

A Method of Rate-based Behavior Recognition

Kunmin Yeo*, Whui Kim*, Seungtae Kim**

*Electronics and Telecommunication Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon, Korea

**Agency for Defence Development, Yuseong P.O. Box 35, Daejeon, Korea

kunmin@etri.re.kr, khsunkh@etri.re.kr, st6556@add.re.kr

Abstract—This paper presents a method to significantly improve the accuracy of behavior recognition and overcome the physical limitations of recognition speed. It performs learning based on variable-sized behavior times for the same behavior, and suggests a method to match the learned frame rate to the inference model. Since it takes a lot of time to construct a skeletal feature map provided as the input of the inference model, in this paper, by pre-calculating and configuring the feature points of each video frame and storing them in a queue, it is possible to avoid repetitive calculations and reduce the input configuration time of the inference model by calculating only the feature points of new video frames. With the implementation of the present invention, inference for various frame rates is possible, and the accuracy of behavior recognition can be increased through frame rate matching.

Keywords— Behavior, Action, Human skeleton, Object detection, Human pose

I. INTRODUCTION

Among behavior recognition modalities in [1]-[5] and [8], skeleton-based behavior recognition has received significant attention in recent days due to its behavior-focusing nature and compactness [1]-[4]. graph convolutional networks (GCN) [1]-[3] have been one of the most popular techniques. Additionally, GCNs considers a human skeleton point (joint) at each image frame as a node. Each edge connects adjacent nodes according to the spatial and time domains. Convolution layers are then used to the graph with space and time domains. Due to performance and simplicity for behavior recognition based on skeleton representation, GCNs have been thought as a popular approach when constructing and realizing skeleton-based behavior recognition [2].

While significant results have been studied, GCN-based approach are limited in important aspects: (1) robustness, (2) interoperability, and (3) scalability [4].

In order to recognize behaviors based on AI, unlike simple object detection, it must pass through several complex models. That is, it can be composed of a step of detecting a target object from a video or camera input transmitted in real time, a step of extracting a skeleton from the detected object, and a step of finally recognizing the final behavior by interpreting several video frames in time. Since it goes through several AI models, it takes a lot of time, and there is a problem in processing it in real time. If the processing time increases, detailed behavior recognition cannot be performed, and it ultimately acts as a cause of lowering the performance of

behavior recognition accuracy. Another problem is that the time between the start and end points of each behavior, that is, the behavior time, is different from each other, and even for the same behavior, the behavior time varies depending on the characteristics of the person. In addition, depending on the speed of the video frame, the inference model needs to match the input configuration according to the learned frame speed.

This paper is designed to solve these two problems, and aims to improve the accuracy and speed of behavior recognition in time-variant real-time images. In other words, the present invention aims to present a method that simultaneously overcomes the two problems mentioned above, that is, the problems of accuracy and inference speed.

II. RATE-BASED BEHAVIOR RECOGNITION

A rate-based behaviour recognition is addressed with training and inference schemes. Since we focus on behaviour determination, the object (human) detection and extraction skeleton, which can be achieved with [6] and [7] respectively, are out of scope of this paper.

A. Training Scheme

Figure 1 shows the fixed bit rate-based activity recognition training process.

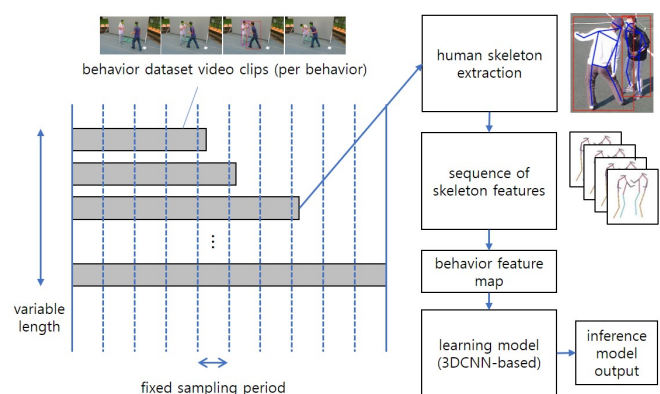


Figure 1. Training scheme

A dataset for a specific behavior consists of video clips of various lengths. This is because each person has different behavior times for the same behavior. Video frames are sampled at fixed intervals for video clips of different lengths. Skeletons are extracted and accumulated from each frame to

construct a behavior feature map. The behavior feature map is constructed in the form of a 2D image by reflecting the probability distribution from the skeleton coordinates of each video frame. In other words, for one skeleton coordinate, a specific area is expressed by imaging it around the corresponding coordinates (skeleton feature). The skeleton features of each frame are accumulated for multiple video frames that constitute a specific behavior to produce a behavior feature map for the corresponding behavior. The same process is performed for all target behaviors. The behavior feature map is a basic unit that characterizes a specific behavior and includes time-varying information of the behavior. The behavior feature map is input to the train model to produce an inference model. The train model is constructed based on 3DCNN that reflects the accumulated time-varying situation of the behavior [4]. The resulting inference model is valid for a fixed sampling period applied to the dataset, and the corresponding sampling period must be used during inference.

B. Behavior Determination Scheme

Figure 2 shows the fixed-rate-based behavior recognition decision process. The image is input from a real-time camera, a human body is detected, and the skeleton of the corresponding bounding box is extracted [6]-[7]. The skeleton coordinates are converted into skeleton features and sequentially accumulated in the sliding window queue. A skeleton feature is expressed as a 2D image that is probabilistically spread over a certain area from the skeleton coordinate [4]. Then, the behavior feature map is composed of multiple skeleton features as an input (e.g. tensor) to behaviour inference model which is constructed based on 3DCNN.

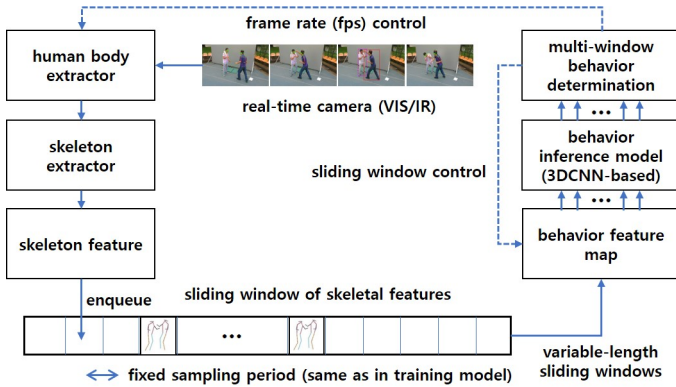


Figure 2. Behavior determination scheme

As mentioned earlier, since the behavior time for one behavior is different for each person and the behavior time for each behavior is also different, multiple inputs with sliding windows of skeleton features of variable size are configured. Here, the range of the sliding window is from the minimum behavior time (N_{min}) to the maximum behavior time (N_{max}) where N_{min} and N_{max} are in terms of the number of frames. If the number of behavior inference models to be driven is M ,

the size of each sliding window is configured to increase by $\text{ceil}((N_{max} - N_{min})/M)$.

The final multi-window-based behavior is determined as the behavior with the maximum confidence score value generated from each behavior inference model. By pre-configuring the skeleton feature and storing them in a queue from the feature points of each image frame at the time of arrival, it is possible to avoid repetitive calculations and reduce the input configuration time of the inference model only by calculating the feature points of new video frames.

III. RESULTS AND ANALYSIS

To verify the proposed method, we target the following four behaviors: violence, faint, invasion, and walking with different distribution of behaviour times. The dataset is collected with VIS and IR cameras in various environment (or backgrounds) as shown Figure 3. The summary of the dataset for training and validation is shown in Table 1.

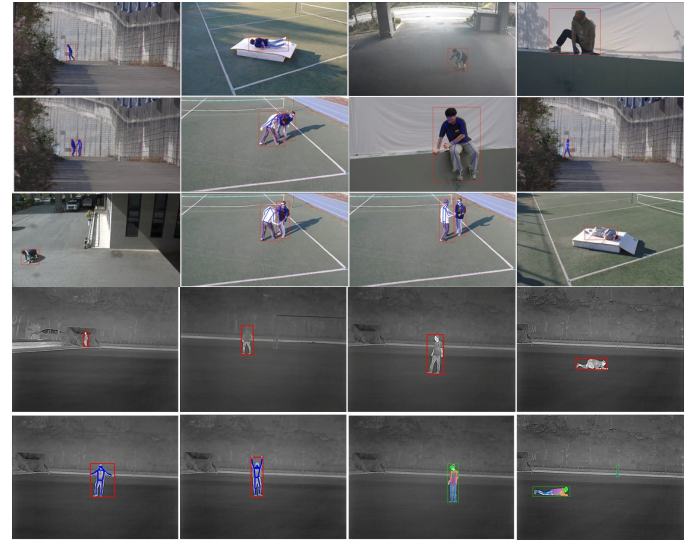


Figure 3. Dataset video clips (with VIS/IR camera)

TABLE 1. DATASET FOR TRAINING AND VALIDATION

Type of behaviors	Number of video clips and behaviour time		
	Training	Validation	Behavior time
Violence	421	1,210	2~5 sec.
Faint	405	902	2~4 sec.
Invasion	403	924	4~8 sec.
Walking	412	951	1~3 sec.

Frame rate is set to 10fps, since this value is sufficient to represent real-time behaviors. The total number of behaviour inference models M is set to 15 to account for real-time processing of multiple inference models. As a result, the processing time of the inference model is measured as 0.011s when the size of skeleton feature image is 56x56 (width x height) pixels.

Figure 4 illustrates the performance of the rate-based behavior recognition with the size of skeleton feature image. As shown in the figure, with increasing values of the size of

skeleton feature image, the mean AP increases. At the skeleton feature size of 35x35 pixels, the mean AP approaches to 0.9781, which seems a reliable number in order to efficiently recognize certain classes of behaviors.

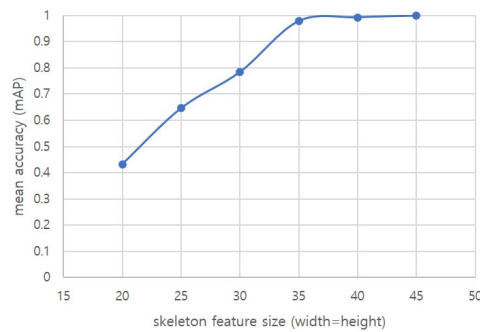


Figure 4. Performance of rate-based behavior recognition

IV. CONCLUSIONS

This paper is a technical field regarding action feature pre-processing for image-based behavior inference. In order to recognize behavior in real-time images that change over time, a video learning method and a feature map construction method for inference are provided. Through the methodology of the paper, the accuracy and speed of behavior recognition can be significantly improved. Also, the implementation of the paper enables the behaviour inference for various frame rates. Finally, a further study may be needed in order to accommodate such recognition environments as large number of persons engaged.

ACKNOWLEDGMENT

This research was supported by the Institute of Civil Military Technology Cooperation funded by the Defense Acquisition Program Administration and Ministry of Trade, Industry and Energy of Korean government under Grant 22-SN-EC-16.

REFERENCES

- [1] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," CVPR, pp. 183–192, 2020.
- [2] Z. Qin, Y. Liu, P. Ji, D. Kim, L. Wang, A. Saeed, and T. Gedeon, "Fusing higher-order features in graph neural networks for skeleton-based action recognition," IEEE Trans. Neural Networks and Learning Systems, 4783–4797, 2024.
- [3] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," ICCV, pp. 13359–13368, 2021.
- [4] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2959 – 2968, 2022.
- [5] N. Nida, M. Yousaf, A. Irtaza, and S. Velastin, "Video augmentation technique for human action recognition using genetic algorithm," ETRI Journal, vol. 44, pp. 327–338, 2022.
- [6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv:1506.01497, 2015.
- [7] K. Sun, B. Xiao, D. Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," CVPR, pp. 5693–5703, 2019.
- [8] H. Han, H. Zeng, L. Kuang, X. Han, and H. Xue, "A human activity recognition method based on Vision Transformer," Scientific Reports, vol. 14, 2024.