Neurocomputing 158 (2015) 246–256

Contents lists available at ScienceDirect

Neurocomputing

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

A supervised dictionary learning and discriminative weighting model for action recognition

Jian Dong [a,b], Changyin Sun [a,b,*], Wankou Yang [a,b,c]

[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 21 July 2014
Received in revised form
17 December 2014
Accepted 10 January 2015
Communicated by L. Shao
Available online 13 February 2015

Keywords:
Dictionary learning
Local Fisher Discrimination
Supervised sparse coding
Discriminative weighting model
Multiple Kernel Learning
Action recognition

ABSTRACT

In this paper, we propose a supervised dictionary learning algorithm for action recognition in still images followed by a discriminative weighting model. The dictionary is learned based on Local Fisher Discrimination which takes into account the local manifold structure and discrimination information of local descriptors. The label information of local descriptors is considered in both dictionary learning and sparse coding stage which generates a supervised sparse coding algorithm and makes the coding coefficients discriminative. Instead of using spatial pyramid features, sliding window-based features with max-pooling are computed from coding coefficients. And then a discriminative weighting model combining a max-margin classifier is proposed using the features. Both the weighting coefficients and model parameters can be jointly learned using the same way in Multiple Kernel Learning algorithm. We validate our model on the following action recognition datasets: Willow 7 human actions dataset, People Playing Music Instrument (PPMI) dataset, and Sports dataset. To show the generality of our model, we also validate it on Scene15 dataset. The experiment results show that only with single scale local descriptors, our algorithm is comparable to some state-of-the-art algorithms.

1. Introduction

Starting at about the year of 2006, human action recognition in still images has been an active research topic in Computer Vision and Pattern Recognition [1]. It has quite a few useful applications, such as image annotation, behavior based image retrieval, video frame reduction and video-based action recognition. Similar to video-based action recognition, it is also applicable to human computer interaction. The general problem of human action recognition includes both localization and classification of actions in images or videos [2]. Without considering the localization problem, this paper just focuses on the classification problem of actions in still images.

Most works on human action recognition either rely on accurate or approximate estimation of human pose [3–5]. Desai et al. [3] present a novel approach to model human pose, together with interacting objects, based on compositional models of local visual interactions and their relations. Yao et al. [4] propose a mutual context model to jointly model objects and human poses in human–object interaction activities. However, it also has been demonstrated that action recognition can be achieved without solving the difficult problem of pose estimation [6–8]. Both of the works in [6,7] first extract dense SIFT descriptors and then learn a vocabulary through k-means clustering. Finally, bag-of-features is used to represent all the images. These works have shown that the co-occurrences of some features can effectively identify actions, e.g. a co-occurrence of ‘camera’ and ‘hand’ in a small region is likely to indicate that ‘taking photos’ is happening. The bag-of-features algorithm followed by pooling technique is an effective way to model these co-occurrences. However, in the above two works, the vocabulary is not constructive for local SIFT descriptors and the representations are not discriminative.

Inspired by these algorithms that do not solve pose estimation problem and the recent development of dictionary learning algorithms, in this paper, we propose a supervised dictionary (i.e. vocabulary) learning algorithm followed by a discriminative weighting model. The locality or geometrical information of descriptors have shown to be important for classification or representation [9–11], thus, instead of using global representations, we use local descriptors to learn the supervised dictionary. An objective function with Local Fisher Discrimination item is proposed to make the learned dictionary both constructive and discriminative. One problem for using local descriptors is that the similar, even the same, local descriptors may belong to different classes, as illustrated in Fig. 1(a). In order to slightly solve this problem, we propose a supervised sparse coding algorithm (SSC) to encode new features,

[*] 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.2015.01.024
0925-2312/© 2015 Elsevier B.V. All rights reserved.

as illustrated in Fig. 1(b), through considering the number of neighbors belonging to different classes. Pooling technique is usually used after encoding all the local descriptors, however, the pooling features in different subregions have different discriminations for classification. Thus, we propose a discriminative weighting model to weight different sliding-window based features. We combine this discriminative weighting model into max-margin model which can be efficiently solved using Multiple Kernel Learning framework.

The details of the proposed method are shown in Fig. 2. We first extract dense SIFT features for all the training images and parts of the SIFT features are randomly selected as Template Features. The Template Features are then fed into the proposed Local Fisher Discrimination Dictionary Learning algorithm. After dictionary learning is finished, the learned dictionary and the Template Features are combined to encode all the new SIFT features using the proposed supervised sparse coding algorithm. Subsequently, the sliding window-based pooling features are used to represent all the training images. Finally, the proposed discriminative weighting model is learned using Multiple Kernel Learning framework. Each window-based feature has a weight coefficient to show its importance.

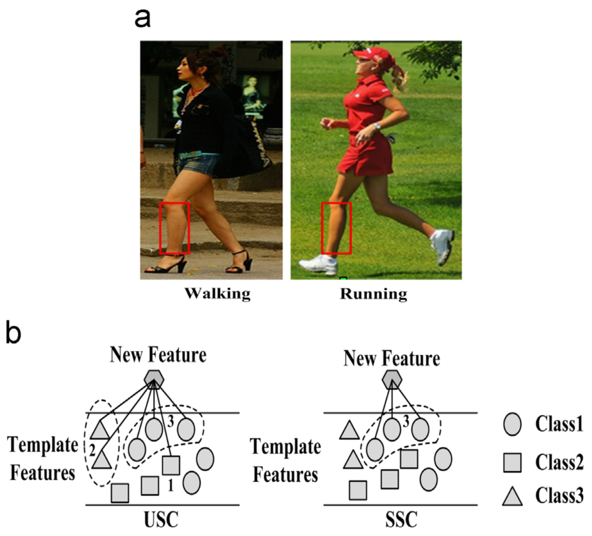

Fig. 1. Illustration of similar local descriptors belonging to different classes and the proposed supervised sparse coding algorithm, see (a). In (b), USC encodes the new feature according to its 6 neighbors in Template Features with one in Class 2, two in Class 3 and three in Class 1. SSC encodes the new feature just according to the Template Features from Class 1.

The contributions of this paper are summarized as follows:

- A Local Fisher Discrimination Dictionary Learning algorithm based on local descriptors is proposed. We propose a supervised sparse coding algorithm for new features using Template Features and their sparse codes. Our experiment results show that our model favors this supervised sparse coding.
- A discriminative weighting model based on sliding window-based features is proposed. Both the weighting coefficients and model parameters can be efficiently solved using Multiple Kernel Learning framework.
- Using only a single scale local descriptors, our algorithm outperforms some state-of-the-art algorithms with multiple scale local descriptors. That is to say, comparing with those algorithms that use features with 7 different scales, just about 1/7 of their local descriptors are encoded in our algorithm.

The remainder of this paper is organized as follows. Section 2 gives some related work. Sections 3 gives the details of the proposed method including the Local Fisher Discrimination Dictionary Learning algorithm, the supervised sparse coding algorithm and the discriminative weighting mode. Section 4 gives the experiment results to validate our algorithm. Section 5 concludes this paper.

2. Related Work

Generally, dictionary learning algorithm can be divided into unsupervised dictionary learning algorithm and supervised dictionary learning algorithm. The proposed algorithm in this paper belongs to the latter one. Thus, in this part, we give some related work on supervised dictionary learning algorithm. Based on the types of features used to learn the dictionary, we roughly classify the supervised dictionary learning algorithm into global representation-based dictionary learning algorithm and local descriptors-based dictionary learning algorithm. Fig. 3 shows two examples of these two algorithms.

2.1. Global representation-based dictionary learning

Jiang et al. [12] propose a label consistent K-SVD (LC-KSVD) algorithm to learn a discriminative dictionary for sparse coding. A new label consistent constraint called ‘discriminative sparse-code error’ is introduced and combined with the reconstruction error and the classification error to form a unified objective function. Li et al. [13] apply Fisher discriminant function to coding coefficients to

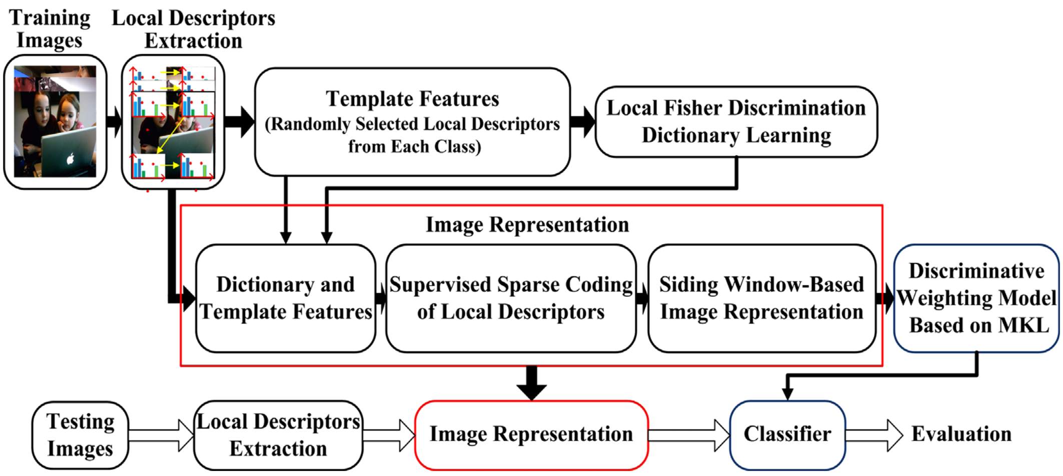
Training Images
Local Descriptors Extraction
Template Features (Randomly Selected Local Descriptors from Each Class)
Local Fisher Discrimination Dictionary Learning
Image Representation
Dictionary and Template Features
Supervised Sparse Coding of Local Descriptors
Siding Window-Based Image Representation
Discriminative Weighting Model Based on MKL
Testing Images
Local Descriptors Extraction
Image Representation
Classifier
Evaluation

Fig. 2. Flowchart of the proposed method in this paper.