APPLICATION OF SELF-ORGANIZING MAPS AND AUTOMATIC IMAGE SEGMENTATION TO 101 OBJECT CATEGORIES DATABASE

Jorma Laaksonen, Ville Viitaniemi and Markus Koskela

Helsinki University of Technology
Neural Networks Research Centre
P.O.BOX 5400, FI-02015 TKK, Espoo, Finland
{jorma.laaksonen,ville.viitaniemi,markus.koskela}@tkk.fi

ABSTRACT

In this paper, we study how well our PicSOM CBIR system is able to find prototypical image segments based on image-level keywords and automatic image segmentation. We also study different methods for focusing a given keyword on a particular image segment. Both these processes are based on the Self-Organizing Map's ability to map image segments which are mutually similar according to a specific image feature in nearby map units. In addition, the PicSOM system can automatically use and weight multiple different features in parallel. In the automatic image segmentation applied to the images of the 101 Object Categories database, a fixed number of segments have been extracted from each image. This leads in many cases to oversegmentation, but our experiments show that the system's ability to find prototypical segments is not severely impaired. On the other hand, it is clear that the process of keyword focusing would benefit from more precise segmentations.

1. INTRODUCTION

The conventional employment of content-based image retrieval (CBIR) systems has been targeted at interactive use where the task of the system is to return to the user interesting or relevant images from an unannotated database. In this paper, however, we are concerned with methods that can automatically solve the keyword versus image segment correspondence problem in a keyword-annotated database of salient objects. This way, we can obtain more focused representations of the object categories specified for the database, when in each image the location of the object is solved.

The technique we use is based on the PicSOM CBIR system [10], where *relevance feedback* is used as a method for *query refinement*. In this work, we have replaced the relevance feedback with keyword-type annotations given to the images in the 101 Object Categories database [5]. After automatic stages of segmentation, feature weighting and keyword focusing, the system is able to present a set of typical image segments that can be regarded as visual counterparts of the semantic concepts provided to the system in the form of keywords.

Due to the general difficulty of robust image segmentation, the images are often purposely oversegmented rather than undersegmented. It will be interesting to examine how well the semantic information presented on the image level really can be focused on specific image segments even in such circumstances. This will be illustrated by experiments with one selected object category.

The paper is organized as follows. In Section 2 we will first discuss some aspects of CBIR systems on a general level and then in Section 3 our PicSOM system in more detail. Section 4 addresses the problem of automatic image segmentation and Section 5 presents our view on how segmentation can be used in extracting semantic contents of images. The data of our experiments will be described in Section 6 and the actual experiments and their results in Section 7. Conclusion will be drawn and future directions discussed in Section 8.

2. CONTENT-BASED IMAGE RETRIEVAL

One popular method to improve the retrieval performance in CBIR systems is to employ relevance feedback from the user in intra-query learning. The relevance feedback from the user can also be recorded online and later analyzed offline to reveal semantic relations between visual objects. In our earlier works [8, 9], we have shown that this user interaction information can be used as a statistical feature to improve online retrieval even without any semantic postprocessing.

Another important and rising technique in CBIR is the utilization of automatic and model-less or assumption-free

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 2000–2005.

segmentation methods for the images. In this scheme one is addressed with the question on how the relevance feedback and potentially existing annotations or keywords given to whole images can be focused on particular image segments. If both the segmentation and focusing problems could be solved simultaneously and successfully, many of the persisting contemporary challenges in computer vision could be settled.

Two conceptually opposite alternatives for the use of keywords or other textual information in CBIR systems exist. The first one is the technique commonly used in general-purpose WWW search engines, such as Google Image Search, where images are located on the basis of their surrounding texts on web pages. In this way, texts and keywords act merely as pointers to images in online textual queries. The second alternative is more challenging and will be studied in this paper as an extension to our earlier experiments addressing online CBIR. More precisely, keyword or image category information is here used offline to extract prototypical segments from the images sharing a common keyword. This will be feasible if the automatic segmentation will be successful enough and we are provided with enough images in the category.

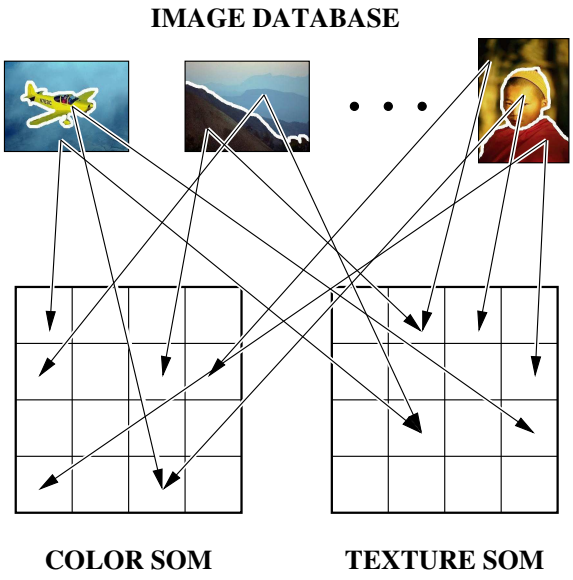

Fig. 1. An example of using two parallel SOM indices for segmented images. The color and texture SOMs are trained with image segments and each segment is connected to its BMU on each SOM.

3. PICSOM SYSTEM

The PicSOM [10] image retrieval system is a framework for research on content-based image retrieval. As the name implies, PicSOM uses the Self-Organizing Map (SOM) [7] as its basic image indexing method, although other clustering methods are also supported. The SOM 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.

3.1. Multiple Self-Organizing Maps

The PicSOM system is fundamentally based on using several parallel SOMs trained with different feature data simultaneously in image retrieval. The features are usually comprised of statistical visual data such as the MPEG-7 content descriptors [6]. Any additional vectorial data can, however, be used to train corresponding SOMs and thus be used in image retrieval. Furthermore, SOM indices can be constructed either from whole images or certain subobjects, such as image segments. On image segment SOMs, the items to be organized on the SOM are not the images themselves but the segments. However, since relevant images, not the segments, are in many applications the actual target of retrieval, the link between the image and its segments is preserved. In that way the combined response for the parent images can still be determined from those of their child segments.

After training the SOMs, the map units are connected with the database images or their appropriate segments. This is done by locating the BMU for each image or segment on each SOM. As a result, the different SOMs impose different similarity relations on the images and the system thus inherently uses multiple features for image retrieval. An illustration with two parallel SOMs trained for image segments is presented in Figure 1.

3.2. Relevance feedback with image segment features

During an online retrieval session with PicSOM, the system presents to the user a set of images of which she marks the ones she considers relevant, and the remaining ones are implicitly regarded as non-relevant. These relevance assessments are then propagated to all segments of the respective images. As the next step, the SOM units are awarded a positive score or response value for every relevant image segment mapped in them resulting in an attached positive impulse. Likewise, associated non-relevant segments result in negative scores and impulses. If the total numbers of relevant and non-relevant marked segments are $N^+(n,m)$ and $N^-(n,m)$ at nth query round on mth SOM, the positive

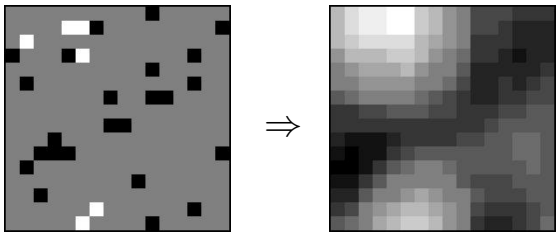

Fig. 2. An example of how a SOM surface is convolved with a tapered window function. On the left, segments of images selected and rejected by the user are shown with white and black marks, respectively. On the right, the convolution result, where relevance information is spread around the centers.

and negative scores are simply the inverses:

$$x_+(n) = \frac{1}{N^+(n,m)} \text{ and } x_-(n) = -\frac{1}{N^-(n,m)} . \quad (1)$$

For each SOM, these values are mapped from the segments of the shown images (and thus rated either as positive or negative) to the corresponding BMUs where the response values are then summed. This way, we obtain a zero-sum sparse value field on every SOM in use.

Due to the topology preservation of the SOM, segments which are 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 (both positive and negative) provided by the user also to the neighboring map units of those BMUs on the SOMs. This can be done by convolving the sparse value fields with tapered (eg. triangular or Gaussian) window functions. Figure 2 illustrates how the positive and negative responses, displayed with white and black map units, respectively, are first mapped on a 16×16-sized SOM and how these responses are expanded in the convolution.

As the response values of the parallel indices are mutually comparable, we can determine a global ordering for determining the overall best candidate segments and images. By locating the corresponding segments in all indices, we get their scores with respect to different feature extraction methods. The total qualification values for the candidate segments are then obtained simply by summing the corresponding responses. For images, their segment-wise values are further summed to form the image-level qualification values. 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 qualification values 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 any explicit weighting.

4. AUTOMATIC IMAGE SEGMENTATION

Image segmentation partitions the image area into segments. The aim is to do the partitioning so that it would be helpful in further image analysis. In an ideal case the segments would directly correspond to the real-world objects present in the image. In practice it is virtually impossible to achieve such a *complete segmentation* in an unsupervised manner as the processes of segmentation and complete understanding of image contents are intrinsically intertwined. In practice one has to settle for *partial segmentations*, where the images are partitioned into regions that are homogeneous in terms of some visual property, such as color or texture.

Recognizing the fact that one is not going to be able to solve the automatic image segmentation problem in full, it is still hoped that we can produce partial segmentations that are good enough to be helpful for CBIR purposes [11]. In our earlier experiments, eg. with the wide-domain Corel image database, most of the images have been "natural" in the sense that there might not be any particular salient object in the image or its background is heterogeneous. Our image segmentation methods have consequently been tuned towards the processing of this kinds of images. Now in the 101 Object Categories database's narrower-domain images there is always a clearly distinguishable object in the middle of the image whose background is often quite homogeneous. The nature of these kinds of images is therefore a bit "artificial".

Figure 3 displays one example from the database. The background is nontrivial, but the automatic segmentation has been quite successful. The segment marked "1" corresponds accurately to the lobster and the other seven segments cover the rest of the image. On the other hand, Figure 4 displays another lobster image which is clearly oversegmented. This is a natural consequence of the fixed number of extracted segments and the simplicity of the image, as there does not exist anything other than the lobster in the image. As a result, the lobster is segmented in parts denoted "1", "3", "4" and "7", background in "0"and "6", whereas segment "2" spans both the object and the background and segment "5" is part of the animal's shadow. The details of the used segmentation method will be discussed in Section 6.2.

5. IMAGE SEGMENTATION AND SEMANTICS

The relation between image segmentation and semantic concepts has become a subject of recent intensive study in the field of CBIR. The goal has been named as "image-to-word transformation" [13], "matching words and pictures" [1], "image auto-annotation" [12], "automatic image captioning" [14], or "automatic image annotation" [4], de-

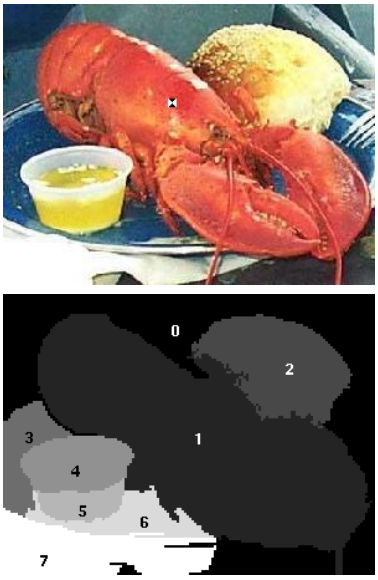

Fig. 3. An image of a lobster and the result of its automatic segmentation in segments marked as '0'–'7'. A reference point for the lobster's neck has been manually marked in the original image and can be seen to reside in segment '1'.

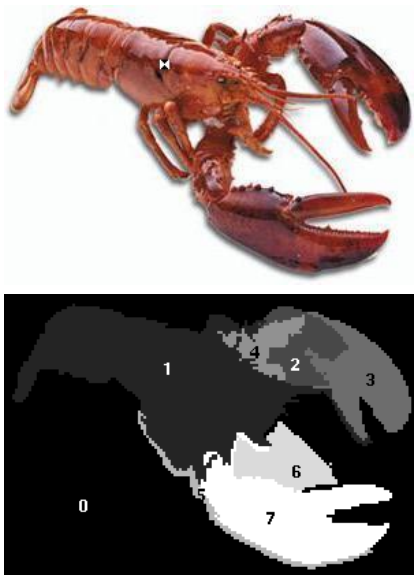

Fig. 4. Another lobster image and the result of its automatic segmentation in segments marked as '0'–'7'. Oversegmentation is visible both in the object and the disjoint background.

pending on the selected viewpoint and the specific tasks the authors have been addressing. Various different technical methods and their combinations have been applied, including co-occurrence statistics [13], Expectation Maximization (EM) [1], Support Vector Machines (SVM) [4], Latent Semantic Analysis (LSA) [12], and Markov random fields (MRF) [2].

According to our knowledge, Self-Organizing Maps have not earlier been used for studying the interplay of image segments and semantic concepts. The usability of SOMs in general CBIR has, however, been demonstrated by our earlier studies and comparisons, eg. in [10, 15]. What then would make the SOM an efficient tool for the semantic analysis of image segments? We believe that the PicSOM system's ability to use different feature extractions simultaneously and to weight them automatically is a feature not shared by many other techniques. In that process, segments which depict the background or otherwise meaningless or arbitrary parts of the image can be regarded as additive noise, whose weight will be reduced in comparison to that of the meaningful segments. This favorable behavior is a direct consequence from the PicSOM system's processing principle, where mutually similar and densely mapped relevant images and image segments strengthen each other and thus in the end dominate over mutually dissimilar ones mapped sparsely on the SOM surfaces.

In this paper, we want to study whether our assumption about the usability of the PicSOM system for extracting semantic information from keyword-annotated images is valid. The experiments to be presented in Section 7 will contain four steps. First, we will study how the segments of images belonging to a certain semantic category are mapped on SOM surfaces. Second, we will examine two different ways of focusing keyword annotations from the image level to the segment level. In the first method, all segments of all images sharing the keyword are first marked relevant on all SOM maps. In this situation we will have many *false positive* segments marked as relevant but no *false negatives* since none of the semantically relevant segments will be missed. After the convolutions, each segment obtains a qualification value which indicates how prototypical representative it is for that keyword. For each image, its segments can then be ordered in the order of descending qualification value. When the least prototypical segments are progressively being discarded, we obtain an operation curve where the number of false positive segments is decreasing while false negative segments increase in their number. The second method is otherwise similar, but the convolutions are repeatedly performed every time after the least representative positive segment has been discarded from each image. In this way, the process is more gradual or really focusing, where the least trustworthy and least probable segments are being iteratively neglected.

Third, we will try to find prototypical image segments for a semantic concept expressed with a pair of keywords. Also in this case the keywords have been given on the image level, so an indication of the system's ability to focus the keywords can be seen in the results.

6. EXPERIMENTAL DATA

6.1. Database

In the following experiments, we use the *101 Object Categories* database [5] of the PASCAL *Visual Object Classes Challenge*[1]. 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 image-wise annotations.

6.2. Image segmentation

The images in the database were segmented in two steps. In the first step isodata variant of K-means algorithm [16] with a K value 15 was used to compute an oversegmentation based on the RGB values of the pixels. This step typically resulted in a few thousand separate segments.

In the second step the segments were merged. The region distance in the CIE L*a*b* color space [3] d_{LAB} was used as the basis for the merging criterion. In addition, the multi-scale edge strength e between the regions was also taken into account. The final merging criterion C was weighted with a function of the sizes $|r_i|$ of the to-be-merged regions r_i:

$$\mathrm{C}(r_1, r_2) = \mathrm{s}\left(r_1, r_2\right)\left(d_{LAB}\left(r_1, r_2\right) + \lambda \mathrm{e}\left(r_1, r_2\right)\right), \quad (2)$$

where

$$\mathrm{s}(r_1, r_2) = \min(|r_1|, |r_2|, a) + b \quad (3)$$

is the size-weighting function and λ, a and b are parameters of the method. The merging was continued until eight regions were left.

Prior to the segmentation, the images were scaled to width of 150 pixels and the original image sizes were restored after it. As the result of the segmentation we thus had a database 82773 visual entities, of which 9197 were images and 73576 image segments. We then determined a subset of 43256 image segments with the additional requirement that the segments in that subset were not allowed to touch the outer borders of the image. The use of this subset is motivated by the nature of the 101 Object Categories collection where the objects are mostly salient, located in the middle of the image and not extending to the borders.

6.3. Features

Seven different features were used to describe the visual content of the segments. Simple color and texture features were included in the feature set, partly due to the reason that the mapping between feature spaces and visually perceptible object properties was thus kept understandable and the experimental results were easier to interpret. The CIE L*a*b* color coordinates themselves and also their first three central moments [17] were used as color features. Texture was described using a feature that compares the YIQ color space Y-values of pixels to their 8-neighbors.

Five MPEG-7 content descriptors [6], Edge Histogram, Region Shape, Color Layout, Dominant Color and Scalable Color, were used as somewhat more sophisticated features. A separate TS-SOM was trained for each feature with levels containing 4×4, 16×16, 64×64 and 256×256 map units.

Features from all the extracted 82773 image segments were used in training the SOMs. Unlike our previous experiment setting, we now did not use whole image features and SOMs at all.

7. EXPERIMENTS AND RESULTS

For performing the experiments we needed ground truth data for the locations of the objects of some semantic category or class in the images. We selected the "lobster" category, examples of which were already displayed in Figures 3 and 4. In those figures, manually specified references points for the object can be seen. This ground truth information was used solely in measuring the PicSOM system's performance and was not available to the system itself when performing the automatic segmentation and focusing tasks.

In general it can be seen that the lobster images are much more often over- than undersegmented, ie. the animal appears in more than one segment. This results from the fact that we have always extracted the fixed number of eight segments while the backgrounds of the images are in many cases very homogeneous and there really are no other visible entities but the lobster itself.

7.1. Class distributions

In the first experiment we studied how the image segments of the "lobster" category are distributed on the SOM surface in the case of the low-level CIE L*a*b* central color moment feature. In the study, we applied the four combinations of two additional options. First, we used either all image segments or only those that do not touch the image borders. This information is naturally available as a byproduct of the automatic image segmentation. The restriction to use only the non-border segments can be motivated by noting that additive noise from background segments can be suppressed by this way. The second option was to use only those lobster segments which contain the mark positioned in the neck of the crustacean when we annotated that particular image category. This kind of auxiliary information is normally not available for the system and can therefore be

[1] http://www.pascal-network.org/challenges/VOC/

used only to study what could be gained if it were.

Figure 5 shows the surface distributions of the classes on the 256×256-sized bottom level SOMs after a convolution with a triangular-shaped window of radius 12 units. Darker shades represent denser distributions in the respective parts of the map. It can be seen that the data distribution in the color moment feature space is roughly three-modal when all segments are used. When the border segments are filtered out, the fraction of the distribution located in the top left corner of the SOM is suppressed. It can easily be verified that these segments depict the white background existing in many of the lobster images. We can further notice that the distribution becomes even sharper when only the explicitly marked segments are used. The difference is, however, not crucial, and the true distribution can sufficiently be approximated with that of all non-border segments.

The result of this experiment is thus that the distributions of the segments from different semantic classes can be seen concentrated on specific locations on different SOM surfaces. Some of them will certainly not be as clear as in this case where the characteristic color of the lobster was a key factor. The concentration seems to take place even though the segment data for a class contains seven times more false positive samples than true positives. However, the distribution of the false positives is less concentrated than that of the true positives, as expected, and can be further suppressed by using only the non-border segments. The effect of the false positives might therefore be regarded at least to some extent as negligible noise.

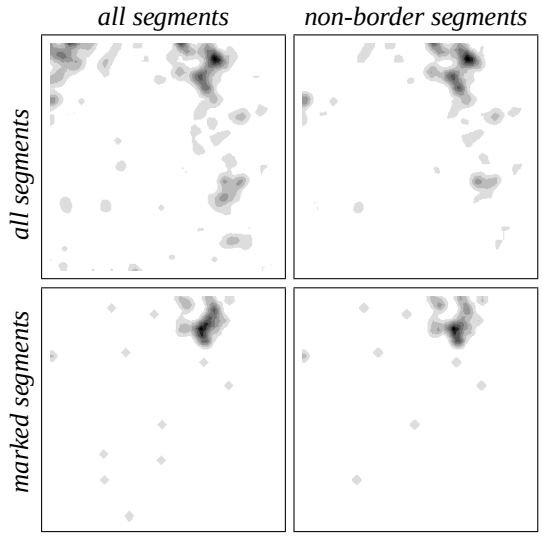

Fig. 5. Distributions of lobster image segments on the color moment SOM. The four distributions differ in the way how the mapped segments have been selected. Top left: all 328 segments; Top right: 193 segments not touching image borders; Bottom left: only the pointed segment from each of 41 images; Bottom right: 33 pointed non-border segments.

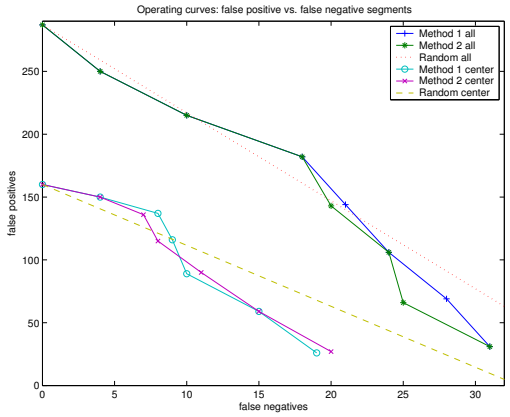

Fig. 6. Curves of false positive versus false negative keyword focusings on segments of lobster images.

7.2. Keyword focusing

In the second experiment we studied how well a keyword given on the image level can be focused on one of the automatically extracted image segments. As the starting point we again had the "lobster" category. We ran the experiment a total of four times. In the first two experiments we used all 41 lobster images and their 328 image segments, whereas in the latter two we used only the 193 segments which did not touch the borders of the images. In that case the effective size of the lobster category was reduced to 33, because in eight lobster images the segment containing the manually-marked reference point was a border segment.

For the focusing process we used two methods. In both methods the "lobster" keyword was originally assigned to all eight or fewer image segments of the images known to portray a lobster. Then the number of segments assigned the "lobster" attribute was sequentially reduced towards one. For each lobster image its segments were sorted in the decreasing order or qualification value produced by the PicSOM system by using the sum value from all the seven feature types.

The two focusing methods differed in the way how the segments with the smallest qualification values, ie. the least prototypical ones, were gradually rejected in the focusing process. In the method "1", the original convolution values were used in all steps whereas in the method "2" the convolutions were calculated again every time when the least representative segment of each image had been removed. The latter method thus results in a more iterative and gradual type of process.

Figure 6 shows how the number of false positive segments decreased when the least prototypical segments were discarded in the focusing process. It can be seen that the number of false negatives was meanwhile steadily increasing. It can be seen that both the methods "1" and "2" are

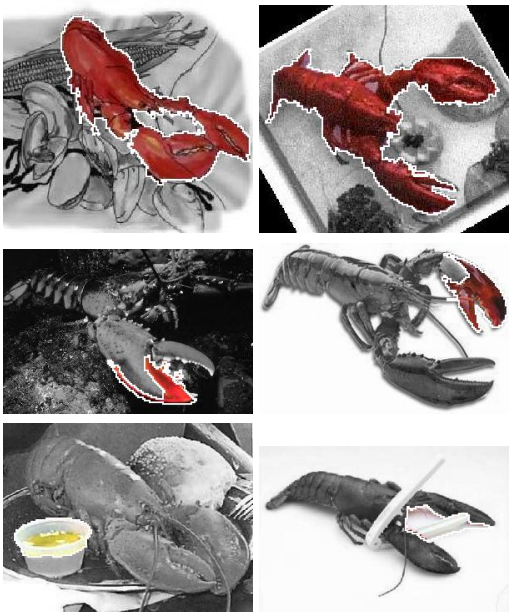

Fig. 7. Some lobster image segments which the keyword focusing process has interpreted to depict a lobster. In the top row the results are successful, in the middle row partially successful and in the bottom row failures. The segment area is bordered and shown in color, the surroundings are in grey.

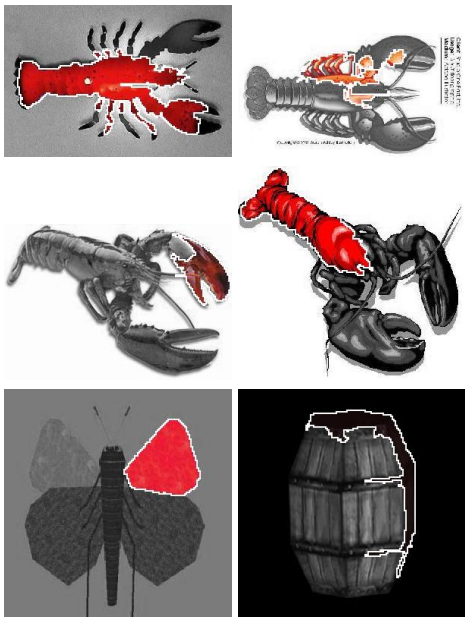

Fig. 8. The best image segments matching the keyword "lobster". The first four displayed segments had the largest score value and belonged to the "lobster" class. The last two segments were those with the largest score and not belonging to that class. The prototypical segment is shown bordered in color, others in grey.

to some extent able to focus the "lobster" keyword on one particular image segment more accurately than by random association shown as a reference method. In the final stage when the keyword is given to one segment only, the methods produce 31 false negative segments, ie. 76% of the total of 41 images, when all segments are used. When only non-border segments are counted, the final false negative count is 20, ie. 61 % of the total of 33. The false positive rates are all about half of that of random selection.

The result of this experiment does not reveal significant difference between the performances of the keyword focusing methods "1" and "2". One might still argue that with a larger category the iterative method "2" could be expected to show better performance. It is anyway clear that the result motivates to use only non-border segments for databases like this one, where the semantic content of an image is always a specific object located in the middle of a homogeneous background.

Figure 7 shows six image segments where the keyword focusing process has positioned the lobster keyword in its last round. The two top images are clearly successful and the detected segment covers the animal quite perfectly. In the middle row the success is partial, mostly due to oversegmentation resulting from a too homogeneous background. In the bottom row the process has found clearly wrong segments outside the animal.

7.3. Prototypical image segments

In the third experiment we wanted to find out how well the PicSOM system can extract the most representative image segments for the semantic category of "lobster". This was performed so that the lobster segments not touching image borders were marked as relevant on all seven bottom-level TS-SOMs. Then the convolutions with a triangle-shaped window of radius four map units was performed. After that the scores for each non-border segment on all maps were summed and the segments were ordered in the order of descending score.

Figure 8 shows first the four most prototypical image segments found. It can be seen that none of them are full lobster images due to the oversegmentation. Otherwise it is visible that the segments depict some characteristic part of a lobster, eg. its body, legs or claw. The last two segments were not placed among the most representative ones, but were the best ones not belonging to the lobster category. One can note that the color of the first one matches the characteristic color of a lobster, whereas the shape of the second segment might resemble that of lobster's claw or leg.

The result of this experiment reveals that the typical segments extracted from the lobster images really depict the lobster and not its surroundings. It is also clear that oversegmented images are more typical than correctly or underseg-

mented ones. The last part of the experiment shows that the segments which are typical for a lobster are not necessarily red in their color. This indicates that also other feature types than color are involved in the segment selection process.

8. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have demonstrated how the Self-Organizing Maps of the PicSOM CBIR system can be used to extract prototypical segments from image classes. Such image classes can be constructed from keyword annotations or from records of online user interaction with the CBIR system. It is worth to note that such semantic classes are in both cases defined on the image level and the system is still able to automatically focus the semantic information on specific image segments.

The experiments described in this work were performed with real-world data and truly automatic image segmentations, but were still merely preliminary "proof-of-concept" studies. More detailed analyses will be needed to compare the results of our approach with ones presented in the open literature. In such comparisons, both the recall–precision performance of the normal CBIR usage and the accuracy of the automatic segmentation subsystem should be studied.

The results of our experiment can be summarized by stating that prototypical segments for an image category could be extracted despite oversegmentation. On the other hand, the keyword focusing process was suffering from the inaccurate image segments. We have plans to ease this situation by extracting features from the image segments hierarchically. This will allow the most reliable segmentation to be determined during the process and thus the effective number of segments will adapt itself to the task.

9. REFERENCES

[1] K. Barnard, P. Duygulu, N. Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research, Special Issue on Machine Learning Methods for Text and Images*, 3:1107–1135, February 2003.

[2] P. Carbonetto, N. Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proceedings of the Eight European Conference on Computer Vision*, Prague, May 2004.

[3] Supplement No. 2 to CIE publication No. 15 Colorimetry (E-1.3.1) 1971: Official recommendations on uniform color spaces, color-difference equations, and metric color terms, 1976.

[4] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 540–547, New York, NY, October 2004.

[5] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *Proceedings of the Workshop on Generative-Model Based Vision*, Washington, DC, June 2004.

[6] ISO/IEC. Information technology - Multimedia content description interface - Part 3: Visual. 15938-3:2002(E).

[7] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer-Verlag, third edition, 2001.

[8] 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)*, pages 72–79, Angers, France, April 2003.

[9] M. Koskela, J. Laaksonen, and E. Oja. 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)*, pages 508–516, Dublin, Ireland, July 2004.

[10] J. Laaksonen, M. Koskela, and E. Oja. PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. *IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing*, 13(4):841–853, July 2002.

[11] V. Mezaris, H. Doulaverakis, R. M. B. Otalora, S. Herrmann, I. Kompatsiaris, and M. G. Strintzis. A test-bed for region-based image retrieval using multiple segmentation and the MPEG-7 eXperimentation Model: The Schema Reference System. In *Proceedings of 3rd International Conference on Image and Video Retrieval (CIVR 2004)*, pages 592–600, Dublin, Ireland, July 2004.

[12] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 275–278, Berkeley, CA, 2003.

[13] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[14] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos. Automatic image captioning. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, June 2004.

[15] M. Rummukainen, J. Laaksonen, and M. Koskela. An efficiency comparison of two content-based image retrieval systems, GIFT and PicSOM. In *Proceedings of International Conference on Image and Video Retrieval (CIVR 2003)*, pages 500–509, Urbana, IL, USA, July 2003.

[16] R. J. Schalkoff. *Pattern Recognition: Statistical, Structural and Neural Approaches*. John Wiley & Sons, Ltd., 1992.

[17] M. Stricker and M. Orengo. Similarity of color images. In *Storage and Retrieval for Image and Video Databases III (SPIE)*, volume 2420 of *SPIE Proceedings Series*, pages 381–392, San Jose, CA, USA, February 1995.