Scenique: A Multimodal Image Retrieval Interface[*]

Ilaria Bartolini, Paolo Ciaccia
DEIS
University of Bologna, Italy
{i.bartolini, paolo.ciaccia}@unibo.it

ABSTRACT

Searching for images by using low-level visual features, such as color and texture, is known to be a powerful, yet imprecise, retrieval paradigm. The same is true if search relies only on keywords (or *tags*), either derived from the image context or user-provided annotations. In this demo we present Scenique, a multimodal image retrieval system that provides the user with two basic facilities: 1) an *image annotator*, that is able to predict keywords for new (i.e., unlabelled) images, and 2) an integrated query facility that allows the user to search for images using *both* visual features and tags, possibly organized in semantic *dimensions*. We demonstrate the accuracy of image annotation and the improved precision that Scenique obtains with respect to querying with either only features or keywords.

Keywords

Multi-structural Databases, Semantic Dimensions, Visual Features.

1. INTRODUCTION

The advent of digital photography has enormously increased the demand of tools for effectively managing huge amounts of color images. Among such tools, those providing similarity-search functionalities are essential if one wants to provide users with the possibility of looking for images whose visual content is similar to a given, so-called *query*, image. Even if this content-based approach can be completely automatized, it is known to yield imprecise results because of the semantic gap existing between the user subjective notion of similarity and the one implemented by the system.

The alternative to content-based retrieval is to look for images by using text-based techniques. Towards this end, several solutions have been proposed in recent years, such as the image search extensions of Google[1] and Yahoo[2], which consider the original Web context (e.g., file name, title, surrounding text) to infer the relevance of an image, as well as systems like flickr[3], which rely on user-provided *tags*. However, in both cases, the accuracy of the results is highly variable, since it heavily depends on the precision and the completeness of the manual annotation process (in the case of flickr) and it is completely uncorrelated with respect to the visual image content (in the case of Google and Yahoo).

In this demonstration we present Scenique (Semantic and ContENt-based Image QUErying), a multimodal image retrieval system whose major aim is to provide users with an *integrated* query facility that allows images to be searched by means of *both* visual features and semantic tags, thus taking the best of the two approaches. The model of Scenique is based on the multi-structural framework proposed in [2]. In particular, each image is viewed as a set of regions, from which color and texture can be automatically extracted, and a set of tags. Tags can be organized in so-called (classification) *dimensions*, which take the form of *tag trees*. Each dimension, such as `location`, is thought to be as a particular coordinate to describe the content of an image. When dimensions are defined by the user, Scenique predicts tags for each specified dimension.

Searching for images in Scenique can take three basic forms, as better explained in the following: content-based only, tag-based only, and integrated. In the demo we will show how the quality of retrieval depends on the chosen query modality.

2. ARCHITECTURE AND PRINCIPLES

The Scenique architecture is mainly composed by a *Feature DB* storing color and texture feature vectors that are automatically extracted from images, and by a *Tag DB* which stores the current tags defined for each image. A tag occurrence is actually a specific node in a *tag tree*, each tag tree representing the organization of tags for a specific *dimension*. As an example, the tag `animal` could be a node in the tag tree of the `subject` dimension. Note that, in principle, the same tag can appear in different tag trees, which allows to discriminate between the different usages and/or meanings different tag occurrences can have. For instance, the tag `Italy` might appear as a node for the `location` dimension (used to organize photos according to the place they have been shot) as well as a node in the `sport` dimension (which only applies to photos related to sport events).

By default, each tag is initially a node of a generic, unstructured, `default` dimension. User-defined dimensions

[*] This work is partially supported by a Telecom Italia grant.

[1] Google image: http://images.google.com/

[2] Yahoo image: http://images.search.yahoo.com/

[3] flickr: http://www.flickr.com/

AVI '08 Napoli, Italy
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$5.00.

can be defined to fit specific needs. For instance, in order to organize photos according to their main `subject`, a corresponding dimension can be defined and structured by creating the nodes `person` and `animal`; then the node `animal` can be split into the three nodes `mammal`, `bird`, and `fish`. The node `mammal` can be further specialized into nodes `bear`, `horse`, etc.

Scenique is based on the multi-structural framework [2], that consists of a set of objects, together with a *schema* that specifies a classification of the objects according to multiple distinct criteria (i.e., the dimensions). In such a way, the user can define several dimensions, with the aim to organize images from different points of view, and, at query time, browse images through the tag trees as well as formulate composed tag-based queries. This is exemplified in Figure 1, where the dimensions `subject` and `place` are conjunctly used to look for "sea animal" images. Within the reference model, a set of operations, such as the *meet* (or logical `AND`) and the *join* (or logical `OR`) are defined. In this way, the user formulates compound queries by means of logical expressions (e.g., `(sea AND animal)`).

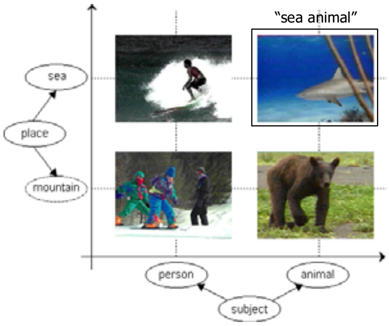

Figure 1: A compound search based on the `place` and `subject` dimensions.

Queries submitted by the user are managed by the *Query Processor*, which supports three main query modalities: *content*-based (C), *tag*-based (T), and *content&tag*-based (CT), respectively. With modality C, the user is looking for images that are similar, from a visual features point of view, to a specific query image. In particular, the Query Processor provides support for k nearest neighbor (k-NN) queries: Given a query image, it ranks images according to a specific similarity criterion and returns the k images with highest similarity score. Queries of type T are formulated using the available dimensions. In the simplest case, the retrieval is based on the resolution of user-provided logical expressions, that relies upon the exact match between selected tags and image associated tags. More interesting queries are derived when the *parent-child* relations between nodes of the tag trees (e.g., "the bear is a mammal") are exploiting by the Query Processor to improve the quality of the results. By supposing that the user is looking for `bear` images, the result provided by the Query Processor might include not only images with the tag `bear`, but also images annotated with the tag `asiatic_brown_bear`, because, in this case, the Query Processor takes the advantage of the relation "the asiatic_brown_bear is a bear". With the same aim, lexical ontologies, such as WordNet[4], can be exploited instead of user-defined dimensions. This allows to deal with the case when provided keywords do not belong to any dimension. Finally, with *content&tag*-based queries, the Query Processor combines C and T modalities by returning images in the intersection of both the C and T results first, followed by images in the T list only and, finally, by images in the C result only.

The user can also take advantage of the *Annotator* component of Scenique to obtains tags for unlabelled images, for each specified dimension, so as to properly characterize their semantic content. Here we summarize the main idea of the image annotation process (for a complete description, please refer to [1]). Annotation is modelled as a nearest neighbor problem on image regions. The set R containing the k-NN regions of each region of a new image is first determined. The initial set T of tags for the new image equals the tags included in images containing regions in R; each tag in T is also given a frequency score f. However, tags in T might include unrelated, or even contradictory, terms. To overcome such limit, we exploit the pairwise *term correlation* by associating to each couple of tags a correlation score c. In particular, we reduce the cardinality of T by combining the scores f and c. To this end, we build an undirected and weighted graph G whose nodes correspond to tags in T with the highest values of f, whereas the weights are the f values. An edge between two nodes is added if their correlation score c exceeds a fixed threshold value. Starting from the graph G, we derive the set of final tags that are both *affine* to the new image and that share a *semantic correlation* among themselves by determine the maximum subset of fully connected nodes.

3. DEMONSTRATION

Let us illustrate a usage example of Scenique. In our system each image is automatically segmented into a set of homogeneous regions which convey information about color and texture features. Each region corresponds to a cluster of pixels and is represented through a 37-dimensional feature vector. With respect to regions comparison the Bhattacharyya metric is used. In the demo we will show results obtained on an image database of annotated images extracted from the Corel image collection.

First of all, the user builds several dimensions by means of the graphical tag tree editing functionalities offered by the GUI. Then she formulates the *content&tag* query "`(sea AND animal)`" by also supplying to the system her favorite *fish* image. Scenique returns images according to the integration rule above described. Depending on her preferences, the user can refine the tags associated to the returned images or assign new ones to images coming from the content-based retrieval only. Finally, for a new photo, she is interested in annotating it. Among the terms predicted by the system, each one associated to the proper dimension, the user can refine them by deleting wrong tags and/or by adding missing terms, depending on the precision of the provided result.

4. REFERENCES

[1] I. Bartolini and P. Ciaccia. Imagination: Accurate Image Annotation Using Link-analysis Techniques. In *AMR 2007*, Paris, France, July 2007.

[2] R. Fagin, R. V. Guha, R. Kumar, J. Novak, D. Sivakumar, and A. Tomkins. Multi-structural Databases. In *PODS 2005*, Baltimore, USA, June 2005.

[4] WordNet: http://wordnet.princeton.edu/