Pattern Recognition 47 (2014) 2689–2701

Contents lists available at ScienceDirect

Pattern Recognition

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

# Laplacian group sparse modeling of human actions

Xiangrong Zhang [a,*], Hao Yang [a], L.C. Jiao [a], Yang Yang [a], Feng Dong [b]

[a] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, P.O. Box 224, Xi'an 710071, China
[b] Department of Computer Science & Technology, University of Bedfordshire, Luton LU1 3JU, United Kingdom

ARTICLE INFO

*Article history:*
Received 14 April 2013
Received in revised form
7 February 2014
Accepted 12 February 2014
Available online 20 February 2014

*Keywords:*
Action recognition
High-level representation
Laplacian group sparse coding
Structural information

ABSTRACT

Recently, many local-feature based methods have been proposed for feature learning to obtain a better high-level representation of human behavior. Most of the previous research ignores the structural information existing among local features in the same video sequences, while it is an important clue to distinguish ambiguous actions. To address this issue, we propose a Laplacian group sparse coding for human behavior representation. Unlike traditional methods such as sparse coding, our approach prefers to encode a group of relevant features simultaneously and meanwhile allow as less atoms as possible to participate in the approximation so that video-level sparsity is guaranteed. By incorporating Laplacian regularization the method is capable to ensure the similar approximation of closely related local features and the structural information is successfully preserved. Thus, a compact but discriminative human behavior representation is achieved. Besides, the objective of our model is solved with a closed-form solution, which reduces the computational cost significantly. Promising results on several popular benchmark datasets prove the efficiency and effectiveness of our approach.



## 1. Introduction

Human action recognition in videos is a very interesting and challenging problem in the area of computer vision. It can be applied to solving many reality problems such as video surveillance, smart camera system and human computer interaction [1–3].

Due to the reliability under noise and illumination changes, may approaches based on spatio-temporal interest points have been developed by researchers to describe human behavior. Most of these approaches extract sparse features from 3D space-time volumes corresponding to the local regions where the human motion varies significantly. Since the detected interest points are able to capture the motion patterns of different actions, local features extracted from the detected regions can be employed to represent human actions in video sequences. Descriptors like HOG [4], HOF [4] and HOG3D [5] have proved to be very efficient in representing human actions and such spatio-temporal local feature based approaches have achieved relatively good performance on many simple datasets. However, there still exist many difficulties to represent complex human actions in reality:

(1) The amount of local features is pretty large when dealing with complex dataset, especially under the condition with camera motion and viewpoint varying. In these cases the representation of human actions is more likely to take in redundant information. But we cannot be sure that every feature makes equal contribution to describing the movement of body part. In other words, achieving efficient representation is very desirable when it comes to dealing with large dataset.
(2) The relationship among local features in the same video is ignored by many previous approaches. The hidden structure hidden behind human actions is quite important to discriminate different actions but difficult to explore. As shown in Fig. 1, local features from similar frames are more likely to become neighbors since they have captured similar motion patterns. While it seems difficult to describe those similarities, the locality based characteristic of an action clip plays quite an important role in the following step of classification.

Therefore, we need to learn a proper representation of human actions that is able to exploit the structural information and meanwhile does not cost too much space to store. To meet this end, a compact but discriminative high-level representation is very desirable. The procedure of building high-level representation from local features is suggested to be called feature learning in [6]. Compared with high-level representation, the local features can be also called raw features [7]. One of the traditional and common used techniques for feature learning is Bag-of-Features (BoF) [4,8]. BoF is a vectorizing technique which maps raw features to the nearest codeword created by *k*-means clustering. It is easy to apply BoF to modeling simple periodic actions. However, for

* Corresponding author. Tel.: +86 29 8202279; fax: +86 29 8201023.
*E-mail address:* xrzhang@mail.xidian.edu.cn (X. Zhang).

http://dx.doi.org/10.1016/j.patcog.2014.02.007
0031-3203 

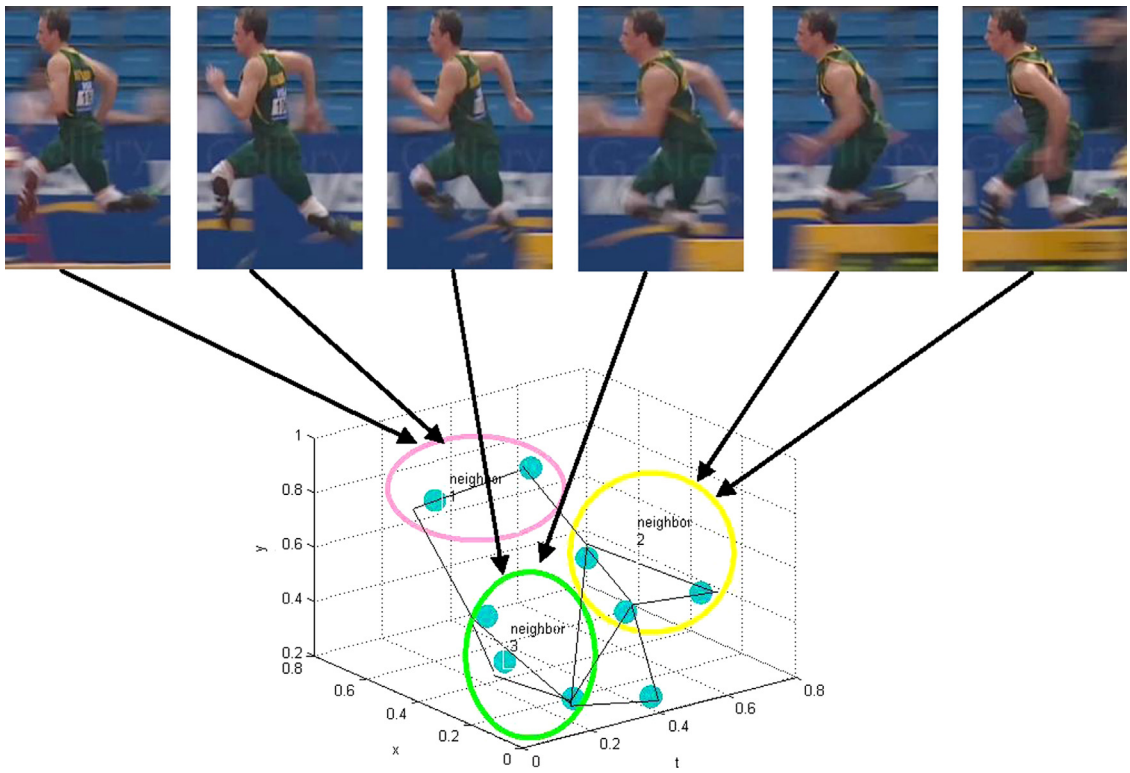


**Fig. 1.** Frames from the same video sequences may illustrate similar motion patterns, which means there should be similarities among their corresponding spatio-temporal interest points.

complex datasets BoF may require a large number of codewords which will lead to very redundant and high dimensional representation [26].

Recently, sparse coding has drawn lots attention in the areas of denoising [9,10], feature learning and classification [11–14]. It usually learns an over-complete dictionary from raw features. The dictionary has the same function as codebook used in BoF techniques. But different from BoF, sparse coding prefers that each feature is approximated by a weighted sum of several atoms rather than simply assigned to the nearest codeword. In this way, raw features can be represented more accurately with less approximation error. Liu etc. [38] utilize sparse coding to represent each frame as a linear combination of all frames in the training sequences. And the test sequence is classified by majority voting of all the frames. The method can help lifting the performance of sparse coding to some extent compared with BoF.

However, sparse coding only guarantees a compact representation for each raw feature but not the whole video clip. No matter whether the features are from the same video sequences, they are considered independently and encoded one by one. In other words, sparse coding only promotes region-level sparsity but not the video-level. Besides, it is hard to mining the structural information, such as similarities and spatio-temporal dependencies among different regions within the same video clip, if we only rely on simple sparse coding to gain high-level representation of human actions. But the structural information can be an important clue to distinguish ambiguous actions.

To address the above issues, we propose a novel method, Laplacian group sparse coding (LGSC), to help improving the performance of activity recognition. The first benefit is the structural information is implicitly encoded by incorporating the Laplacian regularization. As a fact, local features from the same video sequences are usually more closely related to each other than the ones from different video sequences or even different classes. So it is very helpful to strengthen the discriminative power of high-level descriptors if the relationship among local features can be preserved. Once interest points are detected, LGSC constructs a similarity based graph structure of local features from the same group. The similarity graph aims to describe properly how near the descriptor vectors of local features can be. In this way, the relationship among local features from the same video sequences is successfully preserved. As shown in Fig. 2, the structure hidden behind group of local features can be reflected by the representation achieved by our LGSC framework. Otherwise, ordinary methods such as sparse coding seem to ignore the hidden structural information which may lead to less distinctiveness.

Moreover, our work fully absorbs the advantages of group sparse coding (GSC) [7]. The meaning of group sparsity can be explained in two aspects. On one hand, closely related local features can make up of a group. Namely local features from the same video sequences or even the same class which show more concurrency are encoded simultaneously. On the other hand, as an extension of GSC, LGSC also promotes as less atoms as possible participant in the sparse coding of the whole group of features. If an atom is selected to represent some local feature from a video sequences, then it is considered that local features from the same group can be encoded by this atom too without much additional computational cost. By combining these two kinds of feature learning pattern, a representation with video-level sparsity can be attained by our LGSC. This also makes LGSC different from Laplacian sparse coding [16]. Besides, to make the model more natural, we add a non-negative constraint to the reconstruction coefficients.

The main contribution of our work can be summarized as three folds:

- We propose a locality constrained group sparse coding method which can achieve compact but discriminative high-level representation for human actions.
- The optimization of our LGSC can be solved with a closed-form solution. This will significantly reduce the computational demand