Vision-based Fight Detection from Surveillance Cameras

Şeymanur Aktı
Department of Computer Engineering
Istanbul Technical University
Istanbul, Turkey
akti15@itu.edu.tr

Gözde Ayşe Tataroğlu
Department of Computer Engineering
Istanbul Technical University
Idea Technology Solutions R&D Center
Istanbul, Turkey
tataroglu15@itu.edu.tr

Hazım Kemal Ekenel
Department of Computer Engineering
Istanbul Technical University
Istanbul, Turkey
ekenel@itu.edu.tr

***Abstract*—Vision-based action recognition is one of the most challenging research topics of computer vision and pattern recognition. A specific application of it, namely, detecting fights from surveillance cameras in public areas, prisons, etc., is desired to quickly get under control these violent incidents. This paper addresses this research problem and explores LSTM-based approaches to solve it. Moreover, the attention layer is also utilized. Besides, a new dataset is collected, which consists of fight scenes from surveillance camera videos available at YouTube. This dataset is made publicly available[1]. From the extensive experiments conducted on Hockey Fight, Peliculas, and the newly collected fight datasets, it is observed that the proposed approach, which integrates Xception model, Bi-LSTM, and attention, improves the state-of-the-art accuracy for fight scene classification.**

***Index Terms*—Deep learning, action recognition, fight detection**

I. INTRODUCTION

Violence detection has been receiving increasing attention as a research topic, since it has many practical use cases. Since, unfortunately, the violent scenes in movies or media have become common, and since young generation can have access to these media content easily, a group of research activities is on automatic detection of violent activities in media contents. Another main use case is to detect violent activities in public areas, such as underground, streets, buses, hospitals, welfare institutions, etc. in order to automatically warn the public officers and enable quick action against them. Violent activities contain a broad range of activities, for example, vandalism, explosion, and fighting. In this study, we focus on the fight activity. A fight event is defined as two or more people, who are fighting to a degree that must be interfered.

Related approaches consist of two parts as feature extraction and classification. Mainly two different approaches are applied for feature extraction: computing optical flow information of the videos and computing deep convolutional neural networks-based representations. Due to the proven success of convolutional neural networks (CNN) in various computer vision applications, CNN based approaches are highly preferred in recent works. Long Short-term Memories (LSTM) are used for modeling the temporal information, as they find out relationships between the consecutive frames through their memory ability. In summary, CNN + LSTM network is commonly used in action recognition due its high performance.

In this study, in order to enhance the CNN + LSTM based approach for the fight detection task, a modified Xception CNN is trained using the fight scenes. Thus, it is expected that this CNN is more familiar with the input sequences and extracts more relevant features from them. In the classification layer, a novel approach is developed by using Bidirectional LSTM (Bi-LSTM) along with a self-attention layer to improve the performance. Furthermore, a new surveillance camera fight dataset is collected.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related work. In section 3, technical details of the proposed method are explained. Section 4 presents and discusses the experimental results. The obtained results are summarized in section 5 and finally, the paper is concluded in section 6.

II. RELATED WORK

One of the most common deep learning solutions for action recognition is two-stream convolutional networks [1]. In this method two CNNs are used, one for spatial feature extraction, which learns the actions from single images and the other one is for the temporal feature extraction, which learns from the optical flow vectors of multiple frames. Then, outputs of the two networks are combined at the end. Sudhakaran and Lanz preferred to use convolutional LSTM for classification in order to discriminate the spatio-temporal changes between frames in a better way [2].

Xu et al. use attention in image captioning by focusing on the objects that can give important information about what is happening in the scene [3]. Sharma et al. use attention in action recognition for processing the features, which have the largest effect on the output [4]. In this work, GoogLeNet [5] is used for feature extraction and multi-layered deep LSTM with attention mechanism is used for classification. According

[1]https://github.com/sayibet/fight-detection-surv-dataset

978-1-7281-3975-3/19/$31.00 ©2019 IEEE

to experimental results, the attention layer enhances the performance of the LSTM. Song et al. apply LSTM to the skeleton data, where the subjects in video sequences are represented as skeletons to recognize the human actions. Furthermore, they benefit from the attention layer in order to focus on the most active joints of sample skeletons in terms of spatio-temporal changes between frames [6].

Liu et al. introduced a new type of LSTM, which is named as Global Context-Aware Attention LSTM [7]. This new method is developed to perform 3D action recognition on skeleton data and it aims to choose the most informative joints of the samples by using an iterative attention method. Additionally, it evaluates the global context while learning from the frames, differently from the regular 2D LSTM. Dong et al. detected the violent actions between people by using multi-stream CNNs [8]. Firstly, CNNs extract spatio-temporal features, then they add one more stream for learning the acceleration of the videos. Thus, the sequences can be classified considering the activity of the scene [8]. Singh et al. extracted different kinds of features from video sequences through a multi-stream CNN [9]. After detecting the person in the frame, they construct a bounding box on the tracked person and use several streams for taking motion features from both inside of bounding box and general frame. Then the features are fed into a bi-directional LSTM for classifying the actions. Ullah et al. used various CNN architectures to extract features from the frames of video sequences [10]. Features are taken from the second to the last layer of network and classified by a bi-directional LSTM. 3D convolutional neural networks are also utilized for action recognition in video sequences [11]–[14]. Peixoto et al. used 3D CNN and CNN-LSTM for violence detection in videos. Then, they combined the outputs of these two networks with another network which can distinguish the different concepts of the violence [15].

In the literature, there are several publicly available violence detection datasets. For example, Technicolor presents their Hollywood movie dataset that contains violent and non-violent sequences from 31 movies [16]. Peliculas dataset contains various fight and non-fight videos from YouTube or the movies [17]. Hockey dataset includes fight and non-fight videos from ice hockey games [17]. Another dataset is Violent Flows Dataset and it contains multiple violence scenes [18]. UCF-Crimes dataset includes different crime scenarios such as robbery, argon, burglary etc. along with fighting [19]. A recent dataset released in 2019 [20], contains surveillance camera videos with fight instances. To complement these datasets, in this study, a fight dataset is constructed by using the surveillance camera footages from YouTube.

III. PROPOSED METHOD

In the following subsections, feature extraction and classification parts of the proposed method are presented.

A. Feature Extraction Model

Various types of CNN architectures are tested for feature extraction part, such as VGG16 [21] and Xception [22].

VGG16 takes 224 x 224 pixel resolution images as input. It has three fully connected layers at the end. The features are taken from the second fully connected layer. On the other hand, Xception takes 299 x 299 pixel resolution input. The features are extracted from the last global average pooling layer.

Furthermore, one additional CNN is trained for fight detection, which is named as Fight-CNN. Fight and non-fight frames of the video sequences in Hockey dataset are used for training. The trained CNN has the Xception architecture but the last layer is mapped into two classes. Also the kernel size is widened in order to catch more relative features from the fight scenes. The new network with Xception is smaller than the regular model with 11 million parameters. It has two fully connected layers before classification layer and features are extracted from the first fully connected layer.

Before sending the videos for the feature extraction, frames are sampled from video sequences. Uniform sampling is used and 5 or 10 frames from each video are selected. Then, using cubic interpolation these frames are resized to the input size of the network architecture.

B. Classification Model

In the classification part, Bi-LSTM is used, since it can learn the dependency between past and current information. Then, an attention layer is included to determine the significant parts of the input.

1) LSTM: Long Short-term Memory is a method that is used in sequence learning tasks [23]. The memory usage capability of LSTM differs from the regular recurrent neural networks (RNN). Its memory gates in the modules make it possible to keep the necessary information and ignore irrelevant information. The gates choose to pass or throw some parts of the data according to its relevancy by considering the previous data. In other words, the gates in LSTM learn how much the new information depends on the previous information. Therefore, the relationship between the elements of a sequence can be learned. In this case, the data consists of sequence of images and the network can connect the information in frames which are taken at different times from the videos. During this process, the system remembers the previous frame while examining the current frame. The system learns the temporal changes occurring during the video processing and those changes give significant information to recognize the actions.

During the LSTM experiments, an LSTM model with one LSTM layer, three dense (1024, 50, 2) and three activation layers (relu, sigmoid, softmax) are used. At the end of the architecture, softmax layer is used with two classes instead of binary classification by sigmoid. Therefore, the prediction confidences in the output can be observed. So that, mean squared error is used as the loss function which gives better results than the cross entropy loss function.

2) Bi-LSTM: Different from the regular LSTM which has only forward flow in the sequence where the inputs are determined according to the previous information, Bi-LSTM

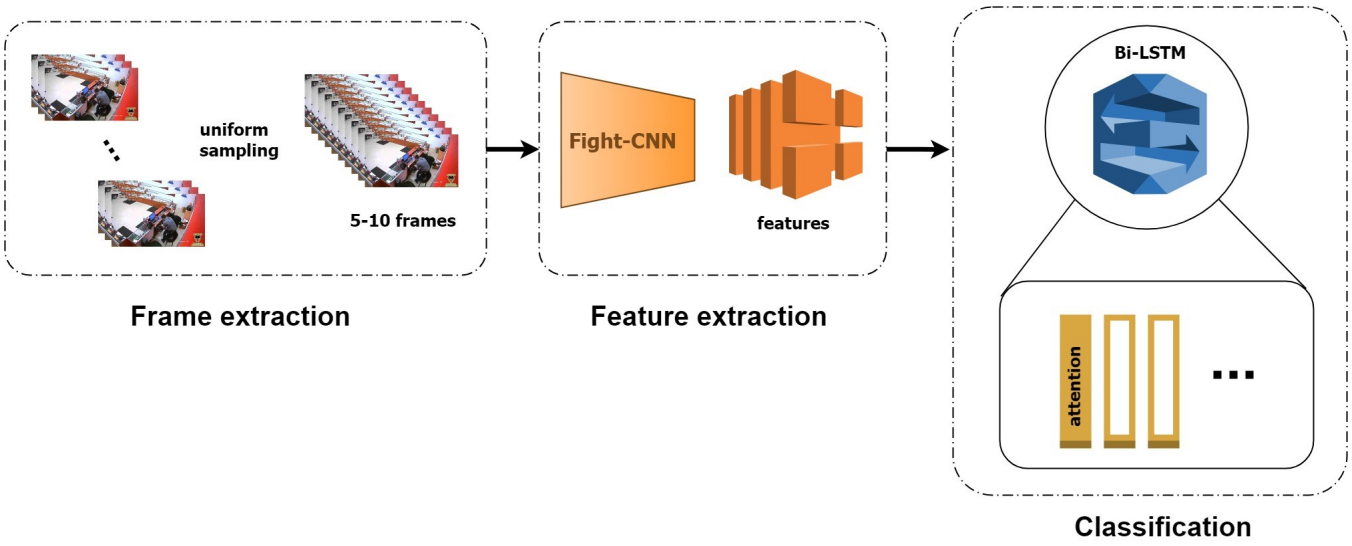

Fig. 1. Overview diagram of the proposed system.

has an additional backward flow [24]. After completing the forward learning, a backward learning is processed starting from the last element to the first element. Therefore, in each cell, both the past and future information is kept and outputs are determined by taking into account this information.

While performing the experiments with Bi-LSTM, the same architecture with regular LSTM is used with an additional Bi-LSTM layer instead of LSTM layer. Besides, dropout is applied in order to reduce overfitting.

3) Attention layer: Attention mechanism is first introduced by Bahdanau et al. in 2014 [25] and generally used in natural language processing in RNNs for deciding how much attention must be given to other words while processing the current word. It is also used in visual problems like image captioning [26]–[28] and object detection [29].

When attention layer is used together with bi-directional LSTMs, it computes weights for each cell to interpret each element in the sequence. The backward and forward layer values of each element is calculated and affect the other elements outputs. Attention layer determines how much each output should be affected by other inputs. After observing both past and future information, it generates a weight matrix and this matrix is used to calculate the outputs.

Self-attention [30] is another type of the attention mechanism, which is used in this study. The authors apply the attention to the input data and try to represent it in a more convenient form by focusing on significant parts of the data while processing the elements in sequence. For instance, the input data in this study is feature vectors from ten frames. The attention layer performs on the input and generates new feature vectors considering the attention matrix and relationships between input vectors. After that, the new feature vectors are sent into the next layers for classification. The overview of the proposed system can be seen in Fig. 1.

TABLE I
NUMBER OF SAMPLES FOR EACH DATASET

Datasets	# fight	# non-fight	# total
Hockey Dataset	500	500	1000
Peliculas Dataset	100	100	200
Collected Surveillance Camera Dataset	150	150	300

IV. EXPERIMENTAL RESULTS

In the following subsections, we first explain the used datasets and the experimental setups. Then, we present and discuss the experimental results.

A. Datasets

1) Hockey Fight Dataset: The dataset contains fight and non-fight scenes from ice hockey games. There are 1000 video samples in total, where 500 of them are fight sequences and other 500 of them are non-fight sequences. Videos are two seconds long and frame sizes are constant. Background of the videos are all similar and they contain background motion.

2) Peliculas Dataset: It includes fight sequences from Hollywood movies, some non-fight scenes from football games, and other events. There are 200 videos in total. 100 of them are fight videos and 100 of them are non-fight videos. Duration of videos are two seconds and size of the frames can differ. Environment and people in the videos are varying, since they are from the movie scenes. These videos also have background motion.

3) Surveillance Camera Fight Dataset: This dataset is collected for this study. Even though there are some fight or violence specific datasets, the main samples in these datasets are taken from movies or hockey games, which correspond to different type of scenes. These datasets can help to learn actions itself, but they are not exactly suitable for the purposed task. The actors in the hockey game scene records look identical and the background itself does not change much.

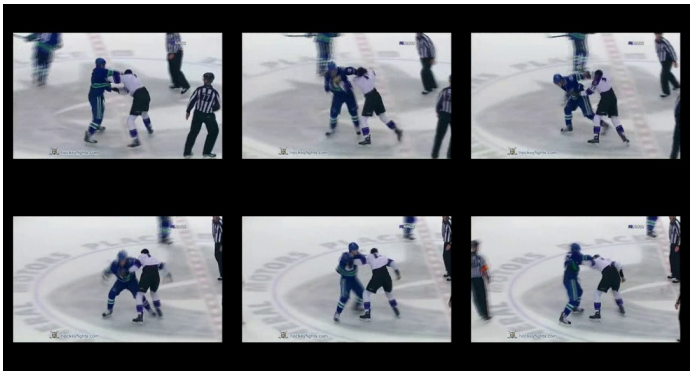

Fig. 2. An example fight scene from Hockey dataset.

Fig. 3. An example fight scene from Peliculas dataset.

Fig. 4. An example fight scene from surveillance camera dataset.

However, in surveillance applications, humans in the scenes always differ and the background of the footage differs for each camera. In movies and hockey games, the background is moving due to filming techniques like zoom in / out. On the other hand, surveillance cameras are mostly still and the background in recordings is more stable. The differences can be observed from Fig. 2, 3, 4. Thus, a new dataset containing the fight / non-fight sequences from surveillance camera footage would complement the existing datasets.

In surveillance camera dataset, there are 300 videos in total, 150 of them are fight sequences and 150 of them are non-fight sequences. The surveillance camera footages are collected from YouTube mostly and some surveillance camera datasets like CamNet [31] and Synopsis dataset [32], [33] are used for extracting non-fight video cuts. After collecting videos, 2-second-long fight / non-fight sequences are cut from them. The videos have different sizes and different number of frames. Therefore, the frames are resized before they are sent to the CNNs. Then uniform sampling is applied by taking into account the total frame number of the videos as seen in Fig. 1. Table 1 summarizes the number of samples in the used datasets.

Fig. 5. Various fight scenarios from the collected dataset.

There are various types of fight scenarios in the dataset such as kick, fist, hitting with an object, and wrestling. Since the security camera footages contain different light and coloring conditions, these variations are also taken into consideration to increase the diversity in the dataset further. In addition, security camera footages from different places are collected like cafe, bar, street, bus, shops, etc. This way, the variety in the dataset is ensured. Fight scenarios are independent from the environment of the surveillance camera as seen in Fig. 5.

This dataset is publicly available and can be accessed through https://github.com/sayibet/fight-detection-surv-dataset.

B. Results

Each experiment is conducted for each three datasets: Hockey, Peliculas, and surveillance camera dataset. For feature extraction part, VGG16 and Xception architectures are tested. In addition, a modified Xception architecture is trained using the fight scenes from Hockey dataset and named as Fight-CNN.

For the classification part, regular LSTMs and Bi-LSTMs are tested along with VGG16 and Xception models. Also the network is augmented by attention layer, which are tested by Xception and Fight-CNN. For each CNN, two classifiers which are Bi-LSTM with attention or Bi-LSTM without attention, are considered. In CNN and LSTM experiments, to observe the effect of number of frames to the accuracy, frame numbers are changed between 5 and 10.

Number of epochs is 20, batch size is 10 for Fight-CNN experiments and 100 for VGG16 and Xception experiments. Datasets are split as 80% for training and 20% for testing. Experimental results are presented in terms of test accuracy in Tables 2-3-4.

Since Fight-CNN is trained with the scenes from Hockey dataset, the test result of the Fight-CNN on Peliculas is not as good as can be seen in Table 2. The Peliculas dataset has little amount of fight scenes samples, so the accuracy is

TABLE II
EXPERIMENTAL RESULTS ON PELICULAS DATASET.

	Peliculas Dataset	
	10 Frames	5 Frames
	accuracy	*accuracy*
VGG16 + LSTM	95%	100%
VGG16 + Bi-LSTM	100%	100%
Xception + LSTM	97.5%	97.5%
Xception + Bi-LSTM	97.5%	97.5%
Xception + Bi-LSTM + attention	**100%**	**100%**
Fight-CNN + Bi-LSTM	77.5%	80%
Fight-CNN + Bi-LSTM + attention	87.5%	90%

TABLE III
EXPERIMENTAL RESULTS ON HOCKEY DATASET.

	Hockey Dataset	
	10 Frames	5 Frames
	accuracy	*accuracy*
VGG16 + LSTM	87.05%	92.5%
VGG16 + Bi-LSTM	92.5%	91%
Xception + LSTM	93.5%	93.5%
Xception + Bi-LSTM	94.5%	95%
Xception + Bi-LSTM + attention	**97.5%**	**98%**
Fight-CNN + Bi-LSTM	95.5	93.5%
Fight-CNN + Bi-LSTM + attention	96%	95%

highly affected by the false predictions. Therefore the standard deviation of accuracy is higher than others. At the end of the training, loss values of Bi-LSTM methods are mostly lower than the regular LSTM models. As it is observed in Table 2, addition of the attention layer significantly increases the accuracy compared to the other approaches.

The Hockey dataset experiments indicate the advantage of Bi-LSTMs over regular LSTMs as seen in Table 3. The attention layer shows its effect again when it is compared with the Xception and Fight-CNN experiments. The results of Fight-CNN along with Bi-LSTM and attention are found to be promising. Since the Xception network that we use in Fight-CNN structured with few parameters, it gives lower accuracy compared with regular Xception network. On the other hand, Fight-CNN contains less number of parameters and extracts features faster than regular Xception network.

As can be seen in Table 4, the results for surveillance camera dataset is not as good as the ones presented for the other datasets. Since the variety of the samples in this dataset is very high, the models cannot easily generalize to this dataset.

TABLE IV
EXPERIMENTAL RESULTS ON COLLECTED SURVEILLANCE CAMERA DATASET.

	Surveillance Camera Fight Dataset	
	10 Frames	5 Frames
	accuracy	*accuracy*
VGG16 + LSTM	62%	61.67%
VGG16 + Bi-LSTM	45%	52%
Xception + LSTM	60%	55%
Xception + Bi-LSTM	63.3%	63%
Xception + Bi-LSTM + attention	69%	68%
Fight-CNN + Bi-LSTM	68.5	70%
Fight-CNN + Bi-LSTM + attention	**71%**	**72%**

The results show that Fight-CNN provides a better feature extraction on the data, when it is compared to Xception model. Since the CNN is familiar with the fight scenes that it is trained with, it can extract the significant features more easily. Again the attention layer increased the accuracy in both regular Xception and Fight-CNN with its focusing ability.

It is observed that the number of frames per video parameter has no direct correlation with the accuracy in most of the cases. However, using five frames per video has less computation load for the feature extraction step compared with using ten frames per video.

V. DISCUSSION

The proposed method has benefited from the CNNs for feature extraction from frames. Two-way learning of bi-directional LSTMs and the attention layers that can also determine the amount of given attention to each part of the sequence are found to improve the accuracy. As a result, proposed method has surpassed the state-of-the-art performance. Additionally, a new model is tested by using Fight-CNN, a modified version of Xception model.

Bi-LSTMs show better performance than regular LSTMs in action recognition, as also stated in related studies in [8], [9]. Also the studies in [3], [4], [6] show that the attention layer improves the performance of sequence learning. This study validates this finding and shows that using Bi-LSTM together with attention is a promising solution to classify fight scenes.

The experimental results also indicate that the more diversity a dataset contains, the more challenging it gets to classify fight scenes. Since the collected surveillance fight dataset contains different types of fight events, from different locations, under different conditions, it poses a significant challenge for the state-of-the-art action recognition systems.

VI. CONCLUSION

The main objective of this study is detecting fight scenes from surveillance cameras in a fast and accurate way. The proposed method which employs attention layer along with Bi-LSTM networks has improved the detection accuracy and provided promising results. Moreover, using a pre-trained Fight-CNN for feature extraction proves its effectiveness on surveillance camera dataset experiments.

Another important contribution of the study is the collected surveillance camera fight dataset, which presents further challenges for automatic fight detection. This surveillance camera dataset can be extended by adding new samples from security camera footages on streets or underground stations.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[2] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–6.

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044*, 2015.

[4] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[6] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Thirty-first AAAI Conference on Artificial Intelligence*, 2017.

[7] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1647–1656.

[8] 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*. Springer, 2016, pp. 517–531.

[9] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1961–1970.

[10] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

[11] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3d convolutional neural networks," in *International Symposium on Visual Computing*. Springer, 2014, pp. 551–558.

[12] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *International Workshop on Human Behavior Understanding*. Springer, 2011, pp. 29–39.

[13] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[14] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, 2016.

[15] B. Peixoto, B. Lavi, J. P. P. Martin, S. Avila, Z. Dias, and A. Rocha, "Toward subjective violence detection in videos," in *ICASSP*, 2019.

[16] C.-H. Demarty, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, and C. Penet, "Benchmarking violent scenes detection in movies," in *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2014, pp. 1–6.

[17] 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, pp. 332–339.

[18] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 1–6.

[19] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[20] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveillance videos," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2019.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*, 2017.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[26] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[27] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.

[28] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 375–383.

[29] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, vol. 162, 2015.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[31] S. Zhang, E. Staudt, T. Faltemier, and A. K. Roy-Chowdhury, "A camera network tracking (camnet) dataset and performance baseline," in *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 365–372.

[32] W.-C. Wang, P.-C. Chung, C.-R. Huang, and W.-Y. Huang, "Event based surveillance video synopsis using trajectory kinematics descriptors," in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, 2017, pp. 250–253.

[33] C.-R. Huang, P.-C. J. Chung, D.-K. Yang, H.-C. Chen, and G.-J. Huang, "Maximuma posteriori probability estimation for online surveillance video synopsis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1417–1429, 2014.