APPLICATION OF AUTOMATIC MOSAIC FOR VIDEO BASED ON YOLO

[1] Shi-Hao Li (黎世豪), [2] Chiou-Shann Fuh (傅楸善)

[1] Department of Biomedical Electronics and Bioinformatics,
[2] Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan,
E-mail: r07945053@ntu.edu.tw fuh@csie.ntu.edu.tw

ABSTRACT

In this paper, we developed a method that automatically adds mosaics to areas of the video that are not suitable for viewing, such as blood, violence and smoking, using the YOLO v3 deep learning model to identify objects in the video, if the identification result contains bloody, violent or smoking areas, then the blur post-processing is carried out. The traditional method of adding mosaics is mostly through manpower, but it takes a lot of time and can cause psychological discomfort to workers. This application can be widely used by social media, and can be extended to applications such as automatic video screening.

Keywords: *Object recognition, mosaic, deep learning*.

1. INTRODUCTION

With the popularity of smartphones and mobile networks, transmission speed of information has been explosive growth, but negative information can be easily spread on the Internet, many parents will show children some children's videos, in order to please them, however, the video is often inserted by the bad guys into bloody or violent fragments that are not suitable for viewing, causing the child to receive the wrong information. This is a very serious matter, because children can't tell whether this is good or bad, and it is possible to let them develop in a bad direction. Nowadays, most of the methods of filtering videos are screened by the administrator of the website or by the public, but this method is very labor-intensive and time-consuming, and usually it is the child who has already seen it before being removed. So we need a system that automatically filters videos.

Many similar applications have been proposed, such as pornographic video identification and classification based on CNN network architecture, using AlexNet [1], GoogLeNet [2] and Ensemble-ConvNet [3], or first detect whether the image contains a lot of skin color [4], consider motion information [5], etc., to identify pornographic and non-pornographic videos, but no one has yet identified the video containing blood, violence or smoking.

Therefore, we proposed a method that automatically adds mosaics to areas that are not suitable for viewing in the video, using the YOLO v3 [6] deep learning model to detect whether the video contains violent, bloody or smoking areas, if there is, the blurring is performed on the selected area. In this way, children can be prevented from directly receiving too stimulating videos, on the other hand, they can be extended to applications such as automatic video screening.

The rest of paper is organized as follows: Section 2 introduces the methods we used to implement this system. Section 3 explains the experiment and results of the implementation. The conclusion is addressed in section 4.

2. METHODS

The algorithm we proposed is shown in Fig 1. First, use the YOLO v3-based object recognition algorithm to find the area containing violence, bloody or smoking in the image and generate the corresponding bounding box, then the area inside the bounding box is blurred.

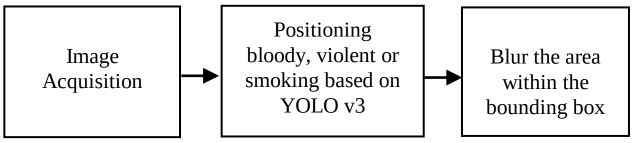


Image Acquisition
Positioning bloody, violent or smoking based on YOLO v3
Blur the area within the bounding box

Fig. 1: Algorithm flow chart.

2.1. Network architectures

There are many good object recognition models, such as R-CNN [7], Fast R-CNN [8], Faster RCNN [9], and SSD [10], etc., but there are a lot of videos on the Internet that need to be filtered, Therefore, we need to take the model with fast detection speed as a priority, and YOLO v3 has this feature, so this paper uses YOLO v3 as the model for object recognition.

Fig 2 shows the network architecture of YOLO v3. The base network is Darknet-53, which has 53 layers. It uses the ResNet structure commonly used in general neural network to solve the gradient problem. At the same time, using the FPN multi-level architecture, the feature layer has changed from single layer 13x13 to multiple layers 13x13, 26x26 and 52x52, from the single-layer prediction of five kinds of bounding boxes to three kinds of bounding boxes in each layer, so that the better target position of the lower layer and the better semantic features of the upper layer can be merged, and predictions can be independently performed at different feature layers, thereby improving the prediction ability of the small object.

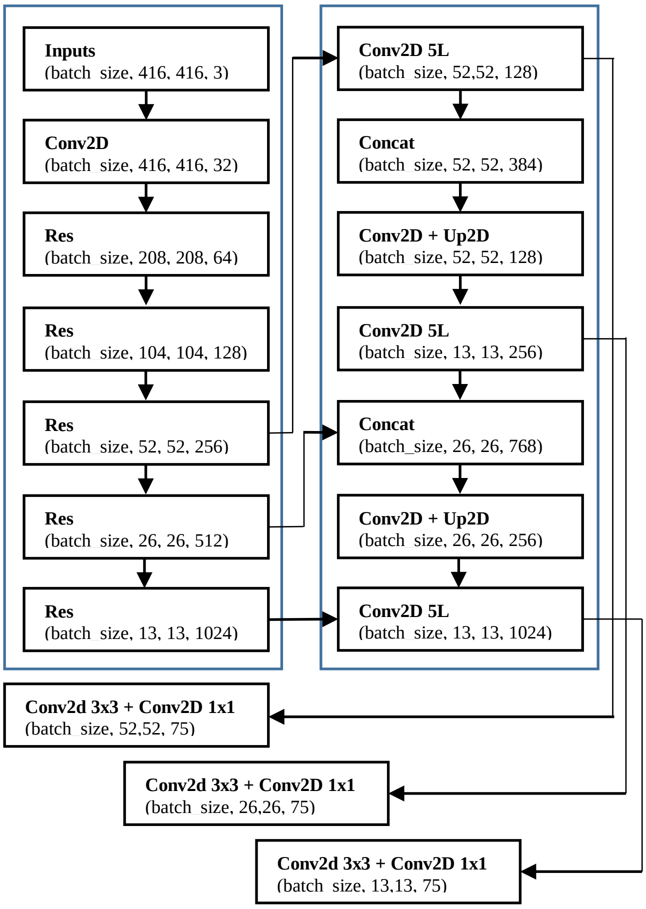

Inputs
(batch size, 416, 416, 3)
Conv2D
(batch size, 416, 416, 32)
Res
(batch size, 208, 208, 64)
Res
(batch size, 104, 104, 128)
Res
(batch size, 52, 52, 256)
Res
(batch size, 26, 26, 512)
Res
(batch size, 13, 13, 1024)
Conv2D 5L
(batch size, 52,52, 128)
Concat
(batch size, 52, 52, 384)
Conv2D + Up2D
(batch size, 52, 52, 128)
Conv2D 5L
(batch size, 13, 13, 256)
Concat
(batch_size, 26, 26, 768)
Conv2D + Up2D
(batch size, 26, 26, 256)
Conv2D 5L
(batch size, 13, 13, 1024)
Conv2d 3x3 + Conv2D 1x1
(batch size, 52,52, 75)
Conv2d 3x3 + Conv2D 1x1
(batch size, 26,26, 75)
Conv2d 3x3 + Conv2D 1x1
(batch size, 13,13, 75)

Fig. 2: YOLO v3 network architecture.

In addition, the category prediction is changed from softmax to logistic classifier, and the matching strategy of each bounding and ground truth becomes 1 to 1, in order to conform to the actual situation that the classification types are not mutually exclusive.

2.2. Dataset

The objects to be identified are bloody, violent and smoking, in order to increase the generality of the model, we have collected pictures in a variety of different situations. As shown in Fig 3 to Fig 5, a total of 1189 pictures were used as training data sets. There are 309 pictures of violence, 380 pictures of bloody, and 500 pictures of smoking as shown in Table 1.

Table 1: Training data set.

Name	Smoke	Blood	Violence	Total
Amount	500	380	309	1189

2.3. Training

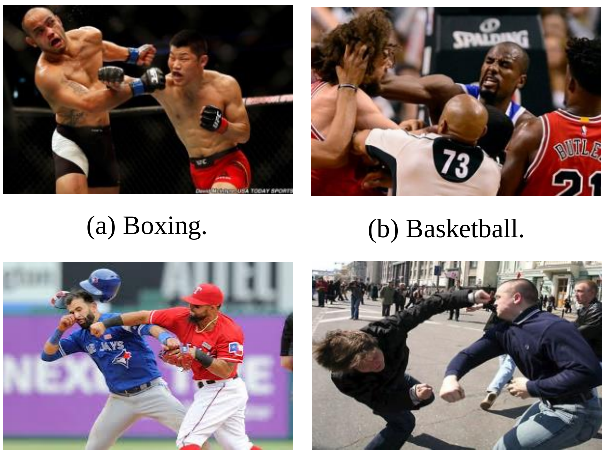
In the YOLO training procedure, the input image is first divided into S×S grid units. If the center of an object frame in the image falls within a certain grid unit, then the unit is responsible for detecting the object (responsible for detecting objects). At the same time, each grid unit simultaneously predicts the position of B bounding boxes and a confidence level. This confidence is not only the probability that the bounding box is the target to be detected, but the bounding box is the product of the probability of the target to be detected multiplied by the IoU of the bounding box and the real position. By multiplying this cross ratio, the accuracy of the predicted position of the bounding box is reflected, as shown in (1):

$$condifence = P(Object) \times IoU_{pred}^{truth} \qquad (1)$$

Each bounding box corresponds to 5 outputs, which are x, y, w, h and the confidence levels mentioned above. Where x, y represents the offset of the center of the bounding box from the boundary of the grid cell where it is located. w, h represents the ratio of the true width and height of the bounding box relative to the entire image. The parameters x, y, w, and h have been bounded to the interval [0, 1]. In addition, each grid cell also produces C conditional probabilities. Note that regardless of the size of B, each grid cell produces only one set of such probabilities. In the non-maximum suppression phase of test, for each bounding box, we should measure (2) whether the box should be retained.

$$confidence \times P(Class_i|Object) = P(Class_i) \times IoU_{pred}^{truth} \qquad (2)$$

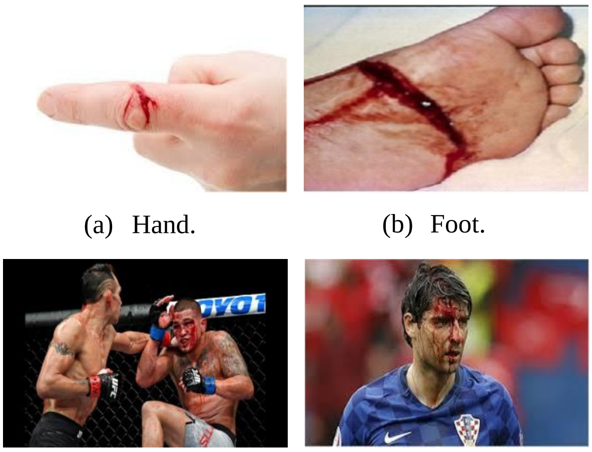
In order to speed up the convergence of the model, the convolution weights used by ImageNet are loaded at the beginning. The network architecture is basically the same as YOLO v3, with only minor adjustments: the last convolution layer has a filter size of 24, the object category to be identified is 3, and the batch size is 32. We use 90% of the data set as training, 10% as validation and the hardware used in training is the GPU provided by Google Colab.



(a) Boxing. (b) Basketball.

(c) Baseball. (d) Street.

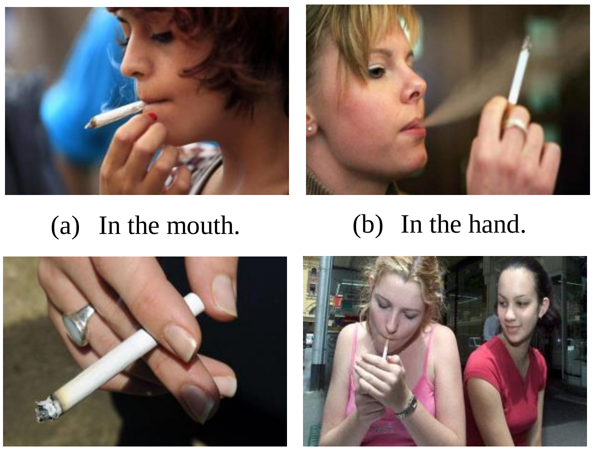
Fig. 3: Violent pictures in the training data set.



(a) Hand. (b) Foot.

(c) Boxing. (d) Football.

Fig. 4: Bloody pictures in the training data set.



(a) In the mouth. (b) In the hand.

(c) Hand close-up. (d) Complex background.

Fig. 5: Smoking pictures in the training data set.

2.4. Model evaluation

The YOLO model evaluation mainly adopts the common indicators of object detection: Precision (3), Recall (4), Accuracy (5) and Detection error rate (6), first define some evaluation indicators：

- TP(c): True Positive in class c, the predicted proposal matches ground true (the type is correct and the overlap is high enough).
- FP(c): False Positive in class c, the predicted proposal does not match the ground true (the type is wrong or the overlap is not high enough).
- TN(c): True Negative in class c, the predicted proposal matches ground true (the type is wrong and the overlap is high).
- FN(c): False Negative in class c, the predicted proposal does not match the ground true (the correct type or overlap is not high enough).

$$Precision = \frac{TP(c)}{TP(c)+FP(c)} \tag{3}$$

$$Recall = \frac{TP(c)}{TP(c)+FN(c)} \tag{4}$$

$$Accuracy = \frac{TP(c)+TN(c)}{TP(c)+FN(c)+TN(c)+FP(c)} \tag{5}$$

$$Detetction\ \ error\ \ rate = \frac{FP(c)+FN(c)}{TP(c)+FN(c)} \tag{6}$$

3. RESULTS

We collected a total of 180 images as test data sets, each of which has 60 images, which can be divided into 30 positive examples and 30 negative examples, as shown in Table 2, and the confusion matrix for each category is shown in Table 3. Precision, Recall, Accuracy, and Detection error rate are shown in Table 4.

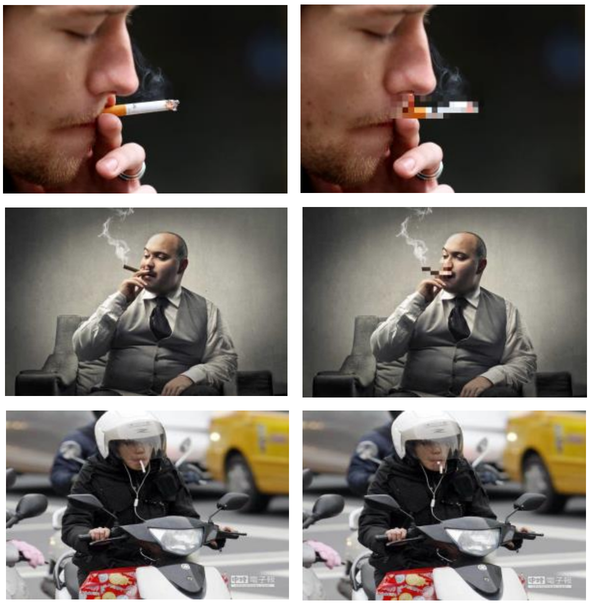
Table 2: Testing data set.

Name	**Smoke**	**Blood**	**Violence**	**Total**
Amount	60	60	60	180

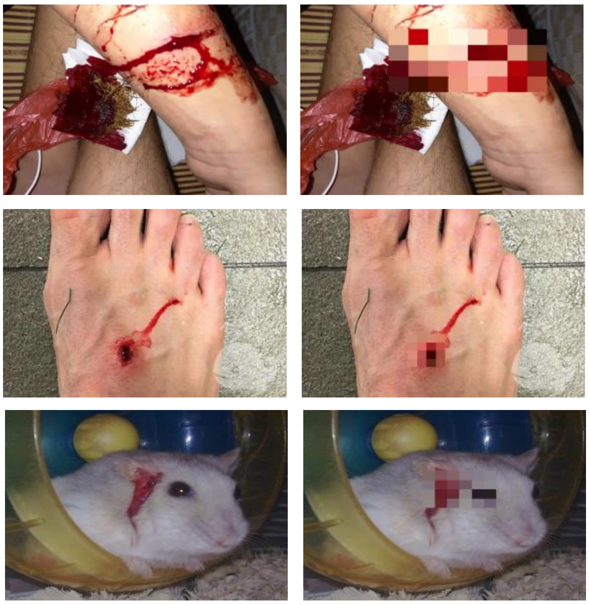
Figure 6 is an image of three positive examples from each of the categories in the test data and correctly identified. We have selected a number of images that are similar but not identical to the test data set to test the generality of the model, among the results of smoking, there are images of smoking cigars and smoking on a motorcycle. In the bloody results, there are images of animal injuries, even if there is no such type of images in the training data set.

Figure 7 shows three negative examples and no recognized images from each of the categories in the test data. In the result of smoking, use an image of a stick with a mouth, such as blowing bubbles, drinking a drink, and blowing arrows. In the bloody results, use images with red areas, such as tomato, lipstick, and watermelon. Finally, the result of violence, using images of physical interactions, such as sweating, backing and high-five.

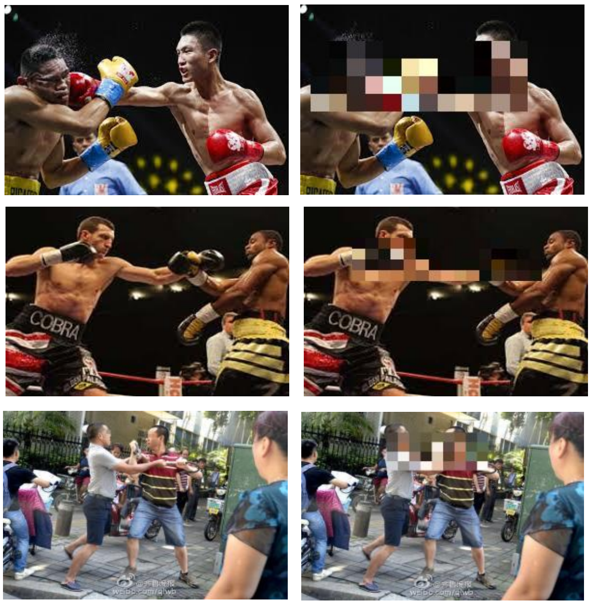
Figure 8 shows three positive examples from each of the categories in the test data but no recognized images. In the result of smoking, there are images in black or striped clothes, and images with very dark backgrounds.



(a) Smoking identification result.



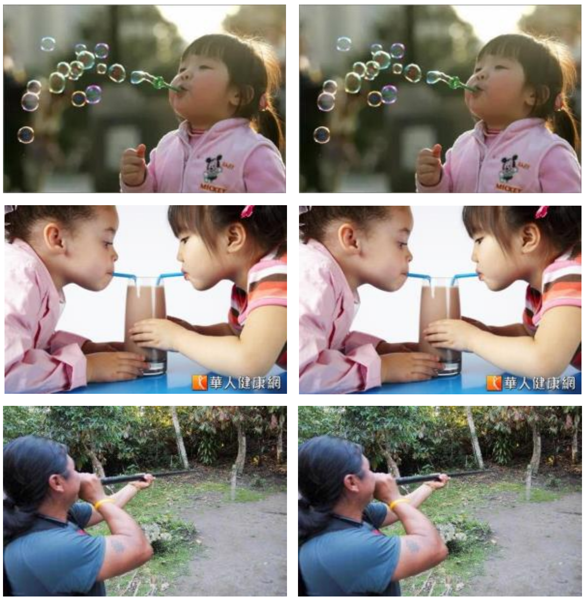
(b) Bloody identification result.



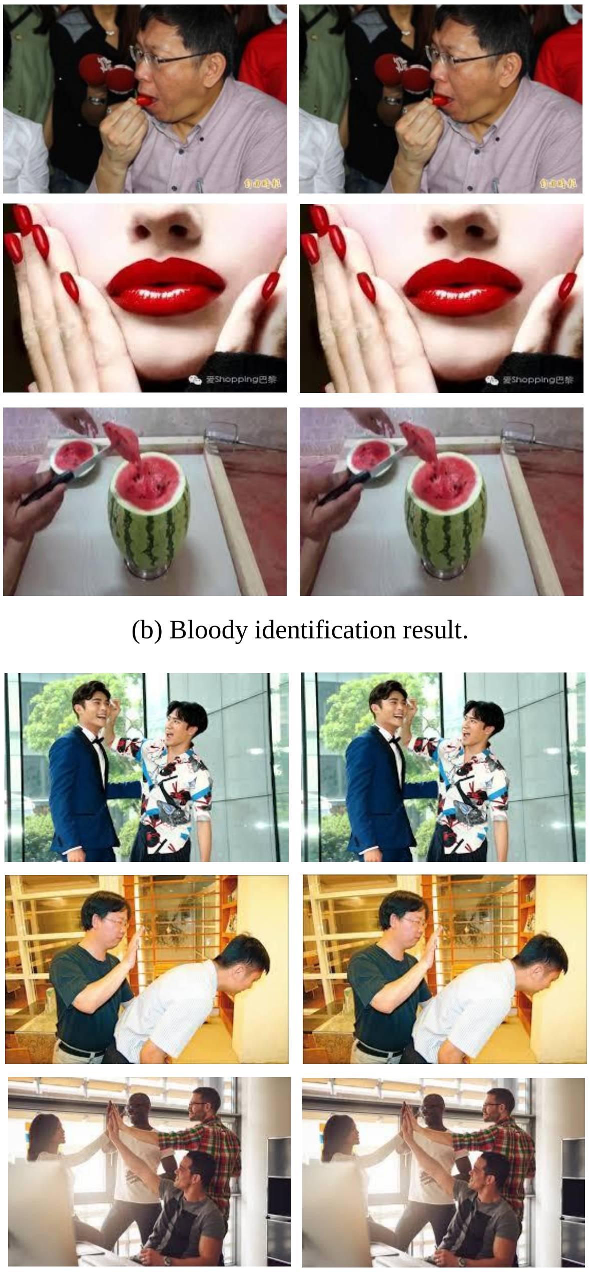
(c) Violence identification result.

Fig. 6: Positive examples in the test data and correctly identified: (Left-side) Original image, (Right-side) Identification result.

We think that the background interferes with the identification. In the bloody results, there are nosebleeds, large bloodshed and drops. Images of blood, which we believe are mostly caused by images of wounds in the training dataset. In the result of violence, there are many people or two people who have disputes, but because the fight contains dynamic information, it is difficult to define violence.



(a) Smoking identification result.



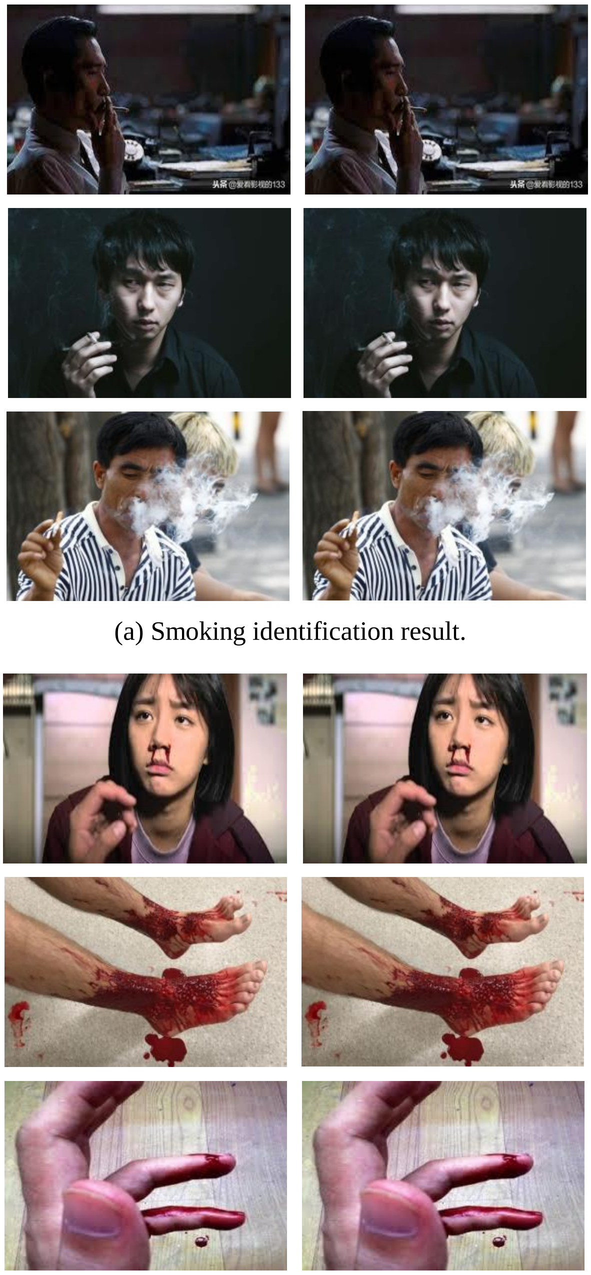
(b) Bloody identification result.

(c) Violence identification result.

Fig. 7: Negative examples in the test data but not identified: (Left-side) Original image, (Right-side) Identification result.

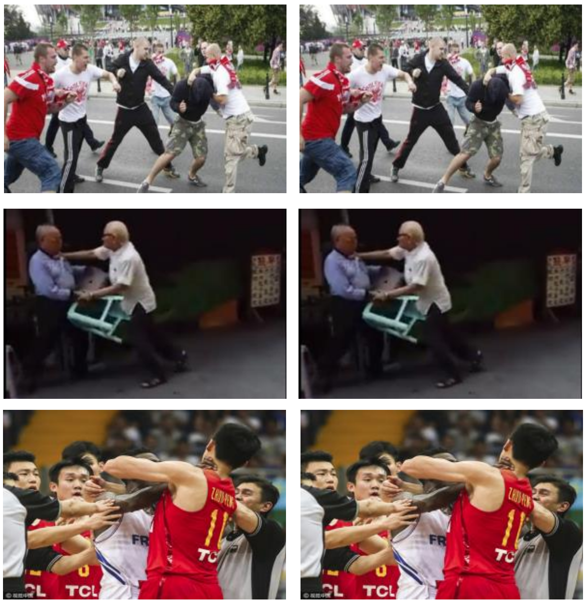
Figure 9 shows three negative examples of each category from the test data but the recognized images. In the result of smoking, there are images of students writing in chalk, drinking drinks, and eating lollipops, but because they hold white sticks with their hands or with their mouths, they are very close to smoking. In the bloody results, there are images of hands wearing red bracelets, eating red apples, and people being splashed by tomato juice. We think that the training data set is not enough.

In the result of the violence, there are images of two people shaking hands or dancing, and images of many people queuing under the bridge, but as mentioned in the previous paragraph, the fight contains dynamic information, and it is difficult to define violence.



(a) Smoking identification result.

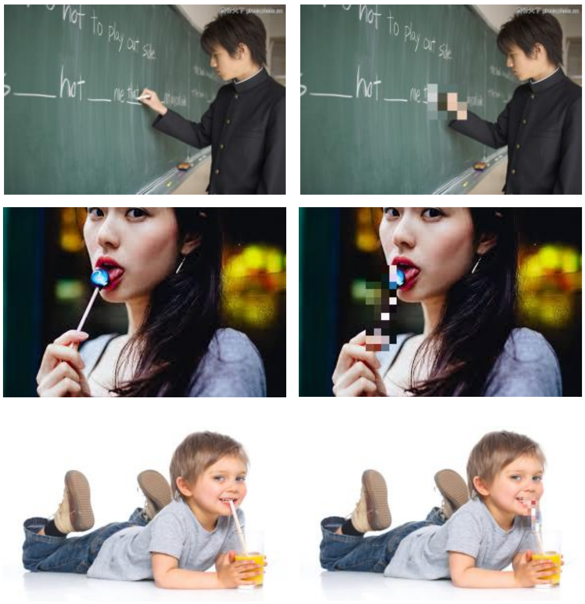
(b) Bloody identification result.



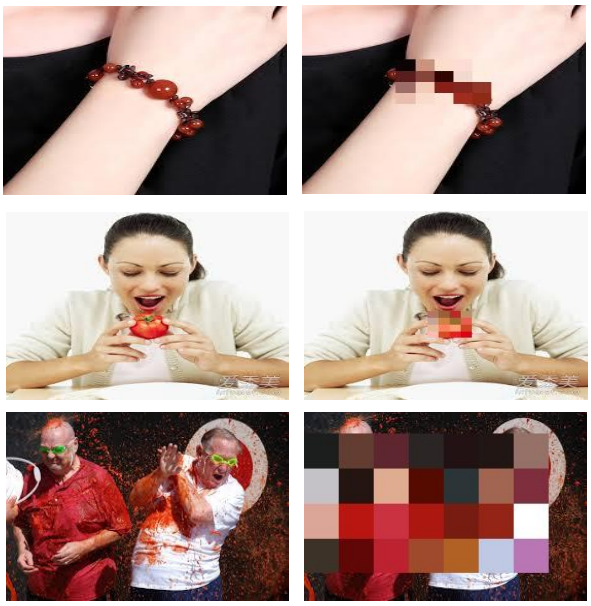
(c) Violence identification result.

Fig. 8: Positive examples in the test data but not identified: (Left-side) Original image, (Right-side) Identification result.

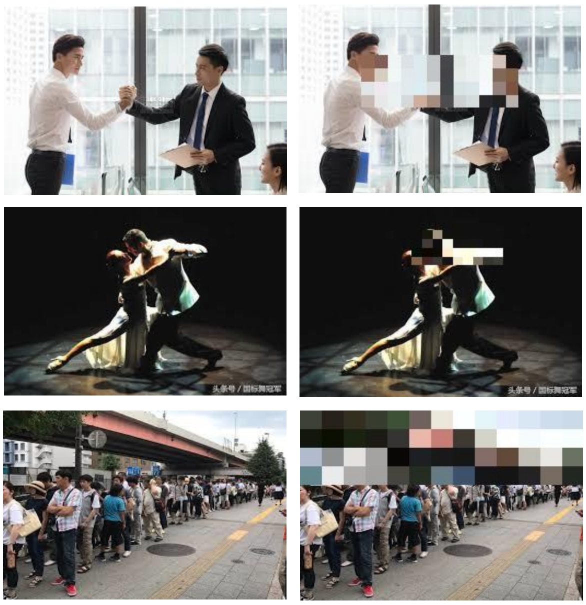
First of all, from the results obtained in Table 4, it can be known that the recall rate of the smoking category is about 20% higher than the precision, which means that the picture that is not actually smoked is easily recognized as smoking because the smoke is a white strip. It is easy to confuse with similar objects such as straws or lollipops, and the bloody category picture, the precision is about 20% higher than the recall rate, which means that the actual bloody pictures are not easily recognized as bloody, because the bleeding is recognized, which is equivalent to detecting whether there is a red area, but we don't want to see any red areas that are considered bloody.



(a) Smoking identification result.



(b) Bloody identification result.



(c) Violence identification result.

Fig. 9: Negative examples in the test data but identified: (Left-side) Original image, (Right-side) Identification result.

Finally, the category of violence, the recall rate is only 33%, which means that many pictures that are actually violent are not recognized as violence, and the error rate is also the highest, because it is actually difficult to identify pictures with dynamic components from static images. It is also difficult for us to define the point in time when violence occurs.

Table 3: Validation index of testing data set.

Class	TP	FP	TN	FN	Total
Smoke	25	13	17	5	60
Blood	17	4	26	13	60
Violence	10	3	27	20	60
Sum	52	20	70	38	180

Table 4: Measure results of testing data set.

Class	Precision	Recall	Accuracy	Error
Smoke	65.8%	83.3%	70.0%	60.0%
Blood	81.0%	56.7%	71.7%	56.7%
Violence	76.9%	33.3%	61.7%	76.7%
Avg.	74.6%	57.8%	67.8%	64.4%

Class	Specificity	Sensitivity
Smoke	43.3%	16.6%
Blood	13.3%	43.3%
Violence	10.0%	66.6%
Avg.	22.2%	42.2%

4. DISCUSSION

In identification of smoking categories, cigarette is easy to confuse with long sticks. If these long sticks are also added to the training data set, the recognition rate may be improved. In identification of bloody categories, blood is easily confused with the red area. It is necessary to convert the image into gray scale or other pre-processing before identification, which should effectively increase the generalization of the model, the identification of the violence category is easy to cause misjudgment or unconfirmed. Different from the former two categories are static object recognition, the difference between frame and frame is not high, at most the relative position offset, but the latter is a fast dynamic change, the frame-to-frame change is very large, and it is difficult for us to distinguish the difference between the physical interaction and the fight from the static information. For example, it is impossible to predict whether the two people want to shake hands or fight, a possible improvement is to use the optical flow method to track objects, or to consider facial expressions, effective training data amplification, and so on.

5. CONCLUSION

In this paper, we have developed methods that automatically add mosaics to areas of the video that are not suitable for viewing, such as blood, violence, and smoking. Object recognition is performed on the video by the YOLO V3 deep learning model. If the identification result contains bloody, violent, or smoking areas, then the blur post-processing is carried out. The result of this paper is only a prototype, and there are still many areas for improvement. If a certain recognition accuracy is achieved, this application can be widely used by social media, and can be extended to applications such as automatic video screening.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, G. Hinton, “Imagenet classification with deep convolutional neural networks,” In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc, 2012

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, “Going deeper with convolutions,” CoRR, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1-9.

[3] M. Moustafa. “Applying deep learning to classify pornographic images and videos,” arXiv preprint arXiv:1511.08899, 2015

[4] K. Zhou, L. Zhuo, Z. Geng, J. Zhang and X. G. Li, "Convolutional Neural Networks Based Pornographic Image Classification," 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), Taipei, 2016, pp. 206-209.

[5] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha, “Video pornography detection through deep learning techniques and motion information, Neurocomputing,” Volume 230, 2017, Pages 279-293, ISSN 0925-2312.

[6] J. Redmon, A. Farhadi “YOLOv3: An Incremental Improvement,” Computer Vision and Pattern Recognition,2018, arXiv:1804.02767

[7] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 580-587.

[8] R. Girshick, ”Fast r-cnn,” [C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448. fig

[9] S. Ren, K. He, R. Girshick, J. Sun, ”Faster R-CNN: Towards real-time object detection with region proposal networks,” [C]//Advances in neural information processing systems. 2015: 91-99.0

[10] L. Wei, A. Dragomir: SSD: Single Shot MultiBox Detector.arXiv preprint arXiv:1512.02325v5,2016