Spatially Organized Visualization of Image Query Results

Gianluigi Ciocca[a], Claudio Cusano[a], Simone Santini[b], and Raimondo Schettini[a]

[a]Università degl Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy
[b]Universidad Autónoma de Madrid, C/ Tomas y Valiente 11, 28049 Madrid, Spain

ABSTRACT

In this work we present a system which visualizes the results obtained from image search engines in such a way that users can conveniently browse the retrieved images. The way in which search results are presented allows the user to grasp the composition of the set of images "at a glance". To do so, images are grouped and positioned according to their distribution in a *prosemantic feature space* which encodes information about their content at an abstraction level that can be placed between visual and semantic information. The compactness of the feature space allows a fast analysis of the image distribution so that all the computation can be performed in real time.

Keywords: Image retrieval, content-based image analysis, image browsing

1. INTRODUCTION

We present here a system for the visualization and browsing of the results obtained by querying image search engines such as flickr®or Google images. Usually, on-line image collections are organized by metadata such as captions and tags which allows the users to perform searches by means of textual queries. The results are presented as collections of "pages" where thumbnails, generated from the retrieved images, are placed on a regular rectangular grid, sorted according to some predefined criterion (relevance, time stamp...). Although this approach is highly space-efficient, it may present some drawbacks. The order in which the images are displayed may not be very obvious to the user, who has to cycle through several pages to find an appropriate image. Moreover, groups of very similar images are often placed on the same page which would then convey a small amount of information. In practice, a single page is seldom enough to communicate the general composition of the whole set of retrieved images.

Several works addressed the problem of visualizing sets of images on the basis of the associated metadata.[1–3] However, metadata are already taken into account by the search engines. On the other hand, visual content is usually ignored. Therefore it may be considered as an additional source of information which can be exploited to provide a more convenient and efficient way to browse the results of the queries.

Content-based approaches have been used for the visualization of large database of images.[4–6] Since indexing is performed off-line, very powerful methods can be used to extract rich descriptions of the content of the images, and to analyze their distribution in the feature space.

In this work we applied a real-time, content-based approach which queries an image search engine, and displays the results in such a way that the user can understand "at a glance" the composition of the set of retrieved images.

The proposed system heavily relies on image features which combine low-level visual properties with automatically extracted semantic information about the content of the images. These features, which we called *prosemantic* (from "towards the meaning"),[7,8] are compact 56-dimensional feature vectors. The small dimensionality of the feature space allows for an efficient analysis of the distribution of the images. As a result, the whole visualization process can be performed in real time.

Images are displayed on a plane in such a way that similar images (either visually or semantically) are grouped in the same region of the plane. Images are not allowed to cover each over: those which would overlap are recursively grouped into clusters which can be explored independently.

Gianluigi Ciocca: ciocca@disco.unimib.it, Claudio Cusano: claudio.cusano@disco.unimib.it, Simone Santini: simone.santini@uam.es, Raimondo Schettini: schettini@disco.unimib.it

Multimedia on Mobile Devices 2011; and Multimedia Content Access: Algorithms and Systems V, edited by David Akopian, Reiner Creutzburg, Cees G. M. Snoek, Nicu Sebe, Lyndon Kennedy, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7881, 788116 · © 2011 SPIE-IS&T · CCC code: 0277-786X/11/$18 · doi: 10.1117/12.876258

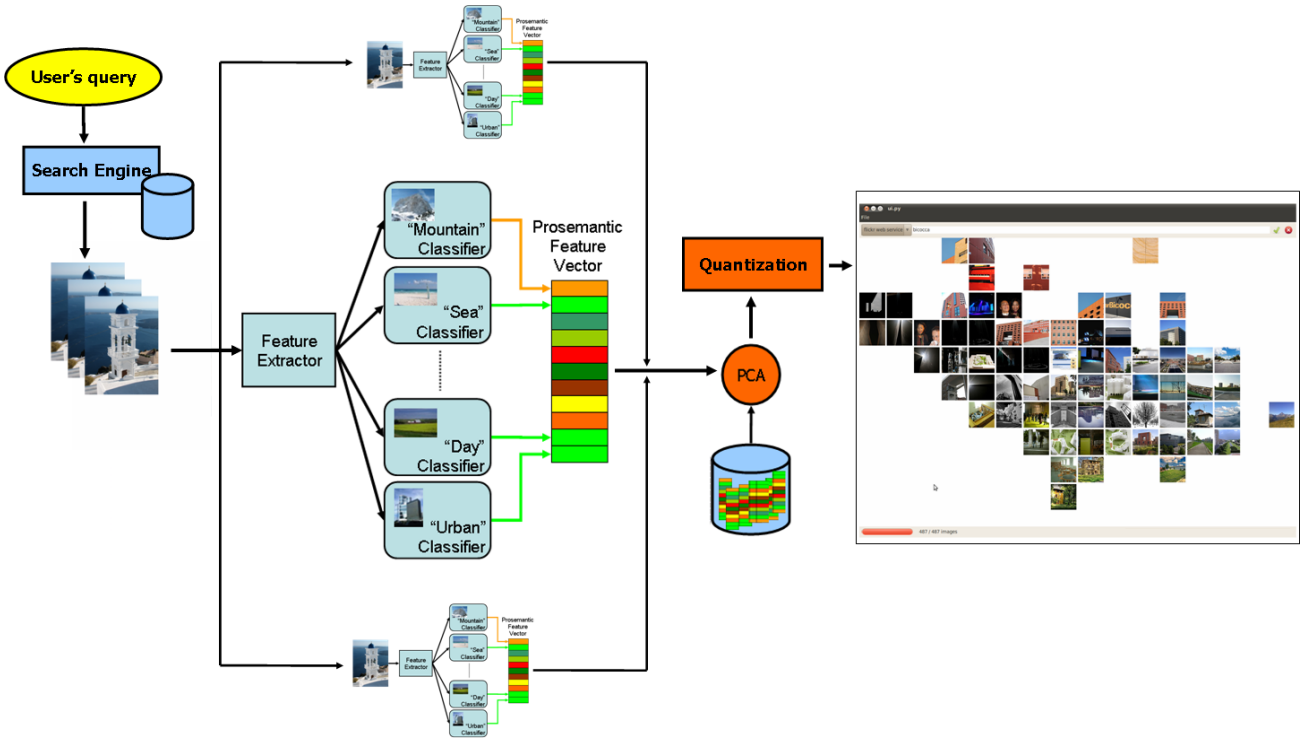

Figure 1. Scheme of the architecture of the proposed system.

2. SYSTEM ARCHITECTURE

The execution of each query is composed of three major steps:

- query processing;
- image description;
- thumbnail placement and visualization.

Figure 1 shows a more detailed overview of the system. The steps in the pipeline are performed asynchronously on each image, which is shown to the user as soon as its processing is finished.

In the query processing step the text entered by the user is sent to an image search engine such as Google images, flickr®... The search engine returns a set of images whose content is described by a set of prosemantic features.

2.1 Image Description

We originally introduced prosemantic features in a previous work[7] and we further explored them in another paper.[8] Briefly, the prosemantic approach try to retain the advantages of using image classifiers (the possibility of introducing semantics) without the disadvantages (a system based on classifier is useful only inasmuch the queries are made on the categories on which it was trained).

images are first described by a set of content ("low-level") features. These features are used as input to an array of 56 soft classifiers, trained to recognize a set of 14 partially overlapping classes. The output of the classifiers forms a 56-dimensional vector which characterizes the content of the image. The four low-level features are:

- first two statistical moments of the components in the LUV color space, computed on the image divided into 9×9 blocks;

- a eight-bin edge direction histogram, computed on the luminance image subdivided into 8×8 blocks;
- a color histogram obtained by a $8 \times 8 \times 8$ uniform quantization of the RGB color space;
- a bag-of-features descriptor obtained by quantizing SIFT descriptors according to a codebook of 1096 visual words.

The 14 classes considered are: "animals", "city", "close-up", "desert", "flowers", "forest", "indoor", "mountain", "night", "people", "rural", "sea", "street", and "sunset". A Support Vector Machine has been trained for each combination of a class and a low-level feature. The "soft" output of the resulting 56 SVMs forms the prosemantic feature vector. The use of soft classification scores, together with the choice of independently process each low-level feature, places prosemantic features in an intermediate level between low-level representation and semantic annotation.

In the experimentation reported in our previous works[7,8] prosemantic features demonstrated to perform significantly better for image retrieval than low-level features. In our opinion these performance derives from the capability of prosemantic features of allowing a better match against users' intuition about the similarity of the images. In fact, reasoning about low-level features requires some basic understanding of image processing. On the basis of the results obtained we decided to investigate the application of prosemantic features on different tasks, such as visualization and browsing (as is the case here).

In this work we modified the features to allow their real-time extraction. More in detail, we considered only three low-level features and we computed them on a subsampled version of the images (the thumbnails returned by the search engine). We decided to drop the bag-of-features descriptor because of its high computational cost and its low reliability when computed on low resolution images. Therefore, in this case the prosemantic feature vectors consist of 42 components.

2.2 Image visualization and browsing

Images have been described as points in a 42-dimensional prosemantic space. Unfortunately, computer monitors are limited to two dimensions only. Therefore, we need to reduce the dimensionality of the space. For this purpose, we decided to adopt a standard, widely used technique: Principal Component Analysis (PCA).

We investigated two different approaches: in the first the PCA basis is computed on-line on the prosemantic features extracted on the downloaded images; in the second approach the basis has been precomputed on a large dataset of images (here, we used the training set of the Pascal VOC 2007 challenge[9]). We observed that the precomputed basis selects two principal components which roughly represent the open/closed and natural/man-maid dichotomies. The on-line approach produces less predictable results, since the basis is specific for each query.

The visualization in the subspace of the first two principal components allows showing to the users the result of their queries "at a glance". However, this visualization strategy is far from optimal: some regions of the space can result too crowded with an unwanted overlapping of the images which can also be completely hidden under their neighbors.

In order to simplify the presentation of the images we applied a quantization to the transformed prosemantic space. Neighbor images are recursively grouped in clusters, and only a representative example of the cluster is shown to the user (who can select the cluster to explore its content). The quantization process works as follows:

- a set of seed points is generated according to a hexagonal grid which uniformly subdivides the plane of the first two principal components;
- images are grouped into clusters according to the nearest seed point;
- for each cluster, the nearest image is selected as representative;
- the representative images are shown at the coordinates of the corresponding seed point.

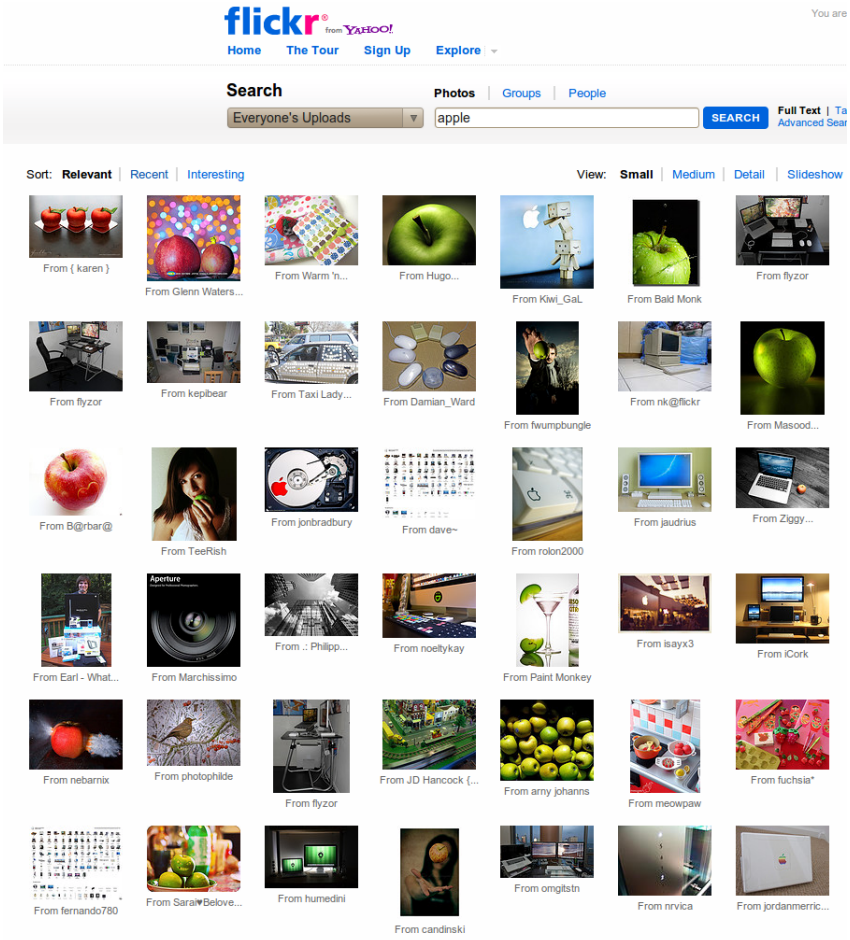

Figure 2. Results for the query "apple" submitted to the flickr® service.

Note that some seed point may correspond to an empty cluster; in this case an empty space will be left on the screen.

Figure 2 shows the search results of flickr® for the query "apple", sorted by decreasing relevance. Figure 3 shows the first 200 images visualized following our strategy. The downloaded images can be placed in different semantic categories: images representing fruits have been placed on the top while images with a technological theme have been placed on the bottom part of the screen.

Even if our system has been designed to visualize the results of image queries, it can be used to explore any collection of images. Figure 4 shows, for instance, the output obtained on the 120 images supplied for the "HP Challenge 2010: High Impact Visual Communication".

The proposed strategy has been implemented in a prototype application. The prototype submits the queries to the flickr® web service. We chose to use flickr® for its high popularity, and for its ease of integration with external applications. The adoption of others search engines (such as Google images) should be straightforward.

The application processes each image independently in parallel. An image is displayed when its processing is complete.

The application implements the two different visualization strategies presented in Section 2.2, together with more traditional viewing paradigms (e.g. a grid of thumbnails). During the demonstration attendees will be allowed to submit their own queries, and the results will be shown in real time.

Figure 3. Visualization of the 200 images which correspond to the query "apple". Overlapping images have been grouped into clusters (singleton clusters are displayed without the indication of their size).

Figure 4. Visualization of the 120 images of the"HP Challenge 2010: High Impact Visual Communication".

3. CONCLUSIONS

We presented here a system for the visualization of the results obtained from image search engines. The proposed strategy is content-based, where the content of the images is represented by prosemantic features which combine low-level visual properties with automatically extracted semantic information. The user can submit any textual query to the system and the retrieved set of images is displayed in such a way that its composition can be understood "at a glance".

Our future plans about this work include an extensive evaluation of the proposed strategy. We are also investigating how it would be possible to take into account ancillary information (time, captions...) as well.

REFERENCES

[1] Dontcheva, M., Agrawala, M., and Cohen, M., "Metadata visualization for image browsing," in [*ACM Symposium on User Interface Software and Technology*], (2005).

[2] Yee, K., Swearingen, K., Li, K., and Hearst, M., "Faceted metadata for image search and browsing," in [*Proceedings of the SIGCHI conference on Human factors in computing systems*], 408–415 (2003).

[3] Chang, M. and Leggett, J., "Collection understanding through streaming collage," in [*Proc. of the Information Visualization Interfaces for Retrieval and Analysis (IVARA) Workshop, associated with the Joint Conference on Digital Libraries*], (2004).

[4] Nguyen, G. and Worring, M., "Interactive access to large image collections using similarity-based visualization," *Journal of Visual Languages & Computing* **19**(2), 203–224 (2008).

[5] Ryu, D., Chung, W., and Cho, H., "PHOTOLAND: a new image layout system using spatio-temporal information in digital photos," in [*Proceedings of the 2010 ACM Symposium on Applied Computing*], 1884–1891 (2010).

[6] Achanta, R., Shaji, A., Fua, P., and Süsstrunk, S., "Image summaries using database saliency," in [*ACM SIGGRAPH ASIA 2009 Posters*], 1 (2009).

[7] Ciocca, G., Cusano, C., Santini, S., and Schettini, R., "Pro-semantic features for content-based image retrieval," in [*Proc. of 7th Workshop on Adaptive Multimedia Retrieval*], **In Press** (2009).

[8] Ciocca, G., Cusano, C., Santini, S., and Schettini, R., "Halfway through the semantic gap," *Information Sciences* ((submitted)).

[9] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A., "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results." http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.