Signal Processing: Image Communication 67 (2018) 161–170

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Adaptive features selection for expert datasets: A cultural heritage application

Dorian Michaud [a,b,*], Thierry Urruty [a], Philippe Carré [a], François Lecellier [a]

[a] CNRS, Univ. Poitiers, XLIM, UMR 7252, F-86000, Poitiers, France
[b] Quadra Informatique, 68 Rue du Docteur Eloy, F-59133, Phalempin, France

ARTICLE INFO

Keywords:
Cultural heritage collection
Content based image retrieval
Visual saliency
Weighting scheme

ABSTRACT

Image Retrieval is still a very active field of image processing as the number of available image datasets continuously increases. One of the principal objectives of Content-Based Image Retrieval (CBIR) is to return to user the most similar images to a given query with respect to their visual content. Our work fits in a very specific application context: indexing small expert image dataset, e.g. cultural heritage images, with no prior knowledge on the images. Because of the image complexity, one of our contributions is the choice of effective descriptors from literature placed in direct competition. Two strategies are used to combine features: a psycho-visual one and a statistical one. In this context, we propose an automatic and adaptive framework based on the well-known bags of visual words and phrases models that select relevant visual descriptors for each keypoint to construct a more discriminative image representation. Experiment results show the adaptiveness and the performance of our framework on "generic" benchmark datasets and on two cultural heritage datasets.

1. Introduction

In the last decades, the success of smartphones and other mobile devices capable of taking and sharing photos instantaneously is closely linked to the exponential increase of image datasets. A huge number of those images are everyday life photos where the end user could be any one of us. But there are also expert images. In this paper, we focus on "expert image datasets", which are of interest for domain expert end-users. The expert end-users can be clinicians, historians, digital curators, numismatists, etc. Those expert datasets may have quite heterogeneous contents (e.g. historic images of persons, constructions, etc.) or more specific contents (e.g. datasets of old coins, butterflies, etc.). The particularities of such datasets have to be considered in the indexing tools that will help them to manage their data for further image exploitation stages as retrieval or browsing.

This paper focus on indexing expert image collections, and particularly cultural heritage image datasets which has become a topic of major interest for experts and researchers. Indeed, implementing digital and long-term preservation strategies, supporting open cultural heritage data are major topics for numerous countries.

Cultural heritage datasets contain very heterogeneous content as paintings, sculptures, etc. Picard et al. in [1] have proven that the classical approaches of CBIR do not provide satisfying results on this type of data. So, there is an urgent need to propose new methodologies to help expert users manage their data.

In this specific application context, we focus our research on image datasets with no prior knowledge. Thus, the scope of this paper is indexing image collection with Content-Based Image Retrieval (CBIR) methods. One objective of CBIR research is to create a discriminative visual signature describing the visual content of each image. To do so, the Bags of Visual Words model [2] (BoVW) has become popular. It aims at representing image visual features into a simple multidimensional vector. These vectors are the image signatures, i.e. histograms of the visual word occurrences but lacks discriminative power [3,4]. More recently, models based on human brain obtained very good results on computer vision tasks such as image classification or object recognition. These models are called Convolutional Neural Network (CNN) and outperform results obtained with "classical" schemes [5,6]. Other recent papers include semantic knowledge in addition to visual content. Thus, cross-modal retrieval using deep learning has become very popular [7,8]. These new methodologies have a tremendous appetite for learning which is impossible in our specific context of small expert datasets with no prior knowledge.

In this paper, we present our unsupervised framework, which aims at combining automatically the information from various local descriptors. Our main contribution consists in the selection of the most relevant

* Corresponding author.
E-mail address: dorian.michaud@univ-poitiers.fr (D. Michaud).

https://doi.org/10.1016/j.image.2018.06.011
Received 2 February 2018; Received in revised form 7 May 2018; Accepted 21 June 2018
Available online 23 June 2018

visual features by keypoint in a smart way. Indeed, we select some of the keypoints according to their importance to both the human visual system and the dataset itself. Thus, the chosen set of descriptors should provide relevant texture, shape, edge and colour information to obtain a more discriminative image representation. To strengthen the discriminative power of keypoints, we propose to introduce two strategies in our unsupervised approach. First, we introduce a psycho-visual model in two different steps of our image representation framework: (i) to discard irrelevant keypoints before starting the indexing step and (ii) to weight the importance of each keypoint during the signature construction step. This process gives more importance to salient keypoints and discredits the others. Secondly, a statistical approach is used to choose for each keypoint the combination of local descriptors providing the best information with respect to the dataset by using a particular weighting scheme. Weighting schemes and Information Gain models have been widely used in the Information Retrieval field [9]. They increase the importance of a term within a document (in our case, an image) by weighting each visual feature by a value that evolves with the number of occurrences within the dataset.

Another contribution of the proposed framework is its efficiency. Indeed, our framework reduces the number of used keypoints with respect to the visual saliency information. Furthermore, the obtained image signatures are sparser, which reduces the retrieval complexity.

To evaluate the performance of our contribution, we compare results obtained on well-known "generic" datasets. The BoVW model with the different descriptors, including deep features, and their concatenation will be the baseline approaches. We also compare our performance against other deep learning frameworks. Linked to our specific context and the motivation of this work, we present an evaluation of the proposed framework on two cultural heritage datasets: ROMANE 1K, which is a collection of Romanesque art images with heterogeneous content (paintings, sculptures, ...) and the Coin Collection Online Catalogue which is a numismatic dataset of Byzantine coins.

This article is structured as follows: first, we present the state of the art in Section 2. It describes the following processes and models: BoVW and selected improvements, CNN based literature, and a brief overview of weighting schemes in image retrieval and visual saliency methodologies. Then, Section 3 gives an overview of our proposal. Section 4 presents the experiments on "generic" image datasets and on two different cultural heritage image collections, and discusses the findings of our study. Finally, Section 5 concludes and gives some perspectives.

2. Related works

In this section, we first present the BoVW model, few inspired improvements and Bags of Visual Phrases model approaches. After introducing CNN models and transfer learning approaches, we expose a brief study of literature approaches concerning visual saliency and weighting schemes in the context of CBIR.

2.1. Classical retrieval frameworks

Inspired by the Information Retrieval field, Csurka et al. [2] introduced the classical BoVW model. The main idea of this methodology is to cluster descriptors of image patches and use this clustering to obtain a high-dimensional visual vocabulary. This vocabulary is used to construct image signatures. Then, during the retrieval process, a distance metric is used to measure similarity between the signatures. The first step is the detection and the description of image patches. Several descriptors can be used as SIFT, DAISY, HOG, ... [10,11]. An algorithm is then used to assign descriptors to a set of predetermined clusters, i.e. the visual vocabulary. This assignation allows the construction of bags of keypoints: by counting number of patches assigned to each cluster, a histogram of the visual words occurrences is constructed. The size of the vocabulary should be carefully chosen: it should be large enough to distinguish relevant changes, but it should be small enough according to the learning dataset size to distinguish irrelevant variations such as noise.

More recently, Jégou et al. [12] have proposed the VLAD representation. This approach can be seen as an accumulation of distances between keypoints descriptors and the different cluster centres. Delhumeau et al. in [13] proposed an extension of this method by using PCA for every part of VLAD vector. A hierarchical VLAD was also proposed by Eggert et al. in [14].

Other methods focus on efficiency like Bags of Visual Phrases (BoVP) model which is an extension of BoVW model. The BoVP model groups the keypoints to better represent small regions. This method preserves the geometry of objects inside images. Many ways exist to construct phrases: using sliding window [15], grouping the keypoints with their nearest neighbours [3], or by regions [4]. Finally, with the BoVP model, the image is represented by a histogram of visual phrases that is proved to be more discriminative than the BoVW model however computationally more expensive.

In this paper, we present our framework inspired by these methodologies with specific procedure of features selection and weighting by keypoint.

2.2. Convolutional neural network

In recent years, deep convolutional neural networks provide great performances in a lot of computer vision and image processing tasks [5,6]. This innovative approach is based on the work of Yann LeCun et al. [16] which proposed the first modern CNN architecture. Alex Krizhevsky et al. in [17] presented a deep CNN called AlexNet. This model achieved a top-5 error of 15.4% in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012. It is trained on ImageNet [18] data, which contained over 15 million annotated images from over 22,000 categories. The network consists of convolutional layers, max-pooling layers, dropout layers (5 for each of them), and 3 fully connected layers and it was used for classification with 1000 possible categories. Data augmentation and dropout used in AlexNet are part of numerous recent literature frameworks mentioned below. Karen Simonyan and Andrew Zisserman proposed two years later in [19] the VGG-Net model for ILSVRC 2014. Their network obtained a top-5 error of 7.3%. The main contribution of this network is the reduction of parameters by using smaller size of convolutional filters than the other well-known networks. Another model, GoogLeNet, proposed by Szegedy et al. in [20] stands out from the usual CNN. Indeed, GoogLeNet, was one of the first CNN model that does not use the classical sequential structure. The authors proposed a new module, which does not stack convolutive and pooling layers on top of each other but proceeds these operations in parallel. This module is called Inception. GoogLeNet achieved a top-5 error of 6.7% in ILSVRC 2015. Microsoft Research Asia proposed during the same challenge the ResNet model which consists of a very deep CNN of 152 layers in [21]. This model outperformed the other well-known models by achieving a top-5 error of 3.6%.

Faced with the excellent performances offered by these models, researchers approach this field in different ways. Instead of recreating new models of CNN, several authors have proposed to adapt them to specific problems [22–24]. In [24], Pittaras et al. proposed to compare three different fine-tuning strategies in order to investigate the best way to transfer the parameters of popular deep CNNs. Transfer learning approach allows to adapt a network trained for one task to another more specific one. Gando et al. [22] fine-tuned a deep CNN for distinguishing illustrations from photographs. Another example is the work of Jung and Hong in [23]. They designed a deep network for pedestrian detection. In the field of image retrieval, recent papers like SPoC [25] and Neural Codes [26] proposed by Babenko Yandex et al. obtained pretty good results on well-known datasets.

The drawback of these methods is the amount of training data needed to adapt and tune the numerous parameters. In our context with no