Information Sciences 281 (2014) 635–647

Contents lists available at ScienceDirect

Information Sciences

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

Action recognition based on overcomplete independent components analysis

Shengping Zhang [a,*], Hongxun Yao [a], Xin Sun [a], Kuanquan Wang [a], Jun Zhang [b], Xiusheng Lu [a], Yanhao Zhang [a]

[a] *School of Computer Science and Technology, Harbin Institute of Technology, China*
[b] *School of Computer Science and Information, Hefei University of Technology, China*

ARTICLE INFO

Article history:
Available online 8 January 2014

Keywords:
Action recognition
Feature learning
Independent component analysis
Overcomplete representation

ABSTRACT

Existing works on action recognition rely on two separate stages: (1) designing hand-designed features or learning features from video data; (2) classifying features using a classifier such as SVM or AdaBoost. Motivated by two observations: (1) independent component analysis (ICA) is capable of encoding intrinsic features underlying video data; and (2) videos of different actions can be easily distinguished by their intrinsic features, we propose a simple but effective action recognition framework based on the recently proposed overcomplete ICA model. After a set of overcomplete ICA basis functions are learned from the densely sampled 3D patches from training videos for each action, a test video is classified as the class whose basis functions can reconstruct the sampled 3D patches from the test video with the smallest reconstruction error. The experimental results on five benchmark datasets demonstrate that the proposed approach outperforms several state-of-the-art works.

1. Introduction

Action recognition is an important research topic in the computer vision community. In the past decades, it has been attracting increasing interest due to its many applications in sport video analysis [72], animation synthesis [58,59] and intelligent video surveillance [44,1]. The goal of action recognition is to automatically recognize ongoing activities from a video. Although it is effortless for human beings to recognize a specific action in complex scenes such as a crowded airport, it is extremely difficult for a computer. Even if the scene is constrained to be simple, e.g., a single person captured by a static camera, most existing action recognition systems can only recognize a small number of simple actions, which is still far from becoming useful in practical situations.

From the perspective of pattern recognition, an action recognition system usually consists of two main stages: feature representation and classification. The purpose of feature representation is to compute a feature vector from a video to represent its appearance. Using this feature representation method, a collection of feature vectors is extracted from training videos of all classes and can then be used to train a multi-class classifier. After extracting the feature vector from a new test video, the trained classifier can be used to classify the test video as one of the predefined action classes. In recent years, many interesting feature representation and classification methods have been proposed [45,57,51,62,13,71,50,64] for different recognition tasks. In the field of action recognition, a huge number of works have also been proposed [20,26,24,46,29,41,63].

[*] Corresponding author. Tel.: +86 45186416485.
E-mail address: s.zhang@hit.edu.cn (S. Zhang).

0020-0255/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.12.052

Most of these methods focus on feature representation and usually follow a classical pipeline [48] that divides the feature representation into three steps: local feature extraction, vector quantization and global feature pooling. Kmeans is widely used in the vector quantization step [29]. In feature pooling step, a bag-of-words model is usually adopted to combine local features to form the final feature representation [6]. The majority of the existing work focuses on how to extract powerful features in the local feature extraction step.

In recent years, low-level hand-designed features have been heavily employed in action recognition, motivated by their successes in object recognition in still images. The most successful hand-designed features include Gabor responses [45], SIFT [33], HOG [10], GLOH [36], SURF [3] and trajectory features [47]. To describe the appearance of a video, a 3D detector is usually used to locate the interest points with rich features. Well-known feature detectors include Harris3D [25], Cuboid [11] and Hessian [53]. After detecting an interest point, a 3D descriptor is used to describe the appearance of the 3D patch centered at the detected point. The widely used 3D descriptors are Cuboid [11], HOG/HOF [26], HOG3D [24], Extended SURF [53], slow features [69] and others [37,43]. Note that these 3D descriptors are usually extended from their 2D versions used in object recognition in still images.

Hand-designed features need much prior knowledge, therefore, they cannot adapt well to different video data. In addition, the research purpose of computer vision is to approach and even outperform human intelligence. Human beings have the ability of learning knowledge from the external word. Therefore, in contrast to hand-designed features, learning features from video data is more important for developing intelligent action recognition systems. Recently, unsupervised feature learning methods have been successfully used in object recognition [49,7,62,61], image classification [57,60,71] and visual tracking [30,65,70,67,66]. Among these methods, sparse coding has been attracting increasing interest because the responses of sparse coding have similar properties with receptive fields of simple cells in visual cortex [38,39]. Independent component analysis (ICA) [5,19], a specific case of sparse coding when constraining the number of basis functions to equal the feature dimension, also shows similar response properties with simple cells in visual cortex and achieves success in face recognition [2] and action recognition [29]. Very recently, some sophisticated learning methods such as convolutional GRBM [46], dynamic topic model [16], 3D convolutional neural networks [22], temporal bayesian networks [68], temporal actom model [12] and Gaussian mixture model [56] have also been used in action recognition.

The features learning methods mentioned above are only for modeling the appearance of a video. The resulting representation has to be fed into a classifier to perform the final recognition task. Although many methods report reaching desired performance, the ability of the learning methods is not fully exploited. An interesting question is whether we can directly exploit the learning methods to perform classification? A few works have been proposed along this line for abnormal behavior detection in a video [9] or novel topic detection in document [23]. In [9], a normal dictionary is learned by sparse coding from training videos containing only normal behaviors. The sparse reconstruction cost over the normal dictionary is used to measure the normalness of a test video. Similarly, Kasiviswanathan et al. [23] also learns a normal dictionary to represent the documents containing only normal topics and then uses the sparse reconstruction cost over the normal dictionary to measure the novelness of a test document. Both of these works attempt to solve a binary classification problem. They cannot be used for multi-class classification, e.g., action recognition studied here. In this paper, we propose an approach to action recognition based on reconstruction error over an overcomplete ICA dictionary. Unlike Le et al. [29], our method uses ICA to perform classification rather than extracting features for a SVM classifier. In particular, we use the newly proposed overcomplete ICA to train a set of basis functions using 3D patches densely sampled from training videos for each class. For a new test video, we densely sample a set of 3D patches from it and reconstruct them using the basis functions learned for each class. The test video is classified as the action class whose basis functions can reconstruct the test video with the smallest error. The conceptual diagram of the proposed method is shown in Fig. 1.

The motivation of the proposed method is that videos from an action class contain specific intrinsic features, which can be used to distinguish them from others. By densely sampling local 3D patches from training videos for all action classes, the class label of a test video can be inferred from the underlying features in the 3D patches sampled from it. When learning a set of overcomplete ICA basis functions from the sampled 3D patches from an action class, these basis functions capture the underlying features in the patches. Intuitively, if the new video comes from the same action class, the learned basis functions should reconstruct it with minimal error. Therefore, we can exploit this property to classify the test video. In contrast to previous methods, which extract feature representation to describe the video and then classify its feature representation by a SVM [17,52,14] or AdaBoost classifier, our method exploits the statistical properties of the underlying features learned by overcomplete ICA for classification. The proposed method is more intuitive and experimental results verify that it outperforms several state-of-the-art works.

The rest of the paper is organized as follows. In Section 2, we review some related works on feature representation for action recognition. Section 3 presents a detailed description of the proposed action recognition method. Experimental results are reported and analyzed in Section 4. We conclude the paper in Section 5.

2. Related works

In this section, we review some related works on feature representation for action recognition including both the hand-designed features and learning based features.