Pattern Recognition 76 (2018) 537–548

Contents lists available at ScienceDirect

Pattern Recognition

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

Sketch-based image retrieval with deep visual semantic descriptor

Fei Huang[a], Cheng Jin[a], Yuejie Zhang[a,*], Kangnian Weng[b], Tao Zhang[b], Weiguo Fan[c]

[a] School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[b] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China
[c] School of Accounting and Information Systems, Virginia Tech, Blacksburg, VA, USA

ARTICLE INFO

Article history:
Received 5 February 2017
Revised 21 August 2017
Accepted 30 November 2017
Available online 1 December 2017

Keywords:
Sketch-based image retrieval (SBIR)
Deep learning
Deep visual semantic descriptor
Sketch-like transformation
Re-ranking optimization
Multiple feature fusion
Accelerated hierarchical K-means clustering

ABSTRACT

Sketch-based Image Retrieval (SBIR) has received a lot of attentions recently. In this paper we aim to enhance SBIR with deep visual semantic descriptor and related optimization mechanisms. Our scheme significantly differs from other earlier work in: 1) A feature representation via deep visual semantic descriptor is established to bridge the gap between sketches and images, which can encode both low-level local features and high-level semantic features; 2) A clustering-based re-ranking optimization is introduced to further improve SBIR by dynamically adjusting the correlations of images in the ranking list. The main contribution of our work is that we effectively apply the deep visual semantic descriptor to enable deep sketch-image matching, which has provided a more reasonable base for us to fuse local low-level visual features with high-level semantic features by determining an optimal correlated mapping. Our experiments on a large number of public data have obtained very positive results.

1. Introduction

With the rapid developments of Internet and mobile devices, more efficient image retrieval with billions of images available both online and offline becomes an important research topic [1,2]. Content-based Image Retrieval (CBIR) mainly uses existing RGB images as queries to search similar images. Text-based Image Retrieval (TBIR) simply utilizes plain texts as queries. However, a user's image search intention is often complicated, which cannot be easily formulated as a simple image query or an exact keyword query that could meet his/her real needs. Thus, query-by-freehand-sketch, i.e., Sketch-based Image Retrieval (SBIR), has become a more precise and convenient manner for image users to express what they need when lacking query images or textual descriptions, especially with the popularity of touch screen devices [3,4].

Sketches, unlike typical images, only contain main strokes or lines of target images and lack rich texture attributes and luminance, which result in the inherent ambiguity for given sketch queries among different users [5,6]. Although SBIR has been extensively studied since recent years [7,8], there are still remaining issues to be addressed simultaneously: 1) more effective visual conversions to improve the mutual resemblance between sketches and images; 2) more comprehensive feature descriptor constructions to bridge both the visual appearance gap and deep representation gap between sketches and images; and 3) more precise ranking optimization techniques to determine better corresponding relations between sketches and images [9–11]. To address the first issue, it's important to establish a robust visual transformation mechanism that can convert a general colorful image into a sketch-like form reasonably. To address the second issue, it's critical to learn an intermediate expression form and a deep representation mechanism to capture accurate visual and semantic feature information. To address the third issue, we need to develop an optimal ranking scheme with high accuracy but low cost, which can efficiently exploit the inter-related correlations among visual semantic attributes of sketches and images.

Inspired by the above observations and the success of deep learning in CBIR [12,13], in this paper we aim to employ deep learning to enhance SBIR with deep visual semantic descriptor and related optimization mechanisms. Moreover, how to integrate multiple information sources in SBIR for large-scale annotated images is still a very challenging issue. It is hard to provide a common base for the inter-related correlations between sketches and images due to the visual and semantic gap. Our work effectively applies a deep visual semantic descriptor to enable deep sketch-image matching, which has provided a more reasonable base for us to fuse local low-level visual features with high-level semantic features by determining an optimal correlated mapping. Such an SBIR framework with deep visual semantic descriptor can be treated as a bi-media correlation distribution over deep representations of

* Corresponding author.
E-mail address: yjzhang@fudan.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.patcog.2017.11.032
0031-3203/

sketches and images, i.e., to create more effective pairwise mapping and measure how they are correlated. The main contributions of this paper that significantly differ from other earlier works are the following four aspects: a) A sketch-like transformation for colorful images is established based on the contour detection and screening with the non-maxima suppression algorithm, in which the edge points that humans find important on colorful images are preserved to greatly improve the general sketch-image resemblance; b) A deep visual semantic descriptor is created to bridge the expression gap between sketches and images, which encodes both deep low-level local features and high-level semantic features and can be highly discriminative in making distinctions among dissimilar sketch-image pairs; c) A special clustering-based re-ranking optimization is introduced to further improve SBIR by accurately characterizing and dynamically adjusting the inter-related correlations between sketches and images in the ranking list; d) A new real-time SBIR framework is built by fusing the above transformation, representation and ranking patterns, which not only enables users to present on the sketch query panel whatever they imagine in their minds, but also returns the most similar images to the picture in users' minds. We have obtained positive results in our experiments to demonstrate our proposed framework.

The remainder of the paper is organized as follows. Section 2 reviews some related works. In Section 3 we describe in detail the enhanced SBIR framework on integrating multimodal visual and textual cues to construct deep visual semantic descriptor. Section 4 discusses the experimental settings, results, comparisons and analyses on two datasets. Finally, conclusion remarks and future works are given in Section 5.

2. Related works

Sketch-based image retrieval is not a novel task. It first emerged in the 1990s and has received sustained attentions since 1992 [14]. Early research studies mainly focused on exploiting unsupervised methods to extract low-level visual features for sketches and images [15]. Traditional SBIR approaches can be mainly divided into two categories, namely, color-blob-based [16,17] and line-art sketch [18,19]. The color-blob-based method first uses the predefined textures and colors as coarse features. Then, the shape and spectral information within a sketch is exploited [16,17]. Different from the sketch considered in the color-blob-based method, the line-art sketch only contains sparse strokes or lines, and their spatial arrangement is normally encoded as a feature signature for a given sketch query. Among various line-art sketch approaches, the most popular one is the Bag-of-Visual-Words (BoVW) framework, which has been proved to be more effective in SBIR [18]. Similar to the Bag-of-Words (BoW) model in text retrieval, BoVW describes an image as visual words and normally uses the *k*-means clustering to create a visual codebook, and then each image is mapped to a histogram of fixed length [19]. Eitz et al. created a large image dataset for SBIR and further proposed two new descriptors based on the BoVW framework for large-scale image retrieval, i.e., the Spark feature and SHOG (Sketched Feature Lines in the Histogram of Oriented Gradients) feature [19]. Cao et al. developed a real-time SBIR system by indexing 2 million images, named EdgelIndex. This method used Chamfer Distance to measure the distance between the sketch query and images, with a novel indexing method based on the inverted index to speed up the computation [9]. However, it didn't handle with position or scale variations well. Hu et al. introduced Gradient Field-HOG (GF-HOG), an enhanced version of the general HOG, to extract local features of sketches and then embedded it to the BoVW framework [20]. Such a combination has appeared effective in improving the whole SBIR performance. In recent years, more discriminative features and matching schemes have been proposed for SBIR. Bui et al. extended the color attributes to sketches, fused with the shape descriptor, and finally proposed an efficient inverse-index to speed up large-scale image retrieval [21]. Jin et al. proposed a novel visual region descriptor which actually combined the local HOG and spatial distribution of global interest points, and used the dynamic gridding strategy for matching [22]. Qian et al. enhanced the traditional SBIR by the re-ranking and relevance feedback, in which they applied a two-step visual feature verification to reduce the number of false positive results and optimized the final results [23]. Yu et al. introduced a specific database of shoes and chairs, and developed a deep triplet-ranking model for the instance-level SBIR [24]. However, these research works are all based on low-level or mid-level visual features, and cannot fully express sparse sketches and colorful images. Moreover, the ambiguities of freehand sketches and different drawing styles may weaken the representation capability of visual features.

Although more supervised methods are explored for image retrieval recently, most of them are proposed for CBIR and rarely for SBIR [12,13,28]. With the great strides of deep learning in the area of computer vision, especially with Convolutional Neural Networks (CNNs) leading the obvious progress in image classification [25], CNN features, as a global representation, are increasingly exploited in image retrieval and exhibit the beneficial boosting for SBIR [26]. Babenko et al. retrained a CNN model on several datasets which were similar to queries and extracted features for retrieval [12]. A very important factor for their work is that the features extracted from the retrained model retain the high-level semantic information from the original image. Lin et al. introduced a deep learning framework to learn binary hash codes for fast image retrieval, which was comparable to several state-of-the-art hashing algorithms [27]. Ng et al. explored the features in different layers of a deep network for image retrieval and found that the deeper layers lost the local features which were important for the instance-level image retrieval [28]. Liu et al. added text queries as the semantic information to remove the sketch ambiguity, and used SHOG as the visual feature to calculate the visual similarity and the co-occurrence probability of query-texts and images as the semantic similarity, and then obtained the final similarity by the linear fusion of these two kinds of similarity [29].

Unfortunately, all of these existing approaches have not yet provided good solutions for the following crucial issues, which are tightly coupled with each other.

(1) *Narrowing the Visual Appearance Difference between Sketches and Images* – To improve the resemblance of sparse sketches and colorful images, most existing methods attempt to transfer colorful images to the sketch-like forms through a simple edge detector, such as Canny Detector [30]. The shape of target object is composed of several important outlines in general perception of users, which means that only those significant contour lines are drawn when sketching an object. However, most edge detectors cannot preserve the significant contours while suppress the noisy ones, which heavily affect the correspondences between sketches and images. Meanwhile, freehand sketches are usually much sparser than those edge maps. Thus it is very important to handle such asymmetric attributes between sketches and images to achieve better retrieval performance. (2) *Mining Multiple Attribute Information for Deep Feature Representation* – Most existing methods are generally based on low-level visual features, which cannot fully express sparse sketches and colorful images and may ignore the possible semantic information involved in sketches and images. Such visual features are sensitive to noises, distortions and positions or scale variations of sketches and images to some extent. Furthermore, a lot of Internet images are annotated with tags, which can provide relatively rich semantic information for more discriminative representation and mitigate the vulnerability of visual feature information. However, to the best of our knowledge, no existing