State-of-the-art violence detection techniques in video surveillance security systems: A systematic review

Batyrkhan Omarov [Corresp., 1, 2, 3, 4] , **Sergazi Narynov** [1] , **Zhandos Zhumanov** [1, 4] , **Aidana Gumar** [1, 5] , **Mariyam Khassanova** [1, 5]

[1] Alem Research, Almaty, Kazakhstan

[2] International University of Tourism and Hospitality, Turkistan, Kazakhstan

[3] Suleiman Demirel University, Almaty, Kazakhstan

[4] Al-Farabi Kazakh National University, Almaty, Kazakhstan

[5] Asfendiyarov Kazakh National Medical University, Almaty, Kazakhstan

Corresponding Author: Batyrkhan Omarov
Email address: Batyrkhan.Omarov2@kaznu.kz

We investigate and analyze methods to violence detection in this study to completely disassemble the present condition and anticipate the emerging trends of violence discovery research. In this systematic review, we provide a comprehensive assessment of the video violence detection problems that have been described in state-of-the-art researches. This work aims to address the problems as state-of-the-art methods in video violence detection, datasets to develop and train real-time video violence detection frameworks, discuss and identify open issues in the given problem. In this study, we analyzed 80 research papers that have been selected from 154 research papers after identification, screening, and eligibility phases. As the research sources, we used five digital libraries and three high ranked computer vision conferences that were published between 2015 and 2021. We begin by briefly introducing core idea and problems of video-based violence detection; after that, we divided current techniques into three categories based on their methodologies: conventional methods, end-to-end deep learning-based methods, and machine learning-based methods. Finally, we present public datasets for testing video based violence detectionmethods' performance and compare their results. In addition, we summarize the open issues in violence detection in videoand evaluate its future tendencies.

State-of-The-Art Violence Detection Techniques in Video Surveillance Security Systems: A Systematic Review

Batyrkhan Omarov[1-4], Sergazi Narynov[1], Zhandos Zhumanov[1,4], Aidana Gumar[1,5], and Mariyam Khassanova[1,5]

[1] Alem Research, Almaty, Kazakhstan;
[2] International University of Tourism and Hospitality, Turkistan, Kazakhstan;
[3] Suleiman Demirel University, Turkistan, Kazakhstan;
[4] Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[5] Asfendiyarov Kazakh National Medical University, Almaty, Kazakhstan

Corresponding Author:
Batyrkhan Omarov
Al-Farabi 71, Almaty, 050050, Kazakhstan
Email address: batyahan@gmail.com

Abstract

We investigate and analyze methods to violence detection in this study to completely disassemble the present condition and anticipate the emerging trends of violence discovery research. In this systematic review, we provide a comprehensive assessment of the video violence detection problems that have been described in state-of-the-art researches. This work aims to address the problems as state-of-the-art methods in video violence detection, datasets to develop and train real-time video violence detection frameworks, discuss and identify open issues in the given problem. In this study, we analyzed 80 research papers that have been selected from 154 research papers after identification, screening, and eligibility phases. As the research sources, we used five digital libraries and three high ranked computer vision conferences that were published between 2015 and 2021. We begin by briefly introducing core idea and problems of video-based violence detection; after that, we divided current techniques into three categories based on their methodologies: conventional methods, end-to-end deep learning-based methods, and machine learning-based methods. Finally, we present public datasets for testing video based violence detection methods' performance and compare their results. In addition, we summarize the open issues in violence detection in video and evaluate its future tendencies.

Keywords: Violence detection; Machine learning; Deep learning; Computer Vision; Artificial Intelligence; Video features; Datasets

1. INTRODUCTION

39 Surveillance and anomaly detection have become more important as the quantity of video data has
40 grown rapidly (Feng et. al., 2021). When compared to regular activity, such aberrant occurrences
41 are uncommon. As a result, creating automated video surveillance systems for anomaly detection
42 has become a need to reduce labor and time waste. Detecting abnormalities in films is a difficult
43 job since the term "anomaly" is often imprecise and poorly defined (Yang et. al., 2018). They differ
44 greatly depending on the conditions and circumstances in which they occur. Bicycling on a
45 standard route, for example, is a typical activity, but doing so in a walk-only lane should be noted
46 as unusual. The uneven internal occlusion is a noteworthy, yet difficult to explain characteristic of
47 the abnormal behavior. Furthermore, owing to its large dimensionality, resolution, noise, and
48 rapidly changing events and interactions, video data encoding and modeling are more challenging.
49 Other difficulties include lighting changes, perspective shifts, camera movements, and so on
50 (Yazdi et. al., 2018).

51 Violence detection is one of the most crucial elements of video-based anomaly detection (Khan
52 et. al., 2019). The usage of video cameras to monitor individuals has become essential due to the
53 rise in security concerns across the globe, and early detection of these violent actions may
54 significantly minimize the dangers. A violence detection system's primary goal is to identify some
55 kind of aberrant behavior that fits under the category of violence (Mabrouk et. al., 2019).

56 If an event's conduct differs from what one anticipates, it is considered violent. A person striking,
57 kicking, lifting the other person, and so on are examples of such anomalies (Shao et. al., 2017).
58 Item in an unusual place, odd motion patterns such as moving in a disorganized way, abrupt
59 motions, fallen objects are all examples of violent occurrences (Munn et. al., 2018).

60 Since human monitoring of the complete video stream is impractical owing to the repetitive nature
61 of the work and the length of time required, automated identification of violent events in real-time
62 is required to prevent such incidents (Tripathi et. al., 2018).

63 Many scholars considered various methods to improve violence detection performance. Using a
64 comprehensive literature review, various techniques of detecting violence from surveillance
65 camera videos are examined and addressed in depth in this study.

66 The main goal of this review is to provide an in-depth, systematic overview of the techniques for
67 detecting violence in video. Various techniques of detecting violence in video and aggressive
68 behavior have been developed during the past decade. These techniques must be classified,
69 analyzed, and summarized. To perform a systematic literature review, we created basic search
70 phrases to find the most relevant studies on the detection of violent behavior accessible in five
71 digital libraries as ScienceDirect, IEEEXplore digital library, Springer, Wiley, Scopus databases,
72 and well-known conferences in the computer vision area as Conference on Computer Vision and
73 Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), and
74 European Conference on Computer Vision (ECCV).

75 Research highlights of this systematic review are described as follows:

76   • Review of state-of-the-art violence detection methods highlighting their originality, key
77   features, and limitations.

78   •   Study of ranking and importance of video feature descriptors for detecting violence in
79   video.
80   •   Exploration of datasets and evaluation criteria for violence detection in video.
81   •   Discussion of limitations, challenges, and open issues of the video-based violence detection
82   problem.
83   The remainder of this review is split into five parts. The research methodology of the current
84   review is described in Section II. The fundamental idea and main concepts of violence detection
85   in videos are discussed in Section III. The methods of violence detection in videos are explored
86   in-depth in Section IV. Video feature descriptors and their significance are described in Section
87   V. Worldwide datasets to train the models for violence detection are discussed in Section VI.
88   Evaluation metrics that were used to test violence detection methods are described in Section VII.
89   Challenges and open issues in violence detection are discussed in Section VIII. The last section
90   concludes the review by discussing trends, future perspectives, and open problems of violence
91   detection.
92

93 ## 2. RESEARCH METHODOLOGY

94   Both qualitative and quantitative analytic techniques were integrated and used in this systematic
95   literature review (Ramzan et. al., 2019; Lejmi et. al., 2019). Table 1 demonstrates inclusion and
96   exclusion criterias of the collected studies. We included two criteria as inclusion and six criteria
97   as exclusion criteria.
98

99   **Table 1. Inclusion and exclusion criteria.**
100

101   Fig. 1 illustrates the stages of the review as well as the number of articles that were included and
102   eliminated. For a collection of articles, search terms and query strings were determined. The query
103   string consists of three search terms: "Video" and "Violence detection" with the logic operator
104   "AND" between them. Science Direct, IEEE Xplore, Springer, and Wiley libraries, and "Scopus"
105   abstract and citation database were utilized. In addition, we asked for publications from CVPR,
106   ICCV, and ECCV conferences. Articles published between 2015 and 2021 were considered in our
107   study. 58 recordings were eliminated during the screening step, 16 records were eliminated during
108   the Eligibility stage due to a lack of complete texts, duplication. As a result of the article collection,
109   80 research articles were included in the systematic literature review.
110

111

112   Figure 1. Systematic literature review flowchart.
113

114   In our systematic literature review, we put four research questions. Table 2 demonstrates research
115   questions and their motivations.
116

117   **Table 2. Research questions and their motivations.**

2.1 Data Analysis

In this subsection, we provide general analysis to the obtained results. Fig. 2 demonstrates year-wise distribution of the papers that dedicated to violence detection. Fig. 2a illustrates distribution of violence detection papers from January, 2016 to November, 2021. As the figure shows, the interest to the given problem increases year by year. Fig. 2b presents distribution of applied methods in the selected papers. As it is illustrated in the figure, in 2016-2017 machine learning methods were popular in video violence detection problem. Moreover, we can observe decreasing of conventional methods usage, and increasing trend of deep learning based techniques.

Fig. 3 demonstrates percentage of each method usage. SVM is consistently applied in the detection of violence in video occupying 24% of all methods used. Conventional methods that were used in 2015 to 2018, take about one fifth of all the applied methods. In machine learning, four algorithms frequently used in violence detection, in detail k nearest neighbors (2%), Adaptive boosting (4%), Random forest (7%), and k-means (2%). Increasing of deep learning techniques in video based violence detection can be associated with the increasing of computational performance of equipment. 43% of all the applied methods use deep learning for violence detection problem. Convolutional neural networks is the most applied method for the given problem.

a) Year-wise violence detection papers distribution   b) Year-wise violence detection papers by applied methods

Figure 2. Year-wise violence detection papers distribution

Figure 3. Distribution of violence detection methods

3. CONCEPTS

The main objective of a violence detection system is to identify events in real-time so that hazardous situations may be avoided. It is, nevertheless, essential to comprehend certain key principles. Fig. 4 depicts the fundamental stages of video-based violence detection methods.

3.1 Action Recognition

Action recognition is a technology that can identify human actions. Human activities are categorized into four groups based on the intricacy of the acts and the number of bodily parts engaged in the action. Gestures, actions, interactions, and group activities are the four categories (Aggarwal et. al., 2011). A gesture is a series of motions performed with the hands, head, or other body parts to convey a certain message. A single person's actions are a compilation of numerous gestures. Interactions are a set of human activities involving at least two people. When there are two actors are involved, one should be a human and the other may be a human or an object. When

156 there are more than two participants and one or more interacting objects, group activities involve
157 a mix of gestures, actions, or interactions (Aggarwal et. al., 2011).
158
159 Figure 4. Fundamental stages of video-based violence detection.
160
161 **3.2 Violence Detection**
162 The detection of violence is a specific issue within the larger topic of action recognition. The goal
163 of violence detection is to identify whether or not violence happens in a short amount of time
164 automatically and efficiently. Automatic video identification of human activities has grown more
165 essential in recent years for applications such as video surveillance, human-computer interaction,
166 and video retrieval based on content (Poppe, 2010; Sun & Liu, 2013).
167 The goal of violence detection is to identify whether or not violence happens automatically and
168 effectively. In any case, detecting violence is a tough task in and of itself, since the notion of
169 violence is subjective. Because it possesses features that distinguish it from generic acts, violence
170 detection is a significant problem not just at the application level but also at the research level.
171
172 # 4. CLASSIFICATION OF VIOLENCE DETECTION TECHNIQUES
173 In everyday life, violence is defined as suspicious occurrences or actions. The use of computer
174 vision to recognize such actions in surveillance cameras has become a popular issue in the area of
175 action recognition (Naik & Gopalakrishna, 2017). Scientists presented various approaches and
176 methods for detecting violent or unusual occurrences, citing the fast rise in crime rates as an
177 example of the need for more efficient identification. Various methods for detecting violence have
178 been developed in past few years. Based on the classifier employed, violence detection methods
179 are divided into three categories: violence detection using machine learning, violence detection
180 using SVM, and violence detection using deep learning. Because SVM and deep learning are
181 extensively employed in computer vision, they are categorized independently. Tables explain the
182 specifics of each technique. The techniques are given in the order in which they were developed.
183 A methodology for detecting objects and a method for extracting features are also discussed.
184
185
186 **4.1 Violence Detection Using Machine Learning Techniques**
187 In this subsection, we review violence detection techniques that applied classical machine learning
188 techniques. In Table 3, we summarize different classification techniques for violence detection in
189 videos by indicating object detection, feature extraction, classification, the applicability of the
190 methods for different types of scenes, and their evaluation parameters when using in different
191 datasets. Further, we describe each technique in detail.
192
193 **Table 3. Violence detection techniques that use machine learning.**
194

195   In the field of computer vision, action recognition has now become a relevant research area.
196   Nevertheless, most researches have concentrated on relatively basic activities such as clapping,
197   walking, running, and so on. The identification of particular events with immediate practical
198   application, such as fighting or general violent conduct, has received much less attention. In certain
199   situations, such as prisons, mental institutions, or even camera phones, video surveillance may be
200   very helpful. A new technique for detecting violent sequences was suggested by Gracia et al.. To
201   distinguish between fighting and non-fighting episodes, features derived from motion blobs are
202   utilized. The proposed method was assessed using three different datasets as "Movies" dataset with
203   200 video clips (Bermejo et. al., 2011), the "Hockey fight" dataset that consists of 1000 video clips
204   (Nievas et. al., 2011), and the UCF-101 dataset of realistic action videos collected from Youtube
205   (Soomro et. al., 2012). The proposed method was compared with other five related methods as
206   Bag of Words (BoW) (Wang et. al., 2021) using scale-invariant feature transform (MoSIFT) (Chen
207   & Hauptmann, 2009) and STIP (Ushapreethi & GG, 2020) features, Violent Flows (ViF) method
208   (Souza & Pedrini, 2017), Local Motion method (Zhang et. al., 2019), also variant v-1 and variant
209   v-2 methods that applied KNN, AdaBoost, and Random Forest classifiers. Although the proposed
210   technique falls short from a perspective of performance, it has a much quicker calculation time,
211   making it suitable for practical uses.
212   Automatically detecting aggressive behaviors in video surveillance situations such as train
213   stations, schools, and mental institutions is critical. Previous detection techniques, on the other
214   hand, often extract descriptors surrounding spatiotemporal interesting spots or statistic
215   characteristics in motion areas, resulting in restricted capacities to identify video-based violent
216   activities efficiently. Wang et. al. present a new technique for detecting violent sequences to solve
217   this problem (Zhou et. al., 2017). To begin, the motion areas are divided into segments based on
218   the distribution of optical flow fields. Second, we suggest extracting two types of low-level
219   characteristics to describe the emergence and dynamics of violent behaviors in the motion areas.
220   The Local Histogram of Oriented Gradient (LHOG) descriptor (Dalal and Triggs, 2005) derived
221   from RGB pictures and the LHOF descriptor (Dalal et. al., 2006) extracted from optical flow
222   images are the suggested low-level features. Finally, to remove duplicate information, the collected
223   features are coded using the Bag of Words (BoW) model, and a specific-length vector is produced
224   for each video clip. Finally, SVM is used to classify the video-level vectors. The suggested
225   detection technique outperforms the prior approaches in three difficult benchmark datasets,
226   according to experimental findings.
227   We chose to start at a basic level to describe what is often present in film with violent human
228   behaviors: jerky and unstructured motion, that is due to the fact that aggressive occurrences are
229   difficult to quantify owing to their unpredictability and sometimes need high-level interpretation.
230   In order to capture its structure and distinguish the unstructured movements, a new problem-
231   specific Rotation-Invariant feature modeling MOtion Coherence (RIMOC) was suggested (Ribeiro
232   et. al., 2016). It is based on eigenvalues calculated locally and densely from second-order statistics
233   of Histograms of Optical Flow vectors from successive temporal instants, then embedded into a
234   spheric Riemannian manifold. In a poorly supervised way, the proposed RIMOC feature is utilized

235   to develop statistical models of normal coherent movements. Events with irregular mobility may
236   be identified in space and time using a multi-scale approach combined with an inference-based
237   method, making them ideal candidates for aggressive events. There is no special dataset available
238   for violence and aggressive behavior detection. A big dataset is produced for this goal, which
239   comprises of sequences from two distinct sites: an in-lab fake train and a genuine underground
240   railway line, real train, and then four datasets are formed: fake train, real train, real train station,
241   and real-life settings. These datasets are used in the trials, and the findings indicate that the
242   suggested approach outperforms all state-of-the-art methods in terms of ROC per frame and false-
243   positive rate.
244   Yao et al. present a multiview fight detection technique based on optical flow statistical features
245   and random forest (Yao et. al., 2021). This technique may provide fast and reliable information to
246   cyber-physical monitoring systems. Motion Direction Inconsistency (MoDI) and Weighted
247   Motion Direction Inconsistency (WMoDI), two new descriptors, are developed to enhance the
248   performance of current techniques for films with various filming perspectives and to address
249   misjudgment on nonfighting activities like jogging and chatting. The motion regions are first
250   marked using the YOLO V3 method, and then the optical flow is calculated to retrieve descriptors.
251   Finally, Random Forest is utilized to classify data using statistical descriptor features. The
252   experiments were performed using CASIA Action Dataset (Wang et. al., 2007) and the UT-
253   Interaction Dataset (Zhang et. al., 2017). All films of fighting, as well as 15 additional videos in
254   five categories, were chosen from the CASIA Action Dataset. The findings demonstrated that the
255   proposed approach improves violence detection accuracy and reduces the incidence of missing and
256   false alarms, and it is robust against films with various shooting perspectives.
257   Fast face detection (Arceda et. al., 2016) is developed to accomplish the objective of identifying
258   faces in violent videos to improve security measures. For the initial step of violent scene
259   identification, the authors utilized the ViF descriptor (Keçeli et. al., 2017) in conjunction with
260   Horn-Schunck (Keçeli et. al., 2017). Then, to enhance the video quality, the non-adaptive
261   interpolation super-resolution algorithm was used, followed by the firing of the Kanade-Lucas-
262   Tomasi (KLT) face detector (Alex Net, 2021). The authors used CUDA to parallelize the super-
263   resolution and face detection algorithms in order to achieve a very fast processing time. The Boss
264   Dataset (Bas et. al., 2011) was utilized in the tests, as well as a violence dataset based on security
265   camera footage. Face detection yields encouraging results in terms of area under the curve (AUC)
266   and accuracy.
267   For years, computer vision researchers have been exploring how to identify violence. Prior studies,
268   on the other hand, are either shallow, as in the categorization of short-clips and the single scenario,
269   or undersupplied, as in the single modality and multimodality based on hand-crafted
270   characteristics. To address this issue, the XD-Violence dataset, a large-scale and multi-scene
271   dataset with a total length of 217 hours and 4754 untrimmed films with audio signals and poor
272   labels was proposed (Wu et. al., 2020). Then, to capture different relations among video snippets
273   and integrate features, a neural network with three parallel branches was proposed: a holistic
274   branch that catches long-range dependencies using similarity prior, a localized branch that captures

275 local positional relations utilizing proximity prior, and a score branch that dynamically captures
276 the closeness of predicted score. In addition, to fulfill the requirements of online detection, the
277 proposed approach incorporates an approximator. Authors use the frame-level precision-recall
278 curve (PRC) and corresponding area under the curve (average precision, AP) (Wu et. al., 2020)
279 instead of the receiver operating characteristic curve (ROC) and corresponding AUC (Yoganand
280 & Kavida, 2018; Xie et. al., 2016) because AUC typically shows an optimistic result when dealing
281 with class-imbalanced data, whereas PRC and AP focus on positive samples (violence). The
282 proposed approach beats other state-of-the-art algorithms in the publicly available dataset created
283 by authors. Furthermore, numerous experimental findings indicate that multimodal (audio-visual)
284 input and modeling connections have a beneficial impact.

285 Most conventional activity identification techniques' motion target detection and tracking
286 procedures are often complex, and their applicability is limited. To solve this problem, a fast
287 method of violent activity recognition is introduced which is based on motion vectors (Xie et. al.,
288 2016). First and foremost, the motion vectors were directly retrieved from compressed video
289 segments. The motion vectors' characteristics in each frame and between frames were then
290 evaluated, and the Region Motion Vectors (RMV) descriptor was produced. To classify the RMV
291 to identify aggressive situations in movies, a SVM classifier with radial basis kernel function was
292 used in the final step. In order to evaluate the proposed method, the authors created VVAR10
293 dataset that consists of 296 positive samples and 277 negative samples by sorting video clips from
294 UCF sports (Xie et. al., 2016), UCF50 (Reddy & Shah, 2013), HMDB51 (Kuehne et. al., 2011)
295 datasets. Experiments have shown that the proposed method can detect violent scenes with 96.1%
296 accuracy in a short amount of time. That is why the proposed method can be used in embedded
297 systems.

298 Most of the research in the field of action recognition has concentrated on people identification
299 and monitoring, loitering, and other similar activities, while identification of violent acts or
300 conflicts has received less attention. Local spatiotemporal feature extractors have been explored
301 in previous studies; nevertheless, they come with the overhead of complicated optical flow
302 estimates. Despite the fact that the temporal derivative is a faster alternative to optical flow, it
303 produces a low-accuracy and scale-dependent result when used alone (Li et. al., 2018). As a result,
304 a cascaded approach of violence detection was suggested (Febin et. al., 2020), based on motion
305 boundary SIFT (MoBSIFT) and a movement filtering method. The surveillance films are examined
306 using a movement filtering algorithm based on temporal derivatives in this approach, which avoids
307 feature extraction for most peaceful activities. Only filtered frames may be suitable for feature
308 extraction. Motion boundary histogram (MBH) is retrieved and merged with SIFT (Lowe, 2004)
309 and histogram of optical flow feature to create MoBSIFT descriptor. The models were trained
310 using MoBSIFT and MPEG Flow (MF) (Kantorov & Laptev, 2014) descriptors using AdaBoost,
311 RF, and SVM classifiers. Because of its great tolerance to camera motions, the suggested
312 MoBSIFT surpasses current techniques in terms of accuracy. The use of movement filtering in
313 conjunction with MoBSIFT has also been shown to decrease time complexity.

314   In computer vision, Lagrangian theory offers a comprehensive set of tools for evaluating non-
315   local, long-term motion information. Authors propose a specialized Lagrangian method for the
316   automatic identification of violent situations in video footage based on this theory (Senst et. al.,
317   2017). The authors propose a new feature based on a spatio-temporal model that utilizes
318   appearance, background motion correction, and long-term motion information and leverages
319   Lagrangian direction fields. They use an expanded bag-of-words method in a late-fusion way as a
320   classification strategy on a per-video basis to guarantee suitable spatial and temporal feature sizes.
321   Experiments were conducted in three datasets as "Hockey Fight" (Nievas et. al., 2011), "Violence
322   in Movies" (Bermejo et. al., 2011), "Violent Crowd" (Hassner et. al., 2012), and "London
323   Metropolitan Police (London Riots 2011)" (Cheng & Williams, 2012) datasets. Multiple public
324   benchmarks and non-public, real-world data from the London Metropolitan Police are used to
325   verify the proposed system. Experimental results demonstrated that the implementation of
326   Lagrangian theory is a useful feature in aggressive action detection and the classification efficiency
327   rose over the state-of-the-art techniques like two-stream convolutional neural network (CNN,
328   ConvNet), ViF, HoF+BoW with STIP, HOG+BoW with STIP, etc. in terms of accuracy and ROC-
329   AUC measure.
330

4.2 Violence Detection Techniques Using SVM

332   The methods for detecting violence using the SVM as a classifier are described in-depth here.
333   A collection of SVM based violent incident recognition methods is shown in Table 4. SVM is a
334   supervised learning method that is used to tackle classification issues. We display data on (number
335   features) dimension space in SVM and distinguish between two groups. SVM is a popular
336   technique in computer vision since it is robust and takes quantitative information into account. It
337   is used to do binary classification jobs. Kernel is the foundation of SVM. Kernel is a function that
338   transforms data into a high-dimensional space in which the issue may be solved. The lack of
339   transparency in the findings is a significant drawback of SVM (Auria & Moro, 2007). SVM-based
340   techniques for detecting violence are now described in full separately.
341

342   **Table 4. Violence Detection Techniques Using SVM.**
343

344   A new method for identifying school violence was proposed (Ye et. al., 2020). This technique uses
345   the KNN algorithm to identify foreground moving objects and then uses morphological processing
346   methods to preprocess the identified targets. Then, to optimize the circumscribed rectangular frame
347   of moving objects, a circumscribed rectangular frame integrating technique was proposed. To
348   explain the distinctions between school violence and everyday activities, rectangular frame
349   characteristics and optical-flow features were retrieved. To decrease the feature dimension, the
350   Relief-F and Wrapper algorithms were applied. SVM is used as a classifier, and 5-fold cross-
351   validation was conducted. The results show 94.4 percent precision and 89.6 percent accuracy. In
352   order to improve recognition performance, a DT–SVM two-layer classifier is created. Authors
353   utilized boxplots to identify certain DT layer characteristics that can differentiate between

354   everyday activities and physical violence. The SVM layer conducted categorization for the
355   remaining activities. The accuracy of this DT–SVM classifier was 97.6 percent, while the precision
356   was 97.2 percent, indicating a considerable increase.
357   Surveillance systems are grappling with how to identify violence. However, it has not received
358   nearly as much attention as action recognition. Existing vision-based techniques focus mostly on
359   detecting violence and make little attempt to pinpoint its location. To tackle this problem, Zhange
360   et. al. presented a quick and robust method for identifying and localizing violence in surveillance
361   situations to address this issue (Zhang et al., 2016). A Gaussian Model of Optical Flow (GMOF)
362   is suggested for this purpose in order to extract potential violent areas, which are adaptively
363   modeled as a departure from the usual crowd behavior seen in the picture. Following that, each
364   video volume is subjected to violence detection by intensively sampling the potential violent areas.
365   The authors also propose a new descriptor called the Orientation Histogram of Optical Flow
366   (OHOF), which is input into a linear SVM for classification to differentiate violent events from
367   peaceful ones. Experimental results on violence video datasets like "Hockey" (Nievas et. al.,
368   2011), "BEHAVE" (Blunsden & Fisher, 2010), "CAVIAR" (Caviar, 2004) have shown the
369   superiority of the proposed methodology over the state-of-the-art descriptors like MoSIFT and
370   SIFT, HOG, HOF, and Combination of HOG and HOF (HNF), in terms of detection accuracy,
371   AUC-ROC, and processing performance, even in crowded scenes.
372   With more and more surveillance cameras deployed nowadays, the market need for smart violence
373   detection is steadily increasing, despite the fact that it is still a challenging subject in the study. In
374   order to recognize violence in videos in a realistic manner, a novel feature extraction method
375   named Oriented VIolent Flows (OViF) was proposed by Gao et. al., (Gao et. al., 2016) . In
376   statistical motion orientations, the proposed method fully exploits the motion magnitude change
377   information. The features are selected using AdaBoost, and the SVM classifier is subsequently
378   trained on the features. Experiments are carried out on the "Hockey" and "Violent-Flow" (Xu et.
379   al., 2018) datasets to assess the new approach's performance. The findings indicate that the
380   suggested technique outperforms the baseline methods LTP and ViF in terms of accuracy and
381   AUC. Furthermore, feature and multi-classifier combination methods have been shown to help
382   improve the performance of the violence detector. The experiment results demonstrate that the
383   combination of ViF and OViF using AdaBoost with a combination of Linear-SVM surpasses the
384   state-of-the-art on the Violent-Flows database. The final best violence detection rates are 87.50%
385   and 88.00% on Hockey Fight and Violent-Flows separately using ViF + OViF with Adaboost +
386   SVM.
387   One of the most important stages in the development of machine learning applications is data
388   representation. Data representation that is efficient aids in better classification across classes.
389   Deepak et. al. investigate Spatio-Temporal Autocorrelation of Gradients (STACOG) as a
390   handmade feature for extracting violent activity characteristics from surveillance camera videos
391   (Deepak et. al., 2020). The proposed strategy is divided into two stages: (1) Extraction of STACOG
392   based Features features; (2) Discriminative learning of violent/non-violent behaviors using an
393   SVM Classifier. Two well-known datasets were used to test the proposed approach. The Hockey

394 fight dataset (Nievas et. al., 2011) contains 1000 video clips and the Crowd Violence Dataset. The
395 proposed "STACOG features + SVM" model shown 91.38% accuracy in violence detection
396 overcoming state-of-the-art methods like HOF+BoW, HNF+BoF, ViF+SVM, BiLSTM, GMOF,
397 and others.
398 In video processing, aggression detection is critical, and a surveillance system that can operate
399 reliably in an academic environment has become a pressing requirement. To solve this problem, a
400 novel framework for an automatic real-time video-based surveillance system is proposed (Al-
401 Nawashi et. al., 2017). The proposed system is divided into three phases during the development
402 process. The first stage is preprocessing stage that includes abnormal human activity detection and
403 content-based image retrieval (CBIR) in the event that the system identifies unusual student
404 behavior. In the first stage, students are registered by entering their personal data including first
405 name, second name, birthday, course, student id card, and photos. The entered data is stored in a
406 central database for conducting a search when abnormal actions are detected. The video is then
407 turned into frames in the second step. Motion objects are detected using a temporal-differencing
408 method, and motion areas are identified using the Gaussian function. Furthermore, a form model
409 based on the OMEGA equation is employed as a filter for identified items, whether human or non-
410 human. SVM is used to classify human behaviors into normal and abnormal categories. When a
411 person engages in abnormal behavior, the system issues an automated warning. It also adds a
412 method to get the identified item from the database using CBIR for object detection and
413 verification. Finally, a software-based simulation using MATLAB is performed, and experimental
414 findings indicate that the system performs simultaneous tracking, semantic scene learning, and
415 abnormality detection in an academic setting without the need of humans.
416 Kamoona et. al. proposes a model-based method for anomaly identification for surveillance video.
417 There are two stages to the system (Kamoona et. al., 2019). Multiple handcrafted features have
418 been presented on this platform. Deep learning techniques have also been used to extract spatial-
419 temporal characteristics from video data, such as C3D features (Sultani et. al., 2018), as well as
420 anomaly detection using SVM. The next phase is behavior modeling. In this phase, SVM is trained
421 using a Bag of Visual Word (BOVW) to learn the typical behavior representation.
422 Song et. al. proposes a new framework for high-level activity analysis based on late fusion and
423 multi-independent temporal perception layers, which is based on late fusion (Song et. al., 2018).
424 It is possible to manage the temporal variety of high-level activities using this approach. Multi-
425 temporal analysis, multi-temporal perception layers, and late fusion are all part of the framework.
426 Based on situation graph trees (SGT) and SVM, authors create two kinds of perception layers
427 (SVMs). Through a phase of late fusion, the data from the multi-temporal perception layers are
428 fused into an activity score. To test the proposed method, the framework is applied to the detection
429 of violent events by visual observation. The experiments are conducted applying three well-known
430 databases: BEHAVE (Blunsden & Fisher, 2010), NUS–HGA (Zhuang et. al., 2017), and a number
431 of YouTube videos depicting real-life situations. The tests yielded an accuracy of 70.2% (SVM),
432 and 87.2% (SGT) in different datasets, demonstrating how the proposed multi-temporal technique
433 outperforms single-temporal approaches.

434 (Vashistha et. al., 2018) utilized Linear SVM to categorize incoming video as violent or non-
435 violent, extracting important characteristics like centroid, direction, velocity, and dimensions.
436 Their approach took into account two feature vectors, i.e. ViF and the Local Binary Pattern (LBP).
437 Because calculating LBP or ViF individually it takes less time than combining these feature
438 vectors, their study found that combining LBP and ViF did not offer substantial direction for future
439 development.

4.3 Violence Detection Techniques Using Deep Learning

442 The methods for detecting violence that utilizes deep learning algorithms in the suggested
443 frameworks are described in-depth here. Convolutional Neural Netowks (CNN) (Zhang et. al.,
444 2019) and its imrovemebts are widely used in violence detection in video. Table 5 shows a
445 collection of recognition techniques that are based on deep learning. Neural networks are the
446 foundation of deep learning. Using additional convolutional layers, the method is utilized to
447 categorize the violent recognition based on the data set and retrieved features. Now, techniques for
448 detecting violence that utilize deep learning algorithms are discussed in depth individually.
449 While most studies have focused on the issue of action recognition, fighting detection has received
450 much less attention. This skill may be very valuable. To build complicated handmade
451 characteristics from inputs, most techniques require on domain expertise. Deep learning methods,
452 on the other hand, may operate directly on raw inputs and extract necessary features automatically.
453 As a result, Ding et. al. created a new 3D ConvNets approach for video violence detection, which
454 does not need any previous information (Ding et. al., 2014). The convolution on the collection of
455 video frames is computed using a 3D CNN, and therefore motion information is retrieved from the
456 input data. The back-propagation technique is used to obtain gradients and the model has been
457 trained to apply supervised learning. Experimental validation was carried out in the context of the
458 "Hockey fights" (Nievas et. al., 2011) dataset to assess the approach. The findings indicate that
459 the approach outperforms manual features in terms of performance.

Table 5. Violence detection using deep learning techniques.

463 Campus violence is a worldwide social phenomenon that is the most dangerous sort of school
464 bullying occurrence. There are various possible strategies to identify campus violence as AI and
465 remote monitoring capabilities advance, such as video-based techniques. Ye et. al. combine visual
466 and audio data for campus violence detection (Ye et. al., 2021). Role-playing is used for campus
467 violence data collection, and 4096-dimension feature vectors are extracted from every 16 frames
468 of video frames. For feature extraction and classification, the 3D CNN is used, and overall
469 precision of 92.00 percent is attained. Three speech emotion datasets are used to extract mel-
470 frequency cepstral coefficients (MFCCs) as acoustic features: CASIA dataset (Wang et. al., 2007)
471 that has 960 samples, Finnish emotional dataset (Vaaras et. al., 2021) that consists of 132 samples,
472 and Chinese emotional dataset (Poria et. al., 2018) that has 370 samples.

473 An enhanced Dempster–Shafer (D–S) algorithm is proposed to handle the problem of evidence
474 dispute. As a result, recognition accuracy reached 97%.
475 To address the issue of large-scale visual place identification, the NetVLAD architecture is
476 presented, where the goal is to rapidly and correctly identify the location of a supplied query image
477 (Arandjelovic et. al., 2016). NetVLAD is a CNN-based approach for weakly supervised place
478 recognition. In this work, three major contributions are presented. First, for the location
479 recognition problem, CNN architecture is created that can be trained in a direct end-to-end way.
480 The central feature of this approach, NetVLAD, is a novel generalized VLAD layer inspired by
481 the widely used picture format "Vector of Locally Aggregated Descriptors (VLAD)." The layer
482 may be easily integrated into any CNN model and trained using backpropagation. The second
483 contribution is the construction of a training method based on a novel weakly supervised ranking
484 loss, to learn architectural parameters in an end-to-end way using Google Street View Time
485 Machine pictures showing the same locations over time. Finally, using the Pittsburgh (Torii et. al.,
486 2013) and Tokyo 24/7 (Torii et. al., 2015) datasets, the authors demonstrated that the proposed
487 architecture outperforms non-learned image representations and off-the-shelf CNN descriptors on
488 two difficult place recognition benchmarks, and outperforms current state-of-the-art image
489 representations on standard image retrieval benchmarks.
490 A real-time violence detection system is presented (Fenil et. al., 2019), which analyzes large
491 amounts of streaming data and recognizes aggression using a human intelligence simulation. The
492 system's input is a massive quantity of real-time video that feeds from various sources, which are
493 analyzed using the Spark framework. The frames are split and the characteristics of individual
494 frames are retrieved using the HOG function in the Spark framework. The frames are then labeled
495 based on characteristics such as the violence model, human component model, and negative model,
496 which are trained using the BDLSTM network for violent scene detection. The data may be
497 accessed in both directions via the bidirectional LSTM. As a result, the output is produced in the
498 context of both past and future data. The violent interaction dataset (VID) is used to train the
499 network, which contains 2314 movies with 1077 fights and 1237 no-fights. The authors also
500 generated a dataset of 410 video episodes with neutral scenes and 409 video episodes with
501 violence. The accuracy of 94.5% in detecting violent behavior validates the model's performance
502 and demonstrates the system's durability.
503 Mu et. al. are presented a violent scene identification method based on acoustic data from video
504 (Mu et. al., 2016). CNN in two ways: as a classifier and as a deep acoustic feature extractor. To
505 begin, the 40-dimensional Mel Filter-Bank (MFB) is used as the CNN's input feature. The video
506 is then divided into little pieces. To investigate the local features, MFB features are split into three
507 feature maps. Then CNN is utilized to represent features. The CNN-based features are applied to
508 construct SVM classifiers. Then the violent scene detection process is applied to each frame of
509 video. After that, detection is generated by applying maximum or minimum pooling to the segment
510 level. Experiments are conducted using the MediaEval dataset (Demarty et. al., 2014), and the
511 findings indicate that the proposed approach outperforms the fundamental techniques in terms of
512 average precision: audio alone, visual solely, and audio learned fusion and visual.

513 A new deep violence detection approach based on handcrafted techniques' distinctive
514 characteristics was presented (Mohtavipour et. al., 2021). These characteristics are linked to
515 appearance, movement speed, and representative images, and they are supplied to a CNN as
516 spatial, temporal, and spatiotemporal streams. With each frame, the spatial stream teaches the
517 neural network how to recognize patterns in the surroundings. With a modified differential
518 magnitude of optical flow, the temporal stream included three successive frames to learn motion
519 patterns of aggressive behavior. Furthermore, the authors developed a discriminative feature with
520 a new differential motion energy picture in the spatio-temporal stream to make violent behaviors
521 more understandable. By combining the findings of several streams, this method includes many
522 elements of aggressive conduct. The proposed CNN network was trained using three datasets:
523 Hockey (Nievas et. al., 2011), Movie (Bermejo et. al., 2011), and ViF (Rota et. al., 2015). The
524 proposed method beat state-of-the-art approaches in terms of accuracy and processing time.
525 Sudhakaran and Lanz proposed a deep neural network for detecting violent scenes in videos
526 (Sudhakaran & Lanz, 2017). To extract frame-level characteristics from a video, a CNN is applied.
527 The frame-level characteristics are then accumulated using LSTM that uses a convolutional gate.
528 The CNN, in combination with the ConvLSTM, can capture localized spatio-temporal
529 characteristics, allowing for the analysis of local motion in the video. The paper also proposed
530 feeding the model neighboring frame differences as input, pushing it to encode the changes in the
531 video. In terms of recognition accuracy, the presented feature extraction process is tested on three
532 common benchmark datasets as "Hockey" (Nievas et. al., 2011), "Movies" (Bermejo et. al., 2011),
533 and "Violent-Flows" (Xu et. al., 2018). Findings were compared to those produced using state-of-
534 the-art methods. It was discovered that the suggested method had a promising capacity for
535 identifying violent films prevailing state-of-the-art methods as three streams + LSTM, ViF, and
536 ViF+OViF.
537 To identify violent behaviors of a single person, an ensemble model of the Mask RCNN and LSTM
538 was proposed (Naik & Gopalakrishna, 2021). Initially, human key points and masks were
539 extracted, and then temporal information was captured. Experiments have been performed in
540 datasets as Weizmann (Blank et. al., 2005), KTH (Schuldt et. al., 2004), and own Dataset
541 respectively. The results demonstrated that the proposed model outperforms individual models
542 showing a violence detection accuracy rate of 93.4% in its best result. The proposed approach is
543 more relevant to the industry, which is beneficial to society in terms of security.
544 Typical approaches depend on hand-crafted characteristics, which may be insufficiently
545 discriminative for the job of recognizing violent actions. Inspired by the good performance of deep
546 learning-based approaches, propose a novel method for human violent behavior detection in videos
547 by incorporating trajectory and deep CNN, that includes the advantage of hand-crafted features
548 and deep-learned features (Meng et. al., 2021). To assess the proposed method, tests on two distinct
549 violence datasets are performed: "Hockey Fights" (Nievas et. al., 2011) and "Crowd Violence"
550 (Song et. al., 2019) dataset. On these datasets, the findings show that the proposed approach
551 outperforms state-of-the-art methods like HOG, HOF, ViF, and others.

552 Rendón-Segador et. al. present a new approach for determining whether a video has a violent scene
553 or not, based on an adapted 3D DenseNet, for a multi-head self-attention layer, and a bidirectional
554 ConvLSTM module that enables encoding relevant spatio-temporal features (Rendón-Segador et.
555 al., 2021). In addition, an ablation analysis of the input frames is carried out, comparing dense
556 optical flow and neighboring frames removal, as well as the effect of the attention layer, revealing
557 that combining optical flow and the attention mechanism enhances findings by up to 4.4 percent.
558 The experiments were performed using four datasets, exceeding state-of-the-art methods, reducing
559 the number of network parameters needed (4.5 million), and increasing its efficiency in test
560 accuracy (from 95.6 percent on the most complex dataset to 100 percent on the simplest), and
561 inference time (from 95.6 percent on the most complex dataset to 100 percent on the simplest
562 dataset) (less than 0.3 s for the longest clips).
563 Human action recognition has become a major research topic in computer vision. Tasks like violent
564 conduct or fights have been researched less, but they may be helpful in a variety of surveillance
565 video situations such as jails, mental hospitals, or even on a personal mobile phone. Their broad
566 applicability piques interest in developing violence or fight detectors. The main feature of the
567 detectors is efficiency, which implies that these methods should be computationally quick.
568 Although handcrafted spatio-temporal characteristics attain excellent accuracy for both
569 appearance and motion, extraction of certain features remains prohibitive for practical uses. For
570 the first time, the deep learning paradigm is applied to a job using a 3D CNN that accepts the
571 whole video stream as input. However, motion characteristics are critical for this job, and utilizing
572 full video as input causes noise and duplication in the learning process. A hybrid feature
573 "handcrafted/learned" framework was developed for this purpose (Serrano et. al., 2018). The
574 technique attempts to get an illustrative picture from the video sequence used as an input for feature
575 extraction, using Hough forest as a classifier. 2D CNN is then utilized to categorize that picture
576 and determine the sequence's conclusion. Experiments are carried out on three violence detection
577 datasets as "Hockey" (Nievas et. al., 2011), "Movie" (Bermejo et. al., 2011), and "Behavior" (Zhou
578 et. al., 2018). The findings show that the suggested approach outperforms the various handmade
579 and deep learning methods in terms of accuracies and standard deviations.
580 Two-stream CNN architecture, as well as an SVM classifier, is proposed (Xia et. al., 2018). Feature
581 extraction, training, and label fusion are the three phases of the method. Each stream CNN employs
582 an Imagenet VGG-f architecture that has been pre-trained. The first stream collects visual
583 information from successive frame differences, whereas the second stream extracts motion data.
584 Then, using sight and motion information, two SVM classifiers are trained. Finally, a label fusion
585 technique is used to get the detection result. The primary benefit of this technique is that it takes
586 very little time to process. However, since this technique can not identify aggressive behaviors
587 amongst individuals at close range, it is difficult to detect violence in large groups. Accattoli used
588 two-stream CNNs in a similar way (Accattoli et. al., 2020). To capture long temporal information,
589 they suggest combining CNNs with better trajectories. To extract geographical and temporal
590 information, they utilize two VGG-19 networks. Video frames are used to extract spatial
591 information, while dense optical flow pictures are used to retrieve temporal information.

592 In smart cities, schools, hospitals, and other surveillance domains, an improved security system is
593 required for the identification of violent or aberrant actions in order to prevent any casualties that
594 may result in social, economic, or environmental harm. For this purpose, a three-staged end-to-
595 end deep learning violence detection system is presented (Ullah et. al., 2019). To minimize and
596 overcome the excessive processing of use-less frames, people are first identified in the surveillance
597 video stream using a lightweight CNN model. Second, a 16-frame sequence containing identified
598 people is sent to 3D CNN, which extracts the spatiotemporal characteristics of the sequences and
599 feeds them to the Softmax classifier. The authors also used open visual inference and neural
600 networks optimization tools created by Intel to optimize the 3D CNN model, which transforms the
601 training model into intermediate representation and modifies it for optimum execution at the end
602 platform for the ultimate prediction of violent behavior. When violent behavior is detected, an
603 alarm is sent to the closest police station or security agency so that immediate preventative
604 measures may be taken. The datasets "Violent Crowd" (Hassner et. al., 2012), "Hockey" (Nievas
605 et. al., 2011), and "Violence in Movies" (Bermejo et. al., 2012) are used in the experiments. The
606 experimental findings show that the proposed approach outperforms state-of-the-art algorithms
607 such as ViF, AdaBoost, SVM, Hough Forest, and 2D CNN, sHOT, and others in terms of accuracy,
608 precision, recall, and AUC.

609

610 ## 5. VIDEO FEATURES AND DESCRIPTORS

611 This section goes through the feature descriptors that violence detection papers utilized in their
612 research as well as other recent state-of-the-art descriptors.

613 The fundamental components for detecting activity from the video are video features. The dataset
614 and characteristics collected from video to evaluate the pattern of activity have a direct impact on
615 the methodology's accuracy. For example, in combat situations, the movement of various objects
616 increases faster. The movement of objects in a typical setting is normal and not too rapid. The
617 direction of item movement in relation to time and space is also utilized to investigate unusual
618 occurrences. Table 6 lists all of the features that were utilized in the research.

619 A number of scholars, such as Lam et al. (Lejmi et. al., 2019), have worked hard to identify fights
620 and physical violence. Previous studies (Aggarwal and Ryoo, 2011; Poppe, 2010) used blood or
621 explosions as signals of violence, but these cues are seldom alarming. One study recently
622 developed a feature that offers strong multimodal audio and visual signals by first combining the
623 audio and visual characteristics and then exposing the combined multi-modal patterns statistically
624 (Sun & Liu, 2013). Multiple kernel learning is used to increase the multimodality of movies by
625 combining visual and audio data (Naik & Gopalakrishna, 2017). Audio-based techniques, on the
626 other hand, are always constrained in real life due to the lack of an audio channel.

627 The problem of detecting violent interactions is basically one of action recognition. The objective
628 is to extract characteristics that may describe the sequences throughout the battle using computer
629 vision technology. Handcrafted features and learning features are the two types of features
630 available.

631

632   **Table 6. Video features were used in the selected studies.**

633

634   Hand-crafted features. Human-designed features are referred to as hand-crafted features. In action
635   recognition, Space-Time Interest Points (STIPs) (Gracia et. al., 2015) and Improved Dense
636   Trajectories (iDTs) (Bermejo et. al., 2011) are often employed. Deniz et al. proposed a new
637   approach for detecting violent sequences that utilizes severe acceleration patterns as the primary
638   characteristic and applies the Radon transform on the power spectrum of successive frames to
639   identify violent sequences (Nievas et. al., 2011). Yang et. al. presented additional characteristics
640   derived from motion blobs between successive frames to identify combat and non-fight sequences
641   recently (Yang et. al., 2018). A robust and understandable method based on motion statistical
642   characteristics from optical flow pictures was presented in similar works (Yazdi et. al., 2018;
643   Soomro et. al., 2012). Zhang et al. use a GMOF to identify potential violence areas and a linear
644   SVM with OHOF input vectors to discover fight regions (Zhang et. al., 2019). This kind of
645   approach based on hand-crafted features is simple and effective for a small-scale dataset, but when
646   used to a large dataset, its deficiencies are exposed, resulting in slow training times, large memory
647   consumption, and inefficient execution.
648   Learning features. Deep neural networks learn features, which are referred to as learning features.
649   Physical violence detection based on deep learning has made significant progress because of the
650   increase in computing power brought on by GPUs and the gathering of large-scale training sets.
651   Two-stream ConvNets were created (Wang et. al., 2021) and comprise spatial and temporal nets
652   that use the ImageNet dataset (Chen & Hauptmann, 2009) for pre-training and optical flow to
653   explicitly capture motion information. Tran et al. used 3D ConvNets (De Souza, 2017) trained on
654   a large-scale supervised dataset to learn both appearance and motion characteristics. Zhang et al.
655   recently used long-range temporal structure (LTC) neural networks to train a movie and found that
656   LTC-CNN models with increasing temporal extents enhanced action identification accuracy
657   (Zhang et al., 2019). However, because of the computational complexity, these techniques are
658   restricted to a video frame rate of no more than 120 frames. The temporal segment network (Shao
659   et. al., 2017) used a sparse temporal sampling approach with video-level supervision to learn valid
660   information from the whole action video, attaining state-of-the-art performance on the two difficult
661   datasets HMDB51 (69.4 percent) and UCF101 (94.2 percent).

662

663   **5.1 Histogram of Oriented Gradients (HOG)**
664   HOGs are a feature descriptor for object identification and localization that can compete with
665   DNN's performance (Dalal & B. Triggs, 2005). The gradient direction distribution is utilized as a
666   feature in HOG. Because the brightness of corners and edges vary greatly, calculating the gradient
667   together with the directions may assist in the detection of this knowledge from the images.

668

669   **5.2 Histogram of Optical Flow (HOF)**
670   A pattern of apparent motion of objects, surfaces, and edges is produced as a result of the relative
671   motion between an observer and a scene. This process is called Optical Flow. The histogram of

672 oriented optical flow (HOF) (Dalal et. al., 2006) is an optical flow characteristic that depicts the
673 series of events at each point in time. It is scale-invariant and unaffected by motion direction.
674

5.3 SPACE – Time Interest Points

676 Laptev and Lindeberg and Laptev proposed the space–temporal interest point detector by
677 expanding the Harris detector. A second-moment matrix is generated for each spatiotemporal
678 interest point after removing points with high gradient magnitude using a 3D Harris corner detector
679 (Laptev and Lindeberg, 2004; Laptev, 2005). This descriptor's characteristics are used to describe
680 the spatiotemporal, local motion, and appearance information in volumes.
681 Space-Time Interest Points (STIP) is a space-time extension of the Harris corner detection
682 operator. The measured interest spots have a significant degree of intensity fluctuation in space
683 and non-constant mobility in time. These important sites may be found on a variety of geographical
684 and temporal scales. Then, for 3D video patches in the vicinity of the recognized STIPs, HOG,
685 HOF, and a combination of HOG and HOF called HNF feature vectors are retrieved. These
686 characteristics may be utilized to recognize motion events with high accuracy, and they are
687 resistant to changes in pattern size, frequency, and velocity.
688

5.4 MoSIFT

690 MoSIFT (Chen & Hauptmann, 2009) is an extension of the popular SIFT (Lowe, 2004) image
691 descriptor for video. The standard SIFT extracts histograms of oriented gradients in the image.
692 The 256-dimensional MoSIFT descriptor consists of two portions: a standard SIFT image
693 descriptor and an analogous HOF, which represents local motion. These descriptors are extracted
694 only from regions of the image with sufficient motion. The MoSIFT descriptor has shown better
695 performance in recognition accuracy than other state-of-the-art descriptors (Chen and Hauptmann,
696 2020) but the approach is significantly more computationally expensive than STIP.
697

5.5 Violence Flow Descriptor

699 The violence flow, which utilizes the frequencies of discrete values in a vectorized form, is an
700 essential feature descriptor. This is different from other descriptors in that instead of assessing
701 magnitudes of temporal information, the magnitudes are compared for each, resulting in much
702 more meaningful measurements in terms of the previous frame (Zhang et. al., 2019). Instead of
703 looking at local appearances, the similarities between flow-magnitudes in terms of time are
704 investigated.
705

5.6 Bag-of-Words (BoW)

707 The Bag-of-Words (BoW) method, which originated in the text retrieval community (Laptev,
708 2004), has lately gained popularity for a picture (Lewis, 1998) and video comprehension (Csurka
709 et. al., 2004). Each video sequence is represented as a histogram over a collection of visual words
710 in this method, which results in a fixed-dimensional encoding that can be analyzed with a
711 conventional classifier. The cluster centers produced via k-means clustering across a large

712 collection of sample low-level descriptors are usually described as the lexicon of visual words in
713 a learning phase (Lopes et. al., 2010).
714

5.7 Motion boundary histograms

716 By measuring derivatives independently for the horizontal and vertical components of the optical
717 flow, Dalal et al. developed the MBH descriptor (Dalal et. al., 2006) for human detection. The
718 relative motion between pixels is encoded by the descriptor. Because MBH depicts the gradient of
719 the optical flow, information regarding changes in the flow field (i.e. motion boundaries) is
720 preserved while locally constant camera motion is eliminated. MBH is more resistant to camera
721 motion than optical flow, making it better at action detection.
722

5.8 Vector of Locally Aggregated Descriptors

724 Soltanian, et al. presented a state-of-the-art VLAD descriptor (Soltanian et. al., 2019). VLAD
725 varies from the BoW image descriptor in that it records the difference between the cluster center
726 and the number of SIFTs allocated to the cluster, rather than the number of SIFTs assigned to the
727 cluster. It inherits parts of the original SIFT descriptor's invariances, such as in-plane rotational
728 invariance, and is tolerable to additional changes like picture scaling and clipping. VLAD retrieval
729 systems typically do not utilize the original local descriptors, which is another variation from the
730 conventional BoW method. These are employed in BoW systems for spatial verification and
731 reranking (Jégou et. al., 2010; Jegou et. al., 2008), but for extremely big picture datasets, they need
732 too much storage to be kept in memory on a single machine. VLAD is comparable to the previous
733 Fisher vectors (Philbin et. al., 2007) in that they both store features of the SIFT distribution given
734 to a cluster center.
735 VLAD is made up of areas taken from an image using an affine invariant detector and characterized
736 with the 128-D SIFT descriptor. The nearest cluster of a vocabulary of size k is then given to each
737 description (where k is typically 64 or 256 so that clusters are quite coarse). The residuals (vector
738 discrepancies between descriptors and cluster centers) are collected for each of the k clusters, and
739 the k 128-D sums of residuals are concatenated into a single k 128-D descriptor (Philbin et. al.,
740 2007). VLAD is comparable to other residual descriptors such as Fisher vectors (Arandjelovic &
741 Zisserman, 2013) and super-vector coding (Perronnin & Dance, 2007).
742

5.9 MoBSIFT

744 The MoBSIFT descriptor is a mixture of the MoSIFT and MBH descriptors. The two main stages
745 of the MoBSIFT method are interest point identification and feature description. The video is
746 converted into a few interest points once interest points are detected, and a feature description is
747 done locally around these interest points.
748 By combining the MBH (Dalal, et. al., 2006) with the movement filtering technique, the MoSIFT
749 descriptor (Chen and Hauptmann, 2009) is enhanced in both accuracy and complexity. Camera
750 motion is a significant issue for any system since motion data is considered an essential signal in
751 action detection. MBH is thought to be a useful feature for avoiding the effects of camera motion.

752 It is suggested that movement filtering be used to decrease complexity by excluding most
753 nonviolent movies from complicated feature extraction (Zhou et. al., 2010).
754
6. DATASETS
756 The real-world datasets that are utilized to evaluate the proposed violence detection methods are
757 described in this section. The specifics of all datasets linked to the violence are summarized in
758 Table 7.
759
760 **Table 7. Datasets for Violence Detection in Video.**
761
7. EVALUATION PARAMETERS
763 In this section, we describe evaluation parameters that are used to test the performance of physical
764 violence detection systems.
765 To evaluate the effectiveness of classification algorithms, the following indicators are usually
766 used: accuracy, completeness (also called True Positive Rate, TPR), and F-measure. After the
767 classification, it is possible to obtain four types of results: True Positive (TP), True Negative (TN),
768 False Positive (FP), and False Negative (FN). Let us explain the meaning of these terms: True
769 Positive: Its value represents the number of instances that have been correctly classified as violent.
770 False Negative: Its value represents the number of neutral videos that have been misclassified as
771 violent programs. False positive: Its value represents the number of normal classes that have been
772 misclassified as violent. True negative: Its value represents the number of normal classes that have
773 been correctly classified as normal. Then, to estimate the accuracy, use the following definition
774 and formula.
775
Accuracy
777 Accuracy is the ratio of the number of correct predictions to the total number of test samples. It
778 tells us whether a model is being trained correctly and how it may perform generally. Nevertheless,
779 it works well if only each class has an equal number of elements.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

780
Precision
782 Precision tells how often a prediction is correct when the model predicts positive. So precision
783 measures the portion of positive identifications in a prediction set that were actually correct.

$$precision = \frac{TP}{TP + FP}, \tag{2}$$

784 **Recall**
785 Recall is the number of correct positive results divided by the number of all relevant samples so
786 recall represents the proportion of actual positives that were identified correctly.

$$\Pr = \frac{TP}{TP + FN},\tag{3}$$

787   F1-score

788   F1 Score is a measure of the test's accuracy It is the Harmonic Mean between precision and recall.
789   The value of the F1 Score can be between 0 and 1. When the F1 score is equal to 1, the model is
790   considered to work perfectly.

$$F_{measure} = \frac{2 \times precision \times recall}{precision + recall},\tag{4}$$

791   AUC-ROC

792   When predicting the probability, the greater we can get the true positive rate (TPR) at a lower false
793   positive rate (FPR), the better the quality of the classifier. Therefore, we can introduce the
794   following metric that evaluates the quality of the classifier that calculates the probability of an
795   object belonging to a positive class:

$$AUC = \int_0^1 TPRdFPR,\tag{5}$$

796   This value is the area under the ROC curve. Here, $AUC \in [0,1]$. The ROC curve is a graphical tool
797   for evaluating the accuracy of binary classification models. It allows to find the optimal balance
798   between sensitivity and specificity of the model, which corresponds to the point of the ROC curve
799   closest to the coordinate (0,1), in which sensitivity and specificity are equal to 1, when there are
800   no false-positive and false-negative classifications.
801
802   False alarm

803   False Alarm: volume of false classes relative to the sum of classes.

$$FA = \frac{FP}{TN + FP},\tag{6}$$

804   Missing alarm

805   Missing Alarm (MA): volume of false classes relative to the sum of classes.

$$MA = \frac{FN}{TP + FN},\tag{7}$$

806

807   8. CHALLENGES TO VIOLENCE DETECTION IN VIDEO

808

809   Due to the many challenges faced when capturing moving people, detecting aggressive conduct is
810   tough. The major problems that need to be addressed are discussed below:

811

812   8.1 Dynamic Illumination Variations

813   When studying images with shifting light, which is a prevalent feature of realistic surroundings,
814   tracking becomes problematic. When collecting video at night, outside cctv are exposed to

815 environmental lighting variations, which might result in low contrast, making content
816 interpretation complicated. Zhou, et al. [43] shown a better susceptibility to light fluctuation. The
817 authors employed the LHOG descriptor, that is derived from a cluster of nodes. The
818 LHOG features consequence from color space. In this research, two ways to deal with
819 light variation were utilized: initially, the authors cut the block pitch in half, resulting in a half-
820 block overlap. The normalization process is then carried out per LHOG. The motion magnitude
821 photos are used to derive the LHOF, which catches real-time data. Furthermore, the adaptive
822 background subtraction technique [2] provides a dependable way for dealing with illumination
823 variations, as well as repeated and long-term scenario alterations.
824

8.2 Motion Blur

826 This is a difficult challenge for optical flow-based motion estimates to solve: Parts of human body
827 as head, arms, legs, elbows, and shoulders are mathematical feature points that create distinctive
828 abstraction of different stances. Deniz et al. proposed an approach for monitoring high
829 accelerations that does not require tracking [3]. Extreme acceleration causes visual blur, making
830 tracking less accurate or impossible, according to the researchers. Camera movement may, in fact,
831 produce picture blur. They used a deconvolution primary processing to get rid of it. To deduce
832 global motion between each couple of successive frames, the phase of correlation approach is
833 applied initially. If global motion is identified, the predicted slope and length of displacement are
834 utilized to create a Point Spread Function and deconvolve next frame applying Lucy-Richardson
835 iterative deconvolution technique.
836

8.3 Presence of a Non-stationary Background

838 Since low-resolution movies often feature background movement caused by camera movement or
839 changes in illuminance, noise correction is required. Whereas the amplitude of the optical flow
840 vector is a very powerful signal for detecting the degree of movement, and the flow direction may
841 offer additional motion information, [8] uses the optical flow technique for motion analysis. The
842 optical flow among each adjacent frames is computed using a background motion-resilient
843 technique. The backdrop normally moves at a consistent pace in response to the camera movement.
844 Human actions are more prone to move in irregular patterns. This enabled noise from background
845 movement to be filtered. A 3*3 Gaussian kernel is used to minimize noise, a histogram
846 equalization is used to disperse pixel intensities across a greater contrast range, and a background
847 subtraction utilizing Mixture of Gaussians is used to exclude items not linked to the actors. The
848 overwhelming background components may obstruct action recognition predictive performance.
849 To address this problem, we must consider both localizing action samples and reducing the effect
850 of backdrop video. Because of the camera movement, Wang et al. [34] noted a significant degree
851 of horizontal movement in the backdrop. They proposed adding warped optical flow as an input
852 modality, inspired by enhanced dense trajectory [32]. By predicting the homography matrix and
853 accounting for camera motion, they were able to extract them. That helps to concentrate on the
854 performer by removing the background motion. The temporal segment network architecture

855 receives additional aggregating capabilities in [33]. This is a good way to emphasize crucial pieces
856 while reducing ambient noise.
857

8.4 Non-professionally produced content

859 The volume of non-professionally generated material has expanded dramatically in recent years.
860 People can take images and videos everywhere and at any time owing to the widespread use of
861 cameras and cellphones. Social media and online forums are used to disseminate the material. Only
862 a little amount of study has been done to explore the scene structure of non-professionally
863 generated material, despite the fact that journalists often use amateur content, particularly in news
864 coverage on TV channels and news sites on the Internet. For example, if no professional team was
865 presented at an occurrence or if one was submitted but missed a potentially interesting event. We
866 see possibilities for a new discipline of social media-focused video scene identification algorithms.
867 Films published on social networks sharing sites are typically brief, with added handheld camera
868 movements, and the quality of two scenes might vary greatly. The new problem is not recognizing
869 scenes in these short films, but rather identifying a scene (situation) that occurred in real life among
870 a large number of videos on a social media sharing platform that incorporates footage from many
871 sources and of varying quality. A scene may be displayed from several perspectives, exposing
872 more information, and so providing a better experience for the viewer, by mixing footage from
873 different individuals. Chu et al. [14, 15] and Del Fabro et al. [18, 19] both suggest interesting ideas
874 that are connected to this problem.
875

8.5 Few publicly available datasets

877 Because this is a relatively new study topic, there are much fewer publically accessible datasets
878 for violence detection in video. Furthermore, the data imbalance between positive and negative
879 samples precludes the implementation of supervised models. The lack of ground truth data as well
880 as the ambiguous character of the anomalies makes it difficult to develop end-to-end trainable deep
881 learning models (Lloyd et. al., 2018). The modeling is further confounded by the large variation
882 within positive instances (anomalous occurrences may include a wide range of distinct instances,
883 despite the fact that most training data is restricted) (Zhu et. al., 2018). As a result, there is a need
884 for appropriate standards to assess the methods employed for violence detection in videos
885 (Constantin et. al., 2020).
886

8.6 Computational and time-consuming cost

888 In general, the time-consuming phase of feature representation during video violence detection,
889 which is both computationally and time-consuming, serves as a significant barrier for the
890 implementation of violence detection in real applications (Constantin et. al., 2020; Popoola &
891 Wang, 2012). As a result, the majority of current algorithms for detecting violence in videos have
892 significant time and space complexity costs. As a result, these techniques are unsuitable for
893 practical uses (Vu et. al., 2020). As a result, for improved feature extraction and description, more
894 dense and deeper DNN models are required. In addition, simple, effortless, and more effective

895 methods for detecting violence are required. However, the high dimensional structure coupled with
896 a non-local change between frames increases the complexity of the algorithms used to identify
897 video abnormalities.
898
899 ## 9. DISCUSSION
900
901 Despite considerable advancements in the video-based physical detection area, certain restrictions
902 remain, making it more complex and demanding. In reality, selecting the features that make a
903 moving item is a challenging task since it has a major impact on the behavior's description and
904 analysis. It's problematic, for instance, to describe the action when the scene's backdrop changes
905 often or when new items appear unexpectedly in the scene. Furthermore, the appearance of the
906 moving object may be affected by a variety of variables, including clothes (dress, suit, footwear,
907 etc.) and scene location (outdoor/indoor, etc.). In order to collect meaningful information regarding
908 an object's behavior, features that are resistant to scene modifications (rotating, occlusion,
909 blurring, cluttered backdrops, etc.) and less susceptible to changes in the object's appearance must
910 be used. Furthermore, most algorithms for detecting violence assume that the moving item is
911 located in front of the camera. In fact, though, the point of view is arbitrary. There are some studies
912 that utilize several cameras to record various perspectives of the moving item and then unite them
913 to circumvent this constraint. Even if such methods are efficient and provide excellent results, they
914 are complex, time-consuming, and unsuitable for practical uses. On the other hand, depending on
915 the context in which the action is done, as well as the time and location of the action, the observed
916 behavior may have many interpretations. Hugging, for example, is a common everyday activity
917 for most people, but hitting, kicking, use fighting techniques are considered aberrant conduct that
918 must be alarmed. On the other hand, depending on the context in which the action is done, as well
919 as the time and location of the action, the observed behavior may have many interpretations. To
920 address the constraints stated above, suggested methods utilized a massive quantity of training data
921 that included all permissible situations. To cope with huge amounts of data, cloud computing has
922 become popular, since it enables sophisticated algorithms, such as deep learning, to operate
923 efficiently on larger datasets. In reality, because of their deep structures, the usage of deep learning
924 techniques has exploded in order to achieve considerably more learning capacity.
925 Among the 80 reviewed papers, 4 articles (5%) are classified as review articles (Shidik et. al.,
926 2019; Ramzan et. al., 2019; K. Pawar and V. Attar, 2018; Sreenu and Durai, 2019). The article
927 reviews 220 papers that published between 2010 and 2019. It provides a review about methods,
928 frameworks, techniques in video-based intelligent violence detection systems. While it asserts that
929 its primary contribution is a thorough overview of the state-of-the-art of intelligence video
930 surveillance, several significant shortcomings should be highlighted. Intelligence video
931 surveillance was considered superficially without delving into the essence of the methods used.
932 Moreover, the paper included only journal papers into their study that excludes state-of-the-art
933 researches of high value conferences as ICCV [], ECCV[], ICML [], CVPR []. Thirdly, main accent

934 directed to video surveillance as protection, object detection, activity recognition, traffic
935 controlling, and disaster and accident monitoring that far from specific violence detection area.
936 Next review (Ramzan et. al., 2019) explores violence detection techniques using machine learning
937 methods. The review is not systematic as it was written in a free form. Authors classify video-
938 based violence detection into three categories as machine learning based, SVM based and
939 convolutional neural network based violence detection techniques. These strategies are discussed
940 in depth, as well as their advantages and disadvantages. Additionally, thorough tables detail the
941 datasets and video attributes that are employed in all procedures and play a critical part in the
942 violence detection process.
943 This article aims at studying and analyzing deep learning techniques for video-based anomalous
944 activity detection. As outcome of the study, the graphical taxonomy has been put forth based on
945 kinds of anomalies, level of anomaly detection, and anomaly measurement for anomalous activity
946 detection. The focus has been given on various anomaly detection frameworks having deep
947 learning techniques as their core methodology. Deep learning approaches from both the
948 perspectives of accuracy oriented anomaly detection and real-time processing oriented
949 abnormality detection are compared. This study also sheds light upon research issues and
950 challenges, application domains, benchmarked dataset and future directions in the domain of deep
951 learning based anomaly detection (K. Pawar and V. Attar, 2018).
952 The next section of the evaluation delves further into object identification, action recognition,
953 crowd analysis, and ultimately violence detection in a crowd setting (Sreenu and Durai, 2019). The
954 article superficially examines research in the indicated areas without a deep disclosure of the
955 methods used. Authors present four main approaches as a classification of abnormal behavior:
956 Hiden Markov Model, Gaussian Mixture Model (GMM), optical flow, and STT. Also, 13 datasets
957 for the violence detection were shown. Nevertheless, as the purpose of the paper is related to
958 different domains, the presented datasets also dedicated to different applications of deep learning
959 in video surveillance. There are only four popular datasets in violence detection as CAVIAR,
960 BEHAVE, Movie, and Hockey dataset. The rest are not directly related to the detection of violence.
961 Therefore, to the authors' knowledge, in the present literature, there is still a lack of a more formal
962 and objective systematic review that is specifically focused on "Violence Detection Techniques in
963 Video Surveillance Security Systems" and analyses it from multiple perspectives. The
964 contributions that are made in this work, as mentioned in Section I, can be used to address this
965 issue.
966
967

CONCLUSION

969 With the increasing development of surveillance cameras in many areas of life to watch human
970 behavior, the need for systems that automatically identify violent occurrences increases
971 proportionally. Violent action detection has become a prominent subject in computer vision
972 attracting new researchers. Many academics have suggested various methods for detecting such
973 actions in videos. The primary aim of this systematic review is to examine the most recent studies

in the field of violence detection. The various types of video violence detection techniques, which perform using machine learning, SVM, and deep learning were examined in this study. First, we looked at the most common techniques for extracting and describing features. Furthermore, all datasets and video characteristics were utilized in all techniques, as well as those that play a critical part in the identification process are documented in thorough tables. The accuracy of object identification, feature extraction, and classification methods, as well as the dataset utilized, are all factors. Following that, we gave a thorough review of descriptors for violence detection. In addition, we discussed the most difficult datasets and assessment criteria for video violence detection methods. Finally, we discussed challenges, open issues, and future directions for violence detection in video. Our research may help to highlight the strategies and procedures for detecting violent behavior from surveillance videos.

Acknowledgements

This research was supported by the grant "Development of artificial intelligence enabled software solution prototype for automatic detection of potential facts of physical bullying in educational institutions" funded by the Ministry of Education of the Republic of Kazakhstan. Grant No. IRN AP08855520.

References

Accattoli, S., Sernani, P., Falcionelli, N., Mekuria, D. N., & Dragoni, A. F. (2020). Violence detection in videos by combining 3D convolutional neural networks and support vector machines. Applied Artificial Intelligence, 34(4), 329-344, doi: https://doi.org/10.1080/08839514.2020.1723876

Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Comput. Surv. (CSUR) 43(3), 16:1–16:43 (2011), doi: https://doi.org/10.1145/1922649.1922653

Alex Net, [online] Available: https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet.

Al-Nawashi, M., Al-Hazaimeh, O. M., & Saraee, M. (2017). A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments. Neural Computing and Applications, 28(1), 565-572, doi: https://doi.org/10.1007/s00521-016-2363-z

Arceda, V. M., Fabián, K. F., Laura, P. L., Tito, J. R., & Cáceres, J. G. (2016). Fast face detection in violent video scenes. Electronic Notes in Theoretical Computer Science, 329, 5-26, doi: https://doi.org/10.1016/j.entcs.2016.12.002

Arandjelovic, R., & Zisserman, A. (2013). All about VLAD. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1578-1585).

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5297-5307).

Auria, L., & Moro, R. (2007). Advantages and disadvantages of support vector machines. Credit Risk Assessment Revisited: Methodological Issues and Practical Implications, 49-68.

1014   Bas, P., Filler, T., & Pevný, T. (2011, May). " Break our steganographic system": the ins and outs
1015   of organizing BOSS. In International workshop on information hiding (pp. 59-70). Springer,
1016   Berlin, Heidelberg.
1017   Bermejo E, Deni O, Bueno G, Sukthankar R. (2011) Violence detection in video using computer
1018   vision techniques. 14th Int. Congress on Computer Analysis of Images and Patterns, pp. 332-339.
1019   Blank M., Gorelick L., Shechtman E., Irani M., and Basri R.. Actions as space-time shapes. ICCV,
1020   2005.
1021   Blunsden, S., & Fisher, R. B. (2010). The BEHAVE video dataset: ground truthed video for multi-
1022   person behavior classification. Annals of the BMVA, 4(1-12), 4.
1023   Chandel, H., & Vatta, S. (2015). Occlusion detection and handling: a review. International Journal
1024   of Computer Applications, 120(10).
1025   Chen, M., Hauptmann, A.: MoSIFT: Recognizing human actions in surveillance videos. Tech.
1026   rep., Carnegie Mellon University, Pittsburgh, USA (2009)
1027   Cheng, T., & Williams, D. (2012). Space-time analysis of crime patterns in central London.
1028   International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,
1029   39, 47-52.
1030   Constantin, M. G., Stefan, L. D., Ionescu, B., Demarty, C. H., Sjoberg, M., Schedl, M., & Gravier,
1031   G. (2020). Affect in multimedia: Benchmarking violent scenes detection. IEEE Transactions on
1032   Affective Computing, DOI: https://doi.org/10.1109/TAFFC.2020.2986969
1033   Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004, May). Visual categorization
1034   with bags of keypoints. In Workshop on statistical learning in computer vision, ECCV (Vol. 1,
1035   No. 1-22, pp. 1-2).
1036   Dalal N. and Triggs B., "Histograms of oriented gradients for human detection," Proceedings of
1037   the Conference on Com-puter Vision and Pattern Recognition, San Diego, California, USA, pp.
1038   886–893, 2005, DOI: https://doi.org/10.1109/CVPR.2005.177
1039   Dalal N., Triggs B., and Schmid C., "Human detection using oriented histograms of flow and
1040   appearance," European Con-ference on Computer Vision, 2006, DOI:
1041   https://doi.org/10.1007/11744047_33
1042   Datta, A., Shah, M., & Lobo, N. D. V. (2002, August). Person-on-person violence detection in
1043   video data. In Object recogni-tion supported by user interaction for service robots (Vol. 1, pp. 433-
1044   438). IEEE, https://doi.org/10.1109/ICPR.2002.1044748
1045   De Souza, F., & Pedrini, H. (2017, October). Detection of violent events in video sequences based
1046   on census transform histogram. In 2017 30th SIBGRAPI Conference on Graphics, Patterns and
1047   Images (SIBGRAPI) (pp. 323-329). IEEE, DOI: https://doi.org/10.1109/SIBGRAPI.2017.49
1048   Deniz, O., Serrano, I., Bueno, G., & Kim, T. K. (2014, January). Fast violence detection in video.
1049   In 2014 international confer-ence on computer vision theory and applications (VISAPP) (Vol. 2,
1050   pp. 478-485). IEEE.
1051   Deepak, K., Vignesh, L. K. P., & Chandrakala, S. (2020). Autocorrelation of gradients based
1052   violence detection in surveillance videos. ICT Express, 6(3), 155-159,
1053   https://doi.org/10.1016/j.icte.2020.04.014

1054   Demarty, C. H., Ionescu, B., Jiang, Y. G., Quang, V. L., Schedl, M., & Penet, C. (2014, June).
1055   Benchmarking violent scenes detection in movies. In 2014 12th International Workshop on
1056   Content-Based   Multimedia   Indexing   (CBMI)   (pp.   1-6).   IEEE,   DOI:
1057   https://doi.org/10.1109/CBMI.2014.6849827
1058   Ding, C., Fan, S., Zhu, M., Feng, W., & Jia, B. (2014, December). Violence detection in video by
1059   using 3D convolutional neural networks. In International Symposium on Visual Computing (pp.
1060   551-558). Springer, Cham, DOI: https://doi.org/10.1007/978-3-319-14364-4_53
1061   Enzweiler, M., Eigenstetter, A., Schiele, B., & Gavrila, D. M. (2010, June). Multi-cue pedestrian
1062   classification with partial oc-clusion handling. In 2010 IEEE computer society conference on
1063   computer   vision   and   pattern   recognition   (pp.   990-997).   IEEE,   DOI:
1064   http://dx.doi.org/10.1109%2FCVPR.2010.5540111
1065   Febin, I. P., Jayasree, K., & Joy, P. T. (2020). Violence detection in videos for an intelligent
1066   surveillance system using MoBSIFT and movement filtering algorithm. Pattern Analysis and
1067   Applications, 23(2), 611-623, DOI: https://doi.org/10.1007/s10044-019-00821-3
1068   Fehrman, B., & McGough, J. (2014, April). Handling occlusion with an inexpensive array of
1069   cameras. In 2014 Southwest Symposium on Image Analysis and Interpretation (pp. 105-108).
1070   IEEE, DOI: https://doi.org/10.1109/SSIAI.2014.6806040
1071   Feng, J., Liang, Y., & Li, L. (2021). Anomaly Detection in Videos Using Two-Stream
1072   Autoencoder with Post Hoc Interpreta-bility. Computational Intelligence and Neuroscience, 2021,
1073   DOI: https://doi.org/10.1155/2021/7367870
1074   Fenil, E., Manogaran, G., Vivekananda, G. N., Thanjaivadivel, T., Jeeva, S., & Ahilan, A. (2019).
1075   Real time violence detection framework for football stadium comprising of big data analysis and
1076   deep learning through bidirectional LSTM. Computer Networks, 151, 191-200, DOI:
1077   https://doi.org/10.1016/j.comnet.2019.01.028
1078   Fisher   R.,   CAVIAR   test   case   scenarios.   (2004).
1079   http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/. Accessed 12 August 2021, DOI:
1080   https://doi.org/10.1109/TCSVT.2020.2991191
1081   Fu, E. Y., Leong, H. V., Ngai, G., & Chan, S. (2015, December). Automatic fight detection based
1082   on motion analysis. In 2015 IEEE International Symposium on Multimedia (ISM) (pp. 57-60).
1083   IEEE, DOI: https://doi.org/10.1109/ISM.2015.98
1084   Gao, Y., Liu, H., Sun, X., Wang, C., & Liu, Y. (2016). Violence detection using oriented violent
1085   flows. Image and vision com-puting, 48, 37-41, DOI: https://doi.org/10.1016/j.imavis.2016.01.006
1086   Hara, K., Kataoka, H., & Satoh, Y. (2017). Learning spatio-temporal features with 3d residual
1087   networks for action recognition. In Proceedings of the IEEE International Conference on Computer
1088   Vision Workshops (pp. 3154-3160).
1089   Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd
1090   behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision
1091   and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 16–21 June 2012; pp. 1–6,
1092   DOI: https://doi.org/10.1109/CVPRW.2012.6239348

1093 Jegou, H., Douze, M., & Schmid, C. (2008, October). Hamming embedding and weak geometric
1094 consistency for large scale image search. In European conference on computer vision (pp. 304-
1095 317). Springer, Berlin, Heidelberg, DOI: https://doi.org/10.1007/978-3-540-88682-2_24
1096 Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010, June). Aggregating local descriptors into a
1097 compact image representation. In 2010 IEEE computer society conference on computer vision and
1098 pattern recognition (pp. 3304-3311). IEEE, https://doi.org/10.1109/CVPR.2010.5540039
1099 Kamoona, A. M., Gostar, A. K., Bab-Hadiashar, A., & Hoseinnezhad, R. (2019, October).
1100 Sparsity-Based Naive Bayes Approach for Anomaly Detection in Real Surveillance Videos. In
1101 2019 International Conference on Control, Automation and Information Sciences (ICCAIS) (pp.
1102 1-6). IEEE, https://doi.org/10.1109/ICCAIS46528.2019.9074564
1103 Kang, S. B., Szeliski, R., & Chai, J. (2001, December). Handling occlusions in dense multi-view
1104 stereo. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and
1105 Pattern Recognition. CVPR 2001 (Vol. 1, pp. I-I). IEEE,
1106 https://doi.org/10.1109/CVPR.2001.990462
1107 Kantorov, V. and Laptev, I. "Efficient feature extraction, encoding and classification for action
1108 recognition". In Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2014.
1109 Keçeli, A. S., & Kaya, A. Y. D. I. N. (2017). Violent activity detection with transfer learning
1110 method. Electronics Letters, 53(15), 1047-1048.
1111 Khan, S. U., Haq, I. U., Rho, S., Baik, S. W., & Lee, M. Y. (2019). Cover the violence: A novel
1112 Deep-Learning-Based approach towards violence-detection in movies. Applied Sciences, 9(22),
1113 4963, https://doi.org/10.3390/app9224963
1114 Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011, November). HMDB: a large
1115 video database for human motion recognition. In 2011 International conference on computer vision
1116 (pp. 2556-2563). IEEE, https://doi.org/10.1109/ICCV.2011.6126543
1117 Laptev, I. (2005). On space-time interest points. International journal of computer vision, 64(2),
1118 107-123, https://doi.org/10.1007/s11263-005-1838-7
1119 Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: First International
1120 Workshop, SCVMA, pp. 91–103. Prague (2004), https://doi.org/10.1007/11676959_8
1121 Lejmi, W., Khalifa, A. B., & Mahjoub, M. A. (2019, September). Challenges and methods of
1122 violence detection in surveillance video: A survey. In International Conference on Computer
1123 Analysis of Images and Patterns (pp. 62-73). Springer, Cham, https://doi.org/10.5120/21264-3857
1124 Lewis, D. D. (1998, April). Naive (Bayes) at forty: The independence assumption in information
1125 retrieval. In European conference on machine learning (pp. 4-15). Springer, Berlin, Heidelberg,
1126 https://doi.org/10.1007/BFb0026666
1127 R. Leyva, V. Sanchez, C.-T. Li, Video anomaly detection with compact feature sets for online
1128 performance, IEEE Transactions on Image Processing 26 (7) (2017) 3463–3478,
1129 https://doi.org/10.1109/TIP.2017.2695105
1130 Li, D., Yao, T., Duan, L. Y., Mei, T., & Rui, Y. (2018). Unified spatio-temporal attention networks
1131 for action recognition in videos. IEEE Transactions on Multimedia, 21(2), 416-428,
1132 https://doi.org/10.1109/TMM.2018.2862341

1133 Lloyd, K., Marshall, D., Moore, S. C., & Rosin, P. L. (2016). Detecting violent crowds using
1134 temporal analysis of GLCM tex-ture. arXiv preprint arXiv:1605.05106,
1135 https://doi.org/10.1007/s00138-017-0830-x

1136 Lopes, A. P. B., Valle Jr, E. A. D., de Almeida, J. M., & de Araújo, A. A. (2010). Action
1137 recognition in videos: from motion capture labs to the web. arXiv preprint arXiv:1006.3506.

1138 Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International
1139 journal of computer vision, 60(2), 91-110, https://doi.org/10.1023/B:VISI.0000029664.99615.94

1140 Mabrouk, A. B., & Zagrouba, E. (2017). Spatio-temporal feature using optical flow based
1141 distribution for violence detection. Pattern Recognition Letters, 92, 62-67,
1142 https://doi.org/10.1016/j.patrec.2017.04.015

1143 Mabrouk, A. B., & Zagrouba, E. (2018). Abnormal behavior recognition for intelligent video
1144 surveillance systems: A review. Expert Systems with Applications, 91, 480-491,
1145 https://doi.org/10.1016/j.eswa.2017.09.029

1146 Mahmoodi, J., & Salajeghe, A. (2019). A classification method based on optical flow for violence
1147 detection. Expert systems with applications, 127, 121-127,
1148 https://doi.org/10.1016/j.eswa.2019.02.032

1149 Meng, Z., Yuan, J., & Li, Z. (2017, July). Trajectory-pooled deep convolutional networks for
1150 violence detection in videos. In International Conference on Computer Vision Systems (pp. 437-
1151 447). Springer, Cham, https://doi.org/10.1007/978-3-319-68345-4_39

1152 Mohtavipour, S. M., Saeidi, M., & Arabsorkhi, A. (2021). A multi-stream CNN for deep violence
1153 detection in video sequences using handcrafted features. The Visual Computer, 1-16,
1154 https://doi.org/10.1007/s00371-021-02266-4

1155 Mu, G., Cao, H., & Jin, Q. (2016, November). Violent scene detection using convolutional neural
1156 networks and deep audio features. In Chinese Conference on Pattern Recognition (pp. 451-463).
1157 Springer, Singapore, https://doi.org/10.1007/978-981-10-3005-5_37

1158 Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018).
1159 Systematic review or scoping review? Guidance for authors when choosing between a systematic
1160 or scoping review approach. BMC medical research methodol-ogy, 18(1), 1-7,
1161 https://doi.org/10.1186/s12874-018-0611-x

1162 Naik, A. J., & Gopalakrishna, M. T. (2021). Deep-violence: individual person violent activity
1163 detection in video. Multimedia Tools and Applications, 80(12), 18365-18380,
1164 https://doi.org/10.1007/s11042-021-10682-w

1165 Naik, A. J., & Gopalakrishna, M. T. (2017). Violence detection in surveillancevideo-a survey.
1166 International Journal of Latest Research in Engineering and Technology (IJLRET), 11-17.

1167 Nievas E. B., Suarez O. D., García G. B. and Sukthankar R., "Violence detection in video using
1168 computer vision techniques", International Conference on Computer Analysis of Images and
1169 Patterns, pp. 332-339, 2011, https://doi.org/10.1007/978-3-642-23678-5_39

1170 Niknejad, H. T., Kawano, T., Oishi, Y., & Mita, S. (2013, June). Occlusion handling using
1171 discriminative model of trained part templates and conditional random field. In 2013 IEEE

Intelligent Vehicles Symposium (IV) (pp. 750-755). IEEE, https://doi.org/10.1109/IVS.2013.6629557

F. Perronnin and D. Dance. Fisher kernels on visual vocabularies for image categorization. In Proc. CVPR, 2007, https://doi.org/10.1109/CVPR.2007.383266

Popoola O. P., Wang K., Video-based abnormal human behavior recognition-a review, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (6) (2012) 865–878, https://doi.org/10.1109/TSMCC.2011.2178594

Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. 28(6), 976–990 (2010), https://doi.org/10.1016/j.imavis.2009.11.014

Philbin J., Chum O., Isard M., Sivic J., and Zisserman A.. Object retrieval with large vocabularies and fast spatial matching. In Proc. CVPR, 2007, https://doi.org/10.1109/CVPR.2007.383172

Ramzan, M., Abid, A., Khan, H. U., Awan, S. M., Ismail, A., Ahmed, M., ... & Mahmood, A. (2019). A review on state-of-the-art violence detection techniques. IEEE Access, 7, 107560-107575, https://doi.org/10.1109/ACCESS.2019.2932114

Reddy, K. K., & Shah, M. (2013). Recognizing 50 human action categories of web videos. Machine vision and applications, 24(5), 971-981, https://doi.org/10.1007/s00138-012-0450-4

Rendón-Segador, F. J., Álvarez-García, J. A., Enríquez, F., & Deniz, O. (2021). ViolenceNet: Dense Multi-Head Self-Attention with Bidirectional Convolutional LSTM for Detecting Violence. Electronics, 10(13), 1601, https://doi.org/10.3390/electronics10131601

Rota, P., Conci, N., Sebe, N., & Rehg, J. M. (2015, September). Real-life violent social interaction detection. In 2015 IEEE International Conference on Image Processing (ICIP) (pp. 3456-3460). IEEE, https://doi.org/10.1109/ICIP.2015.7351446

Ribeiro, P. C., Audigier, R., & Pham, Q. C. (2016). RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance. Computer vision and image understanding, 144, 121-143, https://doi.org/10.1016/j.cviu.2015.11.001

Schuldt C., Laptev I., and Caputo B.. Recognizing human actions: A local SVM approach. ICPR, 2004, https://doi.org/10.1109/ICPR.2004.1334462

Senior, A., Hampapur, A., Tian, Y. L., Brown, L., Pankanti, S., & Bolle, R. (2006). Appearance models for occlusion handling. Image and vision computing, 24(11), 1233-1243, https://doi.org/10.1016/j.imavis.2005.06.007

Senst, T., Eiselein, V., Kuhn, A., & Sikora, T. (2017). Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation. IEEE transactions on information forensics and security, 12(12), 2945-2956, https://doi.org/10.1109/TIFS.2017.2725820

Serrano Gracia, I., Deniz Suarez, O., Bueno Garcia, G., & Kim, T. K. (2015). Fast fight detection. PloS one, 10(4), e0120448, https://doi.org/10.1371/journal.pone.0120448

Serrano, I., Deniz, O., Espinosa-Aranda, J. L., & Bueno, G. (2018). Fight recognition in video using hough forests and 2D convolutional neural network. IEEE Transactions on Image Processing, 27(10), 4787-4797, https://doi.org/10.1109/TIP.2018.2845742

1211   Shao, Z., Cai, J., & Wang, Z. (2017). Smart monitoring cameras driven intelligent processing to
1212   big surveillance video data. IEEE Transactions on Big Data, 4(1), 105-116,
1213   https://doi.org/10.1109/TBDATA.2017.2715815

1214   Shidik, G. F., Noersasongko, E., Nugraha, A., Andono, P. N., Jumanto, J., & Kusuma, E. J. (2019).
1215   A systematic review of intelligence video surveillance: Trends, techniques, frameworks, and
1216   datasets. IEEE Access, 7, 170457-170473, https://doi.org/10.1109/ACCESS.2019.2955387

1217   Soltanian, M., Amini, S., & Ghaemmaghami, S. (2019). Spatio-Temporal VLAD Encoding of
1218   Visual Events Using Temporal Ordering of the Mid-Level Deep Semantics. IEEE Transactions on
1219   Multimedia, 22(7), 1769-1784, https://doi.org/10.1109/TMM.2019.2959426

1220   Soomro K, Zamir A, Shah M (2012) A dataset of 101 human action classes from videos in the
1221   wild. CRCV TR. Technical report.

1222   Soomro, K., & Zamir, A. R. (2014). Action recognition in realistic sports videos. In Computer
1223   vision in sports (pp. 181-208). Springer, Cham, https://doi.org/10.1007/978-3-319-09396-3_9

1224   Song, D., Kim, C., & Park, S. K. (2018). A multi-temporal framework for high-level activity
1225   analysis: Violent event detection in visual surveillance. Information Sciences, 447, 83-103,
1226   https://doi.org/10.1016/j.ins.2018.02.065

1227   Song, W., Zhang, D., Zhao, X., Yu, J., Zheng, R., & Wang, A. (2019). A novel violent video
1228   detection scheme based on modified 3D convolutional neural networks. IEEE Access, 7, 39172-
1229   39179, https://doi.org/10.1109/ACCESS.2019.2906275

1230   Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada
1231   Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations.
1232   arXiv preprint arXiv:1810.02508

1233   Sreenu, G., & Durai, M. S. (2019). Intelligent video surveillance: a review through deep learning
1234   techniques for crowd analysis. Journal of Big Data, 6(1), 1-27.

1235   Sudhakaran, S., & Lanz, O. (2017, August). Learning to detect violent videos using convolutional
1236   long short-term memory. In 2017 14th IEEE International Conference on Advanced Video and
1237   Signal Based Surveillance (AVSS) (pp. 1-6). IEEE, https://doi.org/10.1109/AVSS.2017.8078468

1238   Sultani W., Chen C., and Shah M., "Real-world anomaly detection in surveillance videos," in The
1239   IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

1240   Sun, Q., Liu, H.: Learning spatio-temporal co-occurrence correlograms for efficient human action
1241   classification. In: 2013 IEEE International Conference on Image Processing, pp. 3220–3224.
1242   IEEE, September 2013, https://doi.org/10.1109/ICIP.2013.6738663

1243   Torii A., Sivic J., Pajdla T., and Okutomi M.. Visual Place Recognition with Repetitive Structures.
1244   In CVPR, 2013.

1245   Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., & Pajdla, T. (2015). 24/7 place recognition by
1246   view synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern
1247   Recognition (pp. 1808-1817).

1248   Tran, A., & Cheong, L. F. (2017). Two-stream flow-guided convolutional attention networks for
1249   action recognition. In Proceedings of the IEEE International Conference on Computer Vision
1250   Workshops (pp. 3110-3119).

1251 Tripathi, R. K., Jalal, A. S., & Agrawal, S. C. (2018). Suspicious human activity recognition: a
1252 review. Artificial Intelligence Review, 50(2), 283-339, https://doi.org/10.1007/s10462-017-9545-
1253 7

1254 Ullah, F. U. M., Ullah, A., Muhammad, K., Haq, I. U., & Baik, S. W. (2019). Violence detection
1255 using spatiotemporal features with 3D convolutional neural network. Sensors, 19(11), 2472,
1256 https://doi.org/10.3390/s19112472

1257 Ullah, F. U. M., Obaidat, M. S., Muhammad, K., Ullah, A., Baik, S. W., Cuzzolin, F., ... & de
1258 Albuquerque, V. H. C. (2021). An intelligent system for complex violence pattern analysis and
1259 detection. International Journal of Intelligent Systems, https://doi.org/10.1002/int.22537

1260 Ullah, F. U. M., Muhammad, K., Haq, I. U., Khan, N., Heidari, A. A., Baik, S. W., & Albuquerque,
1261 V. (2021). AI assisted Edge Vision for Violence Detection in IoT based Industrial Surveillance
1262 Networks. IEEE Transactions on Industrial Informatics, https://doi.org/10.1109/TII.2021.3116377

1263 Ushapreethi, P., & GG, L. P. (2020). Skeleton-based STIP feature and discriminant sparse coding
1264 for human action recognition. International Journal of Intelligent Unmanned Systems,
1265 https://doi.org/10.1108/IJIUS-12-2019-0067

1266 Vaaras, E., Ahlqvist-Björkroth, S., Drossos, K., & Räsänen, O. (2021). Automatic Analysis of the
1267 Emotional Content of Speech in Daylong Child-Centered Recordings from a Neonatal Intensive
1268 Care Unit. arXiv preprint arXiv:2106.09539.

1269 Vashistha, P., Bhatnagar, C., & Khan, M. A. (2018, March). An architecture to identify violence
1270 in video surveillance system using ViF and LBP. In 2018 4th International Conference on Recent
1271 Advances   in   Information   Technology   (RAIT)   (pp.   1-6).   IEEE,
1272 https://doi.org/10.1109/RAIT.2018.8389027

1273 Vu H., Phung D., Nguyen T. D., Trevors A., Venkatesh S., Energy-based models for video
1274 anomaly detection, arXiv preprint arXiv:1708.05211

1275 Wang, P., Wang, P., & Fan, E. (2021). Violence detection and face recognition based on deep
1276 learning. Pattern Recognition Letters, 142, 20-24, https://doi.org/10.1016/j.patrec.2020.11.018

1277 Wang Y., Huang K. and Tan T., "Human activity recognition based on R transform", IEEE CVPR,
1278 2007, https://doi.org/10.1109/CVPR.2007.383505

1279 Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In Proceedings of
1280 the IEEE international con-ference on computer vision (pp. 3551-3558).

1281 Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2018). Temporal
1282 segment networks for action recognition in videos. IEEE transactions on pattern analysis and
1283 machine intelligence, 41(11), 2740-2755, https://doi.org/10.1109/TPAMI.2018.2868668

1284 Wojek, C., Walk, S., Roth, S., & Schiele, B. (2011, June). Monocular 3d scene understanding with
1285 explicit   occlusion   reasoning.   In   CVPR   2011   (pp.   1993-2000).   IEEE,
1286 https://doi.org/10.1109/CVPR.2011.5995547

1287 Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., & Yang, Z. (2020, August). Not only look, but
1288 also listen: Learning multimodal violence detection under weak supervision. In European
1289 Conference on Computer Vision (pp. 322-339). Springer, Cham, https://doi.org/10.1007/978-3-
1290 030-58577-8_20

Wu, P., Liu, J., Li, M., Sun, Y., Shen, F.: Fast sparse coding networks for anomaly detection in videos. Pattern Recogn. 107515 (2020), https://doi.org/10.1016/j.patcog.2020.107515

Wu, P., Liu, J., Shen, F.: A deep one-class neural network for anomalous event detection in complex scenes. IEEE Trans. Neural Networks Learn. Syst. 31, 2609– 2622 (2019), https://doi.org/10.1109/TNNLS.2019.2933554

Q. Xia, P. Zhang, J. Wang, M. Tian and C. Fei, "Real time violence detection based on deep spatio-temporal features", Biometric Recognition (CCBR), pp. 157-165, 2018, https://doi.org/10.1007/978-3-319-97909-0_17

Xie, J., Yan, W., Mu, C., Liu, T., Li, P., & Yan, S. (2016). Recognizing violent activity without decoding video streams. Optik, 127(2), 795-801, https://doi.org/10.1016/j.ijleo.2015.10.165

Xu, X., Wu, X., Wang, G., & Wang, H. (2018, December). Violent video classification based on spatial-temporal cues using deep learning. In 2018 11th International Symposium on Computational Intelligence and Design (ISCID) (Vol. 1, pp. 319-322). IEEE, https://doi.org/10.1109/ISCID.2018.00079

Xu K., Jiang X., Sun T., Anomaly detection based on stacked sparse coding with intraframe classification strategy, IEEE Transactions on Multimedia 20 (5) (2018) 1062–107, https://doi.org/10.1109/TMM.2018.2818942

Yang, M. Y., Liao, W., Cao, Y., & Rosenhahn, B. (2018). Video event recognition and anomaly detection by combining gaussian process and hierarchical dirichlet process models. Photogrammetric Engineering & Remote Sensing, 84(4), 203-214, https://doi.org/10.14358/PERS.84.4.203

Yao, C., Su, X., Wang, X., Kang, X., Zhang, J., & Ren, J. (2021). Motion Direction Inconsistency-Based Fight Detection for Multiview Surveillance Videos. Wireless Communications and Mobile Computing, 2021, https://doi.org/10.1155/2021/9965781

Yazdi, M., & Bouwmans, T. (2018). New trends on moving object detection in video images captured by a moving camera: A survey. Computer Science Review, 28, 157-177, https://doi.org/10.1016/j.cosrev.2018.03.001

Ye, L., Wang, P., Wang, L., Ferdinando, H., Seppänen, T., & Alasaarela, E. (2018). A combined motion-audio school bullying detection algorithm. International Journal of Pattern Recognition and Artificial Intelligence, 32(12), 1850046, https://doi.org/10.1142/S0218001418500465

Ye, L., Wang, L., Ferdinando, H., Seppänen, T., & Alasaarela, E. (2020). A Video-Based DT–SVM School Violence Detecting Algorithm. Sensors, 20(7), 2018, https://doi.org/10.3390/s20072018

Ye, L., Liu, T., Han, T., Ferdinando, H., Seppänen, T., & Alasaarela, E. (2021). Campus Violence Detection Based on Artificial Intelligent Interpretation of Surveillance Video Sequences. Remote Sensing, 13(4), 628, https://doi.org/10.3390/rs13040628

Yoganand, A. V., & Kavida, A. C. (2018). Face detection approach from video with the aid of KPCM and improved neural network classifier. Multimedia Tools and Applications, 77(24), 31763-31785, https://doi.org/10.1007/s11042-018-6191-2

1330 Zhang, C., Xu, J., Beaugendre, A., & Goto, S. (2012, May). A KLT-based approach for occlusion
1331 handling in human tracking. In 2012 Picture Coding Symposium (pp. 337-340). IEEE,
1332 https://doi.org/10.1109/PCS.2012.6213360
1333 Zhang, T., Jia, W., Yang, B., Yang, J., He, X., & Zheng, Z. (2017). MoWLD: a robust motion
1334 image descriptor for violence detection. Multimedia Tools and Applications, 76(1), 1419-1438,
1335 https://doi.org/10.1007/s11042-015-3133-0
1336 Zhang J., Chen C., Xiang Y. and Zhou W., "Semi-supervised and compound classification of
1337 network traffic", Proc. WDCS, pp. 617-621, 2012,
1338 Zhang, Y., Girgensohn, A., & Tjahjadi, Y. (2019, August). Activity forecasting in routine tasks by
1339 combining local motion trajectories and high-level temporal models. In 2019 IEEE International
1340 Conference on Computational Science and Engineering (CSE) and IEEE International Conference
1341 on Embedded and Ubiquitous Computing (EUC) (pp. 447-452). IEEE,
1342 https://doi.org/10.1109/CSE/EUC.2019.00089
1343 Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J., & He, X. (2016). A new method for violence
1344 detection in surveillance scenes. Multimedia Tools and Applications, 75(12), 7327-7349,
1345 https://doi.org/10.1007/s11042-015-2648-8
1346 Zhou, P., Ding, Q., Luo, H., & Hou, X. (2017, June). Violent interaction detection in video based
1347 on deep learning. In Journal of physics: conference series (Vol. 844, No. 1, p. 012044). IOP
1348 Publishing, https://doi.org/10.1088/1742-6596/844/1/012044
1349 Zhou X., Yu K., Zhang T., and Huang T. Image classification using super-vector coding of local
1350 image descriptors. In Proc. ECCV, 2010, https://doi.org/10.1007/978-3-642-15555-0_11
1351 Zhu, Y., Lan, Z., Newsam, S., & Hauptmann, A. (2018, December). Hidden two-stream
1352 convolutional networks for action recognition. In Asian conference on computer vision (pp. 363-
1353 378). Springer, Cham, https://doi.org/10.1007/978-3-030-20893-6_23
1354 Zhuang, N., Yusufu, T., Ye, J., & Hua, K. A. (2017, May). Group activity recognition with
1355 differential recurrent convolutional neural networks. In 2017 12th IEEE International Conference
1356 on Automatic Face & Gesture Recognition (FG 2017) (pp. 526-531). IEEE,
1357 https://doi.org/10.1109/FG.2017.70
1358 Zhang, Q., Zhang, M., Chen, T., Sun, Z., Ma, Y., & Yu, B. (2019). Recent advances in
1359 convolutional neural network acceleration. Neurocomputing, 323, 37-51,
1360 https://doi.org/10.1016/j.neucom.2018.09.038
1361 Zhou, P., Ding, Q., Luo, H., & Hou, X. (2018). Violence detection in surveillance video using
1362 low-level features. PLoS one, 13(10), e0203668, https://doi.org/10.1371/journal.pone.0203668

Figure 1

Systematic literature review flowchart

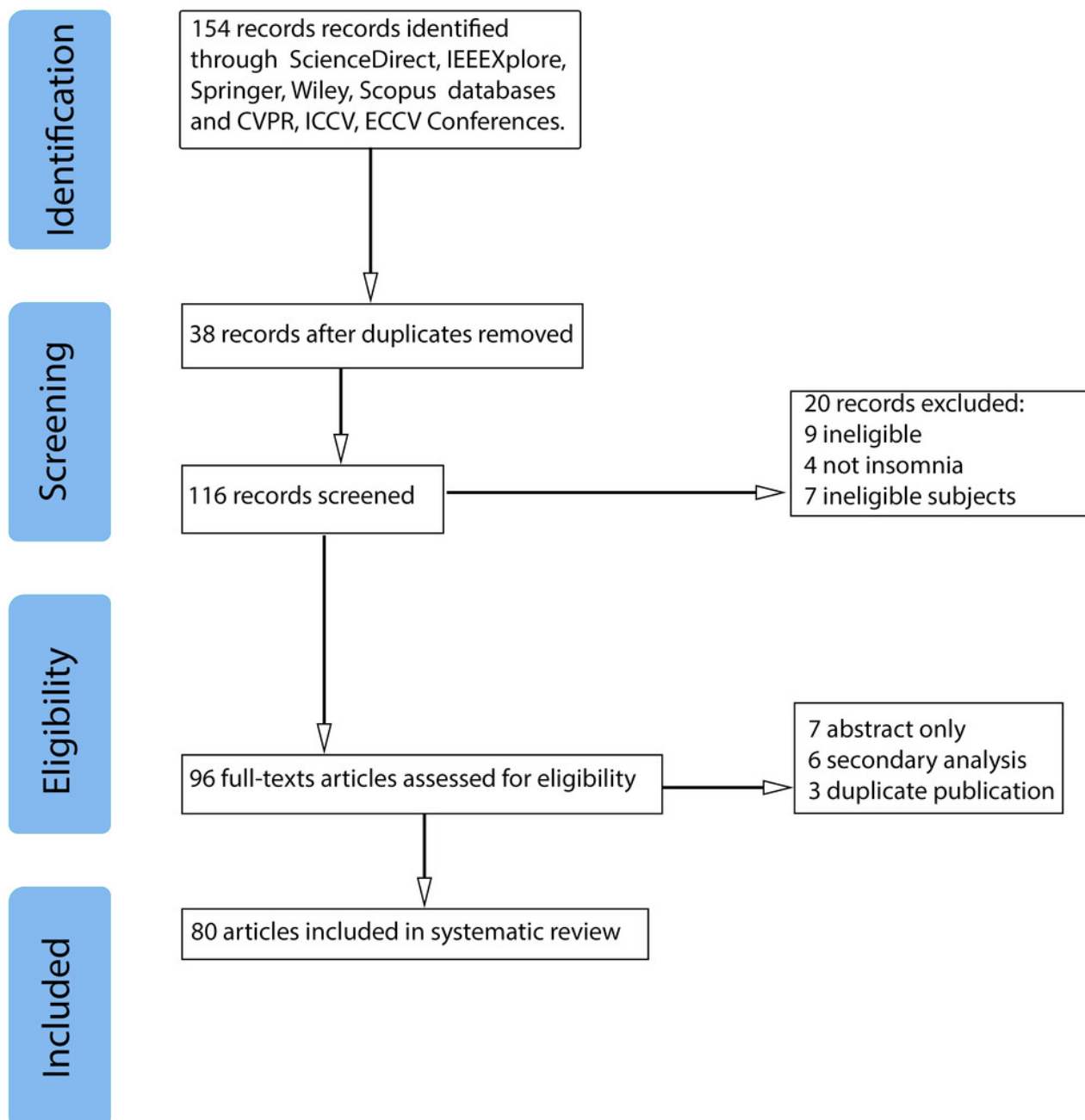



Figure 2

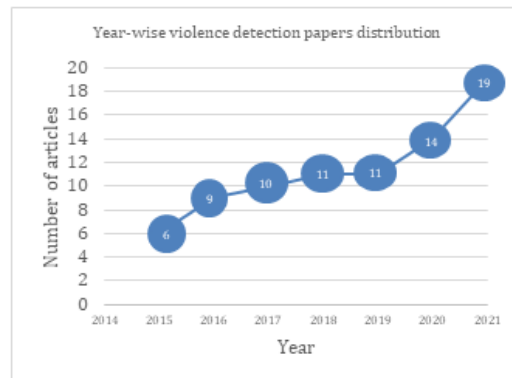
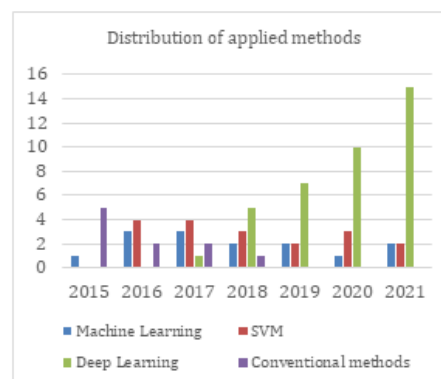
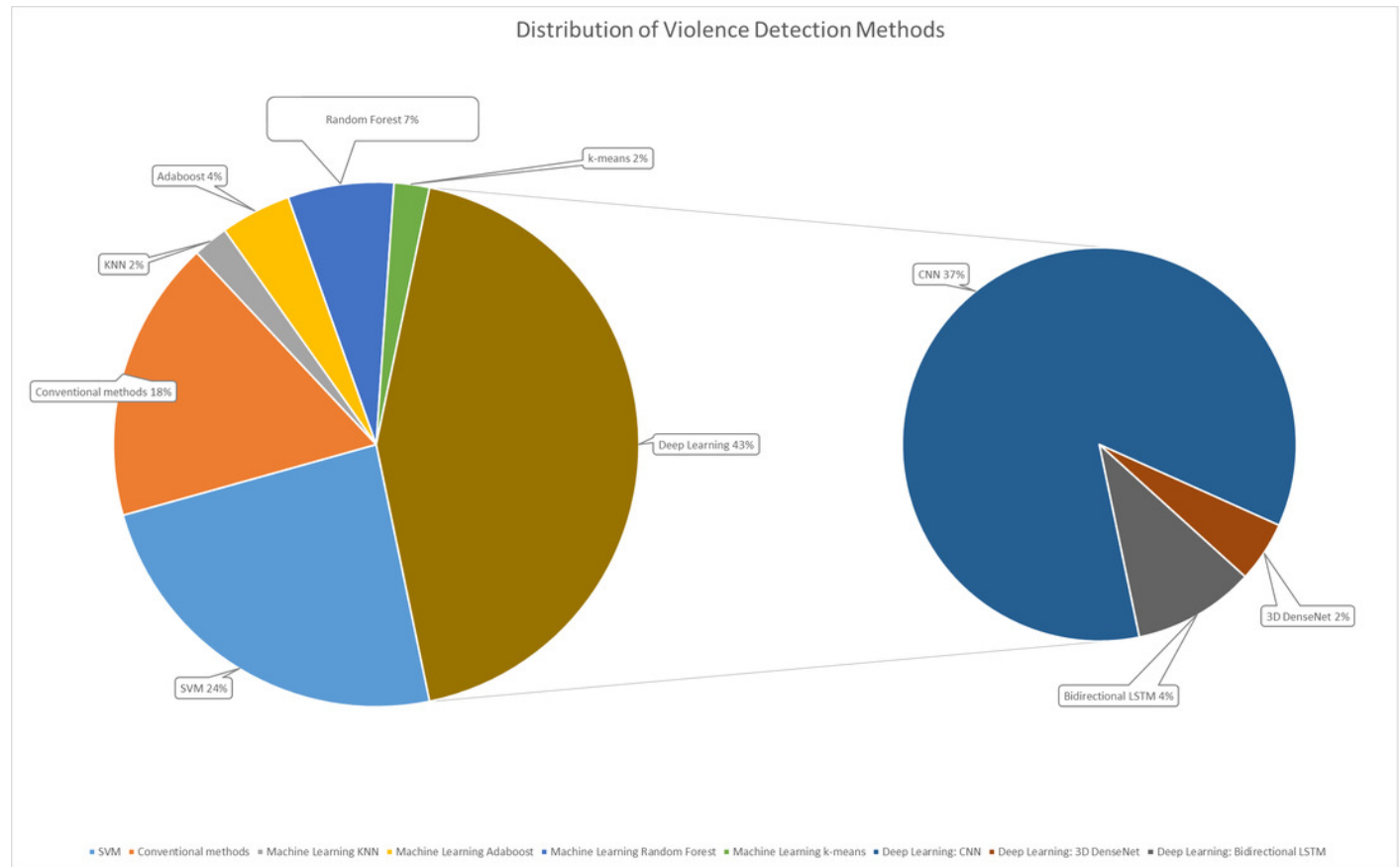
Year-wise violence detection papers distribution

a) Year-wise violence detection papers distribution

b) Year-wise violence detection papers by applied methods

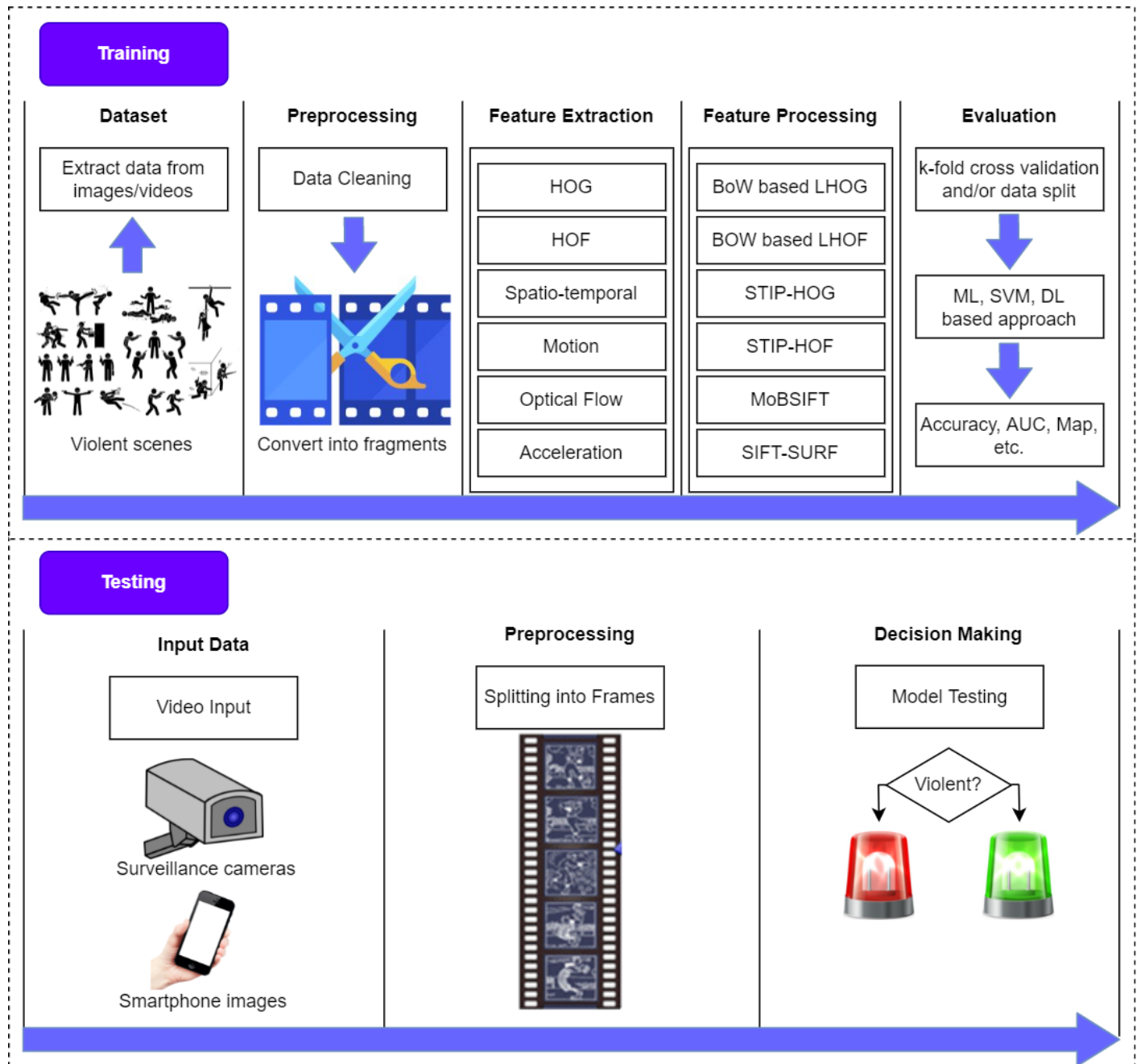
Figure 3

Distribution of violence detection methods

Inclusion and exclusion criteria

1   **Table 1. Inclusion and exclusion criteria.**

I/E	Criteria	Explanation
Inclusion	Review paper	The paper proposes different types of reviews as literature review, systematic review, survey, etc.
	Research paper	The paper aims to solve specific research problems related to video surveillance security systems.
Exclusion	Duplicated papers	The same paper that appears multiple times
	Non-research papers	The paper is not a research article. It might be Editorial notes, comments, etc.
	Non-related papers	The topic under study goes beyond the research context of this work
	Non English papers	The paper is not written in English
	Implicitly related papers	The paper does not directly express the research focus on video surveillance security systems.
	Non research paper	The paper is not a research paper. It might be editorial notes, comments, etc.

Figure 4

Fundamental stages of video-based violence detection

Research questions and their motivations

1 **Table 2. Research questions and their motivations.**

ID	Research Question	Motivation
RQ1	What kind of video based violence detection techniques are applied in state-of-the-art researches?	Identify state-of-the-art methods and techniques in intelligent video surveillance
RQ2	What kind of video features and descriptors are used in video-violence detection?	Identify commonly used and state-of-the-art features and descriptors in video violence detection
RQ3	What datasets are used to train models for video-violence detection	Identify datasets commonly used in intelligent video surveillance
RQ4	What challenges and open questions exist to identify violence in videos?	Identify challenges and open issues in intelligent video surveillance

Violence detection techniques that use machine learning

1 **Table 3. Violence detection techniques that use machine learning.**

Gracia et. al., 2015	Motion Blob acceleration measure vector method for detection of fast fighting from video	Ellipse detection method	An algorithm to find the acceleration	Spatio-temporal features use for classification	Both crowded and less crowded	Accuracy about 90%
Zhou et. al., 2018	FightNet for Violent Interaction Detection	Temporal Segment Network	Image acceleration	Softmax	Both crowded and uncrowded	97% in Hockey, 100% in Movies dataset
Ribeiro et. al., 2016	RIMOC method focuses on speed and direction of an object on the base of HOF	Covariance Matrix method STV based	Spatio-temporal vector method (STV)	STV uses supervised learning	Both crowded and uncrowded	For normal situation 97% accuracy
Yao et. al., 2021	Multiview fight detection method	YOLO-V3 network	Optical flow	Random Forest	Both crowded and uncrowded	97.66% accuracy, 97.66 F1-score
Arceda et. al., 2016	Two step detection of violent and faces in video by using ViF descriptor and normalization algorithms.	Vif object recognition CUDA method and KLT face detector	Horn shrunk method for Histogram	Interpolation classification	Less crowded	Lower frame rate 14% too high rate of 35% fs/s 97%
Wu et. al., 2020	HL-Net to simultaneously capture long-range relations and local distance relations	HLC approximator	CNN based model	Weak supervision	Both Crowded and uncrowded scene	78.64%
Xie et. al., 2016	SVM method for recognition based on statistical theory frames	Vector normalization method	Macro block technique for features	Region motion and descriptor for video	Crowded	96.1% accuracy

			extractions	classification		
Febin et. al., 2020	A cascaded method of violence detection based on MoBSIFT and movement fltering	MoBSIFT	Motion boundary histogram	SVM, random forest, and AdaBoost	Both Crowded and uncrowded scene	90.2% accuracy in Hockey, 91% in Movies dataset
Senst et. al., 2017	Lagrangian fields of direction and begs of word framework to recognize the violence in videos	Global compensation of object motion	Lagrangian Theory and STIP method for extract Motion features	Late fusion for classification	Crowded	91% to 94% accuracy

Violence Detection Techniques Using SVM

1 **Table 4. Violence Detection Techniques Using SVM.**

Ye et. al., 2020	A Video-Based DT–SVM School Violence Detecting Algorithm	Motion Co-occurrence Feature (MCF)	Optical flow extraction	Crowded	97.6%
Zhang et. al., 2016	GMOF framework with tracking and detection module	Gaussian Mixture model	OHFO for optical flow extraction	Crowded	82%-89% accuracy
Gao et. al., 2016	Violence detection using Oriented ViF	Optical Flow method	Combination of ViF and OViF descriptor	Crowded	90%
Deepak et. al., 2020	Autocorrelation of gradients based violence detection	Motion boundary histograms	Frame based feature extraction	Crowded	91.38% accuracy in Crowd Violence; 90.40% in Hockey dataset
Al-Nawashi et. al., 2017	Framework includes preprocessing, detection of activity and image retrieval. It identifies the abnormal event and image from data-based images.	Optical flow and temporal difference for object detection CBIR method for retrieving images.	Gaussian function for video future analysis	Less crowded	97% accuracy
Kamoona et. al., 2019	Sparsity-Based Naive Bayes Approach for Anomaly Detection in Real	Sparsity-Based Naïve Bayes	C3D feature extraction	Both crowded and uncrowded	64.7% F1 score; 52.1% precision; 85.3% recall in UCF dataset

	Surveillance Videos				
Song et. al., 2018	SGT-based and SVM-based multi-temporal framework to detect violent events in multi-camera surveillance.	Late fusion	Multi-temporal Analysis (MtA)	Variety fight scenes from minimum two to maximum fifteen people include various movements	78.3% (SGT-based, BEHAVE), 70.2% (SVM-based, BEHAVE), 87.2% (SGT-based, NUS–HGA), and 69.9% (SGT-based, YouTube)
Vashistha et. al., 2018	An architecture to identify violence in video surveillance system using ViF and LBP	Shape and motion analysis	ViF and Local Binary Pattern (LBP) descriptors	Both crowded and non-crowded scenes	89.1% accuracy in Hockey dataset, 88.2% accuracy in Violent-Flow dataset

Violence detection using deep learning techniques

1   **Table 5. Violence detection using deep learning techniques.**

Ding et. al., 2016	Violence Detection using 3D CNN	3D convolution is used to get spatial information	Backpropagation method	Crowded	91% accuracy
Arandjelovic, 2016,	Deep architecture for place recognition	VGG VLAD method for image retrieval	Backpropagation method for feature extraction	Crowded	87%-96% accuracy
Fenil et. al., 2019	Framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM	Bidirectional LSTM	HOG, SVM	Crowded	94.5% accuracy
Mu et. al., 2016,	Violent scene detection using CNN and deep audio features	MFB	CNN	Crowded	Approximately 90% accuracy
Mohtavipour et. al, 2021,	A multi-stream CNN using handcrafted features	A deep violence detection framework based on the specific features (speed of movement, and representative image) derived from handcrafted methods.	CNN	Both crowded and uncrowded	
Sudhakaran	Detect	CNN along	CNN	Crowded	Approximatel

and Lanz, 2017	violent videos using ConvLSTM	with the ConvLSTM		y 97%
Naik and Gopalakrishna, 2021	Deep violence detection framework based on the specific features derived from handcrafted methods	Discriminative feature with a novel differential motion energy image	CNN	Both crowded and uncrowded
Meng et. al., 2017	Detecting Human Violent Behavior by integrating trajectory and Deep CNN	Deep CNN	Optical flow method	Crowded	98% accuracy
Rendón-Segador et. al., 2021	ViolenceNet: Dense Multi-Head Self-Attention with Bidirectional Convolutional LSTM	3D DenseNet	Optical flow method	Crowded	95.6%-100% accuracy
Xia et. al., 2018	Violence detection method based on a bi-channels CNN and the SVM.	Linear SVM	Bi-channels CNN	Both crowded and uncrowded scenes	95.90 ± 3.53 accuracy in Hockey fight, 93.25 ± 2.34 accuracy in Violence crowd
Meng et. al., 2020	Trajectory-Pooled Deep Convolutional Networks	ConvNet model which contains 17 convolutionpool-norm layers	Deep ConvNet model	Both crowded and uncrowded	92.5% accuracy in Crowd Violence, 98.6% in

		and two fully connected layers			Hockey Fight dataset
Ullah, et. al., 2019	Violence Detection using Spatiotemporal Features	Pre-train Mobile Net CNN model	3D CNN	Crowded	Approximately 97% accuracy

Video features were used in the selected studies

1 **Table 6. Video features were used in the selected studies.**

Mabrouk et. al., 2018	Motion, space and time
Lejmi et. al., 2019	Motion blobs, Edges and corner of image
Serrano et. al., 2015	Motion blobs
Zhou et. al., 2017	Optical flow, motion and moving bob
Ribeiro et. al., 2016	Motion, direction and speed
Yao et. al., 2021	MoDI and WMoDI, motion regions marking
Arceda et. al., 2016	Optical flow, Magnitude
Wu et. al., 2020	Optical flow and audio features
Xie et. al., 2016	Motion vector and direction
Febin et. al., 2020	Optical flow, MBH, movement filtering?
Senst et. al., 2017	Spatial, temporal and motion
Ye et. al., 2018	Rectangular frame and optical-flow
Zhang et. al., 2016	Spatiotemporal and motion
Deepak et. al., 2020	Spatio-Temporal Auto-Correlation of Gradients
Al-Nawashi et. al., 2017	Motion region and optical flow
Kamoona et. al., 2019	Multiple handcrafted features
Sultani et. al., 2018	Spatial-temporal, C3D
Song et. al., 2018	Movement, direction and speed
Vashistha et. al., 2018	Speed, direction, centroid and dimensions
Arandjelovic et. al., 2016	Spatiotemporal features
Torii et. al., 2015	Spatio-temporal
Mohtavipour et. al., 2021	Spatial, temporal, and spatiotemporal streams
Rendón-Segador et. al., 2021	Spatio-temporal features
Serrano et. al., 2018	Apperancde, motion, optical flow
Xia et. al., 2018	Direction and motion information
Accattoli et. al., 2020	hand-crafted and trajectory-deep features
Ullah et. al., 2019	Spatiotemporal features
Fu et. al., 2015	Motion, acceleration and magnitude
Deniz et. al., 2015	Spatiotemporal, acceleration and motion
Ye et. al., 2018	Time-domain and frequency-domain
Mabrouk et. al., 2017	STIP, optical flow

Datasets for Violence Detection in Video

1 **Table 7. Datasets for Violence Detection in Video.**

Dataset	Reference	Characteristics	Published year	Reference used
Movies	Bermejo et. al., 2011	200 video clips of totally 6 min	2011	Gracia et. al., 2015; Senst et. al., 2017; Fenil et. al., 2019; Sudhakaran & Lanz, 2017; Ullah et. al., 2019
Hockey	Nievas et. al., 2011	1 000 video clips of totally 27 min	2011	Gracia et. al., 2015; Senst et. al., 2017; Zhang et. al., 2016; Gao et. al., 2016; Deepak et. al., 2020; Ding et. al., 2014; Sudhakaran & Lanz, 2017; Meng et. al., 2017; Serrano et. al., 2018; Ullah et. al., 2019
UCF-101	Soomro et. al., 2012	13 000 clips	2012	Gracia et. al., 2015
CASIA Action	Wang et. al., 2007	8 classes of single person activities that contain 1446 video clips	2007	Yao et. al., 2021
UT-Interaction	Zhang et. al., 2012	20 video sequences with the resolution of 720x480 at 30fps.	2012	Yao et. al., 2021
The Boss	Bas et. al., 2011	10,000 images for training and 1,000 for testing	2011	Arceda et. al., 2016
XD-Violence	Wu et. al., 2020	4574 videos with duration of 217 hours that has 6 types of violent and 9 types of non-violent videos	2020	Wu et. al., 2020
VVAR10	Xie et. al., 2016	296 positive and 277 negative instances	2016	Xie et. al., 2016
UCF Sports	Soomro & Zamir, 2014	-	2014	Xie et. al., 2016
UCF50	Reddy &	50 actions, 100 minute	2010	Xie et. al., 2016

	Shah, 2013	videos		
HMDB51	Kuehne et. al., 2011	6,766 manually annotated videos that divided to 51 classes	2011	Xie et. al., 2016
BEHAVE	Blunsden & Fisher, 2010	200 000 frames	2010	Zhang et. al., 2016; Song et. al., 2018; Serrano et. al., 2018;
CAVIAR	Bins et. al., 2005	28 video clips	2004	Zhang et. al., 2016;
Violent-Flow	(Xu et. al., 2018)	30 video clips	2012	Gao et. al., 2016; Sudhakaran & Lanz, 2017
UCF	Sultani et. al., 2018	128 hours of videos. 1900 surveillance videos that includes 13 types of classes	2018	Kamoona et. al., 2019
Crowd Violence	Song et. al., 2019	-	2012	Deepak et. al., 2020, Meng et. al., 2017;
Finnish emotional	Vaaras et. al., 2021	132 samples	2021	Ye et. al., 2021
Chinese emotional	Poria et. al., 2018	370 samples	2018	Ye et. al., 2021
Pittsburgh	Torii et. al., 2013	-	2013	Arandjelovic et. al., 2016
Tokyo 24/7	Torii et. al., 2015	-	2015	Arandjelovic et. al., 2016
Violent interaction	Fenil et. al., 2019	2314 movies with 1077 fights and 1237 no-fights	2019	Fenil et. al., 2019
MediaEval	Demarty et. al., 2014	10 000 clips	2014	Mu et. al., 2016
Violent-Flows	Rota et. al., 2015	-	2015	Sudhakaran & Lanz, 2017
Weizmann	Blank et. al., 2005	9 actions, 9 clips	2005	Naik & Gopalakrishna, 2021
KTH	Schuldt et. al., 2004	6 actions, 100 clips	2004	Naik & Gopalakrishna, 2021
Violent Crowd	(Bermejo et. al., 2012)	246 short video sequences that video length is varying from 50 to 150 frames.	2012	Ullah et. al., 2019

London Metropolitan Police	Cheng et. al., 2012	-	2012	Senst et. al., 2017