Classification and Temporal Localization of Robbery Events in CCTV Videos through Multi-Stream Deep Networks

Zakia Yahya and Muhammad Muneeb Ullah
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

***Abstract*—Robbery is an open social problem. Towards tackling this problem, we in this paper propose multi-stream deep networks for the classification as well as temporal localization of robbery events in CCTV videos. In our multi-stream architecture, each stream is comprised of a pre-trained 3D ConvNet in combination with LSTM which is followed by softmax. In particular, we investigate three streams based on three different types of input: (a) RGB data, (b) optical flows, and (c) foreground masks. Each stream is trained independently, and the final scores are averaged for predictions.**

To test the approach, we compile a robbery dataset from YouTube, which contains 124 untrimmed CCTV videos. Empirical comparison with several state-of-the-art methods demonstrate the promise of our multi-stream model in both the classification as well as temporal localization tasks.

I. Introduction

Robbery is a global problem. This open problem incurs millions in financial losses [1], [2], as well as casualties [3], annually. Although a growing number of CCTV cameras are being installed at public places (such as airports, banks, shopping malls, etc.), yet the conventional surveillance systems that rely on human operators are inefficient in detecting rare anomalous events, such as robbery, in real-time. This calls for a robust and efficient automated surveillance system, able to accurately detect any robbery attempt from the CCTV footage, and can respond effectively (e.g., raise alarm, lock the vault, call police, release tear gas, etc.) to foil the plot.

Recently, deep architectures have shown promising results in large-scale video analysis tasks [4], [5], [6], [7], [8], [9], [10], [11]. In particular, 3D Convolutional Neural Network (ConvNet) based approaches [12], [13], [14], [15], [11] have been popular owing to their inbuilt capability of modeling spatio-temporal features. Inspired by the promise of deep spatio-temporal models, we in this work, build upon the single-stream C3D+LSTM model [16], and propose a multi-stream C3D+LSTM framework. Here we explore three streams, based on three different but complementary types of input, which are (a) RGB frames, (b) optical flows, and (c) foreground masks. Each stream is comprised of a pre-trained 3D ConvNet (i.e., C3D [14]), meant to extract spatio-temporal features from the input image sequence, followed by Long Short-Term Memory (LSTM) cells [17] to cater for long-term correlations, and a softmax in the end (see Figure 1). We train each stream independently, and average the prediction scores while testing.

For the evaluation purposes, we collect a representative dataset from YouTube, comprising a total of 124 untrimmed CCTV videos of both robbery and normal activities. On our YouTube-Robbery dataset, we conduct two types of experiments: (i) robbery classification, and (ii) temporal robbery localization. In the robbery classification task, the goal is to assign a single label (robbery/no-robbery) to each test video based on the presence/absence of a robbery activity. Whereas, for the temporal localization task, the objective is to detect the presence of a robbery event in each test video, as well as to temporally localize its extent.

For an extensive benchmarking, we implement several state-of-the-art methods. The empirical results demonstrate competitive performance by our three-stream network in comparison to the state-of-the-art. To summarize, we present the following contributions:

- We propose a three-stream C3D+LSTM network which learns spatio-temporal features by exploiting RGB data, optical flows, and foreground masks.
- We collect a robbery dataset from YouTube, which contains 124 CCTV videos for both robbery and normal activities, and termed as *YouTube-Robbery* dataset.
- On the YouTube-Robbery dataset, we empirically demonstrate competitive performance by our three-stream deep network.

Rest of the paper is organized as follows. Section II outlines related work. Section III explains our multi-stream deep neural network approach. Section IV presents the newly created YouTube-Robbery dataset. Section V then reports the experimental results whereas Section VI concludes the paper with a discussion.

II. Related Work

Local hand-crafted feature representations have been successful for human action recognition in video [18], [19], [20], [21]. Among these spatio-temporal features, improved dense trajectories (IDT) [20] have demonstrated excellent performance. IDT employs dense point trajectories with three low-level descriptors: motion boundary histograms (MBH), histograms of optical flow (HOF) and histograms of oriented

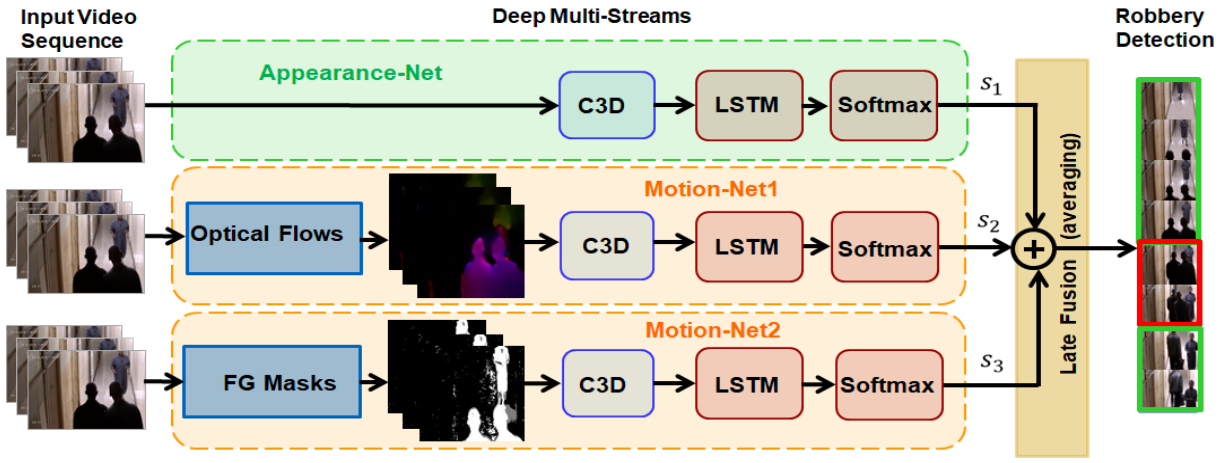

Fig. 1. Deep Multi-Stream model

gradients (HOG). After feature extraction, the local features are encoded by bag-of-features [22] or fisher vectors [23] to produce a global video representation.

Since the seminal work [24], deep learning of visual features has emerged as a promising alternative to the shallow hand-crafted features. Learned feature representations with ConvNets have been extensively investigated in several visual recognition tasks, such as image classification [25], scene recognition [26], object detection [27], and face recognition [28]. Inspired by the phenomenal performance in static images, ConvNets have also been extended to learn spatio-temporal features in videos for human action recognition [12], [4], [14], [11], [15]. In particular, the recent C3D method [14] proposes a 3D ConvNet which learns convolution filters spatially as well as temporally. C3D has shown remarkable performance in video recognition tasks [14], [29], [16].

Another approach [9], [10] is to incorporate two-stream ConvNet architecture for video action recognition. The basic idea is to decompose video into spatial and temporal components by employing RGB and optical flow frames. These components are then fed into two separate deep ConvNets, to learn spatial as well as temporal information about the appearance and motion of objects in a video scene. Each stream is trained separately, and the prediction scores are combined while testing. The two-stream approach has recently been employed in several video action recognition methods [30], [31], [32]. Recently, [33] incorporated a third stream to handle camera ego-motion for first-person activity recognition.

Recent works propose to incorporate LSTM [17] based recurrent neural networks for modeling long and short temporal correlations in video sequences. LSTM together with ConvNets have been successfully employed for classification [6] and activity localization [16] in videos. Specifically, [16] combines C3D [14] and LSTM [17] into a unified framework which learns from RGB frames, and show promising results in the ActivityNet 2016 [34] classification and localization tasks.

III. Multi-Stream Deep Neural Networks

In this section, we elaborate on our proposed framework that is illustrated in Figure 1. On one hand, our approach is inspired by the recent success of single-stream C3D+LSTM method [16], which learns to classify RGB video frames in the ActivityNet Challenge 2016 [34]. On the other hand, our method is motivated by the remarkable performance of two-stream deep models in video understanding [7], [8], [9], [10], [11]. These methods are primarily based on RGB data and optical flows. As a result, we propose to combine the aforementioned two approaches into a multi-stream C3D+LSTM framework. Particularly, we present a three-stream C3D+LSTM model (see Figure 1). Each stream in our deep model applies on a different yet complementary visual modality (RGB, optical flows, and foreground mask), and is trained independently. While testing, the probability scores from the three streams are averaged for making predictions about the robbery activities.

A. The Appearance Net

The appearance net in our multi-stream model (Figure 1) is similar to the one proposed in [16]. In this network, we employ a pre-trained 3D ConvNet called C3D [14] to extract spatio-temporal features from all videos in the dataset. We split the videos into 16-frames clips to fit the input size of the C3D model. Features from the second fully connected layer (i.e., FC6) are then extracted and L2-normalized for each video clip.

Next, we use an LSTM [17] layer, trained with dropout probability $p = 0.5$ and a fully connected layer with a softmax activation. We train the network with the negative log likelihood loss for 100 epochs, with a batch size of 10, where each sample in the minibatch is a sequence of twenty 16-frames video clips. We use RMSprop [35] with a learning rate set to 10^{-5}.

Now, given a sequence of C3D-FC6 features (for each 16-frames clip) extracted from a video, the network returns a sequence of class probabilities. This output is then post-processed to predict the robbery label of the video and its temporal boundary. First, we compute the average of the class

probabilities over all video clips in the video to obtain the robbery prediction for the whole video. We then assign the class with maximum predicted probability as the predicted class label for the video. Second, to temporally localize the robbery activity in the video, we consider all the 16-frames clips having the predicted robbery label and Intersection over Union (IoU) with the ground truth larger than 0.5.

B. The Motion Nets

The two motion nets in our model (Figure 1) have exactly the same settings as that of the appearance net above. The only difference is in the type of the visual input, which in this case is either optical flow or foreground mask rather than RGB. In the case of optical flow, we use a total variation (TV-L1) algorithm [36]. Figure 2 shows some sample frames of a video and their corresponding optical flows.

Fig. 2. Sample frames of an example video and their corresponding optical flows.

Fig. 3. Sample frames of an example video and their corresponding foreground masks.

Inspired by the complementary nature of optical flow as witnessed in the recent two-stream deep networks [7], [8], [9], [10], [11], we in this paper, additionally propose to use foreground masks. For this purpose, we use a mixture of gaussians model [37] for background subtraction. Figure 3 illustrates sample frames of a video and their corresponding foreground masks. It is pertinent to note that the computation of foreground masks is 34 times faster as compared to the optical flow. Moreover, the competitive performance by foreground mask in comparison to optical flow (shown in Section V) makes it an ideal candidate for real-time applications.

IV. The YouTube-Robbery Dataset

For testing and benchmarking our deep multi-stream neural networks, we collect a representative dataset from YouTube, and name it the YouTube-Robbery dataset[1].

A. Video Collection

The YouTube-Robbery dataset is comprised of 124 untrimmed CCTV videos for both robbery and normal activities. All the videos are captured in indoor settings (e.g., banks, shops, etc.). Table I presents some statistics. The 124 videos are divided into two splits: (i) train set and (ii) test set. The train set contains 80 videos, from which 40 videos contain robbery activities, whereas the remaining 40 videos contain normal activities. The test set contains 44 videos, out of which 22 videos contain robbery activities and remaining 22 videos contain normal activities. Figure 4 shows some sample frames of robbery and normal activities from this dataset.

[1]https://github.com/ZakiaYahya/Youtube-Robbery-Video-Dataset

TABLE I
Some statistics of the newly created YouTube-Robbery dataset.

Number of classes	2
Total videos	124
Total duration	3 hours approximately
Average video duration	79.54 seconds
Min. video Duration	8 seconds
Max. video Duration	408 seconds
Average video frames	2060

Fig. 4. Sample frames from the newly created YouTube-Robbery dataset: top three rows illustrate robbery events and bottom three rows show normal activities.

B. Video Annotation

We annotate the videos for two tasks: (i) robbery classification, and (ii) temporal robbery localization. For the classification task, we assign a single label (robbery/no-robbery) to each untrimmed video. Whereas for the temporal localization task, we segment each robbery video into 3-seconds clips, and assign a single label (robbery/no-robbery) to each clip. However, it is important to note that the ground truth temporal extent of a robbery event is until the robber(s) escape from the site. It means that as long as the robber(s) are at the site, even if no robber is visible in the video, the ground truth is labeled as a robbery event.

V. Experimental Results

In this section, we empirically evaluate our multi-stream models on the YouTube-Robbery dataset, and benchmark with several state-of-the-art methods from the literature. For C3D feature extraction and LSTM training, we use a GeForce GTX 980 GPU.

TABLE II
THE EFFECT OF DIFFERENT VISUAL INPUTS (RGB, OPTICAL FLOW, FOREGROUND MASK (FM)) IN THE ROBBERY CLASSIFICATION (@VIDEO) AND LOCALIZATION (@CLIP) TASKS.

Input Stream	C3D+SVM [14] (%)		C3D+LSTM [16] (%)	
	@Video	@Clip	@Video	@Clip
RGB	52.27	51.01	63.64	58.59
FLOW	**70.39**	**67.41**	**72.73**	77.62
FM	59.09	65.27	70.45	**78.40**

TABLE III
EFFECT OF VARYING THE HIDDEN CELL UNITS c OF LSTM LAYER IN C3D+LSTM WITH DIFFERENT INPUT STREAMS.

Hidden cell units (c)	RGB (%)		FLOW (%)		FM (%)	
	@Video	@Clip	@Video	@Clip	@Video	@Clip
64	52.27	58.07	70.45	77.13	65.90	78.25
256	56.82	58.19	68.18	77.37	**70.45**	78.33
512	**63.64**	**58.59**	**72.73**	**77.62**	**70.45**	**78.40**
1024	60.52	58.51	71.34	77.61	67.14	78.33

A. Effect of Visual Input

Since our multi-stream model is based on three different types of visual input (RGB, optical flow, foreground mask), first we are interested in assessing their individual performance in the classification and localization tasks. For this purpose, we choose two different architectures: (i) C3D+SVM [14], and (ii) C3D+LSTM [16].

In the first case (C3D+SVM [14]), we use pre-trained C3D [14] to extract L2-normalized FC6 features for each 16-frames video clip in the dataset. To classify the whole video, we take average of all the FC6 features from its 16-frames video clips, and apply a linear SVM classifier. However, to localize the robbery activity in the video, we apply a linear SVM classifier on each 16-frames FC6 feature. Robbery is detected in a 16-frames video clip if the IoU with the ground truth is larger than 0.5. Whereas, for C3D+LSTM [16], we employ the appearance and motion nets separately, as explained in Section III.

Table II presents the experimental results. Overall, we can see that optical flow and foreground mask induce the best accuracies in the classification as well as localization tasks, regardless of the architecture used. Nevertheless, it is clear that C3D+LSTM outperforms C3D+SVM, especially in the localization task (by about 13%). This is due to the fact that LSTM captures long-term temporal correlations in the video sequences, and consequently, better models the inherent dynamics of robbery activities.

B. Effect of Varying the Number of LSTM Layers and Cells

To select an appropriate configuration of LSTM in C3D+LSTM on our dataset, we have experimented with different values of layers $n = \{1, 2, 3\}$ and cells $c = \{64, 256, 512, 1024\}$, while keeping the remaining parameters fixed as stated in Section III-A. We have seen that $n = 1$ gives the best results. So with $n = 1$, Table III compares the accuracies achieved by varying the number of hidden cell units c. It turns out that overall best results are achieved with $c = 512$.

C. Comparison with the State-of-the-Art

To comprehensively benchmark our proposed three-stream C3D+LSTM model (Section III), we implement several state-of-the-art methods, and present their comparative results in this section.

From the conventional Bag-of-Features paradigm [22], we implement the popular Improved Dense Trajectory (IDT) features [20] with Fisher Vector (FV) encoding [23] in combination with linear SVM. From the deep learning side [11], we implement C3D+SVM [14] and C3D+LSTM [16], as described in Section V-A. Nonetheless, for benchmarking here, we only report their results with RGB input, as originally proposed in [14], [16] respectively. We also concatenate L2-normalized C3D FC6 features with L2-normalized IDT-FV features and combine this video representation with a linear SVM. Further, we implement the two-stream C3D+SVM framework [7], wherein pre-trained C3D [14] is used to extract spatio-temporal features from RGB frames and optical flows, and classification is done using a linear SVM.

Table IV reports all the experimental results. First, we compare the accuracy of our proposed three-stream C3D+LSTM model with the single-stream C3D+LSTM [16] model. Our three-stream C3D+LSTM model outperforms the single-stream C3D+LSTM [16] model in both the classification (by 11%) and localization (by 15%) tasks. Obviously, our model benefits from combining the three complementary channels (RGB, optical flow, foreground mask) in contrast to the single RGB channel in case of C3D+LSTM [16]. This pattern is consistent with the improved accuracy achieved by the two-stream models ([7], [8], [9], [10], [11]), which combine RGB with optical flow only.

Among the implemented methods, IDT-FV+SVM shows competitive performance in both the classification (72.73%) and localization (75.71%) tasks. Whereas, C3D+SVM performs worse in the classification (52.27%) and localization (51.01%) tasks. Moreover, combining IDT-FV with C3D doesn't improve the accuracy further. Two-Stream C3D+SVM outperforms the single-stream C3D+SVM, mainly due to the inclusion of optical flow along with RGB. It seems due to this reason, it even performs better than single-stream C3D+LSTM model.

Our proposed three-stream C3D+LSTM model achieves superior accuracy as compared to state-of-the-art in the case of robbery classification (75.00%). Whereas, it performs competitively in the robbery localization task (73.89%).

VI. CONCLUSION

We have presented a three-stream deep neural network architecture for the classification as well as temporal localization of robbery events in CCTV videos. Each stream in the proposed three-stream architecture is comprised of C3D

TABLE IV
COMPARISON WITH SEVERAL STATE-OF-THE-ART METHODS IN THE ROBBERY CLASSIFICATION (@VIDEO) AND LOCALIZATION (@CLIP) TASKS.

Architecture	@Video Accuracy (%)	@Clip Accuracy (%)
IDT-FV+SVM [20]	72.73	**75.71**
C3D+SVM [14]	52.27	51.01
IDT-FV+C3D+SVM	59.09	56.34
C3D+LSTM [16]	63.64	58.59
Two-Stream C3D+SVM [7]	72.73	64.76
Three-Stream C3D+LSTM (proposed)	**75.00**	73.89

[14] followed by an LSTM [17] with Softmax activations. In particular, we have experimented with three kinds of input streams: RGB, optical flows, and foreground masks. Empirical evaluation on the newly collected YouTube-Robbery dataset demonstrates competitive performance by our three-stream C3D+LSTM model in comparison to several state-of-the-art methods on both the classification and temporal localization tasks.

Due to the limited amount of training data in the YouTube-Robbery dataset, we could not train our multi-stream models end-to-end, and consequently, used pre-trained C3D [14] for feature extraction. In future, we intend to explore end-to-end training on larger datasets (e.g., [38], [39]). Moreover, our multi-stream framework could be explored for transfer learning problems.

REFERENCES

[1] "Bank heist of PKR 20 million; Daily Times, Pakistan," http://dailytimes.com.pk/pakistan/24-Aug-16/bank-heist-of-rs-20-million, 2016.

[2] "Biggest robbery of recent years: private security guards loot PKR 12 million in Karachi; DAWN News, Pakistan," https://www.dawn.com/news/1328175, 2017.

[3] "Security guard sacrifices life to foil robbery bid; Samaa TV, Pakistan," https://www.samaa.tv/pakistan/2017/03/security-guard-sacrifices-life-to-foil-robbery-bid/, 2017.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE CVPR*, 2014.

[5] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. PAMI*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[6] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *IEEE CVPR*, 2015.

[7] A. Diba, A. M. Pazandeh, and L. V. Gool, "Efficient two-stream motion and appearance 3d cnns for video classification," *Workshop on brave new ideas for motion representations in videos, in conjunction with ECCV*, 2016.

[8] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *CoRR*, 2015.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *CoRR*, vol. abs/1406.2199, 2014.

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," *CoRR*, vol. abs/1604.06573, 2016.

[11] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," *In Proc. IEEE CVPR*, 2017.

[12] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. PAMI*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[13] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Proc. ECCV*, 2010, pp. 140–153.

[14] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE ICCV*, 2015, pp. 4489–4497.

[15] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Tran. PAMI*, vol. 40, no. 6, pp. 1510–1517, 2018.

[16] A. Montes, A. Salvador, and X. Giró i Nieto, "Temporal activity detection in untrimmed videos with recurrent neural networks," *1st NIPS Workshop on Large Scale Computer Vision Systems*, 2016.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Journal of Neural Computing and Applications*, pp. 1735–1780, 1997.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies." in *Proc. IEEE CVPR*, 2008.

[19] H. Wang, M. M. Ullah, A. Klser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. BMVC*, 2009.

[20] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE ICCV*, 2013, pp. 3551–3558.

[21] M. M. Ullah and I. Laptev, "Actlets: A novel local representation for human action recognition in video," in *Proc. IEEE ICIP*, 2012, pp. 777–780.

[22] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *CoRR*, vol. abs/1405.4506, 2014.

[23] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. ECCV*, 2010, pp. 143–156.

[24] A. Krizhevsky, llya Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2014.

[26] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. NIPS*, 2014, pp. 487–495.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE CVPR*, 2014, pp. 580–587.

[28] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE CVPR*, 2014, pp. 1701–1708.

[29] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Deep end2end voxel2voxel prediction," *CoRR*, vol. abs/1511.06681, 2015.

[30] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN Features for Action Recognition," in *Proc. IEEE ICCV*, 2015.

[31] G. Gkioxari and J. Malik, "Finding action tubes." in *Proc. IEEE CVPR*, 2015, pp. 759–768.

[32] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. ICML*, 2015, pp. 843–852.

[33] Y. ji Kim, D.-G. Lee, and S.-W. Lee, "First-person activity recognition based on three-stream deep features," in *Proc. IEEE ICCAS*, 2018.

[34] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE CVPR*, 2015, pp. 961–970.

[35] Y. N. Dauphin, H. de Vries, J. Chung, and Y. Bengio, "Rmsprop and equilibrated adaptive learning rates for non-convex optimization," *CoRR*, vol. abs/1502.04390, 2015.

[36] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," in *Proc. 29th DAGM Conference on Pattern Recognition*, 2007, pp. 214–223.

[37] T. Bouwmans, F. E. Baf, and B. Vachon, "Background modeling using mixture of gaussians for foreground detection a survey," in *Recent Patents on Computer Science*, 2008, pp. 219–237.

[38] M. Schedl, M. Sjöberg, I. Mironica, B. Ionescu, V. L. Quang, Y. Jiang, and C. Demarty, "VSD2014: A dataset for violent scenes detection in hollywood movies and web videos," in *International Conference on Content-Based Multimedia Indexing (CBMI)*, 2015, pp. 1–6.

[39] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," *CoRR*, 2018.