Image and Vision Computing 70 (2018) 11–20

Contents lists available at ScienceDirect

Image and Vision Computing

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

Learning deep similarity models with focus ranking for fabric image retrieval☆

Daiguo Deng[a], Ruomei Wang[a], Hefeng Wu[b,*], Huayong He[a], Qi Li[c], Xiaonan Luo[d]

[a] Sun Yat-sen University, Guangzhou 510006, China
[b] Guangdong University of Foreign Studies, Guangzhou 510006, China
[c] Western Kentucky University, Bowling Green, KY 42101, USA
[d] Guilin University of Electronic Technology, Guilin 541004, China

ARTICLE INFO

Article history:
Received 29 March 2017
Received in revised form 21 October 2017
Accepted 10 December 2017
Available online 13 December 2017

MSC:
00-01
99-00

Keywords:
Convolutional neural network
Fabric image retrieval
Metric embedding
Focus ranking

ABSTRACT

Fabric image retrieval is beneficial to many applications including clothing searching, online shopping and cloth modeling. Learning pairwise image similarity is of great importance to an image retrieval task. With the resurgence of Convolutional Neural Networks (CNNs), recent works have achieved significant progresses via deep representation learning with metric embedding, which drives similar examples close to each other in a feature space, and dissimilar ones apart from each other. In this paper, we propose a novel embedding method termed *focus ranking* that can be easily unified into a CNN for jointly learning image representations and metrics in the context of fine-grained fabric image retrieval. Focus ranking aims to rank similar examples higher than all dissimilar ones by penalizing ranking disorders via the minimization of the overall cost attributed to similar samples being ranked below dissimilar ones. At the training stage, training samples are organized into focus ranking units for efficient optimization. We build a large-scale fabric image retrieval dataset (FIRD) with about 25,000 images of 4300 fabrics, and test the proposed model on the FIRD dataset. Experimental results show the superiority of the proposed model over existing metric embedding models.

1. Introduction

Fabric image retrieval, as a special case of generic image retrieval, benefits a wide range of applications, e.g., large-scale clothing searching, online shopping and cloth modeling. Given a query fabric image, a retrieval algorithm is expected to output fabric images that are identical or similar to the query image. It remains a challenging task due to severe variations in illuminations, orientations, scales and wrinkles among fabric images.

Traditional methods for image retrieval mainly involve two critical components: i) design a robust and discriminative image representation, and ii) determine an effective distance or similarity metric for a given image representation. Image representations used in a traditional method are usually hand-crafted, e.g., SIFT [1], GIST [2,3], Bag of Words (BoW) [4], Fisher Vector (FV) [5,6], and VLAD [7]. Although achieving reasonable successes in image retrieval, these methods depend heavily on feature engineering. More seriously, the two components are designed or learned separately, leading to a sub-optimal solution.

Recently, methods were proposed to learn an image representation and a distance or similarity metric jointly [8,9] based on Convolutional Neural Networks (CNN) [10,11], which can be used seamlessly in image retrieval. Specifically, these methods trained a CNN with metric learning embedding. Two simple yet effective metric learning embedding methods are pair [12,13] and triplet embedding [14,15]. These two embedding methods are optimized to pull samples of different labels apart from each other and push samples with the same labels close to each other. The most important advantage of these discriminative models is that they can jointly learn an image representation and semantically meaningful metric, which is more robust against intra-class variations and inter-class confusions.

An image retrieval system aims to find out samples with the same labels against a great many negative ones. But pair and triplet embedding methods model a metric with no more than one negative image as reference, which are extremely rough approximation to a real setting. In this paper, we propose a novel embedding method named *focus ranking*, which can easily be unified into a CNN for a joint optimization. In particular, the proposed model aims to rank a sample

☆ This paper has been recommended for acceptance by Sinisa Todorovic, PhD.
* Corresponding author.
E-mail address: wuhefeng@gmail.com (H. Wu).

https://doi.org/10.1016/j.imavis.2017.12.005
0262-8856/

with the same label (i.e., the matched sample) top over all negative ones. Hence, we penalize any ranking disorder by minimizing the overall cost of the matched samples ranked below any negative ones. At the training stage, we organize training samples into focus ranking units, each of which consists of a probe sample, a matched sample, and a reference set, for an efficient optimization. It learns to rank the matched sample top over all the negative ones in the reference set.

To the best of our knowledge, there are not fabric image retrieval datasets publicly available. We build a large-scale fabric image retrieval datasets (FIRD). It contains 4300 fabrics, each of which has five to ten instances. We divide the FIRD dataset into two partitions, via randomly selecting half of the fabrics as the training set and the rest as the test set. We randomly select 2/5 images as queries for each fabric in the test set and the rest form the retrieval set. Given a query image, we aim to find out the ones of the same fabric at the retrieval set. Extensive experiments on FIRD demonstrate that the proposed model outperforms existing ones by a large margin.

In summary, this paper makes four main contributions to the community.

- We propose a novel focus ranking embedding method, which aims to rank the matched sample top over any negative ones. This method can be easily unified into a CNN to learn a deep similarity model by optimizing the image representation and metric jointly. Compared to existing embedding methods, it can learn more robust and discriminative representation to reduce intra-class variations and enhance inter-class discrimination.
- We apply focus ranking unit generation to model training, which highlights the focus by the negative-positive ratio in the process of training. It can also help increase the diversity of training samples, which is greatly important to learning.
- We build a large-scale fabric image retrieval dataset (FIRD), which contains 4300 fabrics and about 25,000 images for training and test. To the best of our knowledge, it is the first large-scale dataset for this task.
- Extensive experiments demonstrate the superiority of the proposed method over existing metric embedding methods. We also carefully evaluate and discuss key components of our model that improve the performance.

The rest of this paper is organized as follows. Related works are reviewed in Section 2. The proposed model is presented in Section 3, and the FIRD dataset is described in Section 4. Section 5 presents an extensive evaluation and comparison between the proposed method and several state-of-the-art methods, in addition to the component analysis of the proposed method. Section 6 concludes the paper.

2. Related works

In this section, we review three topics most related to this work according to its objective and the technical components. These three topics are traditional descriptors, deep representation learning and distance learning.

2.1. Traditional descriptors

Traditional image retrieval systems mainly focus on designing discriminative hand-crafted image representations for matching two images. Scale-Invariant Feature Transform (SIFT) [1] was used widely in image retrieval, due to its distinct invariance to image translation, scaling, and rotation. Many works applied Bag of Features (BOF) using SIFT features for large scale image retrieval, such as Ref. [4]. They first quantized local feature vectors into visual words, and represented the image by the frequency histogram of visual words. Nister and Stewenius [16] hierarchically quantized local descriptors to a hierarchical vocabulary, which improves not only the retrieval efficiency but also its quality. Jegou et al. [17] presented a contextual dissimilarity measure for accurate and efficient image search, while Fraundorfer et al. [18] proposed a binning scheme for fast hard drive based image search. Jegou et al. [19,20] developed a more precise representation by integrating hamming embedding and weak geometric consistency within the inverted file. Some works also aggregated a set of local descriptors into global ones for large scale image retrieval, e.g., Fisher Vector (FV) [5,6], and Vector of Locally Aggregated Descriptors (VLAD) [7]. Perronnin et al. [21] represented an image as a Fisher Vector, and compressed Fisher Vectors using the binarization technique to reduce their memory footprint and speed up the retrieval. VLAD can jointly optimize dimensionality reduction and indexing algorithms, and thus produce a more compact representation for image retrieval. Although achieving significant progress in visual retrieval, these works depended heavily on hand-crafted features, which were not always optimized for particular tasks.

2.2. Deep representation learning

Recently, deep representation learning has been successfully applied to various computer vision areas, such as image classification [10,22,23], object detection [24–26], pixel-wise image labeling [27,28] and human centric analysis [29,30]. We review some recent works that applied deep learning to image retrieval [8,31-35]. Babenko et al. [31] extracted global features for image retrieval from the fully connected layer of CNN pre-trained with an image classification dataset [36], and demonstrated fine-tuning the network with annotated target images could boost the retrieval performance. Their succeeding work [32] simply aggregated the average and max pooling of the last convolutional layer as image representation and achieved better results than using fully convolutional layer. Perronnin and Larlus [34] proposed a hybrid image retrieval architecture that combined the strengths of supervised deep representation and unsupervised Fisher Vectors [5,6]. However, these methods extracted deep features using the CNN trained with a classification objective function, which merely focused on feature learning and ignored metric learning.

2.3. Distance learning

Distance metric learning (DML) plays an important role in many computer vision tasks such as face recognition, person re-identification and image retrieval. Xing et al. [37] proposed to learn a distance metric from given examples of similar pairs and demonstrated the learned distance metric can improve clustering performance significantly. Zheng et al. [38] formulated distance learning as a probabilistic relative distance comparison model to maximize the likelihood in which the examples from a matched pair have smaller distance than those from a mismatched pair. Mignon and Jurie [39] proposed a new distance learning method with sparse pairwise constraints. The success of deep learning also inspired recent works that applied neural network models to address the distance learning problem [40–42].

The following works are related to our work in spirit of deep metric embedding. Hu et al. [43] presented a new discriminative deep metric learning (DDML) method for face verification in the wild. Bell and Bala [41] learned deep metric for visual search in interior design using contrastive embedding. In Ref. [44], Rippel et al. proposed to maintain an explicit model of the distributions of different classes in representation space and employ it to achieve local discrimination in DML by penalizing class distribution overlap. Wang et al. [8] used a multi-scale CNN trained using triplet embedding for learning fine-grained image similarity directly from the image pixels. Similar ideas