Research on Human Action Recognition Based on Global and Local Mixed Features

Xueping Liu[1,2]
1 Engineering Training Center
Shenyang Aerospace University
Shenyang, China
2 College of Automation Engineering
Nanjing University of Aeronautics and Astronautics
Nanjing, China
liuxueping024@163.com

Yibo Li
College of Automation
Shenyang Aerospace University
Shenyang, China
liyibol@sau.edu.cn

Abstract—In recent years, human action recognition has become a hot topic in the computer visual and pattern recognition fields, which has a wide application prospect, such as intelligent monitoring, human-computer interaction, virtual reality, content-based video retrieval and home robotics services. Although progresses have been made in the existing action recognition technology, it is still in the early stages of recognizing a minority of standard gesture and simple action experimental samples. This paper attempts to realize human action recognition through global features and local features structure variables and mixed feature descriptors reflecting the feature causal relationship, and thus improves the recognition rate of human action. This paper has certain practical significance and application values.

Keywords—Action Recognition; Mixed Features; Global; Local; Cognitive Map

I. INTRODUCTION

The main task of human action recognition lies in that the computer automatically recognizes the meaning of human action in a certain scene. Researches on action recognition are mainly divided into action feature extraction description and recognition, in which the recognition means classification of features according to the descriptive features. The frequently used classification methods consist of template matching method as well as the state-space method[1]. As a necessary precondition for classification, the action feature extraction and description[2-3] is another important aspect of research on action recognition.

Depending on whether it is necessary to model the human body in advance, action recognition research methods can be divided into model-based method and data driven method. Model-based method is composed of two-dimensional and three-dimensional models. The common two-dimensional models include the star-shaped skeleton model proposed by Chen et al [4] and elliptical human body model presented by Ben-Arie et al [5]. Two-dimensional model is generally limited by perspective and occlusion, so researchers have proposed three-dimensional human body geometric model[6-7] to be improved, but such algorithm is features by a huge calculated amount. Kehl et al[8] directly adopted the pose estimation method in three-dimensional spaces, which could describe the physical meaning of human body actions more accurately and it was easy to establish associations between various actions. However, human pose estimation itself is still under study. It has complex algorithm. Besides, the action recognition accuracy depends on the accuracy of the pose estimation. Data-driven method is to extract relevant features mainly based on the human actions, textures, shapes and other information in video images.

According to feature selection methods, action

recognition research methods can be subdivided into the method based on global features and the method based on local features. The method based on global features is to encode the entire action images with abundant information. Efros et al[9] extracted the people-centered region of interest, in order to obtain the optical-flow features for the future objectives within the region. Yilmazet al[10] proposed using spatio-temporal volume (STV) to describe the entire outline of the human body, which not only used the outline information of actions but also contained the spatial information. Nevertheless, the accuracy of global feature description depends on the accuracy of target tracking and background subtraction and other pre-treatment as well as interference sensitive to perspective, occlusion and noises. Based on local features, this paper only selects a part of the human information as the descriptive method, particularly the spatio-temporal interest point based on local features, which can directly detects the key points of significant human action changes from videos through corner detection and filtering processing, takes these key points as the interest points classifying the category of actions, captures the local shapes, appearances, actions and other features in the neighborhood with interest points as centers for description, so as to encode and classify the feature descriptors. This method has less influences on view occlusion, which becomes a hotspot at home and abroad.

II. Related Work

To solve the existing problems of human action recognition, this paper studies the action recognition algorithm based on mixed global and local features. According to two varied action types of in-situ and mobile state, this paper respectively selects the feature types and feature mixed descriptive methods suitable to respective features, besides, establishes the causal relationship between all local features, and improves the recognition accuracy.

In human's daily lives, they have a wide range of actions, which differed in action features. To recognize an action at one time as accurately as possible, it is essential to create feature descriptors with high feature dimensions, abundant information and guarantee of inter-class and intra-class differences. Although this method can improve the recognition rate of actions, the computational complexity increases. However, this problem can be effectively solved by building multiple action models, which is also one of the current main trends in the field of human action recognition. This paper follows the rule of human understanding things from global to local perspectives, and respectively establishes global and local action recognition models. First of all, this paper takes advantage of the global physical feature, and classifies human actions into two distinctive categories depending on whether human move when doing the actions: in-situ action and moving-state action. Secondly, this paper further extracts various local features (such as local shape features, movement features, etc.) reflecting details. Finally, this paper adopts different feature mixed strategies, and selects features for respective intra-class actions with better discrimination from all extracted features, resulting in the final mixed feature descriptor. The contradictions between recognition accuracy and recognition efficiency can be solved through this structured model.

Different types of features have different description perspectives of action states and different scopes of application. As for moving-state actions, such as running, leg/legs jumping ahead and walking are similar in overall shape, which need to strengthen the local actions, appearances and other features of legs; as for in-situ actions, it is necessary to focus on using features reflecting global human body outline changes. The original multi-feature fusion method, each component of feature vectors and number of weight dimension are fixed and inflexible. According to different action types generated from refinement, this paper can flexibly change the mixed feature strategy, select features that can effectively describes the original features of different action types from a number of alternative features, optimize the combination, and generate mixed feature descriptors with two different dimensions and component structures. Each mixed feature descriptor corresponds to a classifier, in order to respectively recognize in-situ and moving-state actions and improve the robustness of the recognition algorithm.

Cognitive map is representation of causal

relationship between features or concepts under a given environment, because it has intuitive knowledge presentation ability and strong reasoning ability, and allows the existence of feedback mechanism, so it is often used for a variety of complex system modeling, which has been extensively applied in the fields of image analysis and understanding, management decisions, failure analysis, analysis of social phenomena and control systems. This paper introduces the cognitive model into the human action recognition, so as to describe each local action feature as the concept node sets of cognitive map, build the cognitive map model, adopt the iterative training and learning, obtain the causal relationship between features, and enhance the reliability of action recognition.

Above all, the process of action recognition comes from easy to difficult and subdivides human's basic daily actions (such as running, jumping, walking, waving and bending, etc.). In the subdivision process, this paper designs different mixed feature descriptors for features of different types of actions, and introduces the temporal and causal relationship descriptors between features, and enhances the recognition rate of all actions, particularly similar actions (such as running and walking).

III. EXPERIMENTAL RESULTS

This paper uses the mixed global and local features to recognize human's basic daily actions, which has reduced the difficulty level of feature description and recognition. The system's global recognition process is shown in "Fig.1". First of all, this paper extracts global shape information of actions. Due to occlusion, light and other factors, although it is impossible to extract the accurate body shape or outline, it is possible to accurately recognize the general region and basic shape features of human by using the existing background subtraction. Accordingly, the displacement changes of horizontal human movement in the image frame sequence is analyzed. Depending on whether human movement position changes, actions are separated into in-situ (waving, bending, boxing, and in situ jumping) and moving-state classes (running, leg/legs jumping ahead and walking, etc.) Secondly, global features of actions are respectively extracted in accordance with two states, and then mixed feature descriptors are established for actions, so as to realize the human action recognition.

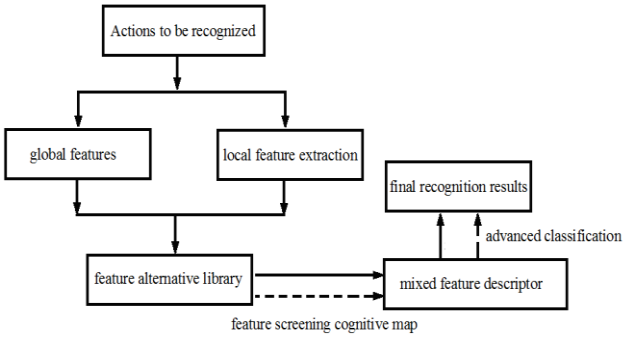

Figure 1. Global Recognition Flowchart

For in-situ actions, changes in the body shape vary greatly; for moving-state actions, there is a small intra-class difference in the global human outline changes. In addition, different human's strides, arm swinging and other actions also vary, resulting in great intra-class differences within the same actions. It is difficult to find a feature descriptor with intra-class and inter-class differences by relying on global outline features. Therefore, it is essential to strengthen the weight with features reflecting local changes (such as changes in the human legs) in action recognition among mixed features. As a result, based on the above features of two actions, this paper respectively adopts different feature mixed strategy to design the mixed feature descriptors, their structures may vary according to the different extracted descriptive feature types, and the weight of each subdivided features can also be changed accordingly. It is thus evident that this part can solve low human action recognition rate.

Mixed feature descriptor is composed of features based on global outline features and local spatio-temporal interest point. The descriptor based on outline features has larger weight, whereas the descriptor based on spatio-temporal interest point has smaller weight. This is because the global outline changes of different actions vary, such as waving and bending, and have greater impacts on action recognition than local changes of hands and feet. Even due to occlusion, light and other reasons, the overall outline features may be missing or inaccurate, differences between different actions can be well separated through compensation of local spatio-temporal interest points. It is unnecessary to adopt more features, increasing the complexity of the algorithm. Local spatio-temporal interest point has been widely used in previous studies, but it has a problem, namely, missing

spatio-temporal association between features. This paper improves by segmenting human regions, and divides human regions into three segments in accordance with the general proportion of human head, body and legs in biology. The lower segment (namely, legs) focuses on refinement changes, which adopts dimensional grids with smaller spatial and temporal scales to equally divide the entire lower region, calculates the local outline features in grids, make statistics for the number of interest points in different grids and mutual spatio and temporal relationship, and thus ultimately recognizes the meaning of actions.

Cognitive map theory can be used for data association rules mining, target reasoning classification and many other aspects. This paper only builds cognitive maps for each local interest point features of actions, in order to obtain a causal relationship between features and increase the robustness of the recognition. Each interest point is regarded as the concept node of cognitive map. According to the temporal positions of interest points in frame sequence images, directed arc of nodes (causal relationship directions) are established, the feature representation vectors of interest points reflect the ideal state of nodes, the weight of directed arc is called cognitive map's weight matrix. The element size in the matrix reflects the intensity of casual relationship between features. Therefore, this paper establishes a cognitive map and obtains the weight matrix of cognitive map. A learning method similar to neural network is adopted, in order to define an initial state matrix for each concept node and define the initial weight matrix, realize the iteration calculation for the state matrix and the weight matrix, until the state value and error under the ideal state are within a defined threshold range, and then the weight matrix represents the causal relationship between various features. Add the casual relationship to the mixed feature descriptor, so as to improve the recognition robustness.

By studying the described global and local mixed feature algorithms, this paper proposes an algorithm based on mixed feature recognizing the meanings of different human actions, and makes simulation tests through KTH, Weizmann, UCF and Hollywood2 and other public video image databases. Indicators to be achieved by the algorithm: (1) the average correct recognition rate of public reference database of KTH, Weizmann with simple backgrounds achieves 93%; (2) the average correct recognition rate of public reference database of UCF, Hollywood2 with complex backgrounds achieves 84%.

IV. CONCLUSION

In conclusion, this paper takes the human action features with global and local mixed feature description as the starting point, studies the multi-feature mixed description based on previous researches, defines a dynamic mixed feature model, and focuses on solving the problem of reduced action distinction degree as a result of information missing, which results from single global or local features. Besides, this paper uses the cognitive map to establish the casual relationship between features, weakens the intra-class differences caused by the same actions done by different persons or the same person at different times with different effects, and improves the robustness of human action recognition. The present study takes advantage of the internationally recognition KTH, Weizmann, UCF, Hollywood2 and other public video action recognition databases as objects of recognition tests, which have achieved good results.

REFERENCES

[1] LI Hong-song,LI Da,"The new research progress in the analysis of human motion," Pattern recognition and artificial intelligence, 2009,(1): 70-78.

[2] ZHU Guang-yu,XU Chang-sheng,HUANG Qing-ming,"Action recognition in broadcast tennis video,"[C] / /Proc of the 18th International Conference on Pattern Recognition.［S.l.］: IEEE Press,2006: 251-254.

[3] WANG Liang, SUTER D,"Recognizing human activities from silhouettes:motion subspace and factorial discriminative graphical model,"［C］/ /Proc of IEEE Conference on Computer Vision and Pattern Recognition.［S.l.］:IEEE Press,2007: 1-8.

[4] Chen, H.S., et al,"Human action recognition using star skeleton,"Proceedings of the 4th ACM international workshop on Video Surveillance and Sensor Networks,"2006:171-178.

[5] Ben-Arie Jezekiel,et al,"Human activity recognition using multidimensional indexing,"IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002,24 (8) : 1091-1041.

[6] Menier, C., Boyer, E., and Raffin, B,"3D skeleton-based body pose recovery,"Proceedings of the 3rd International Symposium on 3D Data Processing, Visualization and Transmission, Chapel Hill (USA), 2006: 389-396.

[7] Fossati, A., et al,"Bridging the gap between detection and tracking for 3D monocular video-based motion capture,"IEEE

Conference on Computer Vision and Pattern Recognition, 2007: 1-8.

[8] Kehl,R., Bray, M. and Van, GoolL,"Full body tracking from multiple views using stochastic sampling,"IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California, USA, 2005: 129-136.

[9] Alexei A. Efros, et al,"Recognizing action at a distance," Proceedings of the International Conference on Computer Vision, 2003:726-733.

[10] Yilmaz Alper, Shah M. ,"Matching actions in presence of camera motion,"Computer Vision and Image Understanding, 2006, 104 (2-3):221-231.