Computer Vision and Image Understanding 000 (2016) 1–13

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Fast action retrieval from videos via feature disaggregation

Jie Qin[a], Li Liu[b], Mengyang Yu[b], Yunhong Wang[a], Ling Shao[b,*]

[a] *Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, China*
[b] *Computer Vision and Artificial Intelligence Group, Department of Computer and Information Sciences, Northumbria University, UK*

ARTICLE INFO

Article history:
Received 10 December 2015
Revised 14 September 2016
Accepted 14 September 2016
Available online xxx

Keywords:
Similarity search
Video retrieval
Feature disaggregation
Learning based hashing

ABSTRACT

Learning based hashing methods, which aim at learning similarity-preserving binary codes for efficient nearest neighbor search, have been actively studied recently. A majority of the approaches address hashing problems for image collections. However, due to the extra temporal information, videos are usually represented by much higher dimensional (thousands or even more) features compared with images, causing high computational complexity for conventional hashing schemes. In this paper, we propose a simple and efficient hashing scheme for high-dimensional video data. This method, called Disaggregation Hashing (DH), exploits the correlations among different feature dimensions. An intuitive feature disaggregation method is first proposed, followed by a novel hashing algorithm based on different feature clusters. Additionally, a kernelized version of DH is proposed for better performance. We demonstrate the efficiency and effectiveness of our method by theoretical analysis and exploring its application on action retrieval from video databases. Extensive experiments show the superiority of our binary coding scheme over state-of-the-art hashing methods.

1. Introduction

With the rapid development of digital media, massive collections of images/videos are created every day. This poses lots of challenging problems to the computer vision community, among which one major challenge is, given an image/video, how to efficiently find its similar ones. For instance, Fig. 1 illustrates a typical scenario of retrieving similar actions from the video database. To address the problem of similarity search, one conventional way is utilizing nearest neighbour search. Tree-based schemes have also been widely studied to improve the efficiency of searching. However, these techniques are not scalable to high-dimensional data as they cannot easily find the approximate structure of the sparse data in the high-dimensional space. To overcome this challenge, a variety of hashing algorithms have been actively studied. The hashing techniques aim at efficiently embedding data from the high-dimensional space to a low-dimensional space, while preserving similarities among data points. After obtaining compact data representations (i.e., binary codes), approximate nearest neighbor (ANN) search can operate with constant or sub-linear time complexity.

Nevertheless, a majority of hashing techniques (Gong and Lazebnik, 2011; Heo et al., 2012; Indyk and Motwani, 1998; Lee et al., 2014; Liu et al., 2011; Wang et al., 2014a; Weiss et al., 2009) have been specifically developed for image retrieval. Compared with images, representations of videos can be more sophisticated and complex. By employing leading feature encoding techniques (e.g., Bag of Words (BoW) (Lazebnik et al., 2006), Vector of Locally Aggregated Descriptors (VLAD) (Jegou et al., 2010), and Fisher Vectors (FV) (Perronnin et al., 2010b)), representations of thousands or even more dimensions will be generated. Although the time complexity of ANN search can be significantly reduced by employing hashing techniques, learning hash functions in the training stage can become quite time-consuming and even intractable when dealing with very high-dimensional video data, which usually involves complex iterative optimization, eigen-decomposition, etc. Besides, a majority of hashing methods are based on linear projections (e.g., random projection Indyk and Motwani, 1998). In terms of mapping the data from the very high-dimensional space to a reduced one, the memory requirements for storing the projection matrix and performing the mapping operation impose heavy burdens. Taking the state-of-the-art hashing method, ITQ (Gong and Lazebnik, 2011), as an example, to reduce 100,000-dimensional features to 10,000-dimensional ones, the memory cost for storing the projection matrix is about 7.5GB and the number of multiplications for coding a test data point is 10^9.

* Corresponding author.
E-mail addresses: qinjiebuaa@gmail.com (J. Qin), li2.liu@northumbria.ac.uk (L. Liu), m.y.yu@ieee.org (M. Yu), yhwang@buaa.edu.cn (Y. Wang), ling.shao@ieee.org (L. Shao).

http://dx.doi.org/10.1016/j.cviu.2016.09.009
1077-3142/© 2016 Published by Elsevier Inc.

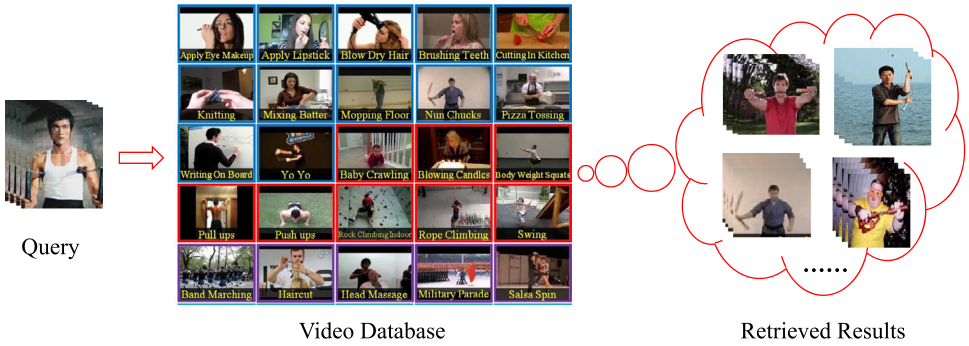

Fig. 1. Action retrieval from videos. Given a video sequence containing a certain type of action as the query, our goal is to retrieve similar action sequences from the video database.

To overcome these shortcomings, in this paper, a novel hashing scheme, namely Disaggregation Hashing (DH), is proposed for high-dimensional video data by exploiting the correlations among different feature dimensions. We start with disaggregating the original high-dimensional feature into several low-dimensional feature clusters/groups where feature values are similar among all training data points. Subsequently, simple and efficient hash functions are learned to embed original feature vectors to compact binary codes separately based on each feature group. Since each group can be embedded independently, a substantial speed up can be guaranteed for learning hash functions in the training stage. Furthermore, in the testing stage, the proposed method only needs to store a D-dimensional projection vector and computing binary codes is of $O(D)$ complexity, where D is the dimension of the original data space. The main contributions of this paper include:

- We propose a novel hashing scheme, namely DH, for high-dimensional video data, which outperforms state-of-the-art methods on three realistic action datasets;
- Feature disaggregation/clustering is proposed to group similar feature dimensions of high-dimensional data by exploiting correlations along the dimensions;
- Hash functions, aiming at preserving the intrinsic data similarity, are learned independently using greedy optimization on each feature group, which can significantly reduce the computational complexity and memory usage in both the training and testing stages;
- A kernelized version (KDH) has also been proposed for better performance.

A preliminary version of this work was presented in Qin et al. (2015). Compared with the conference version, the extensions include a soft feature clustering strategy (Section 3.1), a simple randomized hash function (Section 3.2), a kernelized version of our original method (Section 3.5), and more comprehensive experiments with empirical results including evaluations on a typical image dataset (Section 5.1), and evaluations of variants of our overall hashing framework (Sections 5.3 and 5.4).

The remainder of the paper is organized as follows. We give a brief review of previous hashing methods in Section 2. In Section 3, we introduce the proposed hashing scheme, which includes two steps: i.e., feature disaggregation and hash function learning. The kernel trick is further incorporated into the proposed hashing algorithm. Section 4 introduces the datasets and experimental setup. The experimental results are demonstrated and discussed in Section 5, and we finally draw our conclusions in Section 6.

2. Related work

In order to address the problem of fast approximate nearest neighbor search, a significant amount of work has been focusing on embedding the original data into compact binary codes, while preserving similarity among the original data. Locality Sensitive Hashing (LSH) (Indyk and Motwani, 1998) is one of the most widely employed hashing methods. LSH is mainly based on random projections that project data points that are close in the Euclidean space to similar codes. Many extensions of LSH have also been proposed such as LSH with p-stable distributions (Datar et al., 2004) and shift-invariant kernel-based LSH (Raginsky and Lazebnik, 2009).

However, the performance of LSH-like methods is limited because the data distribution is not taken into account. Thus, a number of data-dependent or learning-based hashing methods have been proposed, aiming at better fitting data distributions to achieve better performance. For instance, Spectral Hashing (Weiss et al., 2009) proposed by Weiss et al. formalizes the hashing task as a particular form of graph partitioning, in which the Laplace–Beltrami eigenfunctions of manifolds are used to determine binary codes. Motivated by Liu et al. (2011) utilize Anchor Graphs to learn compact binary codes by directly approximating the sparse neighborhood graph and its associated adjacency matrix. Another popular hashing approach named iterative quantization (ITQ) (Gong and Lazebnik, 2011) has connections to the Procrustes problem and performs compact binary embedding by minimizing the quantization error of rotating zero-centered data projected by principal component analysis (PCA). Apart from the above approaches which are based on hyperplanes, Heo et al. (2012) introduce a hashing scheme based on hyperspheres to map more spatially coherent data points into a binary code.

As aforementioned, these hashing methods are specifically developed for image search problems, thus not applicable for content-based video retrieval, e.g., action retrieval in videos. In recent years, extensive efforts have been devoted to action recognition (Chen et al., 2015, 2016; Karpathy et al., 2014; Liu et al., 2016; Qin et al., 2016) and detection (Jain et al., 2014; Soomro et al., 2015; Wang et al., 2014b). For instance, Wang et al. (2014b) propose to efficiently detect actions in videos using spatio-temporal tubes (ST-tubes), which are obtained based on mutual information of feature trajectories. Nevertheless, there are few studies on action retrieval in videos. Note that in Ramanathan et al. (2015), a neural network architecture is proposed for action retrieval which extracts relationships between actions through several important cues. However, (Ramanathan et al., 2015) only addresses action retrieval problems in images instead of videos. In this paper, we particularly focus on the retrieval problem of high-dimensional video data. Until now, there have been few works (Gong et al., 2013; Liu et al., 2015c; Xia et al., 2015; Yu et al., 2014) addressing hashing problems on high-dimensional video data. In Gong et al. (2013), compact bilinear projections are utilized instead of a single large