

Learning to Detect Violent Videos using Convolutional Long Short-Term Memory

Swathikiran Sudhakaran^{1,2} and Oswald Lanz²

¹University of Trento, Trento, Italy

²Fondazione Bruno Kessler, Trento, Italy

{sudhakaran, lanz}@fbk.eu

Abstract

Developing a technique for the automatic analysis of surveillance videos in order to identify the presence of violence is of broad interest. In this work, we propose a deep neural network for the purpose of recognizing violent videos. A convolutional neural network is used to extract frame level features from a video. The frame level features are then aggregated using a variant of the long short term memory that uses convolutional gates. The convolutional neural network along with the convolutional long short term memory is capable of capturing localized spatio-temporal features which enables the analysis of local motion taking place in the video. We also propose to use adjacent frame differences as the input to the model thereby forcing it to encode the changes occurring in the video. The performance of the proposed feature extraction pipeline is evaluated on three standard benchmark datasets in terms of recognition accuracy. Comparison of the results obtained with the state of the art techniques revealed the promising capability of the proposed method in recognizing violent videos.

1. Introduction

Nowadays, the amount of public violence has increased dramatically. This can be a terror attack involving one or a number of persons wielding guns to a knife attack by a single person. This has resulted in the ubiquitous usage of surveillance cameras. This has helped authorities in identifying violent attacks and take the necessary steps in order to minimize the disastrous effects. But almost all the systems nowadays require manual human inspection of these videos for identifying such scenarios, which is practically infeasible and inefficient. It is in this context that the proposed study becomes relevant. Having such a practical system that can automatically monitor surveillance videos and identify

the violent behavior of humans will be of immense help and assistance to the law and order establishment. In this work, we will be considering aggressive human behavior as violence rather than the presence of blood or fire.

The development of several deep learning techniques, brought about by the availability of large datasets and computational resources, has resulted in a landmark change in the computer vision community. Several techniques with improved performance for addressing problems such as object detection, recognition, tracking, action recognition, caption generation, etc. have been developed as a result. However, despite the recent developments in deep learning, very few deep learning based techniques have been proposed to tackle the problem of violence detection from videos. Almost all the existing techniques rely on hand-crafted features for generating visual representations of videos. The most important advantage of deep learning techniques compared to the traditional hand-crafted feature based techniques is the ability of the former to achieve a high degree of generalization. Thus they are able to handle unseen data in a more effective way compared to hand-crafted features. Moreover, no prior information about the data is required in the case of a deep neural network and they can be inputted with raw pixel values without much complex pre-processing. Also, deep learning techniques are not application specific unlike the hand-crafted feature based methods since a deep neural network model can be easily applied for a different task without any significant changes to the architecture. Owing to these reasons, we choose to develop a deep neural network for performing violent video recognition.

Our contributions can be summarized as follows:

- We develop an end-to-end trainable deep neural network model for performing violent video classification
- We show that a recurrent neural network capable of encoding localized spatio-temporal changes generates a

better representation, with less number of parameters, for detecting the presence of violence in a video

- We show that a deep neural network trained on the frame difference performs better than a model trained on raw frames
- We experimentally validate the effectiveness of the proposed method using three widely used benchmarks for violent video classification

The rest of the document is organized as follows. Section 2 discusses some of the relevant techniques for performing violent video recognition followed by a detailed explanation of the proposed deep neural network model in Section 3. The details about the various experiments conducted as part of this research are given in Section 4 and the document is concluded in Section 5.

2. Related Works

Several techniques have been proposed by researchers for addressing the problem of violence detection from videos. These include methods that use the visual content [21, 2], audio content [23, 12] or both [33, 1]. In this section, we will be concentrating on methods that use the visual cues alone since it is more related to the proposed approach and moreover audio data is generally unavailable with surveillance videos. All the existing techniques can be divided into two classes depending on the underlying idea

1. **Inter-frame changes:** Frames containing violence undergo massive variations because of fast motion due to fights [28, 5, 4, 8]
2. **Local motion in videos:** The motion change patterns taking place in the video is analyzed [6, 3, 7, 21, 15, 32, 20, 24, 13, 2, 11, 34]

Vasconcelos and Lippman [28] used the tangent distance between adjacent frames for detecting the inter-frame variations. Clarin et al. improves this method in [5] by finding the regions with skin and blood and analyzing these regions for fast motion. Chen et al. [4] uses the motion vector encoded in the MPEG-1 video stream for detecting frames with high motion content and then detects the presence of blood for classifying the video as violent. Deniz et al. [8] proposes to use the acceleration estimate computed from the power spectrum of adjacent frames as an indicator of fast motion between successive frames.

Motion trajectory information and the orientation of limbs of the persons present in the scene is proposed as a measure for detecting violence by Datta et al. [6]. Several other methods follow the techniques used in action recognition, i.e., to identify spatio-temporal interest points and extract features from these points. These include Harris corner detector [3], Space-time interest points (STIP) [7], motion

scale-invariant feature transform (MoSIFT) [21, 32]. Hasner et al. [15] introduces a new feature descriptor called violent flows (ViF), which is the flow magnitude over time of the optical flow between adjacent frames, for detecting violent videos. This method is improved by Gao et al. [11] by incorporating the orientation of the violent flow features resulting in oriented violent flows (OVIF) features. Substantial derivative, a concept in fluid dynamics, is proposed by Mohammadi et al. [20] as a discriminative feature for detecting violent videos. Gracia et al. [13] proposes to use the blob features, obtained by subtracting adjacent frames, as the feature descriptor. The improved dense trajectory features commonly used in action recognition is used as a feature vector by Bilinski et al. in [2]. They also propose an improved Fisher encoding technique that can encode spatio-temporal position of features in a video. Zhang et al. [34] proposes to use a modified version of motion Weber local descriptor (MoIWLD) followed by sparse representation as the feature descriptor.

The hand-crafted feature based techniques used methods such as bag of words, histogram, improved Fisher encoding, etc. for aggregating the features across the frames. Recently various models using long short term memory (LSTM) RNNs [16] have been developed for addressing problems involving sequences such as machine translation [27], speech recognition [14], caption generation [31, 29] and video action recognition [9, 26]. The LSTM was introduced in 1997 to combat the effect of vanishing gradient problem which was plaguing the deep learning community. The LSTM incorporates a memory unit which contains information about the inputs the LSTM unit has seen and is regulated using a number of fully-connected gates. The same idea of using LSTM for feature aggregation is proposed by Dong et al. in [10] for violence detection. The method consisted of extracting features using a convolutional neural network from raw pixels, optical flow images and acceleration flow maps followed by LSTM based encoding and a late fusion.

Recently, Xingjian et al. [30] replaced the fully-connected gate layers of the LSTM with convolutional layers and used this improved model for predicting precipitation nowcasting from radar images with improved performance. This newer model of the LSTM is named as convolutional LSTM (convLSTM). Later, it has been used for predicting optical flow images from videos [22] and for anomaly detection in videos [18]. By replacing the fully-connected layers in the LSTM with convolutional layers, the convLSTM model is capable of encoding spatio-temporal information in its memory cell.

3. Proposed method

The goal of the proposed study was to develop an end-to-end trainable deep neural network model for classifying

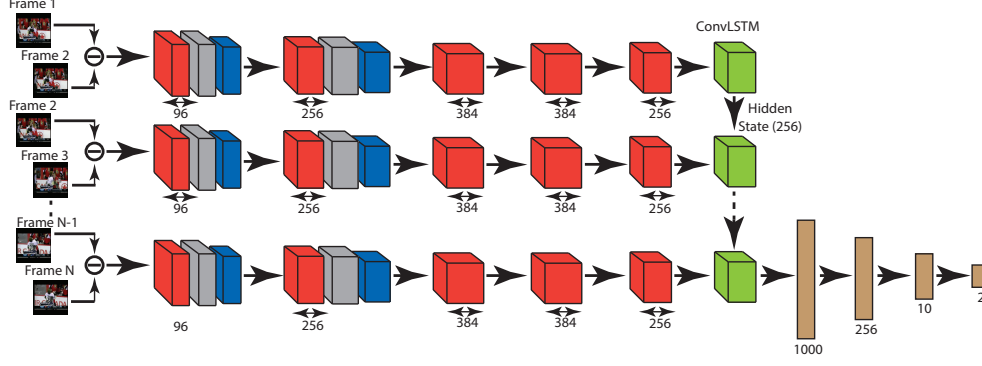


Figure 1. Block diagram of the proposed model. The model consists of alternating convolutional (red), normalization (grey) and pooling (blue) layers. The hidden state of the ConvLSTM (green) at the final time step is used for classification. The fully-connected layers are shown in brown colour.

videos in to violent and non-violent ones. The block diagram of the proposed model is illustrated in figure 1. The network consists of a series of convolutional layers followed by max pooling operations for extracting discriminant features and convolutional long short memory (convLSTM) for encoding the frame level changes, that characterizes violent scenes, existing in the video.

3.1. ConvLSTM

Videos are sequences of images. For a system to identify if a fight is taking place between the humans present in the video, it should be capable of identifying the locations of the humans and understand how the motion of the said humans are changing with time. Convolutional neural networks (CNN) are capable of generating a good representation of each video frame. For encoding the temporal changes a recurrent neural network (RNN) is required. Since we are interested in changes in both the spatial and temporal dimensions, convLSTM will be a suitable option. Compared to LSTM, the convLSTM will be able to encode the spatial and temporal changes using the convolutional gates present in them. This will result in generating a better representation of the video under analysis. The equations of the convLSTM model are given in equations 1-6.

$$i_t = \sigma(w_x^i * I_t + w_h^i * h_{t-1} + b^i) \quad (1)$$

$$f_t = \sigma(w_x^f * I_t + w_h^f * h_{t-1} + b^f) \quad (2)$$

$$\tilde{c}_t = \tanh(w_x^c * I_t + w_h^c * h_{t-1} + b^c) \quad (3)$$

$$c_t = \tilde{c}_t \odot i_t + c_{t-1} \odot f_t \quad (4)$$

$$o_t = \sigma(w_x^o * I_t + w_h^o * h_{t-1} + b^o) \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

In the above equations, ‘*’ represents convolution operation and ‘ \odot ’ represents the Hadamard product. The hidden state h_t , the memory cell c_t and the gate activations i_t , f_t and o_t are all 3D tensors in the case of convLSTM.

For a system to identify a video as violent or non-violent, it should be capable of encoding localized spatial features and the manner in which they change with time. Hand-crafted features are capable of achieving this with the downside of having increased computational complexity. CNNs are capable of generating discriminant spatial features but existing methods use the features extracted from the fully-connected layers for temporal encoding using LSTM. The output of the fully-connected layers represents a global descriptor of the whole image. Thus the existing methods fail to encode the localized spatial changes. As a result, they resort to methods involving addition of more streams of data such as optical flow images [10] which results in increased computational complexity. It is in this context that the use of convLSTM becomes relevant as it is capable of encoding the convolutional features of the CNN. Also, the convolutional gates present in the convLSTM is trained to encode the temporal changes of local regions. In this way, the whole network is capable of encoding localized spatio-temporal features.

3.2. Network Architecture

Figure 1 illustrates the architecture of the network used for identifying violent videos. The convolutional layers are trained to extract hierarchical features from the video frames and are then aggregated using the convLSTM layer. The network functions as follows: The frames of the video under consideration are applied sequentially to the model. Once all the frames are applied, the hidden state of the convLSTM layer in this final time step contains the representation of the input video frames applied. This video representation, in the hidden state of the convLSTM, is then applied to a series of fully-connected layers for classification.

In the proposed model, we used the AlexNet model [17] pre-trained on the ImageNet database as the CNN model for extracting frame level features. Several studies have found

Table 1. Classification accuracy obtained with the hockey fight dataset for different models

Input	Classification Accuracy
Video Frames (random initialization)	94.1 \pm 2.9%
Video Frames (ImageNet pre-trained)	96 \pm 0.35%
Difference of Video Frames (random initialization)	95.5 \pm 0.5%
Difference of Video Frames (ImageNet pre-trained)	97.1 \pm 0.55%

out that networks trained on the ImageNet database is capable of having better generalization and results in improved performance for tasks such as action recognition [25] [19]. In the convLSTM, we used 256 filters in all the gates with a filter size of 3×3 and stride 1. Thus the hidden state of the convLSTM consists of 256 feature maps. A batch normalization layer is added before the first fully-connected layer. Rectified linear unit (ReLU) non-linear activation is applied after each of the convolutional and fully-connected layers.

In the network, instead of applying the input frames as such, the difference between adjacent frames are given as input. In this way, the network is forced to model the changes taking place in adjacent frames rather than the frames itself. This is inspired by the technique proposed by Simonyan and Zisserman in [25] to use optical flow images as input to a neural network for action recognition. The difference image can be considered as a crude and approximate version of optical flow images. So in the proposed method, the difference between adjacent video frames are applied as input to the network. As a result, the computational complexity involved in the optical flow image generation is avoided. The network is trained to minimize the binary cross entropy loss.

4. Experiments and Results

To evaluate the effectiveness of the proposed approach in classifying violent videos, three benchmark datasets are used and the classification accuracy is reported.

4.1. Experimental Settings

The network is implemented using the Torch library. From each video, N number of frames equally spaced in time are extracted and resized to a dimension of 256×256 for training. This is to avoid the redundant computations involved in processing all the frames, since adjacent frames contain overlapping information. The number of frames selected is based on the average duration of the videos present in each dataset. The network is trained using RMSprop algorithm with a learning rate of 10^{-4} and a batch size of 16. The model weights are initialized using Xavier algo-

rithm. Since the number of videos present in the datasets are limited, data augmentation techniques such as random cropping and horizontal flipping are used during training stage. During each training iteration, a portion of the frame of size 224×224 is cropped, from the four corners or from the center, and is randomly flipped before applying to the network. Note that the same augmentation technique is followed for all the frames present in a video. The network is run for 7500 iterations during the training stage. In the evaluation stage, the video frames are resized to 224×224 and are applied to the network for classifying them as violent or non-violent. All the training video frames in a dataset are normalized to make their mean zero and variance unity.

4.2. Datasets

The performance of the proposed method is evaluated on three standard public datasets namely, Hockey Fight Dataset [21], Movies Dataset [21] and Violent-Flows Crowd Violence Dataset [15]. They contain videos captured using mobile phones, CCTV cameras and high resolution video cameras.

Hockey Fight Dataset: Hockey fight dataset is created by collecting videos of ice hockey matches and contains 500 fighting and non-fighting videos. Almost all the videos in the dataset have a similar background and subjects (humans). 20 frames from each video are used as inputs to the network.

Movies Dataset: This dataset consists of fight sequences collected from movies. The non-fight sequences are collected from other publicly available action recognition datasets. The dataset is made up of 100 fight and 100 non-fight videos. As opposed to the hockey fight dataset, the videos of the movies dataset is substantially different in its content. 10 frames from each video are used as inputs to the network.

Violent-Flows Dataset: This is a crowd violence dataset as the number of people taking part in the violent events are very large. Most of the videos present in this dataset are collected from violent events taking place during football matches. There are 246 videos in this dataset. 20 frames from each video are used as inputs to the network.

4.3. Results and Discussions

Performance evaluation is done using 5-folds cross validation scheme, which is the technique followed in existing literature. The model architecture selection was done by evaluating the performance of the different models on the hockey fight dataset. The classification accuracies obtained for the two cases, video frames as input and difference of frames as input, is listed in table 1. From the table, it can also be seen that using a network that is pre-trained on the ImageNet dataset (we used BVLC AlexNet from Caffe model zoo) results in better performance compared to us-

Table 2. Comparison of classification results

Method	Hockey Dataset	Movies Dataset	Violent-Flows Dataset
MoSIFT+HIK[21]	90.9%	89.5%	-
ViF[15]	82.9±0.14%	-	81.3±0.21%
MoSIFT+KDE+Sparse Coding[32]	94.3±1.68%	-	89.05±3.26%
Deniz et al.[8]	90.1±0%	98.0±0.22%	-
Gracia et al.[13]	82.4±0.4%	97.8±0.4%	-
Substantial Derivative[20]	-	96.89±0.21%	85.43±0.21%
Bilinski et al.[2]	93.4	99	96.4
MoIWL[34]	96.8±1.04%	-	93.19±0.12%
ViF+OVIF[11]	87.5±1.7%	-	88±2.45%
Three streams + LSTM[10]	93.9	-	-
Proposed	97.1±0.55%	100±0%	94.57±2.34%

ing a network that is randomly initialized. In this way, we decided to use frame difference as the input and to use a pre-trained network in the model. Table 2 gives the classification accuracy values obtained for the various datasets considered in this study and is compared against 10 state of the art techniques. From the table, it can be seen that the proposed method is able to better the results of the existing techniques in the case of hockey fights dataset and movies dataset.

As mentioned earlier, this study considers aggressive behavior as violent. The biggest problem of considering this definition occurs in the case of sports. For instance, in the hockey dataset, the fight videos consists of players colliding against each other and hitting one another. So one easy way to detect violent scenes is to check if one player moves closer to another. But the non-violent videos also consist of players hugging each other or doing high fives as part of a celebration. It is highly likely that these videos could be mistaken as violent. But the proposed method is able to avoid this which suggests that it is capable of encoding motion of localized regions (motion of limbs, reaction of involved persons, etc.). However, in the case of violent-flows dataset, the proposed method is not able to best the previous state of the art technique (it came second in terms of accuracy). Analyzing the dataset, it is found that in most of the violent videos, only a small part of the crowd is found to be involved in aggressive behavior while a large part remained as spectators. This forces the network to mark such videos as non-violent since majority of the people present in it is found to behave normally. Further studies are required for devising techniques to alleviate this problem involved with crowd videos. One technique that can be considered is to divide the frame in to sub-regions and predict the output of the regions separately and mark the video as violent if any of the regions is outputted by the network as violent.

In order to compare the advantage of convLSTM over traditional LSTM, a different model that consists of LSTM is trained and tested on the hockey fights dataset. The new

Table 3. Comparison between convLSTM and LSTM models in terms of classification accuracy obtained in the hockey fights dataset and number of parameters

Model	Accuracy	No. of Parameters
convLSTM	97.1±0.55%	9.6M(9619544)
LSTM	94.6±1.19%	77.5M(77520072)

model consists of the AlexNet architecture followed by an LSTM RNN layer. The output of the last fully-connected layer (fc7) of AlexNet is applied as input to an LSTM with 1000 units. The rest of the architecture is similar to the one that uses convLSTM. The results obtained with this model and the number of trainable parameters associated with it are compared against the proposed model in table 3. The table clearly shows the advantages of using convLSTM over LSTM and the capability of convLSTM in generating useful video representation. It is also worth mentioning that the number of parameters that are required to be optimized, in the case of convLSTM, is very much less compared to LSTM (9.6M vs 77.5M). This helps the network to generalize better without overfitting in the case of limited data. The proposed model is capable of processing 31 frames per second on an NVIDIA K40 GPU.

5. Conclusions

This work presents a novel end-to-end trainable deep neural network model for addressing the problem of violence detection in videos. The proposed model consists of a convolutional neural network (CNN) for frame level feature extraction followed by feature aggregation in the temporal domain using convolutional long short term memory (convLSTM). The proposed method is evaluated on three different datasets and resulted in improved performance compared to the state of the art methods. It is also shown that a network trained to model changes in frames (frame difference) performs better than a network trained using frames as inputs. A comparative study between the tradi-

tional fully-connected LSTM and convLSTM is also done and the results show that the convLSTM model is capable of generating a better video representation compared to LSTM with less number of parameters, thereby avoiding overfitting.

References

- [1] E. Acar, F. Hopfgartner, and S. Albayrak. Breaking down violence detection: Combining divide-et-impera and coarse-to-fine strategies. *Neurocomputing*, 208:225–237, 2016. [2](#)
- [2] P. Bilinski and F. Bremond. Human violence recognition and detection in surveillance videos. In *AVSS*, 2016. [2](#), [5](#)
- [3] D. Chen, H. Wactlar, M.-y. Chen, C. Gao, A. Bharucha, and A. Hauptmann. Recognition of aggressive human behavior using binary local motion descriptors. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2008. [2](#)
- [4] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su. Violence detection in movies. In *International Conference on Computer Graphics, Imaging and Visualization (CGIV)*, 2011. [2](#)
- [5] C. Clarin, J. Dionisio, and M. Echavez. Dove: Detection of movie violence using motion intensity analysis on skin and blood. Technical report, University of the Philippines, 01 2005. [2](#)
- [6] A. Datta, M. Shah, and N. D. V. Lobo. Person-on-person violence detection in video data. In *ICPR*, 2002. [2](#)
- [7] F. D. De Souza, G. C. Chavez, E. A. do Valle Jr, and A. d. A. Araújo. Violence detection in video using spatio-temporal features. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2010. [2](#)
- [8] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim. Fast violence detection in video. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014. [2](#), [5](#)
- [9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. [2](#)
- [10] Z. Dong, J. Qin, and Y. Wang. Multi-stream deep networks for person to person violence detection in videos. In *Chinese Conference on Pattern Recognition*, 2016. [2](#), [3](#), [5](#)
- [11] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu. Violence detection using oriented violent flows. *Image and Vision Computing*, 48:37–41, 2016. [2](#), [5](#)
- [12] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis. A multi-class audio classification method with respect to violent content in movies using bayesian networks. In *IEEE Workshop on Multimedia Signal Processing (MMSP)*, 2007. [2](#)
- [13] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T.-K. Kim. Fast fight detection. *PloS one*, 10(4):e0120448, 2015. [2](#), [5](#)
- [14] A. Graves, N. Jaitly, and A.-r. Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013. [2](#)
- [15] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *CVPR Workshops*, June 2012. [2](#), [4](#), [5](#)
- [16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. [2](#)
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. [3](#)
- [18] J. R. Medel and A. Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*, 2016. [2](#)
- [19] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, 2016. [4](#)
- [20] S. Mohammadi, H. Kiani, A. Perina, and V. Murino. Violence detection in crowded scenes using substantial derivative. In *AVSS*, 2015. [2](#), [5](#)
- [21] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar. Violence detection in video using computer vision techniques. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 2011. [2](#), [4](#), [5](#)
- [22] V. Pătrăucean, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. In *ICLR Workshop*, 2016. [2](#)
- [23] S. Pfeiffer, S. Fischer, and W. Effelsberg. Automatic audio content analysis. In *ACM International Conference on Multimedia*, 1997. [2](#)
- [24] P. Rota, N. Conci, N. Sebe, and J. M. Rehg. Real-life violent social interaction detection. In *ICIP*, 2015. [2](#)
- [25] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. [4](#)
- [26] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. [2](#)
- [27] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014. [2](#)
- [28] N. Vasconcelos and A. Lippman. Towards semantically meaningful feature spaces for the characterization of video content. In *ICIP*, 1997. [2](#)
- [29] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *ICCV*, 2015. [2](#)
- [30] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. [2](#)
- [31] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. [2](#)
- [32] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao. Violent video detection based on mosift feature and sparse coding. In *ICASSP*, 2014. [2](#), [5](#)
- [33] W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila. Cassandra: audio-video sensor fusion for aggression detection. In *AVSS*, 2007. [2](#)
- [34] T. Zhang, W. Jia, X. He, and J. Yang. Discriminative dictionary learning with motion weber local descriptor for violence detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):696–709, 2017. [2](#), [5](#)