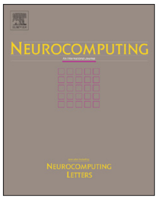
Neurocomputing 207 (2016) 202–212

Contents lists available at ScienceDirect

Neurocomputing

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



SERVE: Soft and Equalized Residual VEctors for image retrieval

Jun Li[a], Chang Xu[b], Mingming Gong[c], Junliang Xing[d], Wankou Yang[a], Changyin Sun[a,*]

[a] *School of Automation, Southeast University, Nanjing 210096, China*
[b] *The Key Laboratory of Machine Perception (Ministry of Education), School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China*
[c] *The Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia*
[d] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

ARTICLE INFO

Article history:
Received 8 December 2015
Received in revised form
27 February 2016
Accepted 18 April 2016
Communicated by Mingli Song
Available online 13 May 2016

Keywords:
SERVE
Manifolds
Multi-graph embedding
Graph ensemble
Image retrieval

ABSTRACT

In the last decade, a wide variety of image signatures, e.g., Bag-of-Visual-Words (BOVW), Fisher Vector (FV), and Vector of Locally Aggregated Descriptor (VLAD), have been developed for effective image retrieval. These image signatures, however, are either computationally expensive or simplified for the purpose of trading accuracy for efficiency. To simultaneously guarantee efficiency and effectiveness, we propose a novel image signature termed Soft and Equalized Residual VEctors (SERVE) which is more discriminatively formulated and maintains higher accuracy. It improves VLAD by encoding the variability in within-cluster feature points into the summation of Residual Vectors (RV) while manifesting superiority in computational efficiency over FV. To find the latent low-dimensional manifolds underlying in the SERVE feature space, we propose to partition the original feature space into separate subspaces by random projections and employ multi-graph embedding to obtain additional performance gain. In particular, we make use of two fusion strategies for graph ensemble to generate a holistic representation. Extensive empirical studies carried out on the three retrieval-specific public benchmarks reveal that our method outperforms existing state-of-the-art methods and provides a promising paradigm for the image retrieval task.

1. Introduction

Content-based image retrieval has become a major research topic in the domains of computer vision and multimedia [1–4]. It has been demonstrated that massive successful real-world image retrieval systems largely rely on powerful mid-level image signatures pooled from distinctive local descriptors [2,5–7].

The early Bag-of-Visual-Words (BOVW) model [2] simply counts the occurrences of visual words in a codebook [8,9], which yields a histogram for holistic representation. Unlike the BOVW, the Fisher Vector (FV) captures the layout of local features by Gaussian Mixture Model (GMM) and obtains the gradient vector by taking derivatives with respect to the Gaussian mixture parameters [10]. Despite suffering from high computational costs, it achieves superior performance in the scenarios of both object categorization and image retrieval [5,10]. As a non-probabilistic simplified version of FV, the Vector of Locally Aggregated Descriptors (VLAD) achieves comparable performance to FV with a much lower computational overhead [6]. It is therefore extensively applicable to fast image search [11,12].

As shown above, either accuracy or efficiency is sacrificed in terms of both FV and VLAD. As a result, devising an efficiently formulated and discriminative signature is still an open problem. In this paper, we develop a novel image signature termed Soft and Equalized Residual VEctors (SERVE). It combines the advantages of both FV and VLAD by producing a more computationally tractable formulation than FV and fully adapting VLAD to a soft probabilistic version. Specifically, it alleviates the quantization error in VLAD aggregation and decreases the computational costs in FV computation without sacrificing discriminating power. Thus, it can be used as an encouraging substitute for the two classical signatures.

Although SERVE can be used as a preferable formulation for image description, it is critical to uncover the intrinsic manifold structures underlying the high-dimensional feature space for the low-dimensional representation. Based on SERVE, we propose in this paper to decompose the original high-dimensional feature

[*] Corresponding author.
E-mail addresses: lijunautomation@gmail.com (J. Li), changxu1989@gmail.com (C. Xu), gongmingnju@gmail.com (M. Gong), jlxing@nlpr.ia.ac.cn (J. Xing), wankou.yang@yahoo.com (W. Yang), cysun@seu.edu.cn (C. Sun).

http://dx.doi.org/10.1016/j.neucom.2016.04.047
0925-2312/© 2016 Published by Elsevier B.V.

space into separate subspaces by random projections and adopt multi-graph embedding for generating manifold representation. In addition, two graph ensemble techniques for merging multiple graphs built from different subspaces are presented, which leads to effective multi-graph embedding. As an image is generally composed of multiple independent components, multi-graph embedding helps to capture the intrinsic relationship among them, which is beneficial for mining the underlying data manifolds [13]. Thus, SERVE and the fused graph representation compose the two crucial components for image feature generation. The former serves as a well-descriptive intermediate representation while the latter is further developed from this base feature to uncover the latent low-dimensional data structure. Therefore, the discriminative information in both original feature space and the transformed manifolds can be maximally preserved by using the final graph ensemble as image feature.

As recent computer vision systems considerably profit from multiview data fusion, in implementation, we adopt three heterogeneous low-level features, i.e. RootSIFT, LCS and LBP, for a comprehensive description of the visual cues in an image. The three features lead to the respective SERVE signatures and multi-graph embedding is performed accordingly for manifold approximation. Then, classical fusion strategies are used to merge the three-view manifold representations for further performance improvement. The comparative study implemented on the three public benchmarks reveals the superiority of SERVE over the state-of-the-art image representations. In particular, it offers higher performance boosts combined with multi-graph embedding.

The major contributions of this work are summarized as follows:

- We propose a novel and discriminative signature called SERVE for image retrieval.
- We impose multi-graph embedding on SERVE for manifold approximation by random projections.
- We utilize two effective graph ensemble schemes to yield fused graph representation.

The processing pipeline of our approach is illustrated in Fig. 1. In the following sections, we will first review the related work (Section 2) and then elaborate the generation of our SERVE signature (Section 3). Next, SERVE is integrated into the framework of multi-graph embedding in the image retrieval scenario (Section 4). Experimental results and analyses will focus on a comparison of SERVE with FV and VLAD, as well as the comparison of overall image retrieval performance and a number of competing methods (Section 5). Lastly, we conclude our paper and give an outlook on our future work (Section 6).

2. Related work

State-of-the-art image retrieval systems benefit from discriminative image representation and manifold approximation. In this section, we will briefly review related works from both aspects.

2.1. Discriminative image representation

Represented as a set of orderless codewords, the BOVW model has become a paradigm for real-world object and image retrieval. In the context of large-scale retrieval tasks, visual codewords can be efficiently structured offline and indexed by inverted files [1,14]. Fast image search can be achieved by traversing the inverted file, and distinctive similarity evaluation is performed by using the TF-IDF weighting scheme [15]. Despite its simplicity, the BOVW model still suffers from a number of limitations, e.g. negative evidence [16], visual word burstiness [17] and asymmetrical dissimilarity [18]. In addition, BOVW is merely a histogram representation in which the spatial layout of local features is not encoded. Consequently, a large body of work has been conducted to embed the spatial contexts of features into the BOVW framework for robust geometric verification [19–21]. Furthermore, supplementary post-processing steps, e.g. relevance feedback [22,23], re-ranking [24,25] and query expansion [26], dramatically enhance the retrieval performance.

As a sophisticated alternative to BOVW, FV encodes high-order statistics by using generative Gaussian Mixture Models to model the generating process of the local features. It is in spirit the concatenation of the component-wise gradient vectors produced by taking derivatives with respect to the Gaussian mixture weight, mean and covariance. FV was initially proposed in scene classification [10] and then extended to large-scale image retrieval application by being compressed into compact binary codes [5]. Furthermore, it has been improved for large-scale image classification by incorporating dual normalizations and spatial pyramid [27]. Analogously, Jégou et al. invented a simplified image feature called VLAD [6]. As a non-probabilistic form of FV, VLAD computes the residual vectors for each visual word and aggregates them into a single vector. It rivals FV in terms of retrieval performance, while having much lower computational complexity. As the original VLAD formulation is far from optimal, substantial advances have been made to improve the performance of VLAD [11,12,28,29].

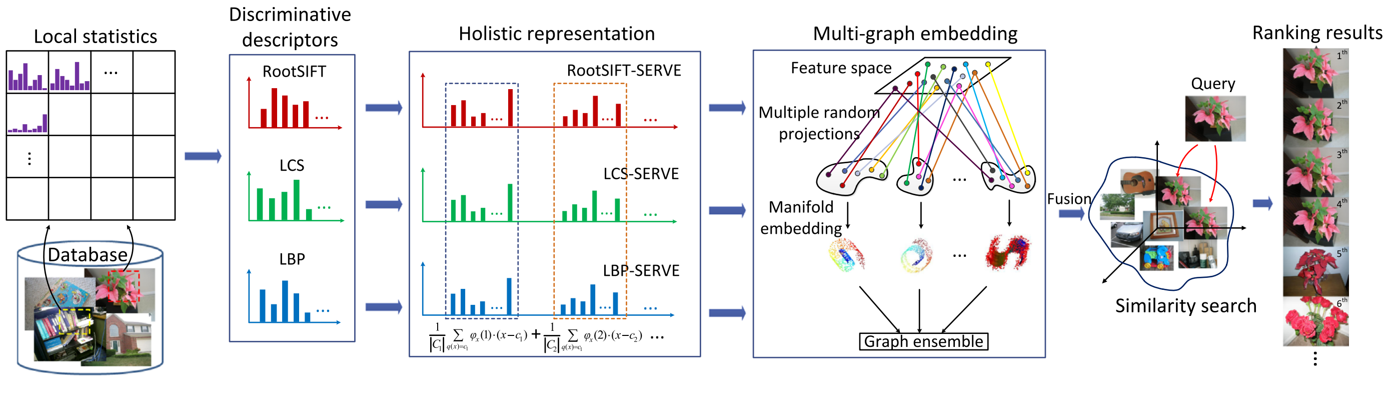


Fig. 1. Overview of our approach for effective image retrieval. We start by extracting low-level features and aggregating them into respective SERVE signatures. Then, multi-graph embedding is performed for manifold approximation which involves the following three steps. First, the original high-dimensional feature space is partitioned into separate subspaces by multiple random projections. Next, spectral embedding is carried out on all subspaces for manifold approximation. Lastly, the final representation is derived by using graph ensemble method.