Pattern Recognition Letters 000 (2017) 1–6

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Metric learning via feature weighting for scalable image retrieval

Xiaoming Lv, Fajie Duan*

State Key Laboratory of Precision Measuring Technology & Instruments, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Metric learning
Multiple features
Image retrieval

ABSTRACT

Two dominant image retrieval schemes are based on local features indexed by an inverted index and global features indexed by compact hashing codes. They both demonstrate excellent scalability, but distinct strength for image retrieval. This motivates us to investigate how to fuse these two search schemes, to further enhance the retrieval effectiveness. Thus, we propose a novel metric learning method, namely Metric Learning via Feature Weighting (MLFW), to effectively fuse different features. MLFW learns the distance metric on individual feature as well as the weights of different features in a joint framework, to combine the learned distance obtained from all the individual feature and the early fusion. Furthermore, we design an efficient solution to optimize the objective function. Extensive experimental results conducted on real-life datasets show that the proposed MLFW outperforms the state-of-the-art methods in terms of search quality.

1. Introduction

Image retrieval based on visual features has long been a major research theme due to the many applications such as the web image search [28] and near-duplicate image detection [27]. Image retrieval is typically accomplished in three steps: feature extraction, searching and post-processing. Over the last few years, a long stream of research efforts have been made to improve these three components [4,11].

From the perspective of image representation and search scheme, most of the successful scalable image retrieval algorithms fall into two categories [38]: (1) quantized local features [24] indexed by an inverted index [19]; and (2) global features [28] indexed by compact hashing codes [20,30]. These two approaches demonstrate distinct strengths in finding visually similar images. Vocabulary tree based methods have advantages in identifying near duplicate images or regions since local features are particularly capable of attending to local image patterns or textures, but they are less robust to the similar textures [28]. On the other hand, global signatures can retrieve the candidates that appear alike at a glance, but they are sensitive to changes in contrast, brightness, scale, rotation, camera viewpoint, and so on [28]. The complementary descriptive capability of local and global features naturally raises the question of how to integrate their strengths to yield more satisfactory retrieval results.

Although both lines of retrieval methods using single feature have been extensively studied, there is not much research effort focusing on the fusion of multiple features for the task of image retrieval. Fusion mechanisms can be grouped into two types which are feature-level fusion and decision-level fusion. In the feature-level fusion [35], a combination of multiple features at the input stage is used to obtain a single feature. The late fusion algorithms [34] first generate separate results from different features, and then combine these results together by different strategy. The correlations among features are not considered. One problem with feature-level fusion for image retrieval task is that it will lose indexing property of each feature. For example, local features are usually followed by BoW search scheme and global features are usually indexed by compact binary codes. They either simply concatenate the features together with equal weight and each part is indexed separately, or combine them into a new feature, but the indexing property of each feature is lost. When different features are fused, their indexing property might be lost. On the other hand, the performance of decision-level fusion is prone to be degraded by a bad-performed feature.

For post-processing, after the image retrieval return a list of images, a potential re-ranking techniques can be used to reorder the initial results. But this procedure is not mandatory, and considering the time limitation, there are not one or some universal methods. Some representative methods are geometric verification [39], query expansion [25] and relevance feedback [41].

In this paper, we focus on the feature extraction and searching. More specifically, we propose a novel distance metric learning method on multiple features, namely Metric Learning via Feature

* Corresponding author.
E-mail addresses: lvxiaoming1@gmail.com (X. Lv), fjduan@tju.edu.cn (F. Duan).

http://dx.doi.org/10.1016/j.patrec.2017.09.026
0167-8655/© 2017 Published by Elsevier B.V.

Weighting (MLFW), to effectively fuse different features for image retrieval. MLFW learns the distance metric on individual feature as well as the weights of different features in a joint framework, to combine the learned distance obtained from all the individual feature and the early fusion. Furthermore, an efficient optimization method is designed to solve the objective function for efficient retrieval.

It is worth noting the following contributions:

- We propose a novel distance metric learning method (MLFW) by fusing multiple features. MLFW can learn the distance metric on individual feature as well as the weights of different features in a joint framework. In this way, the learned distance obtained from all the individual feature are combined and weighted to gain a more accurate data-specific distance metric.
- We design an efficient optimization method to solve the objective function.
- Extensive experimental results demonstrate the superiority of our proposed method.

The remainder of this paper is organized as follows. In Section 2, we discuss the related work. MLFW is presented in Section 3. Extensive experiment results are given in Section 5. Lastly, we draw a conclusion in Section 6.

2. Related work

Most of the scalable image retrieval algorithms fall in two threads: indexing local features by a vocabulary tree and hashing holistic features by binary codes. Their strengths and limitations as well as possible ways to combine them are briefly reviewed below.

2.1. Global feature with compact hashing

As introduced in [29], holistic features such as color histograms and GIST are indexed by locality sensitive hashing [1], resulting in highly compact binary codes (e.g., 128 bits), which can be efficiently compared with a large database using the Hamming distance. The scalability and performance have been improved by spectral graph partitioning and hashing [30] and incorporating the pairwise semantic similarity and dissimilarity constraints from labeled data. As suggested in [13], a random rotation on the PCA-projected features, which is optimized by iterative quantization, achieves surprisingly good performance. These methods leveraging compact hashing of holistic features are efficient in computation and memory usage. However, holistic features tend to be less invariant than local features, and are in general more sensitive to image transformations induced by illumination changes, scaling and pose variations. In practice, the focus on aggregated image statistics rather than fine details results in images that appear roughly similar but the retrieval precision is often lower compared to local feature based methods. Just as its name, the global features can only catch some global information about the images, which is not concrete, and usually cannot deal with the change in illumination, viewpoint. So if only use global features, the retrieval results are usually not satisfied. And now almost all the researchers pay attention to local features.

2.2. Local feature with quantization

Image retrieval based on the BoW of local invariant features [24] has been significantly scaled up by using vocabulary trees [19] which contain millions of leaf nodes attached with inverted indexes. This method demonstrates an excellent scalability in computation and precision, although it is memory consuming. It has been further improved by a spatial verification by RANSAC [22]; the query expansion [25]; using Hamming embedding and weak geometry constraints [39]; constructing high-order features [33]; and indexing relative spatial positions [40] among local features. Since images are essentially delineated by local invariant features, these methods are effective in handling image scaling, rotation, and partial occlusions, leading to a very high precision in near-duplicate image retrieval. However, if no near-duplicate image regions exist in the database, large areas of similar textures may confuse these retrieval methods and lead to irrelevant candidate images and unsatisfactory user experience.

Table 1
Notations.

Notation	Description
n	the number of data points
m_v	the dimension for the vth feature
p_v	the reduced dimension for the vth feature
x_i, x_i^v	a data point and its vth feature
y_i	the label for a data point
α, α_v	the weights, and the weight of the vth feature
M	the learned distance metric
W_v	the linear projection on the vth feature
β	the penalty parameter
λ	the regularization parameter

2.3. Multiple feature fusion

Given that multimedia data can be represented by multiple features, it has now become a trend to properly combine the evidences derived from different features [31]. In the field of machine learning, researchers have proposed many multi-feature fusion algorithms. Representative works include Canonical Correlation Analysis (CCA) [14], two-view support vector machines, i.e., SVM-2k [10] and their variants [16]. These algorithms have been applied to various applications, such as object recognition, image annotation and image-audio clustering, demonstrating satisfactory performance. However, these algorithms require a large amount of labeled data for training, which are often expensive and seldom available.

In the field of multimedia similarity retrieval, lots of task-specific methods are designed. In [37], a query specific late fusion strategy is proposed. In multiple feature hashing (MFH) [26], multiple features are mapped into common Hamming space for fast near-duplicate video retrieval. A group of hash functions are learned for multiple features by preserving the local structure information of each individual feature and also globally considering the local structures for all the features. Next, we give the details of our proposed method.

3. Multi-view metric learning

In this section, we present the details of MLFW. The general framework for MLFW is shown in Fig. 1, which comprises of two phases: the training phase and the testing phase.

3.1. Notations and definitions

Given a training dataset of n data points which are represented as d-dimensional vectors $\{x_i\}_{i=1}^n \in \mathbb{R}^d$, let X denote the matrix representing the whole dataset as a $d \times n$ matrix, i.e., $[x_1, \ldots, x_n]$ where each column represents a data point. Suppose that there are l classes. We use $y_i \in \{1, 0\}^l$ to represent the label for a data point x_i, where the jth element of y_i, denoted as y_{ij}, is 1 if x_i belongs to the jth class, or 0 otherwise. Let Y denote the $n \times l$ matrix whose ith row is the label for the ith data point, i.e., $[y_1, \ldots, y_n]^T$. For simplicity, a list of notations used in this paper are shown in Table 1.

Please cite this article as: X. Lv, F. Duan, Metric learning via feature weighting for scalable image retrieval, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.09.026