Integrating Semantic and Visual Facets for Browsing Digital Photo Collections[⋆]

Ilaria Bartolini and Paolo Ciaccia

DEIS, Università di Bologna - Italy
{i.bartolini,paolo.ciaccia}@unibo.it

Abstract. Managing photos by using low-level visual features is a powerful, yet imprecise, organization paradigm. The same is true if only keywords (or *tags*) are used. In this paper we present a new browsing and search system, named Scenique, that allows the user to manage her photo collections by using *both* visual features and tags, all homogeneously organized into a set of (visual and semantic, respectively) hierarchical facets. We present the basic principles of Scenique, describe the building blocks of its software architecture, and provide evidence of its effectiveness, as evaluated by a set of real users.

1 Introduction

Thanks to the wide dissemination of digital and phone cameras, it is nowadays extremely easy for any ordinary user to collect photos. Since taking and storing pictures is almost priceless, the size of personal digital photo collections is consequently growing at an unprecedented rate, which demands for management tools with advanced functionalities. Among them, effective browsing and searching instruments are essential in order to avoid getting lost within a large photo repository. To this end, current solutions provide a variety of heterogeneous techniques, ranging from content-based search (which relies on low-level similarity features, such as color and texture) to keyword-based retrieval (exploiting *tags* used to annotate images), and from topic-based hierarchies to advanced graphical interfaces based on novel metaphors [1] so as to help the user in navigating through large image collections.

It is a fact that none of these techniques alone is able to reach satisfactory performance levels. For instance, even if the 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 [2]. Similarly, text-based techniques, as exemplified by the image search extensions of Google and Yahoo!, and by systems like Microsoft's Photo Gallery, Google Picasa, and Yahoo's Flickr, yield a highly variable retrieval accuracy. This is due to the imprecision and the incompleteness of the manual annotation process (in the case of Photo Gallery, Picasa, and

[⋆] This work is partially supported by a Telecom Italia grant and the CoOPERARE MIUR Project.

Flickr), or to the poor correlation that often exists between surrounding text of Web pages and the visual image content (for the case of Google and Yahoo!).

Similar problems plague current browsing systems. In this case, a hierarchical organization is commonly adopted to support browsing tasks on top of large image datasets, with most of the existing systems using a single hierarchy. The inadequacy of this approach has been demonstrated by systems like Flamenco [3], where *multi-faceted hierarchies* allow users to explore a data collection across multiple, orthogonal classification schemas.

In this paper we present Scenique (Semantic and ContENt-based Image QUErying), a novel multi-faceted image search and browsing system.[1] Scenique represents an effective step towards providing users with an *integrated* environment that allows photos to be searched and explored using *both* tags and visual features, thus taking the best of the two approaches. The unique features of Scenique can be summarized as follows:

- Automatically-extracted low-level features can be used both to search for and to browse through images. For the latter, Scenique allows for the definition of so-called *visual facets*, in which photos are automatically organized into a hierarchical structure that the user can refine to better fit her purposes.
- Text-based search and browsing relies on tags, which are contextualized into a set of so-called *semantic facets*. This allows for traditional keyword-based search (e.g., `Italy AND water`) as well as for more accurate, facet-oriented, queries (e.g., `sport//Italy AND landscape//water`). Tags can be assigned to a photo either manually or even by means of a semi-automatic procedure that will suggest tags assigned to similar images (the latter case also avoids huge proliferation and replication of user-defined tags).
- Both semantic and visual facets are accessed through an integrated interface, where visual contents and tags are conjunctly used to focus the search. Similarly, tags and visual features can be integrated also for search purposes, thus leading to superior precision with respect to pure content-based search and to higher recall with respect to simple keyword queries.

The rest of the paper is organized as follows. In Section 2 we briefly describe the model on which Scenique is based. Section 3 presents the software architecture of Scenique and provides details on its basic modules. Section 4 presents the results of an evaluation test of the system, and Section 5 concludes.

2 The Model

Scenique is based on a simplified version of the *multi-structural* framework introduced in [6], which allows objects (photos, in our case) to be organized into a set of orthogonal dimensions, also called *facets*. Each facet can be conveniently viewed as a particular coordinate used to describe the content of a picture and is organized as a tree, where each node is assigned a label. Scenique supports both *semantic* and *visual* facets.

[1] A software demo of a preliminary version of the system has been presented in [5].

In case of semantic facets, node labels are tags, with the root node being tagged with the facet name. A same tag can appear in different facets as well as in different nodes of the same facet, which allows to discriminate between the different usages and/or meanings that different tag occurrences can have. A specific occurrence of a tag t in a tree therefore corresponds to a path in the tree and will be referred in the following as a *semantic tag*. For instance, the tag `Italy` might be used to label a node in the `geographic` facet (used to organize photos according to the place they have been shot) as well as to label a node in the `sport` facet (which only applies to photos related to sport events). Further, in the `sport` facet the tag `Italy` might lead to two (or even more) semantic tags, such as `sport/soccer/Italy` and `sport/basket/Italy`. In order to ensure compatibility with systems and devices that do not consider any tag organization (such as Flickr), the system-defined `default` facet is also provided. The `default` facet is simply a 2-level tree, with the root node being labelled with `default` and all tags appearing as child nodes.

Each photo P can be assigned a variable number of semantic tags. If a facet F is not relevant for P, then no semantic tag from F needs to be used to characterize P's content. On the other hand, P might be bound to multiple semantic tags from the same facet F, if this is appropriate. For instance, a picture with a dog and a cat might be assigned the two semantic tags `subject/animal/dog` and `subject/animal/cat`, both from the `subject` facet. Thus, although a facet provides a mean to classify images, this classification is not exclusive at the instance level, which provides the necessary flexibility to organize images.

The other type of facets supported by Scenique are the visual ones. In this case, the facet is built upon low-level features (such as color, texture, and shape) that are automatically extracted by images and organized into a (visual) tree. Each node in the tree actually corresponds to a cluster of features corresponding to photos sharing similar visual features and is labelled using a *representative photo* of that cluster. As for semantic facets, a photo is not forced to be part of a visual facet. For instance, if a visual facet `face` were defined, then only photo portraying people would be relevant for that facet.

3 System Overview

Figure 1 provides an overview of the architecture of Scenique. The main storage components are the *Photo DB*, the *Feature DB*, and the *Tag DB*.[2] The Photo DB stores global information of indexed photos, such as file system locations and thumbnails for fast visualization. The Feature DB manages automatically extracted features, that can be used to build visual facets and indexes, the latter needed for efficiently supporting content-based queries (see below for more details). The Tag DB stores for each image all its associated semantic tags.

[2] For clarity of presentation, in the following we introduce the three components separately. Actually, in our implementation they reside in the same relational DB.

Five major software components constitute the core of the Scenique architecture: the Semantic Facets Manager, the Visual Facets Manager, the Annotation Processor, the Browsing Processor, and the Query Processor.

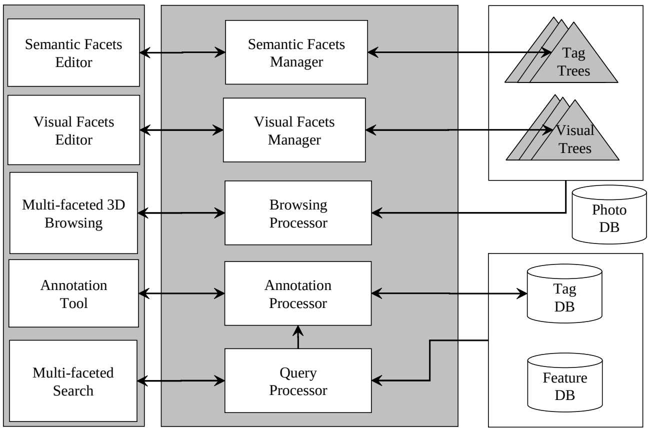


Fig. 1. The Scenique architecture.

The *Semantic Facets Manager* is in charge of the definition of semantic facets. To this end, the user is provided with a graphical interface that allows new dimensions to be defined and a *drag&drop* modality that permits new tags to be added. Further, a command-line modality is also available, in which semantic facets are defined using an SQL-like syntax (e.g., `insert into facet` *subject* `concept` *animal/mammal/dog*).

The *Visual Facets Manager* is derived from our previous PIBE system [7], that was designed to hierarchically cluster photos based on the similarity of their visual features and to provide the user with a set of basic graphical *personalization actions* (such as `fusion`, `split`, etc.) to revise the system-derived clusters. It also includes basic tools to automatically extract visual features from photos. In order to build a visual tree, the Visual Facets Manager recursively applies the partitioning *k-means* algorithm, starting with the whole photo DB and terminating when no more than k pictures are left in a cluster. Each node of the so-derived tree is then labelled with a representative image, that corresponds to the photo closer to the cluster centroid.

The *Annotation Processor* takes advantage of a set of pre-annotated photos to suggest tags for other images. Here we provide only some basic intuition on how tag suggestion works, a detailed description being given in [8]. The annotation process is essentially based on the idea of suggesting those tags that are assigned to photos similar to the target photo P. To this end, a nearest-neighbors search is first performed using low-level features, which determines a set S of images similar to P. For all tags t_j associated to at least one image in S, a frequency score f_j is then computed as the number of images in S annotated with t_j. Then, in order to remove unrelated tags, thus to improve the prediction accuracy, a correlation analysis is performed on each pair of tags (t_i, t_j). The so-resulting correlation scores $c_{i,j}$ are then used to determine whether or not t_i and t_j are connected in a graph G whose nodes are the candidate tags, and where the node of t_j is given weight f_j. Finally, a maximum-weight clique of G is determined, with nodes in the clique determining which are the tags to be suggested. This process can be focused only on the facets of interest, and the user can provide the necessary feedback by confirming correct tags, deleting wrong ones and/or adding other tags.

The *Browsing Processor* provides all the functionalities allowing users to navigate through a photo collection. Its graphical interface, shown in Figure 2, is composed by a 3-D *viewing room* and by a 2-D *facet panel*, that always stay synchronized. The user can start a browsing session by selecting some facets from the available list of semantic and visual dimensions (`animal`, `landscape`, and `feature` in the figure), which leads the corresponding facet trees to be displayed within the 2-D facet panel. By clicking on a node of interest in the facet panel the corresponding axis in the 3-D room is highlighted and related photos are displayed. Image cubes in the 3-D view represent clusters of images, whereas flat images correspond to single photos. The 3-D view is *active*, in that images in it are clickable, thus providing an alternate browsing modality with respect to the 2-D panel.

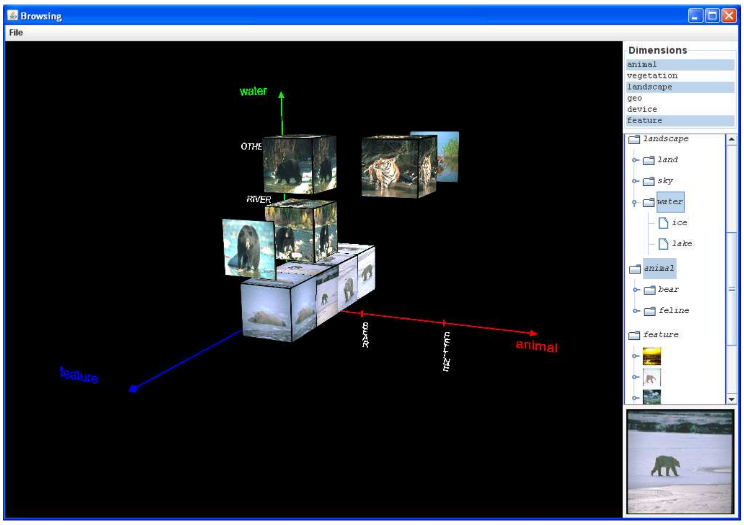
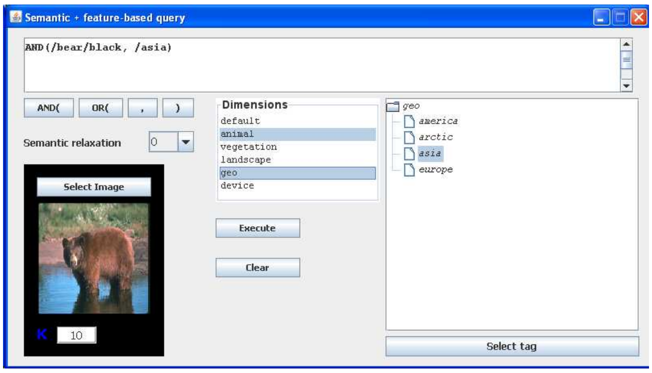


Fig. 2. Multi-faceted 3-D browsing interface.

The *Query Processor* is in charge of managing search requests. Scenique allows for content-based (or visual, V) and tag-based (T) queries, as well as for a combination of them (TV-queries). V-queries look for images similar to a given query image (either selected from a provided sample or input by the user), and are implemented using a k nearest-neighbors algorithm running on top of the Feature DB. For speeding-up query evaluation, features are indexed with M-trees [9]. Queries of type T are formulated using semantic facets and consists of a Boolean expression of semantic tags. Efficient resolution of T-queries is provided by an inverted index built on top of the Tag DB. Note that a T-query consisting, say, of the single term `animal/bear` will also retrieve all the photos that have a semantic tag more specific than the query term, e.g., `animal/bear/brown_bear`. Lexical ontologies (WordNet[3] in our implementation) are used when a search term do not belong to the Tag DB. Finally, the user is also given the possibility to specify some *semantic relaxation*. As an example, given the search term `animal/bear/asiatic_black_bear`, and in order to prevent a possibly low-cardinality result, the user can specify a certain degree of semantic relaxation, which will return also photo higher in the facet tree, i.e., those bound to `animal/bear` if one-level up in the tree is tolerated.

TV-queries, an example of which is shown in Figure 3, are a combination of above described modalities. Ranking of results gives priority to photos matching tags and within the top-k according to visual similarity, then to photos only matching tags, and finally to photos with only a good visual similarity.



Semantic + feature-based query

AND(/bear/black, /asia)

Semantic relaxation 0

K 10

Fig. 3. Formulating a tag&visual-based query.

4 System Evaluation

We implemented Scenique in Java JDK 5.0 and tested it on a database of 5,000 annotated images extracted from the Corel collection. Each photo was automatically segmented into a set of homogeneous regions which convey information

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

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 (see [10] for more details).

We conducted a user study over a set of 10 participants (5 males and 5 females) to evaluate the effectiveness of Scenique and the usability of its visual tools. The system setup for the user study was as follows. Several facets (among which `default`, `animal`, `vegetation`, `landscape`, `geographic`, and `device`) were available to the users for searching and exploring the photo collection.

Participants ranged from 25 to 50 years (average 35.8) and were all expert users of the Web and multimedia search engines. After a demonstration of the Scenique functionalities, the users were given a fixed time of 20 minutes to freely play with the system. Finally, they were requested to fill a questionnaire, using 0 as "strongly disagree" and 4 as "strongly agree" for each question.

Question	Agreement	Average
	0 1 2 3 4	Rating
1) "*I like the Scenique system*"	0 0 0 2 8	3.80
2) "*The multi-faceted paradigm helps me in searching and browsing*"	0 0 0 7 3	3.30
3) "*I found the integration of semantic tags and visual descriptors helpful*"	0 0 1 1 8	3.70
4) "*The expressive power in formulating requests satisfies me*"	0 1 0 2 7	3.50
5) "*I found visual tools intuitive and easy-to-use*"	0 1 0 1 8	3.60
Total average		3.58

Table 1. Mean user satisfaction ratings.

Table 1 shows that there is an overall positive agreement from all participants (3.58/4 score on average) on all the questionnaire statements. In particular, all users liked Scenique (question (1)) and, more important, believed that an integrated use of semantic annotations and visual descriptors is vital to get the retrieval process truly effective (question (3)). Most of the participants found the multi-faceted paradigm helpful for their tasks (question (2)) and they judged the expressive power in formulating requests to the system more than sufficient (question (4)). With respect to the Scenique user interface, most of the participants found it very intuitive and easy-to-use; only one user rated 1 and commented: "*I found the user interface intuitive in general; however, I would have liked the possibility to exchange the labels of the axes space in the 3-D viewing room rather than having to rotate the entire space*".

Such preliminary results are extremely encouraging, especially since the participants were not familiar with the image collection. We believe that results would be even better when using personal photo collections.

5 Conclusions

In this paper we introduced Scenique, a novel multi-faceted image search and browsing system for effectively managing personal photo collections. With Scenique an user can define multiple facets (exploiting both semantic tags and visual descriptors) with the aim to organize her photos under different points of view, and navigate through them in an integrated environment. In order to quickly locate images of interest, the user can also formulate queries combining tags and visual descriptors. Feedback provided by a set of real users testifies that Scenique is an effective system and that its GUI is intuitive and easy-to-use.

With respect to other systems that make use of both visual features and text annotations for searching and browsing image collections, see e.g., [11, 12], Scenique emphasizes the importance of integrating contextualized tags and visual descriptors to focus the search. Visual features are also the key for implementing a semi-automatic procedure able to suggest tags to photos of interest.

In order to improve the usability of Scenique we are currently considering techniques able to automatically induce hierarchical facets from text annotations, such as those provided by systems like Flickr. Although partial results along this way have been obtained (see e.g., [4]), the general problem is still unsolved. We are further studying suitable interfaces for importing predefined taxonomies.

References

1. Porta, M.: Browsing Large Collections of Images through Unconventional Visualization Techniques. In: Proceedings of AVI 2006, 440–444
2. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12) (2000) 1349–1380
3. Yee, K.P., Swearingen, K., Li, K., Hearst, M.A.: Faceted Metadata for Image Search and Browsing. In: Proceedings of ACM CHI 2003, 401–408
4. Dakka, W., Ipeirotis, P.G., Wood, K.R.: Automatic Construction of Multifaceted Browsing Interfaces. In: Proceedings of ACM CIKM 2005, 768–775
5. Bartolini, I., Ciaccia, P.: Scenique: A Multimodal Image Retrieval Interface. In: Proceedings of AVI 2008, 476–477
6. Fagin, R., Guha, R.V., Kumar, R., Novak, J., Sivakumar, D., Tomkins, A.: Multi-structural Databases. In: Proceedings of the ACM PODS 2005, 184–195
7. Bartolini, I., Ciaccia, P., Patella, M.: Adaptively Browsing Image Databases with PIBE. Multimedia Tools and Applications **31**(3) (2006) 269–286
8. Bartolini, I., Ciaccia, P.: Imagination: Exploiting Link Analysis for Accurate Image Annotation. Adaptive Multimedia Retrieval: Retrieval, User, and Semantics (LNCS) **4918** (2008) 322–44
9. Ciaccia, P., Patella, M., Zezula, P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In: Proceedings of VLDB 1997, 426–435
10. Ardizzoni, S., Bartolini, I., Patella, M.: Windsurf: Region-Based Image Retrieval Using Wavelets. In: Proceedings of IWOSS 1999, 167–173
11. Smith, G., Czerwinski, M., Meyers, B., Robbins, D., Robertson, G., Tan, D.S.: FacetMap: A Scalable Search and Browse Visualization. IEEE Transactions on Visualization and Computer Graphics **12**(5) (2006) 797–804
12. Quack, T., Mönich, U., Thiele, L., Manjunath, B.S.: Cortina: A System for Large-scale, Content-based Web Image Retrieval. In: Proceedings of the ACM MM 2004, 508–511