Campus Bullying Detecting Algorithm Based on Surveillance Video

Liang Ye[1,2,3(✉)], Susu Yan[1,3,4], Tian Han[2,4], Tapio Seppänen[2], and Esko Alasaarela[2]

[1] Harbin Institute of Technology, Harbin 150080, China
yeliang@hit.edu.cn
[2] University of Oulu, 90014 Oulu, Finland
[3] Science and Technology on Communication Networks Laboratory, Shijiazhuang, China
[4] Harbin University of Science and Technology, Harbin 150080, China

Abstract. In recent years, more and more violent events are taking place in campus life. Campus bullying prevention is already the focus of current education. This paper proposes a campus bullying detecting algorithm based on surveillance video. It can actively monitor whether students are being bullied on campus. The authors use Openpose to extract bone information from video. According to the coordinate information of bone points, they extract static and dynamic features. Support vector machine (SVM) is used to classify different actions. The recognition accuracy of the classification model is 88.57%. In this way, the campus surveillance camera is able to realize real-time monitoring of bullying behavior. It is conducive to the construction of a harmonious campus environment.

Keywords: Campus bullying · bone points · Openpose · Support vector machine

1 Introduction

Campus bullying has become a common social phenomenon [1]. It has caused great harm to the society and education in many countries. In many school bullying cases, the bullied hide the fact that they were bullied because of fear. They do not inform their parents or teachers in time, which leads to more and more bullying. As a result, bullied students have a serious psychological trauma. The research of deep learning and computer vision are more and more mature. However, most of the application scenarios of computer vision are limited to industry, transportation and commerce.

Through a literature survey [2, 3] in the field of computer vision, it is known that there are relatively few researchers who are researching on the application of campus scenes. The Kinect device [4] developed by Microsoft can capture the dynamic posture of two-dimensional human body. It leads to a lot of research work on gesture recognition techniques based on Kinect. However, the system is highly dependent on the Kinect device. It increases the hardware cost of the system. This research designs a bullying detecting method based on Openpose. It uses ordinary cameras to collect the

S. Shi et al. (Eds.): AICON 2020, LNICST 356, pp. 309–315, 2021.
https://doi.org/10.1007/978-3-030-69066-3_27

campus surveillance video of students, and uses Openpose to obtain the bone points of the human bodies from the video. It can remove redundant information in the image. Only the bone information of human bodies is retained for later action recognition. Thus, the information of an image is significantly reduced, and it is convenient for data transmission. The collected video information is transmitted to the background for processing. Distances and angles between bone points are used to extract motion features. Support vector machine is used to classify actions. The following sections will describe the algorithm in details.

2 Campus Bullying Detecting Algorithm

2.1 Bone Points of Human

In this study, bone points are used to identify human actions. It overcomes the influence of external environmental factors such as light changes and clothing changes. Human bone points represent the positions of human bodies (heads, limbs and trunks) in two-dimensional coordinates. It not only shows the local shapes of human bodies, but also describes the topological information. Therefore, it is of great significance to use the bone information of human bodies to recognize human actions.

In fact, human action is mainly the relative movement of human bones around their own joints. The human body is made up of 206 bones, which can be divided into three parts, namely skull, trunk and limb bones. There are 29 skulls, 51 trunk bones and 126 limb bones. It is unnecessary to use all the bones of the human body for action reconstruction. Instead, the authors choose the simple bone points model. In this study, the coco data set [5] is used as the annotation model of bone points. Table1 gives the specific bone points and their corresponding labels.

Table 1. Bone points and labels of the coco model.

Label	0	1	2	3	4	5	6	7	8
Bone	Nose	Neck	Right shoulder	Right elbow	Right wrist	Left shoulder	Left elbow	Left wrist	Right hip
Label	9	10	11	12	13	14	15	16	17
Bone	Right knee	Right ankle	Left hip	Left knee	Left ankle	Left eye	Right eye	Left ear	Right ear

2.2 Bone Points Detection Based on Openpose

Openpose multi-person pose estimation model was proposed by researchers from Carnegie Mellon University [6]. The model uses a deep neural network to extract the original feature xtraction is an important step map of the image. The model input is divided into two branches. In one branch, a convolution neural network (CNN) is used to predict the heat map of human joint points. In the other branch, another CNN is used to obtain the partial affinity domain of all the connected joint points. The whole

network diagram is shown in Fig. 1. The authors use the coco human skeleton model to extract bone information from key frames in the video. Set the key points numbered 0, 1, 2, 3, 4, 5, 6, 7, 8, 15, 16, and 17 as the activation points in practical application (Fig. 2), and obtain the coordinates of bone points as raw data for post-processing.

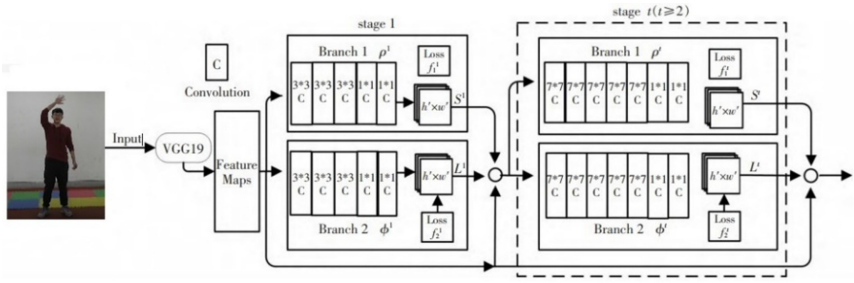
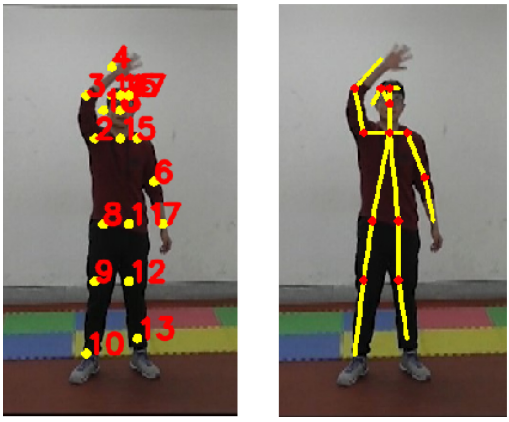


Fig. 1. Structure diagram of the whole network. The overall network architecture is divided into six stages. The upper branch is responsible for predicting the positions of bone points. The lower branch is responsible for predicting the affinity region between bone points. It can improve the accuracy of bone point prediction after multi-stage operation.



(a) (b)

Fig. 2. Output map of human bone points. The main function of Openpose is divided into two parts. (a) The first part is to identify the joints of the human body in the input image. (b) The second part is to connect the corresponding joint points belonging to each person.

2.3 Feature Extraction and Classification

Feature extraction is an important step in human action recognition. Firstly, the authors catalog the collected video data samples. Violent behaviors are regarded as bullying and marked as "positive samples". Daily-life behaviors are regarded as non-bullying actions and marked as "negative samples". There are totally 83 positive samples and 118 negative samples. Motion features can be expressed as the coordinate information of human joint points. In this research, both static and dynamic features of human bodies are extracted for bullying recognition.

Static features are divided into distance features and angle features. The distance feature is extracted by calculating the distance between two joint points. Because the picture is two-dimensional, the authors use the distance formula of two-dimensional space to calculate the distance between joint points as,

$$D_{i,j}^{s} = |p_i^s - p_j^s| \quad i \neq j \tag{1}$$

where p_i^s and p_j^s represent the coordinates of different joint points in the same action sequence frame, respectively, and $D_{i,j}^{s}$ is the Euclidean distance between joint points.

The authors use the coordinates of three points to calculate the angle. Calculate the angle by the cosine theorem,

$$\theta = \arccos\left(\frac{c^2 + a^2 - b^2}{2ac}\right) \tag{2}$$

where a, b, and c are the lengths of the three sides. Human actions differ in both time and space. By analyzing multiple consecutive images, the authors summarize the change rules of human body postures. They establish a more accurate topology for each joint point. The displacement vector of joint points on time series is an important feature, which is also called the dynamic feature. An action sequence can be represented by a series of continuous skeleton frames. The displacement vector sequence can be expressed as follows:

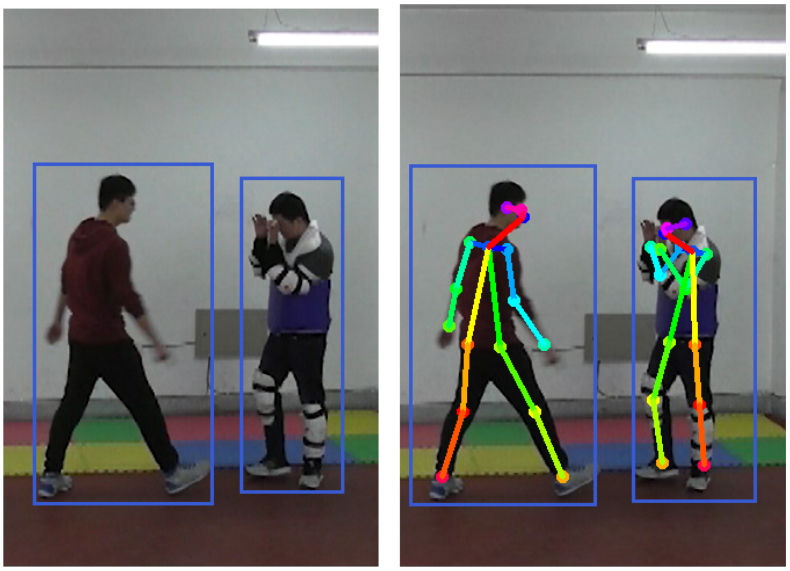
$$\omega_i^s = \left|\frac{p_i^{s+1} - p_i^s}{\Delta T}\right| \quad 1 < s < \tau \tag{3}$$

where p_i^s represents the coordinates of the joint points in the s frame of the action sequence, and ΔT is the time interval between the two frames.

Table 2 shows the extracted features. Campus bullying is commonly attended by multiple persons, and the interaction of these persons is relatively strong. Therefore, the authors use circumscribed rectangular frames to separate different persons in one image. Figure 3 shows the circumscribed rectangular frame target separation and bone analysis.

Table 2. Distance features, angle features, and dynamic features.

Features	Specific features	Quantity
Distance features	Hand - hand	1
	Foot - foot	1
	Hand - waist	2
	Hand - shoulder	2
	Hand - knee	2
	Knee - knee	1
	Elbow - knee	2
	Hand - foot	2
Angle features	Hand - elbow- shoulder	2
	Foot - knee- waist	2
	Elbow - shoulder-neck	2
	Knee - waist - neck	2
Dynamic features	Hands	2
	Feet	2
	Waists	2



(a) (b)

Fig. 3. Target separation and bone extraction. (a) Target separation with circumscribed rectangular frames. (b) Bone extraction.

As mentioned above, label the feature sequences extracted from campus bullying fragments as positive, and those from daily-life actions as negative. Thus, this study is a 2-class classification. The authors choose SVM [7] for classification. Based on the minimization loss function, it looks for a hyperplane to distinguish samples of different classes. The authors used five-fold cross validation to estimate the classification performance, and Table 3 shows the confusion matrix.

Table 3. Confusion matrix of campus bullying detection (%).

	Bullying (predicted)	Non-bullying (predicted)
Bullying (real)	**86.72**	13.28
Non-bullying (real)	9.59	**90.41**

According to Table 3, 86.72% of bullying actions were accurately identified as violent actions, and 90.41% of daily actions were recognized as non-violent actions.

Then the authors calculated the four indexes of accuracy, precision, recall, and F1-score. Table 4 shows the results.

Table 4. Four indexes of campus bullying recognition performance.

Indexes	Accuracy	Precision	Recall	F1-Score
Value (%)	88.57	90.04	86.72	88.35

Finally, the proposed campus bullying detection algorithm gets an average accuracy of 88.57%, which shows a promise for detecting campus bullying events with bone information.

3 Conclusions

Campus bullying is a common social phenomenon in many countries. To detect campus violence, this research proposes a campus bullying detecting method using bone information extracted from surveillance video images. Firstly, the authors use the Openpose model to extract the human bone points in the video. Then, they use the coordinate relationship of bone points to extract features. SVM is used to classify violent actions. Moreover, video data can facilitate managers to confirm the occurrence of violence. Finally, the proposed method gets an average accuracy of 88.57%, which shows a promise for detecting campus bullying events with surveillance cameras.

Acknowledgements. This paper was funded by the National Natural Science Foundation of China under grant number 41861134010, the Key Laboratory of Information Transmission and

Distribution Technology of Communication Network (HHX20641X002), National Key R&D Program of China (No. 2018YFC0807101).

References

1. Sung, Y.-H., Chen, L.-M.: Double trouble: the developmental process of school bully-victims. Child. Youth Serv. Rev. **91**(01), 279–288 (2018)
2. Hammami, S.M., Alhammami, M.: Vision-based system model for detecting violence against children. MethodsX **2**(4), 7–8 (2020)
3. Hao, M., Cao, W.H., Liu, Z.T.: Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. Neurocomputing **12**(07), 390–391 (2020)
4. Li, G., Li, C.: Learning skeleton information for human action analysis using Kinect. In: Signal Processing: Image Communication, vol. 115814 (2020)
5. Newell, A., Huang, Z., Deng, J.: Associative embedding: end-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems, pp. 2274–2284 (2017)
6. Cao, Z., Simon, T., Wei, S E.: Realtime multi-person 2D pose estimation using part affinity fields. In: IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society (2017)
7. Accattoli, S., Sernani, P., Falcionelli, N.: Violence detection in videos by combining 3D convolutional neural networks and support vector machines. Appl. Artif. Intell. **34**(4), 202–203 (2020)