Neurocomputing 173 (2016) 110–117

Contents lists available at ScienceDirect

Neurocomputing

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

Evaluation of local spatial–temporal features for cross-view action recognition

Zan Gao [a,b], Weizhi Nie [c], Anan Liu [c,*], Hua Zhang [a,b]

[a] *Key Laboratory of Computer Vision and System, Tianjin University of Technology, Ministry of Education, Tianjin 300384, China*
[b] *Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, China*
[c] *School of Electronic Information Engineering, Tianjin University, 300072, China*

ARTICLE INFO

Article history:
Received 20 March 2015
Received in revised form
19 June 2015
Accepted 3 July 2015
Available online 10 August 2015

Keywords:
Cross-domain learning
Human action recognition
Spatial–temporal feature

ABSTRACT

Local spatial–temporal feature-based representation is extremely popular for human action recognition. Many spatial–temporal salient point detectors and descriptors have been proposed. Although the promising results have been achieved for action recognition recently, there still exist two severe problems: (1) there is lack of systematic evaluation of local spatial–temporal features on cross-view action recognition; (2) there is lack of a baseline method especially for the task of cross-view action recognition, which can adaptively bridge different feature spaces from multiple views for this cross-domain task. In this paper, we evaluate four popular spatial–temporal features (STIP, Cuboids, MoSIFT, HoG3D) with the framework of transferable dictionary pair learning. This framework can first learn one transferable dictionary pair in both unsupervised and supervised settings. Then, training samples in the source view and testing samples in the target view can be represented with corresponding source and target dictionary respectively to get sparse feature representations, which is used to training classifier for action recognition. In this way, it can map the features from different views into the same feature space to handle the cross-domain task. The evaluation of four spatial–temporal features and the framework of transferable dictionary pair learning are implemented on the popular multi-view human action dataset, IXMAS. The comparative experiments against the representative methods further demonstrate the superiority of this framework on cross-view human action recognition.

1. Introduction

Human action recognition is a hot research topic in computer vision and pattern recognition due to its wide and significant applications in video surveillance, human computer interaction multimedia retrieval and so on. Human action recognition is always a hard work in computer vision, because human actions produce spatiotemporal patterns of appearance or motion, which lead that it is hard to extract robust feature for human action representation.

Many approaches were proposed to handle this problem [1–4]. These methods based on space–time interest points together with bag-of-words models have demonstrated impressive levels of performance. These approaches do not rely on preprocessing techniques, background modeling or body-part tracking, but also robust to spatiotemporal shifts and scales, background clutter and multiple motions in the scene. Although these approaches are quite successful in recognizing actions captured from single view, they have more limitation when the viewpoint changes. The major reason lies in the appearance of actions that may be drastically different when observed from different views. Consequently, the action models learned using samples in one view are less discriminative for recognizing action in a different view.

Recently, many approaches have been proposed to address the problem of cross-view action recognition [5,6]. Several geometry-based approaches were employed for this task. In order to impose fundamental matrix constrains for view-invariant action recognition, epipolar geometry for point correspondences between actions was proposed in Yilmaz et al. [1]. A view-invariant matching method based on epipolar geometry between actor silhouettes without tracking and explicit point correspondences was proposed in Haq et al. [3]. Rao et al. [7] 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. Although these approaches can obtain good accuracy, they highly depend on body joints detection and tracking, which are very difficult for dynamic human action. Another kind of approaches performs 3D reconstruction to take multi-view information for multi-view action recognition. The Action Net model was constructed in Lv

[*] Corresponding author.
E-mail address: anan0422@gmail.com (A. Liu).

http://dx.doi.org/10.1016/j.neucom.2015.07.105
0925-2312/© 2015 Elsevier B.V. All rights reserved.

et al. [8] to represent 3D shapes for action recognition. 3D model from multi-view inputs for action recognition was reconstructed in Li et al. [9]. To build from multi-view points to model actions, three dimensional occupancy grids were proposed in Weinland et al. [10]. To encode the shape and motion information of actors observed from multiple views, a 4D view-invariant action feature extraction was developed in Yan et al. [11]. However, since pre-setup of multi-view cameras to get strict alignment between views was required in 3D recognition, and at the same time, they also need to find the best match between a 3D model and a 2D observation over a large model parameter space. What is more, the computationally expense limits their application in practice.

Transfer learning methods are proposed for cross-view action recognition and achieve success [12–15,5]. These transfer learning based methods have emerged to adapt the action knowledge learned on one or more views (source view) to another different view (target view) by exploring the statistical connections between them. A split-based feature was developed in Farhadi et al. [12], which was learned by Maximum Margin Clustering algorithm. This split-based feature was first generated in one view, then a predictor was utilized to get the split-based feature in the other view for action recognition. However, since feature-to-feature correspondence to train a classifier to obtain split-based features at the frame-level was required, their work is computationally expensive. In order to learn bilingual-words from two individual codebooks belonging to two different views, a bipartite graph-based method was developed in Liu et al. [13] and then a bag-of-bilingual-words (BoBW) model was used to transfer action models between two views. However, only the codebook-to-codebook correspondence in two views was exploited in this method, which cannot ensure that pairs of videos taken at two views have similar feature representation. In order to derive a correlation subspace as a joint representation from different bag-of-words models at different views and incorporate a corresponding correlation regularizer into the formulation of support vector machine, canonical correlation analysis was employed in Huang et al. [14]. In order to connect action descriptors between different views, “virtual views” was proposed in Li et al. [15], where each virtual view is associated with a linear transformation of the action descriptor, and the sequence of transformed descriptors can be utilized to compare actions from different views. Zheng et al. [5] proposed the most encouraging method, where a transferable dictionary pair by forcing two sets of pair action videos in the pair views to have the same sparse representations was learned. Thus, the action model which was learned in one view can be directly applied to predict test videos in the target view. In [5], the author extracts the Cuboid feature which is a popular local space–time feature as the low-level feature. Followed by [13], the global shape-flow feature is used as a complement to the local space–time feature. Such low-level feature will then be used for the transferable dictionary pair learning as well as the view-invariant feature discovering.

In recent years, many different space–time feature detectors and descriptors have been proposed in the past few years. However, there is still lack of systematic evaluation of them on cross-view action recognition. Moreover, there is lack of a baseline method, like the BoW+SVM method for the single-view action recognition, for the cross-view action recognition. In this paper, we evaluate performance of four kinds of local space–time feature on IXMAS multi-view human action dataset: (1) the STIP feature was proposed in [16] by Laptev et al. which consists of Harris3D detector and HOG/HOF descriptor; (2) Dollar et al. proposed the Cuboid detector along with the Cuboid descriptor in [17]; (3) the MoSIFT feature is proposed by Chen et al. The MoSIFT detector finds the distinctive points in spatial domain with the well-known SIFT algorithm and then detects spatio-temporal interest points with (temporal) motion constraints. Then HOG/HOF descriptor is used to describe the detected points; and (4) for the HOG3D feature, we use the combination of Harris3D detector and HOG3D descriptor. The HOG3D descriptor was proposed by Klaser et al. It is based on histograms of 3D gradient orientations and can be seen as an extension of the popular SIFT descriptor to video sequences. Since the transferable dictionary pair learning method [5] has achieved promising results on cross-view action recognition, we select this framework as the baseline method for this task. The evaluation of four spatial-temporal features and the framework of transferable dictionary pair learning are implemented on the popular multi-view human action dataset, IXMAS. The comparative experiments against the representative methods further demonstrate the superiority of this framework on cross-view human action recognition.

The rest of the paper is organized as follows. In Section 2, we introduce the framework of testing method and four popular spatial–temporal features in brief. The baseline method will be detailed in Section 3. The experimental results will be shown in Section 4. Finally, Section 5 concludes this paper.

2. Cross-view action recognition

Although multi-task algorithms [18,19] have been proposed for multi-view action recognition or other tasks, it is different from cross-view action recognition, and in multi-task algorithms the training and testing are done on the same view, but the objective of the proposed approach is to recognize an unknown action from one target view using training data taken from one source view (or more than one source view). For pairwise views, we first learn one transferable dictionary pair using unsupervised and supervised settings. Then, we represent training samples in the source view and testing samples in the target view using corresponding source and target dictionaries respectively. We obtain sparse feature representations by using the transferable dictionary pair. These sparse feature representations are regarded as the view-invariant feature. Based on them, a classifier is trained and utilized to recognize unlabeled testing samples. Through the steps presented above, we realize the object for cross-view action recognition.

In order to evaluate the performance of this method, we select some popular local feature descriptors, such as STIP, Cuboids, MoSIFT and HoG3D. Although CNN feature [20] also can obtain good performance, since only few action samples can be obtained in our setting, but when training the CNN, we need large scale training samples to obtain good CNN parameters, thus, in our experiments, we do not evaluate the proposed algorithms by CNN features. In addition, feature selection [19,21] and feature fusion [22] can also improve the accuracy, but in our experiments, only the original performances of spatio-temporal features are assessed. In all cases we use the original implementation and parameter settings provided by the authors. The detailed introduction of these local features is as follows:

- *STIP*: The STIP feature consists of Harris3D detector and HOG/HOF descriptor. We follow the method described in [16] to extract local spatio-temporal features and used the code released by the author. The code is an extension of Harris 3D detector proposed in [23] which detects spatio-temporal points of interest. HOG/HOF descriptor are then computed to characterize the 3D spatio-temporal video patch extracted at each interest point. The 3D video patch is divided into $3 \times 3 \times 2$ cells and 4-bin histogram of gradient orientations (HOG) and 5-bin histogram of optical flow (HOF) are computed in each cell and the normalized histograms of HOG (72 dimensions) and HOF (90 dimensions) are concatenated. One thing to note is that the