ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413686

VIOLENCE DETECTION IN VIDEOS BASED ON FUSING VISUAL AND AUDIO INFORMATION

Wen-Feng Pang *Qian-Hua He* * *Yong-jian Hu* *Yan-Xiong Li*

School of Electronic and Information Engineering
South China University of Technology, Guangzhou, China

ABSTRACT

Determining whether given video frames contain violent content is a basic problem in violence detection. Visual and audio information are useful for detecting violence included in a video, and are usually complementary; however, violence detection studies focusing on fusing visual and audio information are relatively rare. Therefore, we explored methods for fusing visual and audio information. We proposed a neural network containing three modules for fusing multimodal information: 1) attention module for utilizing weighted features to generate effective features based on the mutual guidance between visual and audio information; 2) fusion module for integrating features by fusing visual and audio information based on the bilinear pooling mechanism; and 3) mutual Learning module for enabling the model to learn visual information from another neural network with a different architecture. Experimental results indicated that the proposed neural network outperforms existing state-of-the-art methods on the XD-Violence dataset.

Index Terms*—Violence Detection, Co-Attention, Information Fusion, Mutual Learning*

1. INTRODUCTION

With the development of information technology, a huge amount of video data is generated every moment, and detecting violence, such as fighting, shooting, and explosions in these videos is important for public security. Since manual detection can only be employed for limited number of videos, automated violence detection is a meaningful study and could be applied in many fields, such as intelligent surveillance, assessing videos uploaded to mobile applications, and prison guard robots.

Visual and audio information have been combined in different tasks. For example, Oh *et al.* [1] presented a method for reconstructing a face by using a short audio recording, while Ginosar *et al.* [2] tried to convert speech signals to gestures. Kazakos *et al.* [3] combined the RGB, optical flow, and audio features for egocentric action recognition. Aytar *et al.* [4] employed video information to help their model to learn sound representation. Such methods validate that visual and audio information have a strong correlation and complement each other. In the violence detection task, audio information in violent events is important; for example, fighting is often accompanied by shouting and swearing, or there will be a loud crash when a car accident occurs. In cases where the target may be blocked by obstacles or blurred by dim light, and audio information will be a good supplement. Since existing studies mainly detected violence using only visual information, we explored violence detection based on the fusion of visual and audio information.

Some early studies [5, 6, 7] have applied multimodal information to violence detection; however, these studies trained with small-scale datasets such as [8] and simple hand-crafted features such as haar-like features caused weak generalization and low stability. Moreover, such methods cannot be transferred to the task explored in the present paper directly because they utilized video shots during the testing phase. Recent studies [9, 10, 11] mainly utilized visual information for detecting violence. Hanson *et al.* [9] and Sudhakaran *et al.* [10] applied deep learning models and utilized adjacent frame differences as visual input features, while Xu *et al*. [11] proposed motion scale-invariant feature transform (MoSIFT) combing histogram of gradients (HoG) and histogram of flow (HoF) descriptors obtained from RGB data and optical flow. Sultani *et al.* [12] applied visual features extracted from a 3D-convolutional network (C3D) [13] and multiple instance learning (MIL) [14] in their proposed method. Recently, [15] built a large-scale dataset XD-Violence for multimodal learning in the violence detection task. They also proposed a HL-Net leveraging self-attention [16] and GCN [17]. HL-Net aims to extract the feature dependency at the temporal dimension but lacks consideration of fusing visual and audio features, which only concatenates them as input features.

Compared with HL-Net, we mainly focus on exploring methods for fusing visual and audio information. Specifically, we utilize three modules with different focuses and functions to fuse the multimodal information. The attention module utilizes the weighted features obtained by mutual guidance between visual and audio information to generate more representative features. The fusion module integrates visual and audio features by the Hadamard product based on the bilinear pooling method. The mutual learning module forces the model to extract more useful visual information from a neural network with different architecture, further improving the proposed neural network performance.

The main contribution of this paper is that a neural network containing three complementary modules for fusing visual and audio information was proposed. The experimental results validate that each module is effective, and the module combination boosts the performance compared to those of existing methods. Additionally, some experiments were conducted to further understand the module combination methods.

*Corresponding author

The work was supported by the National Nature Science Foundation of China (Grant No. 61571192, 61771200);

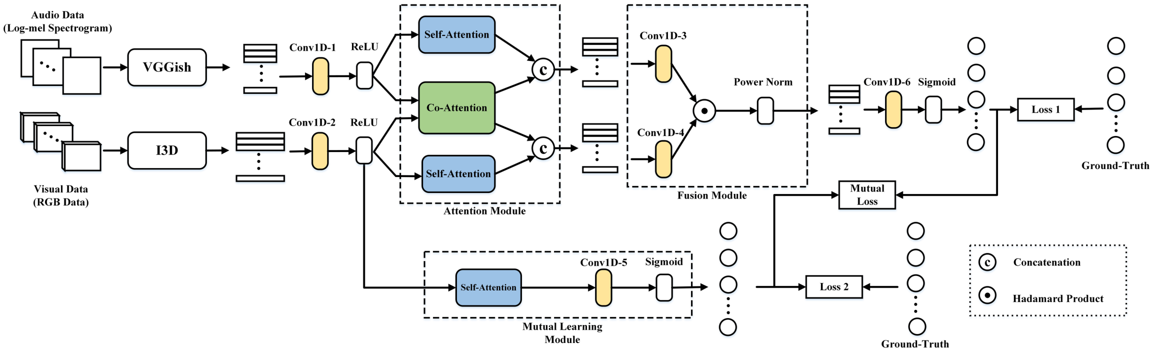

Figure 1. The overall architecture of the proposed neural network.

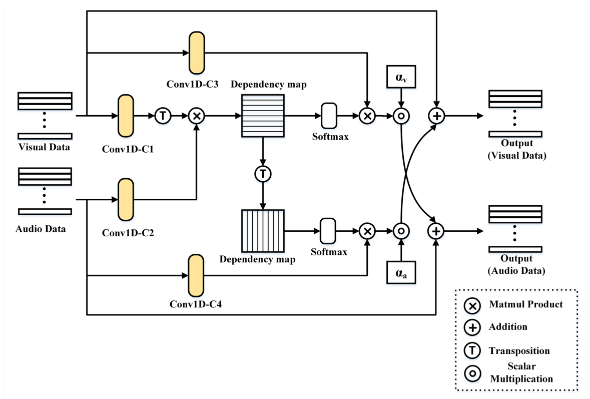

Figure 2. The architecture of the co-attention module.

2. METHODS

The overall architecture of our proposed neural network is shown in Figure 1. Following [15], visual features and audio features are first extracted from the pretrained models I3D [18] and VGGish [19], respectively. After that, the two types of features are fused by three different modules shown in Figure 1. Instead of simply stacking fusion modules, each module in the proposed neural network has different functions and concentrates on different fusion targets. Especially, since the dimension of a visual feature (1024) is different from an audio one (128), two 1D-convolutional layers (Conv1D-1 and Conv1D-2 in Figure 1) with kernel size 3 are utilized to unify the dimension number before the feature is fed to the attention module.

2.1. Attention module

The attention module assigns weights to features in a video and audio sequence and then generates effective features by integrating these weighted features. Inspired by [20], we applied co-attention, shown in Figure 2, to extract dependency between visual and audio information. We denote $X_v \in R^{C \times T}$ and $X_a \in R^{C \times T}$ as the visual and audio input features, where T is the total time steps, and C is the number of visual and audio feature dimensions. The dependency map A is defined as follows:

$$A = X_v^T W_1^T W_2 X_a , \qquad (1)$$

where $W_1 \in R^{C \times C}$ and $W_2 \in R^{C \times C}$ are trainable parameters implemented by 1D-convolutional layer with kernel size 1 (Conv1D-C1 and Conv1D-C2 in Figure 2) for translating the input feature to a new feature space. Map A demonstrates the dependency between each time step of the visual and audio feature sequence, which guides each other's feature integration based on their interactions. For example, in a video with an explosion, the explosion sound feature will assign relevant video features (e.g., video features involving flash) higher weights and generate more reprehensive features; also, the video feature involving flames will guide the model to pay attention to the audio features with an explosive sound. Such mutual guidance improves the effectiveness of visual and audio features. The outputs of co-attention are defined as follows:

$$X_{cv} = \alpha_a W_4 X_a \xi(A^T) + X_v ;\ X_{ca} = \alpha_v W_3 X_v \xi(A) + X_a , \qquad (2)$$

where ξ is the softmax function for normalizing the weights in map A and A^T by column. Parameters α_v and α_a are scalars for scaling attention features. Similar to the formulation (1), $W_3 \in R^{C \times C}$ and $W_4 \in R^{C \times C}$ are trainable parameters for abstracting the corresponding features.

In addition to co-attention, since visual and audio features can detect violence independently, the self-attention method is applied to process both types of features. The visual features generated by co-attention and corresponding self-attention modules are concatenated, and the same procedure is applied to audio features. In summary, the final output of the attention module is:

$$X'_v = f_c(X_{cv}, X_{sv});\ X'_a = f_c(X_{ca}, X_{sa}) , \qquad (3)$$

where f_c is a concatenation operation for concatenating input features in channel-wise, and $X_{sv} \in R^{C \times T}$ and $X_{sa} \in R^{C \times T}$ are the outputs of the corresponding self-attention modules.

2.2. Fusion module

In this study, fusion module is realized based on multimodal low-rank bilinear (MLB) [21] pooling. MLB is modified from the bilinear pooling method, which applies Hadamard product instead of an outer product to reduce the feature dimension. Compared with attention module generating new visual and audio features based on mutual guidance between visual and audio information, fusion module fuses multimodal information directly, and the generated features contain both visual and audio information. Considering that the Hadamard product's output magnitude may be dramatic, we append power normalization after the Hadamard product [22]. The fusion module output X_f can be achieved as:

$$\begin{aligned} z &= W_v X'_v \circ W_a X'_a ; \\ \mathrm{X}_f &= sign(z)\left|z\right|^{0.5} \end{aligned} . \qquad (4)$$

In the above formulation, W_v and W_a are the trainable parameters implemented by 1D-convoultional layer with kernel size 3, and $\circ$ is the Hadmard product.

We also tried to add an L2 normalization function to normalize the outcomes in time dimension, but this greatly decreased the performance. We argue that unlike the classification task, L2 normalization is not suitable for the current task demanding the model output score in each time step, as L2 normalization function will destroy the correlation between generated features at each time-step, making it difficult for the last 1D-convoluatonal layer (Conv1D-6 in Figure 1) with kernel size 3 to learn the pattern of these features.

2.3. Mutual learning module

The mutual learning module aims to enable the master branch (the stack of attention and fusion module) to learn visual information from a different branch. We did not construct another whole neural network like [23] but built a branch separated from Conv1D-2, shown in Figure 1, for reducing the computing and memory cost. This separated branch is implemented by the mutual learning module, which contains only a self-attention module. Considering that visual features contain more information than audio features in the violence detection task and always yield better performance when training a model with single modal information, we feed visual features instead of audio features to the mutual learning module. Since our task is to output violence scores for each time step, mean square error (MSE) loss ($L_{mut}(Y_1,Y_2)=\|Y_1-Y_2\|_2^2$) is utilized rather than the Kullback-Leibler divergence (KD) loss [24]. This is because KD loss is mainly used in classification problems that require a model utilizing normalization function such as softmax function before producing the final output.

In practical applications, we can flexibly remove the mutual learning branch and only retain the master branch for inference during the determination phase to reduce computational and memory consumption.

2.4. Training and determination methods

In this study, we treat the violence detection task as a weakly supervised detection problem where only video-level labels are offered in the training set, and frame-level labels prediction is required by the task, and MIL is applied in the training phase. Following [12, 15], we take each video as a bag and each video segment as an instance; a video including any violence is a positive bag (labeled as 1), and a video excluding violence is a negative bag (labeled as 0). In this study, K instances with the highest scores are employed to calculate the training loss with the MSE function:

$$L_b(l,Y)=\frac{\|l-f_{K-\max}(Y)\|_2^2}{K}, \quad (5)$$

where Y=[y_1,y_2,...,y_T] is the prediction (violence score) for a branch and $l\in R^K$ is the ground-truth vector of the corresponding input segment. Since the Sigmoid function is set before producing final output, y_i in Y ranges from (0, 1). The function $f_{K\text{-max}}$ is to select the K highest instances in a violence score sequence.

When training the overall model, we combine the prediction losses and mutual learning loss with γ:

$$L=L_b(l,Y_{mas})+L_b(l,Y_{ml})+\gamma L_{mut}(Y_{mas},Y_{ml}), \quad (6)$$

where Y_{mas} and Y_{ml} are the predictions for the master branch and mutual learning module, respectively.

Table 1. Comparison of different normalization methods in the fusion module.

Normalization Method	AP(%)
None (ReLU)	80.23
Power Normalization	80.90
L2 Normalization	77.25
Power + L2 Normalization	78.22

During the determination phase, the output of each branch is calculated as the violence score directly, and the combination violence score with weight β is defined as a weighted sum:

$$S=(1-\beta)Y_{mas}+\beta Y_{ml}. \quad (7)$$

In practical applications, a threshold can be set in advance. If the violence score is higher than the threshold during the determination phase, the system will make an alarm.

3. EXPERIMENTS

3.1. Dataset

To the best of our knowledge, XD-Violence is the only existing large-scale dataset offering visual and audio data of violent events at the same time. Specifically, XD-Violence consists of 4757 videos (217 hours) and six types of violent events. Since only samples in the testing set are tagged with frame-level labels, weakly supervised methods should be applied to this dataset. In addition to the small scale of other datasets, many datasets, such as [25, 12], do not provide audio data; meanwhile, quite a few audio clips in [26, 27] are silent or background music irrelevant to the video content. Therefore, we evaluate the proposed method on the XD-Violence dataset.

3.2. Implementation details

During the training stage, Adam [28] is utilized as the optimizer with an initial learning rate 5e-3 with cosine annealing algorithm. A total of 30 epochs and a batch size of 64 are set for each model. The values of γ and β are set as the same, and unless otherwise stated, they are set as 5e-2. Parameters α_v and α_a in the co-attention module are set as 1. Notably, for fair evaluation, visual and audio features utilized are the same as [15], which are extracted from a pretrained I3D model and VGGish model. During the testing phase, frame-level average precision (AP) is utilized as the evaluation metric.

3.3. Comparison of normalization methods in the fusion module

Because the violence detection task requires the model to predict violence scores at each time step, the normalization method in fusion module needs to be redesigned. Table 1 shows the performance training when only the master branch is used. The first row in Table 1 is the result without normalization but a ReLU [29] activation after the Hadamard product in the fusion module. When applying power normalization, performance is improved (see data in the second row), thereby proving the validity of reducing the magnitude variation of Hadamard product outcomes. Besides, the performance descends whether L2 normalization is utilized alone or in combination with power normalization. We conclude that L2 normalization destroys the feature relevance at different time steps, making optimization difficult during the training phase.

Table 2. Comparison of different channel extension numbers in the fusion module.

Expansion Multiple	AP(%)
5	80.26
3	80.39
1	80.90
0.5	80.57

Table 3. Comparison of different γ values.

	Ind	0.01	0.05	0.1	0.2
ML Module	77.32	77.65	78.06	77.33	77.35
Master Branch	80.90	80.95	81.28	81.10	80.18
Combination	-	81.39	81.69	81.45	80.56

3.4. Comparisons of different channel extension numbers in the fusion module

Referring to [30], we evaluated the channel extension methods on the convolutional layers before using the Hadamard product (i.e., layer Conv1D-3 and Conv1D-4 in Figure 1). All experimental results were obtained by training with the master branch. The first two rows in Table 2 are the results of expanding the convolutional layer channel by three and five times; after introducing the Hadamard product, the channel number is restored to the original one again with a sum pooling. However, the channel extension brings no improvement to the model whereas keeping the channel unchanged (the third row) results in better performance. We argue that the current dimension number (256) is large enough to represent the video information, and channel extension may cause overfitting that reduces the performance. As shown in the last row of Table 2, we also tried to reduce the number of channels by half, which results in the AP decreasing because of the information loss in input features of the fusion module.

3.5. Comparison of different γ

During the training phase, γ is an important parameter determining the degree to which the two branches learn from each other. As shown in Table 3, compared with the APs of 77.32% and 80.90% in the first column obtained by training the mutual learning module (ML module in Table 3) and master branch independently, the performances of both combined and single branch increase when training the master branch and mutual learning module together with an appropriate γ. Besides, too large of a γ harms the model as it may interfere with each branch's ability to learn independently, and the proposed neural network gets the best performance when the γ value is 0.05.

3.6. Ablation study on different modules

To validate the effectiveness of combining the use of different modules, we conducted ablation experiments. As shown in Table 4, the combination of attention and fusion modules effectively improved performance over utilizing one of them alone, thus indicating these two modules are complementary. The last row shows the master branch performance when combing all three modules. Although the improvement is not much large compared with the combination result of the two modules, it consumes no additional resources as this result is given by removing the mutual learning module during the determination phase.

Table 4. Ablation study on different modules.

Attention Module	Fusion Module	ML Module	AP(%)
●	-	-	80.03
-	●	-	78.23
●	●	-	80.90
●	●	●	81.28

Table 5. Comparison with existing methods on the XD-Violence dataset.

Method	AP (%)
SVM	50.78
OCSVM [31]	27.25
Hasan *et al.* [32]	30.77
Sultani *et al.* [12]	73.20
Wu *et al.* [15]	78.64
Ours (master branch)	81.28
Ours (combination)	81.69

Especially, we also experimented with training audio data and visual data solely, getting 69.96% and 77.32 APs, respectively. These performances are worse than those obtained when using multimodal information (Table 4). Hence, the experimental results indicate that multimodal information is helpful in the violence detection task.

3.7. Comparison with existing methods

As shown in Table 5, we compared the proposed method with some existing methods. The first two rows methods are the methods with hand-crafted features, and their performances show that traditional methods are ineffective in the current violence detection task. The last two rows methods are the proposed methods with the attention, fusion, and mutual learning modules; both the performance of combination of master branch and mutual learning modules and that of single master branch exceed existing state-of-the-art methods. We argue the main reason is that the proposed method focuses on the fusion of audio and visual information, but the existing methods consider it insufficiently.

4. CONCLUSION

This paper explored the fusion methods that combine visual and audio features in the violence detection task. Specifically, we proposed a neural network consisting of three modules (e.g., attention module, fusion Module, mutual learning module) with different properties. The experimental results show the validity of each module, and the combination of the modules effectively boosts the model's overall performance on the XD-Violence dataset. Extensive experiments demonstrate the factors that affect the proposed neural network, such as normalization methods in the fusion module. Modifying the proposed neural network to enhance its real-time characteristic is a direction for future work.

REFERENCES

[1] Oh, T.H., Dekel, T., Kim, C., Mosseri, I., Freeman, W.T., Rubinstein, M., Matusik, W. “Speech2face: Learning the face behind a voice,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7539-7548, 2019.

[2] Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J. “Learning individual styles of conversational gesture.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3497-3506, 2019.

[3] Kazakos, E., Nagrani, A., Zisserman, A., Damen, D. “Epic-fusion: Audio-visual temporal binding for egocentric action recognition,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5492-5501, 2019.

[4] Aytar, Y., Vondrick, C., Torralba, A. “Soundnet: Learning sound representations from unlabeled video,” Advances in neural information processing systems (NIPS), pp. 892-900, 2016.

[5] Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S. “Audio-visual fusion for detecting violent scenes in videos,” Hellenic conference on artificial intelligence, pp. 91-100. Springer, 2010.

[6] Lin, J., Wang, W. “Weakly-supervised violence detection in movies with audio and video based co-training,” Pacific-Rim Conference on Multimedia, pp. 930-935.Springer, 2009.

[7] Zajdel, W., Krijnders, J.D., Andringa, T., Gavrila, D.M. “Cassandra: audio-video sensor fusion for aggression detection,” IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS),. pp. 200-205. IEEE, 2007.

[8] Demarty, C.H., Penet, C., Soleymani, M., Gravier, G. “Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation,” Multimedia Tools and Applications 74(17), 7379-740, 2015.

[9] Hanson, A., Pnvr, K., Krishnagopal, S., Davis, L. “Bidirectional convolutional lstm for the detection of violence in videos,” Proceedings of the European Conference on Computer Vision (ECCV), pp. 0-0, 2018.

[10] Sudhakaran, S., Lanz, O. “Learning to detect violent videos using convolutional long short-term memory,” IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-6. IEEE, 2017.

[11] Xu, L., Gong, C., Yang, J., Wu, Q., Yao, L. “violent video detection based on mosift feature and sparse coding,” Acoustics, Speech and Signal Processing (ICASSP), pp. 3538–3542. IEEE, 2014.

[12] Sultani, W., Chen, C., Shah, M. “Real-world anomaly detection in surveillance videos,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6479-6488, 2018.

[13] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. “Learning spatiotemporal features with 3d convolutional networks,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4489-4497. 2015.

[14] Zhou, Z.H., “Multi-instance learning: A survey,” Department of Computer Science & Technology, Nanjing University, Tech. Rep, Mar. 2008.

[15] Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., & Yang, Z. “Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision,” In European Conference on Computer Vision (ECCV), pp. 322-339. 2020.

[16] Wang, X., Girshick, R., Gupta, A., He, K. “Non-local neural networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 7794-7803, 2018.

[17] Kipf, T.N., Welling, M. “Semi-supervised classification with graph convolutional networks,” arXiv preprint arXiv:1609.02907, 2016.

[18] Carreira, J., Zisserman, A. “Quo vadis, action recognition? a new model and the kinetics dataset,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6299-6308, 2017.

[19] Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C.,Plakal, M., Platt, D., Saurous, R.A., Seybold, B., *et al.* “Cnn architectures for large-scale audio classification,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131-135. IEEE (2017)

[20] Nguyen, D. K., & Okatani, T. “Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6087-6096, 2018.

[21] Kim, J. H., On, K. W., Lim, W., Kim, J., Ha, J. W., & Zhang, B. T. “Hadamard product for low-rank bilinear pooling,” arXiv preprint arXiv:1610.04325, 2016.

[22] Gao, Y., Beijbom, O., Zhang, N., & Darrell, T. “Compact bilinear pooling,” In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 317-326, 2016.

[23] Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. “Deep mutual learning,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4320-4328, 2018.

[24] Hinton, G., Vinyals, O., & Dean, J. “Distilling the knowledge in a neural network” arXiv preprint arXiv:1503.02531, 2015.

[25] Nievas, E.B., Suarez, O.D., Garca, G.B., Sukthankar, R. “Violence detection in video using computer vision techniques,” International Conference on Computer Analysis of Images and Patterns. pp. 332-339, Springer, 2011.

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

[27] Perez, M., Kot, A.C., Rocha, A. “Detection of real-world fights in surveillance videos,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2662-2666. IEEE, 2019.

[28] Kingma, D. P., & Ba, J. “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[29] Xu, B., Wang, N., Chen, T., & Li, M. “Empirical evaluation of rectified activations in convolutional network,” arXiv preprint arXiv:1505.00853, 2015.

[30] Yu, Z., Yu, J., Fan, J., & Tao, D. “Multi-modal factorized bilinear pooling with co-attention learning for visual question answering,” In Proceedings of the IEEE international conference on computer vision (CVPR), pp. 1821-1830, 2017.

[31] Scholkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C. “Support vector method for novelty detection,” Advances in Neural Information Processing Systems (NIPS), pp. 582–588, 2000.

[32] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S, “Learning temporal regularity in video sequences” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–742,