Neurocomputing 191 (2016) 286–295

Contents lists available at ScienceDirect

Neurocomputing

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

Heterogeneous discriminant analysis for cross-view action recognition

Wanchen Sui, Xinxiao Wu*, Yang Feng, Yunde Jia

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, PR China

ARTICLE INFO

Article history:
Received 10 August 2015
Received in revised form
9 December 2015
Accepted 15 January 2016
Communicated by Yongzhen Huang
Available online 18 February 2016

Keywords:
Cross-view action recognition
Transfer learning
Discriminant analysis
Heterogeneous domain adaption

ABSTRACT

We propose an approach of cross-view action recognition, in which the samples from different views are represented by features with different dimensions. Inspired by linear discriminant analysis (LDA), we introduce a discriminative common feature space to bridge the source and target views. Two different projection matrices are learned to respectively map the action data from two different views into the common space by simultaneously maximizing the similarity of intra-class samples, minimizing the similarity of inter-class samples and reducing the mismatch between data distributions of two views. In addition, the locality information is incorporated into the discriminant analysis as a constraint to make the discriminant function smooth on the data manifold. Our method is neither restricted to the corresponding action instances in the two views nor restricted to a specific type of feature. We evaluate our approach on the IXMAS multi-view action dataset and N-UCLA dataset. The experimental results demonstrate the effectiveness of our method.

1. Introduction

Human action recognition in videos plays an important role in computer vision due to its wide applications in human–computer interaction, smart surveillance, and video retrieval. In order to accurately recognize human actions, lots of approaches focus on developing effective action representation, such as 2D shape matching [1–3], optical flow patterns [4], spatio-temporal interest points [5–7], and trajectory-based descriptors [8–10]. Especially, dense trajectories-based methods have achieved impressive results on a variety of datasets [11,12]. These methods are effective for action recognition from a single viewpoint. However, the problem of viewpoint changes has posed a real challenge to human action recognition for the fact that the same action appears quite different when observed from different views. Both the data distribution and the feature space can vary drastically from one view to another. As a result, action models learned in one view tend to be incapable of the recognition in another different view [13–16].

Recently, lots of efforts have been made towards the problem of cross-view action recognition. A number of geometry-based approaches are motivated to perform by using the geometry measurement of body joints [17–20] or inferring 3D models of human subjects [21–23], usually requiring robust joint estimation which is still a challenging task. Another group of approaches tries to compute view-invariant human action representations that are stable across different viewpoints, such as temporal self-similarity matrix descriptors [24], temporal dynamics representation [25]. Several methods [26–32] have resorted to transfer learning, which constructs the mappings or connections to bridge the gap between different views. Methods [26,27,31,32] rely on either feature-to-feature correspondence or video-to-video correspondence to transfer knowledge across views. Methods [28,30] require action features of the same type in different views. However, the corresponding data and homogeneous features in both views are not always available easily. In [29], Wu et al. proposed an iterative optimization algorithm to learn a common subspace for cross-view action recognition over heterogeneous feature spaces. Their method has less restrictions except that each action sample must be represented by a sequence of image features.

In this paper, we present a new transfer learning approach for cross-view action recognition in heterogeneous feature spaces, called Heterogeneous Linear Discriminant Analysis (*HLDA*). Our method is neither restricted to the corresponding action instances in the two views nor restricted to action features of the same type. Moreover, in this work, each action sample is represented by a commonly used feature vector. All these make our method more general than the existing ones. Specifically, in order to effectively utilize these features, we are encouraged to align the features from the two views via a discriminative common space, where the action samples captured from different viewpoints can be compared directly and the data from different classes can be separated well.

Our paper focuses on the construction of the common space. We aim to learn two different projection matrices to respectively map the data from the source and target views to the common feature

* Corresponding author.
E-mail address: wuxinxiao@bit.edu.cn (X. Wu).

http://dx.doi.org/10.1016/j.neucom.2016.01.051
0925-2312/© 2016 Elsevier B.V. All rights reserved.

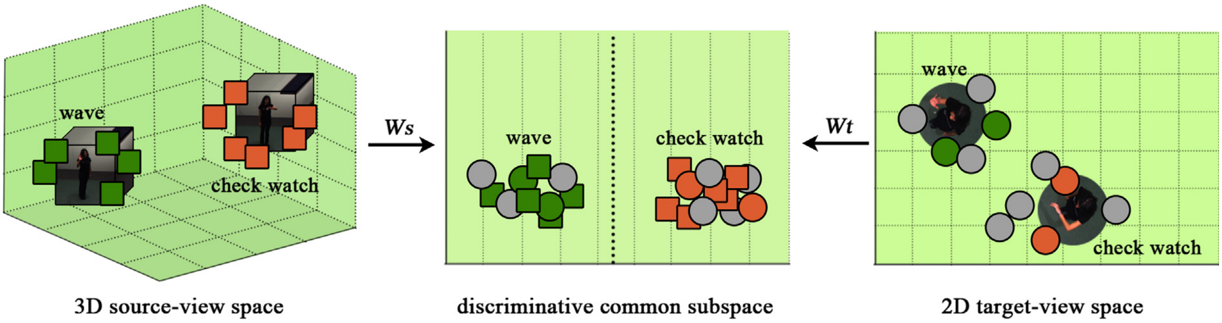

Fig. 1. An illustration of our framework. Samples from different views are represented by features with different dimensions, painted as different shapes (i.e., square and circle). Samples from different classes are denoted by different colors (i.e., green and red). The gray ones are unlabeled data. *Ws* and *Wt* are the projection matrices respectively mapping the heterogeneous data from two views to the derived common subspace. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

space. The two projections are learned by simultaneously maximizing the variance of intra-class samples, minimizing the variance of inter-class samples. Taking Fig. 1 for example, given the 3D data as source-view data and the 2D data as target-view data, we propose to explore the common subspace where the data points belonging to different classes (denoted in different colors) are well-separated from each other, while those from the same class (denoted in the same color) are closely related to each other. In order to reduce the mismatch between data distributions of different views, we also add an effective nonparametric criterion into the objective function. As Fig. 1 shows, the two distributions are similar in the projected subspace, even though they look quite different in the original 3D and 2D spaces. Moreover, a valid locality constraint is incorporated into the discriminant analysis, which preserves the local manifold structure and makes the discriminant function as smooth as possible on the data manifold. It is designed to make the neighbor data points in original space still close to each other in the new projection subspace, which keeps the similarity of the original neighbor data by using the labeled and unlabeled data. This framework can be naturally generalized to the corresponding kernel version using the kernel trick [33], called Heterogeneous Kernel Discriminant Analysis (*HKDA*), which leads to better performance.

The rest of the paper is organized as follows. Section 2 describes the recent works related to our approach. The proposed HLDA method and HKDA are introduced in Sections 3 and 4, respectively. Extensive experimental results are presented in Section 5, followed by conclusions in Section 6.

2. Related work

2.1. *Cross-view action recognition*

Recently, several transfer learning based methods have been proposed for cross-view action recognition. Farhadi et al. [26] employed maximum margin clustering (MMC) to generate split-based features in source view, and then transferred the split values to the corresponding frames in target view. Zhang et al. [32] imposed temporal regularization on the traditional MMC. These methods require the feature-to-feature corresponding relation at the frame-level. Liu et al. [27] presented a bipartite-graph-based approach to learn bilingual-words from two view-dependent vocabularies in an unsupervised manner, and then transferred actions from different views by a bag-of-bilingual-words model instead of bag-of-visual-words model. Zheng and Jiang [31] proposed a dictionary learning framework to exploit the video-to-video correspondence, by jointly learning a set of view-specific dictionaries for aligning view-specific features and a common dictionary for modeling view-shared features. Different from these approaches, our method possesses the view-shared action representations without any feature-to-feature correspondence or video-to-video correspondence. Li and Zickler [28] tried to connect the source and target views by a smooth virtual path, which is represented as a sequence of linear transformations of action descriptors. Similarly, Zhang et al. [30] intended to bridge two views via a continuous virtual path keeping all the visual information. These methods require the action features of different views with the same dimension. Jia et al. [34] proposed to transfer the depth information from the source RGB-D database to the target RGB database, and use the additional source information to recognize human actions in RGB videos. It can be applied to cross-view action recognition, but it emphasized the data type of source and target database. Wu et al. [29] extended discriminant-analysis of canonical correlations (DCC) [35] to accomplish cross-view action recognition over heterogeneous feature spaces. In their work, each action sample must be represented by a set of image features, and the framework cannot be combined with some impressive action features, such as spatio-temporal features [5], dense trajectory features [8], while ours relax the restriction of data type. Another distinct difference is that our method permits kernelization, which is necessary for learning the projections with non-linear effect. In addition, our method can easily get the closed-form solution, while theirs find the optimized solution by an iterative optimization algorithm.

2.2. *Transfer learning*

In terms of transfer learning, our view-transfer problem has much in common with the heterogeneous domain adaptation problem, and the methods [36–44] are closely related to our work for constructing an effective common feature space. Shawe-Taylor and Cristianini [36] proposed kernel-based canonical correlation analysis (KCCA) to learn the common feature subspace by maximizing the correlation between the source and target training data in an unsupervised manner. Several approaches [37,38] extend KCCA to deal with the problem of cross-view action recognition. Shi et al. [39] employed spectral embedding to unify the different feature spaces without using any label information of training data. Different from these approaches, our method does not require the simultaneous multi-view observations of the same action instance. Wang and Mahadevan [40] proposed a manifold alignment based approach to project samples into a latent space, simultaneously matching the samples with the same labels, separating the samples with different labels and preserving the topology of each domain. They assumed that the data should have a manifold structure. Kulis et al. [41] tried to learn an asymmetric, non-linear transformation for domain adaptation by a supervised learning approach. Hoffman et al. [42] extended this work to simultaneously learn the transformation of features and the