Image and Vision Computing 50 (2016) 1–13

Contents lists available at ScienceDirect

Image and Vision Computing

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

Action recognition via spatio-temporal local features: A comprehensive study[☆]

Xiantong Zhen[a,c], Ling Shao[a,b,*]

[a] College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, China
[b] Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, United Kingdom
[c] Department of Electronic and Electrical Engineering, The University of Sheffield, United Kingdom

ARTICLE INFO

Article history:
Received 5 September 2013
Received in revised form 24 October 2015
Accepted 19 February 2016
Available online 20 March 2016

Keywords:
Action recognition
Spatio-temporal local features
Feature coding
Bag-of-words
Sparse coding
Fisher kernel
VLAD
NBNN
Match kernels
Performance evaluation

ABSTRACT

Local methods based on spatio-temporal interest points (STIPs) have shown their effectiveness for human action recognition. The bag-of-words (BoW) model has been widely used and dominated in this field. Recently, a large number of techniques based on local features including improved variants of the BoW model, sparse coding (SC), Fisher kernels (FK), vector of locally aggregated descriptors (VLAD) as well as the naive Bayes nearest neighbor (NBNN) classifier have been proposed and developed for visual recognition. However, some of them are proposed in the image domain and have not yet been applied to the video domain and it is still unclear how effectively these techniques would perform on action recognition. In this paper, we provide a comprehensive study on these local methods for human action recognition. We implement these techniques and conduct comparison under unified experimental settings on three widely used benchmarks, *i.e.*, the KTH, UCF-YouTube and HMDB51 datasets. We discuss insightfully the findings from the experimental results and draw useful conclusions, which are expected to guide practical applications and future work for the action recognition community.

1. Introduction

Human action recognition as an active topic in the computer vision community has been extensively researched in the last decades. Most of the existing methods, including both low-level feature extraction and high-level representations, in action recognition are extended from the text and image domains, i.e., the bag-of-word (BoW) model [1]. Local features have shown increasing effectiveness in visual recognition, and local methods based on spatio-temporal local features, *e.g.*, three-dimensional histogram of oriented gradients (HOG3D) [2] and HOGHOF [3], become popular in action recognition since the inventions of spatio-temporal interest points detectors [4–7]. In contrast to holistic representations [8,9], local methods enjoy many advantages such as 1) avoidance of some preliminary steps, *e.g.*, background subtraction and target tracking required in holistic methods, and 2) resistance to background variation and occlusions.

The most widely used local methods, *e.g.*, the bag-of-word (BoW) model [1] and sparse coding (SC) [10–12], have obtained remarkable performance in image and object classification. Recently, refinements of BoW and SC as well as alternative techniques including the soft assignment coding (kernel codebooks) [13], Triangle assignment coding [14], localized soft-assignment coding (LSC) [15] and locality linear-constrained coding (LLC) [16], have been developed to forward the state-of-the-art. However, these developments mostly remain in the image domain, which makes transferring them to the video domain an urgent and promising task.

A simple non-parametric nearest neighbor (NN) based classifier, naive bayes nearest neighbor (NBNN) [17], was proposed in recently. By computing the ‘Image-to-Class’ rather than ‘Image-to-Image’ distance, NBNN is able to avoid quantizing local features in the BoW model. In contrast to learning-based classifiers, the non-parametric NBNN classifier requires no training phase thus no risk of overfitting the parameters. Recently, enhanced versions of NBNN, including the NBNN kernels [18] and the local NBNN [19], have also been developed. The NBNN family have shown excellent effectiveness in image and object recognition.

The Fisher kernel (FK) has recently drawn increasingly attention in the image domain and produced remarkable results for image

[☆] This paper has been recommended for acceptance by Ahmed Elgammal, Ph.D.
[*] Corresponding author.
E-mail address: ling.shao@ieee.org (L. Shao).

http://dx.doi.org/10.1016/j.imavis.2016.02.006
0262-8856/© 2016 Elsevier B.V. All rights reserved.

classification [20–22]. It is shown in a recent study on feature coding [23] that the improved Fisher kernel (IFK), which is also called Fisher vector (FV), outperforms all the other encoding methods on several image datasets. Another important encoding method is the vector of locally aggregated descriptors (VLAD) introduced by Jégou et al. [24,25]. VLAD can be regarded as a simplified non-probabilistic version of Fisher vector and has shown comparable results with IFK.

Match kernels between sets of local features have long been exploited in visual recognition [26,27]. Without relying on any mid-level feature representations, match kernels are able to compute the similarity between sets of unordered local features and have shown the effectiveness in image and object recognition. More importantly, match kernels provides a basic formulation of measuring two sets of local features, based on which local methods are connected. The newly proposed feature coding techniques have been widely used and demonstrated their effectiveness in the image domain, however, their performance on action recognition has not been comprehensively evaluated and compared. Motivated by this, in this paper, we transfer these prevailing techniques from the image domain to the video domain and put them under a unified evaluation framework with the same experimental settings. In contrast to the previous evaluations [5,28–30], we focus on the evaluation of state-of-the-art local methods, *e.g.*, the BoW model, sparse coding, Fisher kernels, VLAD, NBNN and match kernels, based on spatio-temporal local features for human action recognition.

Recently, methods using tracking of trajectories has been used for action recognition which can always outperform those based on STIPs while requiring higher computational complexity [31]. In addition, it is found by Reddy and Mubarak [32] that motion based descriptors are not scalable with respect to the number of action categories, which can be reasonably assumed to also hold for trajectory-based sampling of descriptors. As we concentrate on the comparison of representation methods rather than the overall performance, we follow a standard paradigm for action recognition using local features [28,29], and apply the same feature detection and description steps to all the methods to be evaluated.

1.1. Contributions

We systematically evaluate the performance of representative local methods, some of which have not been used for action recognition yet. Extensive experimental results have been reported on three widely used benchmark action datasets, *i.e.*, KTH, UCF-YouTube and HMDB51. To the best of our knowledge, we, for the first time, pull local methods under a unified setting and conduct a comprehensive study both theoretically and experimentally for action recognition.

The main contributions of this paper lie in the following three aspects: **1)** we have conducted a comprehensive study on state-of-the-art local methods for human action recognition, which serves as a baseline for research in this field; **2)** we provide in-depth analysis and draw impartial conclusions from the findings in the experiments, which offers an important guide for further work on human action recognition; **3)** we provide a timely review on the recent advancement of local methods based on spatio-temporal local features, which can be used as an up-to-date reference for the community of action recognition.

2. Review of local methods

During the last decade, action recognition with local spatio-temporal interest points (STIPs) have been extensively explored. To give an overview of the advancement of local features for human action recognition, we will provide a review of recently developed local methods both within and beyond the BoW model. In the following, we will give a more detailed description of these methods.

2.1. The BoW model

The BoW model is a widely used algorithm for local representations and has proven to be successful in many action recognition tasks. However, local representations also suffer from many limitations. One of the most notorious deficiencies is that it fails to capture adequate structural and temporal information. In order to compensate for the loss of structures in local representations, a lot of methods try to improve local representations by exploring spatio-temporal structural information [33], including context information of each interest point [34,35], relationships between/among spatio-temporal interest points [36–39] and neighborhood-based features [40]. The relationship among visual words in the BoW model and their semantic meaning have also been explored to encode higher-level features [15,41–43]. New local descriptors have also been developed [44,45] to improve the performance of local methods.

Sun et al. [34] proposed to model the spatio-temporal context information in a hierarchical way by exploiting three levels of context, namely, point-level, intra-trajectory and inter-trajectory context. In their work, trajectories are first extracted using Scale Invariant Feature Transform (SIFT). The point-level context is the average of SIFT descriptors extracted at the salient points on the trajectory. Intra-trajectory and inter-trajectory context is modeled by the transition matrix of a Markov process and encoded as the trajectory transition and trajectory proximity descriptors.

In order to capture the most informative spatio-temporal relationship between local descriptors, Kovashka and Grauman [40] proposed to learn a hierarchy of spatio-temporal neighborhood features. The main idea is to construct a higher-level vocabulary from new features that consider the hierarchical neighboring information around each interest point.

Matikainen et al. [36] proposed to express pair-wise relationships between quantized features by combining the power of discriminative representations with key aspects of naive Bayes. The relationship between local features is modeled as the distribution of quantized location differences between each pair of interest points. Two basic features namely STIP-HOG and quantized trajectories are considered.

Gaur et al. [33] modeled the activity in a video as a “string of feature graphs” (SFGs) by treating a video as a spatio-temporal collection of primitive features (e.g., STIP features). They divide the features into small temporal bins and represent the video as a temporally ordered collection of such feature-bins, each bin consisting of a graphical structure representing the spatial arrangement of the low-level features. A video then becomes a string of such graphs and comparing two videos is to match two strings of graphs.

Claiming that the higher-order semantic correlation between mid-level features (e.g., from the BoW representation) is useful to fill the semantic gap, Lu et al. [42] proposed novel spectral methods to learn latent semantics from abundant mid-level features by spectral embedding with nonparametric graphs and hypergraphs. A new semantics-aware representation (i.e., histogram of high-level features) is derived for each video from the original BOW representation, and actions are classified by a SVM with a histogram intersection kernel based on the new representation.

Wang et al. [38] presented a novel local representation by augmenting local features with contextual features, which capture the interactions between interest points. Different from previous work on mining contextual information is considered as spatio-temporal statistics in the 3D neighborhood of each interest point. Multi-scale channels of contextual features are computed and, for each channel, a regular grid is used to encode spatio-temporal information in the local neighborhood of an interest point. Multiple kernel learning is employed to integrate the contextual features from different channels.