Integrating Collaborative Tagging and Emergent Semantics for Image Retrieval

Melanie Aurnhammer
Sony Computer Science Laboratory Paris
6 rue Amyot
75005 Paris, France
melanie@csl.sony.fr

Peter Hanappe
Sony Computer Science Laboratory Paris
6 rue Amyot
75005 Paris, France
hanappe@csl.sony.fr

Luc Steels*
University of Brussels (VUB), AI Lab
Pleinlaan 2
1050 Brussels, Belgium
steels@arti.vub.be

ABSTRACT

In this paper, we investigate the combination of collaborative tagging and emergent semantics for improved data navigation and search. We propose to use visual features in addition to tags provided by users in order to discover new relationships between data. We show that our method is able to overcome some of the problems involved in navigating databases using tags only, such as synonymy or different languages, spelling mistakes, homonymy, or missing tags. On the other hand, image search based on visual features can be simplified substantially by the use of tags. We present technical details of our prototype system and show some preliminary results.

1. FROM IMAGE RETRIEVAL TO EMERGENT SEMANTICS

Searching large databases of images is a well known problem commonly referred to as image retrieval. The main challenge in this field has been quoted as automatically generating high-level descriptions of images. This is often called "bridging the semantic gap" between high-level descriptions a user will naturally look for, and low-level visual features that can be extracted from the data. This problem is known to be hard, if not impossible to solve. In order to provide a high-level description of an image, two requirements need to be fulfilled: first, the objects in an image need to be recognised. This is, however, not yet possible in a general way. Second, starting from the image components, the system would need to be able to infer the meaning of the image. It has been argued though that it is not always possible to break down the meaning of an image into its components. Images in general do not have an intrinsic meaning but the meaning emerges from the interaction with a user and by placing an image into the context of other images. This concept is called *emergent semantics* [7] (see also [3, 8]).

According to this argument, solving the problem of automatically annotating an image in a general way is not feasible, even if it was possible to detect all its objects. It was long assumed that manual annotations for images would not be provided by users because it is a tedious and time consuming process. However, recent developments like Flickr [1] have shown that this is no longer true. Users do tag images, mainly their own, but also those of other users. Reasons for this development might be the motivation that comes along with the exposure of personal photos to others. If a photo is tagged in a reasonable way, others can find it and leave comments that are mostly encouraging. User tagging does, however, not only work for personal data. Sites like Last.fm [2] where user tag music tracks are very successful as well.

These recent developments indicate a shift of the traditional image retrieval paradigm. Provided with the right motivation, users do give high-level descriptions for images which seem to make the need for automatic annotation less relevant. We believe that the new challenge is to combine both user-provided image descriptions and low-level visual features extracted from the data.

2. PERSONAL AND SOCIAL TAGS

Personal tags, introduced by a specific person for himself, are very different from expert tags. Personal tags are mostly associative, high-level, subjective, and inconsistent. Expert tags, on the other hand, are an attempt to be objective and consistent. The underlying assumption of expert tags is that all users have the same interpretation of the data, and the same categorisation. This is, however, not consistent with the concept of emergent semantics.

Personal tags become social tags when exposed to others, as seen on Flickr, for example. Although still associative, subjective, and inconsistent, the mass effect of social tags leads to some form of relevance and coherence. The exposure to the tags and photos of other users creates an implicit feedback between users and the tagging system. It can be observed that over time, the relative frequency of tags used to label a resource tends to approach a constant value [5]. This indicates that collaborative tagging is able to coordinate the actions of Web users to create coherent annotations of the shared photos. This confirms the findings of earlier studies that used computational models to investigate the emergence of a lexicon shared by a population of autonomous agents [9, 10].

*Luc Steels is also with the Sony Computer Science Laboratory Paris, 6 rue Amyot, 75005 Paris, France

WWW2006, May 22–26, 2006, Edinburgh, UK.
.

3. INTEGRATING TAGS AND IMAGE FEATURES

Although a powerful tool to navigate image archives, searching by tags alone has a number of drawbacks. First, people make mistakes while tagging, such as spelling mistakes, or accidential tagging with the wrong tag. Second, synonymy or different languages can only be handled by explicit tagging. Third, there is no solution to cope with homonymy, i.e. to distinguish different meanings of a word. We believe that by combining tags and visual features, we can overcome the disadvantages of both approaches taken on their own.

In order to approach this problem, we first need to consider the relation between tags and visual features. Our first observation is that tags provide, in general, high-level descriptions that cannot be derived from image features. Although some tags can be related to visual features (the tag can be *grounded*) e.g. "red", or "blackandwhite", this is rather an exception than the rule. Second, a personal tag can refer to images that do not possess a common visual content. For example, the tag "paris" might be assigned to a photo of the Eiffeltower, a photo of a restaurant, and a photo of a hotel room. For social tags, this problem is even more severe. Even if there was visual consistency within a set tagged by one user, consistency cannot be generally assumed to be found across several users. A set of images sharing the same social tag represents the union of categories from different users. Since we do not believe that all users share excactly the same categories, visual consistency among images tagged in the same way is not very likely to occur.

Thus, we cannot assume that a tag simply corresponds to a category. Therefore, we do not seek to capture the semantics for tags directly, but for a set of images, or a collection, explicitly selected by the user. A collection is aquired through the actions of a user, such as browsing, searching and selecting images.

4. TECHNICAL DETAILS

We tested our ideas by implementing an interface that combines intuitive navigation by tags and a user-friendly way to search according to visual features. The images and tags used by our system come from real users, downloaded from Flickr. The current version includes about 3000 photographs from 12 randomly chosen users.

In the following, we describe the visual features and the implementation of our retrieval process. Most of the techniques we used reflect either the state-of-the-art in image retrieval or are well-established standards in image analysis and pattern recognition. The idea of our system is to advance neither of these fields but to use the available tools in a new, intuitive, and creative way.

4.1 Features

We intentionally employ simple global features in our system. Rather than trying to recognise objects or even explain the meaning of an image, we seek to measure a certain "atmosphere", or a vague visual pattern, which we believe is possible to capture by low-level image features.

The visual features we used are colour and texture, i.e.

$$F = \{f_i\} = \{\text{colour,texture}\}$$

4.1.1 Colour Features

Comparison of colour histograms is known to be sensitive to small colour variations caused e.g. by lighting conditions. In order to obtain a more robust and simpler measure of the colour distribution, we calculate the first two moments (mean and standard deviation) in RGB colour space. In addition, we use the standard deviation between the means of the three colour channels. Intuitively, this yields a measure for the "colourfulness" of an image. The feature has a value of zero for grey-scale images and increases for images with stronger colours. We map the values to a logarithmic scale in order to distribute them more equally. In total, the colour feature vector has thus seven dimensions.

4.1.2 Texture Features

Texture refers to the properties that represent the surface or structure of an object. In our work, we seek to employ texture features that give a rough measure of the structural properties, such as linearity, periodicity, or directivity of an image. In experiments, we found *oriented gaussian derivatives* (OGD) to be well-suited for our purposes [4]. This feature descriptor uses the steerable property of the OGD to generate rotation invariant feature vectors. It is based on the idea of computing the "energy" of an image as a steerable function.

The features are extracted by a 2nd order dyadic pyramid of OGDs with four levels and a kernel size of 13x13. The generated feature vector has 24 dimensions. The first order OGD can be seen as a measure of "edge energy", and the second order OGD as a measure of the "line energy" of an image.

4.1.3 Feature Integration

The distance between a query image and an image in the database is calculated according to the $l2$ norm (Euclidean distance). We use a linear combination of the distances in the colour and texture spaces to combine both features. In order to give the same initial weight to all features, the values are normalised linearly before calculating the distance. The joint distance d between a database image x_l and a query image s_k over all features spaces f_i is thus

$$d(x_l, s_k) = \sum_{i=1}^{N} w_i d_i, \qquad \text{with} \quad \sum_{i=1}^{N} w_i = 1$$

where N is the number of features in the set F and w is a weighting factor (see section 4.2.3). For the initial query, $w = \frac{1}{N}$.

4.2 Search Process

The search for visually similar images starts with one or more images selected by the user. These initial images can be found through tags. In our implementation, we focussed on a totally user defined process: Not only is the number of selected images left to the user, he is also free in all further actions to take. When the results of the similarity search are displayed, the user can either (1) exclude images, (2) select images for refinement, (3) combine (1) and (2), or (4) simply not take any action. This distinguishes our approach from methods suggested for *relevance feedback* in image retrieval (see e.g. [6]), where the user is forced to take certain actions, such as giving feedback to every retrieved image, or where he has to follow a strict order of interaction.

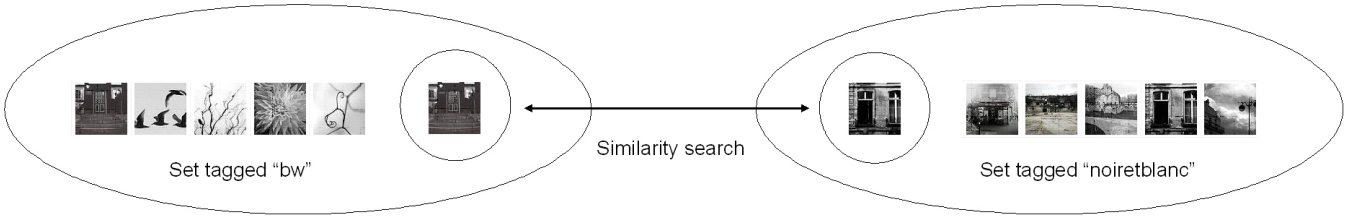

Figure 1: Relating tags in different languages through visual features

4.2.1 Image Selection

In case the user selects several images for his query (multi-image query), we think of these images as representing different classes. Thus, we accept images for retrieval that are similar to one of the query images. An alternative approach would be to average over the selected images which is, however, rarely relevant because the user might select visually distinct images. To give a simple example, a user selection of a yellow and a blue image should not yield green images as a result, but images that are either yellow or blue. Selection of the retrieved images is performed according to the following equation. Let X denote the archive and let x_l denote the l-th image in the archive. Let S denote a set of query images selected by the user. The distance D of x_l to S is then defined by

$$D(x_l, S) = \min_k d(x_l, s_k) \quad (1)$$

where d represents the distance of x_l to an image s_k contained in S, and k denotes the number of query images in S.

4.2.2 Refinement of Results

If the user is not entirely satisfied with the retrieved images, he has the possibility to refine the results. He can choose (1) one or more images as postive examples, or (2) one or more images as negative examples, or (3) combine (1) and (2). In case only positive examples are chosen, these are added to the initial query images and the query is started anew by evaluating Equation 1 and selecting the n closest images. If the user chooses to provide the system with one or more negative examples, the retrieval process becomes a classification problem. The set of all user-selected images can then be seen as prototypes labelled either "positive" or "negative".

It is important to note that the user might choose very different examples for the same label, i.e. he might choose for example, a red image with a very smooth texture, and a green image showing high contrast leaves both as positive examples. Therefore, a parametric classification method is not suited since it assumes the distribution of the underlying density function to be unimodal. In our case, it is a much better choice to employ a non-parametric approach that can be applied for arbitrary distributions and without the assumption that the forms of the underlying densities are known. Furthermore, it is important to ensure a smooth transition between retrieval and classification in order to avoid a drastic change of the results as soon as negative examples are selected.

A method that fulfills these requirements is a simple nearest neighbour classifier. Equation 1 basically defines the distance of an image in the database to a set of query images to be the distance between the test image and its nearest neighbour in the query set. For this reason, nearest neighbour classification is the natural choice to follow similarity retrieval. Let $P^n = \{x_1, \ldots, x_n\}$ denote a set of n labeled prototypes and let $x' \in P^n$ be the prototype nearest to a test point x. Then the nearest neighbour rule for classifying x is to assign it the label associated with x'.

4.2.3 Weight Updating

As mentioned in Section 4.1.3, all features initially have the same weight. This might, however, not correspond to the kind of search the user wants to perform. For example, the user might be looking for a certain colour while the texture of the image is not important to him, or he might be looking for images with predominantly straight lines, while the colour is not relevant. Such preferences can be detected and modelled by analysing the distances in the different feature spaces. We can do this at two levels: first for the global space, i.e. for "colour" or "texture" and then at a smaller scale for all features within a particular feature space f_i.

As a first step, we adapted only the global weights for colour and texture. Considering the set of positive examples, the closer the values in one feature space, the more relevant this feature and the more weight should be given to it. Based on this consideration, we estimate the mean μ of $d_i(x_j, S)$ of the normalised distances in feature space f_i. A good estimation of the weight w_i is thus

$$w_i = \frac{1}{\mu(d_i)}$$

The weights are then normalised according to

$$w_i' = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

In case positive and negative examples are given, the w_i are calculated for both sets separately and then combined by taking the average.

Currently, the weights are recalculated for every iteration. Future work on weight updating will involve taking into consideration temporal relations as well as experiments concerning weight updating within a particular feature space.

5. PRELIMINARY RESULTS

Figures 1 to 3 demonstrate examples where our approach has shown to be successful. Figure 1 illustrates how synonymy of tags, or in this case different languages, can be related by using visual features. The search started with the tag "bw" and a set of images was retrieved including those shown on the left side of Figure 1. Then, a similarity search was launched based on the image on the left side within the circle. The results of this query were visually

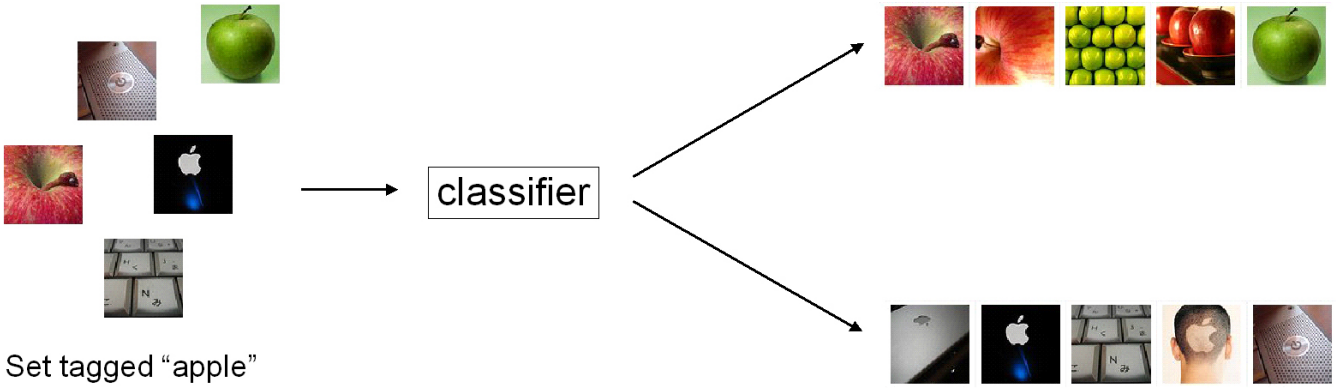

Figure 2: Separating different meanings of the same tag by using visual features

similar images, mostly black and white. Among these results was the image shown in the circle on the right side of Figure 1. Since this image is not tagged “bw” but “noiret-blanc” (french for “black and white”), we can access a new set of images that might not have been found otherwise.

An example showing how homonymy can be tackled by our approach is depicted in Figure 2. On the left side, a subset of the images tagged “apple” by the users is displayed. As we can see, “apple” refers to the fruit apple as well as to Apple computers. By giving examples of both classes, we can build a classifier that separates these two different meanings. It should be noted, that a direct search based on visual featues to retrieve apples (the fruits) would not have been successful because of the high intra-class variability. Only by restricting the context to the tag “apple” it was possible to achieve this good classification result.

Figure 3 shows three examples of visual similarity that has been found by our method as described above. We believe that such visual relations give appealing, sometimes surprising results that lead to interesting new associations and yield additional links between images.

6. CONCLUSIONS

The combination of collaborative tagging and data analysis can overcome problems of both approaches in navigating image archives. We showed how some of the problems inherent to tagging, such as synonymy or hononymy, can be tackled by additionally using visual features. On the other hand, we illustrated how tags can restrict the search space to support classification on visual features. The achieved results would not be possible by relying on visual features only. Furthermore, we showed how low-level similarity extracted by visual features can be used in a creative way that leads to interesting and surprising links between images.

It should be noted, however, that for the first example, the tag can be grounded, and in the second case, the distinction can be made because there is a visual difference between both classes (e.g. red/green vs. grey). In general, it is extremely difficult to develop grounded semantics for tags, particularly for social tagging sites where the tags are very personal, subjective, and inconsistent. However, even for tags that cannot be grounded, enriching the search by visual features gives unexpected, intuitive, and emotionally appealing choices which would never be accessible through tags alone.

7. FUTURE WORK

Future work includes further study of the user’s process of acquiring a collection. His tags, classifiers, and data elements reflect the way the user navigated through the archive. A further step will be to explore how this data can be used to assist the user in tagging. Other future developments are the application of the approach to other media types, especially to music, but also to video data.

8. ACKNOWLEDGEMENTS

This research was carried out and funded by the Sony Computer Science Laboratory in Paris with additional funding from the EU FET project ECAgents (IST-2003-1940) through the Sony Computer Science Laboratory in Paris.

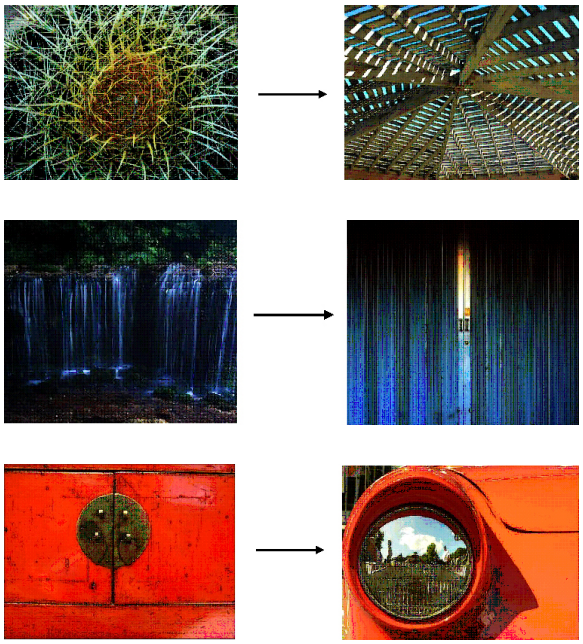

Figure 3: New relations by visual similarity

9. REFERENCES

[1] Flickr – photo sharing. `http://www.flickr.com`.
[2] Last.fm. `http://www.last.fm`.

[3] K. Aberer, P. Cudr-Mauroux, A. M. Ouksel, T. Catarci, M.-S. Hacid, A. Illarramendi, V. Kashyap, M. Mecella, E. Mena, E. J. Neuhold, O. D. Troyer, T. Risse, M. Scannapieco, F. Saltor, L. D. Santis, S. Spaccapietra, S. Staab, and R. Studer. Emergent semantics principles and issues. In *DASFAA*, pages 25–38, 2004.

[4] P. Alvarado, P. Doerfler, and J. Wickel. Axon2 – a visual object recognition system for non-rigid objects. In *Proceedings International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA)*, July 2001.

[5] C. Cattuto. Collaborative tagging as a complex system. Talk given at International School on Semiotic Dynamics, Language and Complexity, Erice, 2005.

[6] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content–based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[7] S. Santini, A. Gupta, and R. Jain. Emergent semantics through interaction in image databases. *IEEE Transactions on Knowledge and Data Engineering*, 13:337–351, 2001.

[8] L. Steels. Emergent semantics. *IEEE Intelligent Systems*, 17(1):83–85, 2002.

[9] L. Steels and P. Hanappe. Interoperability through emergent semantics. a semiotic dynamics approach. *Journal of Data Semantics*, 2006. To appear.

[10] L. Steels and F. Kaplan. Collective learning and semiotic dynamics. In D. Floreano, J.-D. Nicoud, and F. Mondada, editors, *Advances in Artificial Life: 5th European Conference (ECAL 99)*, Lecture Notes in Artificial Intelligence 1674, pages 679–688. Springer-Verlag, 1999.