Detection Violent Behaviors: a Survey

Dalila Durães[1,2], Francisco S. Marcondes[2], Filipe Gonçalves[2,3], Joaquim Fonseca[3], José Machado[2], and Paulo Novais[2]

[1] CIICESI, ESTG, Instituto Politécnico do Porto, Portugal
dad@estg.ipp.pt
[2] Algorithm Center, University of Minho, Braga, Portugal
francisco.marcondes@algoritmi.uminho.pt, {jmac,pjon}@di.uminho.pt
[3] Bosch Car Multimedia, Braga, Portugal
{filipe.goncalves,joaquim.fonseca2}@pt.bosch.com

Abstract. Violence detection behavior is a particular problem regarding the great problem action recognition. In recent years, the detection and recognition of violence has been studied for several applications, namely in surveillance. In this paper, we conducted a recent systematic review of the literature on this subject, covering a selection of various researched papers. The selected works were classified into three main approaches for violence detection: video, audio, and multimodal audio and video. Our analysis provides a roadmap to guide future research to design automatic violence detection systems. Techniques related to the extraction and description of resources to represent behavior are also reviewed. Classification methods and structures for behavior modelling are also provided.

Keywords: Violence Detection · Action Recognition · Video Surveillance · Audio Surveillance · Multimodal Surveillance.

1 Introduction

In recent years, detection and recognition of violence has been studied for several applications, namely in surveillance. In surveillance, the analysis and automatic detection of abnormal, dangerous and violent events is an important field of study [1].

Action recognition and violence detection as been study, with different perspectives, for several disciplines, including psychology, biomechanics and computer vision [1]. In computer vision several studies were performed using different approach. In some approach the technique applied was visual detection. In other approach the technique applied was audio detection. Finally, in other approach, was used audio and visual techniques.

In visual approach a movement has a set of primitive actions and describes a whole-body movement. However, a primitive action is an atomic motion that can be described in terms of a member [2]. The detection of human movements consists of obtaining a set of actions. Finally, several subsequent actions, providing an interpretation of the movement being performed and are called activities [2].

In audio approach the signal produced by the sound of an audio contains a lot of information that only visual data cannot represent, namely: screams, explosions, words of abuse and even sound passages showing some kind of emotion.

In this paper, we conducted a recent systematic review of the literature on this subject, covering a selection of various researched articles. This paper is organized as follows. Firstly, section 2 introduces the concepts of action recognition and violence detection. Then, section 3 presents the sample setup. Next, section 4 presents results and discussion, namely the video approach, audio approach,and multimodal audio and video approach. Finally, section 5 concludes the review by performing a global analysis to the presented review and presenting some future work for this research.

2 Concepts

The goal of an intelligent surveillance system for violence detection has effectively to detected event in real-time to avoid dangerous situations. However, it's necessary to understand some important concepts of surveillance.

2.1 Action Recognition

In a human recognition action recognition is a system that can detect human activities. Types of human activity are classified into four different categories, depending on the complexity of the actions and the number of body parts involved in the action. The four categories are gestures, actions, interactions and group activities [3]. A gesture is a collection of movements made with the hands, head, or members to show a particular meaning. Actions are a collection of multiple gestures performed by a single person. Interactions are a collection of human actions, with at most, two actors. When there are two actor, an actor should be a human and the other can be a person or an object. Group activities are a combination of gestures, actions or interactions in which the number of players is greater than two and there may be one or more interactive objects [3].

2.2 Violence Detection

Violence detection is a particular problem regarding the great problem of the recognition of action. The objective of violence detection is to automatically and effectively determine whether the violence occurs or not within a short period of time. In recent years, the automatic recognition of human actions on videos has become increasingly important for applications such as video surveillance, human-computer interaction and video retrieval based on their content [4, 5].

The purpose of violence detection is automatically and effectively determine if violence occurs or not. Anyway, the detection of violence itself is an extremely difficult problem, since the concept of violence is subjective. Detection of violence is an important issue not only at the application level but also at the scientific level because it has characteristics that differentiate the recognition of generic actions.

3 Sample Setup

To obtain the sample, the research was carried out in February 2020 at NEXT ACM DL (*dlnext.acm.org*). The first query was:

[[All: violence] OR [All: fight] OR [All: aggression]] AND [[All: detection] OR [All: recognition] OR [All: surveillance]] AND [[All: indoor] OR [All: inside] OR [All: interior]

For that query, we have obtained 6641 paper. Then we consider the last five years, the query result as 2055 papers. In addition, if we consider in the title, the keywords *'detection'* or *'recognition'* or *'surveillance'*, the query results is 183 papers. To reduce the query, we introduced the abstract keywords *'review'*, or *'survey'* or *'benchmark'*. The keywords used to obtain the query were:

[[All: violence] OR [All: fight] OR [All: aggression]] AND [[Publication Title: detection] OR [Publication Title: recognition] OR [Publication Title: surveillance]] AND [[All: indoor] OR [All: inside] OR [All: interior]] AND [[Abstract: review] OR [Abstract: survey] OR [Abstract: benchmark]] AND [Publication Date: (02/01/2015 TO 02/29/2020)].

Key words	comp.	abs.	rel.	norm
violence	32	9	22.0	0.5
fight	17	24	5.9	0.13
aggression	41	0	0	0
detection (title)	29	12	29,3	0.67
recognition (title)	24	17	42.5	0.95
surveillance (title)	33	8	19.5	0.44
indoor	28	13	31.7	0.72
inside	25	16	39.0	0.89
interior	39	2	4.9	0.11
review (abstract)	34	7	17.1	0.39
survey (abstract)	30	11	26.8	0.61
benchmark (abstract)	31	10	24.4	0.56
last 5 years	59	18	43.9	1

Fig. 1. Weight of each keyword in the query. First column, *complement*, second, *absolute weight*, third, *relative weight*, and fourth, *normalized weight*.

The query and the keyword weights are shown in Figure 1. The query retrieved 41 results published in the last five years. In the first part of Figure 1, it was presented the query submitted to the ACM DL search engine. In the second part of Figure 1 it was shown the Weight of each keyword in the query. The first column, *complement*, depicts the number of papers retrieved by removing the keyword from the query; second, *absolute weight*, is the difference between the number of papers retrieved by the query and the complement; third, *relative weight*, is the percentage representation for the absolute weight in relation to the number of retrieved papers; and fourth, *normalized weight*, is the max-min representation of relative weight.

4 Results and Discussion

Despite the 41 papers obtained in the query, some papers contained information that is not related to the topic. Reading all the papers, it was possible to separate three different approaches: video, audio, and multimodal audio and video. In the video approach, we have 18 papers from the query; for the audio approach we have 3 papers from the query, and for the multimodal audio and video approach we have 3 papers from the query.

4.1 Video Approach

Automated surveillance occurs if cameras are constantly monitored by a computer, and in real-time, trigger an event when something suspicious happens. To violence detection in video, it is difficult to capture effective and discriminative features as a result of the variations of the human body. The modifications are essentially caused by scale, viewpoint, mutual occlusion and dynamic scenes.

In the last years, it was published several previous surveys about abnormal human behavior recognition [6–8], human detection behavior [9, 10], crowed behavior [11], datasets human recognition [12–14], and foreground segmentation [15]. Also, there are some research of fast violent detection [16–20], multi-features descriptors for human activity tracking and recognition [21], segmentation [22], and vision enhanced color fusion techniques [23].

Mabrouk and Zagrouba [6] study the two main steps composing a video surveillance system, which are the behavior representation and behavior modelling. On behavior representation, it's presented the most popular features for global features (optical flow, and motion information), local features (based on interest points: STIP, CSTIP,MoSIFT or Spatio-temporal volume, cube, blob), widely used for falls detection (shape), adapted for crowd monitoring (texture), and adapted for tracking a single person (object tracking and trajectory extraction). On classifying abnormal behavior recognition methods it's separated in: (i) modelling frameworks and classification methods; and (ii) scene density and moving object interaction. In modelling frameworks and classification methods its made a comparison of classification methods categories: supervised, semi-supervised: rule-based method, semi-supervised: model-based method, and unsupervised methods. Also made a comparison of frameworks and classification methods for abnormal behavior detection. On scene density and moving object interaction its presented a scene density and moving object interaction-based grouping for an uncrowded and crowded scene. On performance evaluation, it's divided into datasets and evaluation metrics. In datasets, it presented available datasets for video surveillance systems evaluation. In evaluation metrics its summarize the performance evaluation results, accuracy, equal error rate, area under, and curve. On existing video surveillance systems its presents a summary of existing video surveillance systems.

Gowsikhaa, Abirami and Baskaran [9] showed a survey of automated human behavior analysis from surveillance videos, which begin to present a map representing the human activity prediction architecture and literature survey of

low-level processing techniques, namely motion detection methods, object classification methods, and motion tracking methods. Then high-level processing techniques are presented: (a) pre-processing and human behaviour recognition and analysis, which compared activities recognized in different works; (b) a sample of semantic descriptions used in the state of art, (c) predicting the activities of a person with respect to an object, (d) performance evaluation, and (e) comparison of low-level and high-level techniques in human behavior analysis. Finally, challenges in human behavior analysis mentioned cavities, human body modelling, handling occlusions, scene classification, person identification, techniques for activity perception, cameras revisited, modelling scenes, standardization, and domain specificity.

Afsar, Cortez, and Santos [10] presented a review of automatic visual detection of human behavior. The authors begin to present an automatic human behavior detection from video keywords by publication year. Also, indicated the techniques used for human behavior detection from video: (i) initialization, (ii) tracking, (iii) pose, and (iv) recognition. The initialization model begins with main approaches and discussion, where it's mentioned comparison of approaches for model initialization. Tracking talks about background segmentation, temporal correspondence, and discussion, which present a comparison of approaches for tracking. Pose estimation mentioned model-free, indirect model use, direct model use, and discussion, where a comparison of approaches for pose estimation is made. Recognition explains scene interpretation, holistic recognition approaches, action primitives and grammar, and discussion, which showed a comparison of approaches for recognition. In addition, a dataset related with human behavior detection, and discussion, where is showed a state-of-the-art-results for human behavior detection datasets. Furthermore, applications with chronological evolution, specifying Human detection using 3D depth images, abnormal activity detection, action recognition from video, player modelling and robotics, pedestrian detection and in-home scenarios, and person tracking and identification.

Maheshwari and Heda [11] present a review of analysis methods for crowd behavior in video surveillance. All the process occur in three phases: (a) video surveillance; (b) crowd analysis; and (c) methods to abnormal behavior detection in crowd scene. (a) In video surveillance, the method consists of three major modules: i) background modeling, ii) blob analysis, and iii) crowd detection and tracking. (b) crowd analysis estimated using three basic steps which are preprocessing, object tracking and event/behavior recognition. Preprocessing step can analyse: pixel-level analysis, texture level analysis, object level analysis, and frame-level analysis. Object tracking step can analyse: region based approach, active contour based approach, feature based approach, and model based approach. Event/behavior recognition step applied two approaches: object based approach and holistic based approach. (c) The methods of abnormal behavior detection in crowded scene can used modeling Gaussian Mixture Modeling (GMM), Social Force Model Method (SFM), Hidden Markov Model (HMM), Correlated Topic

Model (CTM), Markov Random Field (MRF), Sequential Monte Carlo (SMC), and Dynamic Oriented Graph (DOG).

Dubuisson and Gonzales [12] begin for presenting in visual tracking: which dataset for which need. Then presented visual tracking issues: (a) problems inherent to visual tracking, like illumination effects, scene clutter, changes in object appearance, abrupt changes in motion, occlusions, similar appearances, camera motions, appearance/disappearance, and quality of frames; and (b) quantitative evaluation of visual tracking, like: error score, accuracy score, success score, detection scores, curves for comparing tracking algorithms, and robustness of evaluation; and (c) datasets for visual tracking; (d) datasets for scene analysis and understanding based on visual tracking, which present: Human understanding (hand gesture tracking, face tracking, facial expression or emotion, body motion, individual action/activity, behavior, and interactions and social activities), and scene understanding (video surveillance, event, crowd, and sport games).

Table 1. Categories and paper related for video approach.

Categories	**Papers**
Survey	[1–3, 6–8, 10–15]
Human behavior	[6, 9, 10, 16]
Recognition	[1, 2, 8, 10, 11, 13, 20, 21]
Crowed behavior	[7, 11, 20]
Dataset	[6, 8, 10, 12–15]
Segmentation	[8, 10, 11, 13, 15, 16, 18, 20–23]
Violent detection	[16–20]
Tracking	[7, 9, 10, 12, 14, 21]
Color fusion	[23]
Classification	[10, 5–9, 11, 13, 15–17, 19–21, 23]
Features extraction	[7, 9, 15–19, 21, 22]

Mahmood, Khan and Mahmood [23] a review has been examined to provide in detail survey and comparison of night vision imaging techniques. Then it showed theoretical foundations of technical framework and paradigms proposed to enhance and fuse color in night images, namely infra-red spectrum based self-adaptive enhancement, colormap clustering based color transfer method, contrast and color enhancement techniques, region-based color transfer methods, image enhancement based on selective Retinex fusion algorithm, color estimation and sparse representation, histogram based enhancement revisited, and nature inspired models. Next, its presented quantitative analysis: image contrast metric, the gradient metric, phase congruence (PC) metric, color natural metric, and objective evaluation index. Finally, its showed results and comparative analysis.

In the Table 4.1 is identified by category, the papers that the theme.

4.2 Audio Approach

In terms of bandwidth, memory storage and computing requirements, the audio stream is generally much less than the video stream. While standard cameras have a limited angular field of view, microphones can be omnidirectional (providing a spherical field of view). Due to the audio wavelength, many surfaces allow reflections of the acoustic wave, thus allowing the acquisition of audio events even when obstacles are present in the direct path. Illumination and temperature are not problems for audio processing [26].

Souto, Mello and Furtado [24] present, an acoustic scene classification approach involving domestic violence using machine learning. The methodology specifies the parameters used (MFCC, Energy and ZCR), parameter extraction (medium-term parameter sequence processing and short-term processing), and classification (SVM technique). They applied some tests and they used a database audio. After training the classifier, the model obtained were MFCC-SVM classifier, Energy-SVM classifier, and ZCR-SVM classifier.

Rouas, Louradour and Ambellouis [25] referred audio events detection in a public transport vehicle. The first idea is to extract relevant events from the audio stream. So it's necessary to create an automatic audio segmentation, which splits an audio signal into several quasi-stationary consecutive zones, an activity detection algorithm, which aims at skipping silence and low-level noise zones, out of interest and a merging step, to gather successive activity segments. To modelling and classification framework, first, it's obtained features extraction, then it's used a GMM method, after an SVM classifier, then a classification framework.

Crocco, Cristiani, Trucco and Murino [26] showed an audio surveillance review. The authors begin to compare audio data to visual data. Then it explains the background subtraction by monomodal analysis, which presents the features employed in this type of background subtraction. Next, it explains the background subtraction by multimodal analysis, which highlights those methods requiring more or less offline learning. After it's showed the audio event classification, where the taxonomy for the classification methods, with pros and cons, added for each category of approach. Furthermore, source localization and tracking are explained, especially audio source localization: a typology of audio events considered in the literature, features employed in audio event classification, general taxonomy of source localization, the taxonomy for time delay-based localization methods, optimal working conditions for different localization methods, features employed in sound localization; audio-visual source localization; audio source tracking, and audio-visual source tracking. In addition, situation analysis is made: one-layer systems and hierarchical systems. Also, audio features are analysed, namely features employed in situation analysis: time, frequency, cepstrum, time-frequency, energy, biologically/perceptually driven, and feature selection and feature learning. Finally, present some open problems like background subtraction, audio classification, audio source localization and tracking, situation analysis, audio-video fusion, privacy and audio encryption, and adversarial setting.

In the Table 4.2 is identified by category, the papers that referred the theme.

Table 2. Categories and paper related for audio approach.

Categories	**Papers**
Survey	[26]
Human behavior	[24, 25]
Background subtraction, Tracking, fusion methods	[26]
Dataset, Violent detection	[24]
Segmentation	[25]
Features extraction, Classification	[24–26]

4.3 Multimodal Audio and Video Approach

When it's used Audio and video approach in surveillance, its called Multimodal surveillance [28].

Crocco, Cristiani, Trucco and Murino [27] began to explain audio analysis driven violence detection, video analysis driven violence detection, multimodal analysis driven violence detection, and knowledge-based semantics extraction for violence detection. Then some general description of the proposed methodology is presented. Next, audio classification for violence hints like audio class definition, audio feature extraction, and class probability estimation. After, presented visual classification for violence hints, like visual class definition, visual features (motion features, person detection features, and gunshot detection features), and video class probability estimation. In addition, machine learning-based fusion, namely fused feature vector and meta-classifier. Also, explain ontological fusion, which includes the ontological framework, violence ontology definition, visual semantics for violence, audio semantics for violence, video structure ontology, inference engine design, and inference engine implementation. Finally, experimental evaluation with implementation issues, scenario and setup, and classification and detection results are showed.

5 Conclusions

In this paper, we present a survey of violence detection behavior. It was made a review concept of action recognition and violence detection. To carried out the research was created a sample setup and we have obtained a query of 41 papers. Analysing the papers, only 24 was specific with the theme. Also, we concluded that the papers separated in three different approaches: 18 papers were for video approach, 3 papers referred audio approach, and 3 papers, related multimodal audio and video approaches. This three different approach are analyzed independently and in each approach was analysed the methods, techniques and classification used. Besides, in some approaches also was included datasets and fusion methods.

For future works, due to the specificities related to each environment and violence, it can be proposed the exploration of *surveillance inside a vehicle*. Also, *other sensors* might be included, for instance, we can propose a multimodal system to detect violence behavior inside a vehicle. Additionally, a conception of reference architectures incorporated with security measurements is also useful.

Acknowledgments

This work is supported by the European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project nº 039334; Funding Reference: POCI-01-0247-FEDER-039334].

This work has been supported by national funds through FCT – Fundação para a Ciência e Tecnologia through project UIDB/04728/2020.

References

1. Ko, T. (2008). A survey on behavior analysis in video surveillance for homeland security applications. In Applied imagery pattern recognition workshop, 2008. AIPR'08.. 37th IEEE (pp. 1–8). IEEE.
2. Poppe, R. (2010). A survey on vision-based human action recognition. Image and Vision Computing, 28(6), 976–990.
3. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Comput. Surv. (CSUR). 43(3), 16:1–16:43 (2011)
4. Poppe, R. (2010). A survey on vision-based human action recognition. Image and vision computing, 28(6), 976-990.
5. Sun, Q., Liu, H. (2013, September). Learning spatio-temporal co-occurrence correlograms for efficient human action classification. In 2013 IEEE International Conference on Image Processing (pp. 3220-3224). IEEE.
6. Mabrouk, A. B., Zagrouba, E. (2018). Abnormal behavior recognition for intelligent video surveillance systems: A review. Expert Systems with Applications, 91, 480-491.
7. Lopez-Fuentes, L., van de Weijer, J., González-Hidalgo, M., Skinnemoen, H., Bagdanov, A. D. (2018). Review on computer vision techniques in emergency situations. Multimedia Tools and Applications, 77(13), 17069-17107.
8. Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S. (2018). RGB-D-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding, 171, 118-139.
9. Gowsikhaa, D., Abirami, S., Baskaran, R. (2014). Automated human behavior analysis from surveillance videos: a survey. Artificial Intelligence Review, 42(4), 747-765.
10. Afsar, P., Cortez, P., Santos, H. (2015). Automatic visual detection of human behavior: A review from 2000 to 2014. Expert Systems with Applications, 42(20), 6935-6956.
11. Maheshwari, S., Heda, S. (2016, March). A review on crowd behavior analysis methods for video surveillance. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (pp. 1-5).

12. Dubuisson, S., Gonzales, C. (2016). A survey of datasets for visual tracking. Machine Vision and Applications, 27(1), 23-52.
13. Zhang, J., Li, W., Ogunbona, P. O., Wang, P., Tang, C. (2016). RGB-D-based action recognition datasets: A survey. Pattern Recognition, 60, 86-105.
14. Singh, T., Vishwakarma, D. K. (2019). Video benchmarks of human action datasets: a review. Artificial Intelligence Review, 52(2), 1107-1154.
15. Komagal, E., Yogameena, B. (2018). Foreground segmentation with PTZ camera: A survey. Multimedia Tools and Applications, 77(17), 22489-22542.
16. Zhou, P., Ding, Q., Luo, H., Hou, X. (2018). Violence detection in surveillance video using low-level features. PLoS one, 13(10).
17. Deniz, O., Serrano, I., Bueno, G., Kim, T. K. (2014, January). Fast violence detection in video. In 2014 international conference on computer vision theory and applications (VISAPP) (Vol. 2, pp. 478-485). IEEE.
18. De Souza, F. D., Chavez, G. C., do Valle Jr, E. A., Araújo, A. D. A. (2010, August). Violence detection in video using spatio-temporal features. In 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images (pp. 224-230). IEEE.
19. Gao, Y., Liu, H., Sun, X., Wang, C., Liu, Y. (2016). Violence detection using oriented violent flows. Image and vision computing, 48, 37-41.
20. Hassner, T., Itcher, Y., Kliper-Gross, O. (2012, June). Violent flows: Real-time detection of violent crowd behavior. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (pp. 1-6). IEEE.
21. Jalal, A., Mahmood, M., Hasan, A. S. (2019, January). Multi-features descriptors for human activity tracking and recognition in Indoor-outdoor environments. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 371-376). IEEE.
22. Komagal, E., Yogameena, B. (2018). Region MoG and texture descriptor-based motion segmentation under sudden illumination in continuous pan and excess zoom. Multimedia Tools and Applications, 77(8), 9621-9649.
23. Mahmood, S., Khan, Y. D., Mahmood, M. K. (2018). A treatise to vision enhancement and color fusion techniques in night vision devices. Multimedia Tools and Applications, 77(2), 2689-2737.
24. Souto, H., Mello, R., Furtado, A. (2020, January). An acoustic scene classification approach involving domestic violence using machine learning. In Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional (Vol. 16, No. Salvador, pp. 705-716). SBC.
25. Rouas, J. L., Louradour, J., Ambellouis, S. (2006, September). Audio events detection in public transport vehicle. In 2006 IEEE Intelligent Transportation Systems Conference (pp. 733-738). IEEE.
26. Crocco, M., Cristani, M., Trucco, A., Murino, V. (2016). Audio surveillance: A systematic review. ACM Computing Surveys (CSUR), 48(4), 1-46.
27. Perperis, T., Giannakopoulos, T., Makris, A., Kosmopoulos, D. I., Tsekeridou, S., Perantonis, S. J., Theodoridis, S. (2011). Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies. Expert systems with applications, 38(11), 14102-14116.
28. Dedeoglu, Y., Toreyin, B. U., Gudukbay, U., Cetin, A. E. (2008). Surveillance using both video and audio. In Multimodal Processing and Interaction (pp. 1-13). Springer, Boston, MA.