PicSOM—Self-Organizing Image Retrieval With MPEG-7 Content Descriptors

Jorma Laaksonen, *Associate Member, IEEE*, Markus Koskela, and Erkki Oja, *Fellow, IEEE*

***Abstract*—Development of content-based image retrieval (CBIR) techniques has suffered from the lack of standardized ways for describing visual image content. Luckily, the MPEG-7, or formally "Moving Pictures Expert Group Multimedia Content Description Interface" international standard is now 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 CBIR. Our system is named PicSOM and it is based on pictorial examples and relevance feedback (RF). The name stems from "picture" and the self-organizing map (SOM). The PicSOM system is implemented by using tree structured SOMs. In this paper, we apply the visual content descriptors provided by MPEG-7 in the PicSOM system and compare our own image indexing technique with a reference system based on vector quantization (VQ). 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 RF mechanism of PicSOM begins to function, its retrieval precision exceeds that of the reference system.**

***Index Terms*—Content-based image retrieval (CBIR), MPEG-7, query by pictorial example (QBPE), relevance feedback (RF), self-organizing map (SOM), visual content description.**

I. Introduction

CONTENT-BASED image retrieval (CBIR) has been a subject of very intensive research effort for more than a decade [1]–[3]. 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 in interpreting their results. These problems have hindered the researchers from doing comprehensive evaluations of different CBIR techniques. Some noteworthy initiatives have recently been made to overcome these difficulties [4], [5].

In addition, two more points make CBIR systems special. Opposed to such computer vision applications as production quality control systems, operational CBIR systems would be very intimately connected to the people using them. Also, effective CBIR systems call for means of interchanging information concerning images' content between local and remote databases, a characteristic very seldom present, e.g., in industrial computer vision.

We have developed a neural-network-based CBIR system named PicSOM [6]–[8]. The name stems from "picture" and the self-organizing map (SOM). The SOM [9] is used for unsupervised, self-organizing, and topology-preserving mapping from the image descriptor space to a two-dimensional (2-D) lattice, or grid, of artificial neural units. The PicSOM system is built upon two fundamental principles or paradigms of CBIR, namely query by pictorial example (QBPE) [10] and relevance feedback (RF) [11]. In QBPE, 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. On the other hand, in RF 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. Anyhow, the image querying becomes an iterative process where the CBIR system is only a tool in the hands of a human expert. Our conviction is that effective CBIR systems can be built upon these four cornerstones: self-organization, pictorial examples, RF, and iterative interaction.

Until now, there have not existed widely accepted standards for description of the visual contents of images. MPEG-7 [12]–[14], or formally "Moving Pictures Expert Group Multimedia Content Description Interface," is the first thorough attempt in this direction and it will become an international standard of ISO/IEC. 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. Also, all CBIR developers should prepare to accept the challenge to apply their CBIR techniques to tasks that are expressed solely in the terms of the standard. For example, a benchmark study could be organized in which the systems would be given first a large set of MPEG-7 standard image content descriptions to build an image index of the CBIR system. Then, a separate, smaller sample of descriptions would be given and the system should find for each of them the best matching database images in the larger set. The standard also facilitates the change of content information between distributed databases. This brings about valuable benefits if database maintainers can trust on and make use of content descriptors calculated remotely by other database maintainers. In such a situation the actual images need not to be transferred between locations, nor do the descriptors to be recalculated. As MPEG-7 Experimentation Model (XM)

Manuscript received April 19, 2001; revised October 30, 2001. This work was supported by the Finnish Centre of Excellence Programme (2000–2005) of the Academy of Finland, project New Information Processing Principles 44886.

The authors are with the Laboratory of Computer and Information Science, Helsinki University of Technology, 02015 HUT, Finland. (e-mail: jorma.laaksonen@hut.fi; markus.koskela@hut.fi; erkki.oja@hut.fi).

Publisher Item Identifier S 1045-9227(02)04419-3.

[15] 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 nonstandard descriptors with those defined in the MPEG-7 standard and available in XM.

In this paper, we first address the general questions of content-based image retrieval, simultaneous indexing with multiple low-level visual features, pictorial examples, and RF in Section II. Then, we describe the PicSOM system in Section III and summarize the experiences we have collected this far. We also describe a reference CBIR system built within the PicSOM framework. This system will be used as a competitor of our own approach in the experiments. Section IV discusses the visual content descriptors defined in the MPEG-7 standard and their usability for content-based image retrieval. Then, in Section V, we present a series of experiments performed with the PicSOM system and the reference system by using the MPEG-7-defined visual content descriptors. Finally, conclusions are drawn together and future prospects are discussed in Section VI.

II. Content-Based Image Retrieval

In this section, we will briefly review some of the central concepts involved in content-based image retrieval.

A. Image Indexation With Low-Level Features

In a CBIR system implemented with prototype-based statistical methods, each image in the database is transformed with a set of feature extraction methods to a set of lower-dimensional prototype vectors in respective feature spaces. These features can describe, e.g., the colors, textures, and shapes contained in the images. Other types of data can also be used in the similar fashion. Additional useful data can include metadata or keywords describing the images, if available. In a Web image search application, the embedding HTML page and the related hyperlink structure may also be utilized to provide useful information [16].

When the CBIR system tries to find images which are similar to the relevant-marked reference images, it searches for images whose distance to the relevant images is in some sense minimal in any or all of the feature spaces. How the distances in various feature spaces are calculated, weighted, and combined in order to form a scalar value suitable for minimization, leaves a lot of room for different techniques.

The operation of a CBIR system can be interpreted as a series of more or less independent processing stages [8]. For each of these stages, there exist multiple choices and thus a multitude of CBIR systems can be implemented by combining a set of common building blocks.

B. QBPE

With low-level visual features, it is not possible to base a content-based image query on verbal terms like in text-based retrieval. Therefore, other query methods must be applied. One common approach to formulate queries in CBIR is QBPE [10]. With QBPE, the image queries are based on example images shown either from the database itself or some external location. The user classifies these example images as relevant or nonrelevant to the current retrieval task and the system uses this information to select such images the user is most likely to be interested in. Other possible query methods include queries based on user-drawn sketches and image icons representing some common elements found in the database images [3].

C. RF

A CBIR system is generally not able to retrieve the best available images in its first response. As a consequence, satisfactory retrieval results can be obtained only if the image query can be turned into an iterative and interactive process toward the desired image or images. The iterative refinement of a query is known as RF in information retrieval literature [11].

In text-based retrieval, RF can be implemented, e.g., by adjusting the weights of different textual terms when matching the query text with the documents of the database in a vectorial form. RF can be seen as a form of supervised learning to steer the subsequent query rounds by using the information gathered from the user's feedback. The role of the CBIR system is changed by RF from an automatic answering machine to a tool that is being used by a skillful human expert.

D. Multifeature Indexing

With the current state of image processing technology, image retrieval cannot generally be based on matching the user's query with the images in the database on an abstract conceptual level. Therefore, lower-level pictorial features need to be used. This creates the basic problem of CBIR: the gap between the high-level semantic concepts used by humans to understand image content and the low-level visual features used by a computer to index the images in a database. One method to tackle this issue is to use several visual features in parallel and combine their responses in an effective manner. A straightforward way of achieving this is to give appropriate weights to the different features. These weights should be automatically inferred as it is generally a difficult task to explicitly give low-level features such weights which would coincide with the human perception of images [17].

One serious problem with feature weighting is that such techniques treat the feature spaces globally rather than locally. However, a distance measure or feature weighting which is advantageous in the vicinity of a small set of images which are relevant and therefore similar to each other, may not produce favorable results for the rest of the images. Also, rules which are applicable in one part of the feature space are not as such generalizable to handle the whole space. All these phenomena are direct consequences of the inherent nonlinear nature of image similarity [18].

III. PicSOM System

The PicSOM image retrieval system [6]–[8] is a framework for research on algorithms and methods for CBIR. The genuine methodological novelty of PicSOM is to use several SOMs [9] 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 SOMs that provide the most valuable information for each individual 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 RF.

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 its normal human interaction mode, 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 to be described in Section V, the queries are initiated with one database image whose ground truth class is known in advance.

Sections III-A–F will give a brief overview of the PicSOM system. First, Section III-A explains the tree structured SOMs (TS-SOMs). Then, the core technology of RF in PicSOM is described in Section III-B. Next, a more conventional way of implementing RF, based on vector quantization (VQ) and used later as a reference system in the experiments, is presented in Section III-C. The remaining three sections will describe the different features used in PicSOM this far. A more detailed description of the PicSOM system and results of earlier experiments performed with it can be found in [7], [8]. The PicSOM home page including a working demonstration of the system for public access is located at http://www.cis.hut.fi/picsom.

A. Tree Structured SOMs

The main image indexing method used in the PicSOM system is the SOM [9]. 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 2-D grid. The map attempts to represent all the available observations with an optimal accuracy by 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 where similar images, according to the given feature, are located near each other.

The fitting of the model vectors is usually carried out by a sequential regression process, where $t = 0, 1, 2, \ldots, t_{\max} - 1$ is the step index: For each input sample $\mathbf{x}(t)$, first the index $c(\mathbf{x})$ of the best-matching unit (BMU) or the "winner" model $\mathbf{m}_{c(\mathbf{x})}(t)$ is identified by the condition

$$\forall\, i: \|\mathbf{x}(t) - \mathbf{m}_{c(\mathbf{x})}(t)\| \leq \|\mathbf{x}(t) - \mathbf{m}_i(t)\|. \quad (1)$$

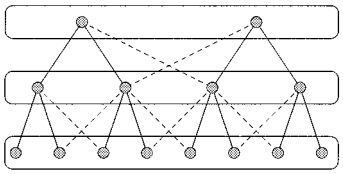

Fig. 1. The structure of a three-level 1-D TS-SOM. The solid lines represent parent–child relations and the dash lines represent neighboring nodes also included in the BMU search space.

The usual distance metric used here is the Euclidean one. After finding the BMU, a subset of the model vectors constituting a neighborhood centered around node $c(\mathbf{x})$ are updated as

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h(t; c(\mathbf{x}), i)(\mathbf{x}(t) - \mathbf{m}_i(t)). \quad (2)$$

Here $h(t; c(\mathbf{x}), i)$ is the "neighborhood function," a decreasing function of the distance between the ith and $c(\mathbf{x})$th nodes on the map grid. This regression is reiterated over the available samples and the value of $h(t; c(\mathbf{x}), i)$ is allowed to decrease in time to guarantee the convergence of the prototype vectors $\mathbf{m}_i$. The large values of the neighborhood function $h(t; c(\mathbf{x}), i)$ in the beginning of the training initialize the network and the small values on later iterations are needed in fine-tuning.

Instead of the standard SOM version, PicSOM uses a special form of the algorithm, the TS-SOM [19], [20]. 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. Unlike most tree-structured algorithms, the search space is not limited to the children of the BMU on the upper level. As each level of the TS-SOM is a normal SOM, the search space can be set to include also neighboring nodes having different parent nodes in the upper level. The structure of a TS-SOM in one-dimensional (1-D) case with three SOM levels illustrated in Fig. 1.

In our experiments in Section V, 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 Fig. 2, 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

Fig. 2. The surface of a 16 × 16-sized TS-SOM level trained with Edge Histogram descriptors.

bottom-left corner. Visually—and also semantically—similar images have thus been mapped near each other on the map.

The hierarchical representation of the image database produced by a TS-SOM 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 layer and then descend to the SOM nodes below the interesting ones to see more of similar images.

B. Self-Organizing RF

Relevance feedback has been implemented in PicSOM by using the parallel SOMs. 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; 2) relevant and irrelevant images mixed; 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. Fig. 3 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. 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 method'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 explicitly weighted as the system automatically takes care of weighting their opinions. In the actual implementation, we search on each SOM a fixed number, say 100, of yet unseen visual labels with 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 SOMs. Twenty images with the highest total qualification values have then been used as the result of the query round.

In our earlier experiments, e.g., [7], [8], [21], only the visual labels of the SOM units on all but the bottommost TS-SOM levels were considered as candidate images to be shown to

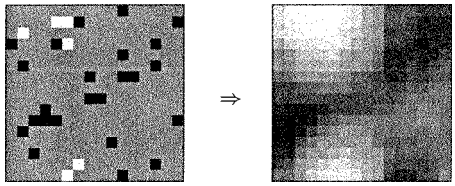

Fig. 3. An example of how a SOM surface, on which the images selected and rejected by the user are shown with white and black marks, respectively, are convolved with a low-pass filter.

the user. 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 V, we have now chosen to consider exclusively the bottommost TS-SOM levels. Therefore, the visual labels of the units have no special role or precedence in the system. This change is motivated by the performance evaluation scheme of Section V-A, in which the queries are always started with one image that certainly belongs to the specified image class. Therefore, one can choose to do a *depth first search* near the initial reference image instead of a *breadth first search* in the whole database.

C. VQ-Based Reference Method

In order to be able to compare PicSOM's performance to other systems, we have built some algorithmic alternatives within our CBIR system. Here we motivate and describe the implementation of a simple VQ-based alternative to using SOMs in implementing RF.

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 exist two alternate quantization techniques, namely scalar quantization (SQ) and VQ. With either of these techniques, the feature vectors are divided into subsets in which the vectors resemble each other. In the case of SQ the resemblance is with respect to one component of the feature vector, whereas resemblance in VQ means that the feature vectors are similar as whole. In our previous experiments [21], we have found out that SQ gives bad retrieval results.

The justification for VQ 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 VQ. When using the model vectors of the SOM units in VQ, one ignores the topological ordering provided by the map lattice and characterizes 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.

For VQ, a well-known method is the K-means or Linde–Buzo–Gray (LBG) vector quantization [22]. According to [21], 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 tradeoff between two objectives, namely clustering and topological ordering. Consequently, we will use LBG quantization in the reference system of the experiments in Section V.

The choice for the number of quantization bins is a significant parameter for the VQ algorithm. Using too few bins results in image clusters too broad 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 Corel database of 59 995 images to be described in Section V-B.

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* of *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 VQ 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. As calculating distance in a possibly very high-dimensional space is a computationally heavy operation, the VQ 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.

D. Visual Feature Maps

We have experimented with various types of low-level visual features. Detailed descriptions of these features can be found in [7], [8], [23]. In this work, we have restricted our feature selection to the ones defined in the MPEG-7 standard. These features will be described in Section IV.

The three common CBIR feature types, enumerated also in the MPEG-7 standard, are color, texture, and shape. For all these types, there exist several different ways to extract feature descriptions. Color features, for example, can be based on color histograms, averaged colors, finding the most common colors in an image according to some color quantization, etc. Also, the division into these three categories is not unambiguous, as some statistical features can be regarded likewise as texture or shape descriptions. One example of such a feature is the Edge Histogram.

The fourth type of visual features applicable to still images is image composition or structure. Capturing the composition of an image from salient objects, however, requires segmentation, which is a difficult task for unconstrained images. Furthermore, extending the discussion to video sequences would introduce time-dependent feature types such as motion.

E. Word Maps

In a specific database application, not connected to the experiments of this paper, we have also used TS-SOMs trained with textual keywords describing some aspects of the images. We collected from the Web a set of images together with the texts, or documents, that embedded the images. Then, after deletion of articles and other common words as well as words that appeared only in very few documents, we formed for each image a binary vector that indicated the presence or absence of each word in the corresponding document. Also, we formed similar presence vectors for all word pairs in the documents. Both types of vectors were quite high-dimensional, so we applied *random projection* [24] to reduce the dimensionality. The resulting vector sets were used in training the word and word pair TS-SOMs. This approach resembles to some extent the WEBSOM [25] system used for interactive browsing of large text document databases.

F. Web Link Relation Map

Recently, we have reported a method that utilizes the hypertext link structure of the World Wide Web [16]. The link features have been used to create a TS-SOM of images, similar to those created with the visual features. In the link relation map, images in neighboring map units are assumed to be semantically correlated not due to their visual similarity but due to their close mutual connection in the Web. The basis of the method consists of a set of basic relations that can take place between two images in the Web. For example, if one image acts as a hypertext link to another image (e.g., as a thumbnail) it can be assumed that the two images are closely related. Also, if two images are situated on the same Web page, it is very likely that they are somehow semantically related. For searching images in the Web, the link feature may thus be a valuable addition to the range of other low-level features.

In the realization of the link relation map, we used SHA-1 message digest algorithm [26] for dimension reduction by random mapping. The URLs of each image and the Web page where the image was found were considered, as well as all URL links found on the page. These URLs and the directory, host, and domain parts of them were first transformed with SHA-1 to pseudorandom numbers of length 32 bits. These bit sequences were then interpreted as concatenations of four eight-bit values. The first eight-bit value was used as an index in the range [0, 255], the second in [256, 511], the third in [512, 767], and the last one in [768, 1024]. The corresponding components in a 1024-dimensional, otherwise zero-valued, sparse vector were then set to value one.

There were 2 622 472 unique URLs or parts of them in the image pages of our Web image database of 1 008 844 images. As the outcome of the SHA-1 digestion, we had that same count, i.e., 2 622 472, 1024-dimensional vectors. We regarded this set of random vectors as an almost orthogonal basis. Each image was then represented by a 1024-dimensional link feature vector obtained as a combination of all the pseudorandom vectors associated with the image and its Web page. These vectors were then used in training a TS-SOM.

IV. MPEG-7

MPEG-7 [12]–[14] 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 application 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 descriptors as well as structures of descriptors and their relationships.

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. Nowadays, audiovisual material is becoming more common and widely used as the needed technologies are becoming easier to use and more available. This development has raised the need for quick and efficient searching techniques for all kinds of multimedia material. MPEG-7 is developed and supported by a wide range of professionals from publishers and digital content creators to intellectual property rights managers, as well as university researchers.

MPEG-7 defines a set of fundamental concepts. Descriptors are used to represent audio–visual features. Descriptors define the syntax and semantics of each feature representation. A single feature, such as color, texture, or shape, may have several descriptors representing different relevant aspects. Description schemes (DSs) specify the structure and semantics of relations between their components, which can be either descriptors or other DSs. Descriptors and DSs are divided into MPEG-7 Visual and MPEG-7 Audio description tools. Generic description tools (descriptors and DSs) which describe neither purely visual data nor audio are referred as multimedia DSs (MDSs). Finally, the description definition language (DDL) is used to specify the existing descriptors and DSs and for defining new ones. DDL is based on W3Cs XML Schema definition language [27].

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. Table I lists the feature types and their applicability to different tasks [13]. It can be seen that MPEG-7 canonizes the old knowledge about color, texture, and shape being the three different types of visual features applicable to automated still image content description.

As a nonnormative part of the standard, a software experimentation model (XM) [15] has been released for public use. The XM software is the framework for all the reference code of the MPEG-7 standard. It implements the normative MPEG-7 components such as descriptors, DSs, and the DDL. 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 this writing, XM is in its version 5.3 and not all descriptors

TABLE I
FEATURE TYPES DEFINED BY MPEG-7 AND THEIR USAGE AREAS [13]

feature type	still images	video	audio
Time		×	×
Shape	×	×	
Color	×	×	
Texture	×		
Motion		×	
Camera motion		×	
Mosaic		×	
Audio features		×	×

have yet been reported to be working properly. Table II lists the visual descriptors applicable for still images and their current availability in the XM.

A set of key or elementary application types are also implemented in the XM software. These include an application for description extraction from media, a search & retrieval application, a media transcoding application, and a description filtering application. Regarding an image retrieval application, the two first application types are clearly the most relevant. The extraction from media application is used to extract descriptions from media input data, i.e., in this case, still images. The search and retrieval application is a simple single-round retrieval application implemented in the XM. The application takes a query description and all descriptions of a media database as input and returns the resulting distance values to the best-matching items in the database with decreasing similarity to the query. As such, the key applications are not suitable as real-world applications. For example, RF or other query improvement is not possible as the key applications do not support user interaction during run time.

V. EXPERIMENTS

This section first addresses the question of performance evaluation in Section V-A. Then, Section V-B describes the image database and ground truth classes we have used in the experiments. Next, Section V-C gives the details of the MPEG-7 visual content descriptors used in the study. Finally, Section V-D presents the results of a comparison involving the MPEG-7 content descriptors. Also, our original PicSOM approach in CBIR is compared with a reference system based on VQ as described in Section III-C.

A. Performance Measures and Evaluation Scheme

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 nonoverlapping, can be formed and then used in the evaluation.

TABLE II
AVAILABILITY OF XM'S VISUAL CONTENT DESCRIPTORS APPLICABLE FOR STILL IMAGES

Color Descriptors	
Dominant Color	available
Scalable Color	available
Color Layout	available
Color Structure	available
GoF/GoP Color	not yet available
Texture Descriptors	
Edge Histogram	available
Homogeneous Texture	not yet available
Texture Browsing	not yet available
Shape Descriptors	
Region-Based Shape	available
Contour-Based Shape	not yet available
Shape 3D	not yet available

All features can be studied separately and independently from others for their capability to map visually similar images near each other. Such an analysis should account both for local and global clustering of image classes as was done, e.g., in [28]. 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 RF mechanism into account. One may note that this type of an approach resembles the search and retrieval application implemented in the MPEG-7 XM.

The evaluation of an entire CBIR system can be mathematically formulated as follows. Let the "correctness" of the outputs of a CBIR system be expressed by a series $\{h_1, h_2, \ldots, h_t, \ldots, h_N\}$ where N stands for the size of the database and $h_t = h(I_t)$ is the Boolean membership value of image I_t in the studied image class $\mathcal{C}$, i.e.,

$$h_t = h(I_t) = \begin{cases} 1, & \text{if } I_t \in \mathcal{C} \\ 0, & \text{if } I_t \notin \mathcal{C}. \end{cases} \tag{3}$$

The series $\{I_1, I_2, \ldots, I_N\}$ specifies the order in which the system presents the images I_t to the user for acceptance or rejection based on their relevance in the query. It is naturally supposed that the verbal class membership criterium and thus also the correctness function $h(I)$ is independent of the presentation order of the images. If the class $\mathcal{C}$ contains $N_{\mathcal{C}}$ images, it holds that

$$N_{\mathcal{C}} = \sum_{t=1}^{N} h_t. \tag{4}$$

The above notation does not make explicit the possibility that the CBIR system may show the user more than one image simultaneously. We assume that the user still glances at these images in the order the system has selected them. This may not always be true, but it simplifies the notations to some extent. If it is chosen that the system shows more than one image at one time, the RF mechanism gets lagged as it does not receive feedback after every successive image but only between consecutive image sets. The size of these image sets can be denoted with N^*.

A simple scalar describing the performance of a CBIR system for a given image class $\mathcal{C}$ can be formed as

$$\tau_{\mathcal{C}} = \frac{1}{NN_{\mathcal{C}}} \sum_{t=1}^{N} th_t \in \left[\frac{\rho_{\mathcal{C}}}{2}, 1 - \frac{\rho_{\mathcal{C}}}{2}\right]. \quad (5)$$

Here, $\rho_{\mathcal{C}}$ is the *a priori* probability of class $\mathcal{C}$, i.e., $\rho_{\mathcal{C}} = N_{\mathcal{C}}/N$. The τ measure coincides with the question "how large portion of the whole database needs to be browsed through until, on the average, the searched image will be found." It can be noted that the $\tau_{\mathcal{C}}$ value can be solved with one pass for the whole class $\mathcal{C}$, i.e., it does not need to be repeated over each image in the class. We have employed this measure in some of our earlier experiments, e.g., [7], [8], [21].

The above kind of evaluation setting becomes, however, meaningless if the size of the database is so large that it is anyway beyond human limits to browse through it exhaustively. In such cases, it must be supposed that the database will not contain just a single best match to the user's request, but that many images will be sufficiently close to what is being searched for. We can therefore assume that there is a count N_T of images the user is willing or has the time for to browse. The system should thus demonstrate its talent within this number of images.

In our current experiments, we have applied this type of nonexhaustive performance evaluation. 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 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 $\rho_{\mathcal{C}} = (N_{\mathcal{C}} - 1)/(N - 1)$. We have chosen to show the evolution of *precision* as a function of *recall* during the iterative image retrieval process.

Recall $\mathcal{R}$ expresses how large portion of the relevant image class has already been shown up to time instance t

$$\mathcal{R}(t) = \frac{\sum_{i=1}^{t} h_i}{N_{\mathcal{C}} - 1} \in [0,1], \quad t = 1, 2, \ldots, N_T. \quad (6)$$

Precision $\mathcal{P}$ indicates the accuracy of retrieval, i.e., how exclusively only relevant images have been retrieved

$$\mathcal{P}(t) = \frac{\sum_{i=1}^{t} h_i}{t} \in [0,1], \quad t = 1, 2, \ldots, N_T. \quad (7)$$

Precision and recall are intuitive performance measures that suite also for the case of nonexhaustive 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 RF mechanism is able to adapt to the class. With an effective RF 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.

B. Database and Ground Truth Classes

We have used images from the Corel Gallery 1 000 000 product [29] 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 rather inconsistent with the keywords. Therefore, we 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 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.
- **planes**, 292 images (0.49%), where all airplane images have been accepted.
- **sunsets**, 663 images (1.11%), where the image has to contain a sunset with the sun clearly visible in the image.
- **houses**, 526 images (0.88%), where the main target of the image has to be a single house, not severely obstructed, and it has to fill at least 1/16 of the image area.
- **horses**, 486 images (0.81%), where the main target of the image has to be one or more horses, shown completely in the image.

C. Content Descriptors

We have used a subset of MPEG-7 content descriptors for still images [15], [30] in a set of experiments with the PicSOM system and its VQ-based competitor of Section III-C. These descriptors were available and working in the XM [15] software of MPEG-7 and they are summarized in Table III.

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 VQ prototypes are based on minimizing a square-form error criterium. Only in the

TABLE III
THE MPEG-7 VISUAL CONTENT DESCRIPTOR USED IN THE EXPERIMENTS. d IS THE DIMENSIONALITY OF THE DESCRIPTOR. THE DESCRIPTOR HAVE BEEN DEFINED IN [30] AND IMPLEMENTED IN [15]

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 ($d = 256$) Bins=256	The descriptor is a 256-bin color histogram in HSV color space, which is encoded by a Haar transform.
Color Layout ($d = 12$) #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 used as s descriptor.
Color Structure ($d = 256$) Bins=256	The image is presented in HMMD color systemand 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) [30] 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.

case of the 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 the Dominant Color descriptor.

D. Results

Our experiments were two-fold. First, we wanted to study which of the four color descriptors in Table III would be the best one to be used together with the one texture and one shape descriptor in the table. Second, we wanted to compare the performance of our PicSOM system with that of the VQ-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. Twenty images were shown at each iteration round, i.e., $N^* = 20$, which resulted in 50 rounds when N_T was set to 1000 images. Both recall and relative precision were recorded after each query iteration. Fig. 4 shows, as a representative selection, the recall-relative precision curves of three of the studied image classes (faces, cars, and planes). Qualitatively similar behavior would be 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),

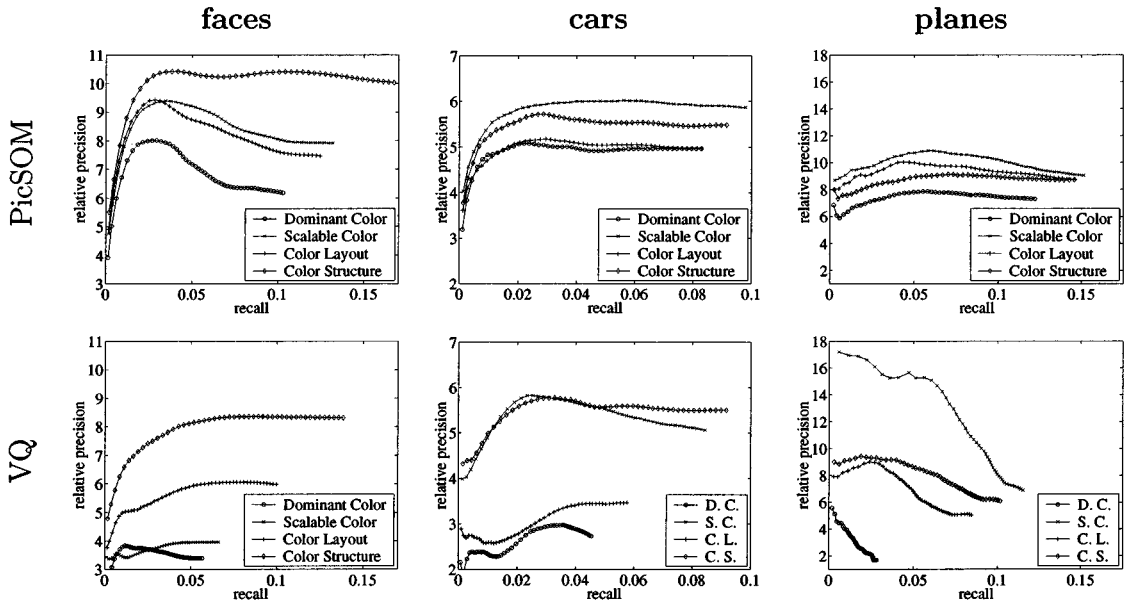

Fig. 4. Recall-relative precision plots of the performance of different color descriptors and the two CBIR techniques. In all cases also Edge Histogram and Region Shape descriptors have been used.

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 than 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 Ds is larger with the VQ method than with PicSOM. This can be observed, e.g., in the planes column of Fig. 4.

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 Fig. 5, where each plot now contains mutually comparable recall-relative precision curves of the two techniques.

It can be seen in Fig. 5 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 Fig. 4 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, 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 RF mechanism of PicSOM is clearly superior to that of VQs. The VQ retrieval has good initial precision but after a few rounds, when PicSOMs RF 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 Fig. 4 and the curves in the upper row of Fig. 5 for an important observation. It can be seen that the PicSOM method is, when using all Ds simultaneously (Fig. 5), able to follow and even exceed the path of the best recall-relative precision curve for the four alternative single color Ds (Fig. 4). This behavior is present in all cases, also with the image classes not shown in Fig. 4, 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 Fig. 5 resemble more the average than the maximum value of the corresponding VQ curves in Fig. 4. As a consequence, the VQ technique is clearly more dependent on the proper selection of used features than the PicSOM technique.

As a final illustration, Fig. 6 shows how the ground truth classes are distributed on the 256×256-sized bottom levels of the six different TS-SOM maps. The distributions are in conformance with our earlier observations concerning Figs. 4 and 5: 1) Of the six classes, sunsets, and planes are clearly best concentrated in only some specific map areas, faces and horses exhibit lesser level of concentration while cars and especially houses are very badly spread. 2) Of the four-color descriptors, Scalable Color is the best one in two cases (planes and horses) and Color Structure is the best one for faces. All the color descriptors cluster the sunsets class well, whereas none of them performs well with cars and

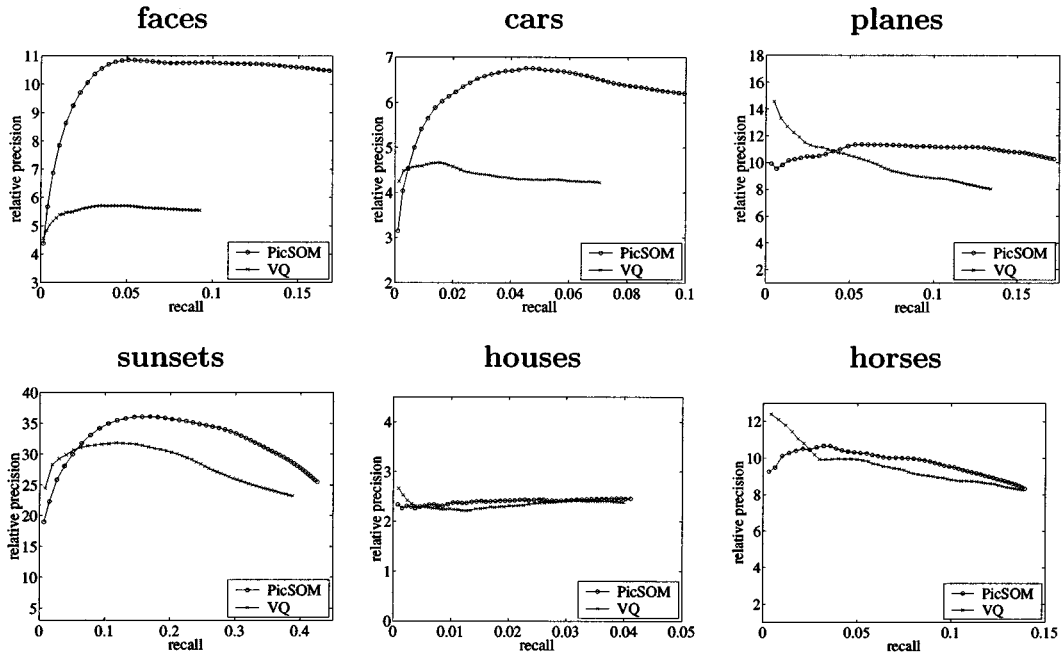

Fig. 5. Recall-relative precision plots of the performance of the two CBIR techniques when all four-color Ds were used simultaneously together with the edge histogram and region shape Ds.

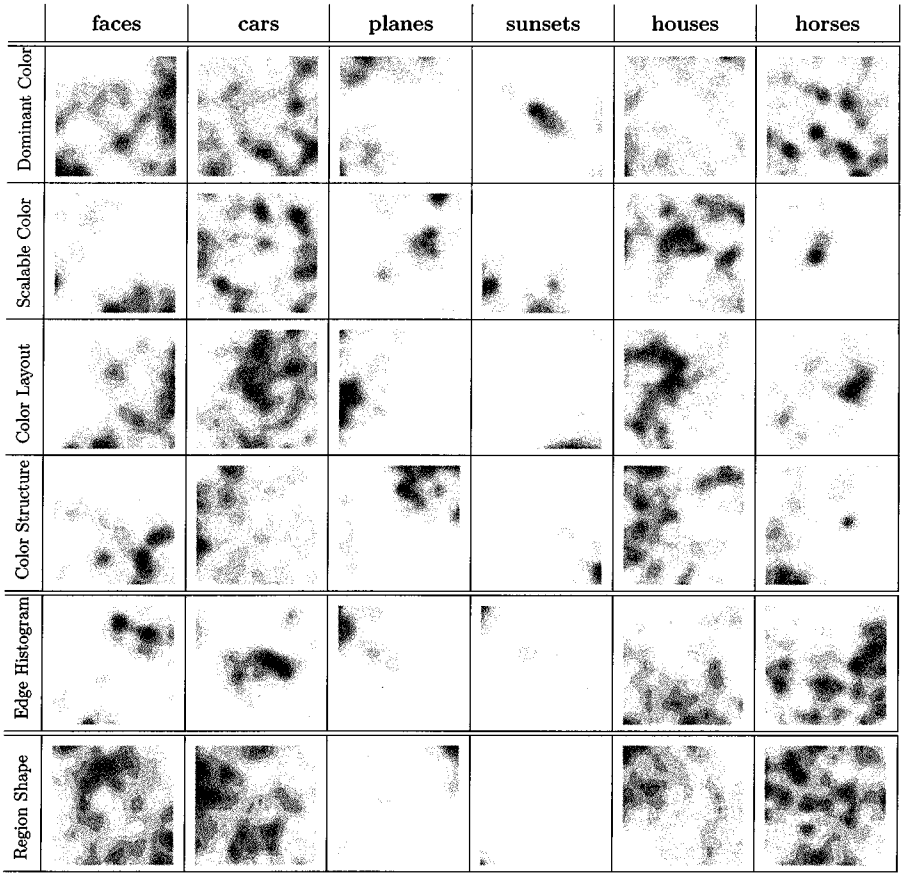

Fig. 6. The distributions of the image classes on the bottom levels of the six TS-SOM maps. The distributions have been low-pass filtered in order to ease the inspection. Darker shades present map areas where the images of the class have been mapped densely.

houses. In addition, one can see that the Edge Histogram texture descriptor is better than other descriptors for faces and cars, whereas the Region Shape descriptor produces the best clusterings for the planes and sunsets classes. Also, one notifies that the sunsets class is an easy one not only for the color but also for the texture and shape features as well.

Overall, convolved class distribution images such as in Fig. 6 are valuable visualizations of the performance of different feature extraction techniques for different image classes.

VI. Conclusion and Future Plans

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

In this paper, we have described our self-organizing 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 SOMs in implementing RF 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 PicSOMs 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 XM 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. This was also confirmed by visual inspection of the distributions of the image classes on the SOMs. That result was no surprise, and it merely emphasizes the need to use simultaneously 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 system focuses on the descriptors which provide the most useful information for the particular query instance.

We also compared the performance of the self-organizing RF technique of PicSOM with that of a VQ-based reference system. The results showed that in the beginning of queries, PicSOM starts with a bit lower precision rate. Later, when its strong RF 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.

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.

References

[1] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Visual Commun. Image Representation*, vol. 10, no. 1, pp. 39–62, Mar. 1999.

[2] A. Del Bimbo, *Visual Information Retrieval*. San Mateo, CA: Morgan Kaufmann, 1999.

[3] M. S. Lew, Ed., *Principles of Visual Information Retrieval*. New York: Springer-Verlag, 2000.

[4] C. H. C. Leung and H. H. S. Ip, "Benchmarking for visual information search," in *Proc. 4th Int. Conf. Visual Inform. Syst. (VISual 2000)*, Lyon, France, Nov. 2000, pp. 442–456.

[5] N. J. Gunther and G. Beretta. (2000) A Benchmark for Image Retrieval Using Distributed Systems Over the Internet: BIRDS-I. HP Labs. [Online]. Available: http://www.hpl.hp.com/techreports/2000/HPL-2000-162.html

[6] J. Laaksonen, M. Koskela, and E. Oja, "PicSOM—A framework for content-based image database retrieval using self-organizing maps," in *Proc. 11th Scandinavian Conf. Image Anal. (SCIA99)*, Kangerlussuaq, Greenland, June 1999, pp. 151–156.

[7] J. T. Laaksonen, J. M. Koskela, S. P. Laakso, and E. Oja, "PicSOM Content-based image retrieval with self-organizing maps," *Pattern Recognition Lett.*, vol. 21, no. 13–14, pp. 1199–1207, Dec. 2000.

[8] J. Laaksonen, M. Koskela, S. Laakso, and E. Oja, "Self-organizing maps as a relevance feedback technique in content-based image retrieval," *Pattern Anal. Applicat.*, vol. 4, no. 2 and 3, pp. 140–152, June 2001.

[9] T. Kohonen, *Self-Organizing Maps*, 3rd ed. New York: Springer-Verlag, 2001, vol. 30, Springer Series in Information Sciences.

[10] N.-S. Chang and K.-S. Fu, "Query by pictorial example," *IEEE Trans. Software Eng.*, vol. 6, pp. 519–524, Nov. 1980.

[11] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, ser. Computer Science. New York: McGraw-Hill, 1983.

[12] The Moving Picture Experts Group MPEG Home Page [Online]. Available: http://www.cselt.it/mpeg/

[13] *Overview of the MPEG-7 Standard (Version 5.0)*, ISO/IEC JTC1/SC29/WG11 N4031, March 2001.

[14] *IEEE Trans. Circuits Syst. Video Technol. (Special Issue on MPEG-7)*, vol. 11, no. 6, June 2001.

[15] *MPEG-7 Visual Part of the Experimentation Model (Version 9.0)*, ISO/IEC JTC1/SC29/WG11 N3914, Jan. 2001.

[16] S. Laakso, J. Laaksonen, M. Koskela, and E. Oja, "Self-organizing maps of web link information," in *Advances in Self-Organizing Maps*, N. Allinson, H. Yin, L. Allinson, and J. Slack, Eds. London, U.K.: Springer-Verlag, 2001, pp. 146–151.

[17] R. W. Picard, T. P. Minka, and M. Szummer, "Modeling User Subjectivity in Image Libraries," M.I.T Media Lab., 382, 1996.

[18] S. Santini and R. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 1–13, Sept. 1999.

[19] P. Koikkalainen and E. Oja, "Self-organizing hierarchical feature maps," in *Proc. IJCNN-90 Int. Joint Conf. Neural Networks*, vol. II, Washington, DC, 1990, pp. 279–285.

[20] P. Koikkalainen, "Progress with the tree-structured self-organizing map," in *Proc. 11th Europ. Conf. Artificial Intell.*, A. G. Cohn, Ed., Aug. 1994.

[21] M. Koskela, J. Laaksonen, and E. Oja, "Comparison of techniques for content-based image retrieval," in *Proc. 12th Scandinavian Conf. Image Anal. (SCIA 2001)*, Bergen, Norway, June 2001, pp. 579–586.

[22] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[23] S. Brandt, J. Laaksonen, and E. Oja, "Statistical shape features in content-based image retrieval," J. Math. Imaging Vision, 2002, to be published.

[24] S. Kaski, "Dimensionality reduction by random mapping: Fast similarity method for clustering," in Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN98), Anchorage, AK, May 1998.

[25] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive text document collection," *IEEE Trans. Neural Networks*, vol. 11, pp. 574–585, May 2000.

[26] Secure Hash Standard (1995). [Online]. Available: http://www.itl.nist.gov/fipspubs/fip180-1.htm

[27] XML Schema Part 0: Primer, W3C Recommendation (2001, May). [Online]. Available: http://www.w3.org/TR/xmlschema-0/

[28] E. Oja, J. Laaksonen, M. Koskela, and S. Brandt, "Self-organizing maps for content-based image retrieval," in *Kohonen Maps*, E. Oja and S. Kaski, Eds: Elsevier, 1999, pp. 349–362.

[29] The Corel Corporation World Wide Web Home Page (1999). [Online]. Available: http://www.corel.com

[30] *Text of ISO/IEC 15 938-3/FCD Information Technology—Multimedia Content Description Interface—Part 3 Visual*, ISO/IEC JTC1/SC29/WG11/N4062, Mar. 2001.

Jorma Laaksonen (S'96–A'97) received the Doctor of Science in Technology degree in 1997 from Helsinki University of Technology (HUT), Finland.

He is presently Senior Research Scientist at the Laboratory of Computer and Information Science, HUT. He is an author of several journal and conference papers on pattern recognition, statistical classification, and neural networks. His research interests are in content-based image retrieval and recognition of handwriting.

Dr. Laaksonen is a Founding Member of the SOM and LVQ Programming Teams and the PicSOM Development Group and a member of the International Association of Pattern Recognition (IAPR) Technical Committee 3: Neural Networks and Machine Learning.

Markus Koskela received the M.Sc.(Tech) degree in 1999 from Helsinki University of Technology (HUT), Finland. He is currently pursuing the D.Sc. degree at the Laboratory of Computer and Information Science, HUT.

His research interests are in content-based image retrieval, neural networks, and image processing.

Erkki Oja (S'76–M'77–SM'90–F'00) received the Dr.Sc. degree in 1977 from Helsinki University of Technology, Finland.

He has been Research Associate at Brown University, Providence, RI, and Visiting Professor at Tokyo Institute of Technology. He is Director of the Neural Networks Research Centre and Professor of Computer Science at the Laboratory of Computer and Information Science, Helsinki University of Technology. He is the author or coauthor of more than 240 articles and book chapters on pattern recognition, computer vision, and neural computing, and three books: *Subspace Methods of Pattern Recognition* (New York: RSP and Wiley, 1983), which has been translated into Chinese and Japanese, *Kohonen Maps* (Amsterdam, The Netherlands: Elsevier, 1999), and *Independent Component Analysis* (Ne York: Wiley, 2001). His research interests are in the study of principal components, independent components, self-organization, statistical pattern recognition, and applying artificial neural networks to computer vision and signal processing.

Dr. Oja is a Member of the editorial boards of several journals and has been in the program committees of several recent conferences including ICANN, IJCNN, and ICONIP. He is a Member of the Finnish Academy of Sciences, Founding Fellow of the International Association of Pattern Recognition (IAPR), and President of the European Neural Network Society (ENNS).