International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 07 July 2024, pp: 763-768 www.ijaem.net ISSN: 2395-5252

A Region-based Convolutional Neural Network for Fighting Detection

Vu Thi Khanh Trinh
Tan Trao University, Tuyen Quang, Vietnam

Date of Submission: 15-07-2024 Date of Acceptance: 25-07-2024

ABSTRACT: Action recognition is one of the fundamental topics and has many practical applications especially related to security and surveillance issues. The fight recognition problem is one of the popular applications today based on action recognition. Along with the development of deep learning, there have been many proposed models that give good results on many different datasets. In this paper, we propose a region-based convolutional neural network for fighting detection. Experimental results show that the proposed model achieves the best performance on two popular datasets: Hockey and Movies. This has proven the effectiveness of the proposed model.

KEYWORDS: Fighting Detection, violence, Convolutional Neural Network, Deep learning.

I. INTRODUCTION

School violence is currently becoming quite common in most countries around the world. According to a report by the United Nations crime prevention agency, each year around the world, about 4-6 million students are directly involved in school violence. This data is increasing, making school violence a common problem in international education.

Therefore, a system or device that identifies fighting behavior can promptly overcome significant consequences for people, especially providing active support for students - preventing school violence. Early identification can help families and schools detect it promptly and help limit the consequences of fights. The need for systems to identify cases of brawls and emergency alerts to parents or other people involved is essential when brawls occur.

As depicted in Figure 1, a brawl recognition system based on the surveillance camera approach includes three main steps as follows:

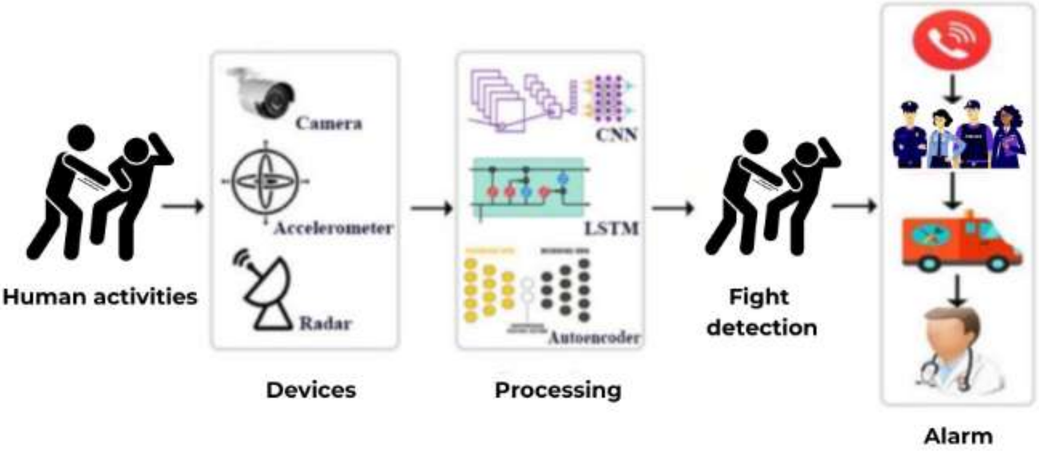

Figure 1. An overview of the fighting detection system

Step 1. Data preprocessing: Videos recorded directly by surveillance cameras will be included in the process of determining the presence of humans in the video. In this step, the function of preprocessing is to synchronize, reduce noise and remove noisy signals, as well as normalize the data to obtain the best signals for the process of identifying fighting behavior (Fight). detection).

Several popular data preprocessing methods can be used, for example: Fourier transform, HOG, image segmentation,...

Step 2. Fight detection: Receive the video processed in Step 1 to analyze and determine whether there was a fight or not. During this Fight detection process, recognition systems often use machine learning algorithms such as SVM, Deep learning, Decision Trees,... to identify fight events in videos. Since the input signal is usually a sequence of frames, thresholding algorithms are rarely applied in this step. This comes from the complexity of the frame sequence being large and not having easily identifiable action features.

Step 3. Alarm generation: In this step, the system will activate the alarm sound and send notifications to the mobile phone to alert supervisors, security guards or request an ambulance to arrive. to take the injured person to the hospital. Techniques commonly applied in this step include: collecting and transmitting GPS location information of the parties involved in the altercation, continuously emitting audio warning signals to help people around easily Locate the victim,...

There have been many machine learning models proposed to solve this problem, such as Nam et al. [1], proposing to identify violent scenes in videos using blood detection and recording the level of violence. movement, as well as sounds characteristic of violent events. Cheng and colleagues [2] study gunshots, explosions, and vehicle braking in audio using a hierarchical method based on Gaussian mixture models and Hidden Markov models (HMM). Giannakopoulos et al. [3] also proposed a violence detector based on audio features.

In recent years, with the development of deep learning, many methods based on neural networks have been proposed to solve the problem e.g., image classification [4, 5], speech recognition [6], natural language processing [7], action recognition [8, 9], video prediction [10], etc. For fighting recognition, in 2015, Serrano et al. proposed a fighting behavior recognition method based on blob motion [11]. Then, Soliman et al. [12] proposed a method to identify fights on video using deep learning techniques, convolutional neural network (CNN) models, and short-length neural networks. term (LSTM). This method uses video frames as input, extracts location features using the VGG-16 model and temporal features using LSTM, and then uses a set of fully connected layers to classify type. In 2022, Qi et al published a two-step training method based on weak surveillance data for a deep video combat detection model. This method is intended to help the model focus on shorter moments during the learning process. The author uses a base network to extract video features and a score generator network to generate prediction scores for video segments. The two-step training method consists of phase one, in which the score generation network is trained on weak surveillance data and the base network is frozen, and phase two, in which both networks are trained together. each other using weak monitoring data generated from phase one [13]. In this paper, we propose a Region-based Convolutional Neural Network to solve the problem of brawl recognition. The framework is built to improve the model's feature learning ability. As a result, the performance of the proposed model improves significantly in the experimental results.

II. THE PROPOSED FRAMEWORK

ResNet (Residual Neural Network) is a deep neural network architecture introduced by Kaiming He and colleagues in 2015 [14]. ResNet was originally designed to solve the problem of suppressing accuracy degradation as deep neural networks become deeper. However, this architecture has proven good for many tasks in the field of computer vision, including object detection problems.

The ResNet network builds on the concept of basic blocks called "Residual Blocks". Each Residual Block consists of convolutional layers and skip connections. The ResNet architecture is based on adding short connections to pass information directly from previous layers to subsequent layers, thereby minimizing information loss and increasing the learning ability of the network (see Table 1).

Table 1. The ResNet architecture

Layer name	**Output size**	**18-layer**	**34-layer**	**50-layer**	**101-layer**	**152-layer**
Conv1	112x112	7x7, 64, stride 2				

Conv2_x	56x56	3x3 max pool, stride 2				
		$\begin{bmatrix}3x3,64\\3x3,64\end{bmatrix}$ x2	$\begin{bmatrix}3x3,64\\3x3,64\end{bmatrix}$ x3	$\begin{bmatrix}1x3,64\\3x3,64\\1x1,256\end{bmatrix}$ x3	$\begin{bmatrix}1x3,64\\3x3,64\\1x1,256\end{bmatrix}$ x3	$\begin{bmatrix}1x3,64\\3x3,64\\1x1,256\end{bmatrix}$ x3
Conv3_x	28x28	$\begin{bmatrix}3x3,128\\3x3,128\end{bmatrix}$ x2	$\begin{bmatrix}3x3,128\\3x3,128\end{bmatrix}$ x4	$\begin{bmatrix}1x3,128\\3x3,128\\1x1,512\end{bmatrix}$ x4	$\begin{bmatrix}1x3,128\\3x3,128\\1x1,512\end{bmatrix}$ x4	$\begin{bmatrix}1x3,128\\3x3,128\\1x1,512\end{bmatrix}$ x8
Conv4_x	14x14	$\begin{bmatrix}3x3,256\\3x3,256\end{bmatrix}$ x2	$\begin{bmatrix}3x3,256\\3x3,256\end{bmatrix}$ x6	$\begin{bmatrix}1x3,256\\3x3,256\\1x1,1024\end{bmatrix}$ x6	$\begin{bmatrix}1x3,256\\3x3,256\\1x1,1024\end{bmatrix}$ x23	$\begin{bmatrix}1x3,256\\3x3,256\\1x1,1024\end{bmatrix}$
Conv5_x	7x7	$\begin{bmatrix}3x3,512\\3x3,512\end{bmatrix}$ x2	$\begin{bmatrix}3x3,512\\3x3,512\end{bmatrix}$ x3	$\begin{bmatrix}1x3,512\\3x3,512\\1x1,2048\end{bmatrix}$ x3	$\begin{bmatrix}1x3,512\\3x3,512\\1x1,2048\end{bmatrix}$ x3	$\begin{bmatrix}1x3,512\\3x3,512\\1x1,2048\end{bmatrix}$
	1x1	Average pool, 1000-d fc, softmatx				
FLOPs		$1.8x10^9$	$3.6 x10^9$	$3.8 x10^9$	$7.6 x10^9$	$11.3 x10^9$

In object detection problems, ResNet networks are often used as part of an overall architecture, such as Faster R-CNN or YOLO. The ResNet network is often used to extract features from input images, which are then fed into classification and regression layers to determine the locations and labels of objects.

A popular structure in object detection using ResNet is Faster R-CNN [15]. In Faster R-CNN, ResNet network is used to extract features from input images. Specifically, ResNet uses several convolutional layers to create a feature map with decreasing size. This feature map is then fed into a subnetwork called Region Proposal Network (RPN) to create potential region proposals containing objects. RPN will propose rectangles (bounding boxes) and calculate the probability that the object lies within each rectangle (see figure 2).

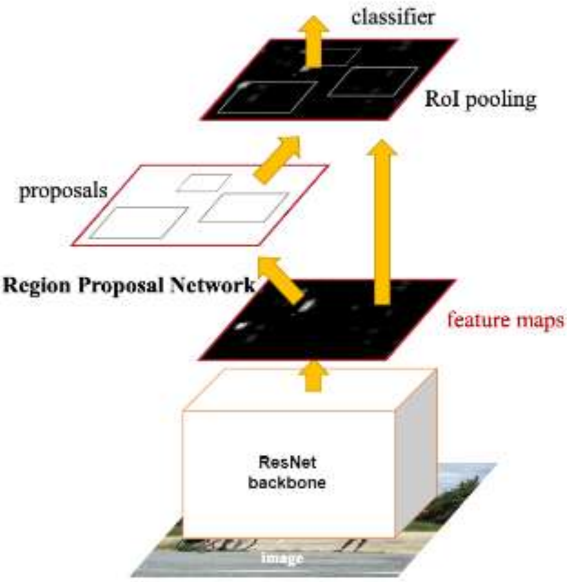

Figure 2. Faster-RCNN model for fighting detection

Faster R-CNN (Region-based Convolutional Neural Networks) is a neural network architecture used in object detection problems. It is considered one of the most advanced and popular methods for object detection and recognition in images. Faster R-CNN includes three main components: a Convolutional Neural Network (CNN) to extract features, a Region Proposal Network (RPN) to propose potential object regions, and a Fully Convolutional Network (FCN) to

Classify and adjust the position of objects. Below is a detailed description of how Faster R-CNN works:

- CNN network (usually ResNet network) is used to extract features from input images. This CNN network can be pre-trained on a large dataset, such as ImageNet, to learn basic features from images.
- The output of the CNN network is a feature map with decreasing size. This feature map contains information about local features of the image, such as edges, texture, and shape.
- RPN is a CNN-based neural network, used to suggest potential object regions in a feature map.
- RPN takes input from the feature map and applies convolutional layers to generate a large number of region proposals. Each proposed region is a rectangle whose size and coordinates are predicted to cover a certain area in the image.
- At the same time, RPN also predicts the probability that each proposed region contains an object and adjusts the size and position of the proposed rectangles.

III.EXPERIMENTATION

3.1. Datasets and Implementation

3.1.1. Datasets

- The Movies [16]: dataset includes 200 videos divided into 100 fight videos and 100 non-fight videos. Fight videos were collected from movie scenes, while non-fight videos were collected from other actions. The backgrounds in the videos are different scenes creating richness in recognition. In particular, non-fight scenes are extracted from the Public Action Recognition dataset [17]. The authors have adjusted the ratio of the videos in the dataset to a uniform size to use for recognition purposes.

Hockey dataset [16]: consists of 1000 videos with resolution 720×576 divided into 500 fight videos and 500 non-fight videos. It is collected from hockey games of the National Hockey League.

3.1.2. Implementation Detail

We apply the stochastic gradient descent employing a momentum of 0.9 for all the above networks. The weight decay is configured as 5×10^{4}. The initial learning rate is set at 0.02, and we adopt the cosine learning rate scheduler which progressively reduces the learning rates from 0.02 to 0. The mini-batch size is established as 64 images. We utilize the RanAug [18] as the data augmentation strategy for all input sample.

3.2. Comparison and Analysis

We experiment with our proposed framework on several datasets including Hocley and Movies. The experimental results are shown in Table 2. In particular, the table presents the accuracy of various methods for two different datasets: Hockey and Movies. Each method's performance is measured and compared to highlight the strengths and weaknesses of each approach in detecting or analyzing specific types of content. As can be seen, the proposed framework achieves the best performance with 94.2% and 99.4% in terms of accuracy on Hockey and Movies, respectively.

Table 2. The performance of our proposed framework compared to state-of-the-art methods on two standard datasets including Hockey and Movies in terms of accuracy.

Method	Hockey	Movies
Bermejo et al [16]	90.90	89.50
Deniz et al [17]	90.10	82.50
Gracia et al [10]	72.50	87.20
Serrano và et al [19]	94.60	99.00
Serrano et al [19]	82.60	98.00
Ours	**94.2**	**99.4**

The method by Serrano và et al. (2018) and Ours show the highest overall accuracy, particularly in the Movies dataset. Specifically, Deniz et al. (2014) and Gracia et al. (2015) display notable differences in their performance across the two datasets, indicating potential specialization or limitations. Our method not only competes with the best existing methods but slightly surpasses them in the Movies dataset, reflecting its robustness and adaptability in various content types. This comparative analysis highlights the strengths and potential areas for improvement in each method, providing insights into their applicability for different types of content analysis.

IV.CONCLUSION

In this paper, we proposed a novel framework based on Region-based Convolutional Neural Network and the ResNet architecture for fighting detection. Specifically, the ResNet is adopted to extract the features from the input sample meanwhile the Region-based Convolutional Neural Network is utilized to provide good prediction on objects specifically fighting objects. The experiment results have shown that the proposed framework achieve the best performance compared to other methods on two standard datasets including Hockey and Movies.

REFERENCES

[1]. Nam, J., Alghoniemy, M., & Tewfik, A. H. (1998, October). Audio-visual content-based violent scene characterization. In Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269) (Vol. 1, pp. 353-357). IEEE.

[2]. Cheng, W. H., Chu, W. T., & Wu, J. L. (2003, November). Semantic context detection based on hierarchical audio models. In Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval (pp. 109-115).

[3]. Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., & Theodoridis, S. (2006). Violence content classification using audio features. In Advances in Artificial Intelligence: 4th Helenic Conference on AI, SETN 2006, Heraklion, Crete, Greece, May 18-20, 2006. Proceedings 4 (pp. 502-507). Springer Berlin Heidelberg.

[4]. Vu, D. Q., Phung, T. T., Wang, J. C., & Mai, S. T. (2024). LCSL: Long-tailed Classification via Self-labeling. IEEE Transactions on Circuits and Systems for Video Technology.

[5]. Cheng, W. C., Mai, T. H., & Lin, H. T. (2023, December). From SMOTE to Mixup for Deep Imbalanced Classification. In International Conference on Technologies and Applications of Artificial Intelligence (pp. 75-96). Singapore: Springer Nature Singapore.

[6]. Tan, H. M., Vu, D. Q., & Wang, J. C. (2023, June). Selinet: a lightweight model for single channel speech separation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[7]. Nga, C. H., Vu, D. Q., Luong, H. H., Huang, C. L., & Wang, J. C. (2023). Cyclic Transfer Learning for Mandarin-English Code-Switching Speech Recognition. IEEE Signal Processing Letters.

[8]. Vu, D. Q., Le, N. T., & Wang, J. C. (2021). Self-supervised learning via multi-transformation classification for action recognition. arXiv preprint arXiv:2102.10378.

[9]. Vu, D. Q., Thu, T. P. T., Le, N., & Wang, J. C. Deep learning for human action recognition: a comprehensive review. APSIPA Transactions on signal and information processing, 12(2).

[10]. Vu, D. Q., & Thu, T. P. T. (2023). Simultaneous context and motion learning in video prediction. Signal, Image and Video Processing, 17(8), 3933-3942.

[11]. Serrano Gracia, I., Deniz Suarez, O., Bueno Garcia, G., & Kim, T. K. (2015). Fast fight detection. PloS one, 10(4), e0120448.

[12]. Soliman, M. M., Kamal, M. H., Nashed, M. A. E. M., Mostafa, Y. M., Chawky, B. S., & Khattab, D. (2019, December). Violence recognition from videos using deep learning techniques. In 2019 ninth international conference on intelligent computing and information systems (ICICIS) (pp. 80-85). IEEE.

[13]. Qi, Z., Zhu, R., Fu, Z., Chai, W., & Kindratenko, V. (2022, October). Weakly supervised two-stage training scheme for deep video fight detection model. In 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 677-685). IEEE.

[14]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[15]. Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.

[16]. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., & Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011,

Seville, Spain, August 29-31, 2011, Proceedings, Part II 14 (pp. 332-339). Springer Berlin Heidelberg.

[17]. Deniz, O., Serrano, I., Bueno, G., & Kim, T. K. (2014, January). Fast violence detection in video. In 2014 international conference on computer vision theory and applications (VISAPP) (Vol. 2, pp. 478-485). IEEE.

[18]. Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 702-703).

[19]. Serrano, I., Deniz, O., Espinosa-Aranda, J. L., & Bueno, G. (2018). Fight recognition in video using hough forests and 2D convolutional neural network. IEEE Transactions on Image Processing, 27(10), 4787-4797.