Semantic Annotation of Image Groups with Self-Organizing Maps*

Markus Koskela and Jorma Laaksonen

Laboratory of Computer and Information Science, Helsinki University of Technology
P.O.BOX 5400, FI-02015 TKK, FINLAND
{markus.koskela,jorma.laaksonen}@hut.fi

Abstract. Automatic image annotation has attracted a lot of attention recently as a method for facilitating semantic indexing and text-based retrieval of visual content. In this paper, we propose the use of multiple Self-Organizing Maps in modeling various semantic concepts and annotating new input images automatically. The effect of the semantic gap is compensated by annotating multiple images concurrently, thus enabling more accurate estimation of the semantic concepts' distributions. The presented method is applied to annotating images from a freely-available database consisting of images of different semantic categories.

1 Introduction

Content-based image retrieval (CBIR) addresses the problem of finding images relevant to the users' information needs, based principally on low-level visual features for which automatic extraction methods are available. Due to the semantic gap, i.e. the weak connection between the high-level semantic concepts that humans associate with images and the low-level features that computers are relying upon, developing this kind of systems has proven to be challenging.

One approach to improve retrieval results is to group somehow similar images together and use these groupings to filter out non-relevant images for the given query. Unfortunately, semantic categorizations often do not exist and they are difficult to produce automatically. Still, low-level classification and, in some cases, also certain semantic categorizations are possible with current automatic methods. Examples of low-level classification are distinguishing photographs from computer-generated graphics [1] and separating color and grayscale images. Certain types of semantic image categories can be distinguished with specialized classifiers which typically perform two-class classifications to the database images [2, 3, 1]. However, constructing such specific detectors for all categories that might appear in real-world images is clearly infeasible.

* 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.

Instead of strict classification, a somewhat more permissive approach is the *automatic annotation* of images (see e.g. [4–8]), where the input images are labeled with any of the available annotations if they fulfill the corresponding criteria. Unlike in classification, we do not assume that the database can be divided to a set of classes but rather that the images having a certain annotation constitute the representation of that semantic concept. Thereby, a single image may contain multiple annotations, and, on the other hand, the annotations may be incomplete, i.e. it is assumed that the database may contain some images of a certain concept that do not have the corresponding annotation. Instead of completely automatic methods, one may also apply *semi-automatic annotation* [9, 12], in which some additional information is used to derive annotations to the images. Recorded user interaction is usually used for this purpose. In many ways, automatic annotation is an inverse to the problem of keyword-based image retrieval, which can be considered as *automatic illustration* of textual concepts.

An even more challenging task is to target the annotations into specific regions in the images, i.e. *region naming*, partly due to the difficulty of robust image segmentation. This is naturally closely related to object recognition, although the approach is again more inexact as model-based recognition of thousands of objects in large image databases remains an unsolved problem.

In this paper, we approach the problem by assessing simultaneously multiple images sharing a semantic concept and jointly annotating the whole group. Our method can be applied to single images as well, but with a larger group of images of a given concept available, the concept's probability distribution can be estimated more accurately. Here, the focus is on annotation of whole images with global features instead of targeting image regions or blobs, so we do not discuss region naming. Since effective image understanding is generally not feasible without segmentation, the global approach is bound to have its limitations, although they can be somewhat alleviated with the use of several examples of the semantic concepts.

The rest of the paper is organized as follows. Our approach on using Self-Organizing Maps in image indexing and retrieval is described briefly in Section 2. In Section 3, we extend the use of multiple image indices from representing online image queries into modeling various semantic concepts and annotating new images automatically. Annotation experiments using a database of 101 object categories is presented in Section 4. Section 5 then concludes the paper.

2 SOMs in Image Indexing and Retrieval

The Self-Organizing Map (SOM) [10] is a powerful tool for exploring huge amounts of high-dimensional data. It defines an elastic, topology-preserving grid of points that is fitted to the input space. It is often used for clustering or visualization, usually on a two-dimensional regular grid. The distribution of the data vectors over the map forms a two-dimensional discrete probability density. Even from the same data, qualitatively different distributions can be obtained by using different feature extraction techniques.

2.1 Multi-Feature Image Indexing

Using the PicSOM system, we have previously studied CBIR with several parallel SOMs trained with separate feature data simultaneously (see e.g. [11, 12]). After training the SOMs, their map units are connected with the images of the database by locating the best-matching map unit (BMU) for each image on each SOM. As a result, the different 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.

The system can also utilize features and indexing methods for different types of image subsets [12]. Certain feature extraction methods are not meaningful for all kinds of images, e.g. extracting color features may be appropriate only to color images, and shape features requiring segmentation are valid for images containing salient objects and not e.g. for landscape or textural images. Also, it may be the case that a certain feature is available only for a portion of the database. Alternatively, the pertinent information of a subset can be contained in set membership, i.e. the subset consists of images having a specific property, such as the presence of a certain automatically detected object.

2.2 Relevance Feedback

During a retrieval session with the PicSOM system, the user marks images that she considers relevant, and the remaining ones are implicitly regarded as non-relevant. As the first step, the SOM units are awarded a positive score for every relevant image mapped in them resulting in an attached positive impulse. Likewise, associated non-relevant images result in negative scores and impulses. Let us denote the cumulative sets of relevant and non-relevant images up to query round r on mth SOM as $\mathcal{D}^+(r,m)$ and $\mathcal{D}^-(r,m)$. As the positive and negative scores, we use the inverses of the cardinalities of the corresponding image sets. Then, for each SOM, these values are mapped from the shown images (rated either as relevant or non-relevant by the user) to their corresponding BMUs where they are summed. Thus, for the kth map unit, we obtain the following response:

$$x[k]_m^r = \frac{1}{|\mathcal{D}^+(r,m)|} \sum_{i \in \mathcal{D}^+(r,m)} \delta(c_m(i),k) - \frac{1}{|\mathcal{D}^-(r,m)|} \sum_{i \in \mathcal{D}^-(r,m)} \delta(c_m(i),k) \qquad (1)$$

where $c_m(i)$ denotes the BMU of the image i on the mth SOM. This way, we obtain a zero-sum sparse value field on every SOM in use.

Due to the topology preservation of the SOM, we are motivated to spread the relevance information provided by the user also to the neighboring map units of the BMUs. This can be done by convolving the sparse value fields in with a two-dimensional tapered window function. For computational reasons, this is implemented as one-dimensional horizontal convolution followed by one-dimensional vertical convolution. Figure 1 illustrates how the positive and negative responses are first mapped on a 16×16-sized SOM to produce the sparse value field and how the responses are expanded in the convolution.

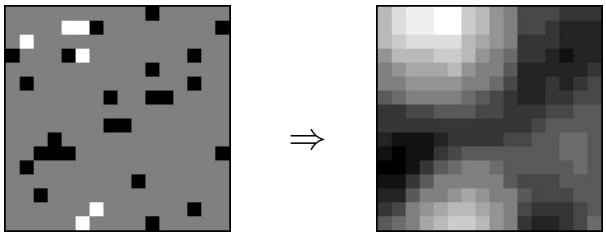

Fig. 1. An example of how a SOM surface is convolved with a window function. Left: the selected and rejected images are shown with white and black marks, respectively. Right: the convolution result, where relevance information is spread around the centers.

2.3 Feature Combination

As the response values of the parallel indices are mutually comparable, we can determine a global ordering and the overall best candidate images. By locating the corresponding images in all SOM indices, we get their scores with respect to different features. The total scores for the candidate images are then obtained by summing up the mapwise values in their BMUs after the convolution.

Content descriptors that fail to coincide with the user's conceptions mix positive and negative user responses in the same or nearby map units. Therefore, they produce lower scores than those descriptors that match the user's expectations and impression of image similarity and thus produce areas or clusters of high positive response. As a consequence, the parallel content descriptors and indices do not need explicit weighting. In image retrieval, this method for combining parallel descriptors automatically has been found out to be able to exceed or at least follow the performance of the best single image descriptors [11].

3 Modeling Semantic Concepts

In addition to the relevant and non-relevant image sets during online processing, the sparse value fields can also be constructed with any other image subsets, such as groups of images with semantically similar content.

3.1 Concept Representation with Class Distributions

Different features' capabilities in mapping semantically similar images near each other in the corresponding feature spaces can be studied visually by considering ground-truth semantic image classes as positive impulses on the sparse value fields. The convolution step is again useful to spread the concept information and also to ease visual inspection on large SOMs, as e.g. in class distribution visualizations shown in [11]. Furthermore, the discrimination abilities of the representations of the classes on the different SOMs can be analyzed quantitatively [13].

These class distributions can be considered as estimates of the true distributions of the semantic concepts in question, not on the original feature spaces, but on the discrete two-dimensional grids defined by the used SOMs. Thereby, instead of modeling the density in the high-dimensional feature spaces, we are essentially performing kernel-based estimation of class densities at the discrete distributions over the SOM surface. Then by enumerating the units of the two-dimensional SOM grid, we can represent the distribution as a vector $\mathbf{x} \in \mathbb{R}^K$ of length equaling the number of SOM units.

As an example, the most representative images of a given semantic concept can be obtained by locating the SOM units, and the images mapped to these units, that have highest responses on the estimated class distribution. Combining the responses of multiple features can be performed similarly as in the retrieval stage (Section 2.3), after which we can obtain the overall most representative image or images of a specific concept regarding all the used feature extraction methods (see Figure 2). Secondly, the shortcomings of different features can be examined by studying the images that yield a strong response on the class distributions but do not share the semantic content in question.

An important source of information about semantic correspondence between images in an unannotated database is the storage of relevance assessments of the system's users for later utilization. The relevance evaluations provided by a user during a query session partition the set of displayed images into classes of relevant and nonrelevant images with respect to that particular query target. The fact that two images belong to the class of relevant images during the same query is a strong cue for similarities in their semantic contents.

3.2 Automatic Annotation of Image Groups

Given an unannotated image or a group of semantically similar images, the goal of automatic annotation is to attach relevant annotations to the input images. For this purpose, some method for estimating the joint distribution of image representations and semantic concepts is required. We utilize an existing ground-truth database for which annotations are available and construct a separate model for every semantic concept present in the training data.

The responses invoked by different concept models on the SOMs can be directly used in automatic annotation. The input image group which we want to annotate is used to construct a class distribution $\mathbf{x}_q$ which is then compared to the existing models of semantic concepts $\mathbf{x}_i$. This approach has the distinct advantage that it inherently supports the annotation of image groups; with more reference images of a given concept available, the estimate of the corresponding distribution can be expected to become more accurate.

In this paper, we experiment with five similarity or distance measures. First of all, whether or not to perform the convolution step on $\mathbf{x}_q$ yields two alternative methods. By carrying out the convolution step we end up measuring the similarity of two estimated probability distributions. The similarity of $\mathbf{x}_q$ and $\mathbf{x}_i$ on the SOM grids can be measured in many ways; e.g. with 1) dot product

$s_{\mathrm{DP}}(\mathbf{x}_q, \mathbf{x}_i)$, 2) Euclidean distance $d_{\mathrm{EU}}(\mathbf{x}_q, \mathbf{x}_i)$, 3) intersection

$$s_{\mathrm{IN}}(\mathbf{x}_q, \mathbf{x}_i) = \frac{\sum_{k=1}^{K} \min(x_q[k], x_i[k])}{\sum_{k=1}^{K} x_q[k]} \ , \tag{2}$$

and 4) Jeffrey divergence

$$d_{\mathrm{JD}}(\mathbf{x}_q, \mathbf{x}_i) = \sum_{k=1}^{K} \left(x_q[k] \log \frac{x_q[k]}{\hat{x}[k]} + x_i[k] \log \frac{x_i[k]}{\hat{x}[k]} \right) \ , \tag{3}$$

where $\hat{x}[k] = (x_q[k] + x_i[k])/2$ is the mean distribution.

Secondly, the input image group can be associated with the semantic concepts that invoke the strongest positive responses on just the BMUs, not the neighborhoods, of the images to be annotated. This leads to measure 5, corresponding to omitting the smoothing convolution operation on $\mathbf{x}_q$ before calculating the dot product between $\mathbf{x}_q$ and $\mathbf{x}_i$.

Regardless of the measure used, the actual value of the similarity measure is an indication of annotation confidence. This can be utilized e.g. by defining an annotation threshold or emphasizing annotations that have high confidence.

4 Experiments

4.1 Database and Settings

In previous works on automatic annotation it has been common to use images from Corel Photo CDs (e.g. [4–8]). These images are of high quality and have been grouped by Corel in thematic groups. Ground-truth keyword annotations are also available for the images. Unfortunately, there is no single uniform Corel image set and thus the Corel databases different research groups possess are usually not identical. In addition, the Corel images are copyrighted and no longer even available. For example, the data set of Barnard et al. [5] has been made available[1], including segmentations and extracted features, but not the original images which we would need in order to properly apply our method to the data. We have also used Corel images in most of our earlier experiments (e.g. [11–13]).

Due to the non-free nature of the Corel database, we decided to use the 101 Object Categories database [14] of the PASCAL object recognition challenge[2] in the following experiments. The database contains 9197 images divided into 101 semantic categories, each containing between 31 and 800 images, and a background class of 520 miscellaneous images. The database has been gathered mostly for object recognition purposes and therefore does not contain detailed imagewise annotations. Still, the provided categorization can be used as a test setting for the annotation approach as well. Images from 16 random categories

[1] http://vision.cs.arizona.edu/kobus/research/data/jmlr_2003/
[2] http://www.pascal-network.org/challenges/VOC/

Fig. 2. Some example images from the 101 Object Categories database. The shown images are the most representative images of the following categories: beaver, electric_guitar, faces_easy, ferry, grand_piano, hedgehog, llama, menorah, pagoda, revolver, rhino, schooner, scissors, starfish, stegosaurus, and stop_sign.

of the database are displayed in Figure 2. Specifically, the shown images are the most representative images of these 16 categories, as defined in Section 3.1.

From each category, ten random images were selected to the test set and the remaining images were used to construct the category model on the SOM indices. Image groups of 10, 5, 2, and 1 images were then annotated by using each of the five measures (Section 3.2) of the similarity between the image group and the category models. All the ten images in the test set were always used in measuring the performance; for image groups smaller than ten, the test images were split into multiple groups and the results are the average of all the respective runs.

As visual features, we used a set of MPEG-7 [15] descriptors suitable for still images, viz. *Scalable Color*, *Dominant Color*, *Color Structure*, *Color Layout*, *Edge Histogram*, *Homogeneous Texture*, and *Region Shape*. These descriptors were extracted from every image in the database and 64×64-sized SOMs ($K = 4096$) were trained for each of them. A triangular window of four map units in length was then used in spreading the responses of the sparse value fields.

4.2 Measuring Annotation Performance

Measuring the performance of automatic image annotation requires some consideration. The straightforward approach is to compare predicted annotations to the manual ones and measure the overlap. In [5], the following measure was used for this purpose:

$$E = \frac{r}{n} - \frac{w}{N-n} \tag{4}$$

where r and w are the numbers of words predicted right and wrong, n is the number of manual annotations for the image and N is the size of the vocabulary. In practice the manual annotations are often incomplete. Appropriate annotations may be missing from individual images, especially ones describing the background of the image or ones being very general, since humans tend to overlook obvious but subsidiary visual cues when describing image content. Synonyms can also be problematic if the annotations were generated without a synonym-free set of allowed keywords. As an example, the supplied annotations for the

Table 1. The results of the annotation experiments for image groups of different sizes. On each cell, the three reported values are *MRR*, N_1 and N_5. In total there were 101 semantic categories and a background category.

group size	1) dot product	2) Euclidean	3) intersection	4) Jeffrey div.	5) no convol.
10	0.755, 64, 90	0.679, 56, 81	0.870, 82, 95	0.868, 83, 93	0.788, 69, 93
5	0.654, 54, 83	0.633, 54, 75	0.720, 62, 86	0.752, 68, 86	0.693, 59, 84
2	0.491, 37, 64	0.512, 41, 64	0.494, 37, 64	0.541, 43, 67	0.518, 40, 66
1	0.391, 27, 52	0.407, 31, 51	0.388, 27, 52	0.403, 29, 53	0.407, 29, 52

Corel database contain distinct annotations such as "automobile" and "car". As a result, the observed annotation performance may be overly pessimistic. When comparing different annotation methods, this is, however, not crucial, since all methods encounter the same missing annotations. The word frequency of the annotations should also be taken into account. Annotating images with general concepts like "sky" or "landscape" is successful with a higher probability than with very specific terms.

In our current experiment setting the situation is more straightforward. Since each image has exactly one correct annotation (i.e. its category) and the word frequency is relatively flat, we can measure the rank of the correct category for each annotation task. In order to be useful for annotation, the rank of the correct category should be low; a high rank can be deemed an annotation failure and the actual rank is inconsequential. Therefore we record the inverses of the ranks and by averaging over the 101 categories, we obtain the *mean reciprocal rank, MRR*. Furthermore, we record the number of categories for which the rank of the correct category is one (N_1) and for which it is less or equal than five (N_5).

4.3 Results

The annotation results for image group sizes 10, 5, 2, and 1 with the five tested similarity or distance measures are shown in Table 1. It can be seen that the size of the image group is a critical factor in annotation performance as increasing the group size improves results considerably in all cases. This behavior was, naturally, to be expected since the probability distributions of the semantic concepts can be modeled more accurately with more reference images available.

The selection of the similarity measure is less crucial. The best results for groups of ten images are obtained using the intersection and Jeffrey divergence measures. With them, all but six[3] and eight[4] categories, respectively, are annotated correctly among the five highest-scoring annotations. For the 16 categories represented in Fig. 2, the five best annotations for groups of ten reference images per category and using the intersection measure are listed in Table 2.

Due to the semantic gap, the performance of single image annotation remained rather poor; less than one third of the single test images were annotated

[3] anchor, ant, barrel, cannon, crab, and wild_cat.

[4] anchor, ant, barrel, cannon, crab, emu, platypus, and wild_cat.

Table 2. Five best annotations for a sample of 16 object categories (see Fig. 2) with ten reference images per category and using intersection as the similarity measure.

category	annotations
beaver	crab, emu, beaver, llama, kangaroo
electric_guitar	electric_guitar, accordion, trilobite, sea_horse, mandolin
faces_easy	faces_easy, faces, dalmatian, lamp, flamingo
ferry	ferry, helicopter, ketch, schooner, laptop
grand_piano	grand_piano, rooster, okapi, mandolin, gramophone
hedgehog	emu, hedgehog, courgar_face, kangaroo, okapi
llama	llama, crocodile_head, elephant, gerenuk, okapi
menorah	menorah, garfield, sunflower, starfish, rooster
pagoda	pagoda, minaret, accordion, trilobite, cellphone
revolver	revolver, stapler, wrench, umbrella, dragonfly
rhino	crocodile, llama, emu, elephant, rhino
schooner	schooner, ketch, buddha, ferry, helicopter
scissors	scissors, snoopy, wrench, pigeon, headphone
starfish	starfish, strawberry, scorpion, sunflower, ant
stegosaurus	stegosaurus, panda, cannon, brontosaurus, octopus
stop_sign	stop_sign, strawberry, flamingo_head, yin_yang, soccer_ball

correctly as the first annotation, among the five highest-scoring annotations the correct one was in about half of the cases. Also, with smaller image groups the differences between the tested similarity methods are less distinctive. Most notably, measuring the responses of the category models directly on the BMUs of the input images (measure 5) seems to work relatively better with small image groups. Overall, Jeffrey divergence seems to perform relatively well on image groups of any size and could thus be used as a default similarity measure.

5 Conclusions and Future Directions

In this paper, we proposed a method for applying multiple SOMs in representing semantic concepts of images and automatic image annotation. The density models for different semantic concepts are produced using an annotated image collection as a ground truth. New image groups are then annotated by comparing them to these concept models on the SOM grids. The presented methods for measuring the similarity between database subsets can also be used for other purposes, e.g. detecting synonyms or similar semantic concepts, and combining such stored user interaction records that had similar semantic query targets.

Due to the weak connection between semantic concepts and low-level visual features, the task of automatic annotation based on global features is bound to have only limited success. By visual inspection of the failed categories, one can observe remarkably high variation and diverse backgrounds. For successful annotation of this kind of images, the method needs to be extended from the image level to subobjects, based either on image segmentation, using fixed image zones or calculating interest points from the images. Especially on the 101 Object

Categories database, the lack of separation of the salient object from the background is a crucial impediment. In any event, even the global approach can reach quite prominent performance by annotating multiple images concurrently. The method presented in this paper is directly applicable to and will undoubtedly be an asset also when dealing with image segments or other subobjects.

The experiments of this paper were carried out using a database consisting of semantic object categories. Further tests and consideration are needed for annotations of different levels of specificity, i.e. by using databases that have imagewise annotations. Such databases should, however, be freely available to researchers to facilitate comparisons of different methods.

References

1. Gevers, T., Aldershoff, F., Geusebroek, J.M.: Integrating visual and textual cues for image classification. In: Proceedings of Fourth International Conference on Visual Information Systems (VISual 2000), Lyon, France (2000) 419–429
2. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: Proc. IEEE International Workshop on Content-Based Access of Image and Video Database, Bombay, India (1998) 42–51
3. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City images vs. landscapes. Pattern Recognition **31** (1998) 1921–1935
4. Chang, E., Goh, K., Sychay, G., Wu, G.: CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Transactions on Circuits and Systems for Video Technology **13** (2003) 26–38
5. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.I.: Matching words and pictures. Journal of Machine Learning Research **3** (2003) 1107–1135
6. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. on Patt. Anal. and Machine Intell. **25** (2003) 1075–1088
7. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proc. 26th ACM SIGIR Conf. on Research and Development in Information Retrieval, Toronto, Canada (2003) 119–126
8. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proc. 26th ACM SIGIR Conf. on Res. and Devel. in Information Retrieval, Toronto, Canada (2003) 127–134
9. Lu, Y., Hu, C., Zhu, X., Zhang, H., Yang, Q.: A unified framework for semantics and feature based relevance feedback in image retrieval systems. In: Proc. 8th ACM Int'l Conf. on Multimedia, Los Angeles, CA, USA (2000) 31–37
10. Kohonen, T.: Self-Organizing Maps. Third edn. Springer-Verlag (2001)
11. Laaksonen, J., Koskela, M., Oja, E.: PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. IEEE Trans. on Neural Networks **13** (2002) 841–853
12. Koskela, M., Laaksonen, J., Oja, E.: Use of image subset features in image retrieval with self-organizing maps. In: Proceedings of 3rd International Conference on Image and Video Retrieval (CIVR 2004), Dublin, Ireland (2004) 508–516
13. Laaksonen, J., Koskela, M., Oja, E.: Class distributions on SOM surfaces for feature extraction and object retrieval. Neural Networks **17** (2004) 1121–1133
14. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In: Proc. Workshop on Generative-Model Based Vision, Wash., DC (2004)
15. ISO/IEC: (Information technology - Multimedia content description interface - Part 3: Visual) 15938-3:2002(E).