Thinking of a System for Image Retrieval

Giovanna Castellano
Università degli Studi di Bari
“Aldo Moro”
via Orabona 4
Bari, Italy
castellano@di.uniba.it

Gianluca Sforza[*]
Università degli Studi di Bari
“Aldo Moro”
via Orabona 4
Bari, Italy
gsforza@di.uniba.it

Alessandra Torsello
Università degli Studi di Bari
“Aldo Moro”
via Orabona 4
Bari, Italy
torsello@di.uniba.it

ABSTRACT

Increasing applications are demanding effective and efficient support to perform retrieval in large collections of digital images. The work presented here is an early stage research focusing on the integration between text-based and content-based image retrieval. The main objective is to find a valid solution to the problem of reducing the so called semantic gap, i.e. the lack of coincidence existing between the visual information contained in an image and the interpretation that a user can give of it. To address the semantic gap problem, we intend to use a combination of several approaches. Firstly, a linking between low-level features and text description is obtained by a semi-automatic annotation process, which makes use of shape prototypes generated by clustering. Precisely, the system indexes objects based on shape and groups them into a set of clusters, with each cluster represented by a prototype. Then, a taxonomy of objects that are described by both visual ontologies and textual features is attached to prototypes, by forming a visual description of a subset of the objects. The paper outlines the architecture of the system and describes briefly algorithms underpinning the proposed approach.

Categories and Subject Descriptors

H [**Information Storage and Retrieval**]

General Terms

Image retrieval

Keywords

Content-based image retrieval, Semantic image retrieval

1. INTRODUCTION

By the end of the last century the question was not whether digital image archives are technically and economically viable, but rather how these archives would be efficient and informative. The attempt has been to develop intelligent and efficient human-computer interaction systems, enabling the user to access vast amounts of heterogeneous image sets, stored in different sites and archives. Additionally, the continuously increasing number of people that should access to such collections further dictates that more emphasis be put on attributes such as the user-friendliness and flexibility of any multimedia content retrieval scheme.

The very first attempts at image retrieval were based on exploiting existing image captions to classify images according to predetermined classes or to create a restricted vocabulary [5]. Although relatively simple and computationally efficient, this approach has several restrictions mainly deriving from the use of a restricted vocabulary that neither allows for unanticipated queries nor can be extended without re-evaluating the possible connection between each item in the database and each new addition to the vocabulary. Additionally, such keyword-based approaches assume either the pre-existence of textual annotations (e.g. captions) or that annotation using the predetermined vocabulary is performed manually. In the latter case, inconsistency of the keyword assignments among different indexers can also hamper performance. Recently, a methodology for computer-assisted annotation of image collections was presented [24].

To overcome the limitations of the keyword-based approach, the use of the visual content has been proposed, leading to Content-Based Image Retrieval(CBIR) approaches [6]. CBIR systems utilize the visual content of images to perform indexing and retrieval, by extracting low-level indexing features, such as color, shape, and texture. In this case, pre-processing of images is necessary as the basis on which features are extracted. The pre-processing is of coarse granularity if it involves processing of images as a whole, whereas it is of fine granularity if it involves detection of objects within an image [1]. Then, relevant images are retrieved by comparing the low-level features of each item in the database with those of a user-supplied sketch or, more often, a key image that is either selected from a restricted image set or is supplied by the user (query-by-example). Several approaches have appeared in the literature which perform visual querying by examples taking into account different facets of pictorial data to express the image contents, such as color [21], object shape [2], texture [14], or a combination of them [8, 18, 20]. Among these, search by matching shapes of image portions is one of the most natural way to pose a query in image databases.

Though many sophisticated algorithms have been designed to describe color, shape, and texture features, these algorithms cannot adequately model image semantics. Indeed, extensive experiments on CBIR show that low-level contents

[*] Corresponding author

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
http://ims.dei.unipd.it/websites/iir10/index.html

often fail to describe the high-level semantic concepts in user's mind [25]. Also, CBIR systems have limitations when dealing with broad content image databases [16]; indeed, in order to start a query, the availability of an appropriate key image is assumed; occasionally, this is not feasible, particularly for classes of images that are underrepresented in the database. Therefore, the performance of CBIR systems is still far from user's expectations.

Summarizing, current indexing schemes for image retrieval employ descriptors ranging from low-level features to higher-level semantic concepts [23]. So far, significant work has been presented on unifying keywords and visual contents in image retrieval, and several hybrid methods exploiting both keywords and the visual content have been proposed [17, 12, 26]. Depending on how low-level and high-level descriptors are employed and/or combined together, different levels of image retrieval can be achieved. According to [7], three levels of image retrieval can be considered:

- Level 1: Low-level features such as color, texture, shape or the spatial location of image elements are exploited in the retrieval process. At this level, the system supports queries like *find pictures like this* or *find pictures containing blue squares.*

- Level 2: Objects of given type identified by low-level features are retrieved with some degree of logical inference. An example of query is *find pictures in which my father appears.*

- Level 3: Abstract attributes associated to objects are used for retrieval. This involves a significant amount of high-level reasoning about the meaning of the objects or scenes depicted. An example of query is *find pictures of a happy woman.*

Retrieval including both Level 2 and Level 3 together is referred to as semantic image retrieval. The gap between Level 1 and Level 2 is known as semantic gap, which 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 situation" [19]. Retrieval at Level 3 is quite difficult, therefore current systems mostly perform retrieval at Level 2, which requires three fundamental steps: (1) extraction of low-level image features, (2) definition of proper similarity measures to perform matching, (3) reducing the semantic gap. Clearly, step (3) is the most challenging one, since it requires providing a link between low-level features (visual data) and high-level concepts (semantic interpretation of visual data).

Currently, various approaches have been proposed to reduce the semantic gap between the low-level features of images and the high-level concepts that are understandable by human. According to [11], they can be broadly grouped into four main categories:

- Use of ontologies [15]. Ontologies can be used to provide an explicit, simplified and abstract specification of knowledge about the domain of interest; this is obtained by defining concepts and relationships between them, according to the specific purpose of the considered problem. This approach exploits the possibility to simply derive semantics from our daily language. Then, different descriptors can be related to the low-level features of images in order to form a vocabulary that provides a qualitative definition of high-level query concepts. Finally, these descriptors can be mapped to high level semantics, based on our knowledge. This approach works fine with small databases containing specifically collected images. With large collections of images with various contents, more powerful tools are required to learn the semantics.

- Automatic image annotation [22]. This approach consists in exploiting supervised or unsupervised learning techniques to derive high-level concepts from images. In particular, supervised learning techniques are used to predict values of a semantic category based on a set of training samples. However, supervised learning algorithms present some disadvantages strictly related to the nature of this kind of technique, that require a large amount of labeled data to provide effective learning results. This represents a problem when the application domain changes and new labeled samples have to be provided. Clustering is the typical unsupervised learning technique used for retrieval purpose. In this approach, images are grouped on the basis of some similarity measure, so that a class label is associated to each derived cluster. Images into the same cluster are supposed to be similar to each other (i.e. having similar semantic content). Thus, a new untagged image that is added to the database can be indexed by assigning it to the cluster that better matches with the image.

- Relevance feedback [13]. This approach concerns the possibility to learn the intentions of users and their specific needs by exploiting information obtained during their interactions with the system. In particular, when the system provides the initial retrieval results, the user judges these by indicating if they are relevant/irrelevant (and eventually the degree of relevance/irrelevance). Then, a learning algorithm is used to learn the user feedback, which will be exploited in order to provide results that better satisfy the user needs.

- Generating semantic templates [27]. This method is based on the concept of visual semantic template that includes a set of icons or objects denoting a personalized view of concepts. Feature vectors of these objects are extracted for query process. Initially, the user has to define the template of a concept by specifying, for example, the objects and their spatial and temporal constraints and the weights assigned to each feature for each object. Finally, through the interaction with users, the system move toward a set of queries that better express the concept in the user mind. Since this method requires the user to know the image features, it could be quite difficult for ordinary users.

Along with state-of-art directions in the field of IR, in this paper we present the idea of an IR system supporting retrieval at Level 2. Precisely, we intend to provide a solution to the problem of semantic gap in IR by designing a methodology based on a combination of several approaches, which is oriented to exploit both the visual and the semantic content of images. This is achieved making use of clustering and visual ontologies. In the following, all the approaches

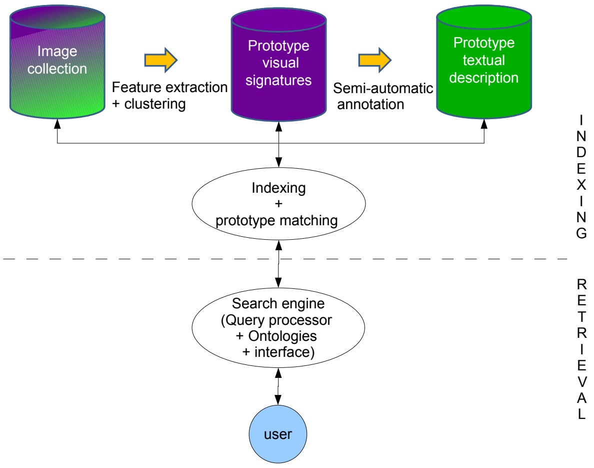


Image collection
Feature extraction + clustering
Prototype visual signatures
Semi-automatic annotation
Prototype textual description
Indexing + prototype matching
INDEXING
Search engine (Query processor + Ontologies + interface)
RETRIEVAL
user

Figure 1: The system architecture.

underpinning the proposed IR methodology are briefly described and the architecture of the system is outlined.

2. OVERVIEW OF THE IR SYSTEM

The proposed system is intended to perform image retrieval by exploiting both the visual and the semantic content of images. As concerns the visual content, in this preliminary phase of the research we focus only on shape content. In fact, we aim to deal with specific domain images containing objects that have a distinguishable shape meaning. Therefore, we assume that indexing and querying are only based on shape matching. The system will allow the user to query the image database not only by shape sketches and by keywords but also by "concepts describing shapes". The general architecture of the proposed IR system is reported in fig. 1.

As it can be seen, several tasks are carried out in order to derive visual and textual features of shapes contained in images. These tasks are:

1. Feature extraction: detecting shapes in images;
2. Clustering: grouping similar shapes into prototypes;
3. Semi-automatic annotation: associating keywords to prototypes;
4. Search.

In the following we describe how each task is carried out.

2.1 Feature extraction

In the proposed system, each image in the database is stored as a collection of objects' shapes contained in it. In order to be stored in the database, every image is processed to identify objects appearing in it. Image processing starts with an edge detection process that extracts all contours in the image. Then, using the derived edges, a shape detection process is performed to identify different objects included in the image and determine their contours. Finally, Fourier descriptors are computed on each contour and retained as visual signatures of the objects in a separate database.

2.2 Clustering

Once all shapes have been detected from images and represented as visual signatures vectors, a set of shape prototypes is automatically defined by an unsupervised learning process that performs clustering on visual signatures (Fourier descriptors) of shapes, so as to categorize similar shapes into clusters. Each resulting cluster C_i is represented by a shape prototype $\mathbf{p}_i$, that is computed by averaging visual signatures of all shapes belonging to the cluster. We intend to apply a hierarchical clustering, in order to generate a hierarchy of prototypical shapes. Each node of the hierarchical tree is associated with one prototypical shape. Root nodes of the tree represent general prototypes, intermediate nodes represent general shapes, leaf nodes represent specific shapes.

During the interaction of the user with the system, the hierarchical tree is incrementally updated. Whenever a new shape is considered (i.e. each time a new image containing relevant object shapes is added to the database), we evaluate its matching against all existing prototypes, from root nodes to *pre-leafs*(final) nodes, according to a similarity measure defined on visual signatures. If the new shape matches a final prototype with a sufficient degree, then the corresponding prototype is updated by averaging the features of shapes that belong to the corresponding cluster [10]. Otherwise, a new prototype is created, corresponding to the new shape.

The use of shape prototypes, which represent an intermediate level of visual signatures, facilitates the subsequent tasks 3. and 4. Actually, prototypes facilitate the annotation process, since only a reduced number of shapes (the prototypical ones) need to be manually annotated. Secondly, the use of prototypes simplifies the search process. Indeed, since only a small number of objects is likely to match any single user query, a large number of unnecessary comparisons is avoided during search by performing matching with shape prototypes rather than with specific shapes. In other words, prototypes acts as a filter that reduces the search space quickly while discriminating the objects.

2.3 Semi-automatic annotation

Once shape prototypes have been derived, a semi-automatic annotation process is applied to associate text descriptions to identified object shapes. The process is semi-automatic since it involves a manual annotation only for prototypes: shapes immediately attached in the hierarchy are automatically annotated, since they inherit descriptions from their prototypes.

Every semantic class that is of interest in the considered image domain (e.g. for ours, glasses, bottles, etc.) will be described by a *visual ontology* (VO), which is intended as a textual description, made of concepts and relationships among them, of the visual content of a prototypical shape [9, 4]. We figure the lexicon used to define the VOs to be as much intuitive as possible, so as to evocate the particular shape it describes. We plan that the system will be supplied of a basic set of domain dependent VOs, one for each considered semantic class.

Of course, different prototypical shapes may convey the same semantic content (e.g., several different shapes may convey the concept of *glass*). We consider such prototypes to belong to the same semantic class. Shape prototypes belonging to the same semantic class will share about the same VO structure, obviously with the appropriate differences.

As an illustrative example, we sketch some possible relationships included in a VO that refers to the semantic class *glass*:

- *wine glass IS SPECIALIZATION OF glass*;
- *bottom IS PART OF wine glass*;
- *wavy shape IS PROPERTY OF bottom.*

The combined use of prototypes and VOs provides a powerful mechanism for automatic annotation of shapes. Every time the user adds a new shape to the database, the system associates the shape to the most similar prototype, which is related to a semantic class and linked to a VO. Thus the new shape inherits all the semantic descriptions associated to the selected prototype in an automatic fashion. Then, a feedback from the user is considered. Namely, the user may accept the choice operated by the system, or reject it. In the latter case, there are two possibilities: the user can select the proper prototype with the related VO from the existing ones, or, if no one can be associated to the shape, the user can create a new prototype (using the new shape) and manually annotate it by modifying the VO incorrectly assigned by the system previously.

2.4 Search

The engine mechanism is designed to allow users to submit sketch-based, text-based and concept-based queries.

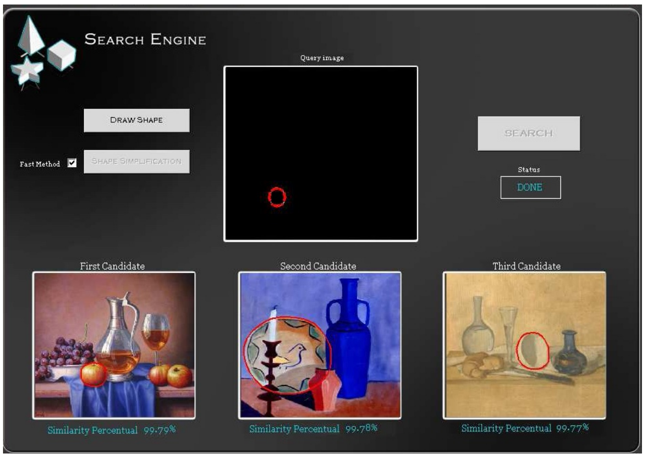
The results of the sketch-based search emerge from a matching between the submitted sample shape and the created prototypes. Precisely, when the user presents a query in the form of an object sketch, the system formulates the query, performing feature extraction by translating that object into a shape model. The extracted query feature is submitted to compute similarity between the query and prototypes first. This is made by considering shapes as points of a feature space. Having characterized each shape as a vector of Fourier descriptors, we simply evaluate dissimilarity between two shapes in terms of Euclidean distance between two vectors of descriptors. Of course, other similarity measures can be considered, encapsulating the human perception of shape similarity (this is an interesting issue that we would like to deepen in future). After sorting the prototypes in terms of similarity, the system returns images containing objects indexed by the prototypes with highest similarities.

The results of the text-based search emerge from a matching between the submitted textual query and textual descriptions associated to prototypes. Namely, when a query is formulated in terms of keywords, the system simply returns images including the objects indexed by the prototypes labeled with that keywords. As before, high-matching prototypes are selected to provide shapes to be visualized as search results.

Finally, when both a visual and textual content are exploited by the user querying the image database, images returned from the two approaches separately, are merged together in a single output set.

3. FIRST STEPS TOWARD THE SYSTEM DEVELOPMENT

In this preliminary phase of the research, only the main functions for tasks 1. and 4. described above have been implemented in the system. For tests during the development of the system, we considered an image database from the art domain. The database, used in other IR works [3] includes digitalized images representing still-object paintings by the Italian artist Giorgio Morandi.



Fast Method ☑

Status: DONE

Similarity Percentual 99.79%

Similarity Percentual 99.78%

Similarity Percentual 99.77%

Figure 2: An initial search engine interface.

As concerns task 1., various image processing tools that are necessary to extract shape features from the image objects have been developed, including edge detection methods, as well as enhancement and reconstruction functionalities. Basic image processing methods were included from the *ImageJ* image analysis software[1], such as thresholding methods (e.g. Canny, Prewitt and Sobel) for automatic detection of objects boundaries lying in images. Having the possibility to act on contrast and brightness properties, the user can adjust the image appearance to refine the extraction of the shapes of objects. The shape identification is made automatically through an edge following algorithm. When the result of shape identification is not satisfying, the user is given the possibility to correct boundaries or to manually draw boundaries directly on the image.

As concerns task 4., the retrieval graphical interface has been developed, that enables users to query the system and to inspect search results (fig. 2). Also, the computation of Euclidean dissimilarity measures for shape prototype matching has been included in the system.

Currently, the system provides also the interfaces for browsing the database and insert new images.

4. CONCLUSIONS

In this paper a preliminary proposal of an IR system has been presented. The system is intended to solve the problem of semantic gap by exploiting clustering and visual ontologies. The use of a visual ontology is motivated by the necessity of reproducing the capacity of a human in describing her visual perception by means of the visual concepts she possesses. From the point of human-computer interaction view, visual ontologies provide a bridge between low-level features of images and visual representation of semantic contained in images. Compared to symbolized ontology, visual ontologies can represent complex image knowledge in a more detailed and intuitive way, so that no expert knowledge is needed to process a complicated knowledge representation of images.

[1] http://rsbweb.nih.gov/ij

The binding created by visual ontologies between image objects and their description, enables the proposed IR system to perform a conceptual reasoning on the collection of images, also when treating with pure content-based queries. Thus, different forms of retrieval become possible with the proposed system:

1. text-based: queries are lexically motivated, i.e. they express objects by their names (keywords);
2. content-based: queries are perceptually motivated, i.e. they express objects by their visual apparency;
3. semantic retrieval: queries are semantically motivated, since they express objects by their intended meaning, i.e. in terms of concepts and their relationships.

Currently, we are continuing to develop the proposed IR system. To this aim, we are looking for the best appropriate clustering algorithm to derive significant shape prototypes and analyzing methods to create visual ontologies.

5. REFERENCES

[1] W. Al-Khatib, Y. F. Day, A. Ghafoor, and P. B. Berra. Semantic modeling and knowledge representation in multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):64–80, 1999.

[2] S. Arivazhagan, L. Ganesan, and S. Selvanidhyananthan. Image retrieval using shape features. *International journal of imaging science and engineering (IJISE)*, 1(3):101–103, 2007.

[3] A. D. Bimbo and P. Pala. Visual image retrieval by elastic matching of user sketches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:121–132, 1997.

[4] M. Bouet and M.-A. Aufaure. *Multimedia Data Mining and Knowledge Discovery*, chapter New Image Retrieval Principle: Image Mining and Visual Ontology, pages 168–184. Springer, 2007.

[5] S. Christodoulakis, M. Theodoridou, F. Ho, M. Papa, and A. Pathria. Multimedia document presentation, information extraction, and document formation in minos: a model and a system. *ACM Trans. Inf. Syst.*, 4(4):345–383, 1986.

[6] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.

[7] J. Eakins and M. Graham. Content-based image retrieval. University of Northumbria Technical Report, 1999.

[8] A. K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29:1233–1244, 1996.

[9] S. Jiang, T. Huang, and W. Gao. An ontology-based approach to retrieve digitized art images. In *WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 131–137, Washington, DC, USA, 2004. IEEE Computer Society.

[10] K.-M. Lee and W. Street. Cluster-driven refinement for content-based digital image retrieval. *Multimedia, IEEE Transactions on*, 6(6):817–827, 2004.

[11] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.*, 40(1):262–282, 2007.

[12] Y. Lu, C. Hu, X. Zhu, H. Zhang, and Q. Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia*, pages 31–37, New York, NY, USA, 2000. ACM.

[13] S. MacArthur, C. Brodley, and C.-R. Shyu. Relevance feedback decision trees in content-based image retrieval. In *IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVLŠ00)*, pages 68–72, 2000.

[14] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(8):837–842, 1996.

[15] V. Mezaris, I. Kompatsiaris, and M. Strintzis. An ontology approach to object-based image retrieval. In *ICIP 2003*, volume II, pages 511–514, 2003.

[16] R. Mojsilovic and B. Rogowitz. Capturing image semantics with low-level descriptors. In *Proc. of ICIP*, pages 18–21, 2001.

[17] M. Naphade, T. Kristjansson, B. Frey, and T. Huang. Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval in multimedia systems. *Image Processing, International Conference on*, 3:536, 1998.

[18] P. Pala and S. Santini. Image retrieval by shape and texture. *Pattern Recognition*, 32:517–527, 1999.

[19] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[20] J. R. Smith and S. Chang. Local color and texture extraction and spatial query. In *Proc. of IEEE Int. Conf. on Image Processing*, volume 3, pages 1011–1014, Sep 1996.

[21] J. R. Smith and S. fu Chang. Tools and techniques for color image retrieval. In *IS&T/SPIE Proceedings, Storage & Retrieval for Image and Video Databases*, volume 2670, pages 426–437, 1996.

[22] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *IEEE Transaction on Image Process*, 10(1):117–130, 2001.

[23] A. Yoshitaka and T. Ichikawa. A survey on content-based retrieval for multimedia databases. *IEEE Trans. on Knowl. and Data Eng.*, 11(1):81–93, 1999.

[24] C. Zhang and T. Chen. An active learning framework for content-based information retrieval. *IEEE Transactions on Multimedia*, 4:260–268, 2002.

[25] X. S. Zhou and T. S. Huang. Cbir: from low-level features to high-level semantics. *Image and Video Communications and Processing 2000*, 3974(1):426–431, 2000.

[26] X. S. Zhou and T. S. Huang. Unifying keywords and visual contents in image retrieval. *IEEE MultiMedia*, 9(2):23–33, 2002.

[27] Y. Zhuang, X. Liu, and Y. Pan. Apply semantic template to support content-based image retrieval. In *Storage and Retrieval for Media Databases*, volume 3972, pages 442–449, 1999.