Computers & Graphics 71 (2018) 77–87

Contents lists available at ScienceDirect

Computers & Graphics

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

Technical Section

# Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression

Tu Bui[a,*], Leonardo Ribeiro[b], Moacir Ponti[b], John Collomosse[a]

[a] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, United Kingdom
[b] Institute of Mathematical and Computer Sciences (ICMC), Universidade de São Paulo, São Carlos/SP, 13566-590, Brazil

ARTICLE INFO

*Article history:*
Received 20 September 2017
Revised 27 November 2017
Accepted 26 December 2017
Available online 4 January 2018

*Keywords:*
Sketch based image retrieval (SBIR)
Deep learning
Cross-domain modeling
Compact feature representations
Multi-stage regression
Contrastive and triplet losses

ABSTRACT

We propose and evaluate several deep network architectures for measuring the similarity between sketches and photographs, within the context of the sketch based image retrieval (SBIR) task. We study the ability of our networks to generalize across diverse object categories from limited training data, and explore in detail strategies for weight sharing, pre-processing, data augmentation and dimensionality reduction. In addition to a detailed comparative study of network configurations, we contribute by describing a hybrid multi-stage training network that exploits both contrastive and triplet networks to exceed state of the art performance on several SBIR benchmarks by a significant margin.

Datasets and models are available at http://www.cvssp.org.



## 1. Introduction

Sketches are an intuitive modality for communicating everyday concepts, and are finding increased application on modern touch-screen interfaces (e. g. on tablets, phones) where gestural interaction is natural. Such devices are now the platform on which the majority of today's visual content is consumed, motivating research into sketch as a medium for searching images and video.

This paper addresses the problem of sketch based image retrieval (SBIR); searching a collection of photographs (images) for a particular visual concept using a free-hand sketched query. We explore SBIR from the perspective of a cross-domain modeling problem, in which a low dimensional embedding is learned between the space of sketches and photographs. Traditionally, SBIR has been addressed using sparse feature extraction and dictionary learning, following the successful application of the same to recognition and search in natural images [1–3]. Deep convolutional neural networks (CNNs) have since gained traction as a powerful and flexible tool for machine perception problems [4], and recently have been explored for SBIR particularly within fine-grain retrieval tasks, e.g. to find a specific shoe within a dataset of shoes [5,6]. Despite early, promising results, it is unclear how suitable embeddings learned by these multi-branch networks are for generalizing across object categories [2,3]. For example, enabling a user to search for visual attributes within datasets containing diverse objects (e. g. a specific furniture form, a spotted dog, or particular building structure); a problem explored more extensively by prior work [2,3,7].

The technical contributions of this paper are two-fold. First, we present a comprehensive investigation of triplet embedding strategies evaluating these against popular SBIR benchmarks (Flickr15k [3], TU-Berlin [2]). In the spirit of recent 'details' papers studying deep networks for object recognition [8], we explore appropriate CNN architectures, weight sharing schemes and training methodologies to learn a low-dimensional embedding for the representation of both sketches and photographs—in practical terms, a space amenable to fast approximate nearest neighbor (ANN) search (e. g. $L^2$ norm) for SBIR. Second, we describe a novel triplet architecture and training methodology capable of generalizing across hundreds of object categories, and show this to outperform existing SBIR methods by a significant margin on leading benchmarks [2,3].

Concretely, we explore several important questions around effective learning of deep representations for SBIR:

1. **Generalization:** Given the diversity of visual concepts in the wild ($\sim 10^5$ categories) and the challenges of annotating large sketch datasets (current best $\sim 10^2$ categories [2]) how well can a network generalize beyond its training to unseen sketched object categories? Are class diversity and volume of exemplars equally important?
2. **Input modality:** SBIR and the related task of sketched image classification variously employ edge extraction as a pre-processing

* Corresponding author.
*E-mail address:* t.bui@surrey.ac.uk (T. Bui).

https://doi.org/10.1016/j.cag.2017.12.006
0097-8493/

step to align the statistics of sketch and photo distributions. Is this a beneficial strategy when learning a SBIR feature embedding?

3. **Architecture:** Recent exploration of SBIR has indicated triplet loss CNNs as a promising archetype for SBIR embedding, however what kind of loss objective should be considered and where, and which weight sharing strategies are most effective? What is the best way to enforce a low dimensional embedding for efficient SBIR indexing?

## 2. Related work and contributions

Sketch based image retrieval (SBIR) began to gain momentum in the early nineties with color-blob based query systems such as Flickner et al. 's QBIC [9] that matched coarse attributes of color, shape and texture using region adjacency graphs. Several global image descriptors for matching blob based queries were subsequently proposed, using spectral signatures derived from Haar Wavelets [10] and the Short-Time Fourier Transform [11]. This early wave of SBIR systems was complemented in the late nineties by algorithms accepting line-art sketches, more closely resembling the free-hand sketches casually generated by lay users in the act of sketching a throw-away query [12]. Such systems are characterized by their optimization based matching approach; fitting the sketch under a deformable model to measure the support for sketched structure within each photograph in the database [13,14]. Despite good accuracy, such approaches are slow and scale at best linearly. It was not until the 2010 decade that global image descriptors were derived from line-art sketches, enabling more scalable indexing solutions.

### 2.1. SBIR with shallow features

Mirroring the success of gradient domain features and dictionary learning methods in photo retrieval, both Eitz et al. [15] and Hu et al. [1] extended Bag of Visual Words (BoVW) to SBIR, also proposing the Flickr15k benchmark [3]. Sparse features including the Structure Tensor [16], SHoG [15], Gradient Field Histogram of Oriented Gradients (GF-HOG) [3] and its extended version [17] are extracted from images pre-processed via Canny edge detection. Chamfer Matching was employed in Mindfinder [18], later adopted by Sun et al. [19] for scalable SBIR indexing billions of images. Qi et al. [20] implemented an alternative edge detection pre-process delivering a performance gain in cluttered scenes. Mid-level features were explored through the HELO and key-shapes schemes of Saavedra and Barrios [7,21,22]. Their latest work [7] uses learned key-shapes and leads the shallow learning approaches.

### 2.2. SBIR with deep networks

SketchANet [23] was among the earliest deep networks for sketch, exploring recognition (rather than search) using a single-branch network resembling a short-form AlexNet [4]. SketchANet forms a component of the very recent work of Bhattacharjee et al. [24], coupled with a complex pipeline including object proposals, and query expansion. Although we also explored SketchANet, and compare with several other contemporary architectures which we show yield superior performance in a triplet framework (Section 4).

An early work exploring multi-branch networks for sketch retrieval (of 3D objects) was the contrastive loss network of Wang et al. [25] which independently learned branch weights to bridge the domains of sketch and 2D renderings of silhouette edges. In a recent short paper, Qi et al. [26] also propose a two-branch Siamese network with contrastive loss. Their results, although comparable with other methods using shallow features, are still far behind state-of-the-art [6,24] by a large margin. As we show later, learning a single function to map disparate domains to the search space appears to under-perform designs where branch weights are learned independently or semi-independently.

**Triplet CNNs** employ three branches [27]: (i) an anchor branch, which models the reference object, (ii) one branch representing positive examples (which should be similar to the anchor) and (iii) another modeling negative examples (which should differ from the anchor). The triplet loss function is responsible for guiding the training stage considering the relationship between the three models. Triplet CNNs have recently been explored for face identification [28], tracking [29], photographic visual search in [27,30] and for sketched queries in order to refine search within a single object class (e. g. fine-grain search within a dataset of shoes) [5]. Similarly, a fine-grained approach to SBIR was adopted by the recent Sketchy system of Sangkloy et al. [6] in which careful reproduction of stroke detail is invited for object instance search. In the former work [5], the authors train one model for each target category, and the embedding is learned using an edgemap extracted from a relatively clutter-free image. They report that using a fully-shared network was better than use two branches without weight sharing. However, the authors in [6] suggest it is more beneficial to avoid sharing any layers in a cross-category retrieval context. Recently, a hybrid design was explored by Bui et al. [31] using the same architecture on both branches but sharing certain layers. However, as their model learns mapping between sketch and edgemap (rather than image directly) its performance is limited. Furthermore, it is still unclear whether triplet loss works better than contrastive loss, with [6,31] supporting the former but [32] claiming the latter. Open questions remain around optimal training methodology, architecture, weight-sharing strategies, and loss functions, as well as the generalization capability of deep models for SBIR.

Our work explores these open questions, and broadens the investigation of deep learning to SBIR beyond intra-class or instance level search to retrieval across multiple object categories. To avoid confusion we hereafter refer as *no-share* or Heterogeneous those multi-branch networks for which there are no shared weights between layers [25]; as *full-share* or Siamese those for which all branches have shared weights in all layers [5,27]; and *partial-share* or Hybrid those for which only a subset of layers are shared. Our contributions for this paper are three-fold:

- A generic multi-stage training methodology for cross-domain learning that leverages multiple loss functions in training shared networks as illustrated in Fig. 1.
- An extensive evaluation of convnet architectures and weight sharing strategies.
- State-of-the-art performance on three standard SBIR benchmarks, outperforming other approaches by a significant margin.

## 3. Methodology

We propose a multi-stage training methodology and investigate several network designs, comparing the Siamese architecture with the Heterogeneous and Hybrid ones. Inspired from [31], we aimed to develop a training strategy for partial sharing networks. However, unlike [31] who employed a single training phase with a single loss function to concurrently train both shared and unshared parts of their sketch-edgemap network, we believe training a sketch-photo network should require more complex procedures. Additionally, we integrate the two most widely used regression functions in deep convnet, the contrastive loss and triplet loss, in our training procedure.

### 3.1. Network architecture

When learning a cross domain mapping between sketch and photo using deep convnet, at least two CNN branches are required