published in the Proceedings of IJCNN'99, Washington, DC, July 1999
available at <URL:http://www.cis.hut.fi/picsom/publications.html>

PicSOM: Self-Organizing Maps for 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

Digital image libraries are becoming more common and widely used as more visual information is produced at a rapidly growing rate. Content-based image retrieval is an important approach to the problem of processing this increasing amount of data. It is based on automatically extracted features from the content of the images, such as color, texture, shape, and structure. We have started a project to study methods for content-based image retrieval using the Self-Organizing Map (SOM) as the image similarity scoring method. Our image retrieval system, named PicSOM, can be seen as a SOM-based approach to relevance feedback which is a form of supervised learning to adjust the subsequent queries based on the user's responses during the information retrieval session. In PicSOM, a separate Tree Structured SOM (TS-SOM) is trained for each feature vector type in use. The system then adapts to the user's preferences by returning her more images from those SOMs where her responses have been most densely mapped.

Introduction

Content-based image retrieval (CBIR) utilizes the visual content of the images in the process of searching and retrieving images from an image database. The aim is to obtain discriminants which correspond as well as possible with the human judgment for image similarity and can be used in conducting image queries. In CBIR, the images are indexed by features directly derived from the visual content of the image. These features usually contain rather low-level information such as the colors, textures, shapes, and spatial relations the image contains.

Most traditional applications in computer vision are usually automatic and self-contained. In CBIR, however, the user is an inseparable part of the process. As the retrieval systems are usually not capable of returning the satisfactory images in their first response to the user, the image query becomes an iterative and interactive process towards the desired image or images. Relevance feedback is a common term in text-based retrieval to describe a form of supervised learning to adjust the subsequent queries using the information gathered from the user feedback [12]. This helps the following rounds of the retrieval process to approximate better the present need of the user.

We have started to study the techniques for utilizing the strong self-organizing power of the Self-Organizing Map (SOM) [7] to facilitate content-based retrieval from image databases. As a part of the project, we have implemented an experimental image retrieval system called PicSOM. PicSOM uses a hierarchical version of the SOM algorithm called Tree Structured Self-Organizing Map (TS-SOM) [8, 9] as the method for retrieving images similar to a given set of reference images. PicSOM supports multiple parallel features and, with a technique introduced in the PicSOM system, the responses from multiple TS-SOMs are combined automatically. The goal is to autonomously adapt to the user's preferences regarding the similarity of images in the database.

Related Work

The best-known system for content-based image retrieval is probably IBM's Query By Image Content (QBIC) [5]. Recently, a number of other CBIR systems, both academic and commercial, have originated, including MIT Media Lab's Photobook [11], NETRA [10], developed in the Alexandria Digital Library project, Virage [1] by Virage Technologies Inc. and Colombia University's VisualSEEk [13].

PicSOM bears resemblance to the WEBSOM [6] document browsing and exploration tool developed at the Neural Network Research Centre at Helsinki University of Technology. WEBSOM is a means for organizing text documents into meaningful maps for exploration and search. It automatically organizes the documents into a two-dimensional grid so that related documents appear close to each other.

Some previous experiments on using the SOM as an indexing tool in content-based image retrieval have been made in [14]. In [4], the SOM was used to classify regions of similar texture in astronomical images in an image retrieval system called ASPECT.

Tree-Structured Self-Organizing Map

The Self-Organizing Map (SOM) [7] is an unsupervised, self-organizing neural algorithm which is widely used to visualize and interpret large high-dimensional data sets. The SOM defines an elastic net of points that are fitted to the input space. It can thus be used to visualize multidimensional data, usually on a two-dimensional grid.

To speed up the search of the best-matching unit (BMU), a variant of SOM called the Tree Structured Self-Organizing Map (TS-SOM) was introduced in [8, 9]. TS-SOM is a tree-structured vector quantization algorithm that uses normal SOMs at each of its hierarchical levels.

The TS-SOM is loosely based on the traditional tree-search algorithm. Due to the tree structure, the number of map units increases when moving downwards on the SOM layers of the TS-SOM. The search space for the best-matching vector on the underlying SOM layer is restricted to a predefined portion just below the best-matching unit of the above SOM. Instead of most tree-structured algorithms, the search space is not limited to the children of the BMU on the upper layer but can be set to include also neighboring nodes having different parents on the upper layer. The tree structure reduces the time complexity of the search from $O(N)$ to $O(\log N)$. The complexity of the searches is thus remarkably lower than if the whole bottommost SOM level had been accessed without the tree structure.

Principle of PicSOM

In PicSOM, the image queries are performed through the World Wide Web. The queries are iteratively refined as the system exposes more images from its database to the user. During the process, PicSOM tries to adapt to the user's preferences regarding to the similarity of images. This is accomplished with the use of separate TS-SOMs for every type of feature vectors extracted from the images. Depending on how close to each other the images accepted by the user are mapped onto a particular SOM layer, the more the system favors the images proposed by that very SOM.

A typical image query with PicSOM begins with the following steps. First, an interested user connects with her web browser to the WWW server providing the search engine. The system presents the user the first set of reference images which are uniformly picked from the top levels of the TS-SOMs in use. The user then selects the subset of images which match her expectations best and to some degree of relevance fit to her purposes. Then, she hits the “Continue Query” button and her browser sends the information on the accepted images back to the search engine.

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. These values are then summed up in their best-matching SOM units in each of the TS-SOM maps. Each SOM layer is then treated as a two-dimensional matrix formed of values describing the user's responses to the contents of the map unit. Finally, the map matrices are low-pass filtered with symmetrical convolution masks in order to spread the user's responses to the neighboring units which, by presumption, contain images that are to some extent similar to the present ones. Starting from the SOM unit having the largest convolved response value, PicSOM retrieves from the database the image whose feature vector is nearest to the weight vector in that unit. If that image has not been shown to the user, it is marked to be shown on the next round. This process is continued with the second largest value and so on until a preset number of new images have been selected. This set is then presented to the user. The iteration is continued until the user is satisfied with one or more images returned by the system.

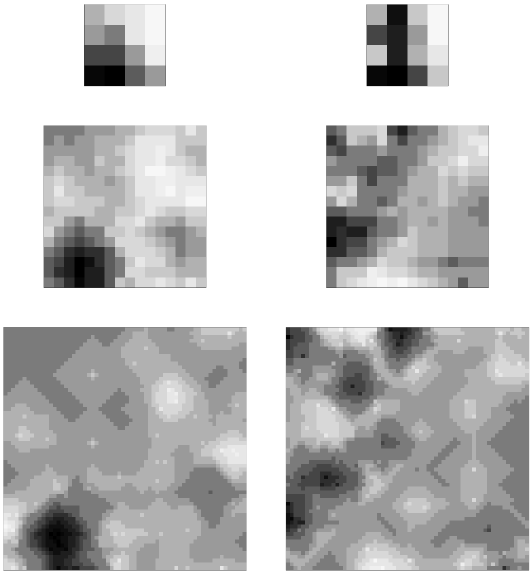
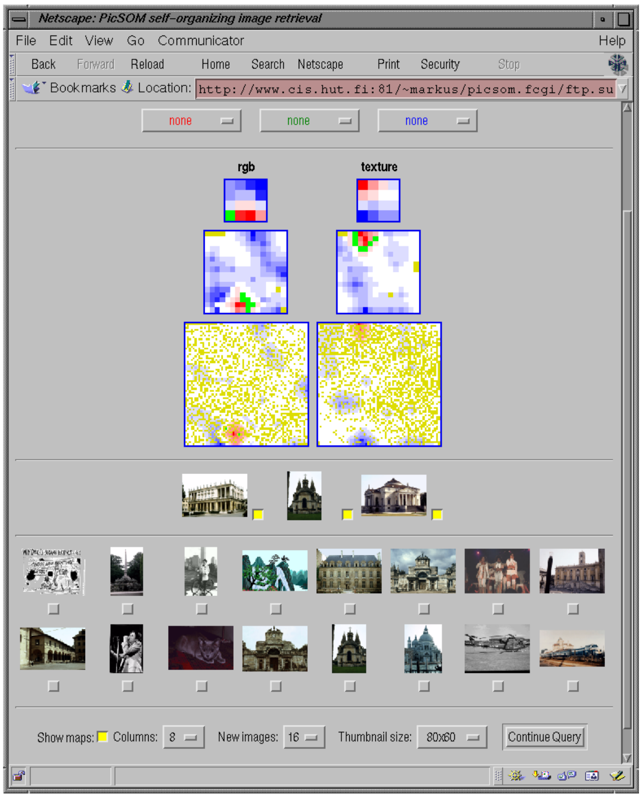


Figure 1: An example of convolved TS-SOMs for color (left) and texture (right) features. Black corresponds to positive and white to negative convolved values.

Figure 1 illustrates two convolved TS-SOMs. The three images on the left represent three map levels of a SOM for RGB color features, whereas the convolutions on the right are calculated on a map of simple textural features. The sizes of the SOM layers are 4×4, 16×16, and 64×64, from top to bottom.


Netscape: PicSOM self-organizing image retrieval
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location:
none
none
none
rgb
texture
Continue Query

Location: http://www.cis.hut.fi:81/~markus/picsom.fcgi/ftp.su

Show maps: Columns: 8 New images: 16 Thumbnail size: 80x80

Figure 2: The WWW-based user interface of PicSOM.

The distinct features do not thus need to be weighted by the user or administrator of the system as required in many other CBIR systems. Instead, PicSOM automatically adapts to the user's particular needs and impression of the similarity of images. Another intrinsic feature in our system is its ability to use any number of reference images, while many other current systems are based on using a single reference image.

The WWW-based user interface is illustrated in Figure 2. 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.

Experiments

We evaluated the usefulness of the PicSOM approach with a set of experiments. We used an image database of 4350 images containing mainly color photographs in JPEG format. This set of images was downloaded from location *ftp://ftp.sunet.se/pub/pictures/*. Three different feature extraction methods were applied to this data and the corresponding TS-SOMs were created.

In the first feature vector set, named *color*, average RGB values were calculated in five zones of the images, corresponding to the left, right, top, bottom, and center parts of the image area. This resulted to 15-dimensional vectors. The second set, named *texture*, was formed likewise in the five zones. This time, each pixel's YIQ luminance was compared to that of its eight neighbors. The estimated probability that the center pixel has larger value than the neighbor was used as a feature. The feature vector was thus 40-dimensional. In the last set [2], named *shape*, the images were first transformed to binarized edge images by using 3 × 3-sized Sobel masks. The edge directions were then discretized to eight values. An 8-bin histogram was formed from the directions in the same five zones as for the other two feature types. This gave rise to a 40-dimensional feature vector.

The TS-SOMs for all features were sized 4 × 4, 16 × 16, and 64 × 64, from top to bottom. During the training, each vector was used 100 times in the adaptation. Each of the TS-SOMs was first used alone and then all the three maps together. A quantitative measure for the image retrieval performance was obtained as follows. First, we selected from our database the subset of images which could be regarded as portraying an aircraft. There were 348 such images and thus *a priori* probability of 8.0 percent. We then implemented an "ideal screener", a computer program which examined the output of the system and marked the images presented by PicSOM either as selected or non-selected according to the hand-picked base truth. The query processing could thus be simulated and performance data collected without any human intervention.

For each of the 348 aircraft images we recorded the total number of images presented by the system until that particular image was shown. From this data, we formed histograms and calculated the average number of shown images needed before hit. After division by 4350, this figure yielded a value $\tau \in (0, 1]$. For values $\tau < 0.5$, the performance of the system was thus better than random picking of images and, in general, the smaller the τ value the better the performance. In our experiments, the system always presented 16 new images to the screening program. The selection of this parameter naturally has effect on the resulting τ value.

As the above-described procedure was repeated for image sets containing buildings (492 items) and human faces (361 items), the results in Table 1 were obtained. The last row gives the unweighted average of the results for the three image sets. The results show that the *shape* features clearly yield better performance than the two other feature types in

	feature types and TS-SOMs used			
images	*color*	*texture*	*shape*	*all*
aircraft	0.307	0.436	0.220	0.261
buildings	0.319	0.312	0.309	0.300
faces	0.372	0.356	0.364	0.362
average	0.333	0.368	0.298	0.308

Table 1: Retrieval performances τ for three distinct features alone and the combined use of all three together in PicSOM.

the case of the aircraft images. For the building and face image classes the performances of the three individual TS-SOMs are more even.

The results for the PicSOM system with the three TS-SOMs used together are in every case either the best or the second best ones. PicSOM system is thus to some extent able to benefit from the existence of multiple feature types. As it is not beforehand known which TS-SOM would perform best for the user's query, the PicSOM approach provides a robust method for using a set of image maps in parallel.

We have also experimented with other types of feature vectors. Also in the light of those results, it seems that if one feature vector type has clearly better retrieval performance τ than the others, it is more beneficial to use that particular TS-SOM alone than to use also the worse-performing maps. Therefore, it is necessary for the proper operation of the PicSOM system that the used features are well balanced, i.e., they should on the average perform quite similarly by themselves.

Conclusions

We have in this paper introduced the PicSOM method for content-based image retrieval and a preliminary quantitative evaluation of its performance. The results of our experiments show that the PicSOM system is able to effectively select from a set of parallel TS-SOMs a combination which either nearly matches the best individual image map or slightly outperforms it in performance.

The features used in this study are certainly yet quite tentative. Therefore, we have started a large series of experiments for selecting a proper and well-balanced set of features to be used in the PicSOM system and our future assessments.

Our future plans also include the use of the Corel [3] database of 60 000 photograph images. Also, we have started to collect images from the World Wide Web and to index them by using PicSOM. The PicSOM 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. Brandt. Use of shape features in content-based image retrieval. Master's thesis, Helsinki University of Technology, 1999. To appear.

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

[4] A. Csillaghy. Neural network-generated indexing features and retrieval effectiveness. In *Proceedings of the Converging Computing Methodologies in Astronomy (CCMA) Conference*, Sonthofen, Bavaria, September 1997.

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

[6] 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.

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

[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] 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.

[10] W. Y. Ma and B. S. Manjunath. NETRA: A toolbox for navigating large image databases. In *IEEE International Conference on Image Processing (ICIP)*, Santa Barbara, California, October 1997.

[11] 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.

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

[13] 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, 1996.

[14] 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.