2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)

Audio-Guided Attention Network for Weakly Supervised Violence Detection

Yujiang Pu
State Key Laboratory of Media Convergence and Communication,
Communication University of China
Beijing, China
pyj2020@cuc.edu.cn

Xiaoyu Wu*
State Key Laboratory of Media Convergence and Communication,
Communication University of China
Beijing, China
wuxiaoyu@cuc.edu.cn

2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE) | 978-1-6654-0886-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCECE54139.2022.9712793

Abstract—**Detecting violence in video is a challenging task due to its complex scenarios and great intra-class variability. Most previous works specialize in the analysis of appearance or motion information, ignoring the co-occurrence of some audio and visual events. Physical conflicts such as abuse and fighting are usually accompanied by screaming, while crowd violence such as riots and wars are generally related to gunshots and explosions. Therefore, we propose a novel audio-guided multimodal violence detection framework. First, deep neural networks are used to extract visual and audio features, respectively. Then, a Cross-Modal Awareness Local-Arousal (CMA-LA) network is proposed for cross-modal interaction, which implements audio-to-visual feature enhancement over temporal dimension. The enhanced features are then fed into a multilayer perceptron (MLP) to capture high-level semantics, followed by a temporal convolution layer to obtain high-confidence violence scores. To verify the effectiveness of the proposed method, we conduct experiments on a large public violent video dataset, i.e., XD-Violence. Experimental results demonstrate that our model outperforms several methods and achieves new state-of-the-art performance.**

Keywords-Violence Detection; multi-modal fusion; weak supervision; multiple instance learning

I. INTRODUCTION

Violence such as abuse, fighting, and gunshots not only adversely affects the physical and mental health of individuals but also poses a serious threat to public security. Therefore, monitoring of public violence has become increasingly important for preventing criminal behavior. However, most existing surveillance systems require manual inspection, and the scarcity of highly qualified security personnel leads to low efficiency and false alarm detection. In addition, monitoring large amounts of video footage for long periods can easily lead to distraction and fatigue due to the limited capabilities of alarm devices. As a result, there is a growing demand for automatic violence detection systems today.

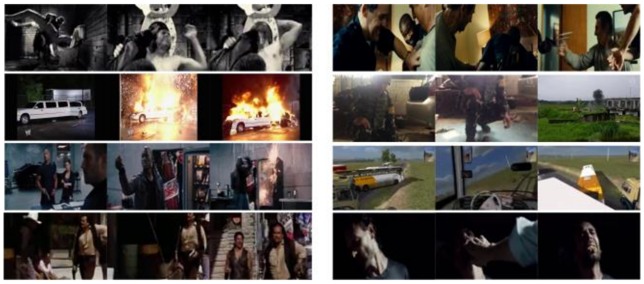

Figure 1. Multiple visual and audio violence in complex scenarios.

In the early days, most work typically uses hand-crafted feature descriptors to construct visual representations of frame sequences. Clarin et al. [1] develop the DOVE system, which detects gore and action violence by analyzing the motion intensity between adjacent frames. In [2], Nievas et al. use the bag-of-words model combining Space-Time Interest Points (SITP) and motion scale-invariant feature transform (MoSIFT) descriptors for fight detection. Similarly, Hassner et al. [3] propose the Violent Flow (ViF) descriptor to detect violence in the crowd. Since visual violence such as abuse, fighting, and riot is generally accompanied by auditory elements such as screaming, explosions, and gunshots, as shown in Fig. 1, it is recognized that audio information plays a complementary role in violence detection. Giannakopoulos et al. [4] utilize several shallow audio features including Zero Crossing Rate (ZCR), spectrograms, and Mel-frequency Spectral Coefficients (MFCC) to detect violence in movie scenes. [5] proposes a two-stage detection process, in which audio and video classifiers are combined in a co-training manner to obtain credible violence prediction scores. Penet et al. [6] explore temporal integration and multimodal information fusion strategy for violence detection. The above methods perform well on simple behaviors and use fewer computing resources, but they rely too much on feature engineering resulting in less robustness.

978-1-6654-0886-8/22/$31.00 ©2022 IEEE

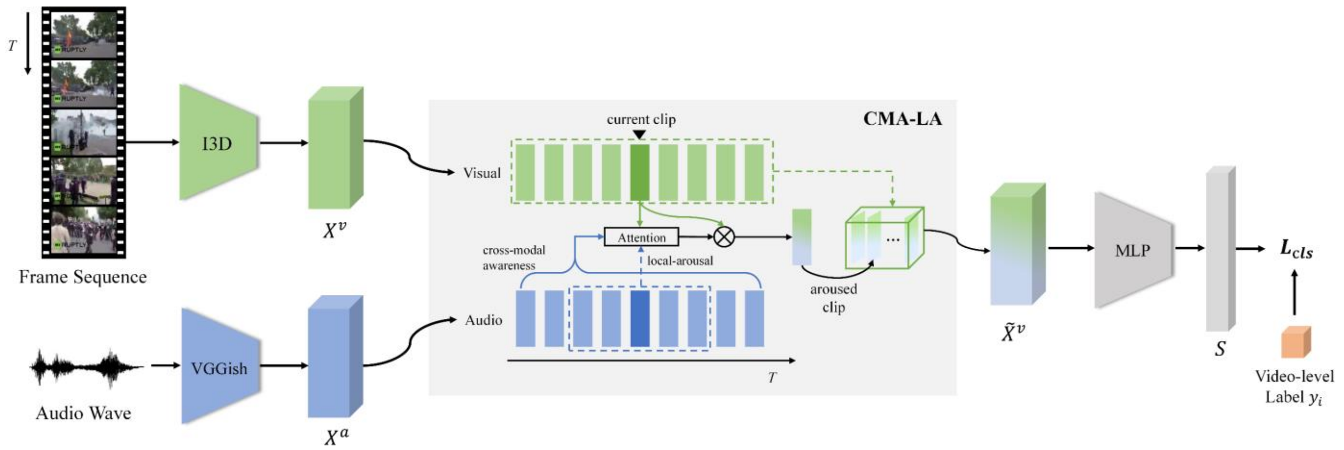

Figure 2. Overview of the audio-guided multimodal violence detection framework.

Recently, deep neural networks have been widely used for violence detection, which attributes to their powerful feature representation capabilities. Fudan-Huawei [7] uses a 2D neural network to extract spatial features of video frames, including RGB features and motion features, after which a long short-term memory (LSTM) network is introduced to capture long-range dependencies. Multiple SVM-based classifiers are fused for a reliable result. Zhou et al. [8] construct a FightNet for detecting complex violence interactions, in which the acceleration field of optical flow is proposed to describe motion properties. Gu et al. [9] use two network structures, P3D and LSTM, to model the correlation of continuous frame sequences, while VGGish is utilized to extract audio features for auxiliary detection. In [10], Wu et al. build an HL-Net in which graph convolution is used to construct the temporal relationships of adjacent video clips, while positional coding is employed to capture local-range dependencies. Most of the above methods use visual information for violence detection, and only simple concatenation or addition operations for audio information, ignoring the contextual correlation between audio and video in the temporal order.

In this paper, we propose an audio-guided attention network for multimodal violence detection. Instead of directly fusing audio information with visual features, it is used for cross-modal awareness to re-calibrate the visual field. A self-adaptive Gaussian-like position prior is then used to activate contextual information and reduce channel-wise redundancy. The enhanced features are fed into a two-layer MLP for semantic encoding, followed by a 1D convolutional layer for temporal causal inference. Extensive experiments on a benchmark dataset, XD-Violence, validate the effectiveness of our approach and achieve competitive results compared to other state-of-the-art methods.

II. Proposed Method

In this section, we describe our multimodal violence detection framework in detail, as shown in Fig. 2. First, two deep neural networks are employed to extract visual and audio features, respectively. Then, a Cross-Modal Awareness Local-Arousal (CMA-LA) module is proposed to implement cross-modal interactions, which further calibrates the video representation over the temporal dimension. Finally, the abstract semantics obtained after MLP is used for violence detection, and our objective function is designed based on multiple instance learning (MIL) under weak supervision.

A. Feature Extraction

Given an untrimmed video, we first divide it into non-overlapping clips, where each clip contains 16 frames. Subsequently, these clips are sent into a pretrained I3D [11] model to extract visual features. Hence, the visual feature is denoted as $X^v \in \mathbb{R}^{T\times D_v}$, where T is the number of clips and D is the feature dimension. For the audio waveform, we divide it into overlapping 960ms segments, where each segment is aligned with the end of the video clip. The VGGish [12] network pre-trained on Audioset [13] is used to extract audio features $X^a \in \mathbb{R}^{T\times D_a}$, corresponding to a feature dimension of 128.

B. Cross-Modal Awareness-Local Arousal Module

After obtaining visual and audio features, modeling the contextual relations of multimodal representations is further explored. Instead of using the common feature concatenation or channel-wise summation, we propose a cross-modal aware-local arousal (CMA-LA) module to achieve enhancement of audio to visual features. In this process, audio is used to generate a global attention map A across modalities, which is expressed as:

$$A = softmax(Q_a K_v^T / \sqrt{D}) \quad (1)$$

where $Q_a = X^a W^q$ and $K_v = X^v W^k$ are two linear projection functions, and D is the dimension of hidden layers. The obtained attention map describes the global correlation between audio and visual features over the temporal dimension, allowing the visual clip to gain global audio perception. Subsequently, we use a self-adaptive position prior to achieve local arousal of the visual representation, which is expressed as:

$$G_{ij} = \exp\left(-|w(i-j)^2 + b|\right) \quad (2)$$

where i and j are the relative distances of the clips, and w b are two learned parameters used to control the neighborhood of the center position and adjust the weight of the current clip, respectively. By adding the local prior G to the global attention map A, we obtain a local-arousal attention map $\tilde{A}$, which calibrates the weights within the neighborhood of the current clip while suppressing information redundancy in the temporal order, as shown in Fig. 3. Thus, the cross-modal awareness (CMA) can be formulated as follows:

$$\tilde{X}^v = LN(X^v + W^h(\tilde{A}X^vW^v)) \tag{3}$$

where W^h and W^v are two linear mapping layers and $LN(\cdot)$ denotes a layer norm operation. A shortcut connection is also used to maintain the original distribution of visual features.

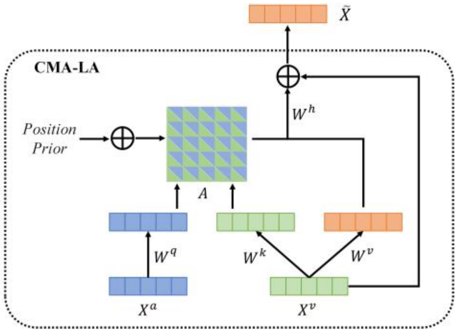

Figure 3. Structure of Cross-Modal Awareness Local-Arousal (CMA-LA)

C. Multilayer Perceptron

Subsequently, we feed the enhanced visual features into an MLP to capture the high-level semantics, which consists of two 1D Convolution layers and a GELU activation. The operation is represented as:

$$X^H = MLP(\tilde{X}) \tag{4}$$

Finally, a temporal convolution layer is used to capture historical observations and obtain prediction scores, which is expressed as follows:

$$y^S = \sigma(WX^H + b') \tag{5}$$

where $\sigma(\cdot)$ denotes a sigmoid function, W denotes a 1×1 convolution of kernel size K, and b' is a bias term. Causal convolution allows us to capture historical observations, which provides a strong discriminative basis for violence detection. Finally, we obtain the violence predictions $y^S = \{y_t\}_{t=1}^{T}$ within a video bag.

D. Object Function

In this paper, violence detection is considered as a multiple instance learning (MIL) task under weak supervision. Following [14][15][10], we use the mean value of the k-max predictions in the video bag as the violence score, where $k = \left\lfloor \frac{T}{q} + 1 \right\rfloor$. The k-max predictions with the highest violence scores in the positive bag are most likely to contain violence, while the k-max predictions in the negative bag are usually hard samples which may lead to false alarm detection. Therefore, our objective function is expressed as:

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N} -y_i \log(\bar{y}^S) \tag{6}$$

where y_i is the video-level annotation, and $\bar{y}^S$ is the average value of the k-max predictions in the video bag.

III. Experiments

To validate the effectiveness of the proposed method, we conduct experiments on a challenging violent video dataset, **XD-Violence** [10], which is also the largest publicly available violence dataset containing both video and audio. The dataset contains a total of 4754 untrimmed videos, with a total duration of 217 hours. Six common types of violence including abuse, car accident, explosion, fighting, riot, and shooting are covered. Following [10], we use the frame-level average precision (AP) as the evaluation metric, which is more sensitive to unbalanced classes (e.g., violence). First, we present implementation details of the experimental setup. Then, we compare the performance with previous state-of-the-art and show a series of ablation studies, which further validate the superiority of the model in this paper.

A. Experimental Setup

The hidden layer in CMA has a dimension of 128, the units of the two-layered MLP are 512 and 128, respectively, with a dropout layer of which rate is 0.1. T is empirically set to 200 and q is set to 16. The Adam optimizer is used to update the network parameters, where the batch size is set to 128 and the initial learning rate is set to 0.0005.

B. Quantitative analysis

Effect of CMA-LA. We first explore the effect of cross-modal awareness (CMA) on model performance, as shown in Table 1. It can be seen that the frame-level AP value is 73.88% in the case of RGB features only, while it significantly improves to 82.15% after introducing the global attention map generated via audio. This suggests that reasonable modeling of contextual associations across modalities is useful and necessary. Meanwhile, we analyze the local-arousal (LA) role of the position prior. Notably, the position prior over temporal dimension shows a 1.39% improvement on a global-aware basis. With a self-adaptive position prior, the attention weights of the current clip are dynamically adjusted, and the position preference from the i^{th} clip to the j^{th} clip is calibrated, further achieving temporal alignment of audio and visual information.

TABLE I. EFFECT OF CMA-LA MODULE ON XD-VIOLENCE

Method	AP (%)
RGB only	73.88
RGB + CMA	82.15
RGB + CMA + LA	83.54

Effect of temporal convolution. Also, we analyze how different kernel sizes of temporal convolution affect the model performance. Fig. 4 shows that the test AP value gradually improves with the increasing of the kernel size until it exceeds a specific value, which indicates that larger convolution kernels may cause overfitting. The best performance is reached on XD-

Violence when K is set to 7, and the corresponding average precision is 83.54%.

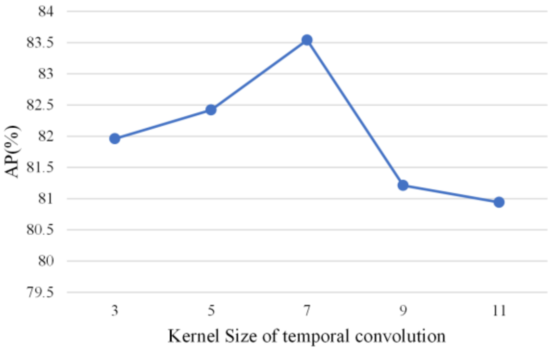

Figure 4. Effect of different kernel size of temporal convolution

Comparison with state-of-the-art methods. Finally, we compare the currently available state-of-the-art methods. It is clear that our model substantially outperforms Wu et al. [10] by 4.9% and Pang et al. [18] by 1.85%, achieving a new state-of-the-art result. The former uses a GCN for temporal modeling of segments and does not explore multimodal fusion. In contrast, Pang et al. use co-attention to fuse information from visual and audio. Unlike explicitly combining visual and audio information, we only use audio as a cross-modal prior to re-calibrate the visual features, and eventually use the enhanced visual features for inference. This also demonstrates that audio has a non-negligible role in violence detection.

TABLE II. FRAME-LEVEL AP PERFORMANCE ON XD-VIOLENCE

Method	**AP (%)**
SVM	50.78
OCSVM [16]	27.25
Hasan et al. [17]	30.77
Sultani et al. [14]	73.20
Wu et al. [10]	78.64
Pang et al. [18]	81.69
Ours	**83.54**

IV. CONCLUSION

In this paper, we propose a novel audio-guided multimodal violence detection framework. First, deep neural networks are utilized to extract visual and audio features, respectively. A global attention map with audio perception is generated through cross-modal awareness (CMA) module, which implicitly learns global audio information. Local Arousal (LA) is subsequently performed on the attention map using a self-adaptive position prior, and the temporal neighborhood of visual clips is re-calibrated to suppress channel-wise noises. To validate the proposed method, we conduct experiments on a challenging violent video dataset, i.e., XD-Violence, and achieve a leading result. Comprehensive experiments indicate the effectiveness of our approach. In the future, we will further explore the role of different position prior, while the multimodal interaction approach also deserves in-depth investigation.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (No.61801441), National Key R&D Program of China (No. 2021YFF0900701) and in part by discipline construction project of "Beijing top notch" discipline (Internet information of Communication University of China).

REFERENCES

[1] Clarin, C., Dionisio, J., Echavez, M., & Naval, P. (2005). DOVE: Detection of movie violence using motion intensity analysis on skin and blood. PCSC, 6:150-156.

[2] Nievas, E. B., Suarez, O. D., García, G. B., & Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In: International conference on Computer analysis of images and patterns. Berlin, Heidelberg. pp. 332-339.

[3] Hassner, T., Itcher, Y., & Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, RI. pp. 1-6.

[4] Giannakopoulos, T., Pikrakis, A., & Theodoridis, S. (2007). A multi-class audio classification method with respect to violent content in movies using bayesian networks. In: 2007 IEEE 9th Workshop on Multimedia Signal Processing. Chania, Crete, Greece. pp. 90-93.

[5] Lin, J., & Wang, W. (2009). Weakly-supervised violence detection in movies with audio and video based co-training. In: Pacific-Rim Conference on Multimedia. Berlin, Heidelberg. pp. 930-935.

[6] Penet, C., Demarty, C. H., Gravier, G., & Gros, P. (2012). Multimodal information fusion and temporal integration for violence detection in movies. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan. pp. 2393-2396.

[7] Dai, Q., Zhao, R. W., Wu, Z., Wang, X., Gu, Z., Wu, W., & Jiang, Y. G. (2015). Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning. In: MediaEval.Wurzen, Germany. pp. 6-10.

[8] Zhou, P., Ding, Q., Luo, H., & Hou, X. (2017). Violent interaction detection in video based on deep learning. In: Journal of physics: conference series. IOP Publishing, 2017, 844(1): 012044.

[9] Gu, C., Wu, X., & Wang, S. (2020). Violent Video Detection Based on Semantic Correspondence. IEEE Access, 8: 85958-85967.

[10] Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., & Yang, Z. (2020). Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: European Conference on Computer Vision. Glasgow. pp. 322-339.

[11] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI. pp. 6299-6308.

[12] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017). CNN architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp) New Orleans, Louisiana. pp. 131-135.

[13] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) New Orleans, Louisiana. pp. 776-780.

[14] Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT. pp. 6479-6488.

[15] Paul, S., Roy, S., & Roy-Chowdhury, A. K. (2018). W-talc: Weakly-supervised temporal activity localization and classification. In:

Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany. pp. 563-579.

[16] Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., & Platt, J. C. (1999). Support vector method for novelty detection. In: NIPS. Denver. pp. 582-588.

[17] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV. pp. 733-742.

[18] Pang, W. F., He, Q. H., Hu, Y. J., & Li, Y. X. (2021). Violence Detection in Videos Based on Fusing Visual and Audio Information. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . Toronto, Ontario. pp. 2260-2264.