MPEG-7 Descriptors in Content-Based Image Retrieval with PicSOM System

Markus Koskela, Jorma Laaksonen, and Erkki Oja

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

Abstract. The MPEG-7 standard is emerging as both a general framework for content description and a collection of specific, agreed-upon content descriptors. We have developed a neural, self-organizing technique for content-based image retrieval. In this paper, we apply the visual content descriptors provided by MPEG-7 in our PicSOM system and compare our own image indexing technique with a reference system based on vector quantization. The results of our experiments show that the MPEG-7-defined content descriptors can be used as such in the PicSOM system even though Euclidean distance calculation, inherently used in the PicSOM system, is not optimal for all of them. Also, the results indicate that the PicSOM technique is a bit slower than the reference system in starting to find relevant images. However, when the strong relevance feedback mechanism of PicSOM begins to function, its retrieval precision exceeds that of the reference system.

1 Introduction

Content-based image retrieval (CBIR) differs from many of its neighboring research disciplines in computer vision due to one notable fact: human subjectivity cannot totally be isolated from the use and evaluation of CBIR systems. This is manifested by difficulties in setting fair comparisons between CBIR systems and in interpreting their results. These problems have hindered the researchers from doing comprehensive evaluations of different CBIR techniques.

We have developed a neural-network-based CBIR system named PicSOM [1, 2]. The name stems from "picture" and the Self-Organizing Map (SOM). The SOM [3] is used for unsupervised, self-organizing, and topology-preserving mapping from the image descriptor space to a two-dimensional lattice, or grid, of artificial neural units. The PicSOM system is built upon two fundamental principles of CBIR, namely query by pictorial example and relevance feedback [4].

Until now, there have not existed widely-accepted standards for description of the visual contents of images. MPEG-7 [5] is the first thorough attempt in this direction. The appearance of the standard will affect the research on CBIR techniques in some important aspects. First, when some common building blocks will become shared by different CBIR systems, comparative studies between them will become easier to perform. As MPEG-7 Experimentation Model (XM) [6] has become publicly available, we have been able to test the suitability of MPEG-7-defined image content descriptors with the PicSOM system. We have thus replaced our earlier, non-standard descriptors with those defined in the MPEG-7 standard and available in XM.

2 PicSOM System

The PicSOM image retrieval system [1, 2] is a framework for research on algorithms and methods for content-based image retrieval. The methodological novelty of PicSOM is to use several Self-Organizing Maps [3] in parallel for retrieving relevant images from a database. These parallel SOMs have been trained with separate data sets obtained from the image data with different feature extraction techniques. The different SOMs and their underlying feature extraction schemes impose different similarity functions on the images.

Every image query is unique and each user of a CBIR system has her own transient view of image similarity and relevance. Therefore, a system structure capable of holding many simultaneous similarity representations can adapt to different kinds of retrieval tasks. In the PicSOM approach, the system is able to discover those of the parallel Self-Organizing Maps that provide the most valuable information for each individual query instance.

A more detailed description of the PicSOM system and results of earlier experiments performed with it can be found in [1, 2]. The PicSOM home page including a working demonstration of the system for public access is located at *http://www.cis.hut.fi/picsom*.

2.1 Tree Structured Self-Organizing Maps

The main image indexing method used in the PicSOM system is the Self-Organizing Map (SOM) [3]. The SOM defines an elastic, topology-preserving grid of points that is fitted to the input space. It can thus be used to visualize multidimensional data, usually on a two-dimensional grid. The map attempts to represent all the available observations with an optimal accuracy by using a restricted set of models.

Instead of the standard SOM version, PicSOM uses a special form of the algorithm, the Tree Structured Self-Organizing Map (TS-SOM) [7]. The hierarchical TS-SOM structure is useful for large SOMs in the training phase. In the standard SOM, each model vector has to be compared with the input vector in finding the best-matching unit (BMU). This makes the time complexity of the search $O(n)$, where n is the number of SOM units. With the TS-SOM one can, however, follow the hierarchical structure and reduce the complexity of the search to $O(\log n)$. This reduction can be achieved by first training a smaller SOM and then creating a larger one below it so that the search for the BMU on the larger map is always restricted to a fixed area below the already-found BMU and its nearest neighbors on the above map.

Fig. 1. The surface of a 16×16-sized TS-SOM level trained with the MPEG-7 *Edge Histogram* descriptor.

In the experiments described in this paper, we have used four-level TS-SOMs whose layer sizes have been 4×4, 16×16, 64×64, 256×256 units. In the training of the lower SOM levels, the search for the BMU has been restricted to the 10×10-sized neuron area below the BMU on the above level. Every image has been used 100 times for training each of the TS-SOM levels.

After training each TS-SOM hierarchical level, that level is fixed and each neural unit on it is given a visual label from the database image nearest to it. This is illustrated in Figure 1, where MPEG-7 *Edge Histogram* descriptor has been used as the feature. The images are the visual labels on the surface of the 16×16-sized TS-SOM layer. It can be seen that, e.g., there are many ships in the top-left corner of the map surface, standing people and dolls beside the ships, and buildings in the bottom-left corner. Visually – and also semantically – similar images have thus been mapped near each other on the map.

2.2 Self-Organizing Relevance Feedback

The relevance feedback mechanism of PicSOM, implemented by using several parallel SOMs, is a crucial element of the retrieval engine. Only a short overview is presented here, see [2] for a more comprehensive treatment.

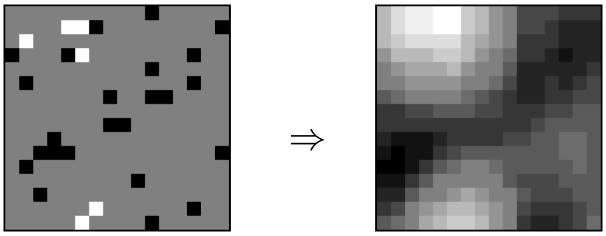

Fig. 2. An example of how a SOM surface, on which the images selected and rejected by the user are shown with white and black marks, respectively, are convolved with a low-pass filter.

Each image seen by the user of the system is graded by her as either relevant or irrelevant. All these images and their associated relevance grades are then projected on all the SOM surfaces. This process forms on the maps areas where there are 1) many relevant images mapped in same or nearby SOM units, or 2) relevant and irrelevant images mixed, or 3) only irrelevant images, or 4) no graded images at all. Of the above cases, 1) and 3) indicate that the corresponding content descriptor agrees well with the user's conception on the relevance of the images. Whereas, case 2) is an indication that the content descriptor cannot distinguish between relevant and irrelevant images.

When we assume that similar images are located near each other on the SOM surfaces, we are motivated to spread the relevance information placed in the SOM units also to the neighboring units. This is implemented in PicSOM by low-pass filtering the map surfaces. All relevant images are first given equal positive weight inversely proportional to the number of relevant images. Likewise, irrelevant images receive negative weights that are inversely proportional to the number of irrelevant images. The overall sum of these relevance values is thus zero. The values are then summed in the BMUs of the images and the resulting sparse value fields are low-pass filtered. Figure 2 illustrates how the positive and negative responses, displayed with white and black map units, respectively, are first mapped on a SOM surface and how the responses are expanded in the convolution. Content descriptors that fail to coincide with the user's conceptions produce lower qualification values than those descriptors that match the user's expectations. As a consequence, the different content descriptors do not need to be explicitly weighted as the system automatically takes care of weighting their opinions.

In the actual implementation, we search on each SOM for a fixed number, say 100, map locations with unseen images having the highest qualification values. After removing duplicate images, the second stage of processing is carried out. Now, the qualification values of all images in this combined set are summed up on all used SOMs to obtain the final qualification values for these images. Then, 20 images with the highest qualification values are returned as the result of the query round.

In the experiments described in this paper, the queries are always started with an image that belongs to the image class in question. Therefore, we neglected the TS-SOM hierarchy and considered exclusively the bottommost TS-SOM levels. This mode of operation is motivated by the chosen query type, since it is justifiable to start the retrieval near the initial reference image. This can be seen as *depth first search*.

However, the hierarchical representation of the image database produced by a TS-SOM is useful in visual browsing. The successive map levels can be regarded as providing increasing resolution for database inspection. In our earlier experiments, e.g. [1, 8, 2], there was no initial example image to start the query with and the queries began with initial *breadth first search* using the visual labels and the TS-SOM structure.

2.3 Vector-Quantization-Based Reference Method

There exists a wide range of distinct techniques for indexing images based on their feature descriptors. One alternative method for the SOM is to first use quantization to prune the database and then utilize a more exhaustive method to decide the final images to be returned. For the first part, there exists two alternate quantization techniques, namely scalar quantization (SQ) and vector quantization (VQ). With either of these techniques, the feature vectors are divided into subsets in which the vectors resemble each other. In the case of scalar quantization the resemblance is in respect to one component of the feature vector, whereas resemblance in vector quantization means that the feature vectors are similar as whole. In our previous experiments [8], we have found out that scalar quantization gives bad retrieval results.

The justification for vector quantization in image retrieval is that unseen images which have fallen into the same quantization bins as the relevant-marked reference images are good candidates for the next reference images to be displayed to the user. Also, the SOM algorithm can be seen as a special case of vector quantization. When using the model vectors of the SOM units in vector quantization, one ignores the topological ordering provided by the map lattice and characterize the similarity of two images only by whether they are mapped in the same VQ bin. By ignoring the topology, however, we dismiss the most significant portion of the data organization provided by the SOM.

A well-known VQ method is the K-means or Linde-Buzo-Gray (LBG) vector quantization [9]. According to [8], LBG quantization yields better CBIR performance than the SOM used as a pure vector quantizer. This is understandable as the SOM algorithm can be regarded as a trade-off between two objectives, namely clustering and topological ordering. Consequently, we will use LBG quantization in the reference system of the experiments.

The choice for the number of quantization bins is a significant parameter for the VQ algorithm. Using too few bins results in too broad image clusters to be useful whereas with too many bins the information about the relevancy of images fails to generalize to other images. Generally, the number of bins should be smaller than the number of neurons on the largest SOM layer of the TS-SOM.

In the experiments, we have used 4096 VQ bins, which coincides with the size of the second bottommost TS-SOM levels. This results in 14.6 images per VQ bin, on the average, for the used database of 59 995 images. Another significant parameter is the number of candidate images that are taken into consideration from each of the parallel vector quantizers. Different selection policies lead again either to *breadth first* or *depth first* searches. In our implementation, we rank the VQ bins of each quantizer in the descending order determined by the proportion of relevant images of all graded images in them. Then, we select 100 yet unseen images from the bins in that order.

After the vector quantization stage, the set of potential images has been greatly reduced and more demanding processing techniques can be applied to all the remaining candidate images. Now, one possible method – also applied in our reference system – is to rank the images based on their properly-weighted cumulative distances to all already-found relevant images in the original feature space. Finally, as in the PicSOM method, we display 20 best-scoring images to the user. In [8], it was found out that the VQ method benefits from this extra processing stage. As calculating distance in a possibly very high-dimensional space is a computationally heavy operation, the vector quantization can thus be seen to act as a preprocessor which prunes a large database as much as it is necessary before the actual image similarity assessment is carried out.

3 Experiments

The performance of a CBIR system can be evaluated in many different ways. Even though the interpretation of the contents of images is always casual and ambiguous, some kind of ground truth classification of images must be performed in order to automate the evaluation process. In the simplest case – employed also here – some image classes are formed by first selecting verbal criteria for membership in a class and then assigning the corresponding Boolean membership value for each image in the database. In this manner, a set of ground truth image classes, not necessary non-overlapping, can be formed and then used in the evaluation.

3.1 Performance Measures and Evaluation Scheme

All features can be studied separately and independently from others for their capability to map visually similar images near each other. These kinds of feature-wise assessments, however, have severe limitations because they are not related to the operation of the entire CBIR system as a whole. In particular, they do not take any relevance feedback mechanism into account. Therefore, it is preferable to use evaluation methods based on the actual usage of the system.

If the size of the database, N, is large enough, we can assume that there is an upper limit N_T of images ($N_T \ll N$) the user is willing to browse. The system should thus demonstrate its talent within this number of images. In our setting, each image in a class $\mathcal{C}$ is "shown" to the system one at a time as an initial image to start the query with. The mission of the CBIR system is then to return as much as possible similar images. In order to obtain results that do not depend on the particular image used in starting the iteration, the experiment needs to be repeated over every image in $\mathcal{C}$. This results in a leave-one-out type testing of the target class and the effective size of the class becomes $N_{\mathcal{C}} - 1$ instead of $N_{\mathcal{C}}$ and the *a priori* probability of the class is $\rho_{\mathcal{C}} = (N_{\mathcal{C}} - 1)/(N - 1)$.

We have chosen to show the evolution of *precision* as a function of *recall* during the iterative image retrieval process. Precision and recall are intuitive performance measures that suite also for the case of non-exhaustive browsing. When not the whole database but only a smaller number $N_T \ll N$ of images is browsed through, the recall value very unlikely reaches the value one. Instead, the final value $\mathcal{R}(N_T)$ – as well as $\mathcal{P}(N_T)$ – reflects the total number of relevant images found that far. The intermediate values of $\mathcal{P}(t)$ first display the initial accuracy of the CBIR system and then how the relevance feedback mechanism is able to adapt to the class. With an effective relevance feedback mechanism, it is to be expected that $\mathcal{P}(t)$ first increases and then turns to decrease when a notable fraction of relevant images have already been shown.

In our experiments, we have normalized the precision value by dividing it with the *a priori* probability $\rho_{\mathcal{C}}$ of the class and call it therefore *relative precision*. This makes the comparison of the recall–precision curves of different image classes somewhat commensurable and more convenient because relative precision values relate to the relative advantage the CBIR system produces over random browsing.

3.2 Database and Ground Truth Classes

We have used images from the Corel Gallery 1 000 000 product in our evaluations. The database contains 59 995 color photographs originally packed with a wavelet compression and then locally converted in JPEG format with a utility provided by Corel. The size of each image is either 384×256 or 256×384 pixels.

The images have been grouped by Corel in thematic groups and also keywords are available. However, we found these image groups and keywords rather inconsistent and, therefore, created for the experiments six manually-picked ground truth image sets with tighter membership criteria. All image sets were gathered by a single subject. The used sets and membership criteria were:

- **faces**, 1115 images (*a priori* probability 1.85%), where the main target of the image is a human head which has both eyes visible and the head fills at least 1/9 of the image area.
- **cars**, 864 images (1.44%), where the main target of the image is a car, at least one side of the car has to be completely shown in the image, and its body to fill at least 1/9 of the image area.
- **planes**, 292 images (0.49%), where all airplane images have been accepted.
- **sunsets**, 663 images (1.11%), where the image contains a sunset with the sun clearly visible in the image.
- **houses**, 526 images (0.88%), where the main target of the image is a single house, not severely obstructed, and it fills at least 1/16 of the image area.

- **horses**, 486 images (0.81%), where the main target of the image is one or more horses, shown completely in the image.

3.3 MPEG-7 Content Descriptors

MPEG-7 [5] is an ISO/IEC standard developed by Moving Pictures Expert Group. MPEG-7 aims at standardizing the description of multimedia content data. It defines a standard set of descriptors that can be used to describe various types of multimedia information. The standard is not aimed at any particular application area, instead it is designed to support as broad a range of applications as possible. Still, one of the main applications areas of MPEG-7 technology will undoubtedly be to extend the current modest search capabilities for multimedia data for creating effective digital libraries. As such, MPEG-7 is the first serious attempt to specify a standard set of descriptors for various types of multimedia information and standard ways to define other descriptions as well as structures of descriptions and their relationships.

As a nonnormative part of the standard, a software Experimentation Model (XM) [6] has been released for public use. The XM is the framework for all reference code of the MPEG-7 standard. In the scope of our work, the most relevant part of XM is the implementation of a set of MPEG-7-defined still image descriptors. At the time of this writing, XM is in its version 5.3 and not all description schemes have yet been reported to be working properly. Therefore, we have used only a subset of MPEG-7 content descriptors for still images in these experiments. The used descriptors were *Scalable Color*, *Dominant Color*, *Color Structure*, *Color Layout*, *Edge Histogram*, and *Region Shape*.

The MPEG-7 standard defines not only the descriptors but also special metrics to be used with the descriptors when calculating the similarity between images. However, we use Euclidean metrics in comparing the descriptors because the training of the SOMs and the creation of the vector quantization prototypes are based on minimizing a square-form error criterium. Only in the case of *Dominant Color* descriptor this has necessitated a slight modification in the use of the descriptor. The original *Dominant Color* descriptor of XM is variable-sized, i.e., the length of the descriptor varies depending on the count of dominant colors found. Because this could not be fit in the PicSOM system, we used only two most dominant colors or duplicated the most dominant color if only one was found. Also, we did not make use of the color percentage information. These two changes do not make our approach incompatible with other uses of *Dominant Color* descriptor.

3.4 Results

Our experiments were two-fold. First, we wanted to study which of the four color descriptors would be the best one to be used together with the one texture and one shape descriptors in the table. Second, we wanted to compare the performance of our PicSOM system with that of the vector-quantization-based variant. We performed two sets of experiments in which the first question was addressed in the first set and the second question in both sets.

We performed 48 computer runs in the first set of experiments. Each run was characterized by the combination of the method (PicSOM / VQ), color feature (*Dominant Color* / *Scalable Color* / *Color Layout* / *Color Structure*) and the image class (**faces** / **cars** / **planes** / **sunsets** / **houses** / **horses**). Each experiment was repeated as many times as there were images in the image class in question, the recall and relative precision values were recorded for each such instant and finally averaged. 20 images were shown at each iteration round, which resulted in 50 rounds when N_T was set to 1000 images. Both recall and relative precision were recorded after each query iteration. Figure 3 shows, as a representative selection, the recall–relative precision curves of three of the studied image classes (**faces**, **cars**, and **planes**). Qualitatively similar behavior is observed with the three other classes as well. The recorded values are shown with symbols and connected with lines.

The following observations can be made from the resulting recall–relative precision curves. First, none of the tested color descriptors seems to dominate the other descriptors and on different image classes the results of different color descriptors often vary considerably. Regardless of the used retrieval method (PicSOM or VQ), *Color Structure* seems to perform best with **faces** and using *Scalable Color* yields best results with **planes** and **horses**. With the other classes (**cars**, **sunsets**, **houses**), naming a single best color descriptor is not as straightforward. The second observation is that, in general, if a particular color descriptor works well for a particular image class, it does so with both retrieval algorithms. Third, the PicSOM method more often obtains better precision then the VQ method when comparing the same descriptor sets, although the difference is rather small. Also, in the end, PicSOM has in a majority of cases reached a higher recall level. The last observation here is, that the difference between the precision of the best and the worst sets of descriptors is larger with the VQ method than with PicSOM. This can be observed, e.g., in the **planes** column of Figure 3.

In the second set of experiments, we wanted to use all the available MPEG-7 visual content descriptors simultaneously. Runs were again made separately for the six image classes and the two CBIR techniques. The results for all classes can be seen in Figure 4, where each plot now contains mutually comparable recall–relative precision curves of the two techniques. It can be seen in Figure 4 that in all cases PicSOM is at first behind of VQ in precision, but soon reaches and exceeds it. In some of the cases (**faces** and **cars**), this overtake by PicSOM takes only one or two rounds of queries. With **planes**, reaching VQ takes the longest time, 11 rounds, due to the good initial precision of VQ, observed also in Figure 3 with the *Scalable Color* descriptor.

Of the tested image classes, **sunsets** yields the best retrieval results as its relative precision rises at best over 30 and, on the average, almost half of all the images in the class are found among the 1000 retrieved images. This is understandable as sunset images can be well described with low-level descriptors,

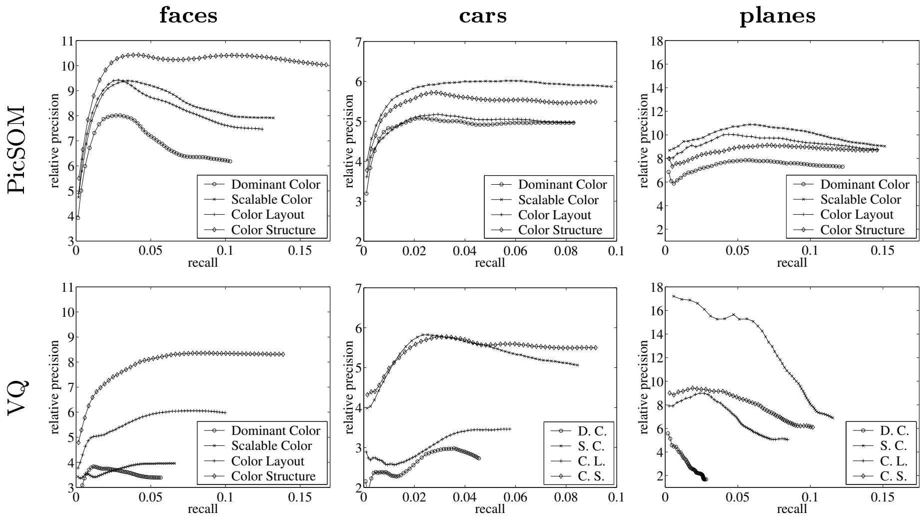

Fig. 3. Recall–relative precision plots of the performance of different color descriptors and the two CBIR techniques. In all cases also *Edge Histogram* and *Region Shape* descriptors have been used.

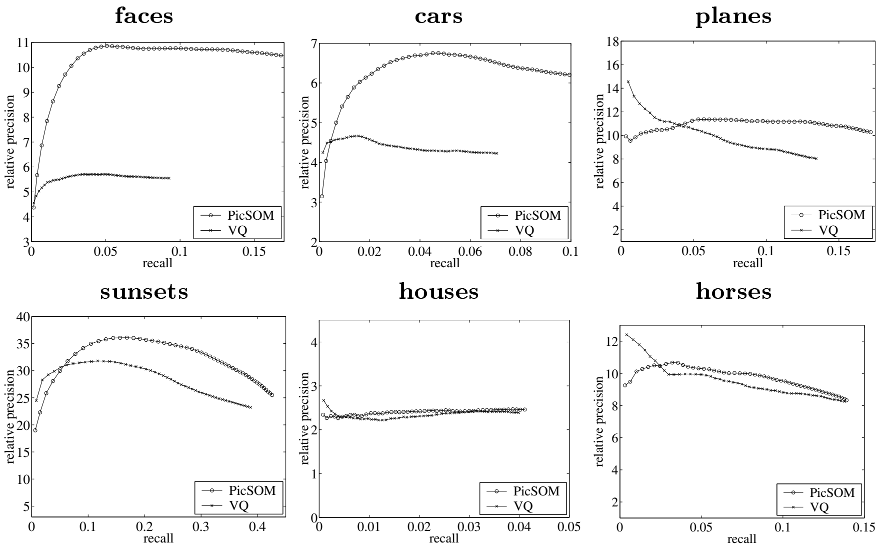

Fig. 4. Recall–relative precision plots of the performance of the two CBIR techniques when all four color descriptors were used simultaneously together with *Edge Histogram* and *Region Shape* descriptors.

especially color. On the other hand, **houses** is clearly the most difficult class, as its precision does not ever rise much above twice the *a priori* probability of the class. This is probably due to the problematic nature of the class as, descriptor-wise, there is not a large difference between the single houses and groups of houses, e.g., small villages.

As the final outcome of the experiment, it can be stated that the relevance feedback mechanism of PicSOM is clearly superior to that of VQ's. The VQ retrieval has good initial precision but after a few rounds, when PicSOM's relevance feedback begins to have an effect, retrieval precision with PicSOM is in all cases higher. The **houses** class can be regarded as a draw and a failure for both methods with the given set of content descriptors.

One can also compare the curves of Figure 3 and the curves in the upper row of Figure 4 for an important observation. It can be seen that the PicSOM method is, when using all descriptors simultaneously (Figure 4), able to follow and even exceed the path of the best recall–relative precision curve for the four alternative single color descriptors (Figure 3). This behavior is present in all cases, also with the image classes not shown in Figure 3, and can be interpreted as an indication that the automatic weighting of features is working properly and additional, inferior, descriptors do not degrade the results. On the contrary, the VQ method fails to do the same and the VQ recall–relative precision curves in Figure 4 resemble more the average than the maximum value of the corresponding VQ curves in Figure 3. As a consequence, the VQ technique is clearly more dependent on the proper selection of used features than the PicSOM technique.

4 Conclusions

In this paper, we have described our content-based image retrieval system named PicSOM and shown that MPEG-7-defined content descriptors can be successfully used with it. The PicSOM system is based on using Self-Organizing Maps in implementing relevance feedback from the user of the system. As the system uses many parallel SOMs, each trained with separate content descriptors, it is straightforward to use any kind of features. Due to PicSOM's ability to automatically weight and combine the responses of the different descriptors, one can make use of any number of content descriptors without the need to weight them manually. As a consequence, the PicSOM system is well-suited for operation with MPEG-7 which also allows the definition and addition of any number of new content descriptors.

In the experiments we compared the performances of four different color descriptors available in the MPEG-7 Experimentation Model software. The results of that experiment showed that no single color descriptor was the best one for all of our six hand-picked image classes. That result was no surprise, it merely emphasizes the need to use many different types of content descriptors in parallel. In an experiment where we used all the available color descriptors, the PicSOM system indeed was able to automatically reach and even exceed the best recall–precision levels obtained earlier with preselection of features. This is a very desirable property, as it suggests that we can initiate queries with a large number of parallel descriptors and the PicSOM systems focuses on the descriptors which provide the most useful information for the particular query instance.

We also compared the performance of the self-organizing relevance feedback technique of PicSOM with that of a vector-quantization-based reference system. The results showed that in the beginning of queries, PicSOM starts with a bit lower precision rate. Later, when its strong relevance feedback mechanism has enough data to process, PicSOM outperforms the reference technique. In the future, we plan to study how the retrieval precision in the beginning of PicSOM queries could be improved to the level attained by the VQ technique in the experiments.

Acknowledgments

This work was supported by the Finnish Centre of Excellence Programme (2000-2005) of the Academy of Finland, project New information processing principles, 44886.

References

1. Laaksonen, J.T., Koskela, J.M., Laakso, S.P., Oja, E.: PicSOM - Content-based image retrieval with self-organizing maps. Pattern Recognition Letters **21** (2000) 1199–1207
2. Laaksonen, J., Koskela, M., Laakso, S., Oja, E.: Self-organizing maps as a relevance feedback technique in content-based image retrieval. Pattern Analysis & Applications **4** (2001) 140–152
3. Kohonen, T.: Self-Organizing Maps. Third edn. Volume 30 of Springer Series in Information Sciences. Springer-Verlag (2001)
4. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. Computer Science Series. McGraw-Hill (1983)
5. MPEG: Overview of the MPEG-7 standard (version 5.0) (2001) ISO/IEC JTC1/SC29/WG11 N4031.
6. MPEG: MPEG-7 visual part of the eXperimentation Model (version 9.0) (2001) ISO/IEC JTC1/SC29/WG11 N3914.
7. Koikkalainen, P., Oja, E.: Self-organizing hierarchical feature maps. In: Proc. IJCNN-90, International Joint Conference on Neural Networks, Washington, DC. Volume II., Piscataway, NJ, IEEE Service Center (1990) 279–285
8. Koskela, M., Laaksonen, J., Oja, E.: Comparison of techniques for content-based image retrieval. In: Proceedings of 12th Scandinavian Conference on Image Analysis (SCIA 2001), Bergen, Norway (2001) 579–586
9. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. IEEE Transactions on Communications **COM-28** (1980) 84–95