Research Article

Motion Direction Inconsistency-Based Fight Detection for Multiview Surveillance Videos

Chuang Yao ¹, **Xiaoyan Su** ¹, **Xuehua Wang** ¹, **Xinyi Kang** ¹, **Jun Zhang** ²,
and **Jiankang Ren** ³

¹School of Economics and Management, Dalian University of Technology, Dalian 116023, China

²Graduate School of Education, Dalian University of Technology, Dalian 116023, China

³School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China

Correspondence should be addressed to Xiaoyan Su; sxy@dlut.edu.cn

Received 6 March 2021; Revised 12 April 2021; Accepted 19 April 2021; Published 17 May 2021

Academic Editor: Keping Yu

Copyright © 2021 Chuang Yao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nowadays, with the increasing number of surveillance cameras, human behavior detection is of importance for public security. Detection of fight behavior using video surveillance is an essential and challenging research field. We propose a multiview fight detection method based on statistical characteristics of the optical flow and random forest. Cyberphysical systems for monitoring can obtain timely and accurate information from this method. Two novel descriptors named Motion Direction Inconsistency (MoDI) and Weighted Motion Direction Inconsistency (WMoDI) are defined to improve the performance of existing methods for videos with different shooting views and solve the misjudgment on nonfight, such as running and talking. First, YOLO V3 algorithm is applied to mark the motion areas, and then, the optical flow is computed to extract descriptors. Finally, Random Forest is used for classification based on statistical characteristics of descriptors. The evaluation results on CASIA dataset demonstrate that the proposed method can improve the accuracy and reduce the rate of missing alarm and false alarm for the detection, and it is very robust against videos with different shooting views.

1. Introduction

With the increasing public demand for security and the development of machine vision technology, research in the field of intelligent monitoring continues to deepen, and many scholars are committed to use the powerful computing power of computers to process surveillance video data to strengthen security control. Nowadays, large-scale monitoring cyberphysical systems surround us and the massive videos obtained from these systems should be analyzed to guarantee public demand for security. Nevertheless, security systems relying on human observers are inefficient, and it may cause missed alarms due to the limited human capability to monitor surveillance video continuously, resulting in an urgent demand to a research of automatic alarm technology for the abnormal behaviors like fighting. And the result of proposed method will be a foundation for the cyberphysical systems to monitor and control of entities in the physical world [1].

As a key step to achieve behavior recognition, behavior analysis can explain and describe the motion of the target by using machine vision. Generally, the behavior analysis methods are categorized in two classes: the methods based on high-level human body structure and the methods based on low-level image information. The information used in high-level methods includes body point model, 2D model of human body [2], and 3D model of human body [3–5]. Moreover, in order to obtain more information, multiview videos or a 3D camera are required for this method. However, it makes the model complicated and computationally intensive. And it is difficult to guarantee the stability of this kind of method due to the immaturity of the human posture estimation technology. On the other hand, in the behavior analysis methods based on low-level image information, motion trajectory [6], shape features [7], texture features [8], optical flow features, and other image information are used to perform the behavior analysis. This kind of method

describes the target's behavior from a macroperspective, and it has the characteristics of simplicity and low complexity. Compared with the methods based on high-level human body structure, it performs better in real-time processing.

Fight as a typical abnormal behavior not only threatens the safety of citizens, but also has a negative influence on public. Therefore, it is crucial to achieve the real-time and reliable fight behavior recognition to assist security agents and ensure safety. The optical flow information extracted from the changes in pixels of the video sequences has satisfactory spatial and temporal characteristics, which is widely used to describe the target's motion tendency in video processing [9]. In recent years, the fighting detection based on the optical flow has been widely concerned. Some researchers detect fight by setting threshold on optical flow features. After that, scholars focus on employing machine learning algorithms for fight detection.

Based on the extracted optical flow information, some detection methods based on descriptors and threshold are proposed. Lin et al. [10] combined texture features and optional flow features to improve recognition rate in crowded scenes. The amplitude-based weighted direction histogram can suppress the direction confusion caused by noise in a small area and the entropy of this histogram can characterize the degree of chaos in the motion. However, it does not consider the changes of the optical flow energy when fight occurs [11]. Xu et al. [12] combined the weighted reconstruction trajectory and histogram entropy to detect fight behavior. The changes in the optical flow energy of the target object can be used to identify abnormal behavior. Nevertheless, it does not make a good distinction between fighting and running [13]. Lin et al. [7] used the motion features and shape features to detect fight, which can reduce the misjudgment rate of anomaly recognition effectively. However, it still has the problem of misjudgment in the conversation behavior. Based on the statistics of the amplitude and direction of optical flow, Huang et al. [14, 15] adopted mean and variance to recognize fight, providing a novel idea for definition of the features. Though the detection approach based on threshold method is simple and performs well in real-time, it performs not very well in videos with different shooting views and can hardly avoid misjudgment on behaviors like running and talking.

Besides, some scholars investigated fighting detection by using machine learning models for purpose of distinguishing fight and other behaviors effectively. Wang et al. [16] proposed a new feature, called Acceleration Violent Flow (AViF), and adopted support vector machine (SVM) and random forest to detect fight behavior. However, this method has a low recognition rate of the fight with insignificant movements such as punching etc. Yang et al. [17] proposed a real-time method based on the optical flow histograms, which calculates a scale descriptor and a rotation invariant feature descriptor histogram of optical flow orientation (HOFO) and uses the SVM classifier for detection. However, this method also has the problem of misjudgment in conversation. Considering the change of optical flow, Gao et al. [18] proposed the oriented violent flows (OVIF) method for violence detection, which adopts AdaBoost and SVM classifiers

for detection. However, this method performed poorly in crowded scenes. Recently, Mahmoodi et al. [19] proposed a new feature descriptor named Histogram of Optical flow Magnitude and Orientation (HOMO) to improve existing violence detection. Xu et al. [20] proposed a localization guided framework which exploits optical flow maps to extract motion activation information for detecting fight actions in surveillance videos. Febin et al. [21] proposed a cascaded method of violence detection based on motion boundary SIFT (MoBSIFT) and movement filtering for action recognition.

Approaches based on machine learning models show great performance in many scenes, but most of them do not consider the robustness in multiview which is essential for fight detection in real scenes. The existing algorithms present some limitations in solving misjudgment problems on running, overtaking, etc. which are similar to fighting in terms of motion characteristics sometimes. In terms of robustness on shooting view of videos and misjudgment on nonfight behaviors, we propose two descriptors, named Motion Direction Inconsistency (MoDI) and Weighted Motion Direction Inconsistency (WMoDI), by analyzing the motion characteristics of fight. In addition, a deep learning framework is used for motion region marking in this paper. We combine three existing descriptors and two novel descriptors, and the final feature with six statistical characteristics of these descriptors are calculated to improve the accuracy and robustness in multiview.

Experimental results on CASIA Action Dataset and UT-Interaction Dataset show that the proposed descriptors are discriminative enough to distinguish fight and nonfight behaviors in videos with different shooting views. In addition, the proposed method has certain performance advantages in accuracy and can recognize fight in different datasets effectively.

In summary, the main contributions of this paper are in the following aspects: (1) the robustness of fight behavior recognition from multiple video perspectives is considered in our proposed method. (2) The misjudgment rate of fight behavior detection can be reduced effectively by using MoDI and WMoDI. Part of this paper has appeared in [22]. As the extension of [22], we improve the method in motion regions marking and descriptors defining and redesign several substantial experiments to evaluate the performance of our method in different scenes.

The remainder of this paper is structured as follows. Section 2 discusses the fighting detection and describes the proposed method. Section 3 introduces the data sets used and provides experimental results. Finally, in Section 4, we summarize our paper and outline some promising future research directions.

2. Fighting Detection

As studies mentioned in Section 1, optical flow as a low-level image information can express the pattern of motion, and based on it, scholars proposed a series of descriptors for fight detection. Compared with nonfight behaviors, e.g., talking and walking, fight behaviors usually have the characteristics of suddenness, violence, and messy movements, which

FIGURE 1: Overview of the proposed detection method.

means it is difficult to define the fight behavior with standard actions [12]. Actually, accelerated and sudden movement is often associated with states of high activation (happiness, anger) [23] which confirms that movement such as punches, kicks, or blows in a frame can be identified through suitable features. But the challenge of this work is to distinguish fight and other behaviors with high speed and sudden movement such as running.

To solve this challenge, we analyze the consistency and intensity of different motions in surveillance videos with different shooting views by calculating the displacement of each corner point. The result shows two typical distinctions. Compared with corner points in nonfighting (talking, walking, and running) videos whose inclination degree of displacement is consistent, corner points in fighting videos show chaotic direction changes. In addition, the velocity amplitude of corner points of nonfight tends to be stable while that of fight is larger and shows irregular changes. Therefore, velocity, acceleration and orientation are considered as the important factors for identifying fight behavior.

In our method, instead of recognizing the body structure and the movement of limbs, we adopt optical flow information to generate descriptors and Random Forest algorithm to identify the fight behavior. Figure 1 shows the overview of the proposed method. Firstly, the state-of-art deep learning model Yolo-V3 is adopted to mark the motion regions of passengers. Then, optical flow vectors are calculated, and a series of descriptors including two novel descriptors are extracted. Next, we smooth the descriptors, compute several statistical characteristics of them, and generate the final feature vector for training by concatenating the statistical characteristics. Finally, Random Forest algorithm is used to classify the frames.

2.1. Motion Regions Marking and Optical Flow Extraction. With the purpose of reducing the impact of noise points, the first step performed in the method is motion region marking. Xu et al. [20] used optical flow to extract motion activation boxes from consecutive frames. Recently, Zhang et al. [24] proposed an improved algorithm based on ViBe (visual background extraction) algorithm [25] to achieve moving target edge detection. In this paper, we employ You Only Look Once (YOLO) V3 [26], which is a deep learning framework based on Darknet-53 [27] and detects objects at three different scales, to mark the minimum circumrectangles of passengers. Figure 2 shows the architecture of it.

After obtaining the motion regions, the optical flow magnitude and orientation are estimated by Lucas-Kanade (LK) algorithm [28] which detects any rapid change of key points at the pixel level. By solving Equation (1), we can get the optical flow vectors of frame I .

$$\begin{bmatrix} I_{x1} & I_{y1} \\ I_{x2} & I_{y2} \\ \vdots & \vdots \\ I_{xn} & I_{yn} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_{t1} \\ I_{t2} \\ \vdots \\ I_{tn} \end{bmatrix}, \quad (1)$$

where I_{xi} , I_{yi} , and I_{ti} are partial differentials with respect to the position (x_i, y_i) and time t evaluated at key point i . The result (u, v) is shown in Equation (2).

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum I_{xi}^2 & \sum I_{xi}I_{yi} \\ \sum I_{xi}I_{yi} & \sum I_{yi}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum I_{xi}I_{ti} \\ -\sum I_{yi}I_{ti} \end{bmatrix}. \quad (2)$$

FIGURE 2: YOLO-V3 network architecture for motion regions detection.

Then, (u, v) is converted to the polar coordinate form as in Equations (3) and (4).

$$\rho_i = \sqrt{(u_i)^2 + (v_i)^2}, \quad (3)$$

$$\theta_i = \arctan \frac{v_i}{u_i}, \quad (4)$$

where (ρ_i, θ_i) represents the velocity magnitude and the orientation of vector i .

2.2. Feature Extraction. Considering input sequences captured from different views in real scenes, we evaluate the robustness of descriptors in multiview. According to Deniz et al. [29], when a sudden movement with high acceleration occurs, velocity magnitude of the key points in motion areas is usually larger than normal behaviors such as walking and talking. Based on that finding, we utilize optical flow to calculate two proposed and three existing descriptors [10–13], which are discriminative enough for fight behavior identification.

Based on the differences between fight and nonfight mentioned above, two novel descriptors are defined to emphasize the change in orientation and magnitude of velocity and ensure the multiview robustness. Next, we give the extraction process of two proposed descriptors.

Firstly, the polar coordinates are divided into 12 intervals as shown in Figure 1, where each interval's size is 30° [30]. The frequency and magnitude of the vectors falling into left intervals and right intervals can be calculated as follows.

$$N_{\text{left}} = \sum_{i=1}^B \delta(a(i) - j)j = 4, 5, 6, 7, 8, 9, \quad (5)$$

$$N_{\text{right}} = \sum_{i=1}^B \delta(a(i) - j)j = 1, 2, 3, 10, 11, 12, \quad (6)$$

$$U_{\text{left}} = \sum_{i=1}^B \rho_i \delta(a(i) - j)j = 4, \dots, 9, \quad (7)$$

$$U_{\text{right}} = \sum_{i=1}^B \rho_i \delta(a(i) - j)j = 1, 2, 3, 10, 11, 12, \quad (8)$$

where B represents the number of key points in motion region R , $a(i)$ is the angle interval in which the vector i falls, j represents the angle interval, and δ is the Kronecker delta function.

Then, the definitions of the proposed descriptors Motion Direction Inconsistency (MoDI) and Weighted Motion Direction Inconsistency (WMoDI) are given.

Definition 1. Given the frequency N_{left} and N_{right} , the MoDI of current frame, denoted by C_M , is defined as

$$C_M = 1 - \frac{|N_{\text{left}} - N_{\text{right}}|}{N_{\text{primary}}}, \quad (9)$$

where N_{primary} is the larger one in N_{left} and N_{right} .

Definition 2. Given the magnitude U_{left} and U_{right} , the WMoDI is defined as

$$C_{WM} = 1 - \frac{|U_{\text{left}} - U_{\text{right}}|}{U_{\text{primary}}}, \quad (10)$$

where U_{primary} is the larger one in U_{left} and U_{right} .

In addition to MoDI and WMoDI, three existing descriptors that are proved to be effective in detecting fight behavior are adopted in the method. Here, we give the definitions of them.

Definition 3. Given the normalized frequency N_{nor_j} , the Optical Flow Direction Entropy is defined as

$$E_{\text{dir}} = - \sum_{i=1}^{12} N_{\text{nor}_j} \log_2 N_{\text{nor}_j}. \quad (11)$$

Definition 4. Given the normalized magnitude H_{nor_j} , the Weighted Optical Flow Direction Entropy is defined as

$$E_H = - \sum_{i=1}^{12} H_{\text{nor}_j} \log_2 H_{\text{nor}_j}. \quad (12)$$

Definition 5. Given the velocity magnitude ρ_i and the number of corner points B , the Average Kinetic Energy is defined as

$$E_K = \frac{\sum_{i=1}^B \rho_i^2}{B}. \quad (13)$$

Normalized frequencies N_{nor_j} and H_{nor_j} are calculated as follows.

$$\begin{aligned} N_{\text{nor}_j} &= \frac{\sum_{i=1}^B \delta(a(i) - j)}{\sum_{j=1}^{12} \sum_{i=1}^B \delta(a(i) - j)}, \\ H_{\text{nor}_j} &= \frac{\sum_{i=1}^B \rho_i \delta(a(i) - j)}{\sum_{j=1}^{12} \sum_{i=1}^B \rho_i \delta(a(i) - j)}. \end{aligned} \quad (14)$$

E_{dir} is calculated based on the optical flow direction histogram, and the higher the value, the more chaotic the motion in current frame. Besides, E_H based on magnitude and orientation of optical flow vector considers the effect of amplitude energy. E_K considers the intensity of action in different behaviors.

Finally, the above descriptors are smoothed according to the continuity of video stream to enhance the regularity and reduce the impact of noise. The FPS of the input video is 25, and we accumulate the descriptors of 10 consecutive frames as the accumulated descriptors of the n th frame (as shown in Equation (15)).

$$\left\{ \begin{aligned} IC_M(n) &= \sum_{n-9}^n C_M, \\ IE_H(n) &= \sum_{n-4}^n E_H, \\ IE_{\text{dir}}(n) &= \sum_{n-9}^n E_{\text{dir}}, \quad IE_{\text{dir}}(n) = \sum_{n-4}^n E_{\text{dir}}, \\ IE_H(n) &= \sum_{n-9}^n E_H, \\ IE_K(n) &= \sum_{n-9}^n E_K. \end{aligned} \right. \quad (15)$$

TABLE 1: Descriptors and statistical characteristics for final feature.

Descriptors	Statistical characteristics
MoDI (proposed in this paper)	
WMoDI (proposed in this paper)	
Optical flow direction entropy	Mean, variance, median, upper quartile, lower quartile, mean absolute deviation
Weighted optical flow direction entropy	
Average kinetic energy	

The last step of feature extraction is to calculate statistical characteristics of accumulated descriptors. Statistical characteristics are commonly used in data analysis. Inspired by the reference [15], we combine the statistical characteristics of accumulated descriptors to generate the final feature vector, and it is fed into the classifier for training. Table 1 gives the descriptors and statistical characteristics used in our method. And Algorithm 1 explains the feature extraction.

2.3. Classification Model. In our method, Random Forest (RF) algorithm is adopted as the classifier. The algorithm was first proposed by Breiman et al. in 2001 [31]. In this paper, bootstrap samples and Classification and Regression Tree (CART) algorithm are utilized to build a decision tree, which deploys the Gini Index, also named Gini impurity, to measure the quality of a split.

Next, we discuss the values of some hyperparameters of classifier in our method. Note that the classifier is trained in CASIA action database, which has been widely used in the research of behavior analysis [10–12, 32, 33]. This data set contains 432 videos, including 15 types of behavior, captured by 3 still cameras in angle, horizontal, and top-down view. And 105 of them are interactive videos, and the remaining 327 are single-person videos.

According to existing work [34], the accuracy of the classifier is mainly affected by the number of trees and the maximum depth of RF. Therefore, the hyperparameters are adjusted from both two aspects. And the accuracy of the classifier is evaluated by using 10-fold cross validation.

2.3.1. The Number of Trees. It is the number of decision trees contained in RF, and it has a great influence on the accuracy. By setting other hyperparameters of the model as fixed, we evaluate the accuracy of the model with different number of trees. Figure 3 shows the accuracy with respect to the number of trees. The x -axis represents the number of trees starting from 10 to 200, whereas the y -axis represents accuracy result for that specific parameter. The result shows that the accuracy improves with the increase of the number of trees, and then, it is gradually stable when the value of x -axis is more than 80. Moreover, the accuracy value reaches the maximum value when the value of x -axis is 110. To this end, the number of trees is set to 110 in the experiments of Section 3.

2.3.2. The Maximum Depth of RF. This parameter is adjusted under the condition that the number of trees is 110. Figure 4 shows the accuracy with respect to the maximum depth of RF, where the x -axis represents the maximum depth of the

Inputs: the video $S(x, y)$
Outputs: the final feature vectors

```

1: for each frame of  $S(x, y)$  do
2:   Compute optical flow vectors  $(u, v)$  by Equation (2)
3:   Convert  $(u, v)$  to the polar coordinate form by Equations (3) and (4)
4:   Compute  $N_{\text{left}}, N_{\text{right}}, U_{\text{left}}, U_{\text{right}}$  by Equations (5), (6), (7), and (8)
5:   if  $N_{\text{left}} > N_{\text{right}}$  then
6:      $N_{\text{primary}} = N_{\text{left}}$ 
7:   else
8:      $N_{\text{primary}} = N_{\text{right}}$ 
9:   if  $U_{\text{left}} > U_{\text{right}}$  then
10:     $U_{\text{primary}} = U_{\text{left}}$ 
11:   else
12:     $U_{\text{primary}} = U_{\text{right}}$ 
13:   Compute MoDI  $C_M$  by Equation (9)
14:   Compute WMoDI  $C_{WM}$  by Equation (10)
15:   Compute descriptors  $E_{\text{dir}}, E_H$ , and  $E_K$  by Equations (11), (12), and (13)
16:   Obtain smoothed descriptors by (15)
17: end for
18: Calculate the statistical characteristics of smoothed descriptors
19: Concatenate the statistical characteristics to generate the final vectors

```

ALGORITHM 1: Feature Extraction algorithm.

FIGURE 3: Accuracy with respect to the number of trees.

decision tree starting from 2 to 30, and the y -axis represents accuracy result for that specific parameter. The result shows that the accuracy increases with the increase of the maximum depth when the maximum depth is under 16. After the parameter reaches 16, the speed of accuracy converging slows down. And the accuracy reaches the maximum value when the parameter is 26. Therefore, 26 is selected as the maximum depth of RF in experiments of Section 3. Table 2 shows the hyperparameters of the RF.

3. Experiments

In this section, the data sets used in our experiments are introduced firstly. Then, in the experiments, we evaluate the

performance of RF classifier. And the robustness of the proposed descriptors is verified by data visualization and ROC curves. Finally, the advantage of the proposed method is shown by comparing with some existing methods.

3.1. Datasets. CASIA Action Dataset and UT-Interaction Dataset [35] are used to evaluate the effectiveness and robustness of our method. From CASIA Action Dataset, all videos of fighting and 15 other videos in 5 categories are selected, and the details are shown in Table 3. With these videos, we verify the performance of the proposed method in different shooting views. Sample frames of fighting in three views of the dataset are shown in Figure 5.

In addition, UT-Interaction Dataset is selected as another dataset, which is taken by an outdoor still camera and contains 6 types of interactive videos. The videos of punching and kicking are used as samples. Figure 6 gives some sample frames.

3.2. Experimental Settings. We evaluate the performance of our classifier in terms of accuracy, missing alarm (MA), false alarm (FA), and F1-score and compare the RF classifier with other representative classifiers including support vector machine (SVM), adaptive enhancement (AdaBoost) and bagging guided aggregation (Bagging). The evaluation parameters are computed based four measures, named True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Their definitions are given in Table 4.

The experimental results are shown in Table 5 and Figure 7. Note that the confusion matrices for each classifier are obtained from 300 frames sampled randomly from CASIA dataset. We can find that the RF outperforms other classifiers in terms of accuracy. In addition, the MA and FA of RF are suitable for fight detection in real scenes. Therefore, we adopt RF as classifier in our method.

FIGURE 4: Accuracy with respect to the maximum depth of RF.

TABLE 2: Hyperparameters of RF.

Parameters	Number of trees	Maximum depth of RF
Value	110	26

TABLE 3: Detail information of CASIA Action Dataset.

Behavior categories	Number of videos	Number of frames
Fighting	$3 \times 4 = 12$	4580
Running	$3 \times 1 = 3$	160
Follow together	$3 \times 1 = 3$	757
Meet apart	$3 \times 1 = 3$	738
Meet together	$3 \times 1 = 3$	992
Overtaking	$3 \times 1 = 3$	694

FIGURE 5: Sample frames in three view of CASIA Action Dataset.

FIGURE 6: Sample “punching” and “kicking” frames of UT-Interaction Dataset.

TABLE 4: Evaluation parameters.

Parameters	Definition
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN} \times 100\%$
Missing alarm (MA)	$\frac{FN}{TP + FN} \times 100\%$
False alarm (FA)	$\frac{FP}{TN + FP} \times 100\%$
F1-score	$2 \times (P \times R / (P + R))$, where $P = TP / (TP + FP)$, $R = TP / (TP + FN)$

TABLE 5: Comparison of classifier performance.

Classifiers	Accuracy	F1-score	MA	FA
SVM	92.00%	92.36%	11.58%	3.67%
AdaBoost	97.00%	96.95%	4.67%	1.33%
Bagging	92.67%	92.14%	1.40%	0.60%
RF	97.66%	97.66%	2.67%	2.00%

FIGURE 7: Confusion matrices for (a) RF, (b) SVM, (c) AdaBoost, and (d) bagging classifier (Label 1: fight (positive), label 0: nonfight (negative)).

FIGURE 8: Proposed descriptors of behaviors in three views. (a) C_M of different behaviors in angle view. (b) C_{WM} of different behaviors in angle view. (c) C_M of different behaviors in horizontal view. (d) C_{WM} of different behaviors in horizontal view. (e) C_M of different behaviors in top down view. (f) C_{WM} of different behaviors in top-down view.

FIGURE 9: ROC curves of different descriptors. SCED: statistical characteristics of existing descriptors; SCAD: statistical characteristics of all descriptors; PD: proposed descriptors; ED: existing descriptors.

TABLE 6: Accuracy of proposed model and existing methods.

Method	Accuracy
Xia et al. [37]	86.4%
Wang et al. [38]	87.9%
MI_ULBP [39]	92.7%
Xu et al. [12]	92.79%
Our method	96.77%

TABLE 7: Detection results with UT-Interaction Dataset.

Action	Punch	Kick
Number of videos	10	10
Number of videos alarmed	9	9

Furthermore, 18 videos of six behaviors shot from 3 different views in CASIA data set are selected to analyze the robustness of the proposed descriptors. Figure 8 gives the quantitative results of descriptors of six behaviors over time. The results show that the descriptors of fight are certainly higher than that of others in most frames. To guarantee that the classifier can recognize fight effectively when the descriptors are not highest and improve the accuracy in frame level, we compute the statistical characteristics of descriptors, which represent the global changes of descriptors.

In order to evaluate the effectiveness of our proposed descriptors and the statistical characteristics used in our method, the Receiver Operating Characteristic (ROC) curves for each descriptor are illustrated in Figure 9. For the model with statistical characteristics of proposed and existing descriptors [10–13, 36], its Area Under Curve (AUC) is 0.976 that outperforms with other descriptors. The results show that statistical characteristics of descriptors are more

discriminative than descriptors and can facilitate classifier distinguish fight and nonfight.

We evaluate the performance of our proposed model against the following state-of-art approaches:

- (i) The method proposed by Xia et al. [37]: it is a multiview interactive based on self-similarity descriptors and graph shared multitask learning
- (ii) The method proposed by Wang and Gao [38]: it is a deep learning based abnormal behavior recognition approach by utilizing interframe information and two-stream convolution neural network
- (iii) MI_ULBP [39]: it combines moment invariants and uniform LBP feature and trains a binary support vector machine pattern classifier for the purpose of recognizing behaviors
- (iv) The method proposed by Xu et al. [12]: it is a human abnormal behavior detection algorithm by combining the weighted reconstruction trajectory and histogram entropy

The accuracy of proposed method and existing methods is shown in Table 6. Our method gives the best result for the CASIA Action Dataset.

In addition, UT-Interaction Dataset is used to verify the applicability. Totally 20 videos of human-human interactions are selected, and half of them contain kicking occurring, and the rest contain punching occurring. Note that each video frames range from 80 to 120, and we suppose that there is fight occurring in the video when the frames predicted as positive are more than thirty percent of total frames. Table 7 shows the results, and it indicates the method proposed in this paper is also applicable to UT-Interaction Dataset and can effectively identify the fight behavior in videos.

In summary, the experimental results above show that the proposed method can effectively recognize fight behavior in videos with different shooting views. The reason mainly includes three aspects. Firstly, the feature vectors proposed in this paper have a high degree of distinction and multiview universality. Secondly, the smoothing process of extracted descriptors reduces the impact of noise and improves the robustness. Thirdly, as an efficient machine learning classifier, Random Forest shows the superiority for this task.

4. Conclusions

In this paper, in order to make the cyberphysical systems perceive the required environment information more effectively, we proposed an effective method for fight behavior detection which can reduce the misjudgment rate on nonfight behaviors and improve the robustness of detection with videos in different shooting views. Based on the analysis of motion characteristic, two novel descriptors MoDI and WMoDI are presented which are extracted from optical flow information. Instead of feeding the descriptors into classifier directly, 5 statistical characteristics are selected to generate the final vector. Experimental results demonstrate that the proposed

detection method performs well in videos with different shooting views, and it has the highest accuracy in CASIA dataset. As part of our future works, we will pay close attention to improve the early-warning capability and extend our method to poor light and even dark scenes.

Data Availability

You can get CASIA Action Dataset in website: <http://www.cbsr.ia.ac.cn/english/action%20databases%20en.asp>. And you can get UT-Interaction Dataset in website: https://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2018YFC0807500), National Natural Science Foundation of China (No. 62072067), National Natural Science Foundation of China (No. 71904022), Social Science Foundation of Liaoning Province (No. L17CTQ002), and Fundamental Research Funds for the Central Universities (No. DUT21JC27).

References

- [1] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: towards survivable cyber-physical systems," in *2008 The 28th International Conference on Distributed Computing Systems Workshops*, pp. 495–500, Beijing, China, June 2008.
- [2] S. Ghazal, U. S. Khan, M. M. Saleem, N. Rashid, and J. Iqbal, "Human activity recognition using 2D skeleton data and supervised machine learning," *IET Image Processing*, vol. 13, no. 13, pp. 2572–2578, 2019.
- [3] W. Ding, K. Liu, E. Belyaev, and F. Cheng, "Tensor-based linear dynamical systems for action recognition from 3D skeletons," *Pattern Recognition*, vol. 77, pp. 75–86, 2018.
- [4] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou, "Fast and robust multi-person 3D pose estimation from multiple views," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 921–929, Long Beach, CA, USA, June 2019.
- [5] X. Wang, L. T. Yang, L. Song, H. Wang, L. Ren, and M. J. Deen, "A tensor-based multi-attributes visual feature recognition method for industrial intelligence," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2231–2241, 2021.
- [6] M. L. T. L. Sandifort, J. Liu, S. Nishimura, and W. Hürst, "An entropy model for loiterer retrieval across multiple surveillance cameras," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pp. 309–317, Yokohama, Japan, June 2018.
- [7] G. Lin, Y. Bai, and W. Zhang, "Recognition of abnormal interactions based on coupled hidden Markov models," *Journal of Southeast University (Natural Science Edition)*, vol. 43, no. 6, pp. 1217–1221, 2013.
- [8] A. F. Villán, "Facial attributes recognition using computer vision to detect drowsiness and distraction in drivers," *Electronic Letters on Computer Vision and Image Analysis*, vol. 16, no. 2, p. 25, 2018.
- [9] H. Chen, S. Ye, O. V. Nedzvedz, and S. V. Ablameyko, "Application of integral optical flow for determining crowd movement from video images obtained using video surveillance systems," *Journal of Applied Spectroscopy*, vol. 85, no. 1, pp. 126–133, 2018.
- [10] Q. Lin and L. Zhang, "Crowded abnormal detection based on GLCM and optical flow," *Computer & Modernization*, vol. 24, no. 3, pp. 114–118, 2014.
- [11] J. Du and L. Xu, "Abnormal behavior detection based on regional optical flow," *Journal of Zhe Jiang University (Engineering Science)*, vol. 45, no. 7, pp. 1161–1166, 2011.
- [12] Z. Xu, Y. Luo, and P. Liu, "Abnormal behavior detection of joint weighted reconstruction trajectory and histogram entropy," *CAAI Transactions on Intelligent Systems*, vol. 13, no. 6, pp. 1015–1026, 2018.
- [13] C. Luo, X. Li, and H. Li, "Research on human abnormal behavior detection based on optical flow energy," *Computer and Network*, vol. 40, no. 21, pp. 71–73, 2014.
- [14] J. Ren, C. Lin, Q. Liu, M. S. Obaidat, G. Wu, and G. Tan, "Broadcast tree construction framework in tactile internet via dynamic algorithm," *Journal of Systems & Software*, vol. 136, pp. 59–73, 2018.
- [15] J. Huang and S. Chen, "Detection of violent crowd behavior based on statistical characteristics of the optical flow," in *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 565–569, Xiamen, China, August 2014.
- [16] C. Wang and S. Fei, "Violence detection in crowded scene," *Industrial Control Computer*, vol. 33, no. 2, pp. 105–106, 2020.
- [17] Z. Yang, T. Zhang, J. Yang, Q. Wu, and L. Yao, "Violence detection based on histogram of optical flow orientation," in *International Conference on Machine Vision*, p. 906718, London, UK, 2013.
- [18] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image & Vision Computing*, vol. 48–49, pp. 37–41, 2016.
- [19] J. Mahmoodi and A. Salajeghe, "A classification method based on optical flow for violence detection," *Expert Systems with Applications*, vol. 127, pp. 121–127, 2019.
- [20] Q. Xu, J. See, and W. Lin, "Localization guided fight action detection in surveillance videos," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 568–573, Shanghai, China, July 2019.
- [21] I. P. Febin, K. Jayasree, and P. T. Joy, "Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm," *Pattern Analysis and Applications*, vol. 23, no. 2, pp. 611–623, 2019.
- [22] X. Wang, C. Yao, X. Su, J. Dong, and Y. Li, "Random forest based multi-view fighting detection with direction consistency feature extraction," in *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (Green Com) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (Smart Data) and IEEE Congress on Cybermatics (Cybermatics)*, pp. 558–563, Rhodes, Greece, November 2020.
- [23] F. A. Pujol, H. Mora, and M. L. Pertegal, "A soft computing approach to violence detection in social media for smart cities," *Soft Computing*, vol. 24, no. 15, pp. 11007–11017, 2019.

- [24] H. Zhang, Y. Qian, Y. Wang, R. Chen, and C. Tian, "A ViBe based moving targets edge detection algorithm and its parallel implementation," *International Journal of Parallel Programming*, vol. 48, no. 5, pp. 890–908, 2019.
- [25] O. Barnich and M. van Droogenbroeck, "ViBe: a universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.
- [26] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, <http://arxiv.org/abs/1804.02767>.
- [27] J. Redmon, *Darknet: Open Source Neural Networks in C*, 2013, <http://pjreddie.com/darknet/>.
- [28] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vancouver, BC, USA, August 1981.
- [29] O. Déniz, I. Serrano, G. Bueno, and T. Kim, "Fast violence detection in video," in *International Conference on Computer Vision Theory & Applications*, Lisbon, Portugal, January 2015.
- [30] W. T. Freeman, "Orientation histograms for hand gesture recognition," in *International Workshop on Automatic Face and Gesture Recognition*, pp. 296–301, Zurich, Switzerland, 1995.
- [31] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [32] L. Ren, Z. Meng, X. Wang, L. Zhang, and L. T. Yang, "A data-driven approach of product quality prediction for complex production systems," *IEEE Transactions on Industrial Informatics*, p. 1, 2020.
- [33] H. Zhu, T. Xie, Y. Guan, C. Deng, and X. Wang, "Hypergraph matching with an entropy barrier function," *IEEE Access*, vol. 7, pp. 16638–16647, 2019.
- [34] C. Hu, Y. Chen, L. Hu, and X. Peng, "A novel random forests based class incremental learning method for activity recognition," *Pattern Recognition*, vol. 78, pp. 277–290, 2018.
- [35] M. S. Ryoo and J. K. Aggarwal, *UT-Interaction Dataset*, 2010, http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html.
- [36] J. Ren, G. Wu, X. Su, G. Cui, F. Xia, and M. S. Obaidat, "Learning automata-based data aggregation tree construction framework for cyber-physical systems," *IEEE Systems Journal*, vol. 12, no. 2, pp. 1467–1479, 2018.
- [37] L. Xia, W. Guo, and H. Wang, "Interaction behavior recognition from multiple views," *Journal of Central South University*, vol. 27, no. 1, pp. 101–113, 2020.
- [38] Z. Wang and B. Gao, "Spatio-temporal fusion convolutional neural network for abnormal behavior recognition," *Computer Engineering and Design*, vol. 41, no. 7, pp. 2052–2056, 2020.
- [39] S. Nigam and A. Khare, "Integration of moment invariants and uniform local binary patterns for human activity recognition in video sequences," *Multimedia Tools and Applications*, vol. 75, no. 24, pp. 17303–17332, 2016.