Information Sciences 281 (2014) 536–558

Contents lists available at ScienceDirect

Information Sciences

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

A data-driven study of image feature extraction and fusion

Zhiyu Wang [a,b], Peng Cui [a,b,*], Fangtao Li [c], Edward Chang [c], Shiqiang Yang [a,b]

[a] Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science, Tsinghua University, Beijing, China
[b] Beijing Key Laboratory of Networked Multimedia, Tsinghua University, Beijing, China
[c] Google Beijing, China

ARTICLE INFO

Article history:
Received 17 January 2013
Received in revised form 24 January 2014
Accepted 9 February 2014
Available online 20 March 2014

Keywords:
Image annotation
Image feature
Web-scale
Fusion

ABSTRACT

Feature analysis is the extraction and comparison of signals from multimedia data, which can subsequently be semantically analyzed. Feature analysis is the foundation of many multimedia computing tasks such as object recognition, image annotation, and multimedia information retrieval. In recent decades, considerable work has been devoted to the research of feature analysis. In this work, we use large-scale datasets to conduct a comparative study of four state-of-the-art, representative feature extraction algorithms: color-texture codebook (CT), SIFT codebook, HMAX, and convolutional networks (ConvNet). Our comparative evaluation demonstrates that different feature extraction algorithms enjoy their own advantages, and excel in different image categories. We provide key observations to explain where these algorithms excel and why. Based on these observations, we recommend feature extraction principles and identify several pitfalls for researchers and practitioners to avoid. Furthermore, we determine that in a large training dataset with more than 10,000 instances per image category, the four evaluated algorithms can converge to the same high level of category-prediction accuracy. This result supports the effectiveness of the data-driven approach. Finally, based on learned clues from each algorithm's confusion matrix, we devise a fusion algorithm to harvest synergies between these four algorithms and further improve class-prediction accuracy.

1. Introduction

Extracting useful features from a scene is an essential subroutine in many multimedia data analysis tasks such as classification and retrieval. Remarkable progress has been made in multimedia computing, computer vision and signal processing in recent decades. Despite this finding, it is still notably difficult for computers to accurately recognize an object or analyze the semantics of a scene. For example, suppose that we want to recognize a piece of white paper in an image. A naive feature we can use is "a white two dimensional rectangle". However, such a feature will not work in most cases because of the following:

1. The paper may be folded.
2. The viewing angle of the piece of paper may not perpendicular, and hence the paper does not appear to be rectangular.
3. Environmental factors such as occlusion and lighting can cause changes in its shape and color.

* Corresponding author at: Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science, Tsinghua University, Beijing, China. Tel.: +86 10 62790810.
E-mail address: cuip@mail.tsinghua.edu.cn (P. Cui).

http://dx.doi.org/10.1016/j.ins.2014.02.030
0020-0255/

The above challenges are all related to *feature invariance* issues. A second challenge is called *feature aliasing* or *feature selectivity*: how well a feature can differentiate one object from the others. For example, the feature "white two-dimensional rectangle" can be used to describe many other objects: a piece of white cloth, a white table, and a white wall, among others. The goal of feature extraction is to find features that are both *invariant* and *selective*.

All traditional feature extraction approaches focus on some specific information in the image. For example, the color-texture codebook (CT) focuses on the statistics of colors and textures in small regions of an image. SIFT focuses on local invariant shapes. Recently, neuro-based approaches such as HMAX and convolution networks (ConvNet) have been proposed to model features according to how the human visual system extracts features. HMAX [45] builds computing models that use the pioneering neuroscience work of Hubel [22]. Hubel's work indicates that visual information is transmitted from the primary visual cortex (V1) through extrastriate visual areas (V2 and V4) to the inferotemporal cortex (IT). The IT, in turn, is a major source of input to the prefrontal cortex (PFC), which is involved in linking perception to memory and action [33]. The pathway from the V1 to the IT (called the *visual frontend*) consists of a number of simple (lower) and complex (higher) layers. The lower layers attain simple features that are invariant to scale, position and orientation at the pixel level. Higher layers can combine simple features to recognize more complex features at the object-part level. Pattern recognition at the lower layers is unsupervised, whereas recognition at the higher layers involves supervised learning. This particular neuroscience-motivated model appears to enjoy at least a couple of advantages: (1) it balances feature selectivity (at lower layers) and invariance (at higher layers) and (2) it models edges of an object and then combines edges to recognize parts of an object and place these features in a hierarchical context. Similar to HMAX, ConvNet is also a neuro-based approach. It differs from HMAX primarily in the way that ConvNet iterates more over the data to learn a model with a deep architecture [39]. This allows for the capture of both the structure and detail of an object.

Herein, we perform comparative evaluation that demonstrates that different feature extraction algorithms have their own set of advantages and excel in different image categories. We provide key observations about why certain algorithms perform better with different image categories. Based on these observations, we establish feature extraction principles and identify several pitfalls for researchers and practitioners to avoid:

1. When training data are insufficient, no scheme performs well. However, because simple algorithms such as CT and SIFT do not require much data to learn model parameters, they may be a better choice when training data are scarce.
2. Increases in the amount of training data correlate with a jump in the accuracy of complex models, such as HMAX and ConvNet. Different feature extraction algorithms enjoy their own advantages, and excel in different image categories.
3. When training data are abundant, all the four algorithms, simple or complex, converge to the same level of accuracy.

The major contributions of this paper are summarized as follows:

1. Through our comparative analysis, we identify pitfalls of past studies: either they did not use enough training data, or their testbed composition already favors a particular feature extraction algorithm.
2. Through our large-scale comparative study, we demonstrate the benefit of employing large training datasets in training, which can make both simple and complex algorithms converge to the same level of accuracy.
3. We devise a fusion algorithm based on learned clues from each algorithm's confusion matrix. Our algorithm harvests synergies between these four algorithms and further improves class-prediction accuracy.
4. We established a large testbed for the research community, namely an annotated dataset of six million PicasaWeb images, which will be released publicly with this paper.[1]

The rest of the paper is organized as follows. Section 2 surveys the related work. Section 3 briefly introduces the four feature extraction algorithms evaluated in this paper. Section 4 details an algorithm that fuses multiple feature extraction methods that we demonstrate can perform better than any individual feature extraction scheme alone. Section 5 explains the setup of our experiments and presents their results. Finally, we offer concluding remarks in Section 6.

2. Related work

The multimedia community has been striving to bridge the *semantic gap* [20,46,62] between low-level features and high-level semantics for decades. (Comprehensive surveys are given in [5,20].) With high-quality image features, fancy applications can significantly improve a user's experience [9,11,30]. One key problem is how to extract powerful features. Numerous feature extraction algorithms have been developed for image annotation [46], as well as machine learning algorithms [14,36,59–61]. Roughly speaking, image features can be grouped into four types: color, texture, local features, and shapes. Color features are the most straightforward image feature and therefore were the first to receive sufficient study. Many color-based image retrieval algorithms have been developed [18]. Typical color features include the color histogram [19], color invariance [17], and color saliency [51]. Texture features, such as local binary patterns (LBP) [34], pyramidal-structured wavelet transforms [37], and tree-structured wavelet transform [8], are another significant set of signals for recognizing ob-

[1] Downloads at https://sites.google.com/site/picasawebdataset/home.