*Published in Proceedings of IR 2001. Oulu, Finland. September 2001.*
*available at <URL:http://www.cis.hut.fi/picsom/publications.html>*

# SELF-ORGANIZING IMAGE RETRIEVAL WITH MPEG-7 DESCRIPTORS

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.** Development of content-based image retrieval techniques has suffered from the lack of standardized ways for describing visual image content. Luckily, MPEG-7 is now emerging as both a 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 our image retrieval system named PicSOM. In this paper, we apply the visual content descriptors provided by MPEG-7 in the PicSOM system and compare our indexing technique with a reference system. The results of our experiments show that the MPEG-7-defined content descriptors can be used as such in PicSOM even though Euclidean distance calculation, inherently used in the PicSOM system, is not optimal for all of them.

## 1. Introduction

Content-based image retrieval (CBIR) has been a subject of very intensive research effort for more than a decade. It 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 interpreting their results, which has partly hindered the researchers from doing comprehensive evaluations of different CBIR techniques.

We have developed a neural-network-based content-based image retrieval system named PicSOM [6, 7]. The name stems from Picture Self-Organizing Map (SOM), the SOM [2] being used for unsupervised, self-organizing, and topology-preserving mapping from the image descriptor space to a two-dimensional lattice. The PicSOM system is built upon two fundamental principles of CBIR, namely *query by pictorial example* (QBPE) [1] and *relevance feedback* (RF) [11].

In query by pictorial example, it is presumed that the user of a CBIR system has no other means of specifying her object of interest but giving or pointing out examples of interesting or relevant images. The user classifies images presented by the system as relevant or non-relevant to the current retrieval task and the system uses this information to select such images the user is most likely to be interested in. On the other hand, in relevance feedback it is assumed that one can build a CBIR system that is able to learn the user's preferences after seeing many enough examples of relevant images. This kind of behavior can be implemented by allowing the user to rank or otherwise evaluate the image outputs from the system. The justification for using relevance feedback is that there is an inherent semantic gap between high-level concepts used by humans and low-level features used by computers in judging image similarity. This gap prevents CBIR systems, in the general case, from providing correct answers as the immediate response and image retrieval becomes an iterative process.

Until now, there have not existed widely-accepted standards for description of visual contents of images. MPEG-7 [10] is the first thorough attempt in this direction and it is scheduled to become an international standard of ISO/IEC in September 2001. 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 descriptors as well as structures of descriptors and their relationships. Although MPEG-7 is not aimed at any particular application area, one of its main applications areas will be searching and retrieving multimedia content. In image retrieval, a natural step towards MPEG-7 is to use MPEG-7 content descriptors suitable for still images in retrieval systems. The standard will undoubtedly have an impact on CBIR system development: First, when some common building blocks will become shared by different CBIR systems, comparative studies between them will become easier to perform. Also, all CBIR developers should prepare to accept a challenge to apply their CBIR techniques to tasks that are expressed solely in terms of the standard. As the MPEG-7 Experimentation Model (XM) [9] has recently become publicly available, we have been able to test the suitability of MPEG-7-defined image content descriptions with PicSOM.

## 2. PicSOM system

The PicSOM image retrieval system [6, 7] is a framework for research on algorithms and methods of content-based image retrieval. The PicSOM home page including a demonstration of the system for public access is located at `http://www.cis.hut.fi/picsom`.

A methodological novelty of PicSOM is to use several Self-Organizing Maps [2] in parallel for retrieving relevant images from a database. These parallel SOMs have been trained with different data sets obtained from the image data with different feature extraction techniques. The most natural choice for data in CBIR are visual features extracted from the images in the database. These features describe eg. the colors, textures, and shapes contained in the images. Other types of data can also be used in similar fashion, eg. in a Web image search application, the embedding HTML page and the related hyperlink structure can also be utilized [5].

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 PicSOM, several SOMs are used in parallel, and the system is able to discover the ones providing most valuable information for each query instance. The goal is to autonomously adapt to the user's preferences regarding the similarity and relevance of images in the database. This can be obtained when the queries are iteratively refined as the system exposes more and more images to the user for evaluation and relevance feedback.

A typical retrieval session with PicSOM consists of a number of subsequent query rounds during which the retrieval is focused more accurately on images resembling the relevant reference images. While in normal human interaction mode, as well as in some previous experiments (eg. [6, 7, 4]), the system presents the user in the beginning of a new query the first set of reference images which have been picked uniformly from the whole database. However, in the experiments described in this paper, the queries are initiated with one image whose ground truth class is known in advance.

### 2.1. Tree Structured Self-Organizing Maps

The main image indexing method used in the PicSOM system is the Self-Organizing Map (SOM) [2]. 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 optimal accuracy using a restricted set of models. As the SOM algorithm organizes similar feature vectors in nearby neurons, the resulting map contains a representation of the database with similar images located near each other. Each feature descriptor is used separately to train its own SOM structure.

Instead of the standard SOM version, PicSOM uses a special form of the algorithm, namely the Tree Structured Self-Organizing Map (TS-SOM) [3]. 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 hierarchy and reduce the complexity 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. In the experiments, we have used TS-SOMs whose level sizes have been $4 \times 4$, $16 \times 16$, $64 \times 64$, and $256 \times 256$ units. In the training of the lower SOM levels, the search for the BMU has been restricted to the $10 \times 10$-sized neuron area below the BMU on the above level.

After training each TS-SOM level, that level is fixed and each map unit on it is given a visual label from the database image nearest to it. This hierarchical representation of the image database can also be utilized in visual browsing. The successive map levels can be regarded as providing increasing resolution for database inspection. When browsing the database, one can search for interesting images on one level and then descend to the SOM nodes below to see more similar images.

## 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 [7] for a more comprehensive treatment.

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 images at all. Of the above cases, 1) and 3) indicate that the corresponding content description agrees well with the user's conception on the relevance of the images. Whereas, case 2) is an indication that the content description 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 in PicSOM implemented 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 then zero. The values are then summed in the BMUs of the images and the resulting sparse value fields are low-pass filtered. Each image used as a visual label on the SOM surface is thus given a qualification value that depends on the local denseness of positive responses on the map and indirectly on the feature extraction's capability to reflect the user's view of image relevance.

In PicSOM, content descriptors that fail to coincide with the user's conceptions always 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 weighted explicitly as the system automatically takes care of weighting their opinions. In the actual implementation, we search on each SOM for a fixed number yet unseen visual labels with the highest qualification values. All duplicate images are then removed from these map-wise image sets and a preset number of the best of the remaining images are then shown to the user as the new reference images. In our earlier experiments, eg. [6, 7, 4], the visual labels of the SOM units on the whole TS-SOM structure were considered as candidate images to be shown to the user. Additionally, to ensure that the whole database is accessible, on the bottommost levels we gave to all the images mapped in each BMU equal precedence in the selection.

In the experiments to be described in Section 3, we have chosen to consider only the bottommost TS-SOM levels. Therefore, the visual labels of the units have no special role or precedence in the system and the TS-SOM hierarchy is neglected. From each SOM, 100 best-scoring images were taken into initial consideration. This change is motivated by the used performance evaluation scheme, in which the queries are always started with one image that certainly belongs to the specified image class. In this setting, starting the retrieval with a *depth first search* near the reference image is justifiable, instead of initial *breadth first search* in the whole database.

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 SOMs. 20 images with the highest total qualification values were then used as the result of the query round.

## 2.3. Vector-quantization-based reference method

In order to be able to compare PicSOM's performance to other systems, we have built an algorithmic alternative within our CBIR system. Here we motivate and describe the implementation of a simple vector-quantization-based alternative for using SOMs in implementing relevance feedback. There are naturally a wide range of techniques for indexing the images based on a feature descriptor. One method is to first use vector quantization (VQ) to prune the database and then to use a more exhaustive method to decide the final images to be returned.

In image retrieval, the justification for VQ is that unseen images which have fallen into the same quantization bins as the relevant-marked reference images are good candidates for the next images to be displayed. Also, the SOM algorithm can be seen as a special case of VQ. When using the model vectors of the SOM units in vector quantization, one ignores the topological order provided by the map lattice and characterize the similarity of two images only by whether they are mapped to the same VQ bin. For vector quantization, a well-known method is the $K$-means or Linde-Buzo-Gray algorithm [8]. According to our previous work [4], using $K$-means yields better performance than the SOM used as a vector quantizer. This is understandable as the SOM algorithm can be regarded as a trade-off between two objectives, namely clustering and topological order, and, by ignoring the topology, we dismiss a significant portion of the data organization provided by the SOM. Thus, we use $K$-means quantization in the reference system.

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 level 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 ($64\times64$). This results in 14.6 images per VQ bin, on the average, for the Corel database of 59 995 images described in Section 3.3. Another parameter is the number of candidate images taken into consideration from each of the parallel vector quantizers. In our implementation, we rank the VQ bins of each quantizer in the descending order given by the proportion of relevant images of all seen images in them. Then, we select 100 yet unseen images, as with the PicSOM method, from the bins in that order.

After the VQ stage, the set of potential images has been greatly reduced and more demanding processing techniques can be applied. One possible method – also applied in this reference system – is to rank the images based on their cumulative distances to all already found relevant images in the original feature spaces. Finally, as in the PicSOM method, we display 20 best-scoring images to the user. In [4], it was found out that the VQ method benefits from this extra processing stage. An alternative method, in which the distances are weighted according to the relative pairwise distances of found relevant images in the feature spaces, was also used in [4]. The motivation for this method is that if the average distance of two relevant images is small, the feature can be considered as well-suited for the current query. In the experiments described in this paper, this weighting did not improve the results and was therefore exluded from the experiment results.

# 3. Experiments

## 3.1. Performance evaluation

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 – 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.

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 class $\mathcal{C}$ is “shown” to the system one at a time as an initial image to start the query with. The system should then return similar images, as much as possible. This results in a leave-one-out type testing of the target class and the effective size of the test class becomes $N_{\mathcal{C}} - 1$ instead of $N_{\mathcal{C}}$ and the *a priori* probability of class $\rho_{\mathcal{C}} = (N_{\mathcal{C}} - 1)/(N - 1)$.

Precision $\mathcal{P}$ and recall $\mathcal{R}$ are intuitive performance measures that suite non-exhaustive use. When not the whole database but only a smaller number $N_T$ 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. The intermediate values of $\mathcal{P}(t)$ display the initial accuracy of the CBIR system and how relevance feedback is able to adapt to the class. It is to be expected that $\mathcal{P}(t)$ first increases and then turns to decrease when a notable fraction of the relevant images have already been shown. Furthermore, 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 above one now relate to retrieval performance that exceeds random browsing.

## 3.2. MPEG-7 content descriptors

MPEG-7 [10] is an ISO/IEC standard aiming 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. It is expected that MPEG-7 will have a similar prominent impact on multimedia content description as the previous MPEG standards on their respective application areas. The MPEG-7 standard – being aimed at describing still and live images and sound – defines many different content descriptors, of which only a part is applicable to still image content description. Regarding still images, it can be noted that MPEG-7 canonizes the old knowledge about color, texture, and shape being the three types of visual features applicable to automated still image content description.

As a nonnormative part of the standard, a software Experimentation Model (XM) [9] has been released in public use. The XM software is the framework for all the reference code of the MPEG-7 standard and it implements the normative MPEG-7 components. In the scope of our work, the most relevant part of XM is the implementation of a set of MPEG-7 defined image descriptors. At the time of writing this paper, XM is in its version 5.3 and not all description schemes have yet been reported to be working properly. Table 1 lists the visual descriptors of MPEG-7 applicable for still images and their current availability in the XM.

We have used a subset of MPEG-7 descriptors in a set of experiments with the PicSOM system and its VQ-based competitor. These descriptors are available in the XM software and summarized in Table 2. Not all of the descriptors are linear by nature and the MPEG-7 standard defines special metrics for the calculation of similarity between such descriptors. However, we use Euclidean metrics in comparing the descriptors as the training of the SOMs and the creation of the VQ 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* of XM is variable-sized, ie. its length depends 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.

## 3.3. 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 photographs and artificial images whose sizes are either 384×256 or 256×384 pixels. The images are grouped in thematic groups and keywords are also available. However, we found these image groups rather inconsistent with the keywords. Therefore, we created for the experiments three manually-picked ground truth image sets with tighter membership criteria. A single subject gathered all image sets. The used sets and membership criteria were:

**faces**, 1115 images (*a priori* probability 1.85%), where the main target of the image has to be a human head which has both eyes visible and the head has to fill at least 1/9 of the image area,

**cars**, 864 images (1.44%), where the main target of the image has to be a car, and 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, and

**planes**, 292 images (0.49%), where all airplane images have been accepted.

It can be hypothesized that shape-describing features would work well with all the three classes. On the other hand, only **faces** and **planes** would be well-describable on the basis of their color content.

## 3.4. Results

Our experiments were two-fold. First, we wanted to study which of the four color descriptors in Table 2 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 24 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**). Each

Table 1. XM's visual content descriptors for still images and their current availability.

| **Color Descriptors** | | **Texture Descriptors** | | **Shape Descriptors** | |
|---|---|---|---|---|---|
| Dominant Color | yes | Edge Histogram | yes | Region-Based Shape | yes |
| Scalable Color | yes | Homogeneous Texture | no | Contour-Based Shape | no |
| Color Layout | yes | Texture Browsing | no | Shape 3D | no |
| Color Structure | yes | | | | |
| GoF/GoP Color | no | | | | |

Table 2. The MPEG-7 visual content descriptors used in the experiments. $d$ is the dimensionality of the descriptor. The descriptors are implemented in [9].

| | |
|---|---|
| **Color Descriptors** | |
| *Dominant Color* $(d = 6)$ | This descriptor is a subset from the original MPEG-7 XM descriptor and is composed of the LUV color system values of the first and second most dominant color. If the XM routine only found one dominant color, it has been duplicated. |
| *Scalable Color* $(d = 256)$<br>Bins=256 | The descriptor is a 256-bin color histogram in HSV color space, which is encoded by a Haar transform. |
| *Color Layout* $(d = 12)$<br>#coeff(Y,Cb,Cr)=(6,3,3) | The image area is divided in 8×8 non-overlapping blocks where the dominant colors are solved in YCbCr color system. Discrete Cosine Transform (DCT) is then applied to the dominant colors in each channel and the coefficients of DCT are used as a descriptor. |
| *Color Structure* $(d = 256)$<br>Bins=256 | The image is presented in HMMD color system and quantized in 256 bins. A 8×8-sized structuring element is slid over the image and the numbers of positions where the element contains each quantized color are used as a descriptor. |
| **Texture Descriptors** | |
| *Edge Histogram* $(d = 80)$ | The image is divided in 4×4 non-overlapping sub-images where the relative frequencies of five different edge types (vertical, horizontal, 45°, 135°, non-directional) are calculated by using 2×2-sized edge detectors for the luminance of the pixels. The descriptor is obtained with a nonlinear mapping of the relative frequencies to discrete values. |
| **Shape Descriptors** | |
| *Region Shape* $(d = 35)$ | 35 Angular Radial Transform (ART) coefficients are calculated within a disk centered at the center of the image's Y channel. A nonlinear mapping is applied to the magnitudes of the complex ART coefficients and the outputs used as a descriptor. |

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 instant and finally averaged. 20 images were shown at each iteration round, which resulted in 50 query rounds as $N_T = 1000$. Recall and relative precision were recorded after each query iteration. Figure 1 shows the resulting recall–relative precision curves in which the averages of the recorded values are shown with symbols and connected with lines.

The following observations can be made from the plots of Figure 1: Overall, none of the tested color descriptors seems to dominate the other descriptors. Regardless of the used retrieval method, *Color Structure* seems to perform best with **faces**, and *Color Structure* and *Scalable Color* seem to be better than the others with **cars**. With **planes**, *Scalable Color* yields best results, especially when using the VQ method. In general, the PicSOM method usually seems to obtain better precision when comparing same descriptor sets. Also, in the end, PicSOM has in a majority of cases reached higher recall level. When using the best descriptor sets with VQ, one can achieve comparable results with PicSOM, but the difference between the best and the worst set of descriptors is larger with the VQ method. However, for some reason, VQ with *Scalable Color* descriptor seems to be initially very suitable for **planes**, exceeding all other results. Even in this case, however, PicSOM reaches a higher final recall value. Of the tested image classes, **cars** is clearly the most problematic one in the light of relative precision.

In the second set of experiments, we used all the available MPEG-7 descriptors simultaneously. Runs were made separately for the three image classes and the two CBIR techniques. The results can be seen in Figure 2 where each plot now contains recall–relative precision curves of the two techniques. It can be seen that in all cases PicSOM is at first behind of VQ in precision, but sooner or later reaches and exceeds it. In two cases (**faces** and **cars**), this overtake by PicSOM takes only one or two rounds of queries. In the third case (**planes**), reaching VQ takes a longer time, due to the good initial precision of VQ, observed also in Figure 1 with the *Scalable Color* descriptor. As the final outcome, in two cases, PicSOM is clearly superior to VQ both in recall and precision. In the third case, the interpretation of the outcome of the experiment depends on the viewpoint of the observer.

One can also compare the curves of Figures 1 and 2 for an important observation. It can be seen that the PicSOM method is, when using all descriptors (Figure 2), in all cases able to follow and even exceed the path of the best recall–relative precision curve for the four alternative color descriptors (Figure 1). This is 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 2 resemble more the average than the maximum value of the corresponding VQ curves in Figure 1. As a consequence, the VQ technique is clearly more dependent on the proper selection of used features than the PicSOM technique.

## 4. Conclusions

The forthcoming MPEG-7 standard does not solve the open questions of content-based image retrieval. Nor does it establish which visual descriptors will be used in future applications. Still, the impact of the standard on the development of CBIR will be considerable. As the standard enables the definition of new types of content descriptions, it will hopefully not restrict the development but only set the frames for it.

In this paper, we have briefly described our CBIR 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 easy to apply any kind of features. Due to PicSOM's ability to automatically weight and

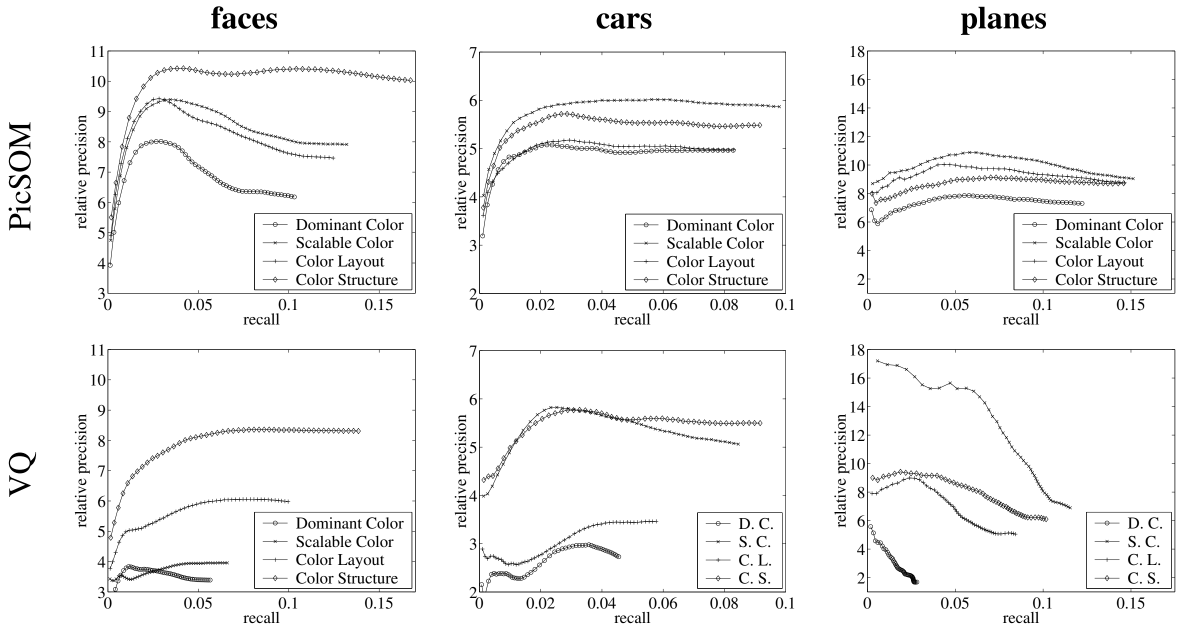

Figure 1. Recall–relative precision plots of the performance of different color descriptors and the two CBIR techniques. In all cases also *Edge Histogram* and *Region Shape* descriptors have been used.

combine the responses, one can make use of any number of content descriptors without needing 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 descriptions.

In the experiments we compared four different color descriptors available in the MPEG-7 XM software. The results showed that no single color descriptor was the best one for all of our three image classes. That result was no surprise, it merely emphasizes the need to use many parallel content descriptors. In an experiment where we used all the available color descriptors, PicSOM indeed was without supervision able to 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 most useful information for this particular

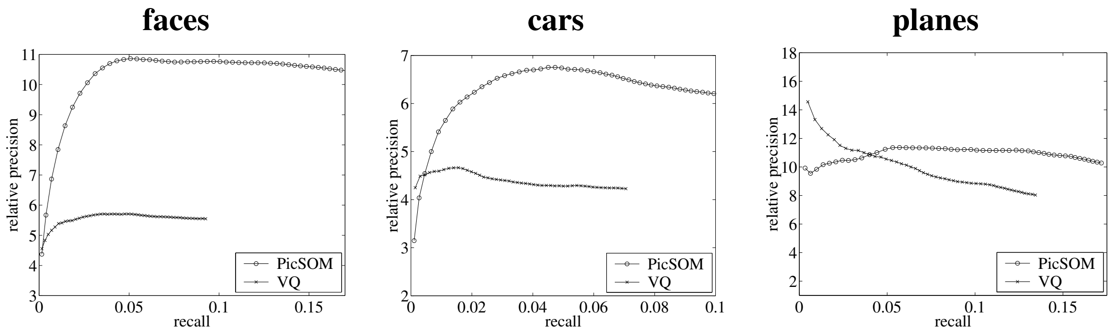

Figure 2. Recall–relative precision plots of the performance of the two CBIR techniques when all four color descriptors were used simultaneously together with the *Edge Histogram* and *Region Shape* descriptors.

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 the relevance feedback mechanism of PicSOM is clearly superior and the PicSOM method produced better final recall values in the experiments, even though the initial precision of VQ was in some cases higher.

As the MPEG-7 XM is not all mature yet, also our experiments are only partially finished. When more MPEG-7 standard content descriptors become implemented in the XM, we will continue the evaluations. Also, we will compare our earlier descriptors with those of the standard, perhaps finding a mixture of them that exceeds in performance both our original and the MPEG-7-defined descriptors alone.

# 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] Chang N-S and Fu K-S (1980) Query by pictorial example. IEEE Transactions on Software Engineering, 6(6):519–524.

[2] Kohonen T (2001) Self-Organizing Maps. Springer-Verlag.

[3] Koikkalainen P and Oja E (1990) Self-organizing hierarchical feature maps. Proc. IJCNN-90, International Joint Conference on Neural Networks, Washington, DC, 279–285.

[4] Koskela M, Laaksonen J and Oja E (2001) Comparison of techniques for content-based image retrieval. Proc. 12th Scandinavian Conference on Image Analysis (SCIA 2001), Bergen, Norway, 579–586.

[5] Laakso S, Laaksonen J, Koskela M and Oja E (2001) Self-Organizing Maps of Web Link Information. In: Allinson N, Yin H, Allinson L and Slack J (eds.) Advances in Self-Organising Maps. Springer.

[6] Laaksonen J, Koskela M, Laakso S and Oja E (2000) PicSOM - Content-based image retrieval with self-organizing maps. Pattern Recognition Letters, 21(13-14):1199–1207.

[7] Laaksonen J, Koskela M, Laakso S, and Oja E (2001) Self-organizing maps as a relevance feedback technique in content-based image retrieval. Pattern Analysis & Applications, 4(2+3):140–152.

[8] Linde Y, Buzo A and Gray R (1980) An algorithm for vector quantizer design. IEEE Transactions on Communications, 28(1):84–95.

[9] MPEG-7 (2001) MPEG-7 visual part of the eXperimentation Model (version 9.0), ISO/IEC JTC1/SC29/WG11 N3914.

[10] MPEG-7 (2001) Overview of the MPEG-7 standard (version 5.0), ISO/IEC JTC1/SC29/WG11 N4031.

[11] Salton G and McGill MJ (1983) Introduction to Modern Information Retrieval. McGraw-Hill.