Neurocomputing 173 (2016) 364–372

Contents lists available at ScienceDirect

Neurocomputing

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

Cluster trees of improved trajectories for action recognition

Quan-Qi Chen*, Yu-Jin Zhang

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 30 June 2014
Received in revised form
13 February 2015
Accepted 19 March 2015
Available online 2 September 2015

Keywords:
Action recognition
Improved trajectories
Fisher vector
Cluster trees
Dimensionality reduction

ABSTRACT

Recently, as an efficient representation of realistic videos, improved trajectory features (ITF) combined with Fisher vector (FV) encoding achieved state-of-the-art results on four challenging datasets concerning action recognition. However, directly integrating it with simple spatio-temporal pyramid (STP) will result in performance degradation. Therefore, in this paper, a novel version of cluster trees model is proposed to improve recognition performance by taking into account spatio-temporal relationships between local trajectory features. We modified and improved cluster trees model to reduce noisy clusters and alleviate intra-class variation. A further advantage of the proposed method is significantly reducing memory storage and computation time by conduct dimensionality reduction on Fisher vectors. Finally, an adaptive kernel is proposed to efficiently compare the variable-size tree representations of two videos for action recognition, which mitigates the risk introduced by noisy cluster tree nodes. Experimental results on four challenging action datasets (i.e., Olympic Sports, Hollywood2, HMDB51 and UCF50) demonstrate the effectiveness and robustness of the proposed approach which outperforms the current state-of-the-art.

1. Introduction

Action recognition in realistic videos has emerged as an active topic in the community of computer vision, whereas it remains challenging due to multiple practical issues, encompassing background clutter, occlusion and view point shifts. Besides, fast camera motion and large intra-class variation also raise severe challenges for action recognition, since significant motion clutter in the background is caused by camera motion. Furthermore, action recognition faces serious computational burden for large-scale video datasets collected from movies [1,2] and web videos [3–5].

A variety of methods have been employed to address this task, among which excellent performance has been observed for local space-time features extracted in videos, since they avoid the tedious steps of pre-processing, such as tracking and segmentation. Among the local space-time features, dense trajectory features (DTF) [6] outperform previous methods on several benchmark datasets. Furthermore, improved trajectory features (ITF) are proposed in [7] by dealing with camera motion and achieve the previous state-of-the-art result.

Recently, high-dimensional features encoding methods such as vector of local aggregated descriptors (VLAD) [8] and Fisher Vector (FV) [9] prove to be the most effective for object recognition and are gradually gaining popularity in video representation [10–12]. Several encoding methods are evaluated in [11] for action recognition, where FV which extends BoF representation by encoding both first and second order statistics reveals the best performance. FV encoding based on ITF exhibits the previous best results on four challenging datasets [7]. However, FV encoding, analogous to BoF, is unable to capture the complex spatio-temporal relationships, since it encodes all trajectory features into a single representation by examining their occurrences, and completely ignores the spatio or temporal structures among these trajectories.

A complementary line of work has focused on sophisticated models that explicitly capture spatio-temporal relationships between local features. Several methods represent actions as a fixed number of cluster parts [13], or a graphical model decomposition [14], which needs bounding box annotations, or video segmentation techniques. To organize the motion components in an unsupervised way, binary cluster tree model is presented in [15], leading to a specific decomposition structure for each video, which is data-driven. However, they applied BoF representation to encode the trajectory features of each cluster tree node, which is not powerful enough since BoF only cares about the counts of different visual words. Besides, noisy clusters often appear in the cluster tree nodes and the number of trajectories extracted from

* Corresponding author.
E-mail addresses: quanqi.chen@gmail.com (Q.-Q. Chen), zhang-yj@mail.tsinghua.edu.cn (Y.-J. Zhang).

http://dx.doi.org/10.1016/j.neucom.2015.03.124
0925-2312/© 2015 Elsevier B.V. All rights reserved.

different videos belonging to the same action class varies a lot, which leads to a huge intra-class differences.

In this paper, a novel cluster tree model method for action recognition is proposed. In the first place, as inspired by the success of the aforementioned methods, we realized that superior recognition results can be achieved by infuse the idea of modeling spatio-temporal structures into the powerful FV encoding scheme. Nevertheless, directly applying the FV encoding on the original cluster tree model is infeasible, since too many cluster notes will be generated by the latter module, and FV as well, will lead to a high dimensional representation. To mitigate these issues, we made several modifications to the original cluster tree model to reduce noisy clusters and alleviate intra-class variation. Specifically, in the process of building cluster tree, only the trajectories belonging to the largest connected component are preserved for the next bi-partition and two adaptive threshold values are utilized to limit minimum and maximum leaf sizes. Such simple adaptations not only reduce the number of nodes, but also endow our model with enhanced robustness.

Furthermore, the high dimensionality involved in FV representation makes them prohibitively computational demanding for practical use, especially in large-scale scenarios. To enhance the efficiency and scalability of the proposed method, Principal Component Analysis (PCA) is applied to reduce the dimensionality of FV, which significantly reduces memory storage and computation time.

Finally, to mitigate the risk introduced by noisy clusters, an adaptive kernel which only accounts the largest similarity cluster node pair is proposed to efficiently compare variable-size tree representations of two videos.

The remainder of this paper is organized as follows. Section 2 reviews some previous work related to this paper. Section 3 presents our action recognition framework in detail. Experimental setup and results are described in Section 4. Finally, conclusions are given in Section 5.

2. Related works

In the past decade, local space-time features have enjoyed the most attention and made significant progress in action and event recognition. Many classical image features have been extended to videos, such as 3D-SIFT [16], HOG3D [17] and extended SURF [18]. A robust CoHOG feature is proposed in [19] for human detection. A recent evaluation of these local spatio-temporal features for action recognition is given in [20], where dense sampling at regular positions in space and time outperforms the space-time interest points (STIP). Among the local space-time features, DTF and motion boundary histogram (MBH) descriptors [6] significantly outperform the previous state-of-the-art results on nine benchmark datasets, especially on real-world videos containing substantial camera motion. DTF can be further improved by taking into account camera motion. The DCS descriptors based on compensated optical flow are designed to increase accuracy for action recognition in [21]. ITF [7] features are proposed by estimating camera motion through computing a homography between frames with RANSAC, which aims to cancel out camera motion from the optical flow.

Once local features are extracted, a BoF representation of these features can be directly obtained for SVM classification. Whereas recently high dimensional features encoding methods such as VLAD [8] and FV [9] prove to be the most effective for object recognition and are gradually gaining popularity in video representation [10–12]. An analysis of several desired properties of FV as well as an extensive evaluation of parameters is presented in [10]. A variety of encoding methods are evaluated in [11] for action recognition, where FV shows the best performance, and VLAD achieves significant improvement on Hollywood2 and HMDB51 datasets over BoF [21]. A proper video encoding scheme [12] built on VLAD can achieve further improvement with negligible computational cost. FV of improved trajectories exhibits the previous best results on four challenging datasets [7].

A complementary line of work has focused on sophisticated models that explicitly capture spatio-temporal structure between local features [22–24]. In [3], action is represented as temporal compositions of motion segments, and a discriminative model is trained for classification by exploiting the temporal structure of human action. As a temporally structured extension of the BoF, the work of [22] organizes the temporal structure of actions as a sequence of histograms of atomic action-anchored features. Meanwhile, binary cluster tree [23] is utilized to model motion components as well, by representing each video as a hierarchy of mid-level motion components, which is data-driven and specific to each video. A recent version of binary cluster tree is described in more detail in [15]. To quantize relative spatio-temporal relationships between pairs of features, a pairwise model is proposed in [24]. Moreover, graphical model structures [14,13] are leveraged to explicitly model multi-scale parts and their spatio-temporal relationships. The method in [14] casts videos into spatio-temporal graphs, with the nodes corresponding to multi-scale video segments, and edges capturing their spatio-temporal relationships. Salient spatio-temporal structure [13] is extracted by forming clusters of trajectories that serve as candidates for the parts of an action. Appearance and motion constraints for the individual parts and pairwise constraints for the spatio-temporal relationships between them are incorporated by a graphical model. A Dynamic Structure Preserving Map (DSPM) [25] is proposed to learn spatio-temporal correlations from feature sets and preserve the intrinsic topologies characterized by different human motions. In [26], a distributed object detection framework (DOD) is proposed by utilizing spatio-temporal correlation, where the process of feature extraction and classification is distributed in the current frame and several previous frames. Saliency-mapping algorithm is employed in [27] to find informative regions and descriptors corresponding to background are pruned. A flexible sliding window method is proposed in [28] where the areas of potential object instances are densely sampled and the areas of non-objects are sparsely sampled.

3. Proposed approach

In this section, we elaborate our framework for action recognition by integrating FV encoding into the cluster tree model. The flowchart of the proposed method is illustrated in Fig. 1. First of all, improved trajectories are extracted as low-level features. Then we divisively cluster the obtained local features of each video into a binary tree, by representing each node as a FV vector. Meanwhile, to endow the proposed method with enhanced efficiency and scalability, PCA is applied to reduce the dimensionality of FV before storing them. Finally, action classification is performed by efficiently comparing cluster trees with an adaptive kernel.

More specifically, we first describe our feature extraction module in Section 3.1. Section 3.2 explains the feature encoding pipeline employed in our experiments. We, then, discuss how we explore the spatio-temporal relationships between trajectory features in Section 3.3. Section 3.4 presents an exposition of our motivation to introduce dimensionality reduction for FV vectors. Finally, the influences of kernels between FV-trees on classification are investigated in Section 3.5.