Neurocomputing 201 (2016) 1–11

Contents lists available at ScienceDirect

Neurocomputing

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

Learning semantic context feature-tree for action recognition via nearest neighbor fusion

Tongchi Zhou, Nijun Li, Xu Cheng, Qinjun Xu, Lin Zhou, Zhenyang Wu

School of Information Science and Engineering, Southeast University, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 22 March 2015
Received in revised form
2 March 2016
Accepted 27 April 2016
Available online 6 May 2016

Keywords:
Feature-tree
Semantic context
Nearest neighbor fusion
Action recognition

ABSTRACT

The spatio-temporal context learnt by the traditional methods for action recognition lacks the semantic meanings and temporal relationships. In order to deal with the drawbacks, we propose a novel semantic context feature-tree model to model the video clip for efficient human action recognition. The proposed method enforces spatio-temporal interest points (STIPs) within an irregular spatio-temporal volume to construct a semantic trees-structured relationship by nearest neighbor fusion. Specifically, we firstly extract STIPs, and moreover, utilize super-pixels to segment the motion image obtained by STIP detection. The points, which fall into super-pixel, are viewed as spatial semantic co-occurring features to represent one body part. Secondly, by patch matching, the point sets which are temporal nearest neighbors are merged into a new node of the next layer to describe the context of one moving part. After matching and associating, the sets of the frame indexes are renewed to group for the next fusion process until the conditions of the recursive process do not satisfy. Using KTH, UCF-YouTube and HOHA action datasets for human action recognition, our representation based on the learnt tree-structured features enhances the discriminative power of action descriptor, and obtains promising results.

1. Introduction

Human action recognition has wide applications in abnormal behavior detection, video copy detection, video retrieval, video surveillance and human computer interface [1–3]. Its research is one of the most popular topics in the field of computer vision and pattern recognition. Over the past few years, although, the methods based on STIPs together with bag-of-visual words (BoVWs) [2–6] have shown promising recognition performance, there still remains a challenging problem because of some factors about cluster background, occlusions, camera movement, and subject with different size, appearance and pose.

Prior efforts for action recognition mainly focus on STIP extraction and its descriptor, because extracting STIP does not require person detection results and is robust to static background clutter, viewpoint and scale change. In recent years, many strategies have been proposed to make video representation more discriminative, such as spatio-temporal context [6–16], local descriptor encoded as a robust representation [17–20], probabilistic latent semantic analysis based learning global semantic tops [21,22] and trajectory-based representation [6,23,24]. In general, a research trend in the field of action recognition has been the emergence of techniques based on co-occurrence statistics to enhance the discriminative power of action prototypes. Usually, the learnt context describes the volumetric region which is predefined balls with fixed sizes or kernels with various shapes [6,7,9,12]. These approaches are based on an assumption that the appropriate neighborhood size is known or uniform. Intuitively, Euclidean distance metric difficultly determines whether interest points belong to the same part or not. Therefore, the co-occurrence statistics by these methods ignore the semantic meanings and lack the discriminating capability.

We observe that human action in video clip organizes the moving processes of body parts with the spatial or temporal configuration. For the similar actions, like the “jog” and “run”, they may be easily confusing if we do not take into account their temporal layouts of local features. Also, without considering multiple temporal scales, the actions, such as “jumping”, “spiking” and “shooting” in UCF-YouTube (UCF-YT) action dataset would be difficult to distinguish, because the same primitive actions of each player are shared at the finest scale, but differ in the compound of these prototypes at a coarser scale. For instance, like “box” action shown in Fig. 1, the limb, head and leg of the performer need to follow the strict temporal orders and spatial layouts in order to convey accurate action meanings. For the moving limbs, the action in the frames of 1, 3, and 5 has the temporal relationships. In 3rd frame, three STIPs around the limbs that belong to the body part have the semantic spatial co-occurrence relationships. Thus, learning co-occurrence statistics within the semantic volumetric region at different temporal scales can enhance the discriminative power of the action representation.

http://dx.doi.org/10.1016/j.neucom.2016.04.007
0925-2312/

Fig. 1. Example action that organizes the body part with the spatial-temporal layouts.

According to the changes of the body movement, several methods have been proposed to model the temporal evolution in action (e.g. trajectory features [6,23,24], geometric temporal context [14], and the parameters of ranking function as video representation [25,26]). In this paper, in order to capture dependencies of STIPs, a feature-tree, which is a "geometric temporal structure" model, is adopted to capture the geometric context at different temporal granularities. In this hierarchical model, our main motivation in designing a feature-tree is more important uses of semantic spatial co-occurrences and temporal scales. The second layer nodes in feature-tree that are the local spatial co-occurrence point sets stand for body parts, and the high nodes capture the contextual information with different temporal scales to represent the changes of body parts.

To model the spatial semantic co-occurrence, we harness super-pixel [27] to obtain the moving part and constrain STIPs to belong to it. To learn the contextual information at the different temporal scales, we perform to fuse the temporal neighbor nodes in a hierarchical way and store them in the complete hierarchy structure. Generally, our method has several advantages. (1) Spatial semantic co-occurrence enhances the discriminative power of the STIPs; (2) in addition to the nodes of the leaf and spatial co-occurrence, nodes of the other layers (high nodes) capture the semantic context at multiple temporal scales; (3) the relationships of nodes at different layers are built by a tree-structured model; (4) the method to match nodes, which utilizes the Sum of Squared Difference (SSD) matching algorithm [28] to establish the relation of each pair of nodes and learn high nodes, is simple and effective. For representing action, we encode the nodes at each layer by Sparse Coding (SC) algorithm [19,20,29,30], then adopt the max-pooling operation to represent video content. To fuse the features at some layers, the two types of early fusion methods [31], Multiple Kernel Learning (MKL) [6,11,31] and concatenated representation, are utilized. We evaluate our method on publicly available action datasets, the KTH, UCF-YT and Hollywood (HOHA), and our method obtains excellent recognition performances. In summary, our contributions include three-fold.

(1) In order to capture the co-occurrence statistics at multiple temporal scales and build their relationships, we build multiple feature-trees by a recursive manner to model each video clip.
(2) We adopt super-pixel to segment the salient motion images and obtain STIP sets that have semantic spatial co-occurrence domains as co-motion body parts.
(3) High nodes that are generated by fusing the low layer associated nodes in temporal neighbor are utilized to describe the context of irregular volumetric region of body parts.

The remainder of this paper is organized as follows. In Section 2, we discuss the related work. In Section 3, the tree-structured context extraction model is presented in detail. The encoded nodes and combined features are stated in Section 4. The experimental results are presented and discussed in Section 5, followed by the conclusion and future work in Section 6.

2. Related work

Numerous STIP based extracting methods have been proposed in the human action recognition community. 3D Harris detector proposed by Laptev in [4] extends the idea of the Harris interest point detector to the spatio-temporal domain and detects the regions having high intensity variation in both space and time as spatio-temporal corners. However, the interest points obtained by this detector are too sparse to characterize many complex videos. Later several other methods for extracting STIPs have been proposed in [5,32–34]. Instead of Gaussian filter in temporal domain, Dollar et al. [5] apply temporal Gabor filters and select regions of high responses to improve the sparse STIP detector. Milanfar et al. [34] extend 2D Local Steering Kernels (2D LSKs) used for object detection to space-time local steering kernels introduced to capture both spatial and temporal geometric structure for action recognition. For the local spatio-temporal feature, many types of descriptors, like HOG [3,4], HOF [3,4,16,35], HOG3D [3,8], 3D Daisy [34], Cuboids [5,34], 3D LSKs [34], and 3D SIFT [37], have been proposed. Among these descriptors, one of the most commonly used in literature, which shows good performance over the various datasets, is the concatenation of Histogram of Oriented Gradients (HOG) and Histogram of Oriented Flow (HOF) [5,35].

To enhance the discriminative power of local features, co-occurrence relationships [7,10,11,13] are proposed to describe the large scale action prototype. Hu et al. [7] propose the descriptive video-phrases and the descriptive video-cliques over the regular Spatio-Temporal Volume (STV) to describe the characteristics of the action at different granularities. Likewise, Bilinski et al. [11] learn neighborhood co-occurrence statistics with the different geometrical arrangements by the rank metric for action recognition.

In order to deal with action style variant, some researchers consider multiple context information for modeling action representation. Wang et al. [16] present a novel representation that captures the contextual interactions between the interest points. This approach utilizes the density observed in each interest point's multi-scale spatio-temporal contextual domain. Zhen et al. [38] propose the multiresolution analysis techniques based spatio-temporal Laplacian pyramid coding for holistic representations. The Laplacian pyramid decomposes spatio-temporal volumes into different levels, then a bank of 3-D Gabor filters is used to extract salient features of each scale. According to the metric of the ranks and distance values, Yuan et al. [14] model the geometric temporal context with directional pyramid co-occurrence of visual words, respectively. Li et al. [39] learn STIP occurrence sequences over the whole video volume and use the motion context derived from STIPs to train Genetic search based random forests for action recognition. Kovashka et al. [8] exploit the hierarchical BoVW model by the rank metric to represent the spatio-temporal layouts at different scales. However, the learnt compound features lack the robustness to the body part and do not capture the changes of each body part. In order to learn semantic context and model the relationships among them, Trichet et al. [13] learn the spatio-temporal context based on the multiscale segmentation volume. The boundaries of segmentation are used to model the co-occurrence domains and hierarchical relationships. Liu et al. [40] adopt the hierarchical partwise bag-of-visual words to encode both local and global visual saliency based on the body structure cue, then utilize part-regularized multitask structural learning to discover both action-specific and action-shared feature subspace. Unlike [8], Brendel et al. [41] adopt agglomerative clustering to initialize a graph, then learn a weighted least-squares graph of the given activity class. The learnt nodes correspond to the multiscale video segments, and the edges capture the relationships of two parts. Wang et al. [35] employ a set of the grammar rules to