Information Sciences 281 (2014) 295–309

Contents lists available at ScienceDirect

Information Sciences

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

Action recognition by spatio-temporal oriented energies

Xiantong Zhen [a,b], Ling Shao [a,b,*], Xuelong Li [c]

[a] College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, Jiangsu 210044, PR China
[b] Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S1 3JD, UK
[c] State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi 710119, PR China

ARTICLE INFO

Article history:
Received 13 September 2013
Received in revised form 7 May 2014
Accepted 14 May 2014
Available online 2 June 2014

Keywords:
Action recognition
Steerable filters
Spatio-temporal oriented energies
Spatio-temporal Laplacian pyramid

ABSTRACT

In this paper, we present a unified representation based on the spatio-temporal steerable pyramid (STSP) for the holistic representation of human actions. A video sequence is viewed as a spatio-temporal volume preserving all the appearance and motion information of an action in it. By decomposing the spatio-temporal volumes into band-passed sub-volumes, the spatio-temporal Laplacian pyramid provides an effective technique for multi-scale analysis of video sequences, and spatio-temporal patterns with different scales could be well localized and captured. To efficiently explore the underlying local spatio-temporal orientation structures at multiple scales, a bank of three-dimensional separable steerable filters are conducted on each of the sub-volume from the Laplacian pyramid. The outputs of the quadrature pair of steerable filters are squared and summed to yield a more robust oriented energy representation. To be further invariant and compact, a spatio-temporal max pooling operation is performed between responses of the filtering at adjacent scales and over spatio-temporal neighbourhoods. In order to capture the appearance, local geometric structure and motion of an action, we apply the STSP on the intensity, 3D gradients and optical flow of video sequences, yielding a unified holistic representation of human actions.

Taking advantage of multi-scale, multi-orientation analysis and feature pooling, STSP produces a compact but informative and invariant representation of human actions. We conduct extensive experiments on the KTH, UCF Sports and HMDB51 datasets, which shows the unified STSP achieves comparable results with the state-of-the-art methods.

1. Introduction

Human action recognition [22,2,9] has been extensively researched in computer vision. Its potential applications can be found in many areas such as visual surveillance, video indexing/retrieval, sport event analysis and human computer interaction. However, action recognition is a challenging task mainly due to difficulties including large intra-class variations (*i.e.*, the same action performed by different actors would differ significantly) and inter-class similarities (*e.g.*, 'running' and 'jogging' appear rather similar). Existing action recognition systems are mainly focused on local and holistic representations.

Local representations using sparsely detected spatio-temporal interest points (STIPs) have dominated in human action recognition in the last decade. The popularity of local methods, *e.g.*, the bag-of-words (BoW) model, results from attractive

* Corresponding author at: College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, Jiangsu 210044, PR China. Tel.: +44 1142225841.
E-mail address: ling.shao@ieee.org (L. Shao).

http://dx.doi.org/10.1016/j.ins.2014.05.021
0020-0255/

advantages, such as being less sensitive to partial occlusions and clutter and requiring no background subtraction or target tracking used in most of holistic methods. Nevertheless, local methods also suffer from some limitations, one of which is the inability to capture adequate spatial and temporal structure information of actions.

Holistic representations directly extract spatio-temporal features from raw video sequences rather than applying sparse sampling by STIP detectors. The advantage of such representations is that they are supplied with the entire spatial and temporal structural information of the human action in a sequence. However, to prevent the interference of background variations, accurate preprocessing steps such as background subtraction, segmentation and tracking are usually required.

In both local and holistic representations, low-level features serve as the fundamental role in human action representations. These features appear in the spatio-temporal volumes at arbitrary orientations and carry important features of actions, *e.g.*, appearance and motion. To detect such features, a set of oriented filter kernels are always applied to each possible orientation, which however would be computationally expensive. Oriented filters have been often used in image processing. Features based on oriented gradients have been widely used and successfully extended from the image domain into video analysis and action recognition [8,27]. In the image domain, Freeman et al. [15] proposed steerable filters to efficiently synthesize filters of arbitrary orientations for linear combinations of basis filters.

Adelson and Bergen [1] introduced a class of models for analysis of human motion mechanisms in which the first stage consists of linear filters that are oriented in space–time and tuned in spatial frequency. The outputs of the quadrature pairs of such filters are squared and summed to give a measure of motion energy. Energy models can be built from elements that are consistent with known physiology and psychophysics, and they permit a qualitative understanding of a variety of motion phenomena.

Extracting oriented spatio-temporal features based on steerable filters for video analysis has been well researched in previous works [49,10,12,11,6]. The use of steerable filters for the spatio-temporal data analysis can date back to the work by Wildes and Bergen [49]. They provided an avenue to perform qualitative analysis of spatio-temporal patterns that capture the underlying salient structures in video sequences. Local energy representations based on the quadrature outputs of the steerable filters were also used in their work, which is deemed as the foundation for analyzing spatio-temporal data using steerable filters.

By extending the two-dimensional steerable filters into three dimensions, Derpanis and Gryn [10] described the details of the construction of the *N*th derivative of Gaussian separable steerable filters in the three-dimensional space. The separable and steerable implementations lead to efficient computation of steerable filters.

In light of the previous work, the local oriented energy representations have been utilized to spatio-temporal grouping [12], efficient action spotting [11] and visual tracking [6]. Derpanis and Wildes [12] adopted the oriented energy representation for grouping raw image data into a set of coherent spatiotemporal regions. This representation describes the presence of particular oriented spatio-temporal structures in a distributed manner to capture multiple oriented structures at a given location. They further designed a descriptor based on the oriented energy measurements for action spotting [11]. Slightly different from [12], in [11] the local energies are calculated based on a third order Gaussian derivative rather than the quadrature outputs of the steerable filters.

In the same spirit, Cannons et al. [6] developed a pixel-wise spatio-temporal oriented energy representation for visual tracking. Distinguished from [12,11], a multi-scale Gaussian steerable filter was used. The representation includes appearance and motion information as well as information about how these descriptors are spatially arranged.

1.1. Motivations

Our work is motivated by the fact that a video sequence with motion could be represented as a single pattern in the X–Y–T space, in which a velocity of motion corresponds to a three-dimensional orientation in this space. Motion information can be extracted by a system that responds to the oriented spatiotemporal energy. In addition, spatio-temporal features reside in different scales and can be extracted by multi-scale analysis. The steerable filters can efficiently perform multiple orientation analysis for videos while the Laplacian pyramid provides an effective multi-scale analysis. By combining the Laplacian pyramid and steerable filters, the STSP can detect non-orthogonal and over-complete features, which also shows the desirable property of shift and rotation invariance. It is a transform that combines the multi-scale decomposition with differential measurements, capturing the oriented structures in spatio-temporal volumes.

Inspired by the success of steerable filters in object classification [32] and video analysis [49], we introduce a novel holistic representation based on the spatio-temporal steerable pyramid (STSP) for action recognition. In contrast to previous holistic methods, our method based on the STSP can to a large extent handle the deficits of holistic representations and provides an informative, compact representation of human actions.

Note that this paper is an extension of the work in [54]. In the current version, we generalize the STSP by extending it from intensity to gradients and optical flow, and comprehensive experiments on the investigation of parameter settings are conducted on more datasets.

1.2. Overview

Given a 3D volume, which in our case can be the intensity volume, optical flows and 3D gradients of a video sequence, a spatio-temporal Laplacian pyramid structure is first constructed. The volume is decomposed into a set of sub-band volumes, which can segregate and enhance spatio-temporal features residing in different scales.