Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Action Recognition Based on Cascade Feature and Multilayer Classifier

Taizhe Tan, Chuhong Li [a)]

Guangdong University of Technology, Guangdong, 510006, China.

[a)] Corresponding author: 253298502@qq.com

Abstract. To Propose a recognition algorithm combines the feature of cascade of HOG/HOF with the multi-layer classifier. By extracting the foreground region of the video sequence and extracting the HOG feature and the HOF feature to makes up the HOG/HOF cascade feature. Then making all the HOG/HOF cascade features form a set of feature vectors. Then, to construct a multi-layer classifier consists of twice self-organizing mapping networks and a layer of supervised neural networks. Finally, all the features are classified to get the result of behavior recognition. The simulation results show the algorithm has a high recognition rate.

Key words: action recognition; multi-layer classifier; cascade feature.

INTRODUCTION

There are two common kinds of recognition methods of human behavior: matching module method and probabilistic method [1-2]. The former one is to convert video fragments or image sequences into a set of modules, and by comparing the similarities between the training modules and the modules to be identified to identify the behavior categories such as MEI(motion energy images)[3], MHI(motion history images)[4], TrajMF [5] and the Action Bank[6] which is hot in recent years, The latter one is to transform human behavior into several states set in the order of time, then use the probability to describe the transferring process between different states, then to identify different behaviors according to the relationship of states changed, such as HMM(Hidden Markov Model)[7], CRF(conditional random field)[8] and so on. Generally speaking, the matching module method is more common and widely used. However, the function of the behavior recognition of this method is greatly affected by the training samples. When the training sample is not sufficient, the behavior recognition performance of the method drops seriously. The probabilistic statistical method has a good recognition effect for some periodic behaviors, such as the recognition of running behavior. When the human behavior is complex and changeable, the recognition performance of this kind of method will decline obviously.

In order to improve the performance of complex behavior recognition, this paper proposes a human behavior recognition method combining the HOG/HOF features and multi-layer classifier. The design thoughts are: Firstly, the foreground region of each frame image is extracted by using the video correlation feature, and the behavior feature is extracted only in the foreground region, which improves the operation efficiency and reduces the interference of background. Then the HOG (histogram of oriented gradients) feature and the HOF (histogram of oriented optical flow) feature are extracted from the foreground region to form the HOG/HOF feature. Then the HOG/HOF concatenated features of all foreground regions in the video segment are constructed into a set of feature vectors to describe the behavioral characteristics of the video segment. Finally, the feature vectors set of the video segment is sent to a multi-layer classifier which is composed of two layers of self-organizing mapping networks and a layer of supervised neural network to recognize the categories of behavior. The specific steps are shown in figure 1.

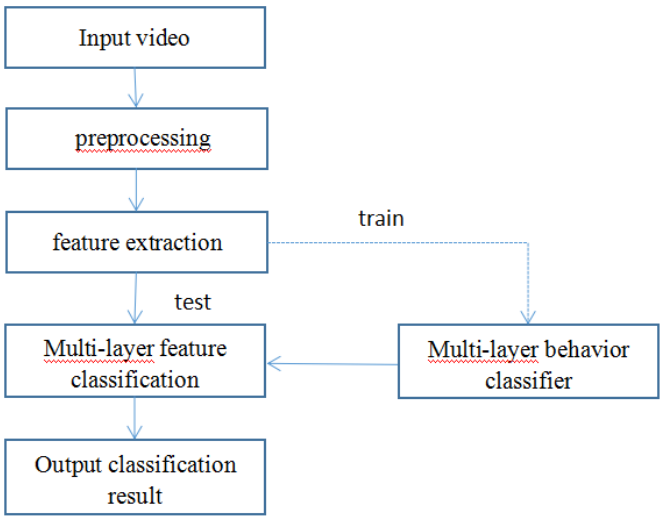

FIGURE 1. The implementation process of the algorithm

FEATURE EXTRACTION AND DESCRIPTION

Before extracting the features, we need to pre-process the video to eliminate background regions that are not interested, which can improve the overall efficiency of behavior recognition, but also reduce the misjudgment caused by these interference areas because of removing the interference regions. In this paper, the background subtraction method is used to carry out the motion reconnaissance mission, and the literature is adopted in detail. The outstanding advantage of the proposed VIBE [9] algorithm is that the background construction speed is fast, only one frame of image can be used to build the background.

Space Time Interest Point(STIP) and MEI are commonly used to describe human behavior. However, the stability of space-time interest points is very poor, which is influenced by the complexity of human behavior itself, the complexity of the scene and the distance from the camera. The motion energy map requires the alignment of human body objects in different video frames. It is suitable for some behavior analysis where the scene is simple and the position of human body is not changed much. This paper mainly focuses on the analysis of complex human behavior, so the behavior feature extraction here should take into account the complex scene and complex human body changes as much as possible. Specifically, the behavioral feature extraction method based on HOG and HOF is used to extract the behavior feature of foreground region. Here the hog feature is used to describe the attitude characteristics of human body in airspace under different behavior conditions, which is very meaningful to distinguish different types of behavior. For example, there are obvious differences in human posture when the behavior of running and diving occurs. OF features are used to describe the changes of human body's attitude in time domain under different behavioral conditions, including the changes of motion direction and velocity of different parts of human body. For example, the movement direction and speed of every part of human body are very different between running behavior and walking behavior. For the foreground region of each frame, we first extract the HOG feature and the HOF feature, and then concatenate them in a single image (HOG features at the front, HOF features at the rear) to obtain the joint features of the foreground region, and the implementation process is shown in figure 2, the details of extracting steps of the Hog feature are in the literature. [10] The details of extracting steps of HOF features are in the literature [11]. The specific steps are shown in figure 2.

In this way, starting from the second frame, a HOG/HOF cascade feature vector can be extracted for each foreground region of each frame image. All the HOG/HOF cascaded feature vectors extracted from the video segment are gathered together to obtain the feature vector set of the video fragment, which is described as

$$X \{xi = |i = 1, 2... K\} \quad (1)$$

Where K is the total number of HOG/HOF cascade eigenvector.

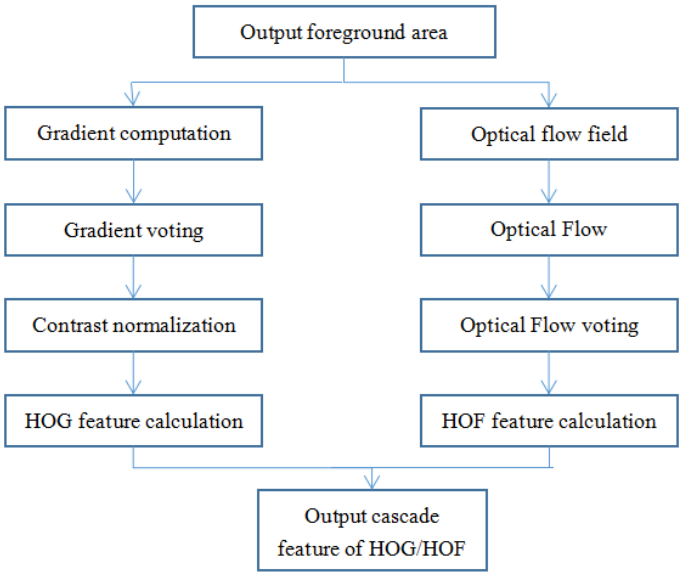

FIGURE 2. The process of extracting cascade feature of HOG/HOF

MULTILAYER CLASSIFIER

The multilayer action classifier in this paper is composed of three layers of neural networks; the first layer and the second layer are self-organizing map networks, while the third layer uses a neural network with supervision.

The First Layer Self-Organizing Mapping Network

The self-organizing map network consists of an input layer and an output layer. The neurons in the input layer and the output layer are linked with full connection. The weight vectors of the output layer neurons are self-adaptively updated during each iteration process in the training stage. The self-organizing mapping network is composed of a set number of neuron grids, forming a fixed topology network. Each neuron n_i corresponds to a weight vector w_i. In the initialization phase, all the elements of the weight vector are randomly assigned to a real number between 0-1 and normalized to a unit vector. When a new input vector is inputted, each neuron represents it in a competitive manner. The neurons that win the competition are called BMU (Best matching unit,). For the input vector x (t), which is inputted by the moment t into the self-organizing mapping network, all of the neurons in the network can receive this input vector. Its corresponding optimal matching unit must satisfy the condition that among all neurons the weight vector n_b of the optimal matching unit is supposed to be most similar with the input vector x(t).

This paper uses cosine measure to describe two vectors (the similarity between the weight vector w_i and input vector x(t)) is represented as

$$R_i = \frac{x(t) \times w_i(t)}{\|x(t)\| \times \|w_i(t)\|} \tag{2}$$

Then, the ordinal number b of the optimum matching unit n_b can be expressed as

$$b = \arg \max_{i=1,\cdots,N} R_i(t) \tag{3}$$

Where N represents the number of neurons in this layer. The optimal matching unit together with its neighbor neurons respond to the input vector. Neighbor neurons will gradually become the special representation unit of the

similar input vector and representing the similar input vector as an ordered mapping feature. Another important characteristic of the self-organizing map networks is their inductive ability, which can be used to identify or describe unknown input vectors.

The Second Layer Self-Organizing Mapping Network

On the basis of the first layer self-organizing neural network, the neurons of the second layer self-organizing mapping network generate new feature vectors by modifying the weight vectors of the current input vectors. Its goal is to increase the similarity between the new feature vector and the original input vector. Concretely, the number of weight vectors needs to be increased. This paper determines to increase the number of weight vectors depending on the distance between the current neuron n_i and the best matching unit n_b. Expressed as

$$\Delta w_i = \gamma(t) G_{ib}(t) \left\| x(t) - w_i(t) \right\| \tag{4}$$

In this function, $\gamma(t)$ represents the learning rate of the current time t, which decreases with time. G_{ib} represents a Gauss function, of which the radius $\sigma(t)$ is decreasing monotonously over time. This Gauss function is used to describe the distance distribution between the neuron n_i and the best matching unit n_b, as

$$G_{ib}(t) = \exp\left(\frac{-d_{ib}^2}{\sigma(t)^2}\right) \tag{5}$$

Where the d_{ib} represents the Euclidean distance between the neuron n_i and the optimum matching unit n_b.

Third Layer Supervised Neural Network

The third layer of supervised neural network is the output layer of the classifier, which is composed of c neurons, and c serves as the categories of behavior. Each neuron n_i corresponds to one weight vector $w_i \in R^n$, among which n is equal to the number of neurons of the second layer. Same as previous description, in the initialization phase, all the elements of the weight vector are randomly assigned to a real number between 0-1. At the time t, each neuron n_i receives an input vector $x_i(t) \in R^N$, which is also the output of No.i neuron of the second layer. After passing through the neuron n_i, $x_i(t)$ adopts the Cosine measure and obtain a classification score which is expressed as

$$y_i = \frac{x_i(t) \times w_i(t)}{\left\| x_i(t) \right\| \times \left\| w_i(t) \right\|} \tag{6}$$

Then, select out the neuron with the highest score in the corresponding classification, and its ordinal number is

$$i = \arg \max_{i=1,\cdots,c} (y_i) \tag{7}$$

That is, the result of the human behavior classification of the current video segment is class i.

EXPERIMENT

In order to evaluate the performance of this algorithm, we compared the algorithm with MHI, TrajMF and Action Bank. The performance of the algorithm is evaluated according to the recognition rate. The datasets we selected is UCF Sports [12] and ADL [13].

Figure 3 shows a comparison of the recognition rates of the four methods tested under two datasets. As can be seen from figure 3, the recognition rate of this method is the highest in the two datasets, and the average recognition

rate index is higher than that of the second ranked algorithm by over 3 percentage points. This shows that this method is superior to the other two methods in the classification of human behavior. Furthermore, figure 4 and 5 show the classification confusion matrix of our algorithm under two datasets. From figure 4, we can see that among the 10 behavior categories in the UCF Sports data set, the method in this paper only has confusion classification for three kinds of behaviors, while the Action Bank method, which has a lower recognition rate, has confusion classification for four kinds of behavior. For example, Action Bank method wrongly classifies some walking behaviors into diving, golfing and swing bench behaviors, while the algorithm in this paper only wrongly classifies some walking behavior into running behaviors, which is because this method combines the attitude characteristics (HOG) of the behavior object spatial domain and motion characteristics (HOF) of time domain, enhancing the ability to distinguish the characteristics of different behaviors, and reducing the confusion phenomenon in the classification of different behaviors. From figure 5, we can see that among the 10 behavior categories in the ADL data set, the method in this paper only has confusion classification for two kinds of behavior, while the Action Bank method with a lower recognition rate will have confusion classification for three kinds of behavior. It can be seen that this method has a strong ability to distinguish different behaviors and a low rate of confusion to complex behaviors.

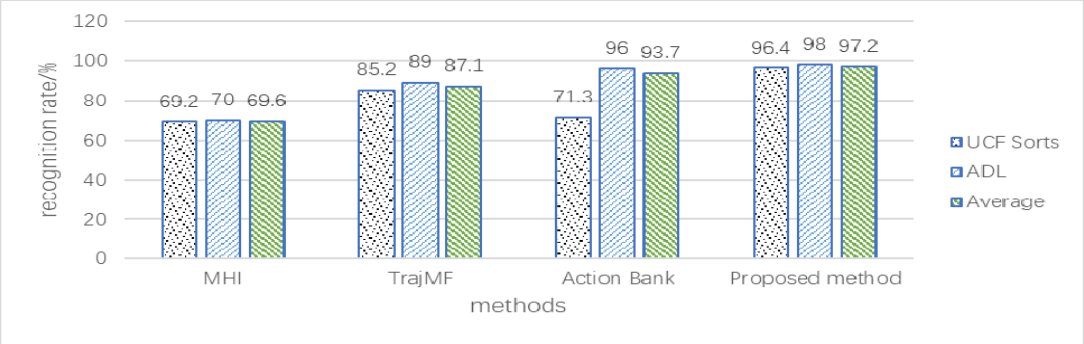

FIGURE 3. Comparison of human behavior recognition rate

answerphone	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dial phone	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Look up phone book	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chop banana	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Peel banana	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.10	0.00
Eat banana	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Eat snack	0.00	0.00	0.00	0.00	0.10	0.00	0.90	0.00	0.00	0.00
Drink water	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Use silverware	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Write on board	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	Answer phone	Dial phone	: up phone book	Chop banana	Peel banana	Eat banana	Eat snack	Drink water	Use silverware	Write on board

FIGURE 4. The confusion matrix of classification result o in UCF Sports

	diving	golfing	kicking	lifting	riding horse	running	skateboarding	swing bench	swing side	walking
diving	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
golfing	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
kicking	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lifting	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
riding horse	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
running	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.14
skateboarding	0.00	0.00	0.00	0.14	0.00	0.00	0.86	0.00	0.00	0.00
swing bench	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
swing side	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
walking	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.89

FIGURE 5. The confusion matrix of classification result o in ADL

CONCLUSION

The paper put forward a recognition algorithm which combines the feature of HOG/HOF with the multi-layer classifier. By extracting the foreground region of the video sequence and extracting the HOG feature and the HOF feature to makes up the HOG/HOF cascade feature. Then making all the HOG/HOF cascade features form a set of feature vectors. Then, to construct a multi-layer classifier consists of twice self-organizing mapping networks and a layer of supervised neural networks. Finally, all the features are classified to get the result of behavior recognition. The experimental results prove that the proposed approach in this paper is more excellent than other state-of-the-art methods.

REFERENCES

1. Guo Zixin, Yi Yang, Li Hanju. Identification of natural environment video behavior based on adaptive feature fusion[J]. Journal of Computer Science,2013,36(11):2330-2339.
2. Li Ruifeng, Wang Liangliang, Wang Ke. Review of human action recognition research[J].Pattern Recognition and Artificial,2014,27(1):35-48.
3. Zhan Xianggan, Liu Haihua, Gao Zhiyong. Recognition human actions using accumulative motion energy image[J]. Journal of South-Central University for Nationalities (Natural Science Edition),2016,35(1):108-113.
4. Thanikachalam V, Thyagharajan K K. Huam action recognition using motion history image and correlation filter[J]. International Journal of Applied Engineering Research, 2015, 10(34):27361-37363.
5. Jiang Y G, Dai Q, Xue X, et al. Trajectory-based modeling of human actions with motion reference points[C]//European Conference on Computer Vision,2012:425-438.
6. Sadanand S. Action bank: A high-level representation of activity in video[C]//IEEE Conference on Computer Vision & Pattern Recognition.2012:123-1241.
7. Rajkumar Saini, Partha Pratim Roy, Debi Prosad Dogra. A segmental HMM based trajectory classification using genetic algorithm[J]. Expert Systems with Applications,2018:169-181.

8. Abidine M B,Fergani B. Evaluating C-SVM,CRF and LDA classification for daily activity recognition [C]//International Conference on Multimedia Computing and Systems,2012:272-277.
9. Barnich O, Van Droogenbroeck M. ViBe: A universal background subtraction algorithm for video sequences[J]. IEEE Transactions on Image Processing,2011,20(6):1709-1724.
10. Kataoka H, Hashimoto K,Iwata K, et al. Extended cooccurrence HOG with dense trajectories for fine-grained activity recognition [C]//Asian Conference on Computer Vision,2014:336-349.
11. Wang T, Snoussi H. Histograms of optical flow orientation for abnormal events detection [C]//IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. IEEE,2013:45-52.
12. Uijling J, Duta I C, Sangineto E, et al. Video classification with densely extracted HOG/HOF/MBH features: An evaluation of the accuracy/computational efficiency trade-off[J]. International journal of Multimedia Information Retrieval,2015,4(1):33-44.
13. Moghaddam Z, Piccardi M. Training initialization of hidden Markov models human action recognition[J]. IEEE Transactions on Automation Science & Engineering,2014,11(2):394-408.