Neurocomputing 158 (2015) 73–80

Contents lists available at ScienceDirect

Neurocomputing

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

Action recognition using direction-dependent feature pairs and non-negative low rank sparse model

Biyun Sheng [a,b], Wankou Yang [a,b,c], Changyin Sun [a,b,*]

[a] School of Automation, Southeast University, Nanjing 210096, China
[b] Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China
[c] Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, Nanjing University of Science and Technology, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 5 October 2014
Received in revised form
6 January 2015
Accepted 31 January 2015
Communicated by Su-Jing Wang
Available online 12 February 2015

Keywords:
Direction-dependent feature pairs
Non-negative low rank sparse model
Direction-specific dictionary
Action recognition

ABSTRACT

In this paper, we propose to use direction-dependent feature pairs (DDFP) to represent actions and a novel non-negative low rank sparse model (NLRM) is developed to encode the features. We summarize our main contributions into three aspects. First, for a video we apply eight different directions to describe the spatio-temporal relations between features, and construct directional feature pairs according to their relative positions. Second, we present a non-negative low rank sparse model which incorporates the low rank term and the non-negative constraint. Our model can not only ensure the consistency of similar DDFP by the low rank term, but also enforce the sparsity of coding coefficients by the modified $l_{2,1}$-norm regularization. Third, we utilize a direction-specific dictionary for each direction and encode DDFP of a specific direction by the corresponding dictionary. A video is finally represented by the concatenation of each direction's pooling result. Experimental results on the KTH, Weizmann and UCF sports dataset show the effectiveness of our proposed framework for human action recognition.

1. Introduction

Human action recognition is an important topic because of its wide applications in video analysis, human interaction, surveillance, and content based retrieval. However, it remains a challenging task due to the significant intra-class variations, occlusion, and background clutter [1].

Among the recent work done for action recognition, using spatio-temporal features together with the bag-of-visual-words (BOVW) framework has attracted much attention [2–4,7]. Features are assigned to the nearest word of a codebook and a histogram of word occurrences is used to represent a video.

The framework is greatly popular for its easy implementation and good performance. However, a feature only describes the region around a single interest point and the context information is discarded. A graph is a useful tool to represent data structure; however, it is tedious to construct the relationship among all of the points and more importantly the similarity measurement between graphs is difficult [8]. Wang et al. [11] propose a novel feature which captures neighboring information and temporal order, while the spatial location relationship is ignored. The failure of capturing spatio-temporal relationship leads to a relatively worse classification accuracy for action recognition.

In order to reduce the reconstruction error of Hard-assignment coding in BOVW model, sparse coding (SC) which approximately constructs the input signal by a few visual words of the dictionary has attracted much attention recently [6,16]. Usually a video is finally represented by the max pooling result of the coding coefficients. However, the coding coefficients of similar features may vary a lot [16], which is unwanted in the coding stage. Besides, a lot of negative and zero coding coefficients are inevitably obtained while encoding features. Since zero (or small positive) elements rather than negative ones are selected during max pooling, some useful information is eliminated which degrades the classification performance. Therefore, it is possible that non-consistent codes generate during coding and information loss emerges during pooling.

In this paper, we present an effective algorithm for action recognition and the main contributions are as follows:

First, motivated by the graph idea, a novel low-level feature named direction-dependent feature pairs (DDFP) is proposed. Local features extracted from a video cannot be taken as independent individuals and they are relevant to each other especially when those are neighbors. It is meaningful for us to preserve the individual power of local features and capture the rich spatio-temporal relationship simultaneously. To construct our features, the values of feature vector and the relative direction of feature pairs are both taken into account. Specifically, a joint feature is obtained by the concatenation of a central feature and one of its

* Corresponding author at: School of Automation, Southeast University, Nanjing 210096, China. Tel.: +86 13913036082.
E-mail address: cysun@seu.edu.cn (C. Sun).

http://dx.doi.org/10.1016/j.neucom.2015.01.064
0925-2312/

neighbors, and then the combination is associated with a direction label according to the spatio-temporal relationship. Therefore, different from the original appearance-only features, the proposed DDFP owns appearance characteristics as well as spatio-temporal context information.

Second, we propose a non-negative low rank sparse model (NLRM) to encode features which is similar to our previous work [25]. DDFP with the same direction label have similarity in some extent. In our model, a low rank term is utilized to solve the non-consistency problem, by which the similarity of coding coefficients for similar features is ensured. Besides, a modified $l_{2,1}$-norm regularization is adopted as the sparse term, in which every column of a matrix is taken as a whole and sparsity of columns is enforced. In order to effectively alleviate information loss during the pooling stage, we add the non-negative constraint on coding coefficients [26].

Third, inspired by the idea of class-specific dictionary, we introduce the direction-specific dictionary in this paper. Given DDFP of R directions, we use R direction-specific dictionaries and then DDFP is represented by its own direction-specific dictionary. By using the direction-specific dictionary, the representations of videos are remarkably more discriminative and robust.

Fig. 1 illustrates the flowchart under our proposed framework. Firstly, we extract local features from videos. To describe the relative position relations between features and capture more context information, DDFP is proposed as the low-level feature. With a direction attribute, every pair of DDFP is assigned to corresponding direction group. Based on DDFP in different groups, direction-specific dictionaries are established. Then, NLRM is presented to learn sparse codes and action representation is the concatenation of each direction's pooling result. Finally, SVM is applied as a classifier. We organize the structure of the paper as follows. Section 2 reviews the related work on context information and coding approaches. Section 3 introduces the new low-level features called DDFP. Section 4 proposes the non-negative low rank sparse model. Section 5 validates our proposed method. Section 6 concludes the paper.

2. Related work

Over the past few years, low-level features have been widely constructed for action recognition in computer vision [2,3,38–41]. However, action representations based on local spatio-temporal features are not able to capture the distribution of points and ignore the rich spatio-temporal context information. To solve the problem, Choi et al. [9] propose Spatial–Temporal Pyramid Matching (STPM) to retrieve videos, which yields the final representation by concatenating the histogram in each grid. However, the high-dimensional vector leads to computational infeasibility and huge storage consumption. Besides, it can only obtain simple spatial and temporal information which is far from enough. Li et al. [12] use Markov chain models to form co-occurrence matrix of visual words. But it has a drawback that the final representation in the form of matrix brings computational complexity. Wang et al. [29] propose Sparse Tensor PCA (STPCA) and further introduce Sparse Tensor Discriminant Color Space (STDCS) [30] to extract discriminative features. Zhao et al. [14] present an optimized version of 3D shape context to encode the layout information. The proposed feature concatenated with the appearance feature such as HOG3D is then adopted to represent a video. The approach can achieve good performance; however, computing two kinds of features is time consuming. Wu et al. [8] use two-graph models to characterize the relationship between features in intra-frames and inter-frames. Then a context-dependent graph kernel is adopted to measure the similarity between graphs. The experimental result heavily relies on the context information in kernel calculation, and each graph can only show either spatial or temporal relationship.

Many authors have tried different methods to encode features for videos or images. In the BOVW model, a high reconstruction error is inevitable due to the restrictive constraint on codes. Then a large number of efficient algorithms have been developed to compute sparse codes [6,15,16]. In the SC model, each input is approximated as a sparse linear combination of the dictionary atoms. Yang et al. [6] learn sparse coefficients of SIFT descriptors followed by multi-scale max pooling. Considering the consistency in sparse representation, Gao et al. [16] propose a similarity matrix to characterize the similarity among local features on the basis of SC model. But it is computationally infeasible and impractical due to a large number of descriptors. A recent trend has been the combination of low-rank and sparse framework. Zhang et al. [17] introduce an ideal regularization term and utilize a supervised learning method to obtain the final representation. Ma et al. [18] integrate rank minimization and class discrimination into sparse representation and successfully apply the algorithm to face recognition. Other approaches have been utilized in a variety of applications [19–21]. Not much work, however, has utilized the low-rank framework to recognize human actions. And these methods have no constraints on the sign of the coding parameters while non-negative components are needed for some particular applications [22,23].

3. Novel low-level features

Traditional features such as HOG/HOF [2] and CUBOID [3] features are only texture or motion descriptions for cuboids which

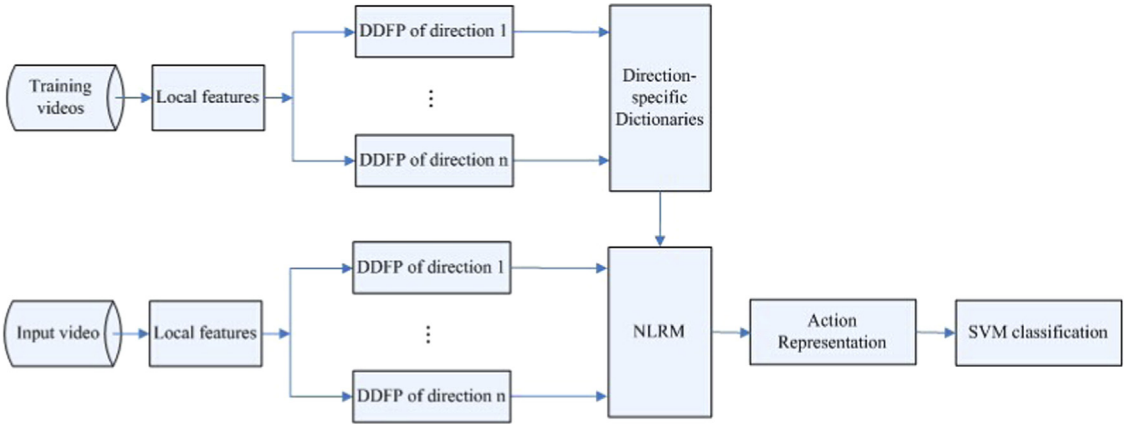

Fig. 1. Flowchart of the proposed action recognition framework.