published in the Proceedings of ICANN'99, Edinburgh, UK, September 1999
available at <URL:http://www.cis.hut.fi/picsom/publications.html>

Application of Tree Structured Self-Organizing Maps in Content-Based Image Retrieval

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
email: {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) [4] 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 which we have given the name PicSOM. It uses the Tree Structured Self-Organizing Map (TS-SOM) [7, 8] as the image similarity scoring method and a World Wide Web browser as the user interface. While PicSOM processes the image selections made by the user, it tries to adapt to the user's preferences regarding the similarity of images. This can be seen as a version of the relevance feedback technique [10].

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 use of TS-SOMs is only one choice for the similarity measure. However, the results we have gained so far, are very promising on the potentials of the TS-SOM method. Some preliminary experiments with the Self-Organizing Map (SOM) [6] in image retrieval have been made in [12]. But, as far as the current authors are aware, there has not been until now notable practical image-retrieval applications based on SOM.

2 Principle of PicSOM

Our method is named PicSOM, which bears similarity to the well-known WEBSOM [5, 11] 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 SOM [6] that automatically organizes documents into a two-dimensional grid so that related documents appear close to each other. Up to now, databases of 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 images. The features may be chosen separately for each specific task and the system may also use keyword-type textual information for the images, if available.

From the user's point of view, 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. Later, there will also be a list of available search strategies, currently only the TS-SOM-based engine has been implemented. 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 images selected by the user with a positive value and the non-selected images with a negative value in its internal data structure. Based on this information, the system then presents the user 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. The user can also deselect images which have earlier been selected. 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. Different kinds of feature vectors can thus be formed for describing the color, texture, shape, 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 added to the system, as long as an equal number of features are calculated from each picture in the database.

In our first experiments we have used simple color and texture features. 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 texture feature vectors in PicSOM are calculated similarly in 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 Tree Structured SOM

The Tree Structured Self-Organizing Map (TS-SOM) [8] is a tree-structured vector quantization algorithm that uses SOMs [6] 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 on the underlying SOM layer in 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 would be accessed without the tree structure. The structure of TS-SOM is illustrated in Figure 1.

The computational lightness of TS-SOM facilitates the creation and use of huge SOMs which, in our PicSOM system, are 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. During the training, each feature vector is presented to the map multiple times and the model vectors stored in the map units are modified to match the distribution and topological ordering of the feature vector space. After the training phase, each unit of the TS-SOMs contains a model vector which may be regarded as the aver-

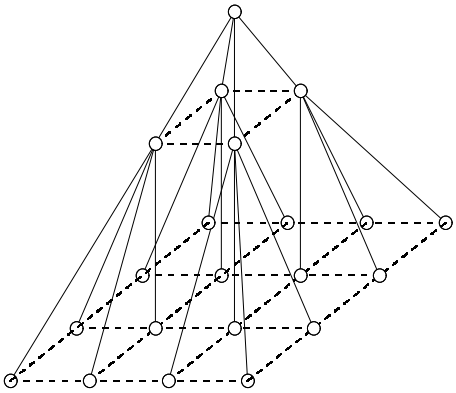

Figure 1: The structure of a three-layer two-dimensional TS-SOM.

age of all feature vectors mapped into that particular 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.

2.3 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 combine the results of multiple maps 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 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 positive and negative responses are normalized so that their sum equals to zero. The mutual relations of positively-marked units residing near to each other can then be enhanced by convolving the maps with a simple low-pass filtering mask. As a result, areas where positively marked images are mapped close to each other 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.

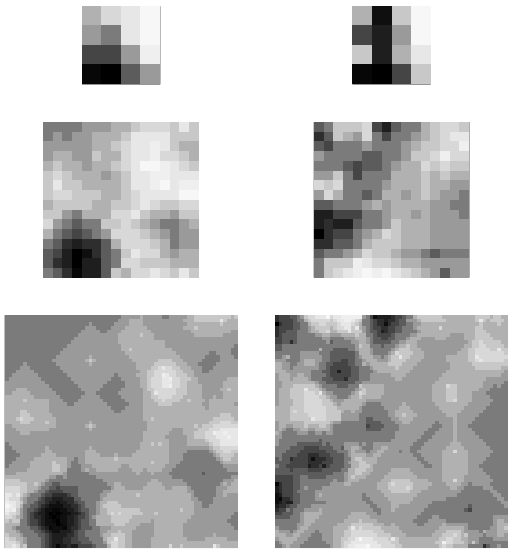

Figure 2: An example of convolved TS-SOMs for color (left) and texture (right) features. Black corresponds to positive and white to negative convolved values.

Figure 2 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

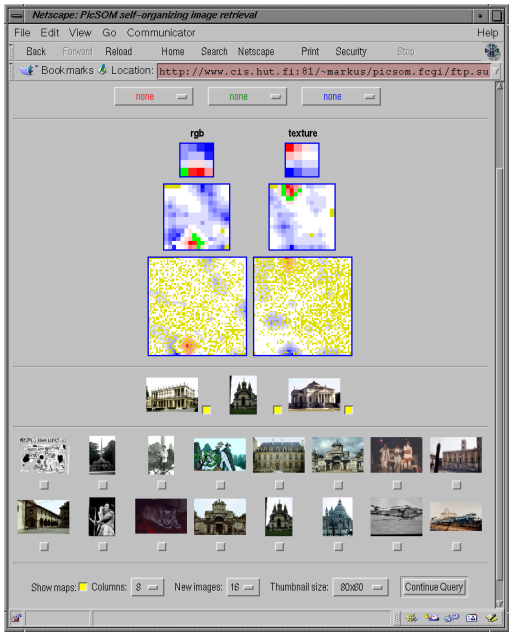

Figure 3: The WWW-based user interface of PicSOM.

user in that particular area of the color feature space.

Initially, the queries begin with a set of reference images picked from the top levels of the TS-SOMs. 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 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.

3 Implementation

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. The WWW-based user interface is illustrated in Figure 3. Below the convolved SOMs, the first set of images consists of images selected on the previous rounds of the retrieval process. These images may be unselected on any subsequent round, thus changing their value from positive to negative. This example shows a query with three images of buildings selected. The next images, separated by a horizontal line, are the 16 best-scoring new images obtained from the convolved units in the TS-SOMs. The user can at any time switch from the iterative queries to examining of the TS-SOM surfaces simply by clicking the map images. Relevant images on the maps can then also be selected for continuing queries.

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 functional interfaces between these processes have then been designed to be open and easily extensible to allow future extensions to the system.

The implementation of PicSOM has three separate modular components: 1) *picsom.cgi* is a CGI/FCGI script written in Perl and handles the requests and responses from the user's web browser. This includes processing the HTML form, updating the information from previous rounds 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 on the next round. It is implemented in C++. 3) *picsomctrltohtml* is again a Perl script. It creates the HTML contents of the new web pages based on the output from the *picsomctrl* program.

4 Experiments

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. Therefore, objective performance comparisons between different feature types and with other retrieval systems are difficult.

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

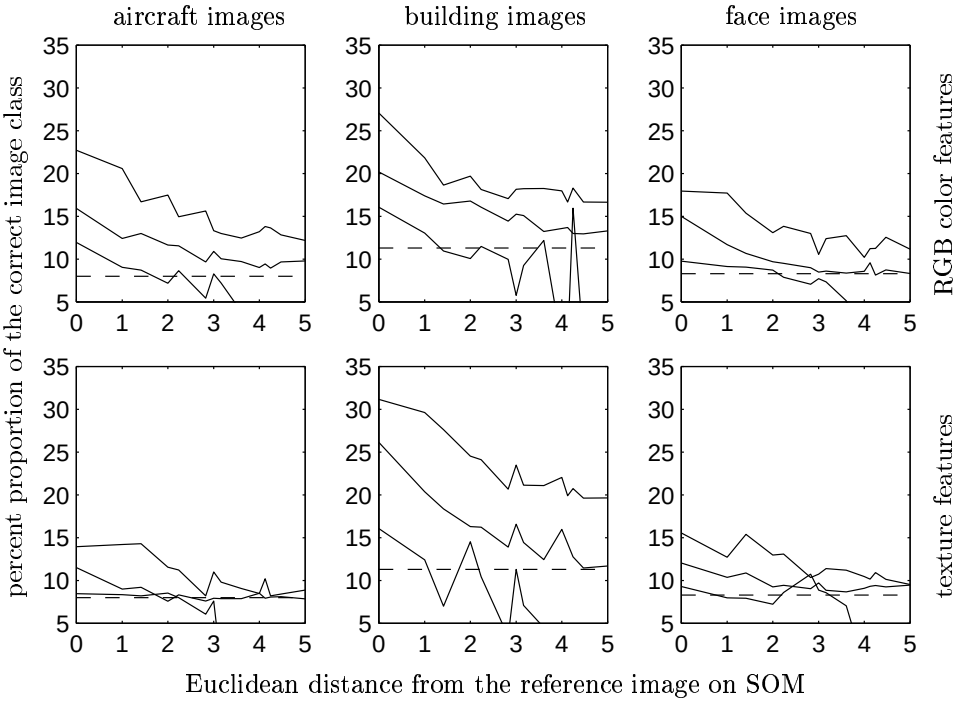

Figure 4: The average proportions of the correct image classes as the function the Euclidean distance from the SOM unit into which a reference image was mapped. Color features were used in the upper row and textural features in the lower. In each plot, the curves present the results for the 64×64, 16×16, and 4×4 maps from top to bottom. The dashed lines display the *a priori* probability.

an image collection residing at the Swedish University Network FTP server, located at *ftp://ftp.sunet.se/pub/pictures/*. Another image source we have started to experiment with is the Corel Gallery [3] with 60.000 photographic images.

PicSOM also supports the utilization of textual class information for the images, if that kind of information is available in the database. For example, the original directory structure of the *ftp.sunet.se* collection has been used to give the images rough textual content classes which we have used as the ground truth for our quantitative assertions.

For the experiments we hand-picked three image subsets, containing aircraft, building, and human face images. Each of the images in these sets was then mapped as a reference image onto the three SOM layers of the RGB feature TS-SOM and the three layers of the texture TS-SOM. The average number of images which belonged to the same image set was then calculated as a function of the Euclidean distance from the reference image on the map. Both TS-SOMs were sized 4×4, 16×16, and 64×64, from top to bottom. Figure 4 shows the retrieval precisions curves for all the three images classes and every SOM layer of the two TS-SOMs. The dashed lines represent the *a priori* retrieval precisions which were 8.0, 11.3, and 8.3 percent for the aircraft, building, and face sets, respectively. It can be seen how the average proportions of relevant images are approximately same for higher and lower level maps when the Euclidean distance in the latter is four-fold. This is a straight consequence of the TS-SOM architecture in which the map dimensions are quadrupled in every downward transition.

These results show that even with a single TS-SOM map and one feature type the Self-Organizing Map algorithm is able to group similar images near to each other. The resulting precision of content-based image retrieval for the different image sets and the two feature types varies. For example, it can be seen that the building set is easier than the other two. Also, this set is bet-

ter retrieved by the texture features whereas the color features perform better for the two other sets. In all the cases, the results for the 64×64-sized maps are well above the *a priori* probabilities.

5 Conclusions

In this paper we have introduced a novel content-based image retrieval system. An important inherent property of the PicSOM system is to use more than one reference image as the input information for retrievals. This feature makes PicSOM different from most other systems, such as QBIC, which use only one reference image at time. Another important characteristic of PicSOM is its ability to use multiple features simultaneously without the need that the user or administrator of the system should manually set their mutual weights.

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 are currently performing massive comparisons between different feature types.

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. In our plans, the system will later be publicly available for demonstration purposes at *http://www.cis.hut.fi/picsom/*.

References

[1] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu. The Virage image search engine: An open framework for image management. In I. K. Sethi and R. J. Jain, editors, *Storage and Retrieval for Image and Video Databases IV*, volume 2670 of *Proceedings of SPIE*, pages 76–87, 1996.

[2] S.-F. Chang, J. R. Smith, M. Beigi, and A. Benitez. Visual information retrieval from large distributed online repositories. *Communications of the ACM*, 40(12):63–69, December 1997.

[3] The Corel Corporation World Wide Web home page, `http://www.corel.com`.

[4] M. Flickner, H. Sawhney, W. Niblack, et al. Query by image and video content: The QBIC system. *IEEE Computer*, pages 23–31, September 1995.

[5] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. 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.

[6] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer-Verlag, 1997. Second Extended Edition.

[7] P. Koikkalainen and E. Oja. Self-organizing hierarchical feature maps. In *Proc. IJCNN-90, Int. Joint Conf. on Neural Networks, Washington, DC*, volume II, pages 279–285, Piscataway, NJ, 1990. IEEE Service Center.

[8] P. Koikkalainen. Progress with the tree-structured self-organizing map. In A. G. Cohn, editor, *11th European Conference on Artificial Intelligence*. European Committee for Artificial Intelligence (ECCAI), John Wiley & Sons, Ltd., 1994.

[9] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. In *Storage and Retrieval for Image and Video Databases II*, volume 2185 of *SPIE Proceedings Series*, San Jose, CA, USA, 1994.

[10] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[11] WEBSOM - self-organizing maps for internet exploration, `http://websom.hut.fi/websom/`.

[12] H. Zhang and D. Zhong. 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.