DETECTION OF REAL-WORLD FIGHTS IN SURVEILLANCE VIDEOS

Mauricio Perez, Alex C. Kot

Nanyang Technological University
School of Electrical and Electronic Engineering
50 Nanyang Ave, Singapore 639798

*Anderson Rocha**

University of Campinas
Institute of Computing
1251 Av. Albert Einstein, Brazil 13083-852

ABSTRACT

CCTVs have since long been used to enforce security, e.g. to detect fights arising from many different situations. But their effectiveness is questionable, because they rely on continuous and specialized human supervision, demanding automated solutions. Previous work are either too superficial (classification of short-clips) or unrealistic (movies, sports, fake fights). None performed detection of actual fights on long duration CCTV recordings. In this work, we tackle this problem by firstly proposing CCTV-Fights[1], a novel and challenging dataset containing 1,000 videos of real fights, with more than 8 hours of annotated CCTV footage. Then we propose a pipeline, on which we assess the impact of different feature extractors, through Two-stream CNN, 3D CNN and a local interest point descriptor, as well as different classifiers, such as end-to-end CNN, LSTM and SVM. Results confirm how challenging the problem is, and highlight the importance of explicit motion information to improve performance.

Index Terms— Video surveillance, Violence detection, Fight events, Activity localization, Deep learning

1. INTRODUCTION

A common technological device for increasing security is Closed-Circuit TeleVision (CCTV), which consists of a system for video surveillance, usually covering different locations, such as public places, schools, shopping malls, residential or commercial areas. In spite of the effort, CCTVs effectiveness is questionable [1], specially because they require sufficient trained supervisors, and human attention capability by itself is limited. Some estimates say that 99% of all surveillance footage generated is in fact never watched [2].

Among the different events that might be relevant for detection and prevention using CCTV, fight is a common event of interest, and that might arise from different situations (e.g., heated discussions, burglary, hate crimes). Consequently, it is a paramount event to have its detection automated. Nonetheless, this problem has been mainly neglected by previous literature. Most of the previous work, although using the surveillance scenario as motivation, focused their solutions on detection of general activities [3, 4, 5, 6]. Moreover, the few works that focus on fight detection do not base their solutions on realistic fights and/or surveillance footage [7, 8, 9, 10], with the great majority focusing on short-clips, instead of long, untrimmed, videos.

Fig. 1. Frames sampled from CCTV-Fights, showcasing the challenging diversity of its fight scenes and conditions.

A possible reason for this research gap might be related to the fact that there is no available dataset, which comprises all these characteristics: real-world fights from surveillance cameras. For example, existing datasets comprise video data from general activity [11], Hollywood movies [12] and faked actions [13]. For that reason, it is paramount to propose a new dataset with these features, that could enable us to design and evaluate a solution better suited for this scenario.

In this regard, our contributions in this paper are twofold: First we introduce the new dataset CCTV-Fights, consisting only of real-world fights, and containing more than 8 hours of CCTV footage temporally annotated. Then we propose an initial methodology to tackle the fight detection problem and that serves as a baseline for future research in the field. We evaluated different feature extraction methods ranging from Deep Learning to Local Interest Points, and also different classifiers – including end-to-end CNN, LSTM and SVM.

*We thank FAPESP DéjàVu grant #2017/12646-3, CAPES DeepEyes grant; and CNPq #304497/2018-5 for the financial support of this research.

[1] http://rose1.ntu.edu.sg/Datasets/cctvFights.asp

2. CCTV-FIGHTS DATASET

Existing datasets for fight detection have a diverse range of characteristics: different nature, duration of the videos, number of videos, purpose, recording source, staged or not, an so on. Table 1 contains a summary of existing datasets in prior art. Yet, none of them can properly cover the scenario of realistic fights recording coming from CCTV surveillance cameras, which is mandatory to evaluate a solution on the paramount problem of fight detection under these circumstances. To overcome this issue, a new dataset was collected, containing $1,000$ videos picturing real-world fights, recorded from CCTVs or mobile cameras: **CCTV-Fights**.

Table 1. Summary of some of the most relevant datasets commonly used for Fight Detection.

	Name	Size	Characteristics
Trimmed	Hockey Fight [7]	1,000 clips	Hockey players
	Movies [7]	200 clips	Trimmed action movies
	Violent-Flows [14]	246 clips	Crowd violence
Untrimmed	VSD [12]	25 movies	Complete Hollywood movies
	RE-DiD [15]	30 videos	Urban fights + Cars/Mobiles
	BEHAVE [13]	4 videos	Acted fights + CCTVs
	CCTV-Fights	1,000 videos	Urban fights + CCTVs/Mobiles

Approx. duration: Clips (2-5 secs) - Videos (20 secs - 5 mins) - Movies (1.5 - 2 hours)

The dataset videos were collected from YouTube, searching with keywords like: CCTV Fight, Mugging, Violence, Surveillance, Physical violence, etc. The fights can contain a diverse range of actions and attributes, for example: punching, kicking, pushing, wrestling, with two persons or more, etc. It was discarded videos that did not came directly from a CCTV recording (e.g., footage made with a mobile camera recording a screen), as well as videos with heavy special effects (e.g., shaded borders, slow-motion). Figure 1 depicts some examples from the collected videos.

This way we managed to acquire 280 CCTV videos containing different types of fights, ranging from 5 seconds to 12 minutes, with an average length of 2 minutes. By itself, this set is a bigger corpus than any of the existing datasets in prior art. Furthermore, we collected additional 720 videos of real fights from other sources (hereinafter referred to as Non-CCTV), mainly from mobile cameras, but a few from car cameras (dash-cams) and drones or helicopters. These videos are shorter, 3 seconds to 7 minutes, with an average length of 45 seconds, but still some have multiple instances of fight and can help the model to generalize better. Table 2 presents a summary of the dataset statistics.

Table 2. Summary of CCTV-Fights statistics. Value in parentheses indicates average number of fight instances per video.

	Videos	Duration (hours)	Fight Instances
All	1,000	17.68	2,414 (2.41)
CCTV	280	8.54	747 (2.67)
Non-CCTV	720	9.13	1,667 (2.32)

Videos were annotated at frame-level, i.e., each fight instance segment in the video was labeled with its exact start and end points. Although the occurrence or not of a fight considering a complete scene or a few seconds split is usually a consensus among most human viewers, the specific begin and end points of the fight is more subjective, being prone to discussion according to different points of view. The following good practices were adopted to overcome this issue:

- Annotation should contain a few seconds extra at the edges of the event;
- Short moments during the fight without strikes or hits, but with the perpetrators still in fighting instance, should still be labeled as positive;
- Long breaks should be labeled as negative, with following fights being considered different instances.

On top of the aforementioned good practices, it is expected that the evaluation metrics should not be so strict over the precise start and end of the fight predictions, introducing flexibility at some degree for the temporal localization. For the experiments, 50% of the videos are used for training, 25% for validation and 25% for testing, randomly selected.

3. METHODOLOGY

We propose the following pipeline, depicted in Figure 2, as a benchmark for this dataset, to represent our first take at solving the CCTV fight detection problem. The pipeline can be split into three specific steps, which will be further discussed bellow, as well as the specific methods utilized at each one of these steps.

Regarding the chosen methods, we decided to rely upon Two-Stream [16], 3D CNN [17] and local interest-points [10].

3.1. Feature Extraction

The first step of the pipeline, **Feature Extraction**, consists of using the RGB information from the frames to extract meaningful features for the task at hand. These features are meaningful if they are discriminative enough for a decision-making method to correctly classify that feature as coming from a

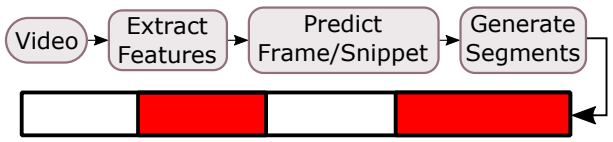

Fig. 2. Proposed Pipeline. The output is represented by the timeline bar, on which the red parts indicate the prediction of the fight segments (starting and ending time).

fight or not. Depending on how these features are generated, it can be used to describe a single frame or a small snippet of the video (e.g., a few sequential frames).

The feature extraction for the two-stream [16] based solution, is performed by using a 2D-CNN architecture for generating two different models, one for the spatial stream of the videos (RGB data of Frames) and another for the temporal stream (Stack of Optical Flows). We aggregate this information in the end by average pooling the scores or by concatenating the features from the last fully-connected layer before feeding it to a classifier.

The 3D-CNN solution [17] consists of a convolutional neural network architecture that enables convolutional on three dimensions. This way it can be explored not only for the spatial correlation within a single frame, but also for the temporal correlation in between a short sequence of frames. It is applied over a stack of sequential frames only, not using optical flow information.

For the local interest-points, we opted to base on Moreira et al.[18], which is one of the most recent papers related to Fight Localization (although only evaluated in general violence on movies). In their work, they used the local features detector and descriptor named **Temporal Robust Features (TRoF)** [18].

3.2. Frame/Snippet Prediction

The next step, **Frame/Snippet Prediction**, denotes the moment in which the classifier of choice will determine whether if the feature comes from a positive case (fight) or a negative case, according to what it has learned before from the training and split of the data. Predictions produced in this step are at the frame or snippet level, represented by a confidence score.

At this stage, each of the chosen features above were paired up with a different technique for prediction, based on their reference works. The two-stream method was applied end-to-end, with the frame/snippet prediction coming from the CNN classification layer. For the 3D CNN, the features from the last fully-convolutional layer before the classification is extracted, then fed to an LSTM for prediction. Finally, the snippet classification of TRoF features (after Fisher Vectors) is done by a linear SVM.

3.3. Segment Generation

The last step, **Segment Generation**, is responsible for aggregating predictions from the previous step to produce well-defined temporal segments for the predicted fight instances. Here, some higher-level intuition on the continuity of an event can be used to smooth the punctual predictions from the frames/snippets and achieve more realistic segments than directly using the scores independently.

Similarly to previous work [19, 17, 10], no specialized method was employed to generate the final segments. A straightforward strategy of smoothing then aggregating was used. The smoothing is a traditional mean filter applied to reduce the impact of noisy prediction scores by using the score information from the neighbors snippets. After smoothing the scores, continuous predictions that satisfy a pre-determined score threshold are merged to create the final segments.

4. EXPERIMENTS

4.1. Implementation Details

For Two-Stream approach, it was used the 2D-CNN architecture VGG16 [20] for frame and flow feature extraction and prediction, fine-tuning with pre-trained weights from ImageNet and UCF101, for the spatial and temporal streams respectively. As suggested in Simonyan and Zisserman [16], ten consecutive flows were stacked and used as input to the temporal stream. With regard to the 3D CNN method, the C3D architecture was utilized as a feature extractor, applying directly the weights learned from Sports-1M dataset [21], and the output from the layer "fc7" as the extracted feature vector. With respect to TRoF, it is used to extract low-level spatio-temporal features from the videos, then the features undergo PCA transformation for whitening and reducing their dimensionality by half. Subsequently, Fisher Vectors (FV) [22] is applied for mid-level generation of features for each snippet of the video.

The features generated through the C3D and TRoF methods were used to describe short snippets of the video, exactly 16 frames, with a stride of 8 frames between the center frame of the snippets – which leads to overlapping snippets. For the TRoF framework, that means that the low-level features pooled during the Fisher Vectors stage will come from the 16 frames comprised in the current snippet. For the C3D architecture, the 16 frames-sized snippet translates into the temporal input size of the network.

For the snippet prediction stage, the LSTM architecture and training hyperparameters were picked by grid-searching and using a validation split for measuring the performance. The SVM hyperparameter is chosen through grid search and cross-validation during the training phase.

To make a more fair comparison between different methods for feature extraction and snippet prediction, all the permutation possibilities were evaluated.

4.2. Results

For evaluating the performance of the chosen methods, the metrics used are mean Average Precision (mAP), such as in MediaEval 2014 edition [23] and the F-measure, based on time duration of predictions and ground truth. As we are dealing with localization, determining whether a predicted segment is a hit or not is not straightforward. If we considered only perfect matches, that would be too strict and could not assess properly the prediction quality. Also, it has to be taken into account that, the exact moment when a fight begins or ends can be subjective, so the evaluation metric should not be so rigid. Therefore, following the protocol from MediaEval, to deem a prediction as a hit (i.e., to contain a fight), it is necessary for it to have at least 50% of its length overlapping with a ground-truth segment. If many (small) segments satisfy this requirement for a same ground-truth segment, only one will be considered as a true positive.

Table 3. The results for the chosen methods on the CCTV-Fights dataset. First column indicates the feature used and second which classifier was used for training and prediction.

Features	Classifier	mAP	F-Measure
Two-Stream	CNN	**79.5%**	75.0%
	SVM	76.6%	72.8%
	LSTM	76.0%	**75.9%**
C3D	SVM	64.5%	58.6%
	LSTM	61.0%	58.1%
TRoF	SVM	69.2%	63.3%
	LSTM	63.8%	63.5%

Table 3 contains the results for the previously described methods, in the CCTV-Fights dataset testing split. The Two-Stream approach is significantly better than the others, regardless of the classifier used, and for both metrics. Using directly the CNN scores output leads to the highest mAP, but the slightest highest F-Measure comes from using the LSTM.

The explanation on why the two-stream approach has higher performance than the others should be related to the use of the explicit motion information. As can be seen in Table 4, when we look at the streams individual performance, the Spatial stream has a much lower performance than the Temporal stream, being equivalent to TRoF. In fact, its fusion with the Temporal information even lowered the performance of the latter by itself.

To look closer at the results, we report Table 5 with the Temporal stream performance split by type of data source: All, Non-CCTV and CCTV. We also targeted at specializing the CNN model by training it with the CCTV data in two different manners: 1-tiered, by training only with these videos; 2-tiered, by first training with all data, then fine tuning using only the CCTV source. As expected, surveillance footage is much more challenging, having a significantly lower mAP and F-measure performance than Non-CCTV. Results also show that using data from multiple sources helps generalize the model better, and that a specialization in a two-stage training leads to a better performance than training solely on the CCTV domain.

Table 4. Results for the Two-Stream approach, separately and combined, using directly the CNN output as classifier.

Features	mAP	F-Measure
Spatial	68.6%	61.0%
Temporal	80.8%	75.3%
Two-Stream	79.5%	75.0%

Table 5. Performance of the Temporal stream, with CNN as classifier, split by training model and type of data source.

Model	Source	mAP	F-Measure
All	All	80.8%	75.3%
	Non-CCTV	85.9%	79.6%
	CCTV	73.7%	66.7%
1-tiered	CCTV	72.1%	63.5%
2-tiered	CCTV	**75.6%**	**67.7%**

5. CONCLUSIONS

In this work, a new and challenging dataset was created for the scenario of real-world fights on surveillance cameras: CCTV-Fights. The dataset might prove invaluable not only because of the obtained volume of data and the important temporal-level annotations, but mainly because there was no standard dataset that completely covered this scenario in previous work. Subsequently we proposed a pipeline for fight detection and localization. Our results shown that the use of explicit motion information (e.g., Optical Flows) has a major positive impact on performance, being significantly superior than the RGB-only methods. Also it is possible to leverage the information coming from Non-CCTV fights, through a 2-tiered model that better generalizes for the CCTV source.

Possible future directions include improving the spatial features, which have not demonstrated positive complementary to the temporal information. A better use of the sequential information at the prediction stage is another interesting aspect, since the LSTM failed to leverage this information. Also, it is possible to design Early Detection methods for this scenario as well, considering the importance of quickly detecting that the fight has started.

6. REFERENCES

[1] Hina Uttam Keval, *Effective design, configuration, and use of digital CCTV*, Phd, University College London, 2009. 1

[2] "VI Dimensions," http://vidimensions.com/, Accessed: 2018-08-16. 1

[3] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem, "Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1914–1923. 1

[4] Shugao Ma, Leonid Sigal, and Stan Sclaroff, "Learning Activity Progression in LSTMs for Activity Detection and Early Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1942–1950. 1

[5] Huijuan Xu, Abir Das, and Kate Saenko, "R-C3D: Region Convolutional 3D Network for Temporal Activity Detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5783–5792. 1

[6] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson, "Encouraging LSTMs to Anticipate Actions Very Early," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 280–289. 1

[7] Enrique Nievas Bermejo, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar, "Violence detection in video using computer vision techniques," *Computer Analysis of Images and Patterns (CAIP)*, vol. 6855, pp. 332–339, 2011. 1, 2

[8] Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, and Vittorio Murino, "Angry Crowds : Detecting Violent Events in Videos," in *Springer European Conference on Computer Vision (ECCV)*, 2016, pp. 1–16. 1

[9] Vu Lam, Sang Phan, Duy Dinh Le, Duc Anh Duong, and Shin'ichi Satoh, "Evaluation of multiple features for violent scenes detection," *Springer Multimedia Tools and Applications (MTAP)*, vol. 76, no. 5, pp. 7041–7065, 2017. 1

[10] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha, "Multimodal Data Fusion for Sensitive Scene Localization," *Information Fusion*, pp. 1–39, mar 2018. 1, 2, 3

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 961–970. 1

[12] Claire Hélène Demarty, Cédric Penet, Mohammad Soleymani, and Guillaume Gravier, "VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation," *Multimedia Tools and Applications*, vol. 74, no. 17, pp. 7379–7404, 2015. 1, 2

[13] S Blunsden and R. B. Fisher, "The BEHAVE video dataset: ground truthed video for multi-person behavior classification," *Annals of the BMVA*, vol. 4, no. 4, pp. 1–11, 2010. 1, 2

[14] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1–6. 2

[15] Paolo Rota, Nicola Conci, Nicu Sebe, and James M. Rehg, "Real-life violent social interaction detection," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 3456–3460. 2

[16] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576. 2, 3

[17] Alberto Montes, Amaia Salvador, Santiago Pascual, and Xavier Giro-i Nieto, "Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks," *CoRR*, vol. 1608.08128, pp. 1–5, 2016. 2, 3

[18] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, and Anderson Rocha, "Temporal Robust Features for Violence Detection," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 1–9. 3

[19] Qi Dai, Zuxuan Wu, Yu Gang Jiang, Xiangyang Xue, and Jinhui Tangz, "Fudan-Njust at mediaeval 2014: Violent Scenes Detection using Deep Neural Networks," in *Proceedings of the MediaEval 2014 Workshop*, 2014, pp. 1–2. 3

[20] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, vol. 1409.1556, pp. 1—-10, 2014. 3

[21] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725–1732. 3

[22] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in *Springer European Conference on Computer Vision (ECCV)*, 2010, pp. 143–156. 3

[23] Mats Sjöberg, Bogdan Ionescu, Yu Gang Jiang, Vu Lam Quang, Markus Schedl, and Claire Hélène Demarty, "The MediaEval 2014 affect task: Violent scenes detection," in *Working Notes Proceedings of the MediaEval Workshop*, 2014, vol. 1263, pp. 1–2. 4