J. Vis. Commun. Image R. 55 (2018) 778–788

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Evolution modeling with multi-scale smoothing for action recognition [☆]

Tingwei Wang [a,b,*], Chuancai Liu [a,e,*], Liantao Wang [c], Bingxian Ma [b], Xingjian Gu [d]

[a] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*
[b] *School of Information Science and Engineering, University of Jinan, Jinan 250013, China*
[c] *College of Internet of Things Engineering, Hohai University, Changzhou 213022, China*
[d] *College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China*
[e] *Collaborative Innovation Center of IoT Technology and Intelligent Systems, Minjiang University, Fuzhou 350108, China*

ARTICLE INFO

Article history:
Received 31 October 2017
Revised 27 July 2018
Accepted 16 August 2018
Available online 18 August 2018

Keywords:
Action recognition
Multi-scale representation
Rank pooling
Evolution modeling
Dynamics

ABSTRACT

The aim of this paper is to model long-term evolution of an action video with temporal multi-scale representation. This task is tough due to huge intra-class variations in motion speed. Most of the existing methods consider evolution modeling and multi-scale feature fusion in two separated phases, which generates sub-optimal representation. To address this issue, this paper proposes a novel method to integrate the evolution modeling and multi-scale representation into a unified framework. The core idea is to introduce a temporal multi-scale smoothing vector, which is used to define how the representations at different temporal scales are combined together for frame smoothing. By formulating the smoothing vector learning, evolution modeling and classifier training jointly, our method can learn a discriminative and flexible representation of multi-scale rather than a single scale or a fixed multi-scale smoothing. Experimental results on three datasets demonstrate the effectiveness of our method.

1. Introduction

Action recognition in videos [1] is one of the most intensively studied problems in computer vision field. It has huge application potential in many vision tasks, e.g., human-computer interaction, sports video analysis and smart home, etc. An action can be regarded as a diverse range of small action primitives with temporal structure. If these distinctive temporal structures in actions are well modeled, the aim of action recognition can be achieved effectively. However, this task is challenging because of the huge intra-class variations in human pose, view angle, spatio-temporal scale and motion speed.

In the literature, three main forms of modeling temporal structure have emerged: (1) spatio-temporal feature extraction [2–5], (2) spatio-temporal part mining [6–10] and (3) spatio-temporal structure learning [11–15]. One of the disadvantages of these methods is that they cannot model the long-term temporal evolution of action video. Recently, Fernando et al. [15,16] proposed the method of rank pooling to capture the latent structure of the video sequence data by considering video-level dynamics. The main idea of their work is to model the long-term evolution of a whole video by a support vector regression (SVR) whose functional parameters are used as the video representation. Rank pooling does not take into account temporal multi-scale representation since it uses all frames to model the chronological ordering between frames. However, it is apparent that the motion speeds vary according to different actors. Thus, the long-term evolution modeling at one single temporal scale is incapable of covering a full range of action speeds and tends to lose information at other temporal scales. In short, it is crucial to model the long-term evolution at multiple temporal scales.

The hierarchical rank pooling, a multi-scale rank pooling scheme, has been proposed in [17], where rank pooling is first conducted on multiple overlapping video segments, then rank pooling is recursively applied on the obtained segment descriptors. This method considers evolution modeling and multi-scale feature fusion in two separated phases, and therefore generates sub-optimal representation. To cope with this problem, we explore multi-scale evolution modeling and integrate the evolution modeling and multi-scale representation into a unified framework. The core idea of our method is to adopt a temporal multi-scale smoothing vector for frame smoothing to define how the representations at different temporal scales are combined together. Our method is inspired by Multi-skip Feature Stacking (MIFS) in [18], where features are extracted simultaneously at several temporal scales from videos. In contrast, by conducting multi-scale smoothing in

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

[*] Corresponding authors at: School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.

E-mail addresses: tingweiwang@ujn.edu.cn (T. Wang), chuancailiu@mail.njust.edu.cn (C. Liu), ltwang@hhu.edu.cn (L. Wang), ise_mabx@ujn.edu.cn (B. Ma).

https://doi.org/10.1016/j.jvcir.2018.08.014
1047-3203/© 2018 Published by Elsevier Inc.

the pooling stage, our method achieves multi-scale modeling in a higher level.

Our method learns a flexible smoothing scheme, avoiding the scale information lost by only one temporal single scale or a fixing multi-scale smoothing. At the meantime, the temporal multi-scale smoothing vector is endowed with discriminative power by a joint training scheme, where we formulate the temporal multi-scale smoothing vector learning, evolution modeling and classifier training into a unified framework and optimize this joint learning problem by an alternative optimization method.

Our contributions in this paper include the following aspects.

(1) By combining the representations at multiple temporal scales together, our method can model various temporal dynamics embedded in videos. In contrast, the original rank pooling models the dynamics of videos only at one single scale.
(2) We integrate the multi-scale evolution modeling into a unified optimization problem, where two structural risk minimization, regression structural risk and classification structural risk, are considered together. Thus, the representations of our method are endowed with discriminative ability.
(3) The multi-scale evolution is modeled in the pooling stage, thus the dimensions of our representation are as same as the original one. This characteristic of compact representation makes our method well suitable for large-scale video action recognition.

The rest of this paper is organized as follows. Section 2 reviews the work related to our methodology. In Section 3, we present our approach in detail, including rank pooling revisited, objective function, optimization and computational complexity analysis. Experimental results are reported in Section 4. Finally we conclude this paper with Section 5.

2. Related work

2.1. Spatio-temporal feature extraction

Much work has been done by using temporal structure for feature extraction task. In [2], 3DSIFT was proposed to depict the gradient magnitude and orientation in 3D instead of 2D space, and weighted histograms are constructed in the 3D neighborhood around interest points. Kläser et al. [3] proposed HOG3D basing on pure spatio-temporal 3D gradients. This feature representation method is robust and cheap while avoiding the problem of singularities by employing regular polyhedrons. Recently, dense trajectory feature was proposed in [4,5], where feature points are tracked in videos based on optical flow after they are densely sampled in each frame. Additionally, motion boundary histogram (MBH) is used to suppresses certain camera motions.

Due to its superior performance, dense trajectories are also integrated into several other methods to capture the motion information more effectively. Ordered trajectories were proposed in [19] to search for trajectories with a longer duration. In [18], the method of MIFS employed a family of differential filters, which are parameterized with multiple time skips, to capture information at multiple temporal scales. Wang et al. proposed trajectory-pooled deep-convolutional descriptors in [20], where convolutional neural networks (CNN) features are aggregated in a 3D volume around a trajectory. In [21], trajectories from action videos are transformed into a sequence of trajectory texture images which are then fed to a CNN to produce macroscopical representation of motion.

To make the extracted features invariant to illuminations, positions, scales [2,3] or location drifting [4], a common characteristic of the above low-level methods is that only local features are extracted in limited sampling neighborhood (e.g. the default 15 frames in [4]). This means that these methods can only capture the evolution in short term rather long one. However, the goal of our method is to model the dynamics in the longest term, i.e., the whole video.

2.2. Spatio-temporal part mining

Data mining technology has also been extensively studied for the discovery of meaningful spatio-temporal patterns [6]. In [22], a variant of AdaBoost training algorithm is introduced to perform the discovery of discriminative mid-level features, each of which is a combination of a set of motion directions at various locations. Jung and Hong [23] used prefixspan algorithm to mine partial sequential patterns, i.e., sequencelets, among primitive actions, and modeled human action by a Bag-of-Sequencelets model which is an ensemble of sequencelets. In [9], exemplar-SVM is employed to learn patch-specific discriminative distance metrics, and a smaller dictionary is selected from the candidate patches according to two criteria: appearance consistency and purity. Wang et al. proposed motion atom and phrase to represent mid-level temporal parts of action in [8], where the phrase is defined as an AND/OR structure on a set of motion atom units. This method utilized Apriori algorithm to mine the representative and discriminative phrases. Hierarchical, spatial and temporal relationships are modeled by a tree structure in [10], and a subgraph mining algorithm is used to mine frequent trees. Zhu et al. [7] proposed a key volume mining deep framework to identify key volumes and conduct classification simultaneously, and utilized “stochastic out” operation for pooling to select higher response volumes with higher probabilities.

After spatio-temporal parts are mined, sum-pooling or max-pooling is often employed for video representations. Therefore, those spatio-temporal relationships among parts are discarded, which limits the representative capacity of mid-level parts, while our method can make full use of the relationships among video frames.

2.3. Spatio-temporal structure learning

To predict the action class from a partially observed video, Kong et al. [24,25] proposed a discriminative multi-scale kernelized model, which captures temporal dynamics of human actions at two temporal scales. Local templates and global templates are used to consider the sequential nature of human actions and capture the history of action information, respectively. Tang et al. [11] proposed a variable-duration HMM to model transitions between states of a video. Barrett and Siskind [26] simultaneously trained the HMM and associated detectors which match the most distinctive temporal subsequences of action video. The appearance filters of key-pose and corresponding temporal location distributions are learned in [13] by a hidden CRF model which is trained by using max-margin criteria. In [12], the compatibility among part filters, part labels and class labels is modeled by a max-margin hidden CRF. By taking both sub-volumes for pooling and human pose types as hidden variables, Ni et al. proposed an adaptive motion feature pooling scheme in [27].

To make inference procedure solvable, most of these methods consider only low-order (e.g., pair-wise compatibility between two neighbor regions [12]) rather than high-order relationship. Besides, since the hidden variable are introduced to represent the discriminative parts, another disadvantage of these methods is that