published in the Proceedings of SCIA'99, Kangerlussuaq, Greenland, June 1999
available at <URL:http://www.cis.hut.fi/picsom/publications.html>

PicSOM – A Framework for Content-Based Image Database Retrieval using Self-Organizing Maps

J.T. Laaksonen, J.M. Koskela, and E. Oja
Laboratory of Computer and Information Science,
Helsinki University of Technology,
P.O.BOX 5400
Fin-02015 HUT, Finland

Abstract

We have developed an image retrieval system which uses Tree Structured Self-Organizing Maps (TS-SOMs) as the method for retrieving images similar to a given set of reference images in a database. It also provides a framework for the research on algorithms and methods for content-based retrieval of images. A novel technique introduced in this paper facilitates automatic combination of the responses from multiple TS-SOMs and their hierarchical levels. The system tries to adapt to the user's preferences in selecting which images resemble each other in the particular sense the user is interested of. This mechanism implements a relevance feedback technique on content-based image retrieval. 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 [12]. Digital image and video libraries are becoming more widely used as more visual information is produced at a rapidly growing rate. The technologies needed for retrieving and browsing this growing amount of information are still, however, quite immature and limited.

Many projects have been started in recent years to research and develop efficient systems for content-based image retrieval. The best-known system is Query By Image Content (QBIC) [4] developed at the IBM Almaden Research Center. Other notable systems include MIT's Photobook [11] and its more recent version, FourEyes [9], the search engine family of VisualSEEk [15], WebSEEk [14], 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 that uses a World Wide Web browser as the user interface and the Tree Structured Self-Organizing Map (TS-SOM) [7, 8] as the image similarity scoring method. The retrieval method is based on the relevance feedback approach [13] adapted from traditional text-based information retrieval. In relevance feedback, the previous human-computer interaction is used to refine subsequent queries to better approximate the need of the user.

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.

As far as the current authors are aware, there has not been until now notable image retrieval applications based on the Self-Organizing Map (SOM) [6]. Some preliminary experiments with SOM have been made previously in [17]. MIT's FourEyes image browser uses Self-Organizing Maps to cluster weights for different features [9].

2 Principle of PicSOM

Our method is named PicSOM, which bears similarity to the well-known WEBSOM [16, 5] 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 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.

PicSOM is intended as a general framework for multi-purpose content-based image retrieval. The system is designed to be open and able to adapt to different kinds of image databases, ranging from small and domain-specific picture sets to large general purpose image collections. 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. In this paper, we describe the PicSOM system in its current form.

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 a small "thumbnail" size. The user then selects the subset of these images which best matches her expectations and to some degree of relevance fits 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 data, 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. Hopefully, the fraction of relevant images increases in each image set 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 calculate the best-scoring similarity results. The feature selection is not restricted in any way and new features can be added to the system later on, as long as an equal number of features are calculated from each picture in the database.

To give an example, consider two simple low-level features, color and texture. Color is a natural and widely-used feature in content-based image retrieval. Common representations for color information in image retrieval include color histograms, color moments, color layouts and the recent color correlograms.

In PicSOM, average R-, G-, and B-values are calculated in five separate regions of the image, as seen in Figure 1. 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.

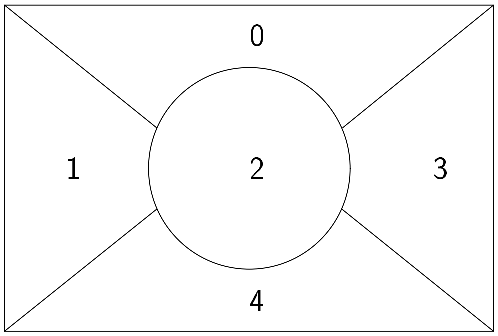

Figure 1: Image regions used calculating color and texture feature vectors.

Texture is an innate property of all surfaces and therefore a suitable feature for image retrieval. Texture features for pattern recognition and computer vision have been researched extensively over the past decades and the achievements in the field include co-occurrence matrices, multi-resolution simultaneous autoregressive (MRSAR) models, shift-invariant eigenvector (EV) models, the Wold decomposition, and wavelets, among others.

The texture feature vectors in PicSOM are calculated separately in the same five regions as the color features and seen in Figure 1. 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 to five eight-dimensional vectors which are combined to one 40-dimensional texture feature vector.

2.2 Tree Structured SOM (TS-SOM)

The Tree Structured Self-Organizing Map (TS-SOM) [7, 8] is a tree-structured vector quantization algorithm that uses Self-Organizing Maps (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 on the underlying SOM level 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 2.

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

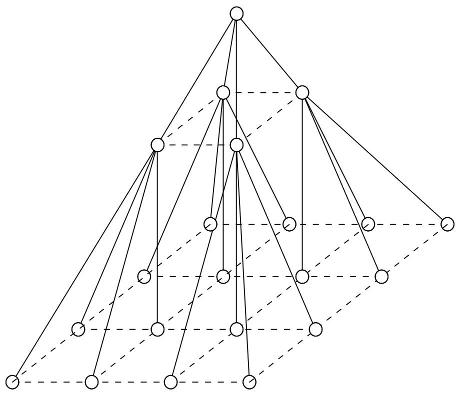

Figure 2: The structure of a three-layer two-dimensional TS-SOM.

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 average of all feature vectors mapped to 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 use the relevance feedback approach. The results of multiple maps then are combined automatically, using the implicit information from the user's responses during the previous rounds of 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 simply by marking on the maps the images shown to the user until now with positive and negative values depending whether she has selected or rejected them, respectively. 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, those areas which have 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 SOM layer increases.

Figure 3 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.

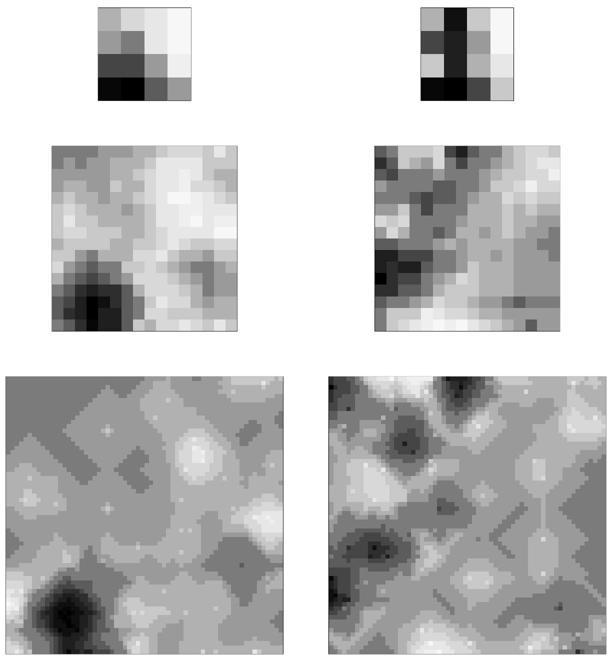

Figure 3: An example of convolved TS-SOMs for color (left) and texture (right) features. Black corresponds to positive and white to negative convolved values.

2.4 Refining Queries

In our current PicSOM implementation, all positive values on all convolved TS-SOM layers are sorted in descending order in one list. Then, a preset number, e.g. 15, of the best candidate images which have not been shown to the user before are output as a new tentative image selection. Image retrieval with PicSOM is therefore an iterative process in which new images get selected or rejected by the user.

Initially, the query begins with a set of reference images picked from the top levels of the TS-SOMs in use. The SOM map units associated with the selected and rejected images get positive and the negative values, respectively. 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 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 uses 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. This also makes the queries on the databases machine independent, because the standard web browsers can be used. The PicSOM search engine and further information on the system is available at *http://www.cis.hut.fi/picsom/*.

Secondly, 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 the inclusions of new features to the system in future.

3.1 User Interface

Figure 4 shows a screenshot of the current web-based PicSOM user interface. On the top of the page, there are

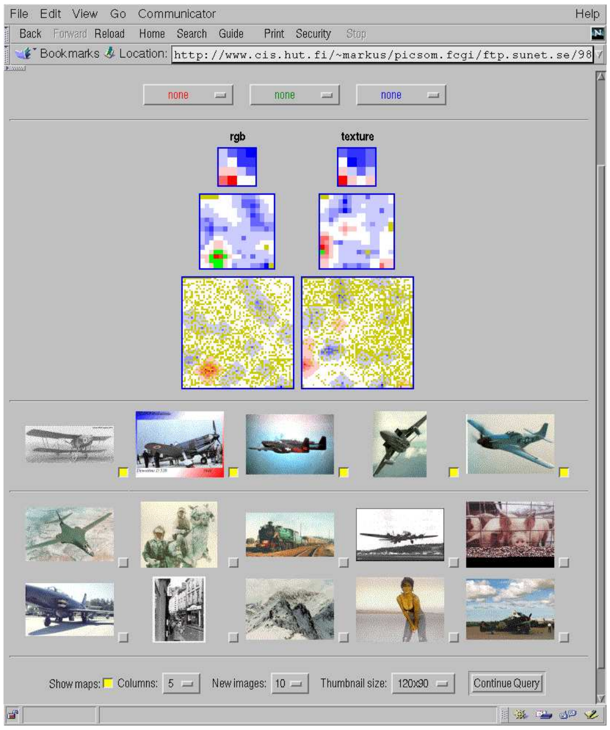

Figure 4: WWW-based user interface of PicSOM. The user has already selected five aircraft images in the previous rounds. The system is displaying the user ten new images to select of.

three pull-down menus for examining class information on the RGB color bands, if that information is available for the particular database. The convolved feature maps are shown next on the page. In this query, three-layer RGB color and texture maps have been used as seen on the labels above the maps. On color terminals, positive map points are seen as blue and negative as red. Darker shades represent stronger responses. White represents zero valued points.

The first row displays images selected on previous rounds of the retrieval process. This example shows a query with five images of aircrafts selected previously as relevant images. The next images are the ten best-scoring new images obtained from the convolved units in the TS-SOMs. It seems that these ten images contain four aircrafts. The relevant images can be selected positive by marking the appropriate checkboxes. Finally, the page has some user-modifiable settings and a "Continue Query" button which submits the new selections back to the search engine.

The user can at any time switch from the iterative queries to examining of the TS-SOM surfaces simply by clicking the map images. A portion of the map around the chosen viewpoint is then shown before the other images. Relevant images on the map surface can then also be selected for continuing queries.

3.2 Parts of the PicSOM System

The current computer implementation of PicSOM has three separate modular components:

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. The performance of the system is improved by using the FastCGI extensions, which enable the one started instance of the script to remain active for the duration of the entire retrieval process.

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 next best-scoring images to be shown to the user in the next round. With the FastCGI extensions, the *picsomctrl* program runs in resident mode, in which one instance of the program handles all the rounds of the query. This removes the burden of initializing the program for each round, as especially reading the TS-SOM data files takes a considerable amount of processing time. The *picsomctrl* program is also used for the required offline calculations, such as the creation and training of the TS-SOMs. The program also includes a special analyse mode used for evaluating the performance of the system.

picsomctrltohtml creates the HTML contents of the new web pages based on the output from the *picsomctrl* program. This separation of user interface generation makes possible to use several user interfaces. The WWW based UI can easily be changed by replacing this component.

Figure 5 illustrates the components of the current PicSOM system and the operations needed in handling the queries. The numbers indicate the normal order of actions.

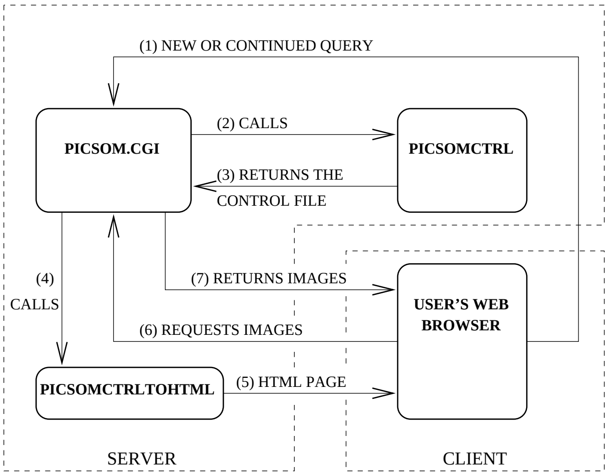

Figure 5: The components of the PicSOM system and the operations performed in handling the queries.

4 Current Databases

Currently, we have made our experiments with an image database of 4350 images. Most of them are color photographs in JPEG format. The images were downloaded from the image collection residing at the Swedish University Network FTP server, located at *ftp://ftp.sunet.se/pub/pictures/*.

To study our method's applicability we need to use larger image databases. For this purpose, we have obtained a photo collection from the Corel Gallery [3]. The collection has nearly 60.000 photographs. The images are all in color and also are identical in size.

PicSOM also supports the utilization of textual class information for the images, if that kind of information is available in the database. In the SUNET database, the original directory structure of the collection has been used to give the images rough textual content classes. Figure 6 shows a tree-form representation of a small subset of the used classes. The classes on child nodes are subclasses of the classes on their father nodes. For instance, {"cars"} $\subset$ {"vehicles"}.

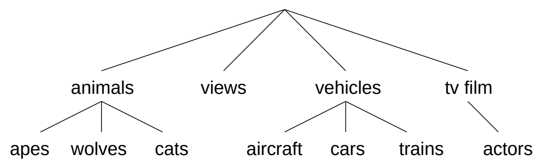

Figure 6: A subset of the used image classes in the *ftp.sunet.se* database.

In the user interface, the convolved TS-SOM map views can be changed to maps colored with this external information of the image content. The three color bands in the RGB color space can be used to visualize the spreads of three individual classes on the maps.

5 Conclusions and Future Plans

Currently, the PicSOM system is at a prototype level, so we have no comparable results on the retrieval performance of the system. The system does show potential and we are confident that it can evolve into a usable and fully functional system for image retrieval. In order to be able to prove that, we need to develop quantitative measures to assess the quality of the responses produced by PicSOM. These same measurement could then be applied to other content-based image retrieval systems to facilitate fare evaluations and comparisons.

Quantitative measures of the image retrieval performance of a system, or any single feature, are problematic due to human subjectivity. There exists no definite right answer to an image query as each user has individual expectations. Therefore, objective comparisons with others systems are difficult. To alleviate the situation, the MPEG organization has started to work on a new standard, called MPEG-7 or "Multimedia Content Desciption Interface" [10], to develop a set of features for image content description. The organization also plans a standard testbed for image retrieval applications.

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, and some more sophisticated texture models. We also plan to compare our features with ones used in other retrieval systems. As the system is designed to be modular and expandable, adding new statistical features is straightforward.

One addition to our system will be shape features which often yield important information on image content. We have made some experiments on using sobel operators to detect edges with different directions on the images. On some image types, they clearly outperform our current color and texture features.

As a vast collection of images is available on the Internet, we have made preliminary plans to use PicSOM as an image search engine for the World Wide Web.

References

[1] J. R. Bach, C. Fuller, A. Gupta, et al. 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] M. Beigi, A. Benitez, and S.-F. Chang. MetaSEEk: A content-based metasearch engine for images. In *Storage and Retrieval for Image and Video Databases VI (SPIE)*, volume 3312 of *SPIE Proceedings Series*, San Jose, CA, USA, January 1998.

[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. 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., August 1994.

[8] P. Koikkalainen and E. Oja. 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.

[9] T. Minka. An image database browser that learns from user interaction. Master's thesis, M.I.T, Cambridge, MA, 1996.

[10] MPEG-7: Context and objectives (version - 10 Atlantic City), October 1998. MPEG 98, ISO/IEC JTC1/SC29/WG11 N2460.

[11] A. Pentland, R. Picard, and S. Sclaroff. 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.

[12] Y. Rui, T. Huang, and S.-F. Chang. Image retrieval: Past, present and future. *Journal of Visual Communication and Image Representation*, 1998. To appear.

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

[14] J. R. Smith and S.-F. Cang. Searching for images and videos on the world-wide web. Technical Report #459-96-25, Columbia University, 1996.

[15] J. R. Smith and S.-F. Chang. VisualSEEk: A fully automated content-based image query system. In *Proceedings of the ACM Multimedia 1996*, Boston, MA, November 1996.

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

[17] 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.