INTER-QUERY RELEVANCE LEARNING IN PICSOM FOR CONTENT-BASED IMAGE RETRIEVAL

Markus Koskela, Jorma Laaksonen, and Erkki Oja

Neural Networks Research Centre, Helsinki University of Technology,
P.O.BOX 5400, 02015 HUT, Finland
{*markus.koskela,jorma.laaksonen,erkki.oja*}*@hut.fi*

ABSTRACT

Content-based image retrieval (CBIR) addresses the problem of finding images that the user wants from unannotated databases, based only on low-level visual features like color or texture that can be automatically derived from the images. Due to the inherently weak connection between the high-level semantic concepts that the user has in mind, and the low-level visual features that the system is using, the performance of CBIR applications often remains quite modest. One method for improving CBIR results is to try to learn the user's preferences with learning methods such as relevance feedback. This learning is essentially intra-query, meaning that learning is started all over again in the beginning of each new query session. However, the relevance information can also be used in long-term or inter-query learning. In this paper, a method for using long-term learning in our PicSOM system is presented. It is shown that the efficiency of the system can be substantially increased by using it in parallel with MPEG-7 visual descriptors.

1. INTRODUCTION

Content-based image retrieval (CBIR) has received considerable research interest in the recent years (for a review, see [1]). The field has matured into a distinct research discipline which differs substantially from text-based information retrieval. The reason is that it is not possible to base image queries on verbal terms like in text-based retrieval, because it is assumed that no associated captions or textual descriptions of the images are available. Therefore, other query methods must be applied.

An especially difficult setting is encountered when the task is to retrieve images from a large database of miscellaneous images, for example, a digital photograph library of scenes, people, etc. Since very few assumptions about the images can be made, only representations of very general nature can be used and the general low-level features used in CBIR are insufficient to discriminate images well on a conceptual level. This creates a fundamental problem: there is a wide gap between the high-level semantic concepts used by a human to understand image content and the low-level visual features used by a computer to index the images in a database.

A possible solution is to learn on-line from the user. The common method for this is *relevance feedback*. In the common approach to formulate queries in CBIR, called *query by pictorial examples*, the image queries are based on example images shown from the database itself. In relevance feedback, the user is asked to rank or otherwise evaluate the query images offered by the system at each round. Based on the relevances given by the user, the system is tuned for example by adjusting the weightings of the features to better comply with the semantic similarity that the user has in mind. When the query proceeds from round to round, the relevance of the offered images gradually improves until the desired image has been found. Then the query stops.

Relevance feedback can be seen as a form of supervised learning to adjust subsequent query rounds by using information gathered from the user's feedback. It is essential that the learning takes place during one query, and the results are erased when starting a new query. This is because the object of the search usually changes from one query to the next, and so the previous relevances have no significance any more. This is therefore *intra-query* learning.

There have been some attempts in CBIR to use relevance feedback also for long-term learning. The history of the previous queries provides information which can also be used in an *inter-query* learning scheme. The basic motivation is that the relevance evaluations provided by the user during the queries partition the set of seen images into relevant and nonrelevant classes with respect to a particular query target. Although the relevant class may change totally from one query to the next, the fact that two images belong to the same relevance class is however a cue for the similarities in their semantic content. In this work, we introduce this idea into our PicSOM system.

This work was supported by the Academy of Finland in the projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *New information processing principles*, the latter being part of the Finnish Centre of Excellence Programme.

The rest of the paper is organized as follows: Section 2 briefly reviews the PicSOM system based on the Tree-Structured Self-Organizing Map. Then, the new inter-query learning scheme based on the relevance feedback history is explained in Section 3. Experimental results and conclusions are given in Sections 4 and 5.

2. PICSOM

Training the Multiple Feature Maps. The PicSOM CBIR system (for a recent review, see [2]) is a framework for research on content-based image retrieval. The methodological novelty of PicSOM, compared to other CBIR systems, is to use several parallel Self-Organizing Maps (SOMs), trained with separate feature types, to index the images in the database.

The well-known SOM [3] defines an elastic, topology-preserving grid of units that is fitted to the input vector space. It can thus be used to visualize multidimensional data, on a usually two-dimensional grid. The map attempts to represent all the available observations with an optimal accuracy by using a restricted set of prototypes, one in each map unit. The key property of the SOM in database indexing is that similar items tend to be mapped to the same or neighboring locations on the map.

In PicSOM, the features are usually comprised of statistical visual data such as the MPEG-7 [4] content descriptors, which we use in this work. To build the maps, these descriptors are extracted from the images in the database and the feature vectors are used to train the maps. Instead of the standard SOM, PicSOM uses a special form of the algorithm, the Tree Structured Self-Organizing Map (TS-SOM) [5].

As a result, the different feature SOMs impose different similarity relations on the images. The system inherently benefits from having a choice between a large set of features, as it may automatically neglect poorly-working ones. This is done by a special implementation of the relevance feedback mechanism.

Relevance Feedback with Self-Organizing Maps. The basic assumption in the PicSOM indexing method is that images similar according to a specific visual feature are located near each other on the corresponding SOM surface. Therefore, we are motivated to spread the relevance information given by the user also to the neighboring map units of the seen images. This is done as follows.

At each round of a query, a number of images from the database are shown to the user. All images indicated by the user as relevant are given equal positive weight, inversely proportional to the number of relevant images. Likewise, nonrelevant images receive negative weights that are inversely proportional to their total number. The overall sum of these relevance values is thus zero. For each SOM, these values are mapped from the images to the corresponding best-matching units (BMU) where they are summed. The resulting sparse value fields on the SOM surfaces are low-pass filtered to produce qualification values for each SOM unit and its associated images.

The low-pass filtering of sparse value fields can be performed by convolving the field with a tapered window function. The total qualification value for each image is finally obtained by summing the corresponding responses at that image location on all SOMs. Then a fixed number of images with the highest total qualification values are given to the user as the result of that query round, and the query continues.

As a consequence, content descriptors that fail to coincide with the user's conceptions mix positive and negative values in nearby map units. Therefore, they produce lower qualification values than those descriptors that match the user's expectations and impressions of image similarity. The different features and the SOMs formed from them do not thus need to be explicitly weighted, as the system automatically takes care of weighting their opinions on the basis of the relevance feedback mechanism.

3. INTER-QUERY LEARNING

One of the classical statistical tools in information retrieval is latent semantic indexing (LSI). The basis for LSI [6] is the vector space model of text documents. A collection of d documents are represented by the words in them by using a $t \times d$ term-by-document matrix $\mathbf{X}$, where t is the number of different words or terms. The element (i, j) of $\mathbf{X}$ represent the relationship of the i-th term to the j-th document; in the simplest case, it is just 1 if the i-th term occurs in the j-th document, 0 otherwise.

Recently, Heisterkamp [7] applied LSI to image databases. Instead of having a set of documents each consisting of words or terms, he considered the images as the vocabulary of the system and the individual queries as documents whose words are the images. In the document (query) vector, the relevance of each term (image) is indicated. Then each row of $\mathbf{X}$ gives the relevance history of one of the images over the consequent queries. If such relevance patterns of two images are similar, then the images must have some *semantic similarity*. It may therefore be reasonable to train yet another SOM, the *relevance map*, using the rows of $\mathbf{X}$ as additional image features. The relevance map can then be used in the operation of PicSOM just like any other feature map. This is our basic idea for inter-query learning in PicSOM.

A problem arises from the high dimension d which now equals the number of image queries in the training data. This may well be in the order of hundreds or thousands and

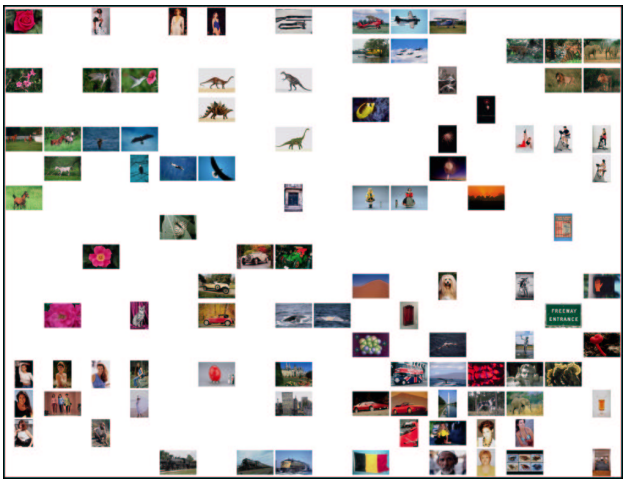

Fig. 1. The 16×16-sized TS-SOM level trained with the relevance feature.

thus excessive for direct usage in SOM training. This is where LSI comes in. In LSI, matrix $\mathbf{X}$ is decomposed by Singular Value Decomposition as follows:

$$\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $t \times r$ and $d \times r$ orthonormal matrices and $\mathbf{S}$ is an $r \times r$ diagonal matrix containing the nonzero singular values of $\mathbf{X}$ on the diagonal, with $r \leq \min(t, d)$ the rank of $\mathbf{X}$. The very high dimensionality of $\mathbf{X}$ can be reduced by selecting only k ($k < r$) dimensions corresponding to the k largest singular values of the diagonal matrix $\mathbf{S}$:

$$\widehat{\mathbf{X}} = \widehat{\mathbf{U}}\,\widehat{\mathbf{S}}\,\widehat{\mathbf{V}}^T \approx \mathbf{X}\,. \tag{2}$$

LSI is applied by considering only k largest singular values of $\mathbf{S}$ and a representation of the data in k dimensions is obtained with $\mathbf{Y} = \widehat{\mathbf{U}}\,\widehat{\mathbf{S}}$.

In our approach, the rows of matrix $\mathbf{Y}$, each corresponding to one image, are treated as a relevance feature of dimensionality k and the corresponding TS-SOM is trained and used in parallel and similarly as the TS-SOMs trained with visual features. The resulting TS-SOM is illustrated in Figure 1, in which the 16×16-sized level of the relevance map is displayed. The sparsity of the map is a direct consequence of the sparsity of the data; images in the same relevance evaluation tend to get mapped into the same map unit. The shown images are the visual labels given to the SOM map units. It can be observed that images with similar semantic content have been mapped near each other on the map. In another study, the user-provided relevance evaluations were shown to be notably similar to hidden annotations [8].

4. EXPERIMENTS

We used an image database containing 59 995 images from the Corel Gallery 1 000 000 product. To run automated tests, we created manually six ground truth image classes: **faces** (1115 images, *a priori* probability 1.85%), **cars** (864, 1.44%), **planes** (292, 0.49%), **sunsets**, (663, 1.11%), **horses**, (486, 0.81%), and **traffic signs**, (123, 0.21%). As visual image features, we used a subset of MPEG-7 [4] content descriptors for still images, viz. *Scalable Color*, *Dominant Color*, *Color Structure*, *Color Layout*, *Edge Histogram*, *Homogeneous Texture*, and *Region Shape*. For these visual features, we trained four-level TS-SOMs with level sizes 4×4, 16×16, 64×64, and 256×256 units. In the training of the lower SOM levels, the search for the BMU has been restricted to the 10×10-sized area below the BMU on the above level. Every image has been used 100 times for training each of the TS-SOM levels.

As for the new relevance feature, the training data consisted of 317 saved query sessions recorded earlier in our laboratory in which 6897 images (11.5% of the database) had been marked relevant at least once. The dimensionality k of the data was reduced to $k = 50$ with LSI as explained in Section 3. Since the relevance feature had non-zero vectors only for 6897 images, the corresponding TS-SOM structure was limited to three levels (64×64=4096 map units on the bottommost level).

Testing was performed automatically, without human interaction. The required relevance feedback was generated by the computer based on the class (faces, cars, etc.) of the desired image. All shown images belonging to the studied class were indicated as relevant and the others non-relevant. The class was not used in any way in choosing the next round of query images.

In our test setting, each image in the studied class is given to the system one at a time as the initial reference image for category search. The system should then return similar images (ie. images belonging to the same class), resulting in a leave-one-out type testing of the target class. The system was set to return 20 images at each round. 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 during a single query session. The system should thus demonstrate its performance within this number of images. We set N_T to 1000 images, resulting in 50 rounds per test query.

As performance index, we chose to show the evolution of *precision* $\mathcal{P}(n)$ as a function of *recall* $\mathcal{R}(n)$ during the iterative image retrieval process, with n the number of shown images. When instead of the whole database, only a smaller number $N_T \ll N$ of images are browsed through, the recall value is very unlikely to reach the value of one. Instead, the final value $\mathcal{R}(N_T)$ – as well as $\mathcal{P}(N_T)$ – reflects the total

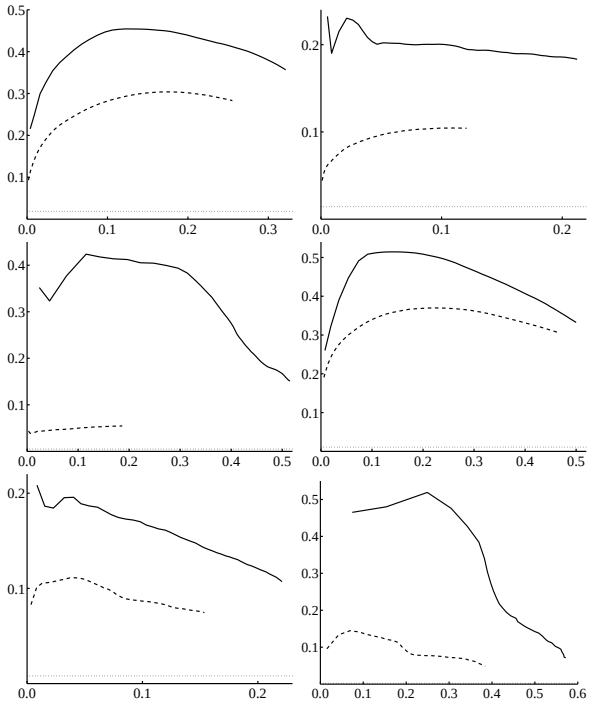

Fig. 2. Partial recall–precision plots using the MPEG-7 descriptors with (solid curve) and without (dashed curve) the user interaction feature. The *a priori* probability of the class is shown with a dotted line. Used classes were, top row, left-to-right: **faces**, **cars**, middle row: **planes**, **sunsets**, bottom row: **horses**, and **traffic signs**.

number of relevant images found that far. The intermediate values of $\mathcal{P}(n),\ n < N_T$ first display the initial accuracy of the CBIR system and then how the relevance feedback mechanism is able to adapt to the class.

As the queries in the used experiment setting are always started with an image that belongs to the image class in question, there is no need for an initial browsing phase. Instead, the retrieval can be initiated in the neighborhoods of the reference image on the bottommost SOM levels (64×64 for the relevance feature, 256×256 for others) as they provide the most detailed resolution. The upper TS-SOM hierarchy is thus neglected after the training phase. In spreading the responses of the sparse value fields, triangular windows of 4 and 8 map units in length were used for the relevance feature and the other features, respectively.

The resulting recall–precision plots are shown in Figure 2. The MPEG-7 descriptors are used in all cases with (solid curves) and without (dashed curves) the relevance feature. It can be seen that the relevance feature considerably improves retrieval precision, even though only 11.5% of the images are included in the feature.

5. CONCLUSIONS

With large databases of heterogeneous images, the retrieval performance of low-level visual features alone often remains quite modest and additional feature types may be needed for acceptable performance. A method for improving the performance based on using automatically recorded user-provided relevance evaluations was presented in this paper. In the PicSOM framework, the user interaction or relevance data is treated similarly as statistical visual features and, after dimensionality reduction, a separate relevance SOM is trained and used in retrieval. The method could also be used for existing keyword annotations. The results of the experiments show that the relevance feature greatly improves the precision of the system without any additional human labor required.

6. REFERENCES

[1] Alberto Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann Publishers, Inc., 1999.

[2] Jorma Laaksonen, Markus Koskela, and Erkki Oja, "PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 841–853, July 2002.

[3] Teuvo Kohonen, *Self-Organizing Maps*, Springer-Verlag, third edition, 2001.

[4] "MPEG-7 overview (version 8.0)," July 2002, ISO/IEC JTC1/SC29/WG11.

[5] Pasi Koikkalainen and Erkki Oja, "Self-organizing hierarchical feature maps," in *Proceedings of International Joint Conference on Neural Networks*, San Diego, CA, 1990, vol. II, pp. 279–284.

[6] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[7] Douglas R. Heisterkamp, "Building a latent semantic index of an image database from patterns of relevance feedback," in *Proceedings of the 16th International Conference on Pattern Recognition (ICPR 2002)*, Quebec, Canada, August 2002, vol. 4, pp. 134–137.

[8] Ingemar J. Cox, Jounama Ghosn, Matt L. Miller, Thomas V. Papathomas, and Peter N. Yianilos, "Hidden annotation in content-based image retrieval," in *Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Libraries (CBAIVL '97)*, San Juan, Puerto Rico, 1997, pp. 76–81.