disjoint sets, 70% for training and 30% for testing. The procedure is also applied on test set and results shows that solution based on subclasses outperforms the existing methods that contains motion features like HOG, Motion Boundary Histogram (MBH) and HOF with Average Precision of 0.303 and Precision at 100 of 0.55 on the MediaEval dataset.

5. *Human Violence Recognition and Detection*

The recognition and detection of violence becomes an important topic for surveillance videos. The basic purpose is to determine if the violence occurs that is recognition and when it happens means detection of violence. Firstly, the extension of Improved Fisher Vectors (IFV) is proposed for video clips [31]. Local features and their spatio-temporal positions are used that allows to represent the video. Then for violence detection, popular sliding window approach is studied. To speed up the approach, summed area table data structure is used and the formulae of IFV is re-formulated. First, local spatio-temporal features are extracted from videos using Improved Dense Trajectories (IDT). Then, video representation for each descriptor are calculated independently like

HOG to represent the video using IFV. Then, linear SVM classifier are used for violence recognition and in the end using an approach of fast sliding window, violence is detected. Extensive evaluation is performed using 4 state-of-the-art datasets of Violent-Flows, Movies and Hockey Fight. Violence-Flow 21 dataset is used for violence detection task. And the results show that proposed approaches perform best over existing approaches like HNF, Jerk, HOF, HOG, Violent Flow (ViF), Histogram of Oriented Tracklets (HOT) based on the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) metrics.

6. *Violence Detection using Oriented Violent Flow*

The demand of market for intelligent violence detection increasing continuously with surveillance cameras, but still it is a challenging research area. Firstly, an innovative feature extraction method named as Oriented Violent Flows (OViF) is proposed for practical detection of violence in videos [32]**.** In statistical motion orientations, it takes full advantage of the motion magnitude change information. AdaBoost is used for the selection of features and then SVM classifier is trained on chosen features. Experiments are performed on Hockey and Violent-Flow database datasets to evaluate the usefulness of proposed method and results show that proposed method performs better than the baseline methods: LTP and ViF in terms of accuracy and AUC. Secondly, feature combination and multi-classifier combination strategies are adopted, and outstanding results are attained. The results of experiment show that utilizing combined features with AdaBoost and Linear-SVM attains improved performance across the state-of-the-art methods on the Violent-Flows benchmark.

7. *Automatic Real-Time Video-based Surveillance System*

In the surveillance system, the detection of suspicious activity plays a vital role. In an environment of academics, there is a crucial need of surveillance system that can perform robustly. That’s why a new framework for a real-time video-based surveillance system that detects automatically is proposed [33]. The work is divided into three phases for the development of system. The preprocessing phase involves the detection of abnormal human activity and content-based image retrieval phase. In the preprocessing phase, all students must register before beginning a course of study and registration involves the collection of personal details and student must submit their own photo to create a student card. The proposed system needs the pictures of students as well in different conditions like anger, fear, sadness etc. to obtain an exact description in terms of Content-Based Image Retrieval (CBIR). These records are saved in a database of CBIR for the case to detect abnormal activity of students. Then in the next phase, the image is converted into frames. Temporal-differencing algorithm is used to detect motion objects and then using the Gaussian function, motion regions are located. Further for the recognized objects that is human or non-human, shape model based on OMEGA equation is used as a filter. Human activities are classified as normal and abnormal activities using SVM. In the case of abnormal activities of human, machine generates automatic warning. Using CBIR, it also inserts the method to retrieve the recognized object from database for the identification and recognition of object. Finally, software-based simulation using MATLAB is done and experimental results show that the system achieves the tracking, semantic scene learning and the detection of abnormality simultaneously in an environment of academics without the involvement of human.

8. *Robust Abnormal Human Activity Recognition*

To detect any abnormal activity with the elderly people and to support the idea of quality and independent living, a robust framework for the recognition of abnormal human activity is proposed [34]. By computing the integration of feature vectors which are Histogram of Oriented Gradients (HOG) and Zernike moments on Average Energy Images (AEI), framework is structured to construct a robust feature vector. The compact representation of video sequences is provided by the formation of AEI without any loss of spatio-temporal information. The combination of Zernike moments and HOG augments inter-class separation, translational and rotational invariance. The depth silhouettes are attained by Microsoft’s Kinect sensor which are utilized to produce clean binary silhouettes by background subtraction making the pre-processing simpler and faster with accuracy. The mutual feature vector dimensions are reduced by applying PCA and to classify the activities, SVM is applied. The proposed work is evaluated on publicly available Kinect Activity Recognition Dataset (KARD) 3D dataset and UR fall detection. The experiments show that the results achieved 94% and 95.22% for UR fall dataset, and KARD dataset in terms of Average Recognition Accuracy (ARA) respectively.

9. *Novel Framework for High-Level Activity Analysis*

A novel framework that is based on late fusion for the analysis of high-level activity using multi-independent temporal perception layers is proposed

[35]. Two kinds of perception layers which is based on SVM and Situation Graph Trees (SGT) are build. The framework comprises of three stages: multi-temporal analysis, multi-temporal perception layers and late fusion. The results attained from the multi-temporal perception layers are fused into an activity score over a step of late fusion. To evaluate the approach, the framework is applied to the detection of violent events in visual surveillance. Three famous datasets: NUS-HGA, Behave and some videos from YouTube are used for the experiments. The experimental results show that multi-temporal framework outperforms the existing single-temporal frameworks in terms of accuracy showing that the use of multi-temporal method has an advantage over single-temporal methods.

10. Real Time Violence Detection

To reliably detect violence actions, manually-selected features are insufficient typically. That's why the model based on bi-channels CNN and SVM for violence detection is proposed [36]. The proposed model comprises of three portions: feature extraction, SVM training and label fusion. Firstly, the structure of bi-channels CNN is used to extract two features. The first feature is the original video frame that is used to extract the features of appearance and second input is the difference of adjacent frames that is used to separate the features of motion. Then, the appearance and motion are adopted as classifier for linear SVM. In the end, the result of detecting violence is attained by using a method of label fusion which integrates the motion information and the appearance information. The experiments are performed to evaluate the proposed method using Hockey Fight and Violent Crowd datasets. The results of experiments show that the proposed method performs better than the existing methods like HOG, HOF, MoSIFT, SIFT, Two-stream in many realistic scenes in terms of accuracy.

11. Breaking Down Violence

The use of automatic means to detect violence in videos is substantial for the analysis of surveillance cameras and law enforcement to maintain the safety of public. Also, it helps for the protection of children from getting into the unsuitable content and helps parents in making better decision about what their children should watch. Although it is a challenging problem as the definition of violence is highly subjective and broad. That's why the detection of violence in video clips without the supervision of human is not only a conceptual but also a technical problem. To overcome the problem, the idea of violence used for CNN by breaking it into more concrete and objective parts is proposed [37]. Firstly to learn the features specifically related to violence like blood, explosions, fights etc., independent networks are used. Then to describe the violence, using such features, distinct SVM classifiers are trained for each concept and then fuse them later in a meta-classification. It is also explored that how to represent the time-based events as network inputs for still images as many images are defined in the form of movement. The experiments are performed to assess the proposed method using the EvalMedia 2013 VSD dataset. The results show that the concept of breaking violence into smaller concepts proved to be an effective solution in terms of MAP@100 and AUC.

12. Automated Detection of Fighting Styles

To classify the videos of martial arts, a recognition method is proposed [38]. In the preprocessing phase to clean out the unnecessary parts of videos, firstly crop and segment the respective dataset. Then for feature extraction, a spatio-temporal interest points are utilized to detect the regions in the videos linked with the movement in a sequence of frames. For classification, training vocabulary of all descriptors are constructed and then converted into clusters. Then each video is represented as a histogram of these clusters. Further, KNN or SVM is applied as a classifier to the videos. The experiments are performed using the videos of Olympics including 2012 London Olympics and the 2014 Nanjing Youth Olympics. The two major classes of videos are selected that is: Judo and Taekwondo. The results of experiments show that the proposed method using SVM as a classifier outperforms existing methods that used KNN and SVM with the RBF kernel as a classifier in terms of accuracy.

Table 6: Violence Detection Techniques using Support Vector Machine

Method	Object Detection Method	Feature Extraction Method	Scene Type	Accuracy %
real-time detection of violence in crowded scenes [28]	ViF descriptor	Bag of features	crowded	88%
Bag of words framework using acceleration for	Background	Ellipse estimation method for	Less	Approx. 90%

action detection [3]	subtraction algorithms	consecutive frames	crowded	
GMOF framework with tracking and detection module [10]	Gaussian Mixture model	OHFO for optical flow extraction	Crowded	82 % - 89 %
Multi model features framework on the base of the subclass [30]	Image CNN And ImageNet	Google Net for feature extraction	Less crowded	98%
To determine the occurrence of violent purpose extended form of IFV (Improved Fisher vector) and sliding windows [31]	Spatial pyramids and grids for object detection	Spatio temporal grid technique for feature extraction	Crowded	96 % - 99 % using different data sets
Violence detection using Oriented Violent Flow [32]	Optical Flow method	Combination of ViF and OViF descriptor	crowded	90%
AEI and HOG combined framework to recognize the abnormal event in visual motions [34]	AEI technique for background subtraction	HOG and spatio- temporal methods to extract features	Both crowded and less crowded	94% - 95%
The framework includes preprocessing, detection of activity and image retrieval. This work identifies the abnormal event and image from data-based images.[33]	Optical flow and temporal difference for object detection CBIR method for retrieving images	Gaussian function for video future analysis	Less crowded	97%
Late Fusion method for temporal perception layers to detect high level activities. Use multiple cameras from 1 to N. [35]	A motion vector method to identify from multiple cameras in two dimensions	SGT MtPL method	Less crowded	98 %
Bi-Channel Convolutional neural network for real time detection [36]	ImageNet method of object detection	VGG-f model for feature extraction	Crowded	91% - 94%
Solve detecting problem by dividing the Objective in depth and clear format using Con Net [37]	Movement detection and TRof Model	BoW approach	Less crowded	96 %
Bag of Words method using the Spatial Temporal method for detection anomalies in the video [38]	Representation of segments and sub segments	Using HOF and HOG for acquiring video features	Less Crowded	84% - 91%

C. Violence Detection Techniques using Deep Learning

Here the techniques of violence detection are discussed in detail that uses algorithms of deep learning in the proposed frameworks. **Table 7** present the list of recognition methods which uses convolutional neural network (CNN, Cov-Net) base classification [39]. Deep learning is based on neural networks. The technique is used to classify the violent recognition on the base of data set and extracted features using more convolutional layers. Now the methods of violence detection that uses the algorithms of deep learning are elaborated in detail separately

1. Violence Detection using 3D CNN

The typical methods for fight detection rely on the domain knowledge to construct the complex handcrafted features from the input. On the other hand, deep models can directly act and automatically extract the features. Ding et.al proposed [40] the novel 3D convNets approach for the detection of violence without using prior knowledge in the videos. 3D CNN is used to compute the convolution on the set of video frames, hence the information about motion is extracted from the input. Model is trained using supervised learning ad gradients are computed using the back-propagation method. Experiments are performed using the Hockey dataset and results show that the proposed method performs superior without relying on handcrafted features in terms of accuracy.

3D CNN用于计算视频帧集上的卷积，因此从输入中提取了有关运动的信息。使用监督学习训练模型，并使用反向传播方法计算梯度

一种基于CNN的弱监督场所识别架构

2. *Deep Architecture for Place Recognition*

To overcome the problem of recognition of large-scale visual places in which the basic job is to rapidly and accurately identify the place of a certain query photographs, CNN based architecture for weakly supervised place recognition was proposed [41]. The proposed method has three principles. Firstly, CNN based architecture is developed which is trained in an end-to-end means to fulfil the place recognition task directly. NetVLAD is a key element of this architecture, it is a new comprehensive VLAD layer which is encouraged by the Locally Aggregated Descriptors' vectors, mostly used for image retrieval. VLAD layer is agreeable to training using back propagation and eagerly pluggable into any architecture based on CNN. Secondly, to acquire the parameters of architecture into an end-to-end means from the images representing the same locations across the time which is taken from Google Street View Time Machine, newly developed weakly supervised ranking loss-based training procedure is developed. Finally, experiments are performed based on the typical place recognition evaluation process using the datasets of Pittsburgh and Tokyo 24/7 freely available and results shows that the proposed architecture performs better than the non-learnt image representations and off-the-shelf descriptors of CNN on two challenging place recognition benchmarks and present state-of-the-art image representation on image retrieval benchmark as well.

3. *Violent Scene Detection using CNN & Deep Audio Features*

Violent scene detection system is proposed [42] that uses CNN built on acoustic information from video clips. CNN is applying in two manner: as a classifier and as an extractor of deep acoustic feature. Firstly, 40-dimensional Mel Filter-Bank (MFB) is utilized as the input feature to the CNN with their delta and delta-delta. Then the video is converted into short chunks. MFB features are divided into 3 feature channels to explore the local features. Then CNN is used for feature representation. CNN-based features are used to built SVM classifiers. Then the detection of violent scene is performed on each chunk of video. Then the detection is produced by max or min pooling on the segment-level detections. Experiments are performed via MediaEval dataset and results show that the proposed method performs better than the baseline methods: audio only, visual only and audio learned fusion and visual in terms of average precision.

4. *Detect Violent Videos using Convolutional Long Short-Term Memory (ConvLSTM)*

Deep neural network based method is proposed [43] to recognize the violence in videos. CNN is used to extract the features from frame level in videos. Then these features are accumulated using a variant of LSTM that uses convolutional gates. The combination of CNN and ConvLSTM can take the localized spatio-temporal features that enables local motion analysis taking place in the videos. It is also proposed to use the adjacent frame differences as input to the model that encode the changes occurred in the videos. Experiments are performed using three popular datasets: Hockey, Movies and Violent-Flows. Results reveal that the proposed model performs better than the state-of-the-art methods like ViF+OViF, ViF, three streams + LSTM and others in terms of accuracy

5. *Detecting Human Violent Behavior By integrating Trajectory and Deep CNN*

The typical methods for violence detection generally depend on hand-crafted features that is not appropriate mostly. Inspired by the performance of deep models for the recognition of human action, an innovative method for the detection of human violent behavior by combining the trajectory and deep CNN is proposed [44] that takes advantage of both hand-crafted features and deep-learned features. Experiments are performed on the two real-world datasets: Hockey Fight and Crowd Violence. The results reveal that the proposed method performs better than the existing methods: HOG, HOF, ViF, and others in terms of accuracy.

6. *Fight Recognition Method*

In computer vision, recognition of actions becomes a significant line of research. The tasks like aggressive behavior or fights comparatively studied less that may be useful in many scenarios of surveillance videos like prisons, psychiatric wards or in personal smartphone as well. Their vast usability creates the interest to develop the violence or fight detectors. The major aspect of the detectors is efficiency means these approaches should be fast computationally. Handcrafted spatio-temporal features achieve high accuracy for both appearance and motion, but the extraction of some features is still prohibitive for the real-world applications. First time, the paradigm of deep learning is applied on the task that uses 3D CNN which takes the full video sequence as an input. But the motion features for this task are crucial and using full video as an input cause noise and redundancy in the process of learning. For that purpose, a hybrid feature "handcrafted/learned"

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2932114, IEEE Access

时空特征和3D CNN

framework proposed [45]. The method firstly aims to get the illustrative image from the video sequence taken as an input for feature extraction and Hough forest is used as a classifier. Then to classify that image and get the conclusion for the sequence, 2D CNN is used. Experiments using three well known datasets of hockey, movie and behave are performed. And results reveal that the proposed method perform better than the different methods of handcrafted and deep learning based on the accuracies and standard deviations.

7. *Violence Detection using Spatiotemporal Features with 3D CNN*

For the recognition of violent activities, an enhanced surveillance system is mandatory for security purposes to avoid the social, economic and ecological damages. For this purpose, the framework of triple-staged end-to-end deep learning violence detection is proposed [46]. Firstly in the surveillance video streams, persons are detected using light-weight CNN modelto overcome and reduce the huge processing of unusable frames. Secondly, an order of 16 frames with detected individuals is passed to 3D CNN, where the spatiotemporal features of these sequences are extracted and fed to the Softmax classifier. Then, the 3D CNN model is optimized using a neural networks optimization toolkit and an open visual inference developed by Intel. Trained model is converted into an intermediate illustration and changes it for execution at the end platform for the final detection of violence. After the violence detection, an alert is transported to adjacent security department or police station to yield the action. Experiments are performed using the Violent Crowd, Hockey and Violence in Movies datasets. The results of experiment reveal that the proposed method performs better than the state-of-the-art methods like ViF, AdaBoost, SVM, Hough Forest and 2D CNN, sHOT and others in terms of accuracy, precision, recall and AUC

Table 7: Violence Detection Techniques using Deep Learning

Method	**Object Detection Method**	**Feature Extraction Method**	**Scene Type**	**Accuracy %**
Violence Detection using 3D CNN [40]	3D convolution is used to get spatial information	Back propagation method	crowded	91%
Deep architecture for place recognition [41]	VGG VLAD method for image retrieval	Back propagation method for feature extraction	Crowded	87% - 96 %
Violent scene detection using CNN & deep audio features [42]	MFB	CNN is used	crowded	Approx. 90%
Detect violent videos using convLSTM [43]	CNN along with the ConvLSTM	CNN model	crowded	Approx. 97 %
Detecting Human Violent Behavior By integrating Trajectory and Deep CNN	Deep CNN	Optical flow method	crowded	98 %
Hough Forest Methodology for reorganization [45]	Object detection using spatio- temporal features	MoSIFT method to extract video features	Less Crowded	84% - 96%
Violence Detection using Spatiotemporal Features with 3D CNN [46]	Pre-train Mobile Net CNN model	3D CNN	crowded	Approx. 97%

V. Video Features

Video feature are the basic elements to detect an activity from video. The accuracy of the methodology directly depends upon the dataset and features that are extracted from video to analyze the ***Table 8*** we present all the features that are used in selected study.

pattern of activity. E.g. in fight scenes the movement of different objects becomes speedier. In the normal scene movement of object are normal and not moving so fast. Direction of object movement and with respect to time and space is also used to analyze the anomalous events. In

Table 8: Video Features used in the selected Studies

Ref	Extracted video features	Ref	Extracted video features
[47]	Motion blobs, Edges and corner of image.	[30]	Motion features (HOG, HOF)
[19]	Motion, direction and speed	[26]	Emotions and images edges
[35]	Movement, direction, speed	[3]	Spatiotemporal, acceleration and motion
[5]	Spatial, temporal and motion stream	[2]	Motion blobs
[8]	Motion, space and time	[41]	Spatiotemporal features
[20]	Optical flow, Magnitude	[1]	Temporal (negative and positive)
[7]	Speed, direction, centroid, movement	[6]	Optical flow, motion and moving bob
[21]	Motion vector and direction	[38]	RGB images, optical flow and acceleration
[10]	Spatiotemporal and motion	[24]	Spatial, temporal and motion
[33]	Motion region and optical flow	[45]	Appearance, motion, optical flow
[11]	Density, direction and speed	[22]	Motion, acceleration and magnitude
[36]	Direction and motion information	[31]	Optical flow, HOF and HOG
[48]	Motion vector	[34]	HOG and motion
[37]	Spatiotemporal	[23]	Motion and postures
[25]	Movement and acceleration		

VI. Datasets

In this section, the real-world datasets are discussed that are used for the evaluation of the proposed techniques. **Table *9*** summarizes the details of all datasets related to the violence.

Table 9: summary of Datasets

Sr No.	Dataset Name	Ref	No. of images/ Clips	Year of Release	Ref used
1.	SBU Kinect Interaction	[49]	21 sets of 8 violent interactions.	2012	[47]
2.	Hockey	[50]	1000 clips	2011	[2], [3], [24], [31], [36], [45]
3.	Movies	[50]	200 clips	2011	[2], [24], [31], [45]
4.	Behave	[51]	200,000 frames	2010	[45]
5.	Caviar	[52]	-	2004	[22]
6	KARD	[53]	-	2017	[34]
7.	Media Eval	[54]	10,000 clips	2015	[30]
8.	UCF 101	[55]	13,000 clips	2012	[5], [6]

使用SVM，CNN和传统的基于机器学习分类的暴力检测方法

Conclusion

With the rapid growth of surveillance cameras in different fields of life to monitor the human activity, also grow the demand of such system which recognize the violent events automatically. In computer vision, violent action detection becomes hot topic to attract new researchers. Indeed, many researchers proposed different techniques for detection of such activities from the video. The basic goal of this systematic review is to explore the state-of-the-art research in the violence detection system. The systematic review delivers details of methods using SVM, CNN and traditional machine learning classification-based violence detection. These techniques are explained in detail and their pros and cons are also deliberated. Moreover, datasets and video features that used in all techniques, which play a vital role in recognition process are listed in comprehensive tables. Accuracy is depending upon the techniques of object recognition, features extraction and classification along with dataset being used. Our study potentially contributes in highlighting the techniques and methods of violence activity detection from surveillance videos.

References

[1] W. Sultani, C. Chen, and M. Shah, “Real-world Anomaly Detection in Surveillance Videos,” *Center for Research in Computer Vision (CRCV), University of Central Florida (UCF)*, 2018.

[2] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T.-K. Kim, “Fast fight detection,” *PloS one*, vol. 10, no. 4, p. e0120448, 2015.

[3] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, “Fast violence detection in video,” in *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, 2014, vol. 2, pp. 478–485.

[4] L. Tian, H. Wang, Y. Zhou, and C. Peng, “Video big data in smart city: Background construction and optimization for surveillance video processing,” *Future Generation Computer Systems*, 2018.

[5] C. Dhiman and D. K. Vishwakarma, “A review of state-of-the-art techniques for abnormal human activity recognition,” *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 21–45, 2019.

[6] P. Zhou, Q. Ding, H. Luo, and X. Hou, “Violent Interaction Detection in Video Based on Deep Learning,” in *Journal of Physics: Conference Series*, 2017, vol. 844, no. 1, p. 12044.

[7] S. Chaudhary, M. A. Khan, and C. Bhatnagar, “Multiple Anomalous Activity Detection in Videos,” *Procedia Computer Science*, vol. 125, pp. 336–345, 2018.

[8] A. Ben Mabrouk and E. Zagrouba, “Abnormal behavior recognition for intelligent video surveillance systems: A review,” *Expert Systems with Applications*, vol. 91, pp. 480–491, 2018.

[9] Z. Mushtaq, G. Rasool, and B. Shehzad, “Multilingual Source Code Analysis: A Systematic Literature Review,” *IEEE Access*, vol. 5, pp. 11307–11336, 2017.

[10] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, “A new method for violence detection in surveillance scenes,” *Multimedia Tools and Applications*, vol. 75, no. 12, pp. 7327–7349, 2016.

[11] M. Alvar, A. Torsello, A. Sanchez-Miralles, and J. M. Armingol, “Abnormal behavior detection using dominant sets,” *Machine vision and applications*, vol. 25, no. 5, pp. 1351–1368, 2014.

[12] X. Li and Y. Shi, “Computer Vision Imaging Based on Artificial Intelligence,” in *2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*, 2018, pp. 22–25.

[13] Q. Li and W. Li, “A Novel Framework for Anomaly Detection in Video Surveillance Using Multi-feature Extraction,” in *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, 2016, vol. 1, pp. 455–459.

[14] G. M. Basavaraj and A. Kusagur, “Vision based surveillance system for detection of human fall,” in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, 2017, pp. 1516–1520.

[15] P. A. Dhulekar, S. T. Gandhe, N. Sawale, V. Shinde, and S. Khute, “Surveillance System for Detection of Suspicious Human Activities at War Field,” in *2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*, 2018, pp. 357–360.

[16] H. Dar, M. I. Lali, H. Ashraf, M. Ramzan, T. Amjad, and B. Shahzad, “A Systematic Study on Software Requirements Elicitation Techniques and its Challenges in Mobile Application Development,” *IEEE Access*, vol. 6, pp. 63859–63867, 2018.

[17] R. Piryani, D. Madhavi, and V. K. Singh, “Analytical mapping of opinion mining and sentiment analysis research during 2000--2015,” *Information Processing & Management*, vol. 53, no. 1, pp. 122–150, 2017.

[18] R. Olmos, S. Tabik, and F. Herrera, “Automatic handgun detection alarm in videos using deep learning,” *Neurocomputing*, vol. 275, pp. 66–72, 2018.

[19] P. C. Ribeiro, R. Audigier, and Q. C. Pham, “RIMOC, a feature to discriminate unstructured motions: application to violence detection for video-surveillance,” *Computer vision and image understanding*, vol. 144, pp. 121–143, 2016.

[20] V. E. M. Arceda, K. M. F. Fabián, P. C. L. Laura, J. J. R. Tito, and J. C. G. Cáceres, “Fast face detection in violent video scenes,” *Electronic Notes in Theoretical Computer Science*, vol. 329, pp. 5–26, 2016.

[21] J. Xie, W. Yan, C. Mu, T. Liu, P. Li, and S. Yan, “Recognizing violent activity without decoding video streams,” *Optik-International Journal for Light and Electron Optics*, vol. 127, no. 2, pp. 795–801, 2016.

[22] E. Y. Fu, H. Va Leong, G. Ngai, and S. Chan, “Automatic Fight Detection in Surveillance Videos,” in *Proceedings of the 14th International Conference on Advances in Mobile Computing and Multi Media*, 2016, pp. 225–234.

[23] R. Nar, A. Singal, and P. Kumar, “Abnormal activity detection for bank ATM surveillance,” in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 2042–2046.

[24] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, “Crowd Violence Detection Using Global Motion-Compensated Lagrangian Features and Scale-Sensitive Video-Level Representation,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2945–2956, Dec. 2017.

[25] E. Y. Fu, M. X. Huang, H. V. Leong, and G. Ngai, “Cross-Species Learning: A Low-Cost Approach to Learning Human Fight from Animal Fight,” in *2018 ACM Multimedia Conference on Multimedia Conference*, 2018, pp. 320–327.

[26] M. Coletto, C. Lucchese, and S. Orlando, "Do Violent People Smile: Social Media Analysis of their Profile Pictures," in *Companion of the The Web Conference 2018 on The Web Conference 2018*, 2018, pp. 1465–1468.

[27] L. Auria and R. Moro, "Advantages and Disadvantages of Support Vector Machines," *Credit Risk Assessment Revisited: Methodological Issues and Practical Implications*, pp. 49–68, 2007.

[28] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1–6.

[29] M. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," 2009.

[30] X. Li, Y. Huo, Q. Jin, and J. Xu, "Detecting Violence in Video using Subclasses," in *Proceedings of the 2016 ACM on Multimedia Conference*, 2016, pp. 586–590.

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

[32] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image and vision computing*, vol. 48, pp. 37–41, 2016.

[33] M. Al-Nawashi, O. M. Al-Hazaimeh, and M. Saraee, "A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments," *Neural Computing and Applications*, vol. 28, no. 1, pp. 565–572, 2017.

[34] C. Dhiman and D. K. Vishwakarma, "High Dimensional Abnormal Human Activity Recognition Using Histogram Oriented Gradients and Zernike Moments," in *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2017, pp. 1–4.

[35] D. Song, C. Kim, and S.-K. Park, "A multi-temporal framework for high-level activity analysis: Violent event detection in visual surveillance," *Information Sciences*, vol. 447, pp. 83–103, 2018.

[36] Q. Xia, P. Zhang, J. Wang, M. Tian, and C. Fei, "Real Time Violence Detection Based on Deep Spatio-Temporal Features," in *Chinese Conference on Biometric Recognition*, 2018, pp. 157–165.

[37] B. M. Peixoto, S. Avila, Z. Dias, and A. Rocha, "Breaking down violence: A deep-learning strategy to model and classify violence in videos," in *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 2018, p. 50.

[38] A. A. Mishra and G. Srinivasa, "Automated detection of fighting styles using localized action features," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 1385–1389.

[39] T. Agrawal, A. Kumar, and S. K. Saraswat, "Comparative analysis of convolutional codes based on ML decoding," in *2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS)*, 2016, pp. 41–45.

[40] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3D convolutional neural networks," in *International Symposium on Visual Computing*, 2014, pp. 551–558.

[41] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

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

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

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

[45] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4787–4797, Oct. 2018.

[46] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network," *Sensors*, vol. 19, no. 11, p. 2472, 2019.

[47] W. Lejmi, A. B. Khalifa, and M. A. Mahjoub, "Fusion Strategies for Recognition of Violence Actions," in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 2017, pp. 178–183.

[48] D. Maniry, E. Acar, F. Hopfgartner, and S. Albayrak, "A visualization tool for violent scenes detection," in *Proceedings of International Conference on Multimedia Retrieval*, 2014, p. 522.

[49] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 28–35.

[50] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *International conference on Computer analysis of images and patterns*, 2011, pp. 332–339.

[51] S. Blunsden and R. B. Fisher, "The BEHAVE video dataset: ground truthed video for multi-person behavior classification," *Annals of the BMVA*, vol. 4, no. 1–12, p. 4, 2010.

[52] R. B. Fisher, "The PETS04 surveillance ground-truth data sets," in *Proc. 6th IEEE international workshop on performance evaluation of tracking and surveillance*, 2004, pp. 1–5.

[53] S. Gaglio, G. L. Re, and M. Morana, "Human activity recognition process using 3-D posture data," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 5, pp. 586–597, 2014.

[54] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[55] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.