LOOK, LISTEN AND PAY MORE ATTENTION: FUSING MULTI-MODAL INFORMATION FOR VIDEO VIOLENCE DETECTION

Dong-Lai Wei[1], *Chen-Geng Liu*[2], *Yang Liu*[1], *Jing Liu*[1], *Xiao-Guang Zhu*[3], *Xin-Hua Zeng*[1*]

[1]Academy for Engineering & Technology, Fudan University, Shanghai, China
[2]School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia
[3]SEIEE, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Violence detection is an essential and challenging problem in the computer vision community. Most existing works focus on single modal data analysis, which is not effective when multi-modality is available. Therefore, we propose a two-stage multi-modal information fusion method for violence detection: 1) the first stage adopts multiple instance learning strategies to refine video-level hard labels into clip-level soft labels, and 2) the next stage uses multi-modal information fused attention module to achieve fusion, and supervised learning is carried out using the soft labels generated at the first stage. Extensive empirical evidence on the XD-Violence dataset shows that our method outperforms the state-of-the-art methods.

Index Terms— Violence detection, multi-modal information, fused attention, deep learning, weak supervision

1. INTRODUCTION

Violence detection is crucial in maintaining social security, which has been researched for years. Nievas *et al.* [1] used bag-of-words framework for fight detection; Hassner *et al.* [2] proposed Violent Flows (ViF) descriptor to distinguish violent events. However, solely using visual information to build a violence detection model is not robust or powerful enough. For example, it is difficult to obtain sufficient information in a surveillance area with obstacles or dim light, and in these cases, audio will be a good supplement. Multi-modal information can provide comprehensive and copious features, which can be more robust and accurate in detecting violence than single-modal information. Therefore, our study is based on the fusion of audiovisual features.

In the research of video anomaly detection, weakly supervised learning becomes the mainstream. Compared to marking each frame with frame-level labels, marking a video-level label is simple, saving much time and human resources. Sultani *et al.* [3] introduce the multiple instance learning (MIL) [4] strategy into the field of video anomaly detection for the first time. However, they only adopt coarse-grained labels, which weaken the learning capability of the model by using rudimentary information. Feng *et al.* [5] and Tian *et al.* [6] develop models on the basis of them. They use 3D-convolutional network (C3D) [7] and two-stream inflated 3D convNet (I3D) [8] to extract visual features, and achieve the best results on two large-scale datasets (UCF-Crime [3] and ShanghaiTech [9]). The above works [3, 5, 6] only focus on single-modal visual features, inspired by them, we develop a weakly supervised method and further transform coarse-grained labels to fine-grained labels, which can fully utilize multi-modal information. Even though the progress has been achieved by other multi-modal methods [10, 11, 12], they have the common flaws that their methods are specifically designed on small-scaled datasets and inadequate utilization of audiovisual features. Wu *et al.*[13] propose XD-Violence large-scaled dataset and HL-Net for violence detection. In their method, the audiovisual features are directly spliced, which lacks the implicit alignment and co-attention between multi-modal information.

To address the above issues, we propose a novel method. First, refining clip-level soft labels is realized through MIL strategies. Next, the Audiovisual Co-attention Fusion (ACF) network is proposed to help the model explore the correlations of multi-modal features. Our contributions are summarized as follows: **1)** we propose a weakly supervised violence detection model targeting multi-modal information. **2)** We propose a multi-modal co-attention mechanism to encourage the model to learn the audiovisual features of violent information. **3)** We conduct extensive qualitative and quantitative experiments. Experiment results on benchmark demonstrate the effectiveness of the ACF network.

2. METHODS

2.1. Overview of the methods

The overall architecture of our method is shown in Figure 1. Similar to [13], we use pretrained I3D [8] and VGGish [14] to extract audiovisual features from original videos and audios, respectively. Given the extracted features, we apply a three-layer MLP to obtain clip-level soft labels. After that,

*Corresponding author.

This work is supported by the Shanghai Key Research Laboratory of NSAI.

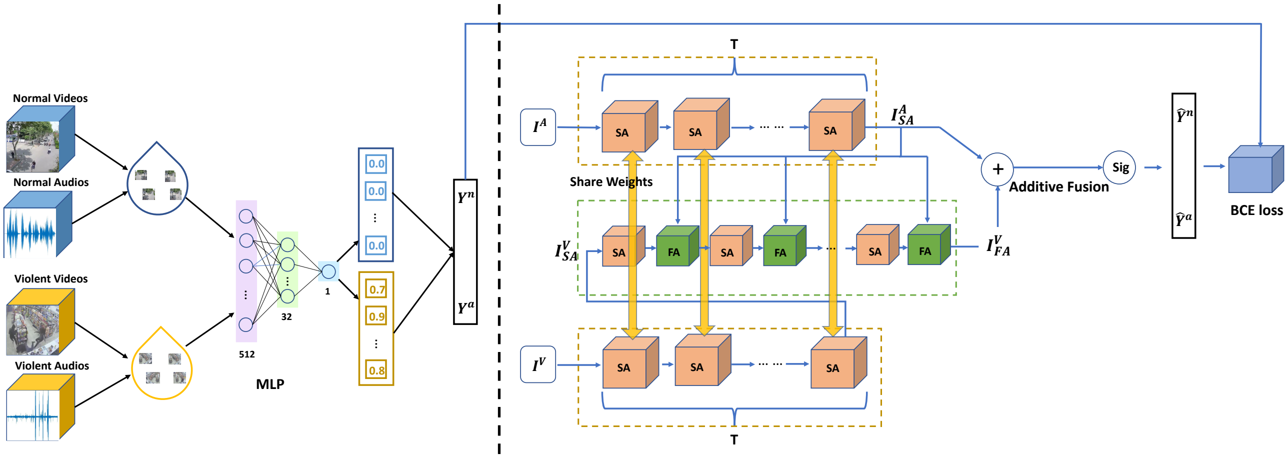


Fig. 1. The overall architecture of the proposed method.

we further introduce the ACF network, which enables multi-modal features to pay mutual attention to achieve the best fusion effect to the greatest extent. The ACF network consists of Single Self-attention (SA) module and Fusion Co-attention (FA) module. Each module contains stacks of SA units and FA units, respectively. Both video and audio features are first sent to the SA module for self-attention processing and then enter the FA module to complete the co-attention enhancement. We use binary cross-entropy loss to train the ACF network supervised by clip-level soft labels of violent information and clip-level annotations of normal information.

2.2. Clip-level labels refinement via MIL

Unlike [3] which assigns coarse-grained labels to each clip, we use a refinement process to create clip-level soft labels. Clip-level soft labels have more fine-grained annotation information, and the ACF network can be better supervised in the next stage with them. We use Bag_v and Bag_n to represent the set of positive and negative bags. The positive bag is a violent video with its associated audio information. In contrast, the negative bag is a normal video with audio. Video segments in bags served as instances. We select the top K pairs of instances with the largest violent score from the sets of bags to calculate the loss. Our overall loss function is:

$$L_{Total} = L_{MIL} + \frac{\lambda}{K} \cdot (\sum_{i=1}^{K} L_{BCE}), \tag{1}$$

where λ working as a regularized hyperparameter. L_{MIL} is a function based on [3], we design as follows:

$$L_{MIL} = \max\big(0, 1 - \max_{(1 \le k \le K)} Bag_v^k + \max_{(1 \le k \le K)} Bag_n^k\big). \tag{2}$$

L_{BCE} is a binary cross-entropy function working on violent instances as regularization. Because the normal instances greatly outnumber violent instances, violent instances are calculated with more attention. The formula is as follows, where y_i is video-level hard label corresponding to k_i :

$$L_{BCE} = -\big[y_i log k_i + (1 - y_i) log(1 - k_i)\big]. \tag{3}$$

Based on this, we can train a shallow clip-level soft label generator to obtain fine-grained labels and provide them to the ACF network for supervised training.

2.3. Audiovisual Co-attention Fusion network

Our ACF network is shown in the right part of Figure 1, mainly composed of SA and FA modules. The ACF network can fuse audiovisual features and effectively detect violent information with these two modules based on attention.

2.3.1. Multi-modal information based on attention

Following the ideas in [15], we use I_k, I_v and I_q to apply attention to the single modal information. The calculation process is as follows:

$$Attention(I_k, I_v, I_q) = softmax(\frac{I_q \cdot I_k}{\sqrt{d}}) \cdot I_v. \tag{4}$$

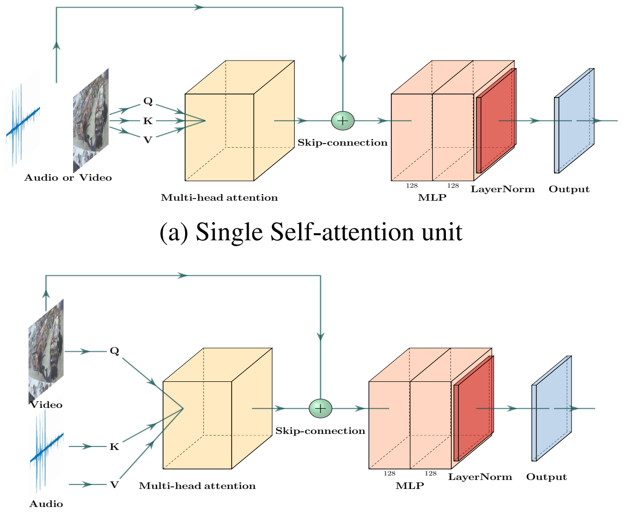
Use I_q to calculate the matching degree with each I_k matrix and obtain the corresponding self-attention scores through the softmax function, where d is the dimension of the input. Based on this weighted summation of all I_v matrices, the self-attention calculation completes in the SA unit. In order to improve the expressivity of the modules, we further adopt multi-head attention [15], which consists of H paralleled heads and W^o is the corresponding parameter matrix:

$$\begin{aligned} Multihead(I) &= Multihead(I_k, I_v, I_q) \\ &= Concat(head_1, head_2, ..., head_H) \cdot W^o, \end{aligned} \tag{5}$$

where I_k, I_v and I_q are calculated from the single modal features (video or audio as K, V and Q) and its corresponding learnable parameter matrix W^k, W^v and W^q:

$$head_i = Attention(KW_i^k, VW_i^v, QW_i^q). \tag{6}$$

We design two basic units for multi-modal information based on the above method to construct SA and FA modules.



(a) Single Self-attention unit

(b) Fusion Co-attention unit

Fig. 2. The details of SA and FA units.

2.3.2. *Single Self-attention and Fusion Co-attention module*

The Single Self-attention module is composed of T cascaded SA units (Figure 2(a)). It is mainly composed of a multi-head attention layer and a fully connected layer, in which residual connections [16] are implemented to obtain multiple levels of features. L and M individually represent fully connected layer and layer normalization [17], and $[\cdot]_T$ means continuous cascaded connection T times:

$$I_{SA}^{V/A} = \left[L\Big(M\big(I^{V/A} + Multihead(I_k, I_v, I_q) \big) \Big) \right]_T . \quad (7)$$

We design two types of SA module based on multi-modal information, weights sharing mode and non-sharing mode. The information of different modalities can perform a parallel calculation by two SA modules. In the sharing mode, different modal information can share common SA module weights.

Similar to the module above, the FA module is composed of T FA units cascaded, where the FA unit is shown in Figure 2(b). Take audiovisual features I_{SA}^{V} and I_{SA}^{A} as a set of examples. Features achieve mutual attention between each other through the multi-head attention layer. The calculation process is as follows:

$$I_{FA}^{V} = \left[L\Big(M\big(I_{SA}^{V} + Multihead(I_{SA_k}^{A}, I_{SA_v}^{A}, I_{SA_q}^{V}) \big) \Big) \right]_T . \quad (8)$$

Overall, the ACF network includes an SA module and an FA module. Passing the SA module, features share the audio and video weights in the module, and self-attention calculated are enhanced. The video features enter the FA module and process with the assistance of the audio features. Mutual attention between multi-modal information in the FA module further augments their correlation to achieve the best fusion effect. Therefore, the two obtained modal features can be merged effectively.

3. EXPERIMENTS

3.1. Implementation details

Datasets. XD-Violence[13] is the largest and most representative multi-modal dataset in the field of violence detection currently. Specifically, it has 2405 violent videos and 2349 non-violent videos (217 hours in total). The dataset covers six common types of violence, including two modal information of videos and corresponding audios.

Training stage. During the first stage, the values of λ and K are 10^{-3} and 7, respectively. For the next stage, we use Adam [18] as the optimizer to train the ACF network, and we set the learning rate to 10^{-4}. A total of 50 epochs and a batch size of 32 are set for training. T and H are hyperparameters with values of 6 and 8. In both stages, we use sigmoid as the activation function to calculate violent scores.

Evaluation metrics. To make a fair comparison with the previous methods [19, 3, 13], we use frame-level average precision (AP) as an evaluation metric. Meanwhile, we provide the area under the curve (AUC) of the frame-level receiver operating characteristics (ROC). According to [3, 20], we also adopt the false alarm rate (FAR) for evaluating robustness.

3.2. Ablation studies

In order to verify the validity of the labels obtained in the first stage, we conducted experiments as shown in Table 1. During the training of the ACF network, we can use video-level hard labels or clip-level soft labels. The experiment shows that the clip-level soft labels significantly improve the performance of the ACF network. Because video-level hard labels only have the rough label information, they can not provide the degree of abnormality and the specific segments information. This seriously affects the model learning, so training the ACF network without clip-level soft labels has a bad performance.

Table 1. Comparisons of different labels.

	FAR (%)	AUC-ROC (%)	AP (%)
$ACF^{w/o\ Soft\ labels}$	67.57	14.42	13.83
ACF	1.12	93.87	80.13

We can select multi-modal features in the ACF network training to enter either the weights sharing SA module or the non-sharing SA module. As shown in Table 2, the weight-sharing module results are better. We argue that in weights sharing, the two modal information based on the same set of parameters can further learn from each other and collaboratively improve model performance. Using weights sharing also saves nearly half of the parameters of the SA module. Multi-modal information can improve the performance

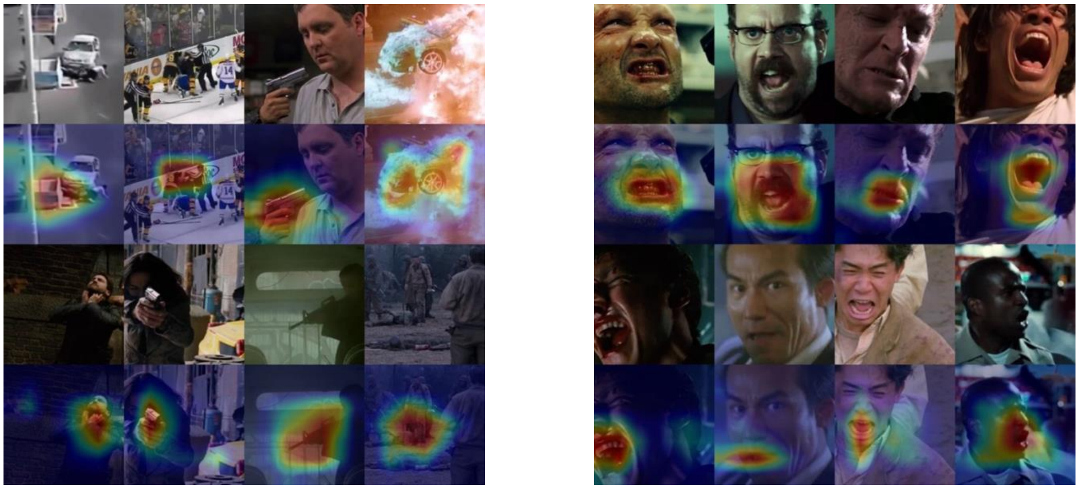


Fig. 3. Visualization results of violence maps (best viewed in color).

Table 2. Comparisons of shared and non-shared SA module.

	FAR (%)	AUC-ROC (%)	AP (%)
$ACF^{w/o\ Sharing\ SA}$	1.51	93.57	79.06
ACF	1.12	93.87	80.13

of the model. In order to prove that multi-modal information has the unique advantages of good complementarity, we conducted ablation experiments on the information of different modalities. Unlike processing multi-modal information, single-modal information only passes through the SA module in the ACF network and does not require fusion operations. The results of Table 3 show that only the use of audio modal information to detect violence is relatively poor. Using modal video information is better than it. This result is in line with our prediction because the information provided by the video modal is much richer than the audio. The multi-modal information can help them complement each other, which significantly improves the model performance.

Table 3. Comparisons of different modal information.

Video	Audio	FAR (%)	AUC-ROC (%)	AP (%)
✗	✓	17.30	75.84	50.74
✓	✗	1.96	91.82	72.09
✓	✓	1.12	93.87	80.13

3.3. Visualization of fused attention results

To further evaluate the performance of our model, we visualize predictions of the ACF network based on the XD-Violence via Grad-CAM [21, 22]. The left part of Figure 3 shows that our model can notice the region where violent events occur. Interestingly, we find that (the right part of Figure 3) when people scream or rage in violent scenes, their mouths are usually noticed with the help of violent audio information. This is a sound proof that the video and audio can get adequate fused attention after passing through the ACF network. It illustrates multi-modal audiovisual information based on fused attention has significant values in violence detection.

3.4. Comparison with existing methods

As shown in Table 4, the outcomes based on traditional methods for detecting violence are relatively poor, and the results obtained by deep learning methods are better. The method we propose outperforms the state-of-the-art method. The main reason for our method working excellent is the good mutual attention and fusion of multi-modal information.

Table 4. Comparisons of existing methods.

Method	AP (%)
SVM	50.78
OCSVM [23]	27.25
Hasan *et al.*[19]	30.77
Sultani *et al.* [3]	73.20
Wu *et al.*[13]	78.64
ACF (ours)	**80.13**

4. CONCLUSION

This paper focuses on the violence detection task with multi-modal information. We propose a two-stage weakly supervised learning method, which pays more attention to the fusion of multi-modal features. The experiment results and visualization process demonstrate the effectiveness of our ACF network. More different types of modal information may be used in future work, and finding a suitable mutual attention fusion method will be our research direction.

5. REFERENCES

[1] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar, “Violence detection in video using computer vision techniques,” in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2011, pp. 332–339.

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

[3] Waqas Sultani, Chen Chen, and Mubarak Shah, “Real-world anomaly detection in surveillance videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[4] Zhi-Hua Zhou, “Multi-instance learning: A survey,” *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, vol. 1, 2004.

[5] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng, “Mist: Multiple instance self-training framework for video anomaly detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14009–14018.

[6] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro, “Weakly-supervised video anomaly detection with robust temporal feature magnitude learning,” *arXiv preprint arXiv:2101.10030*, 2021.

[7] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[8] Joao Carreira and Andrew Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[9] Weixin Luo, Wen Liu, and Shenghua Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 341–349.

[10] Wojtek Zajdel, Johannes D Krijnders, Tjeerd Andringa, and Dariu M Gavrila, “Cassandra: audio-video sensor fusion for aggression detection,” in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 200–205.

[11] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis, and Sergios Theodoridis, “Audio-visual fusion for detecting violent scenes in videos,” in *Hellenic Conference on Artificial Intelligence*. Springer, 2010, pp. 91–100.

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

[13] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang, “Not only look, but also listen: Learning multimodal violence detection under weak supervision,” in *European Conference on Computer Vision*. Springer, 2020, pp. 322–339.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., “Cnn architectures for large-scale audio classification,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 131–135.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.

[18] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[19] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis, “Learning temporal regularity in video sequences,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 733–742.

[20] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei, “Weakly supervised video anomaly detection via center-guided discriminative learning,” in *2020 IEEE International Conference on Multimedia and Expo*. IEEE, 2020, pp. 1–6.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D Parikh, and D Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[23] Bernhard Schölkopf, Robert C Williamson, Alexander J Smola, John Shawe-Taylor, John C Platt, et al., “Support vector method for novelty detection.,” in *Advances in Neural Information Processing Systems*. Citeseer, 1999, vol. 12, pp. 582–588.