Improved Image Retrieval Using Visual Sorting and Semi-Automatic Semantic Categorization of Images

Kai Uwe Barthel, Sebastian Richter, Anuj Goyal and Andreas Follmann

FHTW Berlin, Treskowallee 8, 10313 Berlin, Germany
barthel@fhtw-berlin.de

Abstract. The increasing use of digital images has led to the growing problem of how to organize these images efficiently for search and retrieval. Interpretation of what we see in images is hard to characterize, and even more so to teach a machine such that any automated organization can be possible. Due to this, both keyword-based Internet image search systems and content-based image retrieval systems are not capable of searching images according to the human high-level semantics of images. In this paper we propose a new image search system using keyword annotations, low-level visual metadata and semantic inter-image relationships. The semantic relationships are learned exclusively from the human users' interaction with the image search system. Our system can be used to search huge (web-based) image sets more efficiently. However, the most important advantage of the new system is that it can be used to generate semi-automatically semantic relationships between the images

Keywords: CBIR, Image sorting, SOMs, Relevance feedback, Image semantics

1 Introduction

Over the last decade the amount of digital images has increased tremendously. In order to make use of these images, efficient methods for archiving, organizing, searching and retrieving had to be developed.

Text-based Internet image search systems from Google, Yahoo! or Microsoft mostly use image file names or words from the context of the web page containing the image as keywords. Usually these keyword-based image searches will generate very large result sets, which typically are displayed on separate web pages containing arrangements of about 20 images. The quality and effectiveness of these keyword-based Internet image search systems is quite good if the goal is to find any images that correspond to the keyword. However if the goal is to find images with particular attributes then the performance is rather poor, as the search systems neither know the intention of the searching user nor the semantic relationships of the images that can be found on the Internet. This effect is amplified by homonyms, names similar to the query keyword and misclassified images. When performing a keyword based Internet image search, typically only a tiny fraction of the huge set of result images will be inspected, making the search for a particular image very time consuming or even impossible. A similar effect happens with image databases containing manually annotated images. If images are searched using only one keyword then the result set might be far too large to be inspected; on the other hand for a query with too many keywords, only very few or no images at all might be found.

Presented at "The First International Workshop on Metadata Mining for Image Understanding" (MMIU 2008), VISIGRAPP 2008, 22 - 25 January, 2008, Madeira, Portugal

In this paper we propose an image search system using keyword annotations, low-level visual metadata and semantic inter-image relationships. The semantic relationships are learned exclusively from the human users' interaction with the search system. The proposed system can be used to search huge (web-based) image sets more efficiently. Our system retrieves more images in an initial phase. We use CBIR techniques not to search but to sort these images according to their visual similarity. Using this visually sorted arrangement more images can be displayed simultaneously. Thus the user can identify very quickly images, which are good candidates for the desired search result. In the next step these images will serve as a visual filter for further result images. The filtering will refine the search result and generate more images that are similar to the desired query. Our proposed system dramatically cuts down the time for image retrieval.

However, the most important advantage of the new system is that it can be used to learn semi-automatically semantic relationships between images from the users interaction with the system. These relationships are language independent and can be used to further improve the quality and effectiveness of the image search.

The rest of this paper is organized as follows: Section 1 reviews the principle and current approaches of content-based image retrieval systems. Visual image sorting using self-organizing maps is described. Section 2 presents the proposed strategy and compares our scheme to other approaches. Section 3 describes implementation details and evaluates the new approach. We conclude the paper in Section 4.

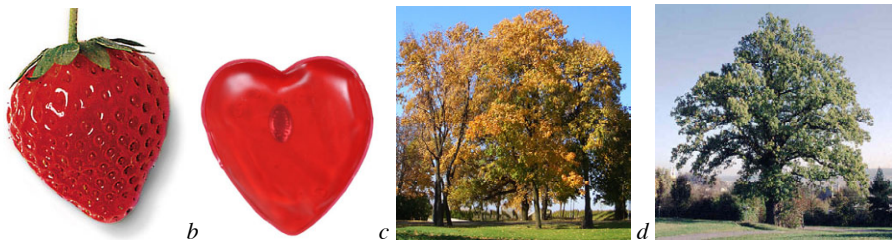
1.1 Content-Based Image Retrieval

In order to avoid manual annotation and to automate the process of image retrieval, *content-based image retrieval* (CBIR) techniques have been developed since the early 1990s. A good overview about the current state of the art of CBIR can be found in [2]. CBIR systems use automatically generated low-level metadata (*features*) to describe the visual statistics of images (like color, texture, and shape) [1, 8].

Low-level CBIR-systems are very well suited to find images that share visual features. These systems rely on the assumption that similar images do also have similar features. This assumption may be correct in many cases, however the opposite case is not necessarily true. Similar features could come from very different images that do not share any semantic similarities (see figure 1, images *a* and *b*).

Despite intense research efforts, the results of CBIR systems have not reached the performance of text based search engines. There are still several unsolved problems: The search for particular images is difficult if no query image is available.

Some approaches do use manually drawn sketches. However, the visual features of these sketched images can differ significantly from those of "real" images.



a *b* *c* *d*

Fig. 1: Problems of CBIR: Low-level CBIR would consider images *a* and *b* similar. Even sophisticated CBIR systems cannot determine the semantic similarity between images *c* and *d*.

Similarities between different kinds of features are measured using different appropriate metrics. It is not clear how these different metrics should be weighted if several features are combined for the search.

Recent approaches (SIFT) have used *interest points* describing significant local features of an image [7]. Interest points have proven to be very effective in finding images containing identical objects although the lightning conditions, the scale and the viewing positions can vary. However, even sophisticated CBIR systems using interest points cannot determine similarities between images that do have similar semantic content but do look different (see figure 1, images *c* and *d*).

The main problem of CBIR systems is the fact that there is an important (semantic) gap between the "content" that can be described with low-level visual features and the description of image content that humans use with high-level semantic concepts. Up until now algorithms cannot achieve high-level semantic understanding of images.

1.2 Image Sorting Using Self-Organizing Maps

In general it is problematic to find useful orderings for larger image sets. Most image management programs allow the sorting of images by size, date, or file format. Sorting images by similarity usually is not possible. Although it is not sure if the problems of current CBIR systems will be solved in the near future, the technique of automatic low-level feature extraction can be used to sort image sets. In the high dimensional feature vector space each image is represented by one vector. The locations of all vectors represent an ordered arrangement of the images. Due to the high number of dimensions, however, this order is unimaginable for human users.

Self-organizing maps (SOMs) are a data visualization technique, which reduce the dimensions of data through the use of self-organizing neural networks [5]. SOMs produce a map of usually one or two dimensions, which group similar data items together. The basic self-organizing map can be visualized as a neural-network array. The nodes of this array are trained and get specifically tuned to various input signal patterns. The learning process of the network is competitive and unsupervised.

The learning or training of a SOM consists of the following procedure: Initially the array of nodes is set up with random vector values. Next a sequence of training vectors is matched against all nodes m_i of the array. For each training vector x the winning (best matching) node m_c (in the sense of minimal distance) is determined.

$$c = \arg\min\left\{ \|x - m_i\| \right\} \tag{1}$$

The vector of the winning node and its neighborhood are updated such, that

$$m_i(t+1) = m_i(t) + h_{ci}(t)\left[x - m_i(t)\right] \tag{2}$$

where t are discrete time steps, h_{ci} is a neighborhood function that decays for increasing values of t and for nodes with larger distance from m_c. This means that the winning node and its neighborhood are adapted to the training vector. After successive training the nodes in the array become ordered as if some meaningful nonlinear coordinate system for the different input features were being created over the network.

SOMs have been used for CBIR [4]. Determining the closest neighbors in the SOM retrieved similar images. This technique was combined with a relevance feedback system. Deng et al. [3] proposed a SOM-based scheme to visualize and compare image collections. In our proposed scheme we use a self-organizing map to automatically sort images according to their visual similarity.

2 The Proposed Strategy

We propose a system that can be used to search extremely large Internet-based image sets more efficiently. In order to overcome the drawbacks of keyword-based image search systems *and* low-level feature-based CBIR systems our approach combines the ideas of high and low-level image retrieval systems. The new system significantly reduces the retrieval time and it can be used to learn (semantic) inter-image relationships from the users' interaction with the system.

Related Work

Previous approaches also proposed to use the combination of high-level semantic and low-level statistical metadata for image retrieval. However most of these systems were mainly focused on the automatic annotation of images.

Wenyin et al. introduced a scheme for semi-automatic image annotation [12]. They used a relevance feedback system that automatically added annotations for positive feedbacks.

Relevance feedback (RF) systems have been proposed to optimize the image retrieval systems. Usually RF is used to adapt the weights of the different features or to modify the search query. If the relevance feedback is based only on few result images, very often the CBIR systems tend to over-adapt to the particular features that were chosen for the feedback. Another problem is that the system does not know why the feedback was given. The system does not know which feature (color, shape or other) was the reason for the positive (or negative) feedback.

Pan et al. propose a graph-based automatic image captioning relying on the similarity of low-level metadata [10]. This approach works nicely for a database containing sets with many similar images taken in the same conditions but it will definitely have problems with different images from independent sources, as they can be found on the Internet. Wang et al. suggest allowing the user to choose from either a semantic or a visual query [13].

Lu et al. [6], and Zhou et al. [14] propose combining semantic keywords and low-level features. By using relevance feedback techniques they assign weighted links between the images and the keywords. However this scheme has the same problem as low-level CBIR systems. The fact that several images have a strong link to the same particular keyword does not mean that humans would consider these images similar. This is particular true for homonyms and names. We avoid this problem by not linking the images to the keywords but linking the images with each other.

Overview of the Proposed System

We will first give an overview of our proposed system. In the rest of the paper the used techniques will be explained in detail. Our system consists of several steps. Figure 2 shows a comparison of conventional Internet-based image retrieval and our proposed scheme. A conventional keyword-based Internet image search will produce a huge set of result images. However only very few of these images will be viewed.

Our system retrieves more images in an initial phase. We use CBIR techniques not to search but to sort these images according to their visual similarity by using a self-organizing map. Using this visually sorted arrangement more images can be displayed simultaneously compared to an unsorted set. Up to several hundred images can be inspected, which in most cases is sufficient to get a good representation of the entire result set. Thus the user can identify quickly those images, which are good candidates for his desired search result. These images will now be used to refine the result.

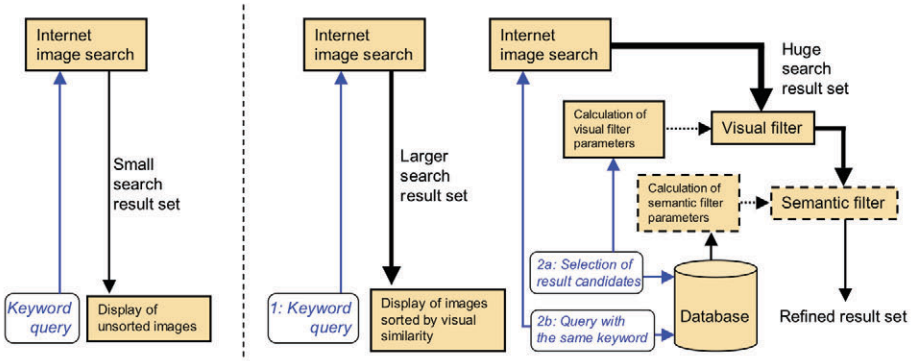


Fig. 2. Conventional Internet image search system (left) and the new proposed scheme (right). User interaction is printed in italics (blue).

Again we use CBIR techniques to filter out those result images that do not share any visual similarities with the candidate images. The filtering will generate more images that are similar to the desired query. This approach helps consistently to reduce the time required for the search for a particular image.

The most important advantage of the new system is that it can be used to learn semantic relationships between images. In order to refine the result the user will make a selection to mark candidate images. First of all the selection of these images can be seen as an affirmation that the keyword and these images actually do match. Even more important, however, is the relationship between the selected images. By making the selection the user expresses the fact, that according to his needs or view these images do share some common semantic meaning in some sense. Although the particular semantic information is unknown to the system, these relationships - when collected over many users - can be used for several purposes that will be described in the next section. Instead of learning the degree of confidence between an image and associated keywords like previous approaches, our new scheme does learn the semantic relationships between the images from the users' interaction with the system.

3 Implementation

Visual Sorting

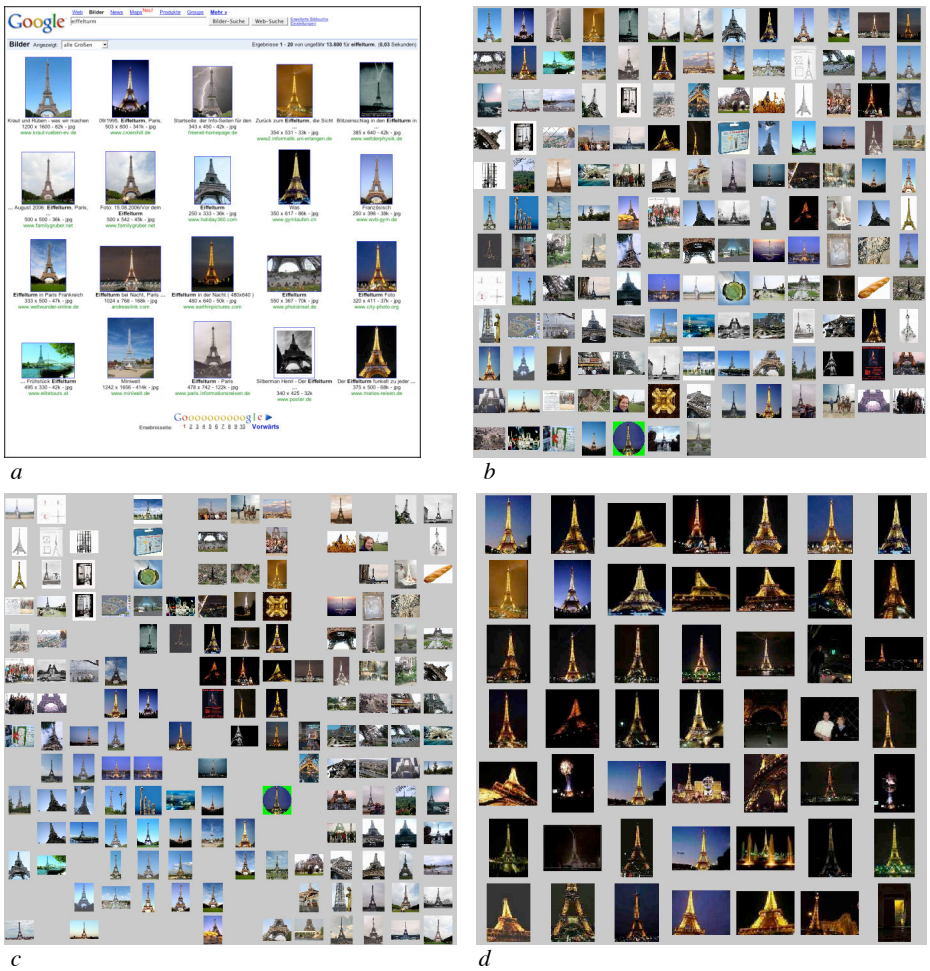
In order to generate visually sorted arrangements of the result images we examined different low-level features and SOM types. It turned out that only features describing color were able to generate sortings that people would consider visually pleasing and useful. In our prototype implementation we use feature vectors derived from the MPEG-7 color-layout descriptor, which can be seen as a heavily compressed thumbnail of 8x8 pixels. We managed to improve the sorting quality by giving the chrominance components a higher weight, by selecting only those DCT-transform coefficients with a strong variance and by omitting their quantization. Using these feature vectors, that can be calculated efficiently, good visual sorting results can be achieved even for very large image sets.

We investigated different SOM-types. In contrast to normal SOMs, the mapping rules had to be modified to ensure that no node position was occupied by more than one image, because otherwise these images would be overlapping when displayed. In our prototype we use a square network with a number of nodes that is larger than the

number of images to be sorted. Using a larger SOM size will result in a visually more pleasing sorting, however the display size for thumbnails will become smaller. We heuristically determined a 30% enlargement of the SOM as a compromise between a good visual sorting quality and a not too small thumbnail size. Best sorting results were achieved with a torus-shaped SOM by connecting the top nodes with the bottom nodes and the left nodes to the right nodes. To achieve a fast execution we used a Batch-SOM with incremental filters. Typically image sets can be sorted in 20 iterations. For 200 images the sorting time is less than 50 ms (Pentium 4, 3 GHz).

Visualization / Selection of desired images

Due to the sorting all images are arranged in such a way that visually similar images will be positioned close to each other. The sorted display makes it much easier to find a particular image within larger image sets.



a *b* *c* *d*

Fig. 3. Google image search using "Eiffel Tower" as keyword. *a*) 20 first result images as displayed in a web browser, *b*) the 150 first result images from the same query unsorted (in the order as they were delivered by Google), *c*) same images like *b* but sorted by similarity. From this set six candidate images (that were taken at night) were selected, *d*) the 49 most similar images filtered from a set of 1000 result images.

Figure 3*a* shows the 20 first result images obtained from Google searching for "Eiffel Tower" images. The first 150 result images are shown on the right in figure 3*b*. In this unsorted display it is difficult to find particular images. In comparison, the visually sorted map is shown below on the left (figure 3*c*).

In a next step the user has to select images that correspond to the desired search result. In order to be able to quickly find desired images the sorted map can be zoomed and panned. Although many more images are displayed compared to conventional image search systems, possible candidate images can more easily be found and marked due to the visually sorted display.

Filtering of keyword-based image retrieval results

When all candidate images have been selected, a new image search is initiated using the same query keyword. At this time more images will be retrieved from the image search system. The marked candidate images will serve to filter-out unwanted images. Again we tested different features for their ability to find images that are visually similar to the candidate images. As mentioned before, even sophisticated CBIR-schemes like SIFT will fail if images are semantical similar but do look different.

We found that a feature vector based on the *multimodal neighborhood signature* [9] is better suited for the visual filtering process. For each image we determine the 16 most representative pairs of neighboring colors. These feature vectors are matched using the *earth movers distance* (EMD) [11].

Let the index of the candidate images be denoted by k. For each new result image with index i the distances between its feature vector x_i and all feature vectors of the candidate images xc_k have to be determined. The minimum of these distances

$$d_i = \min_k \left\{ EMD\left(xc_k, x_i\right) \right\} \tag{3}$$

indicates how similar a new result image is compared with the set of candidate images. After all the distances of the newly retrieved images are determined, the set of the N best matching images will be displayed. Evidently the filtered result will become better if more images are retrieved and filtered. However in our prototype implementation we found that retrieving a number of images, which was five to ten times higher than the number of desired images gave very good results in most cases.

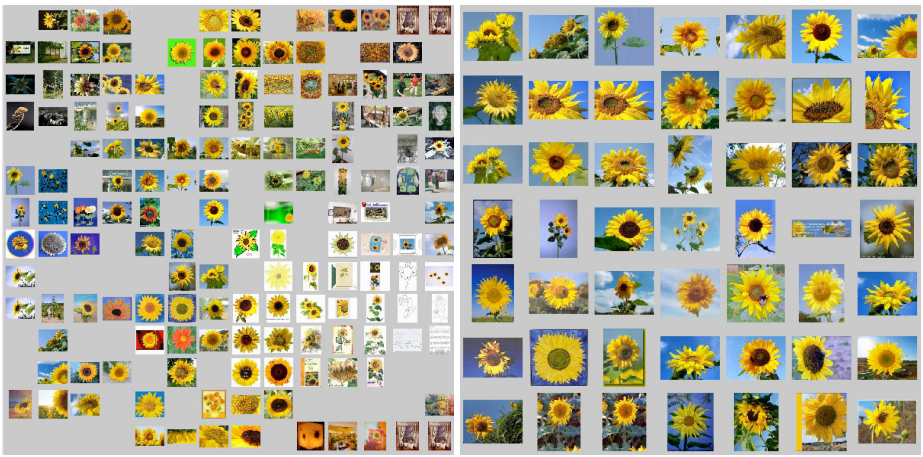


Fig. 4. Left: 150 "sunflower" images (visually sorted), five images with sunflowers on blue backgrounds were chosen as candidate images to filter 1000 "sunflower" result images. The 49 best matches are shown on the right.

An example result generated by our prototype implementation is shown in figure 3. A Google image search was initiated using the keyword "Eiffel Tower". From the first 150 visually sorted images (*c*) six candidate images that showed the Eiffel Tower at night were selected. These images were used to filter a set of 1000 images. The result of the 49 most similar images is shown in (*d*). Figure 4 shows another example where five candidate images (sunflowers with blue backgrounds) were used to filter a total of 1000 sunflower images.

Semi-automatic generation of semantic image relationships

An Internet based keyword image query will retrieve images that somehow are related to that keyword. Many of these images might be good results; however there will also be images from homonyms and proper names. In addition wrongly classified images that semantically do not match the keyword will be retrieved as well.

Previous systems used relevance feedback techniques to learn the degree of confidence between an image and associated keywords. However, this approach has several disadvantages: Firstly, users do not like to give feedback, therefore not all result images will be marked as positive or negative examples. Secondly, it is not clear for the search system why a particular feedback was given. Finally, as mentioned before, homonyms cannot be distinguished in this way.

With our proposed system these problems do not occur. If the number of displayed result images is large enough – which is possible due to the sorted display – then the chance that at least some images will match the desired query will be quite high. In this case – in order to refine the search result – the user will automatically give feedback by selecting matching candidate images. If no matching results were available, the user would rather ask for more images or try a new search with other keywords.

The selection of some images as candidate images can be seen as an affirmation for a match between the keyword and these images. Even more important, however, is the relationship between these selected images. By making the selection of these candidate images the user expresses the fact, that according to his desired search these images do share some common semantic meaning in some sense. Even though the particular semantic relationship is not known to the system, this selection can still be used to model the semantic inter-image similarities.

Our system does store the inter-image relationships by connecting all candidate images with weighted links. The weight represents how often two particular images have been selected together as part of a candidate set. Every time a user makes a new selection of two or more candidate images, the links of these candidate images will be updated. If a particular pair of images had already been selected together before, their link weight will be increased by 1. New images, that had not been selected together before will be linked with an initial weight of 1.

If enough relationships from different sets of candidate images are collected from many users, this collected information can be used to model the degree of semantic similarity of images. Obviously different users will group different sets of candidate image sets according to their needs. However our experiments indicate that this does not affect the ability of the system to learn the semantic relationships between images. Very similar images will have strong link weights, whereas dissimilar images will have very weak link weights or no links at all.

It must be mentioned, that the proposed system has to learn from many users' interactions before estimates of the inter-image similarities can be used to improve image retrieval results. If for a particular keyword search an image never gets selected as candidate, then it obviously should be excluded as result image for this keyword. The proposed scheme does automatically separate homonyms.

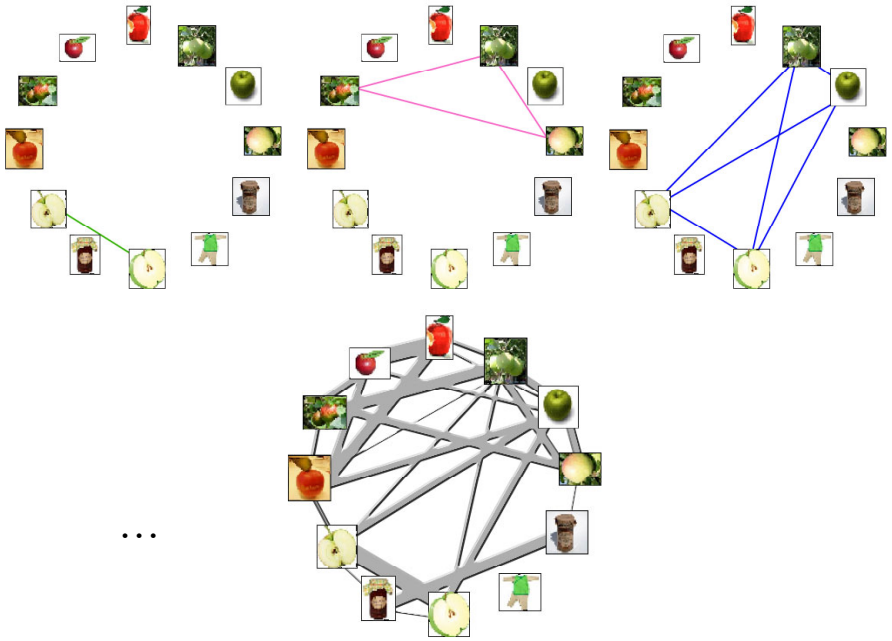


Fig. 5. Top: Three possible candidate sets (apple halves, apples on trees, green apples). The connecting lines indicate the relationships. These relationships can be collected over many users. Below: A map of relationships that our system could build from many different candidate sets. The degrees of similarities are expressed by the thicknesses of the links.

Figure 5 shows an example as it can be generated by our system by combining different candidate sets. Individual candidate sets (top) can be used to build up a network of weighted links of image relationships (below). The thicknesses of the links between the images represent the similarity of the images.

There are two ways in which the described learning of semantic image relationships can be used to improve the proposed image retrieval system: A semantic query by example can be achieved by simply retrieving those images with the strongest links to the query image. In addition the estimated similarities can also be used to perform an additional semantic filtering step as shown in figure 2. This filter will remove images that semantically do not match the candidate set. In another mode it could also be used to add further images that do not match the visual similarity but that do have a high estimated semantic similarity.

4 Conclusion

We have proposed a new image retrieval system that dramatically reduces the time needed to find a particular image in huge image sets as they can be found using Internet image search systems. The new approach combines both visual and semantic metadata. Unlike other approaches the new system does not try to learn the degree of confidence between an image and an associated keyword. We rather propose to model the degree of similarity between images by building up a network of linked images. The weights of the inter-image links are learned from the users' interaction with the system only. For each image search candidate images that are selected from an ini-

tially sorted result set help to refine the result by filtering out non-suiting images. Collecting this information from many users the semantic inter-image relationships of images can be modeled. We have implemented a prototype of an image search system as described before. The results obtained with the new scheme are very promising.

There remain many optimization possibilities for the proposed approach. Our system is using image search results from image retrieval systems like “Google image search”. Currently all images need to be downloaded first before they can be checked if they do conform to the required filtering criteria defined by the candidate set. The system could be much more efficient if it was implemented on the server side.

Another problem is the proper choice of the descriptors to extract the visual features. The proposed color layout descriptor works very nicely for visually sorting arbitrary image set. For the filtering of new result images against the set of candidate images however, further improvements are to be expected with the combination of other features types.

In future work we will provide experimental evaluation and comparison with other schemes, however none of the well-known reference image databases can be used for this purpose. We will set up a database with homonyms and misclassified images as they can be obtained from typical Internet search systems.

References

1. M. Bober: MPEG-7 visual shape descriptors. Special issue on MPEG-7, IEEE Transactions on Circuits and Systems for Video Technology 11/6, 716-719 (2001)
2. Ritendra Datta, Jia Li, James Ze Wang: Content-based image retrieval: approaches and trends of the new age. Multimedia Information Retrieval 2005: 253-262
3. D. Deng, J. Zhang and M. Purvis: Visualisation and comparison of image collections based on self-organised maps, ACSW Frontiers '04: Workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation (2004)
4. J. Laaksonen, M. Koskela, and E. Oja, PicSOM - Self-Organizing Image Retrieval With MPEG-7 Content Descriptors, IEEE Trans. on Neural Networks, Vol. 13, No. 4, (2002)
5. Kohonen, T., Self-Organizating Maps, New York: Springer (1997)
6. Y. L. Lu, C. Hu, X. Zhu, H. J. Zhang, and Q. Yang, “A unified framework for semantics and feature based relevance feedback in image retrieval systems,” in Proceedings of 8th ACM International Conference on Multimedia (MM '00), pp. 31–37
7. David G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.
8. Manjunath B. S. , Ohm J. R. , Vasudevan V. V. , Yamada A. “Color and texture descriptors. Special issue on MPEG-7”, IEEE Transactions on Circuits and Systems for Video Technology, 11/6, 703-715 (2001)
9. G. Matas, D. Koubaroulis and J. Kittler, Colour Image Retrieval and Object Recognition Using the Multimodal Neighbourhood Signature, 6th European Conference on Computer Vision, Dublin, Ireland, June 2000
10. Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, Pinar Duygulu: GCap: Graph-based Automatic Image Captioning (2004)
11. Rubner, Y., Tomasi, C., and Guibas, L. J. 1998 The Earth Mover's Distance as a Metric for Image Retrieval. Technical Report. UMI Order Number: CS-TN-98-86, Stanford Univ.
12. Liu Wenyin, Susan Dumais, Yanfeng Sun, HongJiang Zhang, Mary Czerwinski, Brent Field: Semi-Automatic Image Annotation (2001)
13. Wei Wang, Yimin Wu, Aidong Zhang: *SemView*: A Semantic-sensitive Distributed Image Retrieval System, National Conference on Digital Government Research (2003)
14. X. S. Zhou, T. S. Huang, “Unifying Keywords and Visual Contents in Image Retrieval”, IEEE Multimedia, April-June Issue, 2002.