Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

Sparsity-inducing dictionaries for effective action classification

Debaditya Roy*, Srinivas M., Krishna Mohan C.

Visual Learning and Intelligence Group (VIGIL), Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Kandi, Hyderabad 502285, India

ARTICLE INFO

Article history:
Received 30 August 2015
Received in revised form
1 March 2016
Accepted 8 March 2016

Keywords:
Action Classification
Dictionary Learning
Sparse Representation
Action Bank features

ABSTRACT

Action recognition in unconstrained videos is one of the most important challenges in computer vision. In this paper, we propose sparsity-inducing dictionaries as an effective representation for action classification in videos. We demonstrate that features obtained from sparsity based representation provide discriminative information useful for classification of action videos into various action classes. We show that the constructed dictionaries are distinct for a large number of action classes resulting in a significant improvement in classification accuracy on the HMDB51 dataset. We further demonstrate the efficacy of dictionaries and sparsity based classification on other large action video datasets like UCF50.

1. Introduction

Action recognition is the process of extracting human action patterns from real video streams. It can be used in diverse applications like automated video indexing of huge on-line video repositories like Youtube and Vimeo, analysing video surveillance systems in public places, human-computer interaction, sports analysis, etc. Actions are defined as single-person activities like “walking”, “waving”, “punching”, etc. If the action video contains only one distinct human action, the task is to classify the video into one of the different categories. It has been shown in [1] that both spatial and temporal information are important for action representation. However, features which are shared across action classes are not suitable to build discriminative dictionaries. For example, “running” is a part of both “cricket bowling” and “soccer penalty”. In such a case, the main action (bowling/penalty taking) occupies a small fraction of the entire duration of the video. Hence, it is difficult with just spatio-temporal descriptors to classify such actions with high credibility. Action bank [2] captures the similarity of the video with the class it belongs to and dissimilarity with other classes. Since, running occurs before bowling(or penalty taking), this temporal dependence can be exploited to produce a more unique representation for “cricket bowling” (or soccer penalty) which is useful for classification.

In this work, we construct sparsity-inducing dictionaries built specifically for action classification. Such a sparse dictionary based representation highlights discriminative information about various action classes. Also, these dictionaries distinctly represent the different action classes of HMDB51 dataset. Since dictionary learning has no strict convergence criteria, the dictionaries are trained until reasonable classification performance is obtained. On the HMDB51 dataset which contains many diverse and challenging views of human actions, dictionaries achieve very low misclassification rate.

The rest of the paper is organized as follows. In Section 2 we provide an overview of the various feature descriptors and sparsity based methods which have been applied for action classification. In Section 3, we present the proposed sparsity based classification scheme in detail. In Section 4, we describe the performance of the proposed approach on two large action datasets – UCF50 and HMDB51. Finally, Section 5 gives the conclusion for this work.

2. Related work and analysis

The challenges in action recognition have been studied with great interest in the computer vision community. Schuldt et al. [3] introduced the KTH [4] dataset which consists of six action categories. A support vector machine (SVM) was used for classification with local space-time features. In [5], Kläser et al. presented the histogram of oriented 3D spatio-temporal gradients which is essentially a collection of quantized 2D histograms collected from each frame of the video. Kuehne et al. [6] introduced the HMDB51

* Corresponding author.
E-mail addresses: cs13p1001@iith.ac.in (D. Roy), cs10p002@iith.ac.in (M. Srinivas), ckm@iith.ac.in (C. Krishna Mohan).

http://dx.doi.org/10.1016/j.patcog.2016.03.011
0031-3203/

dataset [7] for action recognition. Features such as histogram of oriented gradients (HOG), histogram of optical flow (HOF) and C2 were extracted and then a radial basis SVM was used for classification. Kliper et al. [8] proposed the use of motion interchange patterns i.e the change of one motion leading to another to describe a distinct action.

Solmaz et al. [9] presented the idea of gist, a global video descriptor which essentially computes the 3-D discrete Fourier transform of a given video clip using 68 3-D Gabor filters placed in 37 and 31 orientations. A trajectory based local descriptor TrajMF was proposed by Jiang et al. [10] which works on top of local feature descriptors like HOG, HOF, etc. and captures global and local reference points to characterize motion information. Wang and Schmid [1] employed the idea of dense trajectories by estimating human motion, accurate camera motion estimation and removing inconsistent matches. In [11], Wu and Hu denoted each action class as an event and assigned a latent variable to it. The crucial motion patterns in each event were then captured using latent models. These latent models were then used to construct latent structural SVMs, max-margin hidden conditional random fields and latent SVMs. Using a latent spatio-temporal compositional model in [12], actions were simplified in terms of spatio-temporal And-Or Graphs.

Recent works like [13] and [14] indicate that self-learned features can be as competitive as manually generated features for action classification. These works focus on convolutional neural networks (CNN) and CNN-based recurrent neural networks (RNN). In [13], consecutive frames of a video were processed through separate CNNs and then the outputs are fused in various configurations to obtain the best possible discriminative representation. Ng et al. [14] combined the outputs of CNNs from 15 or more subsequent frames into a RNN using long short term memory units (LSTM) to obtain a temporal representation. The performance was slightly better than improved dense trajectory features on the UCF101 dataset. A deep parsing based CNN network was proposed in [15] to build an end-to-end relation between the input human image and the structured outputs for human parsing. In [16], images representing humans actions are classified and localized using multiple regions for training a region-based CNN (R-CNN). Lin et al. [17] developed a deep structural model for 3D action recognition. Traditional CNNs were fused with a latent temporal model for representing temporal variation. Regularization was introduced in the form of radius-margin bound for better generalization. A similar architecture is presented in [18]. In [19], handcrafted features were augmented with CNN outputs learnt from various input sources using multiplicative fusion to classify actions. From the literature it can be seen that CNNs can provide a good representation of human actions.

Action bank features are useful for semantic representation of videos proposed by Sadanand and Corso [2]. This representation of videos is achieved by applying 73 spatio-temporal volume detectors on a video clip. There are 205 action templates having an average spatial resolution of approximately 50×120 pixels and a temporal length of 40–50 frames. This contributes to a 14,965-dimensional feature vector for each video clip under consideration. The templates perform classification by detection and give a global description of videos. Action bank produces a single feature vector for an entire video clip which is larger $(14{,}965 \times 1)$ as compared to the number of video clips per class in any of the standard datasets (≈ 100). The resultant matrix is a "fat" matrix $(14{,}965 \times 100)$ which gives rise to an under-complete dictionary learning setting. In this work, we explore sparsity-inducing dictionaries to achieve a discriminative representation of human actions.

Dictionaries have been previously used in literature for action classification. In [20], information maximization was used for building discriminative dictionaries. These dictionaries were used to represent action attributes to classify images representing human actions. Sparse modeling for motion analysis was proposed by Castrodad and Sapiro [21] . Using highly redundant features, a two-level pipeline was built to distinguish human actions. An evaluation of three different dictionary types – shared, class-specific and concatenated for the KTH, Weizmann and Hollywood2 datasets was done in [22]. The study found that the class-specific dictionaries perform better on an average than the shared and concatenated types. In [23], a sparse dictionary was constructed in an on-line manner for each incoming frame. In case of normal activity, consequent frames are related to each other and dictionary update is minimal. However, any abnormal activity would cause a major change in the dictionary. A new descriptor known as locally weighted word context was introduced in [24] which is a context-aware spatio-temporal descriptor. A sparse dictionary based on the descriptor was constructed using the joint $\ell_{2,1}$-norm where each action category share similar atoms in the dictionary.

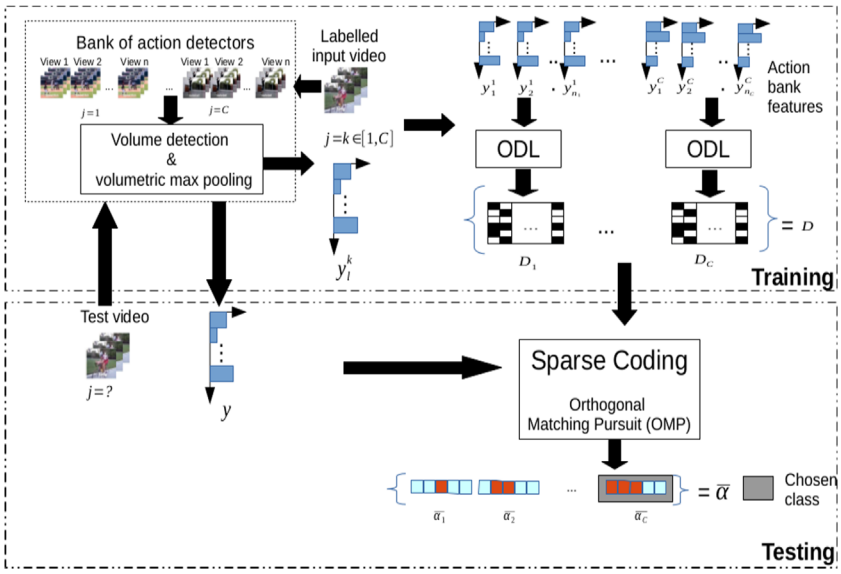

Fig. 1. Flowchart of the proposed approach.

Please cite this article as: D. Roy, et al., Sparsity-inducing dictionaries for effective action classification, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.03.011