Using MPEG-7 Descriptors in Image Retrieval with Self-Organizing Maps

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

Abstract

The MPEG-7 standard is emerging as both a general framework for content description and a collection of specific, agreed-upon content descriptors. We have developed a neural, self-organizing technique for content-based image retrieval. In this paper, we apply the visual content descriptors provided by MPEG-7 in our PicSOM system and compare our own image indexing technique with a reference method based on vector quantization. The results of our experiments show that the MPEG-7 descriptors can be used as such in the PicSOM system.

1. Introduction

Development of content-based image retrieval (CBIR) techniques has suffered from the lack of standardized ways for describing image content. Until now, there have not existed widely-accepted standards for visual content description. MPEG-7 [7] – or "Moving Pictures Expert Group Multimedia Content Description Interface" – is the first thorough attempt in this direction. As the MPEG-7 Experimentation Model (XM) [6] has become available, we have been able to test the suitability of MPEG-7-defined image content descriptors with our PicSOM system. We have thus replaced our earlier, non-standard descriptors with those defined in MPEG-7. In this paper, we present a set of experiments with MPEG-7 descriptors and the PicSOM system.

Most current CBIR systems are based on visual low-level features and *query by example*, where the user specifies her object of interest by pointing out examples of relevant images. As CBIR systems are normally not capable of returning the desired image as their first response, *relevance feedback* [8] is used to improve the results. In relevance feedback it is assumed that the system is able to learn the user's preferences after seeing enough examples of relevant and irrelevant images. This kind of behavior can be implemented by allowing the user to evaluate the outputs of the system.

2. PicSOM system

The PicSOM image retrieval system [3, 4] is a framework for research on methods for content-based image retrieval. The PicSOM home page including a working demonstration of the system for public access is located at *http://www.cis.hut.fi/picsom*.

The main image indexing method used in PicSOM is the Self-Organizing Map (SOM) [1]. The SOM is used for unsupervised, self-organizing, and topology-preserving mapping from the image descriptor space to a two-dimensional lattice, or grid, of artificial neural units. The map attempts to represent the data with an optimal accuracy by using a restricted set of models. Multiple SOMs are used 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. After the training phase, the map units are connected with the images of the database. Therefore, the different SOMs impose different similarity functions on the images. As every image query is unique and each user of a CBIR system has her own transient view of image similarity, a system structure capable of holding many simultaneous similarity representations is desirable as it 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.

Instead of the standard SOM, PicSOM uses a special form of the algorithm, the Tree Structured Self-Organizing Map (TS-SOM) [2]. The hierarchical TS-SOM structure is useful for two purposes. First, it reduces the complexity of training large SOMs by exploiting the hierarchy in finding the best-matching unit for an input vector. Second, as each layer of the TS-SOM is a normal SOM, this hierarchical representation of the image database can be utilized in visual browsing. In the experiments described in this paper, we have used four-level TS-SOMs whose layer sizes have been 4×4, 16×16, 64×64, and 256×256 units.

2.1. Self-organizing relevance feedback

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 towards the desired image or images. In PicSOM, each image seen by the user is graded by her as either relevant or irrelevant. All these images and their associated relevance grades are then projected on all the SOM surfaces. This process forms on the maps areas where there are 1) many relevant images mapped in same or nearby SOM units, or 2) relevant and irrelevant images mixed, or 3) only irrelevant images, or 4) no graded images at all. Of the above cases, 1) and 3) indicate that the corresponding content descriptor agrees well with the user's conception on the relevance of the images.

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 done as follows. All relevant images are first given equal positive weight inversely proportional to the number of relevant images. Irrelevant images receive negative weights in similar fashion. The overall sum of these relevance values is thus zero. For each SOM layer, the values are then mapped to and summed in the SOM units in which the distance from the unit's weight vector to the image's feature vector is at minimum. Finally, the resulting sparse value fields are low-pass filtered to produce qualification values for each SOM unit and its associated images. This process is illustrated in Figure 1.

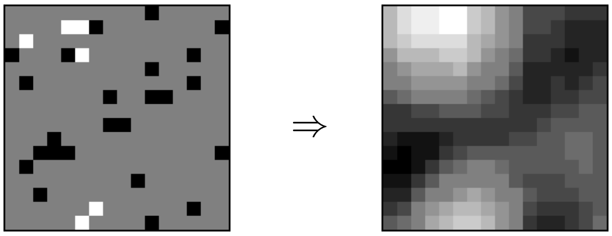

Figure 1. A SOM surface, on which the images selected and rejected by the user are shown with white and black marks, respectively, is convolved with a low-pass filter.

In these experiments, we have considered exclusively the bottommost (256×256) TS-SOM levels. This approach 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.

Content descriptors that fail to coincide with the user's conceptions produce lower qualification values than descriptors matching the user's expectations. Therefore, the different descriptors do not need to be explicitly weighted as the system automatically takes care of weighting their opinions. In the actual implementation, we search on each SOM for a fixed number, here 100, of unseen images with the highest qualification values. After removing duplicates, 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.

2.2. Vector-quantization-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. The justification for vector quantization in image retrieval is that unseen images which have fallen into the same quantization bins as the relevant-marked reference images are good candidates for the next reference images to be displayed to the user. Here we used the well-known K-means vector quantization [5].

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

After the vector quantization stage, the set of potential images has been greatly reduced and more demanding processing techniques can be applied to all the remaining candidate images. Now, one possible method – also applied in our reference system – is to rank the images based on their properly-weighted cumulative distances to all already-found relevant images in the original feature space. As with PicSOM, 20 highest-scoring images are then returned as the result of the query round.

3. Experiments

The performance of a CBIR system can be evaluated in many different ways. Even though the interpretation of the contents of images is always casual and ambiguous, some kind of ground truth classification of images must be performed in order to automate the evaluation process. In the simplest case – employed also here – image classes are formed by first setting verbal criteria for membership in a class and then assigning a Boolean membership value for each image in the database.

3.1. Performance measures

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 reference image to start the query with. The system should then return similar images (ie. images belonging to the same class), as much as possible. This results in a leave-one-out type testing of the target class.

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 both the initial accuracy of the CBIR system and how the relevance feedback mechanism is able to adapt to the class. With relevance feedback, 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 been shown. Furthermore, we have normalized the precision value by dividing it with the *a priori* probability 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. Used database and image classes

We have used images from the Corel Gallery 1 000 000 product in our evaluations. The database contains 59 995 JPEG images. 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 were **faces** (1115 images, *a priori* probability 1.85%), **cars** (864 images, 1.44%), **planes** (292 images, 0.49%), **sunsets**, (663 images, 1.11%), **houses** (526 images, 0.88%), and **horses**, (486 images, 0.81%).

3.3. MPEG-7 content descriptors

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

The MPEG-7 standard defines not only the descriptors but also special metrics for calculating the similarity between images. However, we use Euclidean metrics in comparing the descriptors because the training of the SOMs and the creation of the vector quantization prototypes are based on minimizing a square-form error criterium. Only in the case of *Dominant Color* descriptor this has necessitated a slight modification in the use of the descriptor. Since the original *Dominant Color* descriptor of XM is variable-sized and 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.

3.4. Results

We performed 12 sets of computer runs which were characterized by the used image class (**faces / cars / planes / sunsets / houses / horses**) and the retrieval method (PicSOM / VQ). Each run was repeated as many times as there were images in the image class. The recall and relative precision values were recorded for each such instant and finally averaged. 20 images were shown at each iteration round, which resulted in 50 rounds when N_T was set to 1000 images. Both recall and relative precision were recorded after each iteration round. The results are shown in Figure 2, where each subfigure contains recall–relative precision curves of the two techniques. The recorded values are shown with symbols and connected with lines. It can be seen in Figure 2 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 **horses**), this overtake by PicSOM takes only one round. With some classes, however, the initial precision of VQ is clearly higher and therefore reaching it takes several iteration rounds.

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 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

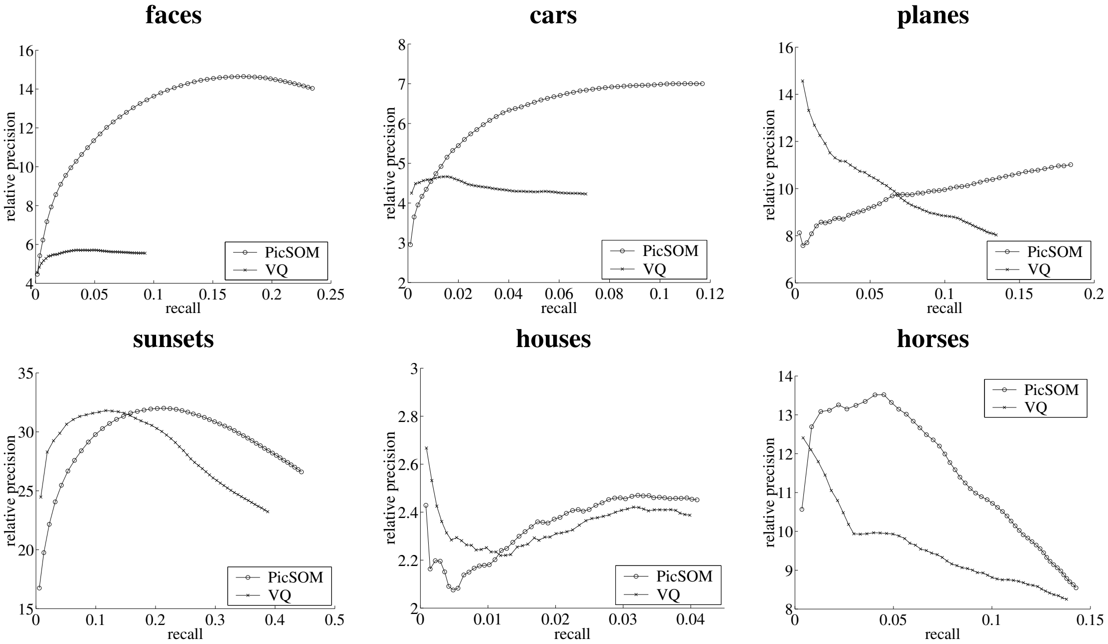

Figure 2. Recall–relative precision plots of the two CBIR techniques for the six image classes.

stays just above twice the *a priori* probability of the class with both methods.

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

4. Conclusions

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

In the experiments we compared the performance of the self-organizing relevance feedback technique of PicSOM with that of a vector-quantization-based reference system. The results showed that in the beginning of queries, PicSOM starts with a bit lower precision rate. Later, when its strong relevance feedback mechanism has enough data to process, PicSOM outperforms the reference technique.

References

[1] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, third edition, 2001.

[2] P. Koikkalainen and E. Oja. Self-organizing hierarchical feature maps. In *Proc. IJCNN-90, International Joint Conference on Neural Networks, Washington, DC*, volume II, pages 279–285, Piscataway, NJ, 1990. IEEE Service Center.

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

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

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

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

[7] Overview of the MPEG-7 standard (version 6.0), December 2001. ISO/IEC JTC1/SC29/WG11 N4031.

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