EMERGENCE OF SEMANTIC CONCEPTS IN VISUAL DATABASES

Jorma Laaksonen, Ville Viitaniemi and Markus Koskela

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

ABSTRACT

Content-based image retrieval (CBIR) systems can be used also for other purposes than online access to unannotated image databases. In particular, when a CBIR system is equipped with an automatic image segmentation subsystem, keyword annotations given on image level can be focused on specific image segments. In this paper, we show that our PicSOM CBIR system is able to reveal semantic knowledge not only from keyword annotations but also from recorded online use of the system. This automatically extracted high abstraction level visual information can then be used to further improve the accuracy of the system and to categorize the objects of the database with semantic concepts. This process, we claim, then helps to bridge the semantic gap between low-level visual features available for computers and the high-level semantic terms used by the humans. The results of the experiments described in this paper support that view.

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 extract semantic information from visual databases. We have developed a technique which enables a CBIR system to produce visual representations for semantic concepts associated with the database's objects.

The technique is based on our PicSOM CBIR system [1], where *relevance feedback* [2] is used as a method for *query refinement*. In this work, we replaced the relevance feedback with keyword-type annotations given to the images in a database. 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. It will be interesting to examine how accurately the keyword information given on the image level is focused on specific image segments.

In the last experiment of this paper, we will study how relevance feedback can be recorded from the user online and later analyzed offline to reveal the essential semantic target of the CBIR query. This approach bridges the use of relevance feedback for query refinement and the keyword annotations given to the images together under the same methodological umbrella. We claim that this kind of online analysis of query refinement can be used both for improving the accuracy of the iterative CBIR process and to uncover the semantic conceptions of the query.

The paper is organized as follows. In Section 2 we will first describe CBIR systems on a general level and then in Section 3 our PicSOM system in more detail. Section 4 then 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 5 and the actual experiments and their results in Section 6. Conclusion will be drawn and future directions discussed in Section 7.

2. CONTENT-BASED IMAGE RETRIEVAL

Content-based image retrieval addresses the problem of finding images relevant to the users' information needs from image databases. As we assume that the images generally do not have textual annotations, the searches are principally based on low-level visual features for which automatic extraction methods are available. The task of developing this kind of systems is very challenging due to the inherently weak connection between the high-level semantic concepts that humans naturally associate with images and the low-level features that the computer is relying upon. This impediment is generally known as the *semantic gap*.

One popular method to improve the retrieval performance in CBIR systems is to employ relevance feedback from the user, i.e. to adjust the subsequent retrieval process by using information gathered from the user's intra-query feedback. Assuming that the feedback is binary, i.e. either acceptance or rejection of the image, it can be seen as a process where the semantic content of an image is being thresholded to the user's binary decision on its relevancy vs. nonrelevancy in that particular query. 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 [3, 4], we have shown that this user interaction information can be used as a statistical feature to improve online retrieval even with-

out any semantic postprocessing.

Another important and rising technique in CBIR is the utilization of automatic and model-less or assumption-free 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 the whole images can be focused on the 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.

As mentioned above, also textual annotations or keywords can be made use of in CBIR. Two conceptually opposite alternatives for the use of the textual information exist. The first one is the technique commonly used in general purpose WWW search engines, 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 further in this paper. Namely, textual information is used offline to extract semantic concepts from the corresponding images. This will be feasible if the automatic segmentation will be successful enough and we are provided with enough images sharing the same keyword annotation.

3. PICSOM SYSTEM

The PicSOM [1, 5] system is a framework for research on content-based retrieval of images and other visual or non-visual information. As the name implies, PicSOM uses the Self-Organizing Map (SOM) [6] as its basic indexing method. Also other clustering methods are supported, for example K-means clustering was experimented with and compared to the SOM in [1]. Instead of the standard SOM version, PicSOM uses a special form of the algorithm, the Tree Structured Self-Organizing Map (TS-SOM) [7]. The hierarchical structure of TS-SOM is useful for two purposes. First, it drastically reduces the complexity of training large SOMs needed for indexing huge databases by exploiting the hierarchy in finding the best-matching map unit (BMU) for an input vector. Second, the hierarchical representation of the image database produced by a TS-SOM can be utilized in browsing and visualizing the images.

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 [8]. 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.

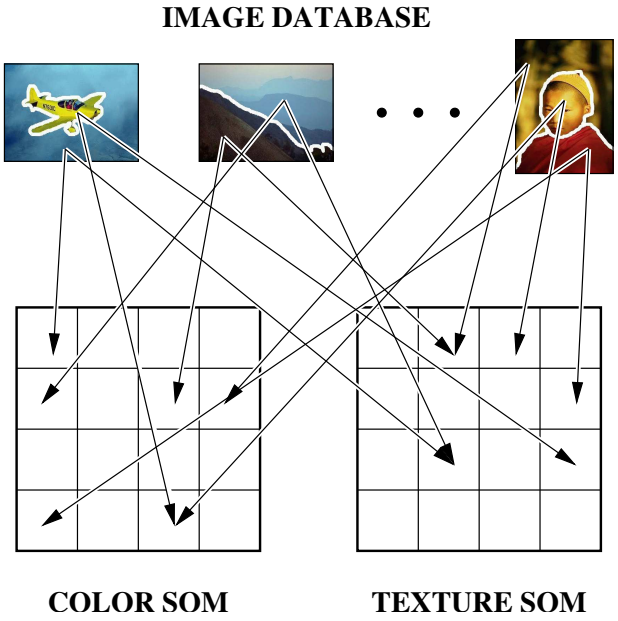

Figure 1. An example of using two parallel SOM indices for segmented images. The color and texture SOMs are trained with image segments and each segment is connected to its BMU on each SOM.

After training the SOMs, their 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 multiple feature indices

The relevance feedback mechanism of PicSOM is a crucial element of the retrieval engine. The basic method is only briefly presented here, see [5] for a comprehensive treatment. During a retrieval session, the user marks images that she considers relevant as positive, and the remaining ones are implicitly regarded as negative. As the first step, the SOM units are awarded a positive *score* or *response value* for every relevant image mapped in them, resulting in an attached positive impulse. Likewise, associated non-relevant images result in negative scores and impulses. If the total numbers of relevant and non-relevant shown images are $N^+(n,m)$ and $N^-(n,m)$ at nth query round on mth SOM, the positive and negative scores are simply the inverses:

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

For each SOM, these values are mapped from the shown images (and thus rated either as positive or negative) and

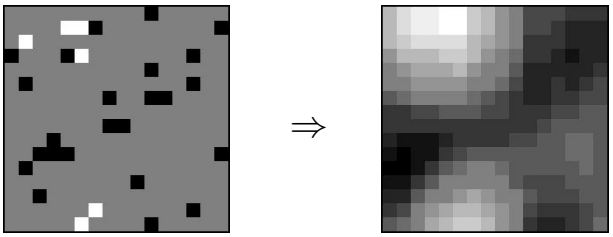

Figure 2. An example of how a SOM surface is convolved with a tapered window function. On the left, images selected and rejected by the user are shown with white and black marks, respectively. On the right, the convolution result, where relevance information is spread around the centers.

their all segments 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, images 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 the BMUs of the shown images on the SOMs. This can be done by convolving the sparse value fields with tapered (e.g. 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 images. By locating the corresponding images in all indices, we get their scores with respect to different feature extraction methods. The total *qualification values* for the candidate images are then obtained simply by summing the corresponding responses. 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.

3.3. Use of 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 or a combination thereof.

Despite not being 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. Assuming that potentially useful segmentations can be produced, a method to utilize the segmentation results in the CBIR system is required.

In the PicSOM system the use of segmentation results is made possible by the general ability of the system to deal with hierarchical objects [9]. Feature indices can be defined to all levels of the hierarchy. Typically the queries have certain level objects as targets. These objects are shown to the user and relevance feedback is obtained for objects on this level. During the query the system first distributes this relevance to feature indices on all hierarchy levels. For each index, qualification values are then calculated. Finally, the values are propagated back in the object hierarchy and combined to target level qualification values.

In case of images and their segmentations the object hierarchy has two levels. On the top level are the whole images, on the bottom level the segments they consist of. Both image and segment levels can have features associated to them, although in the experiments described in Sections 5 and 6 we calculate features only for the segments. Query targets and displayed items are the top level objects, i.e. images. By default, relevance is propagated from the top level so that each segment that belongs to a relevant top level object receives same amount of relevance. After the similarity of image segments has been evaluated, the qualification value is propagated back to the top level by summing together qualification values from all the segments a certain image consists of.

Figure 3 displays one example of an image where the automatic segmentation has been successful enough for utilization in CBIR. A hat in the head of the mannequin on a catwalk can been seen extracted as segment "2". However, the difficult shirt has been split into two segments, "4" and "5".

4. IMAGE SEGMENTATION AND SEMANTICS

The relationship between image segmentation and semantic concepts has become a subject of recent intensive study in the field of CBIR. The goal has been given various names ranging from "image-to-word transformation" [10], "matching words and pictures" [11], "image auto-annotation" [12], "automatic image captioning" [13], to "automatic image annotation" [14], depending 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 [10], Expectation Maximization (EM) [11], Support Vector Machines (SVM) [14], Latent Semantic Analysis (LSA) [12], and Markov random fields (MRF) [15].

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

Figure 3. A catwalk photograph and the result of its automatic segmentation in segments marked as '1'–'8'. A reference point for the hat has been manually marked in the original image. This auxiliary information was not utilized in the segmentation but in performance evaluations.

SOMs in general CBIR has, however, been demonstrated by our earlier studies and comparisons, e.g. in [1, 16]. 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 unique feature not shared by the other techniques. In that process, segments which depict the background or otherwise meaningless 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 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 6 will contain four steps. First, we will study how different semantic classes are mapped on SOMs of different features. When one semantic class represents a low-level defining property such as color and the other a higher-level concept, their distributions on feature maps specific for either colors or shapes should differ qualitatively.

Second, we will examine two different ways of focusing the keyword annotation 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 typical representative it is as a segment for that keyword. For each image, its segments can then be ordered in the order of descending qualification value. When the least representative 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 and genuinely focusing as the least trustworthy and least probable segments are being iteratively neglected.

Third, we will try to find the most representative 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. As our last experiment, we will study genuine user relevance from recorded online interaction. In that way we will see whether the system can reveal the semantic concept the user has had in her mind when performing the query.

5. EXPERIMENTAL DATA

5.1. Database

Our "Models" database consisted of a 900 image subset of the commercial Corel database of color images [17]. The subset images depict people, mostly mannequins on a catwalk, but also other kinds of people in similar types of scenes. The images measure 384×256 or 256×384 pixels in their size.

The semantic image classes in the Models database were defined by presence/absence of objects possessing certain properties. In some of the classes ("red", "blue") object color was used as the defining property. Other semantic classes were defined for higher-level concepts related to clothing such as "hat" and "trousers".

5.2. Image segmentation

The images in the Models database were segmented in two steps. In the first step isodata variant of K-means algorithm [18] 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 [19] 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:

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

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. After the segmentation the original image sizes were restored. As the result of the segmentation we thus had a database of 8100 visual objects, of which 900 were images and 7200 image segments.

5.3. Features

Five different features were used to describe the visual content of the segments. Low-level 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 [20] 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. Three MPEG-7 content descriptors [8], EdgeHistogram, HomogeneousTexture and RegionShape, were used as somewhat more sophisticated features. A separate TS-SOM was trained for each feature with levels containing 4×4, 16×16 and 64×64 map units.

6. EXPERIMENTS AND RESULTS

For performing the experiments we needed ground truth data for the locations of the objects of some semantic class in the images. We selected the "hat" class, an example of which was already displayed in Figure 3. In that figure, a manually specified reference point for the hat 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 it performed the automatic segmentation and focusing tasks.

Some more "hat segments" in the Models database are shown in Figure 4. As can be seen, the segmentation has been only partially successful in some of the cases. In the two cases on the first row, parts of faces and/or hair has been included in the same segment as the hat. The images on the bottom row depicts cases where the hat has erroneously been broken in two segments. The images serve as to demonstrate the difficulty of the automatic image segmentation.

6.1. Class distributions

In the first experiment we studied how two different semantic classes are distributed on the SOM surfaces in the case of two SOMs resulting from different feature extractions. The semantic classes were the "hat" class of quite high abstraction level and the "red" class which should match better with lower-level pixel properties. The two features were the higher-level MPEG-7 RegionShape feature and the lower-level CIE L*a*b* central moment feature.

Figure 5 shows the surface distributions of the classes on the 64×64-sized bottom level SOMs after the convolution with a triangular-shaped window of radius six units. Darker shades represent denser distributions in the respective parts of the map. It can be seen that the color moment

Figure 4. Some automatically extracted hat segments of the Models database. The segment area is bordered and shown in color, the surroundings are in grey.

feature is unable to focus the hat class in any specific part of the SOM. On the other hand, that very same feature maps the segments from images given the keyword "red" very densely in one corner of the map. Opposite observations can be made from the two distributions on the RegionShape SOM.

The result of this experiment is thus that the distributions of the segments from the different semantic classes can be seen to be highly concentrated on different SOM surfaces. This phenomenon seems to be true even though the segment data for the classes 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. The effect of the false positives might therefore be regarded as neglectable noise.

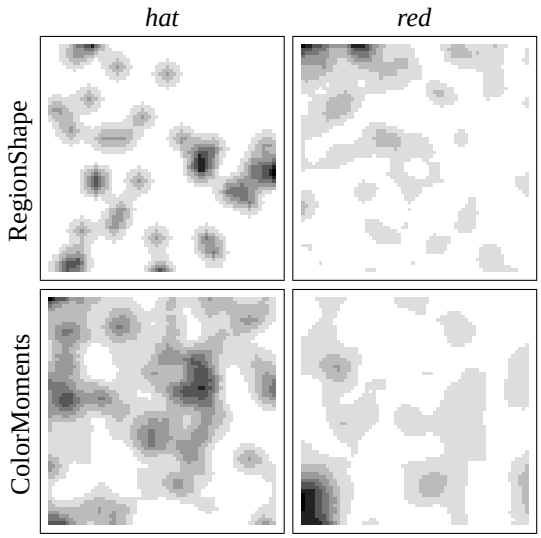

Figure 5. Distributions of image segments extracted from images given the keywords "hat" and "red" on the SOM surfaces of RegionShape and ColorMoments features.

6.2. Keyword focusing

In the second experiment we studied how well a keyword given on image level can be focused on the automatically extracted image segments. As the starting point we had a subset of 34 images selected from the Models database. These images were selected so that (a) the image portrayed a person wearing or holding a hat, and (b) the automatic segmentation had been successful enough to locate the hat at least partially. By the latter condition we mean that the hat or a major part of it had to be a notable part of some image segment. Such segments were seen in Figure 4. This human pre-screening of the studied images can be seen as a way to ease the task of the CBIR system. However, it was here strongly motivated by the small size of the hat image collection, where the spurious effects of noisy segmentation would otherwise been dominant.

We ran two experiments in which the "hat" keyword was originally assigned to all eight image segments of the images known to portray a hat. In both experiments the number of segments assigned the "hat" attribute was then sequentially reduced towards one. The individual segments of each hat image were sorted in the decreasing order of qualification values produced by the PicSOM system as the sum of all the five features used.

The two experiments differed in the way how the segments with the smallest qualification values, i.e. the least representative ones, were gradually rejected in the focusing process. In the first method, the original convolution values were used in all steps, whereas in the second method, the convolutions were calculated again every time when the least representative segment of each image had been removed.

Figure 6 shows how the number of false positives segments decreased when the least representative segments were discarded in the focusing process. It can be seen that the number of false negatives was simultaneously steadily increasing. The final result of the first method can be seen to be worse than that of corresponding random selection of positive segments. The final result of the second method, however, is better than that of the random process.

The result of this experiment supports the idea of using an iterative reduction process in performing keyword focusing from image level to specific segments. The class of hat images was quite small in this experiment and we believe that with a larger semantic class the result would be better. Verifying that will, however, call for laborious manual creation of the ground truth data for the experiment.

6.3. Most representative image segments

In the third experiment we wanted to find out how well the PicSOM system can extract the most representative image segments for a semantic concept specified with a pair of keywords. We used the keywords "trousers" and "red". 129 images had the keyword "trousers" and 126 images were marked as "red" in the Models database. The intersection of these two sets contained 24 images; this does not mean that there were that number of red trousers

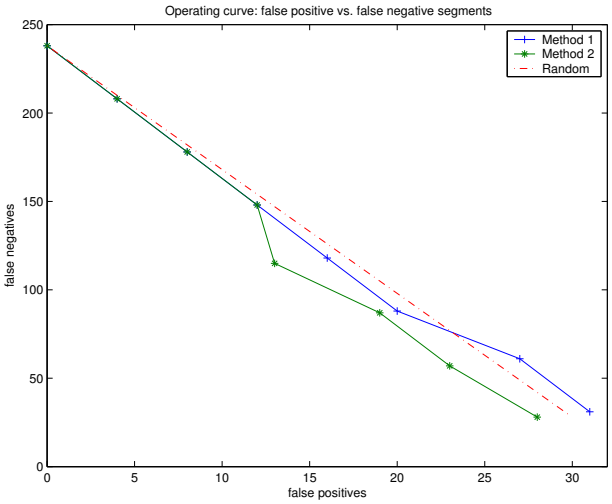

Figure 6. Curves of false negative versus false positive keyword focusings on segments of hat images.

as these two keywords might have been associated with different objects in one image. The precise number of red trousers in the database was manually checked to be 11.

Figure 7 shows the six most representative image segments found. It can be seen that one of them is unquestionably a successful match and another is otherwise perfect, but the automatic segmentation has failed to separate the red trousers and the red shirt. These two images were among the 11 correct ones. On the other hand, the first and the sixth images are failures only with regard to trousers, they still depict red pieces of clothing.

All the six returned images really were included in the "red" class but only the two successful ones in the "trousers" class. Therefore it might be deduced that the color attribute dominates in the result over the shape attribute. This can be interpreted as an indication of the greater difficulty for the PicSOM system in describing semantic shape properties than semantic color properties.

6.4. Revelation of user's semantic target

The last experiment demonstrates how the system could reveal the semantic target of an iterative relevance feedback query performed by a real human user of the system. The user was silently thinking of finding sleeveless shirts in the collection. There was no prior processing made that would have used this same subclass of images. Neither was the suitability of the automatic segmentation ensured in any way for this specific target. During the retrieval session, a total of 140 images were displayed on seven iterative query rounds. The user indicated 11 images as relevant to the query.

In processing the data, we applied one additional constraint by requiring that the image segments were not allowed to touch the image borders. This additional restriction was motivated by our earlier experience that the system would otherwise tend to first find the most typical image backgrounds in the set of relevant-marked images.

Figure 7. The most representative image segments matching the query "red trousers". The representative segment is shown bordered and in colors while others in grey.

Figure 8. Image segments resulting from the user interaction experiment where the user was searching for sleeveless shirts. The representative segment is shown bordered and in colors while others in grey.

Figure 8 shows the six most representative image segments that resulted when the user interaction data was processed. It can be seen that the operation of the system has been successful in four cases out of six. Even the remaining two images depict sleeveless shirts, but the correct image segment has not been successfully located.

7. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have demonstrated how the Self-Organizing Maps of the PicSOM content-based image retrieval system can be used to extract semantic concepts 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 noteworthy that the 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 showed that this focusing was performed correctly in most of the cases.

We believe that similar techniques can prove to be useful also when used in a tighter connection with the automatic image segmentation stage. Information on the class distributions could guide the segmentation algorithm in selecting the most probable segment combinations among those available in the various stages of the process.

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

8. ACKNOWLEDGMENTS

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.

9. REFERENCES

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

[2] Xiang Sean Zhou and Thomas S. Huang, "Relevance feedback for image retrieval: A comprehensive review," *Multimedia Systems*, vol. 8, no. 6, pp. 536–544, April 2003.

[3] Markus Koskela and Jorma 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)*, Angers, France, April 2003, pp. 72–79.

[4] Markus Koskela, Jorma Laaksonen, and Erkki 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)*, Dublin, Ireland, July 2004, pp. 508–516.

[5] Jorma Laaksonen, Markus Koskela, Sami Laakso, and Erkki Oja, "Self-organizing maps as a relevance feedback technique in content-based image retrieval," *Pattern Analysis & Applications*, vol. 4, no. 2+3, pp. 140–152, June 2001.

[6] Teuvo Kohonen, *Self-Organizing Maps*, vol. 30 of *Springer Series in Information Sciences*, Springer-Verlag, third edition, 2001.

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

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

[9] Mats Sjöberg, Jorma Laaksonen, and Ville Viitaniemi, "Using image segments in PicSOM CBIR system," in *Proceedings of 13th Scandinavian Conference on Image Analysis (SCIA 2003)*, Halmstad, Sweden, June/July 2003, pp. 1106–1113.

[10] Yasuhide Mori, Hironobu Takahashi, and Ryuichi 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.

[11] Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research, Special Issue on Machine Learning Methods for Text and Images*, vol. 3, pp. 1107–1135, February 2003.

[12] Florent Monay and Daniel Gatica-Perez, "On image auto-annotation with latent space models," in *Proceedings of the eleventh ACM international conference on Multimedia*, Berkeley, CA, 2003, pp. 275–278.

[13] Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos, "Automatic image captioning," in *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, June 2004.

[14] Jianping Fan, Yuli Gao, and Hangzai 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*, New York, NY, Oct. 2004, pp. 540–547.

[15] Peter Carbonetto, Nando de Freitas, and Kobus Barnard, "A statistical model for general contextual object recognition," in *Proceedings of the Eight European Conference on Computer Vision*, Prague, May 2004.

[16] Mika Rummukainen, Jorma Laaksonen, and Markus 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)*, Urbana, IL, USA, July 2003, pp. 500–509.

[17] "The Corel Corporation WWW home page, `http://www.corel.com`," 1999.

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

[19] "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.

[20] Markus Stricker and Markus Orengo, "Similarity of color images," in *Storage and Retrieval for Image and Video Databases III (SPIE)*, San Jose, CA, USA, February 1995, vol. 2420 of *SPIE Proceedings Series*, pp. 381–392.