Semantic Image Retrieval

C.H.C. Leung
Department of Computer Science, Hong Kong Baptist University, Hong Kong

Yuanxi Li
Hong Kong Baptist University, Hong Kong

INTRODUCTION

The number of Web images is increasing at a rapid rate, and searching them semantically presents a significant challenge. Many raw images are constantly uploaded with little meaningful direct annotation of semantic content, limiting their capacity to be searched and discovered. Unlike in a traditional database, information in an image database is in visual form, which requires more space for storage, is highly unstructured and needs state-of-the-art algorithms to determine its semantic content.

As Web images tend to grow to unwieldy proportions, their retrieval systems must be able to handle multimedia annotation and retrieval on a Web scale with high efficiency and accuracy. With the exception of systems that can identify or detect music, words, faces, irises, smiles, people, pedestrians, or cars, matching is not usually directed toward object semantics. Recent research studies show a large disparity between user needs and technological supply.

BACKGROUND

Vast numbers of Web images are continuously added with few meaningful direct annotations of semantic content, limiting their search and discovery. While some Websites encourage tags or keywords to be included manually, such is far from universal and applies to only a small proportion of images on the Web. Research in image search has reflected the dichotomy inherent in the semantic gap (Hanjalic, Lienhart, Ma, & Smith, 2008), and is divided between two main categories: concept-based image retrieval and content-based image retrieval. The former focuses on retrieval by image objects and high-level concepts, while the latter focuses on the low-level visual features of the image. In order to determine image objects, the image often has to be segmented into parts. Common approaches to image segmentation include segmentation by region and segmentation by image objects. Segmentation by region aims to separate image parts into different regions sharing common properties. These methods compute a general similarity between images based on image properties, (Jing & Baluja, 2008) and common examples of such properties are texture and color where these methods are found to be robust and efficient. Some systems use color, texture, and shape (Vogel, Schwaninger, Wallraven, & Bulthoff, 2007) as attributes and apply them for entire image characterization, and some studies include users in a search loop with a relevance feedback mechanism to adapt the search parameters based on user feedback, while various relevance feedback models and ranking methods for Web search have been developed (Cheng, Jing, Li, Ma, & Jin, 2006). Recent approaches for calculating image similarity using various image features, such as color similarity measure (Chary, Lakshmi, & Sunitha, 2012) can ensure a certain degree of precision for image retrieval within a large image collection. The extracted features could be filled into local (like SIFT, SURF or DAISY) or global (like MPEG-7 color/texture descriptors, Edge Orientation Histogram or GIST) descriptors. Some studies (Wang, Zhao, Yu, & Z. Guang, 2011) uses the Non-Subsampled Contourlet Transform (NSCT) detector combined with a DAISY descriptor to develop a robust interest point matching algorithm for infrared/visible images. As regards to the segmentation of color images, segmentation by object, on the other hand, is widely regarded as a hard problem, which if successful, will be able to replicate and perform the object recognition function of the human vision system; although progress on this front has been slow, some advances in this direction have neverthe-

DOI: 10.4018/978-1-4666-5888-2.ch593

less been made (Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007). Makrogiannis et al. (Makrogiannis, Economou, Fotopoulos, & Bourbakis, 2005) proposed a multiresolution image segmentation scheme based on a graph-theoretic approach. Grady and Schwartz (Grady & Schwartz) treated image segmentation as a linear problem instead of the eigenvector approach to a graph-partitioning problem (Shi & Malik). Wenbing et al. (Wenbing, Hai, & Yimin) developed a robust real-time approach for color image segmentation using the MS segmentation and the normalized cut (Ncut) (Shi & Malik) partitioning methods. Diplaros et al. (Diplaros, Vlassis, & Gevers, 2007) resorted to a spatially constrained EM algorithm to estimate the model parameters. In (Li & Wang, 2008), semantic annotation of images combined with a region-based image decomposition is used, which aims to extract semantic properties of images based on the spatial distribution of color and texture properties. Such techniques have drawbacks, primarily due to their weak disambiguation and limited robustness in relation to object characterization. However, an advantage of using low-level features is that, unlike high-level concepts, they do not incur any indexing cost as they can be extracted by automatic algorithms. In contrast, direct extraction of high-level semantic content automatically is beyond the capability of current technology. Although there has been some effort in trying to relate low-level features and regions to higher level perception, these tend to be for isolated words, and they also require substantial training samples and statistical considerations (Duygulu, Bastan, & Forsyth, 2006). These methods, however, have limited success in determining semantic contents in broad image domains. There are some approaches which exploit surrounding and associated texts in order to correlate and mine these with the content of accompanying images (Li, Chen, Zhang, Lin, & Ma, 2006). Text-based retrieval is often limited to the processing of tags, and no attempt is made to extract a thematic description of the picture. Some research focuses on implicit image annotations which involves an implicit, rather than an explicit, indexing scheme and, in consequence, augments the original indexes with additional concepts that are related to the query (Natsev, Haubold, Tesic, Xie, & Yan, 2007), necessitating the use of some probabilistic weighting schemes.

MAIN FOCUS OF THE ARTICLE

Issues, Controversies, Problems

Some of the most important notions in image retrieval are keywords, terms or concepts. Terms are used both from humans to describe their information need and from the system as a way to represent images. However, current image search systems, such as Yahoo! and Google, use some surrounding text description provided by humans in order to infer semantics. These techniques ignore the meaningful image features which can be extracted via image processing analysis. Further, as most of these images come without explicit semantic tags and those inferred from surrounding text are often unreliable, these models have limited effectiveness at present and they need further development and refinement. There lies the need to automatically infer semantics from raw images to facilitate semantics-based searches.

The effectiveness of image retrieval depends on meaningful indexing; the key problem of image retrieval is to organize them based on semantics. The word ‘semantic’, which frequently appears in the content of this dissertation, is the linguistic interpretation of multimedia objects, such as images and video clips.

From the viewpoint of image annotation and retrieval models, semantics can be textual descriptions attached to images (Li, Shou, Chen, Hu, & Dong, 2008). Moreover, it can be high-level concepts describing the scene or relationships between images in a group that have a particular meaning for a user (Urban & Jose, 2007). The process of assigning semantics to images is referred as image annotation, indexing or tagging in general for multimedia objects. The approach of manual annotation is followed by a number of commercial platforms and collaborative communities of multimedia, such as YouTube, Flickr and Pbase. However, the manual annotation of images is a laborious and error-prone task, as annotations can be biased towards the annotator’s perspective, and it is difficult to define a strategy to ensure annotative consistency. This implicit connection is usually called as the ‘semantic gap’. The concept of the ‘semantic gap’ is proposed and formalized by Smeulders et al. (Smeulders, Worring, Santini, Gupta, & Jain, 2000):