J. Vis. Commun. Image R. 56 (2018) 116–130

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

A new framework of action recognition with discriminative parts, spatio-temporal and causal interaction descriptors ☆

Ming Tong *, Yiran Chen, Mengao Zhao, Weijuan Tian

School of Electronic Engineering, Xidian Univrsity, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 21 April 2018
Revised 7 August 2018
Accepted 2 September 2018
Available online 4 September 2018

Keywords:
Action recognition
Spectral clustering
Discriminative constraint
Action part
Spatio-temporal relationship
Causal relationship

ABSTRACT

To improve action recognition performance, a novel discriminative spectral clustering method is firstly proposed, by which the candidate parts with the internal trajectories being close in spatial position, consistent in appearance and similar in motion velocity are mined. Furthermore, the discriminative constraint is introduced to select discriminative parts. Meanwhile, by fully considering the local and global distributions of data, a new similarity matrix is constructed, which enhances clustering effect. Secondly, the spatio-temporal interaction descriptor and causal interaction descriptor are constructed respectively, which fully mine the spatio-temporal and implicit causal interactive relationships between parts. Finally, a new framework is proposed. By associating the discriminative parts, spatio-temporal and causal interaction descriptors together as the inputs of Latent Support Vector Machine (LSVM), the correlations between action categories and action parts as well as interaction descriptors are mined. Consequently, accuracy is enhanced. The extensive and adequate experiments demonstrate the effectiveness of the proposed method.

1. Introduction

For the past few years, the computer vision has become a new subject and is at a stage of rapid development. As a critical technology of analyzing and understanding massive heterogeneous data of video, action recognition possesses significant academic value, potential business value and huge application prospect, and all of them make it become a research focus and difficulty in the computer vision field rapidly, well then an increasing number of scholars and research institutions have carried out numerous research works successively in the related aspects [1–3]. And consequently, action recognition has been successfully applied in the human-computer interaction field, e.g., intelligent surveillance, intelligent traffic, video retrieval, robot navigation and game entertainment. However, previous action recognition methods usually focus on the design and optimization of low-level features. Although quite a few achievements have been made, they usually merely represent the low-level vision information, and thus their description ability is very limited.

Mining the interactive relationship between action parts is a significantly crucial research line in the field of action recognition [4]. In fact, the part-based action representations have captured the attentions and good graces of many scholars for their abundant semantic information, strong description ability and high correlation to action category. Currently, there are mainly two types of part extraction methods, one of which is based on model learning, and the other is based on trajectory clustering.

The first type of part extraction method based on model learning mainly uses supervised learning method to learn multiple part models, and then mines parts through model matching to describe and explain objects or actions. Based on human pose annotations in 2D image, Bourdev et al. [5] constructed lots of templates and conducted template matching to discover the patches corresponding to the postures of body parts, and thus the poselets of body parts are constructed for human detection and localization. However, the number of poselets is excessively large as to require huge computational consumption. For this reason, Lin et al. [6] proposed the hierarchical Deformable Part Model (DPM). Firstly, global shape models are decomposed into several parts, and a part-template tree model is constructed. Then, pose-adaptive features are extracted and integrated with an affine-invariant tracker to conduct human detection, segmentation and tracking. Because this method adopts the tree structure, it effectively decreases part number and accordingly, the computational complexity is reduced.

☆ This paper has been recommended for acceptance by Dr. Zicheng Liu.

* Corresponding author.
E-mail addresses: mtong@xidian.edu.cn (M. Tong), yiran_chen@stu.xidian.edu.cn (Y. Chen), mazhao_1@stu.xidian.edu.cn (M. Zhao), tianweijuan@stu.xidian.edu.cn (W. Tian).

https://doi.org/10.1016/j.jvcir.2018.09.001
1047-3203/

However, it relies on accurate human detection, and meanwhile the tracking effect is unsatisfactory as well. In this respect, Lan et al. [7] proposed a spatio-temporal model representation, which does not require a reliable detection of motion subject, but treats its location as a latent variable, and then applies a Conditional Random Field (CRF) model with tracking constraints to detect 2D parts frame by frame. But the 2D parts obtained by frame detection only do well in capturing the main postures during the motion process. In order to introduce the temporal information into action detection, Tian et al. [8] generalized DPM from 2D images into 3D spatio-temporal volumes, by which the obtained spatio-temporal parts capture both of the appearance and motion information across several continuous frames, and consequently, the accuracy of action recognition is improved. In summary, there is a significantly practical barrier that a lot of hand-labeled annotation semantics are required for per each part, and meanwhile, the part acquisition heavily relies on the effectiveness of part detection algorithm. Thus, the automatic acquisition of discriminative parts is still a thorny issue.

As a crucial technique for the automatic mining of data distribution and implicit pattern, clustering methods have been widely applied in part extraction. Different from the part extraction method based on model learning, the parts obtained by trajectory clustering method might correspond to atomic actions, specific objects or perhaps random but informative spatio-temporal patches in the video sequences. Lezama et al. [9] achieved the incorporation of long-term motion cues from video frames by conducting trajectory clustering with coherent motion, and the motion information of object at different moments is obtained, and consequently, a good action recognition result is achieved. Yet, how many spatio-temporal patches are demanded to capture all the spatio-temporal motion variations in data is still a challenge. For this purpose, Ravichandran et al. [10] proposed a prior term that the clusters possess low motion discrepancy and take account of motion boundaries, and then automatically achieved the part number by balancing the variation of inter-cluster and the prior term for the clustering centers during the process of clustering. Furthermore, the membership function is adopted as an indicator function of parts to construct a new part representation called superfloxel. In order to obtain a more effective part representation, Zhao et al. [11] presented a framework of unsupervised contextual spectral clustering, in which the inter-image/video context information between visual words is utilized to evaluate the pairwise semantic relationship, and thus the effectiveness of parts is enhanced. For the sake of further highlighting the discriminativeness for action parts, Zhang et al. [12] developed a discriminative multi-scale spatio-temporal patch model, in which the activation ratio between the intra-class and inter-class detections of candidate parts is constrained to effectively mine the parts called actemes with good discriminative ability. However, this metric does not always ensure the high intra-class trigger frequency and low inter-class trigger frequency, and meanwhile, the obtained parts are not of explicit physical meaning.

Though parts mined by the two kinds of cited works above do achieve better performance of action recognition, the interactive relationships between each other are neglected. Subsequently, Brendel et al. [13] presented a video representation method based on spatio-temporal graphs, in which nodes represent multi-scale video segments, and directed edges capture their spatial and temporal relationships. Similarly, Raptis et al. [14] utilized trajectory clustering to obtain parts, and meanwhile captured the relative location relationship between each other. Successively, Yuan et al. [15] employed the graph-based clustering to obtain parts, referred to as activity components, and then introduced a Spatio-Temporal Context Kernel (STCK) method to exploit the relative motion between parts, and consequently, a better performance is achieved in action recognition. Yet, except for the existence of spatio-temporal relationship, abundant causal and other interactive relationships are hidden between parts. Narayan et al. [16] utilized the Granger causality to model the interactive relationships between pairs of parts, and thus a more informative video action representation is obtained. Nevertheless, this method relies on accurate part detection, meanwhile obtains the causal relationship between parts only within the single time segment of action video, and thus neglects the possible causality implied among parts at different motion stages of entire video duration, especially in complex actions.

Based on thorough analysis and research of the existing methods, a new framework of discriminative parts and their interactive relationships for action recognition is proposed, as illustrated in Fig. 1. The major contributions and innovations are summarized as follows. (1) A new discriminative spectral clustering method is proposed, which mines the candidate parts with the internal trajectories being close in spatial position, consistent in appearance and similar in motion velocity, and further adds constraint to obtain the better discriminative parts. (2) The spatio-temporal interaction descriptor and causal interaction descriptor between parts are constructed individually, which fully mine the spatio-temporal interactive relationship and implicit causal interactive relationship in motion patterns between parts. (3) A new framework of action recognition is proposed, which mines the correlations among action categories, action parts and interaction descriptors. Consequently, the action recognition accuracy is further improved.

The following chapters of this paper are presented as follows. Section 2 detailedly describes the proposed new discriminative spectral clustering method. Section 3 constructs two interaction descriptors and a new framework for action recognition. Section 4 shows the relevant experiments and analysis. Section 5 draws the conclusions with direction for future work.

2. Excavation of discriminative parts

Human action is a motion combination of body parts in space and time. In order to separate the parts that are beneficial for action classification, a discriminative spectral clustering method is proposed to obtain a group of parts possessing the better discriminativeness, which could achieve a more effective action representation. This method firstly performs the spectral clustering to automatically obtain candidate parts; then it imposes the

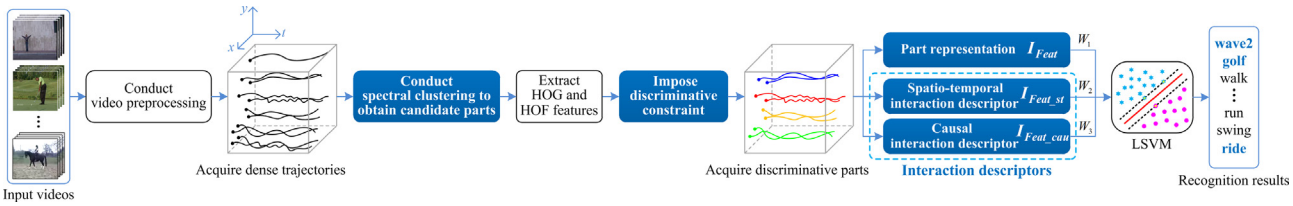

Fig. 1. The proposed new framework for action recognition.