Available online at www.sciencedirect.com

ScienceDirect

CAAI Transactions on Intelligence Technology 1 (2016) 14–29
http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology/

Salient pairwise spatio-temporal interest points for real-time activity recognition

Mengyuan Liu [a], Hong Liu [a,b,*], Qianru Sun [a], Tianwei Zhang [c], Runwei Ding [a]

[a] *Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Peking University, Shenzhen Graduate School, 518055, China*
[b] *Key Laboratory of Machine Perception, Peking University, 100871, China*
[c] *Nakamura-Takano Lab, Department of Mechanoinformatics, The University of Tokyo, 113-8685, Japan*

Available online 8 March 2016

Abstract

Real-time Human action classification in complex scenes has applications in various domains such as visual surveillance, video retrieval and human robot interaction. While, the task is challenging due to computation efficiency, cluttered backgrounds and intro-variability among same type of actions. Spatio-temporal interest point (STIP) based methods have shown promising results to tackle human action classification in complex scenes efficiently. However, the state-of-the-art works typically utilize bag-of-visual words (BoVW) model which only focuses on the word distribution of STIPs and ignore the distinctive character of word structure. In this paper, the distribution of STIPs is organized into a salient directed graph, which reflects salient motions and can be divided into a time salient directed graph and a space salient directed graph, aiming at adding spatio-temporal discriminant to BoVW. Generally speaking, both salient directed graphs are constructed by labeled STIPs in pairs. In detail, the "directional co-occurrence" property of different labeled pairwise STIPs in same frame is utilized to represent the time saliency, and the space saliency is reflected by the "geometric relationships" between same labeled pairwise STIPs across different frames. Then, new statistical features namely the Time Salient Pairwise feature (TSP) and the Space Salient Pairwise feature (SSP) are designed to describe two salient directed graphs, respectively. Experiments are carried out with a homogeneous kernel SVM classifier, on four challenging datasets KTH, ADL and UT-Interaction. Final results confirm the complementary of TSP and SSP, and our multi-cue representation TSP + SSP + BoVW can properly describe human actions with large intro-variability in real-time.

Keywords: Spatio-temporal interest point; Bag-of-visual words; Co-occurrence

1. Introduction

Recently, human action classification from video sequences plays a significant role in human–computer interaction, content-based video analysis and intelligent surveillance, however it is still challenging due to cluttered backgrounds, occlusion and other common difficulties in video analysis. What's worse, intro-variability among the same type of actions also brings serious ambiguities. To tackle these problems, many human action classification methods based on holistic and local features have been proposed [1,2]. Holistic features have been employed in Refs. [3–5], where actions were treated as space–time pattern templates by Blank et al. [3] and the task of human action classification was reduced to 3D object recognition. Prest et al. [4] focused on the actions of human–object interactions, and explicitly represented an action as the tracking trajectories of both the object and the person. Recently, traditional convolutional neural networks (CNNs) which are limited to handle 2D inputs were extended, and a novel 3D CNN model was developed to act directly on raw videos [5].

* Corresponding author. G102-105, School of Computer & Information Engineering Peking University, Shenzhen University Town, Xili, Nanshan District, Shenzhen, Guangdong Province, China. Tel.: +86 (0755)2603 5553.
E-mail addresses: liumengyuan@pku.edu.cn (M. Liu), hongliu@pku.edu.cn (H. Liu), qianrusun@sz.pku.edu.cn (Q. Sun), zhangtianwei5@gmail.com (T. Zhang), dingrunwei@pkusz.edu.cn (R. Ding).
Peer review under responsibility of Chongqing University of Technology.

http://dx.doi.org/10.1016/j.trit.2016.03.001

Comparing with holistic features, local features are robust to shelters which need no pre-processing such as segmentation or tracking. Laptev [6] designed a detector which defines space–time interest points (STIPs) as local structures where the illumination values show big variations in both space and time. Four later local feature detectors namely Harris3D detector, Cuboid detector, Hessian detector and Dense sampling were evaluated in Ref. [7]. Recently, dense trajectories suggested by Wang et al. [8] and motion interchange patterns proposed by Kliper-Gross et al. [9] have shown great improvement to describe motions than traditional descriptors though both need extra computing costs. Besides using content of local features, researches only using geometrical distribution of local features also achieve impressive results for action classification. Bregonzio et al. [10] described action using clouds of Space–Time Interest Points, and extracted holistic features from the extracted cloud. Ta et al. [11] concatenated 3D positions of pairwise codewords which are adjacent in space and in time for clustering. A bag of 3D points was employed by Li et al. [12] to characterize a set of salient postures on depth maps. Yuan et al. [13] extended R transform to an extended 3D discrete Radon transform to capture distribution of 3D points. These methods assume that each local feature equals to a 3D point, and all local features have the only difference of location.

Bag-of-visual words (BoVW) introduced from text recognition by Schuldt et al. [14] and Dollar et al. [15] is a common framework to extract action representation from local features. STIPs are firstly extract from training videos and clustered into visual words using clustering methods. BoVW is then adopted to represent original action by a histogram of words distribution, and to train classifiers for classification. Despite its great success, BoVW ignores the spatio-temporal structure information among words and thus leads to misclassification for actions sharing similar words distribution. To make up for above problem of BoVW, the spatio-temporal distribution of words is explored. Words are treated ***in groups*** to encode spatio-temporal information in Refs. [16–18]. Latent topic models such as the probabilistic Latent Semantic Analysis (pLSA) model are utilized by Niebles et al. [16] to learn the probability distributions of words. Cao et al. [17] applied PCA to STIPs, and then model them with Gaussian Mixture Models (GMMs). A novel spatio-temporal layout of actions, which assigns a weight to each word by its spatio-temporal probability, was brought in Ref. [18]. Considering words ***in pairs*** is an effective alternative to describe the distribution of words. From one point of view, pairwise words which are adjacent in space and in time were explored by Refs. [11,19,20]. Local pairwise co-occurrence statistics of codewords were captured by Banerjee et al. [19], and such relations were reduced using Conditional Random Field (CRF) classifier. Savarese et al. [20] utilized spatial-temporal correlograms to capture the co-occurrences of pairwise words in local spatio-temporal regions. To represent spatio-temporal relationships, Matikainen et al. [21] formulated this problem in a Nave Bayes manner, and augmented quantized local features with relative spatial-temporal relationships between pairs of features. From another point of view, both local and global relationships of pairwise words were explored in Refs. [22,23]. A spatio-temporal relationship matching method was proposed by Ryoo et al. [22] which explored temporal relationships (e.g. before and during) as well as spatial relationships (e.g. near and far) among pairwise words. In Ref. [23], co-occurrence relationships of pairwise words were encoded in correlograms, which relied on the computation of normalized google-like distances.

In this work, the directional relationships of pairwise features are explored to make up the problems of BoVW. It is observed that human actions make huge senses in the directional movement of body parts. From one aspect, the spatial relationships among different parts, which are moving at the same time, are directional. Besides, one part keeps directionally moves from one place to another. Here, a "push" action in Fig. 1 is used to illustrate observations, where green points denote local features. As shown in Frame $t + 1$, the pusher's hands and the receiver's head are moving at the same; meanwhile, the vertical location of hands is lower than the head. The relationship between this type of pairwise motions, which is according to the first observation, is called directional co-occurrence. Crossing from Frame $t - 1$ to Frame t, the pusher's hands keep moving forward. This type of pairwise motions are also directional and reflect the second observation. The observations both indicate the importance of directional information for action representation. Hence the attribute of mutual directions are assigned to pairwise STIPs to encode structural information from directional pairwise motions, generating new features called Time Salient Pairwise feature (TSP) and Space Salient Pairwise feature (SSP).

1.1. Time Salient Pairwise feature

Time Salient Pairwise feature (TSP) is formed from a pair of STIPs which shows "directional co-occurrence" property. In our previous work, [24] and [25] have already employed this property for action recognition. The TSP mentioned in this paper is a refined and expanded version from the conference proceedings paper [24]. TSP is compared with traditional

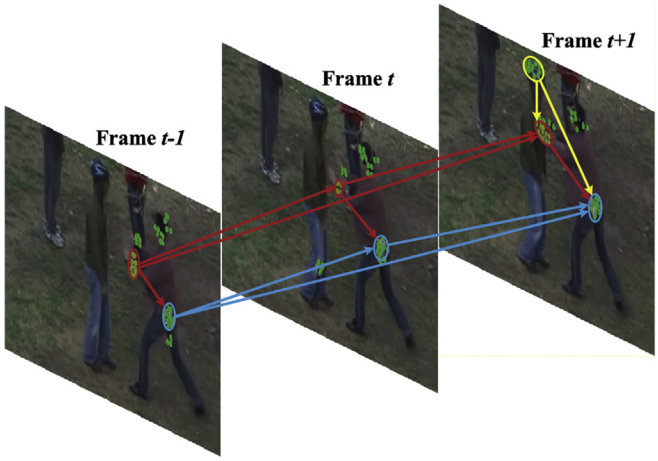

Fig. 1. A "push" action performed by a "pusher" and a "receiver".