Knowledge-Based Systems 84 (2015) 34–45

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Unsupervised video categorization based on multivariate information bottleneck method ☆

Xiaoqiang Yan, Yangdong Ye *, Zhengzheng Lou

School of Information Engineering, Zhengzhou University, Zhengzhou 450052, China

ARTICLE INFO

Article history:
Received 18 June 2014
Received in revised form 23 March 2015
Accepted 28 March 2015
Available online 3 April 2015

Keywords:
Video categorization
Unsupervised learning
Multivariate information bottleneck
Multiple features
Mutual information

ABSTRACT

The integration of multiple features is important for action categorization and object recognition in videos, because single feature based representation hardly captures imaging variations and individual attributes. In this paper, a novel formulation named Multivariate video Information Bottleneck (MvIB) is defined. It is an extensional type of multivariate information bottleneck and can discover categories from a collection of unlabeled videos automatically. Differing from the original multivariate information bottleneck, the novel approach extracts the video categories from multiple features simultaneously, such as local static and dynamic feature, each type of feature is treated as a relevant variable. Specifically, by preserving the relevant information with respect to these feature variables maximally, the MvIB method is able to integrate various aspects of semantic information into the final video partitioning results, and thus captures the complementary information resided in multiple feature variables. Extensive experimental results on five challenging video data sets show that the proposed approach can consistently and significantly outperform other state-of-the-art unsupervised learning methods.

1. Introduction

With the continuing rapid growth of personal video recordings, online video data and broadcast news, unsupervised action categorization and object recognition in video clips have been an active and challenging research area. However, there are two key issues in the task of discovering categories automatically from a collection of unlabeled videos: (1) Because of the cluttered background, camera motion, occlusion, changes of view point and variances of the geometric distribution in videos, robust feature representation extraction remains a difficult problem. Moreover, single feature based representation can hardly capture imaging variations and individual attributes. (2) Automatical method for differentiating videos is also a quite challenging task due to the lack of the ground-truth label information.

For the first key issue, robust feature representation should be achieved before discovering categories from a collection of unlabeled videos. Recently, researchers have found that the quality of feature representation is of great importance to action categorization and object recognition in videos. Therefore, large amounts of feature extraction techniques were proposed. The well-known feature representations include static based on edges and limbs [1–3], shape or form features [4,5]; interest point based representation [6,7]; motion or optical flow patterns [8,9]; non-tensor product wavelet filter banks [10]. However, the capability of single feature is not enough to capture discriminative information, which will make the representations include prejudices caused by single feature. Besides, most of feature extraction approaches described above only consider single aspect of information for the task of video action classification. For instance, certain actions, such as hand clapping, produce a small number of dynamic features since most body parts remain static. While some other actions, such as cycling and horseback riding, are similar in motion features. Therefore, it is difficult to distinguish action categories in the task of unsupervised video categorization only based on single feature. So we strongly feel that various kinds of features are mutually complementary for video action categorization.

A reliable mechanism to learn the action and object categories based on the visual features is the second key issue associated with unsupervised category discovery in videos. However, most current learning approaches are supervised methods, which need the ground-truth label information. Labeling the videos manually is a labor intensive and time consuming process, which often invites subject biases or mistakes by human labelers. So some research efforts have been dedicated to the task of unsupervised object

☆ This work is supported by the National Natural Science Foundation of China under Grant Nos. 60773084 and 61170223 and the Joint Funds of the National Natural Science Foundation of China under Grant No. U1204610.

* Corresponding author. Tel.: +86 13838382185.
E-mail addresses: iexqyan@gmail.com (X. Yan), yeyd@zzu.edu.cn (Y. Ye), iezzlou@zzu.edu.com (Z. Lou).

http://dx.doi.org/10.1016/j.knosys.2015.03.028
0950-7051/

category discovery, such as k-means, PLSA [11], LDA [12], AP [13], SC [14]. These existing unsupervised methods try to learn the object categories from the visual contents in a two-step manner: (1) Building an affinity matrix to reflect the video relations based on visual features. (2) Partitioning the videos into different groups by considering the affinity values. This assumption always limits the performance of aforementioned methods due to the semantic gap between the visual features and their high-level semantic concepts. Besides, most unsupervised approaches used in the domain of video categorization can only cope with single feature.

In this paper, a novel unsupervised video categorization method called Multivariate video Information Bottleneck (MvIB) is proposed, which can partition video clips from multiple cues. Like original multivariate information bottleneck (multivariate IB) model, the novel clustering approach treats pattern structure extraction as data compression. In the video categorization procedure, the proposed method conserves the relevant information from multiple feature cues rather than only one source of feature information. Besides, a information-theoretic optimization is adopted to learn the latent semantic correlations between the videos and their constructive visual words automatically, which can relieve the semantic gap between the visual features and their high-level semantic concepts.

The major contributions of this study are summarized as follows:

- A novel and effective multivariate information bottleneck model is proposed, which extends the original multivariate IB to the task of unsupervised video category discovery.
- The MvIB method can incorporate multiple information cues into the clustering process, which provides an effective solution to integrate multiple features simultaneously.
- An effective information-theoretic optimization method is designed to learn the latent semantic correlations between the videos and their low-level visual features, which alleviates the semantic gap in the current unsupervised learning techniques.

The rest of this paper is organized as follows. In Section 2, the related work is introduced. In Section 3, the basic knowledge is presented about multivariate IB method. In Section 4, details of the MvIB approach are described. In Section 5, extensive experimental results are presented to demonstrate the performance of MvIB. Finally, conclusions are given in Section 6.

2. Related work

Several works have been reported on the integration of features for action category discovery in realistic videos. Neibles et al. [15] proposed a generative method to learn a hierarchical model using both static and dynamic features for action recognition, their results verified the combined features were useful. Liu et al. [16] adopted Fiedler Embedding model to combine local dynamic and spin image features, where the spin image features capture the global pose information. Natarajan et al. [17] analyzed and combined a large set of low-level features that captured appearance, color, motion co-occurrence patterns in web videos. Ikizler et al. [18] proposed multiple instance learning (MIL) framework for human action recognition, which integrates multiple feature channels from several entities such as objects, scenes and human. Kim et al. [19] proposed a new indexing method, which has the ability to capture videos with spatiotemporal information such as time, location, and camera direction. Despite the good recognition performance of the above methods, low-level features are incapable of understanding the hidden semantic structure latent in videos. Often, the aforementioned methods are akin to object recognition and require extra training videos. So these methods may not be applicable for realistic videos due to the difficulty in acquiring good features in unconstrained videos.

Some research efforts have been dedicated to the task of unsupervised category discovery in videos. Bettadapura et al. [20] presented data-driven techniques to augment the BoW model, which allowed for more robust modeling and recognition of complex long-term activities. Beyond the BoW, the discriminative topic learning model has achieved excellent performance on action categorization and object recognition tasks, such as LDA and PLSA [15,21]. Moreover, several methods based on spectral clustering have been applied to the domain of unsupervised image and video categorization by considering multiple features recently. Correlational spectral clustering [22] separates similarity measures for each data representation, and allows for projection of previous unseen data that are only observed in one representation. Heterogeneous image feature integration via multi-modal spectral clustering [23] learns a commonly shared graph Laplacian matrix by unifying different models by considering each type of feature as one modal. Affinity aggregation for spectral clustering [24] seeks for an optimal combination of affinity matrices so that it is more immune to ineffective affinities and irrelevant features.

For the information bottleneck (IB) method [25], Winston et al. [26] utilized the IB method for video reranking and achieved good results, which shows IB is a promising method for semantic learning, but this study solely focused on video search reranking task. Lou et al. [27] extended the original information bottleneck method to multiple-feature version, which aimed to extract the data patterns from multiple feature variables. Therefore, the partition results reflected the hidden patterns provided by multiple types of features simultaneously. But the method presented in [27] was based mainly on original information bottleneck. Note that, the proposed method in this work is a multiple-feature extension of multivariate IB theory.

In this study, we focus on the unsupervised video category discovery issue. Different from all the aforementioned approaches, a novel multiple-feature extensional type of the multivariate information bottleneck [28] method is defined. The MvIB method can integrate multiple visual information into the clustering process, which provides an effective solution to integrate multiple features simultaneously. Moreover, the MvIB method can relief the semantic gap problem effectively by exploiting the correlations between the videos and the visual words. There are many achievements in the field of machine learning and computer vision to cope with multiple sources of features, but most of them need supervision, i.e. the class label information. Note that, the MvIB algorithm is an unsupervised learning method.

3. Multivariate information bottleneck

The original single-sided IB principle involves finding a compressing scheme with a given source variable, X, while preserving the information it maintains about relevant variable Y. This formulation is inherently a-symmetric, only X is compressed while only Y serves as a relevant variable. In order to cope with multiple variables scenario, Slonim et al. presented a general formulation for multivariate extension of the single-sided IB principle, named multivariate information bottleneck (multivariate IB) [28]. The multivariate IB method is an information-theoretic based data analysis method, which treats the pattern extraction as a process of data compression. To define the amount of information that multiple variables contain each other, the concept of *multi-information* is utilized by the multivariate IB method, which is a natural extension of the concept of mutual information. The multi-information