Violence Detection in Surveillance Video-A survey

Article · July 2016

CITATIONS
3

READS
1,762

1 author:

MT Gopalakrishna
SJB Institute of Technology
42 PUBLICATIONS **73** CITATIONS

SEE PROFILE

International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com ǁ PP.11-17

Violence Detection in SurveillanceVideo-A survey

Anuja Jana Naik[1], M.T.Gopalakrishna[2]
[1,2](CSE, K. S. School of Engineering and Management, India)

Abstract: The demand for automatic action recognition systems have increased due to the rapid increase in the number of video surveillance cameras installed in cities and towns. Automatic action recognition system can be effectively used to generate on-line alarm in case of abnormal activities to assist human operators and for offline inspection. Although action recognition problem has become a hot topic within computer vision, detection of violent scenes receives considerable attention in surveillance system which is justified by the need of providing people with safer public spaces. This survey discusses the current state of the art methods and techniques that are being applied for the task of automated violence detection.This survey emphasizes on motivation and challenges of this very recent research area by presenting approaches for violence recognition in surveillance video. This paper aims at being a driving force for researchers who wish to approach the study of violent activity recognition and gather insights on the main challenges to solve in this emerging field.
Keywords: Dataset, Motion Binary pattern, Optical flow, Space Time Interest Point, Violence Detection.

1. Introduction

Video surveillance is an important tool to monitor people's conduct, detect their presence in an area and/or possibly study their actions which can help to prevent, detect and reduce crime. A topic of increasing interest in video surveillance is the categorization of multimedia content based on thepresence of certain human actions. Especially, after the various terrorist attacks in India from past few years detection of violent scenes receives considerable attention in surveillance systems for safety of people in public places. However as the number of cameras increase, number of operators and supervisors needed to monitor surveillance video increases. Their exists an vast and continuously increasing gap between the amount of video data continuously captured by cameras and human tendency to efficiently and intelligently analyse the visual information. Thus some events and activities are missed and suspicious behaviours are not noticed in time to prevent the incidents giving rise to the necessity for automatic understanding of human actions. Recognition of human activities is still challenging task owing to various challenges like the dynamic illumination changes or camera movements in the background scene. Performance ofevent recognition also depends on several factors like changes in camera viewpoint, body shapes and sizes of different actors, different dressing styles and changes in execution rate of activity.

In recent years, there has been a plethora of work on automatic human action recognition but automatically characterizing violence has been comparatively less studied. A few difficulties arise in automatic detection of violence or in general aggressive behaviour due to its subjective nature which imposes some barriers in defining what should be pointed as violence. Also some human behaviours, might be misclassified which appear very similar to aggressive actions. All existing work has addressed the problem by its own definition of violence Thus there is a need to solve these ambiguities to make the system more efficient and robust in real world. However progress is being made ever more rapidly and the demand for automated surveillance continues to increase in areas ranging from crime prevention, public safety and home security to industrial quality control and military intelligence gathering.

Section 2 gives the glimpse of recently published research progress of violence detection methods under a general framework. Section 3 classifies various Violence detection approaches based on descriptors used to represent violent activities. In section 4 state of art datasets that can be used to evaluate novel algorithms are listed. Finally, this paper is concluded with the future directions to work on in this field.

2. Related work

Recent proposed methods for violence detection can be roughly classified into three categories visual based approach, audio based approach and hybrid approach.

2.1 Visual Based Approach

In this approach visual information is extracted and represented as relevant features. Features can be classified aslocal features, global features. Local features could be position,velocity, veins, shape, color and example of global features are average speed, region occupancy, relative positional variations and the relationships between objects and background. Considerable work is done using this approach.Clarin et al. [1]

detects skin and blood pixels in each frame using Kohonen self-organizing map and violent actions involving blood are detected using motion intensity analysis. While previous work addressed only the shot level of videostructure, Chen et al. [2] works on a more semantic-complete scene structure of video.They detect the violent scenes by decomposing the task as action scene detection and bloody frame detection.For action scene detection features like average motion intensity, camera motion ratio, average shot length and shot cut frequency are extracted and are fed into SVM classifier. This method works well for movies but may not work in real scenarios. Recently XinJi et al.[3] have proposed novel violent detection scheme using motion angle, motion intensity, shot & explosion and blood analysis. They have achieved considerably high accuracy.

2.2 Audio based approach

In this approach audio information is used to classify violence. Cheng et al. [4] recognize gunshots,explosions and car-braking in audio using a hierarchical approach based on Gaussianmixture models and Hidden Markov models (HMM). Giannakopoulos et al. [5] used eight frame-level audio features, both from the time and frequency domain, as input to a SVM classifier which classifies the video content with respect to violence. They further extended this work to multi class classification using Bayesian networks [6]. The video content is divided into six classes. Three classes are violent: shots, fights and screams. Recently Md. ZaighamZaheer et al.[7] have proposed a deep learning based scream sound detection approach. MFCC features after interpolation are used as input of the system. The proposed system is experimented using a self-recorded scream database and with controlled and calculated parameters 100 % accuracy is achieved. The problem with this approach is that some videos may not have audio tracks or existing audio features may not relate to violence and can lead to misclassification of scenes.

2.3 Hybrid approach

In the hybrid approach, emphasis is put on combining of visual and audio features. One of the first proposals for violence recognition in video is in Nam et al. [8], which proposed recognizing violent scenes in videos using flame and blood detection and capturing the degree of motion, as well as the characteristic sounds of violent events. Zajdel et al. [9] introduced the CASSANDRA system,which employs motion features related to articulation in video and scream-like cues inaudio to detect aggression in surveillance videos. More recently, Gong et al. [10] propose a violence detector using low-level visual and auditory features and high-level audio effects identifying potential violent content in movies. Giannakopoulos et al. [11] present a methodfor violence detection in movies based on audio-visual information that uses a statistics ofaudio features, average motion and motion orientation variance features in video combinedin a k-Nearest Neighbour classifier to decide whether the given sequence is violent.

As evident from literature most works require audio cues or depend on visual cues like blood to detect violence. It is noted that in dominant applications of surveillance audio or color are not available. In other situations where audio features are available, it can increase false positive rates or miss rates because audio signal is not significant in many violence situations likepushing,attacking with a knife, knocking down someone,etc. Besides, while explosions, blood and running may be useful cues for violence inaction movies, they are rare in real-world situations.

3. Violence Detection Methods

Violence Detection is a core computer vision problem as evidenced by the plethora of papers. Currently, many algorithms have been proposed to recognize various human activities and specifically activities related to violence. This survey focuses mainly on violence detection methods in real scenes based on visual features. Various state of art Violence detection methods can be classified into four groups:

3.1 Optical Flow Based

Optical flow field is a set of the instantaneous velocity vector of all the pixels in the image sequence. It refers to the image grey scale pattern of surface movement, and containsinformation about relevant scene three-dimensional structure change along with foreground motion information. In optical flow field, each pixel corresponds to an optical flow motion vector, and can approximate the motion state of the objects in image sequence, so the corresponding motion information can be extracted from the optical flow field for abnormal behaviour detection.

Optical flow can be described as the motion of pixel I(*x, y, t*) which moves a distance of (*dx, dt*) over *dt*time, this can be expressed as:

$$I(x, y, t)=I(x+dx,y+dy,t+dt) \quad \text{----------(1)}$$

The optical flow equation is derived from the equation (1) is as follows:

$$f_x u+f_y v+f_t=0 \quad \text{----------(2)}$$

Where *fx*, *fy*are pixel intensity gradients and *ft*is the first temporal derivative.

Solving equation (2) we get two flow vector maps *u* and *v* that dictate perceived motion in both the x and y co-ordinate plane.

For each pixel I(x,y,t) the magnitudes M (x,y,t)of these vectors is determined as:

$M(x,y,t)=\sqrt{(U^2+V^2)}$----------(3)

And the Orientation θ of non-zero optical flow vector is determined as:

$$\theta = \begin{Bmatrix} -\pi + \tan^{-1}\left(\frac{v}{u}\right);\ u<0, v<0 \\ \pi + \tan^{-1}\left(\frac{v}{u}\right);\ u<0, v\leq 0 \\ \tan^{-1}\left(\frac{v}{u}\right);\ u>0 \\ -\frac{\pi}{2};\ u=0, v<0 \\ \frac{\pi}{2};\ u=0, v>0 \end{Bmatrix} \text{----------(4)}$$

$$with -\pi < \theta < \pi$$

Optical flow can describe coherent motion of moving objects,which is a good feature for motion detection and tracking. As a consequence, ithas been widely used for action recognition.In [12], optical flow histograms, based on horizontal and vertical directions, were used as action descriptors to address the problem of human action recognition. The directional value of a silhouette is divided into small regions and the normalized direction histogram of optical flow is computed in each region. The motion vector is the values of a histogram in every region respective concatenation. The motion vector is concatenated values of histogram in every region which is smoothed in time domain by moving average to reduce the motion variation and noise.Wang etal. in [13] employed optical flow to obtain densely tracking sampled points, andthen they utilized these dense trajectories to calculate local descriptors for action recognition. Although it is efficient method for action recognition it requires special hardware which makes it computationally expensive.

3.1.1 Violent Flows(ViF)

It is Motion Texture Descriptor that uses Optical Flows for motion estimation. It uses two consecutive frames magnitude maps and forms a single binary map that describes the change in motion in two frames. This is accomplished by taking the difference between two magnitude maps at each pixel and assigning a value of 1 if the resultant difference is greater than a set threshold.

$$\begin{Bmatrix} 1 & if\ |M(x,y,t)-M(x,y,t-1)| \geq \theta \\ 0 & otherwise \end{Bmatrix} \text{----------(5)}$$

A mean magnitude map is computed by summing together a series of binary maps and normalizing the result. The resulting binary map is split into M x N non-overlapping cells. The contents of each cell are used to populate a fixed sized histogram then all cell histograms are concatenated to create the final ViF descriptor.Tal Hassnar et al.[14] proposed a novel descriptor called ViF for real-time crowd violence detection. Disadvantage of ViF is that it does not make use of orientation information of optical flow which is overcome by Yuan Gao et al[15] by proposing novel feature extraction method named Oriented Violent Flows (OViF), which takes full advantage of the motion magnitude change information in statistical motion orientations.

3.1.2 Histogram of oriented optical flow(HOF)

The aim of these descriptors is to provide long term temporal description of motion trajectory and its surrounding motion. The descriptor is a combination of a trajectory pattern and a Histogram of Oriented Flows describing the re-active motions that exist around the path trajectory. To create a Histogram of Oriented Optical Flow, both *u* and *v* optical flow fields must be generated for a given area within two consecutive images. Flow orientations are computed and rounded into 360/*n* evenly spaced angular directions where *n* is the number of histogram bins used to represent the Histogram of Oriented Flows. Each histogram bin is incremented by the magnitude of respective flow. Chen and aggarwal et al[16] proposed a histogram of optical flow for recognizing actions from a far field of view. They used either histogram of orientation of optical flow to represent an action or histogram of magnitude of optical flow. This method may satisfy some unusual events, such as running in a shopping mall, anti-direction moving and so on but the drawback of this work is it is based on available human tracking.Another problem with this method is it has to specify which kind of histogram is to be used for different applications. To overcome these problems Thus Yan Chan et al[17] have proposed new feature optical flow context histogram for violence detection which is a log-polar histogram system which combines the

histogram of orientation and magnitude of optical flow together so no need to specify either of it.Most existing methods focus on detecting violence rather than locating it as it is difficult to precisely locate the regions in the scene where violence is happening. Tao Zhang et al[18] locate the violence location in the scene, which is important for public surveillance. In this method the Gaussian Mixed Model is extended into the optical flow domain in order to detect violence regions. In each region, a new descriptor, Histogram of Optical Flow Orientation (HOFO), is proposed to measure the spatial-temporal features which distinguish between violent and non-violent events. Further linear SVM classifier is used for classification. Traditional motion models do not account for the continuity of motion characteristics between frames especially in crowded scenes thusQiang Wang et al[19].havecome up with a new method for the detection of abnormalities in crowded scenes based on the crowd motion characteristics. A new feature descriptor called hybrid optical flow histogram is proposed in which statistical variances of optical flow in different directions between video frames are computed.The concept of thevisual salient area is introduced, through which only the active areas in the frame are extractedfor the sample data, to reduce the computational burden.

3.2 Space Time Interest Point(STIP)

In Space Time Interest Point (STIP) method interest points are identified by analysing spatial and temporal variations. It is an extension of the Harris-Stephens corner detection method. Around each space time interest point a 3 dimensional volume is extracted; the volume shows how a 2D image segment evolves over time. The size of the 3D volume is based on the detected features scale. The detected interest points are characterized by a high variation of the intensity in space, and non-constant motion in time. These salient points are detected at multiple spatial and temporal scales.

The spatial size of the volume is determined by $2k\sigma$ where σ is the feature scale determined by the STIP detector. The temporal scale is determined using $k\tau$ where τ is the temporal duration of a feature. The parameter k is assigned the value of 9.

The appearance of the 3D cube is identified using a Histogram of Oriented Gradients and the motion is described using a Histogram of Oriented Flows.These features can be used for recognizing motion events with high performanceand they are robust to scale, frequency and velocity variations of the pattern [20].E.Bermejo et al [21] have analysed bag of words framework specifically for fight detection using STIP and SIFT descriptors.Despite encouraging results with high accuracy rate for this specific task, the computational cost of extracting such features is prohibitive for practical applications, particularly in surveillance and media rating systems.

3.3 Motion binary pattern (MBP)

Motion Binary Pattern is a motion texture descriptor; it assumes that motion can be detected from the change in pixel intensities. MBP requires three frames to describe motion. Two difference binary maps are created by comparing frames n with n+1 and n+1 with n+2. At each pixel, a value of 1 is assigned if the pixel intensity on the first frame is greater than the corresponding pixel intensity on the next frame. A 0 (zero) is assigned otherwise. An 'exclusive or' function is applied to the two binary patterns which produces a third combination binary pattern depicting areas of perceived motion across three frames. At each pixel in this new binary pattern the L-1 norm is generated using a 3 by 3 grid. The final binary motion pattern is created by assigning 1 where the L-1 Norm of each 3x3 window is greater than a set threshold and 0(zero) otherwise. Once the final binary pattern has been computed for a set of frames they are added together along the temporal plane to produce a single texture.

Initially Riccardo Mattivi et al [22] applied Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) descriptor in field of human action recognition. They have extended LBP-TOP considering the action at three different frames in XY plane and at different views in XT and YT planes. The authors reached 88.19% accuracy on the KTH dataset and by combining Extended Gradient LBPTOP with Principal Component Analysis they reached 91.25%.Further they extended the work in [23] where Extended Gradient LBP-TOP was combined with Dollar's detection method and have achieved better accuracy. MBP was proposed by F. Baumann et al[24] which combines the advantages of Volume Local Binary Patterns together with static object information and Optical Flow to obtain motion information.

3.4 Grey Level Co-Occurrence Matrix (GLCM)

A GLCM is computed by counting the co-occurring grey level intensity values that a pixel with a grey level value i occurs in conjunction with a second pixel intensity j given a linear spatial relationship between the two pixels. An offset map dictates the spatial relationship between a pixel of interest and its neighbours using vectors given by (θ, d) where θ *is* the orientation and d is the distance between two pixels. All pixel intensities are scaled to n different grey level values from [0:255] before computing GLCM matrix. Once a co-occurrence matrix is created, several statistics can be derived to describe the nature of image texture.

For crowd description Heng Wang et al. in [13] have used Grey Level Co-Occurrence Matrix (GLCM) approach and have proved that haralick' s GLCM features can be used to determine density of a crowd. Using this concept as base Kaelon Lloyd et.al[25] has applied temporal encoding to GLCM texture features for detection of violence in crowd. The method is based on Haralick texture features like Energy, Contrast, Homogeneity and Correlation. They refer to this method as violent crowd texture. After testing several state of the art techniques their proposed method proved to be highly effective in discriminating violence.

4. Datasets

Based on our analyses most challenging datasets for violence detection in the literature are listed below in table 1. Different datasets represent different types of violence seen within city, street and indoor environments.

Dataset	**Classes**	**Videos**	**Properties**
KTH	• Number of action classes = 6. • Actions: walking, jogging, running, boxing, hand waving, and hand clapping.	• 600 videos (192 training, 192 validation, 216 testing). • Resolution = 160x120. • Black and white videos. • Static camera	• Homogeneous indoor/outdoor backgrounds. • Performed by 25 persons, 4 scenes, and 6 actions.
UCF-Sports	• Number of action classes = 9. • Actions: diving, golf swinging, kicking, lifting, horse riding, running, skating, baseball swinging, walking.	• 182 videos • Resolution = 720x480. • Static camera.	• Actions collected from various sports in broadcast television channels. • High intra-class similarity between videos for certain classes such as diving and weight lifting.
Hollywood 2	• Number of action classes = 12. • Actions: answer phone, drive car, eat, fight person, get out car, hand shake, hug person, run, sit down, sit up, stand up. • Number of scene classes = 10. • Scenes: house, road, bedroom, car, hotel, kitchen, living room, office, restaurant, shop.	• 2859 videos • .For actions: 823 training, 884 testing. • For scenes: 570 training, 582 testing. • Resolution = 400-300x300-200 (varies between videos).	• Short sequences from 69 movies. • An expansion of the Hollywood dataset. • Provide videos of scene for both training and testing.
UT-Interaction	• Number of action classes = 6. • Actions: hand shaking, hugging, kicking, pointing, punching, pushing.	• 20 videos. (2 sets, 10 for each set.) • Resolution= 720x480. • Static camera.	• All actions are defined for interactions between two people. • Two sets of video taken from different background environment.
Hockey Fight Dataset	• Actions happening in the ice hockey rink	• 1000 videos (500 violence and 500 non-violence) • Resolution= 720x576 pixels	• Designed for evaluating violence detection system. • Non crowd violence dataset
Violent flows	• Violent actions in crowded places	• **246** videos(123 violence and 123 non- violence) • Resolution=320x240	• Database of real-world, video footage of crowd violence

Table 1. Available Datasets for Violence Detection

5. Conclusion

Human activity recognition is a challenging task due to the flexibility of the human body. The survey provides the brief discussion about the important issues related to detection of violent activities including current issues, methodology. A summary of different datasets that can be used for violence detection is also listed in the survey. Currently, many models have been proposed to recognize various human activities and specifically activities related to violence. In each model there are many limitations like some models do not work in complex environments like scenes involving crowds and large amounts of occlusion in real time which still needs to be considered by researchers in future.However until automatic video surveillance systems provides the same reasoning capabilities as that of monitored person are able to do there is still a lot of research ahead. It is an ultimately endless pursuit and hence continued improvement of such algorithms is important.

REFERENCES

[1]. Clarin C, Dionisio J, Echavez M, Naval P ,*DOVE: Detection of movie violence using motion intensity Analysis on skin and blood*, Technical report, University of the Philippines, 2005

[2]. Chen L, Su C, Hsu H ,Violent scene detection in movies, *International Journal of Pattern Recognition and Artificial Intelligence, vol 25*, 2011,pp: 1161–1172.

[3]. XinJi, Ou Wu, Chenlong Wang, Jinfeng Yang, Visual Feature-Based Violent Video Detection, *In Proceedings of Communication Control and Intelligent System ,IEEE,*2014,pp:619-623.

[4]. Cheng W, Chu W, Ling J, Semantic context detection based on hierarchical audio models, *In Proceedings of the ACM SIGMM Workshop on Multimedia Information Retrieval*, 2003, pp. 109-115.

[5]. T. Giannakopoulos, A. Pikrakis, S. Theodoridis, Violence content classification using audio features, *In Proceedings of the 4th Hellenic Conference on Artificial Intelligence*, Crete, Greece, 2006, pp: 502–507.

[6]. T. Giannakopoulos, A. Pikrakis, S. Theodoridis, A Multi-Class Audio Classification Method With Respect To Violent Content In Movies Using Bayesian Networks, In *Proceedings of IEEE International WorkshopOn Multimedia Signal Processing*, Crete, Greece, 2007, pp: 90–93.

[7]. Md. ZaighamZaheer, Jin Young Kim, Hyoung-Gook Kim, Seung You Na, A Preliminary Study on Deep-Leaning Based Screaming Sound Detection, *In proceedings of 5th International Conference on IT Convergence and Security (ICITCS), IEEE,*2015.

[8]. Nam J, Alghoniemy M, Tewfik A, Audio-Visual Content-Based Violent Scene Characterization, *In Proceedings Of International Conference of Image Processing,IEEE,* 1998, pp: 353-357.

[9]. Zajdel W, Krijnders J, Andringa T, Gavrila D, CASSANDRA: Audio-Video Sensor Fusion For Aggression Detection, *In Proceedings of IEEE Conference on Advanced Video And Signal Based Surveillance (AVSS),* 2007,pp: 200-205.

[10]. Gong Y, Wang W, Jiang S, Huang Q, Wen Gao, Detecting Violent Scenes In Movies By Auditory And Visual Cues, *In Proceedings of the 9th Pacific Rim Conference on Multimedia. Berlin, Heidelberg: Springer-Verlag*, 2008, pp. 317-326.

[11]. Giannakopoulos T, Makris A, Kosmopoulos D, Perantonis S, Theodoridis S, Audio-Visual Fusion For Detecting Violent Scenes In Videos, *In proceeding of 6th Hellenic Conference on AI, Springer-Verlag,* 2008, pp. 91-100.

[12]. KanokphanLertniphonphan , SupavadeeAramvith,Thanarat, H. Chalidabhongse, Human action recognition using direction histograms of optical flow, *11th International Symposium on Communications and Information Technologies (ISCIT),IEEE,* 2011.

[13]. Heng Wang, CordeliaSchmid, Action Recognition with Improved Trajectories, *In proceedings of IEEE International Conference on Computer Vision,* 2013.

[14]. Tal Hassnar, Y. Itcher, O. Kliper-Gross, Violent Flows: Real-Time Detection Of Violent Crowd Behavior, *Computer Vision on Pattern Recognition Workshops (CVPRW*), 2012, pp. 1-6.

[15]. Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, Yi Liu, Violence detection using Oriented violent Flows, *Image and Vision Computing* , 2016.

[16]. C.-C. Chen, J. K. Aggarwal, Recognizing Human Action from a Far Field of View, In *proceedings of IEEE Workshop on Motion and Video Computing(WMVC*), 2009.

[17]. Yan Chen, Ling Zhang, Biyi Lin, Yong Xu, XiaoboRen, Fighting Detection Based on Optical Flow Context Histogram, *In proceedings of Second International Conference on Innovations in Bio-inspired Computing and Applications,IEEE,*2011,pp:95-98.

[18]. Tao Zhang, ZhijieYang,WenjingJia, Baoqing Yang, Jie Yang, Xiangjian He, A New Method For Violence Detection In Surveillance Scenes, *Multimedia Tools and Applications, Springer, vol 75*, 2015,pp7327-7349.

[19]. Qiang Wang, Qiao Ma, Chao-HuiLuo, Hai-Yan Liu, Can-Long Zhang, Hybrid Histogram of Oriented Optical Flow for Abnormal Behavior Detection in Crowd Scenes, *International Journal of Pattern Recognition And Artificial Intelligence, World Scientific ,Vol. 30, No. 2*, 2016,pp:1655007(1-14).

[20]. I. Laptev ,On Space-Time Interest Points, *International Journal of Computer Vision, vol 64, number 2/3,* 2005,pp.107-123.

[21]. E. Bermejo , O. Deniz , G. Bueno, R. Sukthankar, Violence Detection in Video Using Computer Vision Techniques, *InProceedings of Computer Analysis of Images and Patterns*, 2011.

[22]. Mattivi, R., Shao, L, Human Action Recognition Using Lbp-Top As Sparse Spatio-Temporal Feature Descriptor, *Computer Analysis of Images and Patterns (CAIP),Springer-Verlag Berlin Heidelberg*,2009, pp. 740–747.

[23]. Shao, L, Mattivi, R, Feature Detector And Descriptor Evaluation In Human Action Recognition, *In Proceedings of the ACM International Conference on Image and Video Retrieval*, 2010,pp:477-484.

[24]. F. Baumann, J.Liao, A.Ehlers, B. Rosenhahn, Motion Binary Patters for Action Recognition, *In Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM),2014, pp:385-392.

[25]. Kaelon Lloyd, Paul L. Rosin, David Marshall, Simon C. Moore, Detecting Violent Crowds using Temporal Analysis of GLCM Texture,*Computer Vision and Pattern Recognition, ACM*, 2016.