ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

A Survey of Computer Vision based methods for Violent action Representation and Recognition

Febin I P[1], Jayasree K[2]

PG Student (Mtech-CS Image Processing), Department of Computer Engineering, Model Engineering College, Thrikkakara, India[1]

Assistant Professor, Department of Computer Engineering, Model Engineering College, Thrikkakara, India[2]

Abstract: Action Recognition is an active research area in computer vision and among the action recognition tasks Violent action recognition has high importance as it can be used for safety and security enforcement. Design of Intelligent robots and Intelligent Surveillance systems that can automatically detect violences, Violent video indexing systems etc. are the current trends in computer vision and Artificial Intelligence. This paper surveys the different computer-vision based methods available for representing and recognizing violent actions. Core technology beneath every action or activity recognition system is Feature extraction and classification. Feature extraction is the process of fetching those features which are highly discriminative. In case of Violent activity recognition computer vision provides many such feature extraction methods which we can broadly classify as Local and Global features ,Spatial and temporal features. Current trends in Violent action recognition is based on combination of these local, global spatio-temporal features.

I. INTRODUCTION

Violent action representation and recognition can consider as a subcategory of general action recognition researches going on but it has high importance compared to other actions since it is closely related to our safety, security and social well being. Many researches are going on this area as Abnormal activity detection, Violence detection, Fight detection etc. Advancement in computer vision has provided many methods for representing and hence recognising different human activities like walking, sitting, hand waving, jumping etc. Computer vision based action recognition algorithms were used in many abnormal activity detection methods presented so far. It was mainly based on the idea that an abnormal action will be different from the normal simple actions and if any action that is not normal it will be abnormal. But it cannot be considered as a good method, as recognising all normal actions for deciding abnormal action is difficult. Researches on Violent action detections are very helpful in many areas like video indexing for content based video retrieval, Intelligent surveillance camera design, Intelligent robot design etc.
Generally human actions can be be divided into 3 categories:

1. Single person static actions:
It includes actions like sitting, standing etc. Detection of these actions can be done using images or videos as datasets.

2. Single person dynamic actions:
It includes actions like running, walking etc. For detecting these actions we need video datasets which also contains movement information other than the appearance information. Popular datasets using for detection of theses actions are KTH[21] dataset ,Weizmann[22] dataset and IXMAS[23] dataset. They all contain around 6–11 action performed by various actors. They are all not very realistic and share strong simplifying assumptions, such as static background, no occlusions, given temporal segmentation, and only a single actor[6].

3. Multiple person Interactions:
These are comparatively complex to detect. Interactions generally include simple interactions like hugging, handshaking and complex interactions like kicking, punching. UT Interaction dataset[19] can be used for detecting such interactions. All violent actions comes under multiple person Interaction category. Kikking and punching are the basic violent actions complex vigorous fights and wrestling fights are also included in violences.

- **The KTH dataset**[21]

It contains the six actions walking, jogging, running, boxing, hand waving and hand clapping, performed several times by 25 subjects in four different scenarios. Overall it contains 2391 sequences. It has fewer action classes than the other datasets, but the most samples per class. It is hence well suited for learning intensive approaches, e.g. approaches based

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
IJARCCE
nCORETech
LBS College of Engineering, Kasaragod
Vol. 5, Special Issue 1, Feburary 2016

on SVMs. In difference to the two other datasets it does not provide background models and extracted silhouettes, and moreover some of the scenes are recorded with a shaking and zooming camera. Most approaches that evaluate on the KTH dataset are hence based on local features which are best suited to such scenarios.

Fig.1 KTH dataset

- **The Weizmann dataset[22]**

It contains the nine actions running, walking, bending, jumping-jack, jumping-forward-on-two-legs, jumping-in-place-on-two-legs, galloping-sideways, waving-two-hands, waving-one-hand, performed by nine different actors. Overall it contains 93 sequences, all performed in front of similar plain backgrounds, and with a static camera. It is the smallest of the three datasets considered.

Fig.2 Weizmann dataset

- **The IXMAS dataset[23]**

The INRIA XMAS dataset contains the 11 daily-life actions: check watch, cross arms, scratch head, sit-down, get up, turn around, walk, wave, punch, kick, pick-up, performed each three times by 11 non-professional actors. Note that there are two more actors and actions on the dataset's website, but those have not been used by most of the approaches. The actions were filmed with five carefully calibrated and synchronized cameras. Overall it contains hence 429 multi-view sequences, or, if the views are considered individually, 2145 sequences. It also provides background subtracted silhouettes and reconstructed visual hulls. The scenes are recorded in front of simple static studio-like backgrounds. Its main difficulty comes from the changing view point, that is caused by the different camera configurations and the fact that actors freely chose their orientation while performing the actions. Respectively, the dataset is in particular used by view independent approaches.

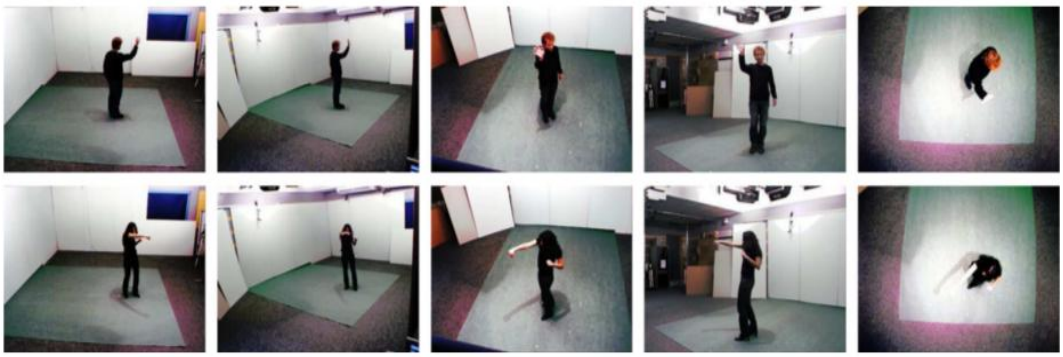

Fig.3 IXMAS dataset

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
IJARCCE
nCORETech
LBS College of Engineering, Kasaragod
Vol. 5, Special Issue 1, Feburary 2016

- **UT-Interaction dataset[19]**

The UT-Interaction dataset contains videos of continuous executions of 6 classes of human-human interactions: shake-hands, point, hug, push, kick and punch. Ground truth labels for these interactions are provided, including time intervals and bounding boxes. There is a total of 20 video sequences whose lengths are around 1 minute. Each video contains at least one execution per interaction, providing us 8 executions of human activities per video on average. Several participants with more than 15 different clothing conditions appear in the videos. The videos are taken with the resolution of 720*480, 30fps, and the height of a person in the video is about 200 pixels.

Fig.4 UT-Interaction dataset (a) Hugging (b) Punching (c) Pushing (d) Kicking

1. Global features:

Global features gives a global representation of the action.Global representations are obtained in a top-down fashion: a person is localized first in the image using background subtraction or tracking. Then, the region of interest is encoded as a whole, which results in the image descriptor. The representations are powerful since they encode much of the information. However, they rely on accurate localization, background subtraction or tracking. Also, they are more sensitive to viewpoint, noise and occlusions. When the domain allows for good control of these factors, global representations usually perform well[20]. Global representations encode the region of interest (ROI) of a person as a whole. The ROI is usually obtained through background subtraction or tracking. Common global representations are derived from silhouettes, edges or optical flow. They are sensitive to noise, partial occlusions and variations in viewpoint. To partly overcome these issues, grid-based approaches spatially divide the observation into cells, each of which encodes part of the observation locally. Multiple images over time can be stacked, to form a three-dimensional space–time volume, where time is the third dimension. Such volumes can be used for action recognition. The silhouette of a person in the image can be obtained by using background subtraction. In general, silhouettes contain some noise due to imperfect extraction. Also, they are somewhat sensitive to different viewpoints, and implicitly encode the anthropometry of the person. Still, they encode a great deal of information. Instead of (silhouette) shape, motion information can be used. The observation within the ROI can be described with optical flow, the pixel-wise oriented difference between subsequent frames[20].

2. Local features:

Local features gives a local representation of the action. Local representations describe the observation as a collection of independent patches. The calculation of local representations proceeds in a bottom-up fashion: spatio-temporal interest points are detected first, and local patches are calculated around these points. Finally, the patches are combined into a final representation. After initial success of bag-of-feature approaches, there is currently more focus on correlations between patches. Local representations are less sensitive to noise and partial occlusion, and do not strictly require background subtraction or tracking. However, as they depend on the extraction of a sufficient amount of relevant interest points, pre-processing is sometimes needed, for example to compensate for camera movements. Accurate localization and background subtraction are not required and local representations are some what invariant to changes in viewpoint, person appearance and partial occlusions. Patches are sampled either densely or at space–time interest points. Space–time interest point detectors Space–time interest points are the locations in space and time where sudden changes of movement occur in the video. It is assumed that these locations are most informative for the recognition of human action[20].

Violence detection methods proposed in earlier days were not based on vision based action clues, instead they used sound and color as the key features which made this method only useful in movies and other extreme violent scenarios. Action recognition and other researches in computer vision had lead us to focus more on to day to day actions.

Initially Violent action recognition was addressed as a Key action detection task by recognising Kicking and punching. Then the focus shifted on to more realistic violences using Descriptor based methods. Other methods for fast detection of violence's was also proposed in recent years to have real-time application[20].

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
IJARCCE
nCORETech
LBS College of Engineering, Kasaragod
Vol. 5, Special Issue 1, Feburary 2016

II. VIOLENCE AS KEY ACTION DETECTION

Detecting key actions like kicking and punching as part of general Interaction detection was the initial step towards violent action recognition. Many methods were proposed for this purpose which includes Background subtraction or Human detection, Feature extraction and Classification. [4] address the problem of detecting human violence in video, such as fist fighting, kicking, hitting with objects. To detect violence this method rely on motion trajectory information and on orientation information of a person's limbs. It defines an Acceleration Measure Vector (AMV) composed of direction and magnitude of motion and defines jerk to be the temporal derivative of AMV.

Initially Background Subtraction is done to acquire the silhouettes of moving bodies. After that it does test to determine if the moving body is a person. Then divide the bounding rectangle of the silhouette horizontally into three equal parts (H1, H2, H3). In order to extract the macroscopic features of a silhouette pattern from each part, get the projection histogram of each part that is obtained by counting the number of black pixels in each column of the silhouette pattern. Then it Fits a person model to silhouette and assign labels. The Mean, Standard deviation and Aspect Ratio of the Silhouette-Bounding Rectangle kind of features are extracted from the histogram and compared with a lookup table for 20 human models to fit person model. This method does Determination of neck and shoulder and then initialization of Head Tracking Box. It tracks head using Color Sum of Squared Differences and compute Acceleration Measure Vector and jerk. It computes Orientation map for arms and legs. Detect violence using objects if there is any object detected also detects non-violent activities like walking, handshakes, object handovers and finger pointing etc.

III. VIOLENCE DETECTION USING DESCRIPTORS

Key action detection methods were evaluated on popular datasets like IXMAS dataset[23], UT-Interaction dataset[19] etc. They are all not very realistic datasets and share strong simplifying assumptions, such as static background, no occlusions, given temporal segmentation and only few actors. Hence these methods can not promise a good performance in more realistic violent datasets. Local feature based or descriptor based methods were proposed to handle realistic datasets in a a better way. Different methods proposed so far compared and used descriptors like SIFT, STIP and MoSIFT along with BoW framework. Fillipe et al. [1] used STIP and BoW with linear-SVM and it also compared with SIFT descriptor and obtained better performance in a newly created dataset. Bermejo et al. [3] compared the use of STIP(HOG), STIP(HOF) and MoSIFT on two kind of datasets. A newly created Hockey dataset, collected 1000 clips of action from hockey games of the National Hockey League (NHL), and Movie dataset, 200-clip collection of scenes from action movies are used for comparing performance.

STIP (Spatio-temporal interest point detector)

Laptev presented a differential operator for simultaneously considering extremas over the spatial and temporal scales[25].Many interest events in videos are characterized by motion variations of image structures over time. In order to retain those important information, the concept of spatial interest points is extended to the spatio-temporal domain. This way, the local regions around the interest points are described with respect to derivatives in both directions (space and time). Initially for finding scale-space interest points convolution of simple model of an image and Gaussian kernel of variance σ_l^2 is done then localizes the interest points. Localizing interest points means to find strong variations of image intensities along the two directions of the image. To detect interest points in the scape-time domain an anisotropic Gaussian kernel $g(\sigma_l^2, \tau_l^2)$ is used over two independent variances σ_l^2 (spatial) and τ_l^2(temporal).
SIFT(Scale Invarient Feature Transform). This algorithm was published by David Lowe in 1999 [24]. SIFT extracts distinctive local features by computing oriented-gradient histograms. This process is accomplished in four principal computation steps, namely, detection of maximas in scale-space, selection of interest points, assignment of orientations to the interest points, and description of the interest points by measuring local gradients in their *neighbourhoods*. So as this descriptor is very sensitive to borders, many noisy features are detected in images with cluttered backgrounds. In this case, focusing on regions belonging uniquely to the object of interest in the scene is an advantage of considering temporal information on detection of interest points[1].

Fig.5 Interest point detection in [1].

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
IJARCCE
nCORETech
LBS College of Engineering, Kasaragod
Vol. 5, Special Issue 1, Feburary 2016

TABLE 1: COMPARISON OF METHODS[1]

Method	Using Oriented Gradient 1000-word codebook		
	%	Violent	Non-Violent
SIFT	Violent	80.09	19.91
	Non-violent	14.65	85.35
STIP	Violent	99.54	0.46
	Non-violent	0	100

400 video dataset , 200 composing each category was used in this experiment. Violence and Non-violence samples found on social networks were collected to make this dataset. In this dataset STIP outperforms the SIFT method.

MoSIFT(Motion SIFT)

Ming-yu Chen and Alex Hauptmann[16] presented a MoSIFT algorithm to detect and describe spatio-temporal interest points. In part-based methods, there are three major steps: detecting interest points, constructing a feature descriptor, and building a classifier. Detecting interest points reduces the whole video from a volume of pixels to compact but descriptive interest points. Therefore, we desire to develop a detection method, which detects a sufficient number of interest points containing the necessary information to recognize a human action. The MoSIFT algorithm detects spatially distinctive interest points with substantial motions. We first apply the well-know SIFT algorithm to find visually distinctive components in the spatial domain and detect spatio-temporal interest points with (temporal) motion constraints. The motion constraint consists of a ' sufficient' amount of optical flow around the distinctive points.

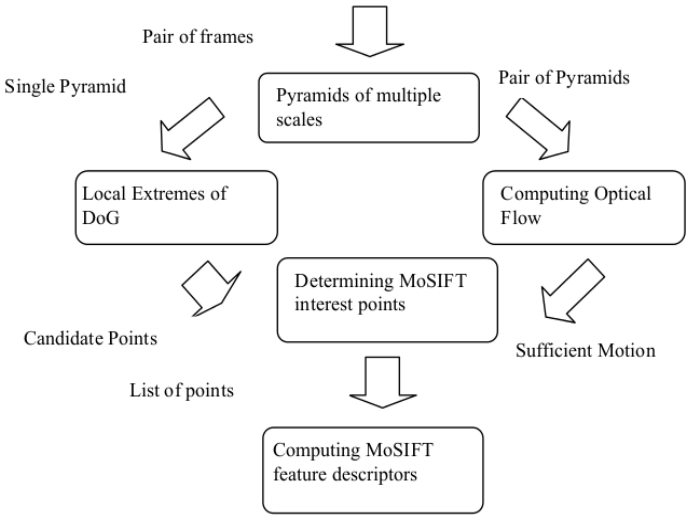

Fig.6 System flow graph of the MoSIFT algorithm. A pair of frames is the input. Local extremes of DoG and optical flow determine the MoSIFT points for which features are described.

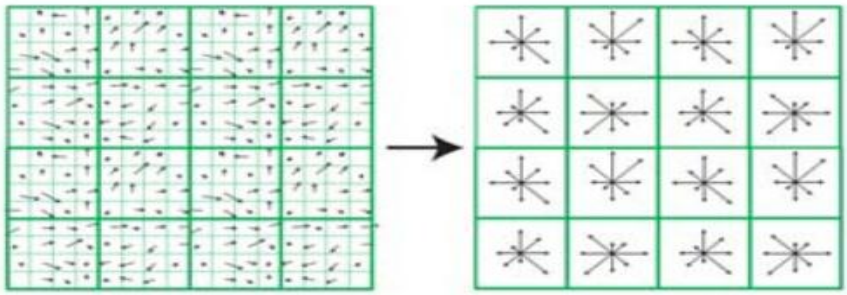

Fig.7 Grid aggregation for SIFT/MoSIFT feature descriptors. Pixels in a neighborhood are grouped into 4x4 regions.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
IJARCCE
nCORETech
LBS College of Engineering, Kasaragod
Vol. 5, Special Issue 1, Feburary 2016

TABLE 2: COMPARISON OF SIFT AND STIP[3]

Method	Vocabulary on 1000-clip hockey		
	50	100	1000
STIP(HOG)+HIK	87.8%	89.1%	91.7%
STIP(HOF)+HIK	83.5%	84.3%	88.6%
MoSIFT+HIK	87.5%	89.4%	90.9%

On this dataset,STIP(HOG) and MoSIFT perform comparably.

TABLE 3: COMPARISON OF SIFT AND STIP[3]

Method	Vocabulary on action movie dataset 200-clip		
	50	100	1000
STIP(HOG)+HIK	44.5%	45.0%	38.5%
STIP(HOF)+HIK	51.2%	56.5%	52.5%
MoSIFT+HIK	76.0%	79.5%	89.0%

In this dataset MoSIFT dramatically outperforming the best STIP under all conditions.

IV. FAST VIOLENCE DETECTION

The presence of large accelerations is key in the task of violence recognition. In [17], body part tracking is considered, and introduced the so called Acceleration Measure Vectors (AMV) for violence detection. In general, acceleration can be in ferred from tracked point trajectories. However, we have to note that extreme acceleration implies image blur, which makes tracking less precise or even impossible. Motion blur entails a shift in image content to wards low frequencies. Such behavior allows to build an efficient acceleration estimator for video. First, we compute the power spectrum of two consecutive frames. It can be shown that, when there is a sudden motion between the two frames, the power spectrum image of the second frame will depict an ellipse(Barlow and Olshausen, 2004). The orientation of the ellipse is perpendicular to the motion direction, the frequencies outside the ellipse being attenuated. Most importantly, the eccentricity of this ellipse is dependent on the acceleration. Basically, the proposed method aims at detecting the sudden presence of such ellipse. In [18], It is hypothesized that motion blobs in fight sequences have a distinct position and shape. Firstly, the absolute image difference between consecutive frames is computed. The resulting image is then binarized, leading to a number of motion blobs. Only the K largest motion blobs are selected for further processing. In order to characterize the K blobs, different measures are computed such as area, centroid, perimeter as well as distances between blob centroids. Experiments show that the method does not outperform the best methods considered. However, it is much faster while still maintaining useful accuracies ranging from 70% to near 98% depending on the dataset.

V. CONCLUSION

Violent action detection from realistic datasets is still an open problem, compared to body pose based key action detection methods, descriptor based methods gave us more promising results in this area. Compared to SIFT, STIP showed us higher performance in [1] but its performance is not enough in the action movie dataset. MoSIFT descriptor provides good results in two completely different data sets and hence shows adaptability to different data. But MoSIFT is computationally expensive. Few fast violence detection methods [17] [18] have been proposed for real-time use but these methods compromise on accuracy.

VI. ACKNOWLEDGMENT

We extend our gratitude to Model Engineering College, Thrikkakara for all the supports provided. We are sincerely thankful to each and everyone helped us in completing this work.

REFERENCES

[1]. Fillipe D. M. de Souza, Guillermo C. Ch avez, Eduardo A. do Valle Jr. and Arnaldo de A. Ara ujo, ”Violence Detection in Video Using Spatio-Temporal Features”, 23rd SIBGRAPI Conference on Graphics, Patterns and Images, 2010.
[2]. Z. Zhang and D. Tao, ” Slow feature analysis for human action recognition”, IEEE Pattern Anal. Mach. Intell., Vol.34,No. 3, March 2012
[3]. E. Bermejo, O. Deniz, G. Bueno, and R. Sukthankar, ”Violence Detection in Video Using Computer Vision Techniques”, CAIP proceedings of the 14th international conference on Computer Analysis of Images and patterns Seville Spain,pp. 332-339, August 2011.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

[4]. A. Datta, M. Shah, and N. Da Vitoria Lobo, "Person-on-person violence detection in video data", in ICPR 02: Proceedings of the 16th International Conference on Pattern Recognition(ICPR02) Volume 1.Washington, DC, USA:IEEE Computer Society, 2002, p. 10433.
[5]. Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, Dimitris Samaras, "Two-person Interaction Detection Using Body-Pose Features and Multiple Instance Learning",
[6]. Daniel Weinland a , Remi Ronfard b , Edmond Boyer c, "A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition", Computer Vision and Image Understanding, October 2010.
[7]. Liang-Hua Chen, Hsi-Wen Hsu, Chih-Wen Su and Li-Yun Wang, "Violence Detection in Movies", 8th International Conference Computer Graphics, Imaging and Visualization, 2011.
[8]. Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, Kyoung Ho Choi, "A Review on Video-Based Human Activity Recognition", computers ISSN 2073431X, June 2013.
[9]. C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: A local svm approach", International Conference on Pattern Recognition, Vol. 3, pp. 3236, 2004.
[10]. L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, Actions as Space-Time Shapes, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2247-2253, Dec.2007.
[11]. D. Weinland, R. Ronfard, E. Boyer, "Free viewpoint action recognition using motion history volumes", Computer Vision and Image Understanding, pp. 249257, 2006.
[12]. Laptev, I. Marszalek, M. Schmid, C. Rozenfeld, B., "Learning Realistic Human Actions from Movies", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage USA, pp. 18, June 2008.
[13]. Feng Wang, Zhanhu Sun, Yu-Gang Jiang, and Chong-Wah Ngo, "Video Event Detection Using Motion Relativity and Feature Selection", IEEE Transactions on multimedia, Vol. 16, No. 5, August 2014.
[14]. Manoranjan Paul, Shah M E Haque and Subrata Chakraborty, "Human detection in surveillance videos and its applications a review", EURASIP Journal on Advances in Signal Processing, 2013.
[15]. Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank "A Survey on Visual Surveillance of Object Motion and Behaviors", IEEE Transaction on systems, Vol. 34, No. 3, August 2004.
[16]. Chen, M., Hauptmann, A.: MoSIFT: Recognizing human actions in surveillance videos. Tech.rep., Carnegie Mellon University, Pittsburgh, USA, 2009.
[17]. O. Deniz , I. Serrano , G. Bueno and T-K. Kim,"Fast violence detection in video",Computer Vision Theory and Applications (VISAPP) International Conference ,Vol. 2 ,2014.
[18]. Serrano Gracia I, Deniz Suarez O, Bueno Garcia G, Kim T-K, "Fast Fight Detection", PLoSONE 10(4): e0120448. doi:10.1371/journal. pone.0120448,2015.
[19]. M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/ SDHA2010/Human_Interaction.html, 2010.
[20]. Ronald Poppe, "A survey on vision-based human action recognition",Image and Vision Computing, vol.28, pp. 976–990, 2010.
[21]. C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: International Conference on Pattern Recognition, pp. 32–36, 2004.
[22]. M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, in: International Conference on Computer Vision, pp. 1395–1402,2005.
[23]. D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding, vol.104, no. 2–3,pp. 249–257,200.
[24]. David G. Lowe"Distinctive image features from scale-invariant keypoints"., International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.
[25]. I. Laptev, "On space-time interest points," Int. J. Comput.Vision, vol. 64, no. 2-3, pp. 107–123, 2005.