Neurocomputing 151 (2015) 544–553

Contents lists available at ScienceDirect

Neurocomputing

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

Single/multi-view human action recognition via regularized multi-task learning

An-An Liu [a], Ning Xu [a], Yu-Ting Su [a], Hong Lin [a], Tong Hao [b], Zhao-Xuan Yang [a]

[a] School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China
[b] College of Life Sciences, Tianjin Normal University, Tianjin 300387, China

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form
27 January 2014
Accepted 8 April 2014
Available online 13 November 2014

Keywords:
Pyramid partwise bag of words
Multi-task learning
Graph structure
Sparsity
Human action recognition

ABSTRACT

This paper proposes a unified single/multi-view human action recognition method via regularized multi-task learning. First, we propose the pyramid partwise bag of words (PPBoW) representation which implicitly encodes both local visual characteristics and human body structure. Furthermore, we formulate the task of single/multi-view human action recognition into a part-induced multi-task learning problem penalized by graph structure and sparsity to discover the latent correlation among multiple views and body parts and consequently boost the performances. The experiment shows that this method can significantly improve performance over the standard BoW+SVM method. Moreover, the proposed method can achieve competing performance simply with low dimensional PPBoW representation against the state-of-the-art methods for human action recognition on KTH and MV-TJU, a new multi-view action dataset with RGB, depth and skeleton data prepared by our group.

1. Introduction

Recently human action recognition has obtained increasing attention owing to its wide range of applications in visual surveillance, human computer interaction, video retrieval, etc. The goal of human action recognition is to automatically analyze ongoing activities from an unknown video [1,2]. When the video only contains the progress of one action, this task is equivalent to classifying the video into the corresponding action category. In more general cases, when the test video contains continuous human activities, we need to localize both spatial and temporal positions of all actions in the video.

Considering the view information available for human action recognition, the state-of-the-art methods can be roughly categorized into two classes, single-view [3–6] and multi-view methods [7,8]. Single-view human action recognition is challenging because of the high variability of appearances, shapes and potential occlusions. Local space–time feature-based methods have demonstrated impressive levels of performance [9–11] because such features can capture local salient characteristics of appearance and motion [12]. Furthermore, they are robust to spatiotemporal shifts and scales, background clutter and multiple motions in the scene. Local space–time feature extraction usually involves two components, local feature detectors and descriptors. The detectors usually select spatiotemporal locations and scales in the video by maximizing specific saliency function. Laptev and Lindeberg [13] proposed Harris3D detector by extending the 2D Harris detector. The spatio-temporal second-moment matrix at each pixel is calculated and utilized for objective function formulation, the local maxima of which denotes the detected local point. Dollár et al. [14] proposed Cuboid detector. The response function is formulated with the filtered image by 2D spatial Gaussian smoothing kernel and a quadrature pair of temporal Gabor filters. Willems et al. [15] proposed the 3D Hessian detector by maximizing the saliency with the determinant of the 3D Hessian matrix. Since local space–time detector can pinpoint the salient positions in one video, feature descriptors [15–20] can be computed to capture the shape and motion in the neighborhoods of selected points with image operation such as spatiotemporal image gradients and optical flow. Dollár et al. [14] proposed the Cuboid descriptor which can be computed by concatenating the gradients at each pixel in one 3D volume into a single vector and then implementing dimension reduction on it with principle component analysis. Laptev et al. [20] proposed HOG/HOF descriptors by computing histograms of spatial gradient and optic flow accumulated in space–time neighborhoods of detected local saliency points. Kläser et al. [21] extended SIFT descriptor and proposed HOG3D descriptor for simultaneous representation of shape and motion by computing the histograms of 3D gradient orientations. With the detected local space–time points and their descriptors, the bag-of-words (BoW) framework [22–24] can be utilized to represent a video as a collection of local features for action classification. Because different experimental settings

http://dx.doi.org/10.1016/j.neucom.2014.04.090
0925-2312/

limited the comparison of related methods, Wang et al. [25] recently gave a comprehensive evaluation of the popular local feature detectors and descriptors for the standard BoW+SVM framework. Large scale experiments have shown promising results with this framework for human action recognition on the dataset captured under controlled environment (Weizmann [26,27], KTH [13,16,28]) and the realistic dataset (UCF Sports [29], Hollywood2 [30], Youtube [31]).

Compared with single-view human action recognition, multiple views can provide more information for human action recognition [32–34] while it is still challenging because of the difficulties in correlation discovery among multiple views [35,36]. To address this problem, epipolar geometry was employed for this task. Yilmaz and Shah [37] employed epipolar geometry for point correspondences between actions to impose fundamental matrix constraints for view-invariant action recognition. Rao et al. [38] showed that the maxima in space–time curvature of a 3D trajectory were preserved in 2D image trajectories, and therefore the 2D trajectories can be utilized to capture the view-independent representation of human actions. These methods highly depend on reliable body joints detection and correspondences which is nontrivial for dynamic human action. Another kind of methods implements 3D reconstruction to take advantage of multi-view information [39]. Gao et al. [40] proposed an effective method for view selection which will facilitate discriminative feature representation and model learning. Lv and Nevatia [41] constructed the Action Net model to represent spatial shapes of human action. Li et al. [42] reconstructed 3D model from multi-view inputs for action recognition. This approach usually needs pre-setup of multi-view cameras and consequently limits their application in practice. Different from the methods aforementioned which explicitly utilize view-based knowledge for action representation, researchers are more interested in view-independent feature learning by view-specific feature transformation [43]. For a pair of given views, Farhadi and Tabrizi [44] extracted features from respective views. Then, the proposed maximum margin clustering was implemented to generate split-based features in one view and a predictor was utilized to obtain the split-based feature for action recognition in the other view. Furthermore, the graph matching and learning methods [45–48] are popular methods to group the view-specific BoW features into visual-word clusters and thereby transformed both view-specific features into the view-independent representation. The emerging problem is that the multi-view information causes the drastic increase of data. The high computation complexity might be solved by the promising GPU or many-core computing [49].

The motivation of our work comes from three existing problems. First, the standard BoW representation has limited descriptive ability because these methods ignore all information about the structure of human body [50]. Although both the hand-crafted hierarchical features [51,52] and feature learning-based methods [9,53] achieved improvement, they cost complex computation for feature representation while cannot discover the meaningful feature subspace [54]. Second, the current methods belonging to the single-task learning have the deficiency in discovering correlation among multiple views compared to the multi-task learning framework [55]. The latent relationship is always heuristically defined [45,44] with prior knowledge. Seldom literatures work on adaptive latent correlation discovery by model learning. Third, the current methods usually treat single-view and multi-view action recognition separately. The single-view based methods usually adopt global BoW representation without the body structure information for formulation. Consequently, they always ignore discovering the latent correlation among partwise visual features. Comparatively, the correlation learning is extremely critical for multi-view methods [56].

In this paper, we propose to formulate partwise BoW representation and further turn the single-task classification into part-induced multi-task classification by discovering the underlying common knowledge for both single and multiple-view human action recognition. Our contribution lies in three-fold:

- *Pyramid partwise BoW* (*PPBoW*) *representation*: Different from the standard BoW [22] which simply computes an orderless BoW representation and its enhanced version, the spatial pyramid BoW [57], which repeatedly subdivides the image and computes histograms of local features at increasingly fine resolutions, we propose the PPBoW representation with different feature pooling strategies depending on the prior knowledge of human body structure.
- *Part-induced multi-task learning*: Since there exists intrinsic correlation among multiple body parts in different views for one action, we formulate the action recognition task as a joint multi-task learning (MLT) problem penalized by graph structure and sparsity to discover the latent correlation. Since different partwise features in multiple views have different characteristics for action representation, we propose the adaptive partwise BoW feature selection-based MTL prediction.
- *Multi-view human action dataset with RGB, depth and skeleton data*: To out knowledge, most human action datasets with RGB, depth and skeleton data are captured in single view [58,59] while there is seldom multi-view human action dataset with data in three modalities. With the rapid spread of Microsoft Kinect, we can simultaneously capture RGB/depth/skeleton data with an affordable device. The depth and skeleton data can compensate RGB data and significant benefit human action recognition. We prepared the multi-view dataset with RGB/depth/skeleton data (MV-TJU) to encourage the research on multi-view human action recognition with multi-modality information.

The comparison experiments show that the proposed method simply with low dimensional feature representation can significantly improve performance over the standard BoW+SVM method. Moreover, the proposed method can achieve competing performance against the state-of-the-art methods for human action recognition.

The rest of paper is structured as follows. In Section 2, we introduce the PPBoW representation. The regularized MTL will be illustrated in Section 3. The experimental method and results will be detailed in Sections 4 and 5 respectively. At last, conclusion are presented.

2. Pyramid partwise bag-of-words (PPBoW) representation

To overcome the inability of the standard BoW which ignores the prior knowledge of human structure for feature construction and simply computes an orderless BoW representation, we proposes the pyramid partwise bag-of-words representation by proceeding through three steps as shown in Fig. 1:

(1) *Spatiotemporal feature extraction*: Local spatiotemporal feature extraction can be fulfilled by two consecutive steps, local saliency point detection and description. Since Harris3D detector and HoG/HoF descriptor have been demonstrated to outperform other detectors and descriptors in the standard BoW+SVM framework [25], we utilized their combination in our work. Then an input video can be represented by a set of local spatiotemporal points, which will be further grouped into different categories depending on the location of different body part centers.
(2) *Part detection*: With the success of part-based model for human detection on RGB data [60] and the release of Kinect for skeleton capturing, we can localize one person in each frame with the seven parts, including head, left and limbs, left and right legs, left and right foots. Furthermore, the bounding-box of human can be conveniently localized by the root