LOCALIZATION GUIDED FIGHT ACTION DETECTION IN SURVEILLANCE VIDEOS

Qichao Xu[1], John See[2], Weiyao Lin[1]*

[1]Dept. of Electronic Engineering, Shanghai Jiao Tong University, China ([*]Corresponding author)
[2]Multimedia University, Malaysia
969915370@sjtu.edu.cn, johnsee@mmu.edu.my, wylin@sjtu.edu.cn

ABSTRACT

Automatic detection of fight behaviors in surveillance videos is an important task for surveillance systems. In this work, we propose a novel localization guided framework for detecting fight actions in surveillance videos. Specifically, we exploit optical flow maps to extract motion activation information, which indicates the location of active regions. Then, a detection guided alignment module is designed to adjust the localized active regions. This approach employs a two-stream based 3D convolution network as the backbone network with a novel motion acceleration representation on the temporal stream. While most existing methods are still evaluated on three benchmark datasets which were not originally collected from surveillance scenarios, we present a novel *Fight Action Detection in Surveillance-videos (FADS)* dataset for this purpose. With a total of 1,520 video clips, the FADS is the largest known dataset in terms of number of surveillance videos with fight scenes. Experimental results on both the benchmark datasets and the FADS show that our proposed localization guided method outperforms state-of-the-art techniques.

Index Terms— fight detection, action localization and recognition, surveillance dataset, group behavior analysis

1. INTRODUCTION

As monitoring of public violence become increasingly important for safety and security, surveillance systems are now widely deployed in public infrastructure and places such as schools, bars and prisons. However, many existing surveillance systems still require human operators and manual inspection. Firstly, the number of well-trained security personnel is usually insufficient; it is common to find a small group of supervisors monitoring tens to hundreds of video feeds with alert systems of minimal functionality. Also, humans are prone to distractions and fatigue after long periods of watching the monitors. Therefore, there is an increasing need today for automatic violence detection systems. In this work, we focus on the task of fight action detection, and aim to design a framework that can automatically detect fight behaviors from surveillance videos in an effective manner.

A naive approach [1, 2] to fight action detection is to extract different types of hand-crafted visual descriptors from local regions or from the whole frame, and proceed to generate a set of visual words by Bag-of-Words (BoW) approach for classification. Nievas et al. [1] attempted to detect fight action with two spatial-temporal descriptors, Space-Time Interest Points and Motion SIFT. In [2], a novel motion feature called Motion Co-Occurrence Feature (MCF) for accurate fight detection was proposed. However, the computational cost of extracting these features is large, making it intractable for real-time applications. To reduce computational time, some researchers focus their attention on finding more efficient feature representations for fight detection. Among them, [3] proposed a motion analysis method which evaluates the size, count, and direction of motion regions. In [4], a simple fast fight detection method extracted motion blobs from the absolute difference of consecutive frames. Despite the fast speed, its detection capability is also severely compromised. In fact, for uncontrolled outdoor environments, most of these features also encode background noise, thus degrading their recognition performance.

Fig. 1. Detection and recognition of fight behaviors in videos captured by surveillance cameras.

In recent years, video action recognition has received increasing attention while achieving very promising performances by taking advantage of the incredible robustness of Convolutional Neural Networks (CNN). Inspired by the success of deep learning approaches in object detection [5], many works [6, 7] utilize object detectors, especially person detectors, to localize regions-of-interest in each frame for action localization. Such methods typically include three submodules: person detection, tracking or linking, and action classification. However, there are several challenges in such approaches. Firstly, since fight events always involve multiple

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00104

persons, person-person interaction should also be considered during the procedure. Secondly, as prerequisite phases, existing detection and tracking algorithms still struggle in real-world, cluttered outdoor environments.

To address these problems, we propose a new localization guided fight action detection framework for realistic surveillance videos. First, to localize all potential active regions where fight actions might occur, several activation boxes are extracted from a motion activation map, which measures the activity level at each position. Then, we cluster all localized proposals around the extracted active regions based on the spatial relation between each pair of human proposals and activation boxes. In this way, we model person-person interaction as a group aggregation problem. An interesting aspect of our method is that failures in the human detector can be circumvented by the availability of regions obtained from activation boxes. In the recognition phase, we adopt the 3D Convolution Network [8] as our backbone using a two-stream framework [9] to fuse visual appearance features and temporal motion features. For the temporal stream, we opt to use the magnitude difference of optical flow maps, which produces better performance compared to primitive optical flow.

Our main contributions are summarized as follows:

1. We introduce a newly established dataset, the *FADS* dataset, for fight action detection on real-world surveillance scenarios. The *FADS* dataset provides realistic fight events captured in 1,520 video clips, and we hope to contribute towards advancing research in this area.
2. We propose a novel localization guided fight action detection framework which is robust and accurate for real-world surveillance systems.
3. We introduce a motion activation map, which is a reliable characterization of velocity information, which can help address common problems in group aggregation and action localization.

2. FIGHT ACTION DETECTION FRAMEWORK

In this section, a novel framework for fight action detection is introduced in detail. As shown in Fig. 2, the overall framework comprises of two branches. Given a video V containing m frames, the localization branch takes every T consecutive frames as input to localize multiple active regions of interest. The recognition branch predicts the action categories of each localized region using a two-stream convolution network.

2.1. Group activity recognition

Since fight event involves multiple persons, it can be considered as a group activity detection and recognition problem. Currently, two popular approaches [10, 11] towards group activity recognition are illustrated in Fig. 3. However, both approachces are not very suitable for fight action detection task.

The hierarchical LSTM model [10] (Fig. 3a) uses a person pooling module to aggregate features of all individual persons into one global feature. However, such high-level pooling strategy discards all spatial relations between the individual persons, which are essential for representing person-person interactions. Another related approach [11] (Fig. 3b) used a human detector to detect human positions and proceeds to apply clustering strategies (such as K-means) to aggregate individual humans into several groups, based on the distance, shape or other metrics. Their approach tends to produce blurred silhouettes caused by quick movements under low resolutions, which is detrimental to clustering.

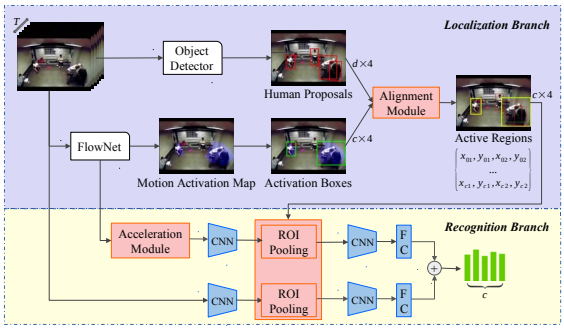

Fig. 2. An illustration of the proposed framework. The upper part is the region localization branch while the bottom part is the action recognition branch. The c active regions (marked by yellow boxes) from the localization branch are fed into a ROI Pooling layer in the two-stream action recognition network. Each region is expressed as a 4-dimensional vector, indicating the coordinates of top left point(x_1, y_1) and bottom right point(x_2, y_2). Best viewed in color.

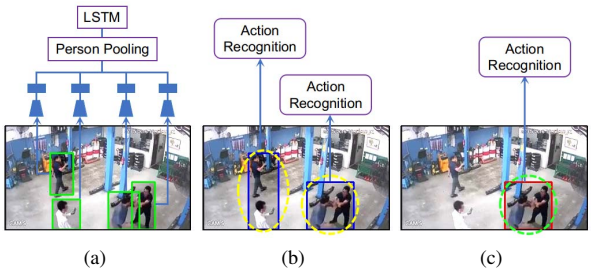

Fig. 3. Illustration of three different group activity recognition models. (a) Hierarchical LSTM model. (b) Distance based group clustering model. (c) Our motion activation map based group aggregation model.

2.2. Active region localization

Considering the limitations mentioned above, we introduce a motion guided active region localization model, which improves region detection as well as group aggregation process.

Optical flow based activation boxes. As the velocity magnitude of fight action is usually larger than of non-fight action, we can infer a fight event based on high response areas in the motion velocity map. With this idea, we utilize optical

flow to extract a motion activation map from T consecutive frames. Specifically, the motion activation map is calculated by averaging the sum of magnitude of optical flow maps. In order to reduce noises caused by illumination changes, we remove all regions with very few fragmented pixels from the map. After that, activation boxes are obtained from the motion activation map by extracting its high response areas.

Detection guided alignment module. In our framework, activation boxes serve as the centers of active action areas. Since motion activation map contains only pixel-level representation and no semantic information, the extracted activation boxes may depict only the moving parts of human bodies. To accurately adjust the active regions to include whole human bodies and to aggregate humans around these regions, a novel detection guided alignment module is designed. As is shown in Fig. 2, the alignment module takes activation boxes and human proposals as inputs and outputs several aligned regions of interest, which are then fed into the ROI pooling layers of the recognition branch. Formally, let c be the number of activation boxes from T consecutive frames and d the number of detected human proposals. An affinity factor δ_{ij} is calculated as the IoU (Intersection over Union) between every pair of B_i and P_j, where B_i is the i-th activation box and P_j is the j-th human proposals. For each activation box B_i, human proposals with a high affinity factor with it are preserved and their unions are marked as Q_i. The final active region $\overline{B_i}$ is then calculated as the union of Q_i and B_i:

$$S_i = \{j \mid \delta_{ij} > 0.6\} \tag{1}$$

$$Q_i = \bigcup_{j \in S_i} P_j \tag{2}$$

$$\overline{B_i} = B_i \cup Q_i \tag{3}$$

Therefore, for each activation center, the detection guided alignment module will generate an aligned active region. A total of c active regions are generated.

Discussions. We argue that compared to other group activity recognition methods (see Fig. 3), by introducing motion based active region localization, our proposed fight action detection framework has three advantages:

1. When the human detector fails to detect some "hard" proposals, the optical flow based activation maps are still able to obtain their regions-of-interest, avoiding potential missed targets (see Fig. 4a).
2. The activation boxes can be naturally treated as the aggregation center of action areas. Such group aggregation strategy is better than distance or shape based clustering methods since it utilizes both appearance and temporal motion information (see Fig. 4b).
3. Since human proposals with little intersection with all activation boxes are usually non-fight targets which will be eliminated during the localization phase, there is no need to infer action labels on these proposals. As a result, the overall computational cost is reduced.

(a) When the detector fails to detect small-size or occluded persons, the activation guided method is still able to identify the correct active regions.

(b) Motion activation boxes can be naturally treated as the aggregation center of group activities.

Fig. 4. Characteristics of the proposed motion based active region localization method.

Fig. 5. **Left:** Original RGB frame. **Middle:** Optical flow map. **Right:** Motion acceleration map.

2.3. Fight Action Recognition

For fight action recognition, we choose the popular 3D Convolution Networks (C3D) [8] as our backbone network, for its good performance and efficiency. Besides, two-stream frameworks [9] have also shown to yield significant improvements in video action recognition. In this work, we leveraged on both ideas, constructing a two-stream C3D network.

Instead of directly using optical flow as the temporal stream, we use an acceleration module to extract motion acceleration information for the temporal stream. Motivated by the Violence Flow (ViF) [12] representation, we propose the use of motion acceleration information, which emphasizes the drastic changes of actions. Formally, the magnitude map of t-th optical flow is denoted as $M(t)$. The motion acceleration map $\mathcal{A}(t)$ (as shown in Fig. 5) is calculated as follows:

$$\delta(t) = M(t) - M(t-1) \tag{4}$$

$$\mathcal{A}(t) = \frac{\delta(t) + 255}{2} \tag{5}$$

Eq. (5) linearly normalizes the difference in optical flow maps to [0, 255]. With T consecutive frames, the two-stream C3D network simultaneously extracts features from T RGB frames and T-2 temporal acceleration frames. An ROI pooling layer is inserted between layers `conv5a` and `conv5b` to focus attention to the active regions found. At the end of the branch, a late fusion strategy is adopted to fuse predictions from both streams. The output c final scores represent the fight action possibilities of each c-th region.

2.4. Implementation Details

In our implementation, we use the pre-trained FlowNet 2.0 network [13] to estimate optical flow and the pre-trained SSD

(a) Movies (b) Hockey (c) UCF101 (d) FADS

Fig. 6. Sample "fighting" frames from the four datasets.

network [5] (VGG-16 backbone) to detect human proposals. Weights of both the two pre-trained networks are fixed. To train a two-stream C3D network, active regions extracted from the localization branch are labeled as fight and non-fight actions. In the evaluation phase, when at least one active region is recognized as a fight, the whole video clip will be classified as a fight scene. The C3D network is optimized by SGD with an initial learning rate of 0.0002 and a weight decay of 0.0005. The learning rate decays by 0.1 every 10,000 iterations (until maximum of 50,000 iterations). Considering the duration of a fight action, we set T as 16. Besides, to better detect fight events, we use a temporal sliding window strategy with an overlap of 4 frames in our experiments.

3. EXPERIMENTS

This section is organized in accordance to the steps in our experiments. Firstly, we describe the four datasets that are evaluated in our experiments. Secondly, we compare our proposed method with several state-of-the art methods. To better verify the effectiveness of our method, we also divide our methods into localization and recognition parts and evaluate the influence of each module independently.

3.1. Datasets

We evaluate our proposed method on four datasets of different characteristics. The first two datasets (Movies and Hockey dataset) [1] were designed particularly for fight action detection. The UCF101 [14] is a large dataset of realistic action videos collected from YouTube. Among its categories, two actions ("Punch" and "SumoWrestling") are representative of fight actions and we used these samples for our evaluation.

It must be mentioned that these three datasets are not collected from real surveillance video footages. Another crowd violence dataset called Violent-Flows dataset [12] has been proposed for the purpose of detecting violent crowd behavior. However, most videos in this dataset were taken by hand with very sharp camera jitters. Thus, we find its samples not consistent with real surveillance-type videos.

In order to evaluate on real surveillance scenarios, we present the newly collected *Fight Action Detection in Surveillance-videos (FADS)* dataset. We collected surveillance based fight video clips from the UCF-Crime [15] dataset and from YouTube. Most of the original videos have a duration of several minutes but contain only several seconds with fight events. From all these videos, we temporally trim 1,520 video clips by reporting the starting and ending timestamps. All clips are captured at 30 *fps* and lasts about 3.5 seconds on average, with a resolution of 320×240 pixels. Among them, 756 clips are manually labeled as fight and 764 clips are labeled as non-fight– a fairly balanced distribution. Our FADS dataset can be considered a "crowd" fight action dataset as the number of people involved could be quite large in most cases, covering a wide range of commonly seen fighting incidences from various scenarios. Compared to the aforementioned three datasets, our FADS dataset is much more challenging. A statistical analysis and comparison between the four datasets is presented in Table 1 and we show some sample "fighting" frames from each of the four datasets in Fig. 6.

3.2. Quantitative Comparison with Prior Methods

We first compare our proposed approach against several state-of-the art methods. Table 2 reports the classification accuracy values of the competing methods for the Movies, Hockey, UCF101 and FADS datasets. For our newly created FADS dataset, the accuracy results of some existing methods [12, 16, 4, 17] are obtained by implementations following the original version in their respective works. From Table 2, we observe that our proposed method is able to outperform existing algorithms in the Hockey, UCF101 fight dataset and FADS dataset. As for the Movies dataset, our classification accuracy (99.8%) is very close to the existing best results (100%) achieved by ConvLSTM [17] and FightNet [18].

The advantages of our proposed method are most significantly reflected in the classification results on the FADS dataset. Compared to other competing approaches, our proposed method exceeded the accuracy of the next best method by $\sim 7\%$. As mentioned earlier, video clips from the Movies, Hockey and UCF101 datasets are mainly trimmed and less complex; hence, existing methods can already approach near-perfect results. But for the more complex FADS dataset, where most of the fight events occur only in a small, specific localized area in the frame, the reported detection accuracy vary greatly across the board. Existing algorithms fail to deal with such crowded environments because the extracted frame-level features contain substantial background noises while the precise action information is not well-emphasized. Our proposed method overcomes these shortcomings by encoding the motion in localized regions and performing recognition on the aggregated grouping of people.

3.3. Analysis of the proposed method

In this section, we aim to verify the effectiveness of our proposed architecture by evaluating the performance of each independent module. In real-world setting, a major portion of surveillance videos contains non-fight events. It is crucial for

Table 1. Comparison between the four datasets

Descriptions	***Movies***	***Hockey***	***UCF101***[1]	***FADS***
Fighting scenarios	action theme movies	ice hockey rink	boxing arena, sumo site	bar, yard, prison, street, hospital, supermarket, room, platform, corridor and carriages
# fight/non-fight videos	100 / 100	500 / 500	276 / 276[2]	756 / 764
Video resolution	720 x 576	360 x 288	320 x 240	320 x 240
Average video duration	1.8s	1.64s	9.6s	3.5s
Is it temporally trimmed?	✓	✓	✓	✓
Is it crowd dataset?	×	×	×	✓
Is it from a CCTV footage?	×	×	×	✓

[1]Only "Punch" and "SumoWrestling" actions from UCF101 are used for fight action evaluation.

[2]The non-fight videos are randomly chosen from videos of all remaining 99 action categories

Table 2. Mean accuracy values (%) of competing methods

Method	***Movies***	***Hockey***	***UCF101***	***FADS***
MoSIFT [1]	86.5	89.5	-	-
ViF [12]	91.3	82.9	84.7	77.2
ViF+OViF [16]	-	87.5	86.6	81.4
Deniz et al. [19]	98.9	90.1	92.4	-
Fast Fight [4]	97.8	82.4	83.5	79.5
STIFV [20]	99.5	93.7	-	-
ConvLSTM [17]	**100.0**	97.1	93.1	86.5
FightNet [18]	**100.0**	97.0	-	-
Ours	99.8	**98.6**	**96.2**	**93.3**

the false alarm rate to be as low as possible. Thus, using the challenging FADS dataset, we evaluate our proposed method based on three metrics: TPR (True Positive Rate), FPR (False Positive Rate) and AUC (Area Under Curve). The values of TPR and FPR are calculated using a threshold of 0.8. We conduct these ablation experiments by evaluating the localization and recognition branches separately.

3.3.1. Evaluation of localization branch

For the evaluation of localization branch, we design four sub-experiments to validate the feasibility of various localization techniques: 1) using the whole frame for action recognition, i.e. no localization applied; 2) using human detector and a distance based K-means clustering algorithm to aggregate groups of human; 3) using the activation boxes without alignment module; 4) using the proposed alignment guided module. For all the four sub-experiments, we use the same two-stream C3D network for the recognition step.

Table 3 and Fig. 7 gives the comparative results and ROC curves of the four localization sub-experiments on FADS dataset. As shown, it is obvious that applying the proposed localization module achieves higher TPR and lower FPR. Several details can be observed from Table 3. Firstly, recognition without localization leads to very poor results for such cluttered videos. Secondly, the performance of "detector + clustering" method is actually worse than using activation boxes without alignment module. The main reason is likely that many fast moving or small-size human targets are not clear in the FADS videos and therefore this is extremely challenging even to a pre-trained human detector. This indicates that the motion and temporal information should be utilized to help improve performance in action localization.

Table 3. Comparison of different localization method

Localization method	***TPR(%)***	***FPR(%)***	***AUC***
whole frame	69.4	9.43	0.72
detector + clustering	83.1	5.10	0.82
activation w/o alignment	89.5	2.32	0.87
activation w/ alignment	91.5	1.43	0.91

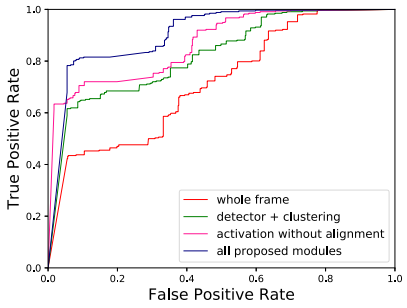

Fig. 7. ROC curves for various localization methods.

3.3.2. Evaluation of recognition branch

For the evaluation of recognition branch, we use the same localization result for ROI Pooling layer and compare the classification results of different feature streams. We first evaluate the impact of the proposed motion acceleration module in the temporal stream. Besides, since there are many newer 3D CNN architectures than the C3D, it is interesting to see how much boost in performance can the acceleration module attain on these advanced architectures. For that, we also conduct experiments using the recently introduced 3D Multi-fiber

Table 4. comparison of different temporal streams

Recognition method	***TPR(%)***	***FPR(%)***	***AUC***
C3D (flow)	84.3	3.45	0.84
C3D (flowAcc)	86.3	3.01	0.85
C3D (RGB + flow)	87.2	3.21	0.86
C3D (RGB + flowAcc)	91.5	1.43	0.91
MF-Net (flow)	90.3	2.33	0.90
MF-Net (flowAcc)	91.6	1.87	0.91
MF-Net (RGB + flow)	92.7	1.95	0.92
MF-Net (RGB + flowAcc)	94.9	0.86	0.95

Network (MF-Net) [21], which has achieved state-of-the-art performance on many action classification datasets.

The results shown in Table 4 demonstrate that in the case of both C3D and MF-Net networks, as well as for both single- and two-stream cases, the replacement of optical flow ('flow') with the accelerated optical flow maps ('flowAcc') in the temporal stream saw an improvement of 1%$\sim$4% in TPR and a reduction of 0.5%$\sim$ 1.8% in FPR. This shows that when identifying fight action scenes, the motion acceleration can be a better representation than primitive optical flow (motion).

4. CONCLUSION

This paper presents several new advances in the task of fight action detection. A novel *FADS* dataset is constructed , which is the largest in terms of size and variety, with challenging crowded conditions. To detect fight events under such scenarios, we propose a novel localization guided framework which produces well-aligned active regions that also aggregates adjacent humans into groupings. To recognize fight actions upon localizing them, we use a two-stream architecture which takes in the localized active regions fed via an additional ROI Pooling layer. We find that motion acceleration provides a better representation than conventional optical flow particularly for identifying fight actions. Experimental results show that our method outperforms state-of-the-art techniques.

5. ACKNOWLEDGEMENTS

This paper is supported in part by: Shanghai 'The Belt and Road' Young Scholar Exchange Grant (17510740100), CREST Malaysia (No. T03C1-17).

6. REFERENCES

[1] Enrique Bermejo Nievas, O.D. Suarez, G.B. García, and Rahul Sukthankar, "Violence detection in video using computer vision techniques," in *CAIP*, 2011, pp. 332–339.

[2] Ersin Esen, Mehmet Ali Arabaci, and Medeni Soysal, "Fight detection in surveillance videos," in *11th Int. Workshop on Content-Based Multimedia Indexing*, 2013, pp. 131–135.

[3] Eugene Yujun Fu, Hong Va Leong, Grace Ngai, and Stephen CF Chan, "Automatic fight detection in surveillance videos," *Int. Journal of Pervasive Computing and Comm.*, vol. 13, no. 2, pp. 130–156, 2017.

[4] Ismael Serrano Gracia, Oscar Deniz Suarez, Gloria Bueno Garcia, and Tae-Kyun Kim, "Fast fight detection," *PloS one*, vol. 10, no. 4, pp. e0120448, 2015.

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, S. Reed, C.-Y. Fu, and A. C Berg, "Ssd: Single shot multibox detector," in *ECCV*. Springer, 2016, pp. 21–37.

[6] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid, "Action tubelet detector for spatio-temporal action localization," *ICCV*, 2017.

[7] Rui Hou, C. Chen, and M. Shah, "Tube convolutional neural network (T-CNN) for action detection in videos," in *ICCV*, 2017.

[8] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[9] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014, pp. 568–576.

[10] Mostafa S Ibrahim, Srikanth Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *CVPR*, 2016, pp. 1971–1980.

[11] Ali Al-Raziqi and Joachim Denzler, "Unsupervised group activity detection by hierarchical dirichlet processes," in *ICIAP*, 2017, pp. 399–407.

[12] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *IEEE CVPRW*, 2012, pp. 1–6.

[13] Eddy Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017, vol. 2, p. 6.

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[15] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," *arXiv preprint arXiv:1801.04264*, 2018.

[16] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu, "Violence detection using oriented violent flows," *IVC*, vol. 48, pp. 37–41, 2016.

[17] Swathikiran Sudhakaran and Oswald Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *IEEE AVSS*, 2017, pp. 1–6.

[18] Peipei Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," in *Journal of Physics: Conference Series*, 2017, vol. 844, p. 012044.

[19] Oscar Deniz, Ismael Serrano, Gloria Bueno, and Tae-Kyun Kim, "Fast violence detection in video," in *Int. Conf. on Comp. Vision Theory and App. (VISAPP)*, 2014, vol. 2, pp. 478–485.

[20] Piotr Bilinski and Francois Bremond, "Human violence recognition and detection in surveillance videos," in *IEEE AVSS*, 2016, pp. 30–36.

[21] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng, "Multi-fiber networks for video recognition," *arXiv preprint arXiv:1807.11195*, 2018.