Summary of continuous action recognition

Jiahui An, Xinrong Cheng, Qing Wang, Hong Chen, Jiayue Li and Shiji Li

Department of Computer Science and Technology, China Agricultural University, Beijing 100083, China
*cheng_xinrong@126.com

Abstract. In the field of human-computer interaction, it is very important for computers to understand human behaviors, so human action recognition is of great significance. But in the current action recognition work, most of them are for the segmented action data. Compared with this, there is less research on continuous action recognition. Therefore, this paper summarizes the research on continuous action recognition, and elaborates on action feature extraction, action classification, and continuous action segmentation.

1. Introduction

Human-computer interaction refers to the interaction between human and computer in a certain way to complete the information exchange between human and computer more efficiently and naturally. People can interact with the computer through different channels such as vision, touch, speech, gestures, expressions, eye movements and other channels.

People hope that computers can become more and more intelligent, and they can "see" the world and "listen" to the world like humans. Among them, computer vision technology can make computers "see" the world like humans. In the field of computer vision, action recognition technology has an important position, it can understand human movements, and better interact with people. It has appeared in video surveillance, gaming, medical and virtual reality fields.

In action recognition, most of them are still in the research of single action. That is, it is assumed that the input actions are independent of each other or have been divided, that is, the start point and end point of the action sequence data must be manually selected [1]. But in the actual situation, there will be some non-core actions before and after an action. If you want to identify the core action, you must manually select the start point and end point of the action sequence data when calibrating the action sequence. This obviously increases the workload, and also adds human factors, making the action data not very accurate. At the same time, it also affects the interaction process between the user and the system, reducing the naturalness of the interaction. Therefore, continuous action segmentation and recognition are required.

The traditional methods of action recognition are divided into several processes: feature extraction, feature integration, and feature classification. In recent years, with deep learning methods, it has also appeared in the field of action recognition. Next, this article will introduce action feature extraction, action feature classification, and continuous action segmentation.

2. Research on action feature extraction

In human action recognition, if we can extract the action features that effectively express the action, it is very important for the result of action recognition, because it directly affects the final result of action recognition. Therefore, it is necessary to analyze the specific situation, select different types of
features according to the different video quality and application scenarios, and choose different methods according to the different types of features selected. More common features can be extracted from the following aspects: static features, dynamic features, space-time mixed features and descriptive features\[^2\].

Static features, as the name implies, mainly describe the state of the human body target when it is relatively still, such as size, contour, shape, edge, etc., showing the overall information of the human body. For static features, a pose estimation method can be used at this time. In order to complete the action recognition, Carlsson et al.\[^3\] performed shape matching between the key frames extracted from the action video and the saved action prototype. Among them, the shape information is some edge data, which is detected by the Canny edge detector. Shotton et al.\[^4\] estimated the 3-dimensional spatiotemporal position of human joints from depth images. Compared with RGB image features, joint points have the advantage of not being blocked by light.

Dynamic features include movement speed, movement direction and trajectory, etc. Commonly used methods include low-level tracking methods and optical flow calculation methods, but the former is extremely error-prone, especially in complex scenarios. In view of this, Efros et al.\[^5\] divided the optical flow field into horizontal and vertical channels, and then divided them into left and right channels, respectively, and used a Gaussian filter to filter the four channels, and finally carried out Normalization realizes action recognition using optical flow descriptors.

Common image processing methods for spatio-temporal features, Davis et al.\[^6\] connect static frames to form action sequence, and then subtract the dynamic sequence background to form a static image to characterize a certain type of action. The static graph can be action energy graph or action history graph, and finally action recognition is performed.

Descriptive features mainly use machine learning methods. Yao et al.\[^7\] took atomic actions, objects and postures as descriptive features. In addition, automatic feature extraction using neural networks has also appeared. Hinton et al.\[^8\] proposed a deep belief network consisting of multiple layers of restricted Boltzmann machines. The RBM model learns the expression of features in an unsupervised manner, and the learning and training process is very efficient. Inspired by Hinton\[^9\], Lei Jun\[^10\] designed a convolution restricted Boltzmann machine to express the statistical structure of the samples in the tracking scene. In the CRBM model, several filters are learned from the data. These filters are actually equivalent to local feature detectors. When the amount of training data is small, the CRBM model can still be effectively trained and produce discriminative features.

3. Research on action recognition method

Current commonly used action recognition methods include: template-based methods, probability-statistic-based methods, and grammar-based methods\[^2\]. The template-based method is intuitive and simple, and judges the action category by comparing the similarity between the target to be detected and the template. Therefore, it lacks certain robustness. Ji et al.\[^11\] used the dynamic time warping method to calculate the degree of similarity between the action to be recognized and the action in the action library.

Probabilistic statistical models represent actions as a continuous sequence of states, and the transition law between states can be expressed with a time transfer function. Shi et al.\[^12\] adopted the Markov model and proposed a Viterbi-like dynamic programming algorithm to segment and identify continuous actions simultaneously. Liu Fen\[^13\] used Kinect sensors to generate human action depth maps, built a three-dimensional human model, used the angle and modulus ratio of action vectors as feature vectors, and used SVM classifiers to classify and recognize human actions. The classification diagram of linear support vector machine is shown in figure 1.
The grammar-based method describes human actions as a series of symbols. Each symbol represents an atomic level action. Action recognition is performed by first identifying the atomic action.

In recent years, with the continuous development of deep learning, it has gradually appeared in the field of action recognition. Yu Hua [14] uses an improved DPM algorithm to extract features, and uses a gradient optimization training CNN model to classify and recognize actions. Li Ting [15] used the improved L-K optical flow method of convolution kernel to extract action features for the problem of continuous action recognition, and used 3D CNN and SVM mixed models to recognize action. Article [16] proposed a new method of action recognition using convolutional neural network (CNN) and deep bidirectional LSTM (DB-LSTM) network to process video data. Depth features are extracted from every six frames of video, and the DB-LSTM network is used to learn the sequence information between frame features. In the forward and backward traversal of DB-LSTM, multiple layers are superimposed together to increase its depth. The model frame diagram is shown in figure 2.

4. Research on continuous action segmentation
In recent years, domestic and foreign researches on the recognition of single actions have made important progress. However, in most application scenarios, actions are not manually collected, labeled and segmented, and are more often complex continuous actions. Therefore, the recognition of continuous and complex actions is becoming more and more important. For the recognition of
continuous actions, it is an important part to segment the collected action time series. The quality of
the segmentation results directly affect the final recognition result.

4.1. Segmentation method
According to the sequence of segmentation and recognition, it can be divided into direct segmentation
method and indirect segmentation method. The direct segmentation method is to segment the action
first and then recognize. For example, Bai Dongtian [17] splits continuous actions into sequences
through a priori knowledge, extracts skeleton information as action features for each sequence, and
combines a hidden Markov model to identify individual actions of the human body. And through
dynamic programming algorithm and a threshold model to get the best action recognition results. One
drawback of this method is that the prior action knowledge divides the continuous action sequence to
reduce the recognition accuracy.

The indirect segmentation method is to recognize while performing action segmentation. For
example, Mao Yijie [18] studied an action segmentation recognition scheme based on the confidence of
support vector machine classification. A method for calculating the classification confidence of the
SVM multiclassifier is proposed. The sliding window is used to obtain the action starting point and
action category.

4.2. Segmentation model
Action segmentation is a kind of sequence analysis. The solution of sequence analysis is sometimes
domain model method, in which HMM model is widely used. The HMM algorithm was first applied
to the field of speech recognition, and recently has also been applied to action recognition. HMMS can
effectively solve the problem of spatial and temporal differences, but HMMS requires a large number
of training sets [1].

Therefore, in combination with the current hot deep learning technology, many researchers use
HMM and neural networks for continuous action segmentation and recognition. Luo Xiaoyu [19] uses a
sliding window-based method to detect the initial segmentation point, and uses deep confidence
networks and hidden Markov models to identify individual actions in continuous actions, and uses
dynamic programming to optimize the initial segmentation point to achieve continuous Segmentation
and recognition of actions. Lei Jun [10] combines the CNN network and the HMM model. It not only
combines the CNN network’s feature learning ability and the HMM model’s sequence dynamic
modeling ability, but also can realize the model training under the condition of weakly labeled samples.
For weakly labeled continuous action videos, the HMM model estimates the implicit sequence of
action categories, which can be used as label information to train the CNN network.

In addition, the conditional random field model is also widely used. Lei Jun [20] proposed a
convolutional neural network and hidden dynamic conditional random field model to solve the
problem of continuous action recognition in video. This method designs and constructs a three-
dimensional CNN network, which automatically learns the action features directly from the original
video data. The LDCRF model is used to model the continuous action, and the tasks of continuous
action recognition and segmentation are completed. The model framework is shown in figure 3.

Figure 3. CNN-LDCRF model framework.
5. Conclusion
Continuous action recognition is a challenging subject. With the development of computer technology, there have been many achievements. Now that video data is so rich, it lays the foundation for a large number of data sets required for deep learning. It is believed that there will be more research on continuous action recognition using deep learning technology in the future.

Acknowledgments
Thanks to my mentor Cheng Xinrong, Wang Qing and Chen Hong. Thanks for their valuable suggestions for my article.

References
[1] Huang, Y. H., Ye, S. Z. (2008) Summary of Continuous Action Segmentation. In: The 14th National Conference on Image Graphics. Fuzhou.
[2] Hu, Q., Qin, L., Huang, Q.M. (2013) Overview of human action recognition based on vision. Chinese Journal of Computers.
[3] Carlsson, C., Carlsson, S., Sullivan, S. (2001) Action recognition by shape matching to key frames. In: Proceedings of the Workshop on Models Versus Exemplars in Computer Vision. Colorado. pp. 1-8.
[4] Shotton, J., Fitzgibbon, A., Sharp, T., et al. (2011) Real-time human pose recognition in parts from a single depth image. In: Proceedings of the IEEE Conference on Recognition. Colorado Springs. pp. 1297-1304.
[5] Efros, A. A., Berg, A. C., Mori, G., Malik, J. (2003) Recognition action at a distance. In: Proceedings of the 9th IEEE International Conference on Computer Vision. Nice. pp. 726-733.
[6] Davis, J. W., Bobick, A. F. (1997) The representation and recognition of action using temporal templates. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Juan. pp. 928-934.
[7] Yao, B., Jiang, X., Khosla, A., et al. (2011) Human action recognition by learning bases of action attributes and parts. In: Proceedings of the IEEE International Conference on Computer Vision. Barcelona. pp. 1331-1338.
[8] Hinton, G. E., Osindero, S., The, Y. W. (2006) A fast learning algorithm for deep belief nets. Neural Computation, 18 (7): 1527–1554.
[9] Hinton, G. E., Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks. Science, 313 (5786): 504–507.
[10] Lei, J. (2017) Research on continuous action recognition method combining deep network and probability graph. National University of Defense Technology.
[11] Ji, R., Yao, H., Sun, X. (2011) Actor-independent action search using spatial temporal vocabulary with appearance hashing. Pattern Recognition, 44 (3): 624-638.
[12] Shi, Q., Cheng, L., Wang, L., et al. (2011) Human action segmentation and recognition using discriminative semi-Markov models. International Journal of Computer Vision, 93 (1): 22-32.
[13] Liu, F., Wu, Z. P. (2019) A Kind of Human Action Recognition Algorithm Based on Kinect and SVM. Modern Computer, 18: 55-58.
[14] Yu, H., Zhi, M. (2019) Human action recognition based on convolutional neural network. Computer Engineering and Design, 40 (04): 1161-1166.
[15] Li, T. (2018) Human body continuous action recognition based on 3D CNN. Harbin Institute of Technology.
[16] Ullah, A., Ahmad, J., Muhammad, K., et al. (2017) Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features. IEEE Access, 6: 1155-1166.
[17] Bai, D.T. (2016) Static gesture and continuous action recognition of upper limbs based on KINECT. Beijing Institute of Technology.
[18] Mao, Y. J. (2017) Research on continuous action recognition of human body based on kinect.
University of Electronic Science and Technology of China.

[19] Luo, X. Y. (2018) Human action recognition based on DBN-HMM. Xi’an University of Technology.

[20] Lei, J., Li, G., Li, S., Tu, D., Guo, Q.(2016) Continuous action recognition based on hybrid CNN-LDCRF model. In: 2016 International Conference on Image, Vision and Computing. Portsmouth. pp.63-69.