Zhirong Yang and Jorma Laaksonen. 2005. Interactive retrieval in facial image database using Self-Organizing Maps. In: Proceedings of the 9th IAPR Conference on Machine Vision Applications (MVA 2005). Tsukuba Science City, Japan. 16-18 May 2005, pages 112-115.

Interactive Retrieval in Facial Image Database Using Self-Organizing Maps

Zhirong Yang and Jorma Laaksonen
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 HUT, Espoo, Finland
{zhirong.yang, jorma.laaksonen}@hut.fi

Abstract

Content-based image retrieval in facial image collections is required in numerous applications. An interactive facial image retrieval method based on Self-Organizing Maps (SOM) is presented in this paper, in which multiple features are involved in the queries simultaneously. In addition, the retrieval performance is improved not only within queries for current user but also between queries by long-term learning from other users' relevance feedback. In that way recorded human intelligence is integrated to the system as a statistical feature. The work constituting this paper has been incorporated into our image retrieval system named PicSOM. The results of evaluation experiments show that the query performance can be substantially increased by using multiple features and the long-term learning.

1 Introduction

Nowadays there exist a lot of databases involving digital facial images or mug shots. Searching for the image or images of a person in such a collection is a frequent operation and required by numerous applications. A number of systems of face recognition or identification have been developed, in which the user usually has to provide an image as initial example. However this is not practical in many situations, e.g. when searching for a photo of a specific criminal only through the recalling of a witness.

An integrated browsing tool for queries without user-provided example image can thus be very useful in facial image retrieval applications. However, designing effective interfaces for this purpose is a challenging task due to the inherently weak connection between the high-level semantic concepts perceived by humans and the low-level visual features automatically extracted by computers. This paradox is known as the *semantic gap*.

The framework for facial image queries has been incorporated in our PicSOM content-based image retrieval (CBIR) system [1]. PicSOM uses Self-Organizing Maps (SOM) [2] for indexing and its flexible architecture is able to accommodate multiple SOMs in parallel. Additionally, the *intra-query feedback* from the user can be recorded and later used in a long-term *inter-query learning* scheme.That way human intelligence is incorporated into the system as a statistical feature.

2 PicSOM CBIR System

2.1 Self-Organizing Map as indexing framework

In CBIR based on the vector space model, statistical features are automatically extracted from the images and represented by multi-dimensional vectors. We employ the Self-Organizing Maps (SOM) [2] as the indexing technique to organize these extracted feature vectors because SOM exhibits strong self-organizing power in unsupervised statistical data analysis.

After training a SOM, its map units are associated with the images of the database by locating the best-matching map unit (BMU) for each image on the two-dimensional discrete SOM grid. The SOM training guarantees the mapping preserves the topology in the original feature space. In image retrieval this means that mutually similar images are connected to the topologically near map units.

Instead of the standard SOM version, PicSOM uses a variational form of the algorithm, the Tree Structured Self-Organizing Map (TS-SOM) [3]. The multi-level structure of TS-SOM reduces the complexity of training large SOMs by exploiting the hierarchy in finding the BMU for an input vector efficiently.

2.2 Use of multiple features

For multiple features, the PicSOM system trains several parallel SOMs with different feature data simultaneously. The multiple SOMs impose different similarity relations on the images. The task of the retrieval system then becomes to select and combine these similarity relations so that their composite would approximate the human notion of image similarity in the current retrieval task as closely as possible.

Figure 1 illustrates the two-stage multi-feature retrieval setting. Each SOM $m = 1, \ldots, M$ is used separately for finding a set $\mathcal{D}_m^\alpha$ of the best image candidates according to that feature. This is especially advantageous when the distances calculated in the different feature spaces are weighted dynamically as in such a case it is not possible to order the images by their mutual distances in advance.

The per-feature image subsets are then combined into a larger set $\mathcal{D}^\beta$ of images which may be further processed in a more exhaustive manner. In our current implementation, the union of the initial sets, $\mathcal{D}^\beta = \bigcup_{m=1}^{M} \mathcal{D}_m^\alpha$.

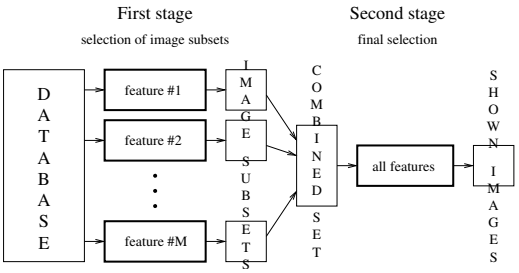

Figure 1. Two-stage structure of the PicSOM system.

3 Relevance Feedback with SOMs

3.1 Intra-query improvement

Despite the semantic gap, satisfactory CBIR results can often be obtained if the image query can be turned into an iterative process toward the desired retrieval target. The PicSOM system presents the user a set of facial images she has not seen before in each round of the image query. The user is then expected to mark the relevant images, and the system implicitly interprets the unmarked images as non-relevant. The SOM units are awarded a positive score for every relevant image mapped in them. Likewise associated non-relevant images result in negative scores. This way, we obtain a sparse value field on every SOM in use.

If a particular SOM unit has been the BMU for many relevant images and for none or only few non-relevant ones, it can be deduced that its content coincides well with the user's opinion. By assumption the neighboring SOM units are similar to it and the images mapped in them can likewise be supposed to be relevant for the user. Low-pass filtering of the sparse value fields is therefore applied on the two-dimensional map surfaces so that strong positive values from dense relevant responses get expanded into neighboring SOM units. On the other hand, weak positive and negative values in the map areas where the responses are sparse and mixed cancel each other out.

Each SOM has been trained with a different feature extraction method and therefore the resulting sparse value fields vary in different SOMs. In some SOMs the positive responses may spread evenly over the map surface, resulting in a seemingly random distribution of impulses. By contrast, in other SOMs the positive responses may densely cluster in certain areas of the map. The latter situation can be interpreted as being an indication on the good performance of those particular features in the current query. The denser the positive responses are the better the feature coincides in that specific area of the feature space with the user's perception on image similarity and relevance.

The relevance values from all the maps are summed for each image and the images with the highest overall scores are displayed to the user in the next round. This results in an iterative improvement for the query process.

3.2 Inter-query learning

Humans evaluate the similarity between faces intelligently, bringing semantic information in the made relevance assessments. For example, in some cases a user is able to track down the target by utilizing partial relevance like mustache, hair style or glasses. Actually the marking actions by users can be seen as hidden annotations of the images which subsequently serve as cues for similarity in their semantic contents. In practice, it turns out that previous user assessments provide valuable accumulated information about image semantics and can be a considerable asset in improving retrieval performance, albeit being static in nature.

In the PicSOM system, we consider the previous user interaction as metadata associated with the images and use it to construct a statistical *user interaction feature*, to be used alongside with the visual features [4]. In the PicSOM framework, this approach has desirable properties since one of the strengths of the system is that it inherently uses multiple features and generally benefits from adding new ones. This way the user interaction data is treated similarly as any other source of information about image similarity without the need of special processing.

The basis for the user interaction feature is the LSI method [5] in the vector space model of textual documents. Suppose we have r queries on a database of n images. First the singular value decomposition is applied on the image-by-query matrix X, preserving k ($k \ll r$) largest singular values: $[\hat{U}, \hat{S}, \hat{V}] = \mathrm{svds}(X, k)$. Thus we obtain a representation of the originally n-dimensional data in k dimensions as the rows of $Y = \hat{U}\hat{S}$. The rows of the matrix Y, each corresponding to one image, are treated as a user interaction feature of dimensionality k and the corresponding SOM is trained and used in parallel and similarly as the SOMs trained with visual features.

4 Experiments

4.1 Database

The research in this paper uses the FERET database of facial images collected under the FERET program [6]. After face segmentation, 2409 frontal facial images (poses "fa" and "fb") of 867 subjects were stored in the database for the experiments. The number of images belonging to one subject varies from one to twenty, the statistics of which are shown in Table 1.

Table 1. Histogram of cardinality of subject classes.

cardinality of subject class	number of subjects
1	2
2	632
3∼4	164
5∼6	42
7∼20	27
average: 2.78	total: 867

4.2 Feature extraction

In addition to the feature of the whole face, PicSOM supports features extracted from any other facial parts including eyes, nose and mouth provided that they can be reliably generated from the images. In our experiments the coordinates of the facial parts (eyes, nose and mouth) were obtained from the ground truth data of the FERET collection,

with which we calibrated the head rotation so that all faces are upright. Afterwards, all face boxes were normalized to the same size, with fixed locations for left eye (31,24) and right eye (16,24) in accordance to the MPEG-7 standard. The box sizes of the face and facial parts are shown in the second column of Table 2.

After extracting the raw features within the boxes mentioned above, we applied singular value decomposition to obtain the eigenfeatures of the face and facial parts [7]. The numbers of principle components preserved are shown in the third column of Table 2. For convenient discussion, we will refer to the five features involved in the experiments by their short names shown in the table.

Table 2. Specification of features used.

feature name	normalized size	eigenfeature dimensions	short name
face	46×56	48	**f**
left eye	24×16	10	**l**
right eye	24×16	10	**r**
nose	21×21	10	**n**
mouth	36×18	13	**m**
user interaction	—	50	**u**

4.3 Evaluation measures

In what follows, each experiment iterates over every one of the 867 subjects. In each loop, the retrieval goal is to search all images depicting that particular subject. 20 images were "displayed" per round and the first set of images was selected in random. In the automated evaluation the sole criteria for relevance of an image was whether it depicted the current subject or not. This scheme was used to avoid the subjective difference in evaluating the similarity between the images. Performance statistics recall, precision and average-precision were recorded after each round:

$$\begin{aligned} \text{recall}[j] &= r[j]/R \\ \text{precision}[j] &= r[j]/n[j] \\ \text{average-precision}[j] &= \frac{1}{R}\sum_{i=1}^{j}\big(cc[i]\cdot\text{precision}[i]\big) \end{aligned}$$

Here $r[j]$ is the cumulative number of relevant images retrieved after j rounds; R is the total number of relevant images in database; $n[j]$ is the cumulative number of total images retrieved after j rounds, i.e. $n[j] = j \cdot 20$; $cc[i]$ is the relevant ratio in i-th round, i.e. the number of relevant images shown in i-th round divided by 20. The overall performance statistics for each round j were obtained by averaging those from the 867 individual subjects.

5 Results

5.1 Single feature experiments

Five experiments were conducted to test the query performance by using every one of the five visual features individually. The averaged recall-precision plots are shown in Figure 2(a). For clarity, only the curve of left eye is displayed while that of right eye is omitted. The result indicates that the face feature (solid) performs the best, followed by eyes (dashed) and mouth (dash-dotted), while nose (dotted) plays the most trivial role in the retrieval. This order matches quite well with the psychological results [8]. The performance curve of a reference algorithm (cross-lined), where the images were chosen randomly in each round, is also shown as the baseline at the bottom. The plots of average-precision in different rounds, shown in Figure 2(b), also reveal the same order of performance.

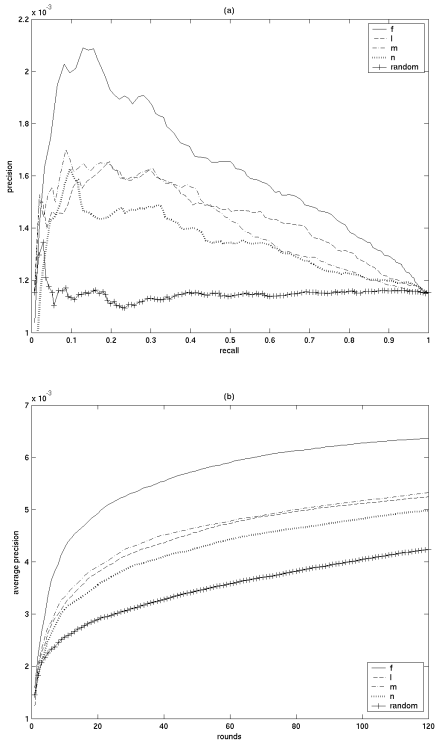

Figure 2. Averaged performance plots using single feature: (a) precisions at different recall levels; (b) average-precision versus rounds.

5.2 Feature combination experiment

In this experiment, we tested the query performance by using all visual features in parallel. Simple summation rule was used in the combination, i.e. each feature was treated equally and their scores from individual features were simply added up, and then the candidates were sorted by their total scores. The performance curves, marked as **flrmn** (dashed) and compared with that of **f** (dash-dotted), are shown in Figure 3. Although the face feature is the best one among the single features, the experiment result demonstrates that query performance can be significantly improved by taking multiple features into account simultaneously.

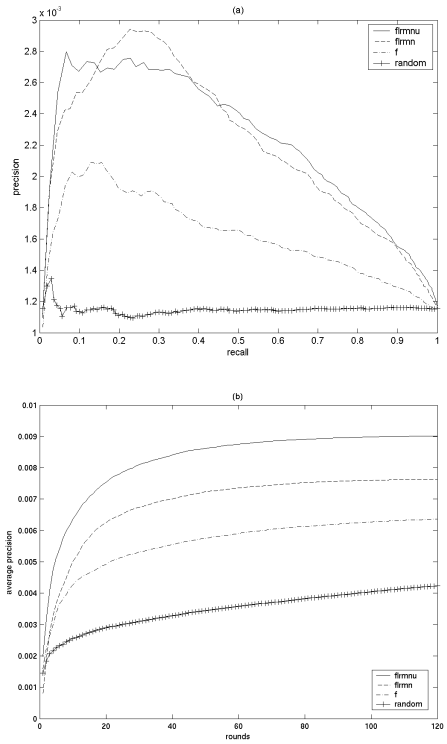

Figure 3. Comparison of averaged performance using feature combination as well as previous user interaction data versus single eigenface feature: (a) precisions at different recall levels; (b) average-precision versus rounds.

5.3 Experiments with user interaction feature

To test the performance of PicSOM in the presence of previous user interaction data, 318 query sessions, in which 1182 images (49% of the database) had been marked relevant at least once, were recorded in our laboratory. The relevance criterion could here consist of any human-available cues to track down one or more images of the target subject. After indexing the data as a new statistical feature with a SOM as described in Section 3.2, another non-interactive query experiment was conducted. The resulting curves are shown as **flrmnu** (solid) in Figure 3. Although in the recall-precision plots the **flrmnu** curve crosses with the **flrmn** curve, the peak of the former appears on the left to the latter, revealing that the user interaction feature is helpful in returning the relevant images earlier. The curves shown in Figure 3(b) confirm the advantage of using the user interaction feature. The average-precision appears to be a good performance indicator for situations where the relevant hits should appear as soon as possible.

6 Conclusions and discussion

An interactive retrieval method based on Self-Organizing Maps has been proposed in this paper to improve the retrieval performance in a facial image database. The method advances by employing multiple features in parallel and by using previous user interaction data as a separate statistical feature. The experiment results with our PicSOM CBIR system have revealed that query performance is substantially increased. No manual work is required during the entire procedure including feature extraction, indexing and query processing. The methods proposed in this paper are not restricted to mug shot images and it is straightforward to extend the methods to images from other fields.

Unlike CBIR systems on general images, the query precision of interactive facial image retrieval suffers from the problem of extremely small class sizes. The negative responses in early rounds provide only little semantic information and, as a result, the iteration performs in a nearly random manner. Consequently many zero pages (i.e. the images in these rounds are all non-relevant) are displayed until the first relevant image emerges. Therefore making the first relevant hit appear earlier is not a trivial task. A potential solution might be utilizing image filtering or partial indexing. This may be done by introducing automatic classification or weighting of the low-level features to enable the system to handle some high-level concepts.

Acknowledgment

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*, a part of the Finnish Centre of Excellence Programme 2000–2005.

References

[1] J. Laaksonen, M. Koskela, and E. Oja, "PicSOM–self-organizing image retrieval with MPEG-7 content descriptors," *IEEE Transactions on Neural Network*, vol. 13, no. 4, pp. 841–853, 2002.

[2] T. Kohonen, *Self-Organizing Maps*, 3rd ed., ser. Springer Series in Information Sciences. Springer, Berlin, 2001.

[3] P. Koikkalainen, "Progress with the tree-structured self-organizing map," in *Proc. of 11th European Conference on Artificial Intelligence, European Committee for Artificial Intelligence (ECCAI)*, 1994.

[4] M. Koskela and J. Laaksonen, "Using long-term learning to improve efficiency of content-based image retrieval," in *Proceedings of Third International Workshop on Pattern Recognition in Information Systems (PRIS 2003)*, France, 2003.

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, no. 6, pp. 391–707, 1990.

[6] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image and Vision Computing J*, vol. 16, no. 5, pp. 295–306, 1998.

[7] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, proceedings of CVPR'91, IEEE Computer Society Conference on*, Maui, HI USA, June 1991, pp. 586–591.

[8] J. Shepherd, G. Davies, and H. Ellis, "Studies of cue saliency," in *Perceiving and Remembering Faces*. Academic Press, London, 1981.