published in the Proceedings of VISual'99, Amsterdam, The Netherlands, June 1999
available at <URL:http://www.cis.hut.fi/picsom/publications.html>

Content-Based Image Retrieval using Self-Organizing Maps

Jorma Laaksonen, Markus Koskela, and Erkki Oja

Laboratory of Computer and Information Science,
Helsinki University of Technology,
P.O.BOX 5400, Fin-02015 HUT, Finland
{jorma.laaksonen, markus.koskela, erkki.oja}@hut.fi

Abstract We have developed an image retrieval system named PicSOM which uses Tree Structured Self-Organizing Maps (TS-SOMs) as the method for retrieving images similar to a given set of reference images. A novel technique introduced in the PicSOM system facilitates automatic combination of the responses from multiple TS-SOMs and their hierarchical levels. This mechanism aims at adapting to the user's preferences in selecting which images resemble each other in the particular sense the user is interested of. The image queries are performed through the World Wide Web and the queries are iteratively refined as the system exposes more images to the user.

1 Introduction

Content-based image retrieval from unannotated image databases has been an object for ongoing research for a long period. Many projects have been started in recent years to research and develop efficient systems for content-based image retrieval. The best-known implementation is probably Query By Image Content (QBIC) [3] developed at the IBM Almaden Research Center. Other notable systems include MIT's Photobook [9] and its more recent version, FourEyes, the search engine family of WebSEEk, VisualSEEk, and MetaSEEk [2], which all are developed at Columbia University, and Virage [1], a commercial content-based search engine developed at Virage Technologies Inc.

We have implemented an image-retrieval system called PicSOM, which tries to adapt to the user's preferences regarding the similarity of images using Self-Organizing Maps (SOMs) [5]. The approach is based on the relevance feedback technique [11], in which the human-computer interaction is used to refine subsequent queries to better approximate the need of the user. Some earlier systems have also applied the relevance feedback approach in image retrieval [8, 10].

PicSOM uses a SOM variant called Tree Structured Self-Organizing Map (TS-SOM) [6, 7] as the image similarity scoring method and a standard World Wide Web browser as the user interface. The implementation of our image-retrieval system is based on a general framework in which the interfaces of co-operating modules are defined. Therefore, the TS-SOM is only one possible choice for the similarity measure. However, the results we have gained so

far, are very promising on the potentials of the TS-SOM method. As far as the current authors are aware, there has not been notable image retrieval applications based on the SOM. Some preliminary experiments with the SOM have been made previously in [13].

2 Principle of PicSOM

Our method is named PicSOM, which bears similarity to the well-known WEBSOM [4, 12] document browsing and exploration tool that can be used in free-text mining. WEBSOM is a means for organizing miscellaneous text documents into meaningful maps for exploration and search. It is based on the SOM which automatically organizes documents into a two-dimensional grid so that related documents appear close to each other. Up to now, databases over one million documents have been organized for search using the WEBSOM system. In an analogous manner, we have aimed at developing a tool that utilizes the strong self-organizing power of the SOM in unsupervised statistical data analysis for image retrieval. The features may be chosen separately for each specific task and the system may also use keyword-type textual information for the images.

The basic operation of the PicSOM image retrieval is as follows: 1) An interested user connects to the WWW server providing the search engine with her web browser. 2) The system presents a list of databases available to that particular user. 3) After the user has selected the database, the system presents an initial set of tentative images scaled to small thumbnail size. The user then selects the subset of images which best match her expectations and to some degree of relevance fit to her purposes. Then, she hits the "Continue Query" button in her browser which sends the information on the selected images back to the search engine. 4) The system marks the selected and non-selected images with positive and negative values, respectively, in its internal data structure. Based on this information, the system then presents a new set of images aside with the images selected this far. 5) The user again selects the relevant images, submits this information to the system and the iteration continues. Hopefully, the fraction of relevant images increases as more images are presented to the user and, finally, one of them is exactly what she was originally looking for.

2.1 Feature Extraction

PicSOM may use one or several types of statistical features for image querying. Separate feature vectors can thus be formed for describing the color, texture, and structure of the images. A separate Tree Structured Self-Organizing Map is then constructed for each feature vector set and these maps are used in parallel to select the best-scoring images. New features can be easily added to the system, as long as the features are calculated from each picture in the database.

In our current implementation, the average R-, G-, and B-values are calculated in five separate regions of the image. This division of the image area increases the discriminating power by providing a simple color layout scheme.

The resulting 15-dimensional color feature vector thus not only describes the average color of the image but also gives information on the color composition. The current texture feature vectors in PicSOM are calculated similarly in the same five regions as the color features. The Y-values of the YIQ color representation of every pixel's 8-neighborhood are examined and the estimated probabilities for each neighbor pixel being brighter than the center pixel are used as features. This results in five eight-dimensional vectors which are combined to one 40-dimensional textural feature vector.

2.2 Self-Organizing Map (SOM)

The Self-Organizing Map (SOM) [5] is a neural algorithm widely-used to visualize and interpret large high-dimensional data sets. The map consists of a regular grid of neurons. A vector m_i, consisting of features, is associated with each unit i. The map attempts to represent all the available observations with optimal accuracy using a restricted set of models. At the same time, the models become ordered on the grid so that similar models are close to each other and dissimilar models far from each other.

Fitting of the model vectors is usually carried out by a sequential regression process, where $t = 1, 2, \ldots$ is the step index: For each sample $x(t)$, first the index $c = c(x)$ of the best-matching unit is identified by the condition

$$\forall i: \quad \|x(t) - m_c(t)\| \leq \|x(t) - m_i(t)\| \; . \tag{1}$$

After that, all model vectors or a subset of them that belong to nodes centered around node $c(x)$ are updated as

$$m_i(t+1) = m_i(t) + h(t)_{c(x),i}(x(t) - m_i(t)) \; . \tag{2}$$

Here $h(t)_{c(x),i}$ is the "neighborhood function", a decreasing function of the distance between the ith and cth nodes on the map grid. This regression is then reiterated over the available samples and the value of $h(t)_{c,i}$ is let to decrease in time to guarantee the convergence of the unit vectors m_i.

2.3 Tree Structured SOM (TS-SOM)

The Tree Structured Self-Organizing Map (TS-SOM) [6, 7] is a tree-structured vector quantization algorithm that uses SOMs at each of its hierarchical levels. In PicSOM, all TS-SOM maps are two-dimensional. The number of map units increases when moving downwards in the TS-SOM. The search space for the best-matching vector of equation (1) on the underlying SOM layer is restricted to a predefined portion just below the best-matching unit on the above SOM. Therefore, the complexity of the searches in TS-SOM is remarkably lower than if the whole bottommost SOM level were accessed without the tree structure.

The computational lightness of TS-SOM facilitates the creation and use of huge SOMs whichare used to hold the images stored in the image database. The

feature vectors calculated from the images are used to train the levels of the TS-SOMs beginning from the top level. After the training phase, each unit of the TS-SOMs contains a model vector which may be regarded as the average of all feature vectors mapped to that unit. In PicSOM, we then search in the corresponding data set for the feature vector which best matches the stored model vector and associate the corresponding image to that map unit. Consequently, a tree-structured hierarchical representation of all the images in the database is formed. In an ideal situation, there should be one-to-one correspondence between the images and TS-SOM units in the bottom level of each map.

2.4 Using Multiple TS-SOMs

Combining the results from several maps can be done in a number of ways. A simple method would be to ask the user to enter weights for different maps and then calculate a weighted average. This, however, requires the user to give information which she normally does not have. Generally, it is a difficult task to give low-level features such weights which would coincide with human's perception of images at a more conceptual level. Therefore, a better solution is to use the relevance feedback approach. Then, the results of multiple maps are combined automatically, using the implicit information from the user's responses during the query. The PicSOM system thus tries to learn the user's preferences from the interaction with her and to set its own responses accordingly.

The rationale behind our approach is as follows: If the images selected by the user map close to each other on a certain TS-SOM map, it seems that the corresponding feature performs well on the present query and the relative weight of its opinion should be increased. This can be implemented by marking on the maps the images the user has seen. The units are given positive and negative values depending whether she has selected or rejected the corresponding images. The mutual relations of positively-marked units residing near each other can then be enhanced by convolving the maps with a simple low-pass filtering mask. As a result, areas with many positively marked images spread the positive response to their neighboring map units. The images associated with these units are then good candidates for next images to be shown to the user, if they have not been shown already. The current PicSOM implementation uses convolution masks whose values decrease as the 4-neighbor or "city-block" distance from the mask center increases. The convolution mask size increases as the size of the corresponding SOM layer increases.

Figure 1 shows a set of convolved feature maps during a query. The three images on the left represent three map levels on the Tree Structured SOM for the RGB color feature, whereas the convolutions on the right are calculated on the texture map. The sizes of the SOM layers are 4×4, 16×16, and 64×64, from top to bottom. The dark regions have positive and the light regions negative convolved values on the maps. Notice the dark regions in the lower-left corners of the three layers of the left TS-SOM. They indicate that there is a strong response and similarity between images selected by the user in that particular area of the color feature space.

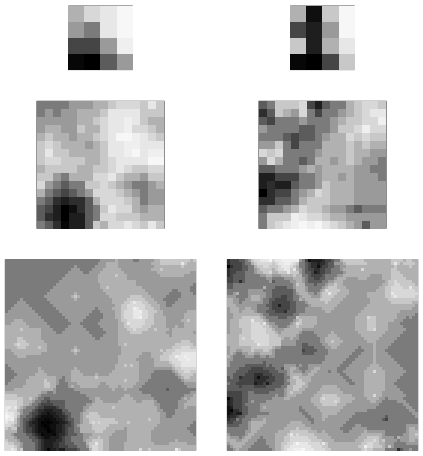

Figure1. An example of convolved TS-SOMs for color (left) and texture (right) features. Black corresponds to positive and white to negative convolved values.

2.5 Refining Queries

Initially, the query begins with a set of reference images picked from the top levels of the TS-SOMs in use. For each reference image, the best-matching SOM unit is searched on every layer of all the TS-SOM maps. The selected and rejected images result to positive and negative values on the best-matching units. The positive and negative responses are normalized so that their sum equals to zero. Previously positive map units can also be changed to negative as the retrieval process iteration continues. In early stages of the image query, the system tends to present the user images from the upper TS-SOM levels. As soon as the convolutions begin to produce large positive values also on the lower map levels, the images on these levels are shown to the user. The images are therefore gradually picked more and more from the lower map levels as the query is continued.

The inherent property of PicSOM to use more than one reference image as the input information for retrievals is important. This feature makes PicSOM differ from other content-based image retrieval systems, such as QBIC, which use only one reference image at a time.

3 Implementation of PicSOM

The issues of the implementation of the PicSOM image retrieval system can be divided in two categories. First, concerning the user interface, we have wanted to make our search engine, at least in principle, available and freely usable to the public by implementing it in the World Wide Web. The use of standard web browser also makes the queries on the databases machine independent. Figure 2 shows a screenshot of the current web-based PicSOM user interface, which can be found at *http://www.cis.hut.fi/picsom/*.

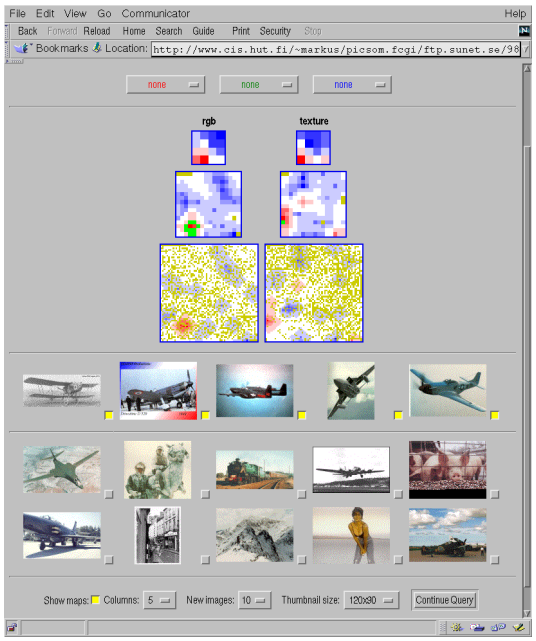

Figure2. WWW-based user interface of PicSOM. The user has already selected five aircraft images in the previous rounds. The system is displaying the user ten new images to select of.

Second, the functional components in the server running the search engine have been implemented so that the parts responsible for separate tasks have been isolated to separate processes. The implementation of PicSOM has three separate modular components: 1) *picsom.cgi* is a CGI/FCGI script which handles the requests and responses from the user's web browser. This includes processing the HTML form, updating the information from previous queries and executing the other components as needed to complete the requests. 2) *picsomctrl* is the main program responsible for updating the TS-SOM maps with new positive and negative response values, calculating the convolutions, creating new map images for the next web page, and selecting the best-scoring images to be shown to the user in the next round. 3) *picsomctrltohtml* creates the HTML contents of the new web pages based on the output generated by the *picsomctrl* program.

4 Preliminary Quantitative Results

Quantitative measures of the image retrieval performance of a system, or any single feature, are problematic due to human subjectivity. Generally, there exists no definite right answer to an image query as each user has individual expectations.

We have made experiments with an image database of 4350 images. Most of them are color photographs in JPEG format. The images were downloaded from

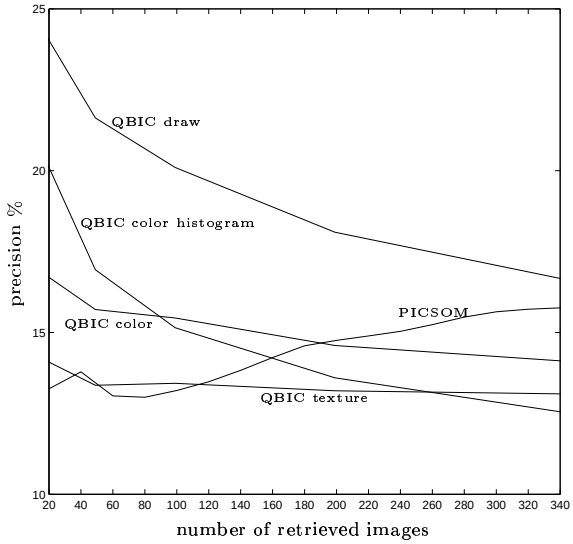

Figure3. Average precisions of PicSOM and QBIC responses when one reference image containing an aircraft is presented to the system. The *a priori* probability for correct response is 8 percent.

the image collection residing at the Swedish University Network FTP server, located at *ftp://ftp.sunet.se/pub/pictures/*.

We evaluated PicSOM's and QBIC's responses when one image containing an aircraft was given at a time as a reference image to the systems. The average number of relevant images was then calculated for both methods by using a hand-picked subset of the *ftp.sunet.se* collection which contained 348 aircraft images. This gave rise to *a priori* probability of 8.0 percent correct responses. Figure 3 shows the average retrieval precisions as functions of the number of images returned from the database. The response of PicSOM was produced by returning 20 best-scoring images in each iteration step and selecting the relevant images among them. For QBIC, separate responses for four features are shown. As could be expected, QBIC's average performance decreases when the number of returned pictures is increased. This is due to the fact that QBIC's response is always based on a single reference image. On the contrary, PicSOM's operation is based on an increasing set of images and the response becomes more accurate as more images are shown. PicSOM's strength is thus in its ability to use multiple reference images and all available features simultaneously.

5 Future Plans

The next obvious step to increase PicSOM's retrieval performance is to add better feature representations to replace our current experimental ones. These

will include color histograms, color layout descriptions, shape features, and some more sophisticated texture models. As the PicSOM architecture is designed to be modular and expandable, adding new statistical features is straightforward. We also need to define more quantitative measures which can be used in comparing the performance of the PicSOM system with that of other content-based image retrieval systems. To study our method's applicability on a larger scale we shall need larger image databases. A vast collection of images is available on the Internet, and we have preliminary plans to use PicSOM as an image search engine for the World Wide Web.

References

1. Bach J. R., Fuller C., Gupta A., et al. The Virage image search engine: An open framework for image management. In Sethi I. K. and Jain R. J., editors, *Storage and Retrieval for Image and Video Databases IV*, volume 2670 of *Proceedings of SPIE*, pages 76–87, 1996.
2. Chang S.-F., Smith J. R., Beigi M., and Benitez A. Visual information retrieval from large distributed online repositories. *Communications of the ACM*, 40(12):63–69, December 1997.
3. Flickner M., Sawhney H., Niblack W., et al. Query by image and video content: The QBIC system. *IEEE Computer*, pages 23–31, September 1995.
4. Honkela T., Kaski S., Lagus K., and Kohonen T. WEBSOM—self-organizing maps of document collections. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6*, pages 310–315. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.
5. Kohonen T. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer-Verlag, 1997. Second Extended Edition.
6. Koikkalainen P. Progress with the tree-structured self-organizing map. In Cohn A. G., editor, *11th European Conf. on Artificial Intelligence*. European Committee for Artificial Intelligence (ECCAI), John Wiley & Sons, Ltd., August 1994.
7. Koikkalainen P. and Oja E. Self-organizing hierarchical feature maps. In *Proceedings of 1990 International Joint Conference on Neural Networks*, volume II, pages 279–284, San Diego, CA, 1990. IEEE, INNS.
8. Minka T. P. An image database browser that learns from user interaction. Master's thesis, M.I.T, Cambridge, MA, 1996.
9. Pentland A., Picard R. W., and Sclaroff S. Photobook: Tools for content-based manipulation of image databases. In *Storage and Retrieval for Image and Video Databases II (SPIE)*, volume 2185 of *SPIE Proceedings Series*, San Jose, CA, USA, 1994.
10. Rui Y., Huang T. S., and Mehrotra S. Content-based image retrieval with relevance feedback in MARS. In *Proc. of IEEE Int. Conf. on Image Processing '97*, pages 815–818, Santa Barbara, California, USA, October 1997.
11. Salton G. and McGill M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
12. WEBSOM - self-organizing maps for internet exploration, `http://websom.hut.fi/websom/`.
13. Zhang H. and Zhong D. A scheme for visual feature based image indexing. In *Storage and Retrieval for Image and Video Databases III (SPIE)*, volume 2420 of *SPIE Proceedings Series*, San Jose, CA, February 1995.