High-Level Features for Image Indexing and Retrieval

Gianluigi Ciocca
Università degli Studi di Milano-Bicocca, Italy

Raimondo Schettini
Università degli Studi di Milano-Bicocca, Italy

Claudio Cusano
Università degli Studi di Pavia, Italy

Simone Santini
Universidad Autónoma de Madrid, Spain

INTRODUCTION

Image indexing for content-based image retrieval is the process of automatically computing a compact representation (numerical or alphanumerical) of some attribute of digital images, to be used to derive information about the image contents. A feature, or attribute, can be related to a visual characteristic, but it may also be related to an interpretative response to an image or to a spatial, symbolic, semantic, or emotional characteristic. A feature may relate to a single attribute or be a composite representation of different attributes. Features can be classified as general purpose or domain-dependent. The general purpose features can be used in any context, while the domain-dependent features are designed specifically for a given application. Every feature is intimately tied with the kind of information that it captures. The choice of a particular feature over another depends on the given application, and the kind (level) of information required.

While more complex and sophisticated general purpose and domain dependent features were being developed, two important issues became evident: the sensory gap and the semantic gap. The sensory gap is the gap between the information of the real world, and the information in a computational description derived from a digital recording of a scene of the world (i.e. low-level features or visual features such as color, texture, etc.). The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given application. As stated by Smeulders et al. (2000), the gap between the pictorial features and the image's semantics makes it difficult for purely low-level content-based retrieval systems to obtain satisfactory results. High-level features try to bridge the semantic gap by embedding in their representation information about the image content. This goal can be pursued by different means from manually annotating the images with texts to exploiting pattern recognition, computer vision and machine learning algorithms.

BACKGROUND

The literature on content based retrieval has become so vast that any attempt at exhaustivity in a article like this would be daunting. Instead, in this section, we will briefly describe those works in the state of the art that deal with the problem of extracting high-level features from the images by following different technical development and general philosophy. Narrowing down the semantic gap can be achieved through the exploitation of different image indexing techniques.

Text/Ontology

The fist technique is text-based meaning that the image content is described in terms of textual keywords. This

DOI: 10.4018/978-1-4666-5888-2.ch585

description can be manually provided to the system or obtained using a vocabulary, an “object-ontology” which provides a qualitative definition of high-level concepts from the low-level features. For example, an image region can be assigned the keyword “sky” if it is an “upper, uniform and blue-collared region.” The “blue” attribute can be itself obtained from low-level features concerning color distribution (e.g. an average RGB colour components between predefined thresholds that characterize a generic sky region). Examples of this kind of descriptors, not limited to colour but applied also to other low-level features, can be found in Ravishankar et al. (1993). Since an ontology is naturally hierarchical, that is, it not only contains a set of concepts but also relations between these concept, an object can receive multiple keywords based on its nature. Moreover, the keywords vocabulary can refers non only to objects within the image but also to the image itself as a whole. In this case, the attributes or concepts of interest are global ones (e.g. the image category such as “sunset,” “landscape,” “indoor,” “close-up,” etc.). An example of how an image can be hierarchically and globally described is ImageNet (Deng et al. 2009), an image database organized according to the WordNet (Miller et al. 1990) hierarchy, in which each node of the hierarchy contains hundreds and thousands of related images. Using a vocabulary of keywords, a query to retrieve similar images, can be performed by using traditional, text-based, information retrieval techniques.

WWW-Based

Other techniques are those based on the capabilities of the Web. For example, Web image retrieval exploit the additional information associated to the Web pages to facilitate semantic -based image retrieval. Examples are the information related to the URLs, ALT-tags or the HTML documents as (for example) in Cai et al. (2004). The rationale of these approaches is that the information “surrounding” an image is somewhat related to it and thus contains keywords meaningful for a semantic retrieval of the images. Usually these approaches are used in the most advanced Web search engines. Interesting works have been done to analyzes and cluster the results returned by conventional Web image search engines in order to return to the user more meaningful results. For example in Cai et al. (2004b) results are grouped into different semantic categories and several images from each category are selected as representative images. Not only textual cues are used but also visual features are extracted from the images. In Wang et al (2008) a framework that tries to learn a new distance measure in the visual space, which can be used to retrieve more semantically relevant images for any unseen query image is proposed. The framework employs a ranking-based distance metric learning method applied to millions of images together with rich textual information to learn the similarity measure. A further possibility, made realistic by the widespread success of ``social networks'' on the Internet, is to use the classification made by other people as semantic prototypes. For example in the work by Cusano et al. (2012), unlabeled pictures are tagged with proposed labels. To formulate the suggestions is exploited the knowledge encoded in how other users have already partitioned their images. This is achieved by using suitable feature representation of the images to model the different classes that the users have collected and by looking for correspondences between the criteria used by the different users. Boosting is used to integrate the information provided by the analysis of multiple users.

Relevance Feedback

To overcome the necessity of manually describing the images content, many of these systems are based on image features derived from computer vision, which can be computed directly and automatically from the images themselves. However, researchers soon realized that simple content-based features could not characterized even the ophtosemantics of images to the degree of generality and sophistication that was required for general-purpose retrieval. In order to come to grips with this problem and to provide satisfactory retrieval performance, hermeneutic solutions were introduced in the retrieval process that take into account the subjectivity of human perception. One of these solutions is *relevance feedback* which relies on the interaction with the user to provide the system with examples of images relevant to the query. The system then refines its result depending on the selected images.

User’s feedback can be exploited in three ways: query point movement, query re-weighting, and class learning. Query point movement improves the estimation of the query by moving it towards the set of the (expected) relevant results. Query re-weighting dynamically updates the weight associated to the query,