Incorporating Frame Image and Frame Sequence into Ensemble Learning Networks to Improve the Accuracy of Physical Bullying-Detecting Model

Runqian Wang
Princeton International School of Mathematics and Science, Princeton, NJ, USA
ray.wang@prismsus.org

Abstract. In this research, a convenient, reliable, and fast system is built for violence detection in schools. This system includes two independent model designed separately for detecting violence in videos and pictures. This ensemble learning structure allows the system to be more accurate and less dependent on the background, or, minimize the influence of changing circumstances. The reason is that the image-identifying network did better on classifying the image, while the video-identifying network can minimize the influence of the background by using optical flow. Therefore, by combining them together the total accuracy increases and the dependency on environment is minimized. The video-identifying model was based on Darknet19, ResNet, optical flow and LSTM. Optical flow could eliminate the background’s influence by extracting moving features. The image-identifying model was based on ResNet. ResNet was tested and selected from 3 different popular networks due to its high accuracy. The system was trained and tested on data from various datasets online as well as different videos and pictures from multiple sources. The final system achieves an accuracy of 90.90% from multiple-sourced videos, and is visualized for a more straight-forward result. The final test shows that the system not only has a high accuracy, but also can process data with an astonishing speed of 75 FPS (frames per second), which is about 3 times faster than normal videos. All of these imply a high practical value of the final system.

1. Introduction
Bullying is one of the most student-threatening factors in campus. Approximately 5 million American students from age 6 to 12 are bullied, which is 20.8% of all the students [1]. It is barely possible to monitor all of these 5 million students by teachers at school. Although most of the schools own camera in their campus, a few of them are really used to monitor students’ harmful behavior, such as bullying. One main reason of that is that there isn’t enough faculties to watch those cameras all day. Thus, it becomes necessary to design a computer system that can immediately identify students’ physical bullying behaviors using videos photographed by the cameras.

Recently, artificial intelligence has become a controversial topic, and many deep learning methods of detecting and identifying different objects have been proposed. Consequently, some research about violence detection were done based on deep learning. For example, Tencent recently announced an A.I. that could use pictures to detect terrorism [2]. Zhou et al. tried to identify violence behavior by using 3D convolutional neural network [3]. Although 3D CNN is accurate, this type of neural network has a
high requirement on hardware, and its calculation speed is slow, which makes it unlikely to be implemented in real world. Serrano et al. used a 2D network for fight recognition [4]. However, its accuracy on Hockey Dataset is 82.6%, which also makes it inapplicable. Almost all of the models only focused on one type of data, either video or picture, and no one had combined them together for a more reliable and precise result. Plus, due to the difference of the background of each camera at different school, it is uneasy to train a neural network which can be applied at all different circumstances. This leads to the idea of using a model that contains optical flow, which is a way to eliminate the influence of the background and focus on the movements of the object. Therefore, all the problems lead to an ensemble system which combines two machine learning models.

**Contributions:**

- A database was built using different sources such as movies, online databases, hockey and sport matches, and manually selected images from www.google.com. 7-fold data augmentation is implemented to produce more training data.
- The ensemble learning of combining two neural networks together in order to achieve a more reliable result. One of the networks studies each single frame in the video, the other studies the relationship between two frames. The two models used in the system were tested and the best one was chosen.
- Achieves a fast speed of 75 FPS on NVIDIA GTX 1080 when calculating the result using deep learning network, providing a high practical value for the system. At the same time, the system could achieve an accuracy of 90.90% under this speed.

2. **Materials**

The dataset used to train the CNN is collected from different sources including hockey videos, movies, UCF101 datasets, hmdb51 datasets, and pictures downloaded from www.google.com.

The total dataset includes 500 fighting videos and 500 non-fighting videos that came from Hockey Fight Dataset [5], 100 fighting videos and 101 non-fighting videos that came from Movies Dataset [5], 448 fighting videos photographed at the playground, fighting videos including punching and kicking and non-fighting videos including clapping, hugging, playing soccer, hand shaking, talking, and waving downloaded from UCF101 [6] and HMDB51 [7] database. All these videos were used to train the video-identifying system.

All of the videos described above were transferred to frames for the training of the picture-identifying system. After data augmentation, there are 193,155 fighting images and 224,665 non-fighting images in total.

Furthermore, there are 100 pictures for fighting scene and 100 pictures for non-fighting scene that were downloaded, inspected and labeled manually for the testing of the picture-identifying system. These downloaded pictures are specialized for campus environment and student fighting and bullying scenes to ensure the reliability of the system in real circumstances.

3. **Method**

3.1. **Targetted CNN System for Identifying Violence in Video**

CNN networks have developed astonishingly recently, resulting in the appearance of many new neural networks. The LSTM network is specialized for studying the relationship between two groups of data, which made it especially suitable for studying sentences. However, a neural network that studies video can also be implemented by using LSTM due to its ability to learning the relation between frames. Here a network which combines Darknet19 [8], ResNet, and LSTM, and uses optical flow to extract the difference between two frames was selected to be the video-identifying system.
This network first requires an input of two frames in the video. Then, the images’ features are extracted using a pre-trained model of Darknet19 which was pre-trained by ImageNet [9] and achieves a top-5 accuracy of 91.2%. These features are then applied to the optical flow algorithm, which is able to output the movements of the objects in the video. After that, the extracted features are put into CNN, which here uses ResNet. Finally, the results are read by the LSTM network, and at last enter the fully connected layers and output the final classification of the current frame.

The LSTM structure often includes 3 gates: an input gate, an output gate, and a forget gate. This structure can partially deal with the vanishing or exploding gradient. It is also relatively insensitive to gap length, which is another advantage of LSTM comparing to typical recurrent neural network (RNN) [10]. A typical LSTM includes formulas and variants in formula 1.

\[
\begin{align*}
    f_t &= \sigma_g \left( W_f x_t + U_f h_{t-1} + b_f \right) \\
    i_t &= \sigma_g \left( W_i x_t + U_i h_{t-1} + b_i \right) \\
    o_t &= \sigma_g \left( W_o x_t + U_o h_{t-1} + b_o \right) \\
    c_t &= f_t \cdot c_{t-1} + i_t \cdot \sigma_c \left( W_c x_t + U_c h_{t-1} + b_c \right) \\
    h_t &= o_t \cdot \sigma_c(c_t)
\end{align*}
\]

Besides from LSTM, optical flow also significantly improves the accuracy of the network. Optical flow can extract the moving objects within two frames. It describes the vector movements of the lights and colors in the video.
Another benefit of this network is the application of ResNet as CNN. This CNN is built of many units called residual units. This unit enables the network to use previous data for the training of the next layer of network. Therefore, this structure is extremely efficient at avoiding the problem of vanishing gradient, and is thus commonly used as a CNN [11]. Its top-5 accuracy reaches 96.43% at 2015 ILSVRC (ImageNet Classification) [9], thus it’s one of the best option for the CNN network.

3.2. The testing of the most suitable CNN network for violence detection in pictures combining transfer learning

In order to improve the accuracy of the whole system, besides the video-identifying system another picture-identifying system is proposed for detecting the violence in single frame. This system needs to fulfill the requirement of fast and accurate. Therefore, based on the requirements, 3 networks are selected to be tested: GoogLeNet, AlexNet, and ResNet. Here DIGITS is used on a NVIDIA GTX 1080 GPU to train all 3 networks based on SGD (stochastic gradient descent) [12].

For GoogLeNet, the reason of choosing it is the “inception” structure of the network. This structure ensures the simplicity of data by using multiple pooling layers. Furthermore, this structure supports a wide network, so the accuracy is also high [13].

As for AlexNet, this is a network proposed in 2012. The structure is simple, which makes it faster than other networks. Although it might not be that accurate, it’s still considered as one of the networks that need to be tested [14].

ResNet is already introduced in part A, so no more discussion is included here.

3.3. Data Augmentation

More training samples were formed from the previous data in order to increase the dataset size and results in a better result. Salt-pepper noise, Gaussian noise, Normal Blur, RGB value variation, Medium
Blur, and RGB value being transfer to HSV value were added to the original images. Consequently, the dataset was amplified by a factor of 7. Data augmentation allows the increase in dataset for both fighting and non-fighting, which significantly enhanced the accuracy of the judgment of the CNN. The CNN was trained using the Interactive Deep Learning GPU Training System (DIGITS) on Ubuntu 16.4 and using a GPU NVIDIA GTX1080 as the hardware.

![Data Augmentation](image)

**Figure 5.** An example of data augmentation

### 3.4. The ensemble learning of combining picture-identifying system and video-identifying system

Due to the complexity of the possible application sites of this system, two systems are combined together for the purpose of getting a more reliable result. By using two different ways of testing, this system could minimize the influence caused by the background and at the same time minimize the possible error caused by misidentifying similar actions with fighting such as hugging. For each frame, two models calculate their result separately, and only if both of them provides a positive result, the final output is “fight”, otherwise it’s “no”.

However, in order to determine whether or not the video is violent, a threshold of frame number is needed so that when the number of consecutive frames that are identified as violent exceeds the threshold, the video is identified as “violent”. This threshold needs to be tested to maintain the highest accuracy. There are 5 thresholds that are selected to be tested: 6, 8, 10, 12 and 14. 18 videos are randomly selected to test the thresholds including 9 fighting videos and 9 non-fighting videos.

This new proposal provides a better system which cannot be easily influenced by the background of the video and also maintain a high accuracy.

![System Diagram](image)

**Figure 6.** An abstract of the system
4. Data Analysis

Figure 7. Loss curve for AlexNet.

Figure 8. Loss curve for GoogLeNet

Figure 9. Loss curve for ResNet

Table 1. Comparison of Accuracies for Different CNNs

| Method    | Accuracy   |
|-----------|------------|
| AlexNet   | 97.01%     |
| GoogLeNet | 98.17%     |
| ResNet    | 99.33%     |
All 3 possible networks for image-identifying model are trained and tested using DIGITS, on NVIDIA GTX 1080. The results show that ResNet has the best accuracy among all 3 of them. Furthermore, considering the loss curve for each network, it’s obvious that AlexNet has a comparatively slower speed of converging, while ResNet and GoogLeNet perform well. Therefore, based on the two reasons, ResNet is chosen to be the image-identifying network.

The accuracy of the proposed video-identifying system is 98.5%, which is high. However, the accuracy is not as high as the ResNet model, which proves that the video-identifying model is less accurate than the image-identifying model, but it is less likely to be influenced by background and can more easily adapt new circumstances.

Table 2. Comparison of Different Threshold for the Combined System

| Threshold (frames) | Accuracy  |
|-------------------|-----------|
| 6                 | 86.36%    |
| 8                 | 86.36%    |
| 10                | 90.90%    |
| 12                | 90.90%    |
| 14                | 81.81%    |

In order to detect violence action, a threshold is needed as mentioned in part D in the method. The result is that 10 frames is the best threshold due to the highest accuracy it achieves. Comparatively, due to the accuracy of 5 and 15 frames are both lower than the accuracy of 10 frames, the 10 frame-accuracy is the peak. Therefore, a threshold of 10 frames is selected and applied in the real system.

For a better visual experience, the system is visualized by writing programs using PyCharm software. The final output would be a video with a colored square at its up-right corner. The block would be green if no violence is detected in this frame, and red otherwise. The current frame number is displayed within that block. More importantly, after the threshold requirement is satisfied, a big red-colored text will appear at the center of the screen: “VIOLENCE”, which will attract attention and thus solve the problem quickly.

The final system is applied in different videos, and the result is that it reaches the highest accuracy of 90.90% in the test. The final system applies ResNet as the image-identifying system, and chooses a network combining Darknet19, optical flow, ResNet and LSTM. Besides from the two networks, the visualization is successful, as shown in Fig.10 and Fig.11, where the color of the up-right block represents the classification of this frame.

Besides from all of the above, the speed of this final network is tested as well by calculating the difference of time between when the system captures a frame and when the system output the result of this frame. After testing, the system could calculate with a speed of 75 FPS. This system could definitely
satisfy the requirement of processing data immediately, which also means a high practical value of this system.

5. Discussion and Conclusion
In this research, a brand new idea of the ensemble learning of combining two neural networks together aiming for a higher accuracy and reliability is proposed. This structure of the system allows a higher reliability as well as a higher accuracy in real circumstances without applying a more complex model which requires expensive hardware. The ResNet-50 is chosen as the image-identifying network after the testing of 3 networks. A network using Darknet19, ResNet, optical flow and LSTM is used as the video-identifying part of the whole system. The image-identifying model could precisely determine whether the image includes fighting scene, but it’s easily influenced by the background of the image. In contrast, the video-identifying system could minimize the influence of the background due to the application of optical flow, thus compensating the weakness of the image-identifying system. By combining the two image-identifying and video-identifying systems, they could minimize the influence of the background and at the same time ensure the high accuracy of identifying motion.

As shown in data analysis, all these two networks have a high accuracy on the validation group comparing to other networks. The video-identifying model uses Darknet19 to extract feature, and applies optical flow to eliminate the influence of the background, and finally using one of the most accurate CNN to get the final result. As for the image-identifying model, 3 popular networks are tested using the same dataset, and the best one is selected to use. Therefore, by combining them together, the system achieves accuracy and reliability.

Last but not least, all the networks are not too over-complicated, unlike some of the contemporarily proposed networks. This provides a low hardware requirement and ensures that this system could be used in many schools. Plus, although not requiring expensive hardware, the calculation speed of the system is fast, with a speed of 75 FPS (frames per second), which satisfies most of the videos photographed by cameras. Therefore, this system owns a high practical value, and could be applied to real life. In future, data could be collected from schools that use this system, and thus the system in schools will become specialized for that school’s environment, and performs a higher accuracy.

References
[1] U.S. Department of Education. Student Reports of Bullying: Results From the 2015 School Crime Supplement to the National Crime Victimization Survey, December 2016.
[2] https://ai.qq.com/product/terror.shtml
[3] ZHOU, Z., ZHU, M. and Yahya, K., 2017. Violence Behavior Detection Based on 3D-CNN. Computer Systems & Applications, 12, p.034.
[4] Serrano, I., Deniz, O., Espinosa-Aranda, J.L. and Bueno, G., 2018. Fight Recognition in video using Hough Forests and 2D Convolutional Neural Network. IEEE Transactions on Image Processing.
[5] Nievas, E.B., Suarez, O.D., Garcia, G.B. and Sukthankar, R., 2011, August. Violence detection in video using computer vision techniques. In International conference on Computer analysis of images and patterns (pp. 332-339). Springer, Berlin, Heidelberg.
[6] Soomro, K., Zamir, A.R. and Shah, M., 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.
[7] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T., 2011, November. HMDB: a large video database for human motion recognition. In Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 2556-2563). IEEE.
[8] Redmon, J. and Farhadi, A., 2017. YOLO9000: better, faster, stronger. arXiv preprint.
[9] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). Ieee.
[10] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8),
pp.1735-1780.

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

[12] Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning representations by back-propagating errors. nature, 323(6088), p.533.

[13] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[14] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).