Psychophysical Evaluation for a Qualitative Semantic Image Categorisation and Retrieval Approach

Zia Ul-Qayyum*, A. G. Cohn, and Alexander Klippel

ziaqayyum@uaar.edu.pk,a.g.cohn@leeds.ac.uk,klippel@psu.edu

Abstract. This paper details the behavioral evaluation of a qualitative image categorisation and retrieval approach using semantic features of images. Content based image retrieval and classification systems are highly active research areas and a cognitively plausible image description can improve effectiveness of such systems. While most approaches focus on low level image feature in order to classify images, humans, while certainly relying on some aspects of low level features, also apply high-level classifications. These high-level classification are often qualitative in nature and we have implemented a qualitative image categorisation and retrieval framework to account for human cognitive principles. While the dataset, i.e. the image database that was used for classification and retrieval purposes contained images that where annotated and therefore provided some ground truth for assessing the validity of the algorithm, we decided to add an additional behavioral evaluation step: Participants performed similarity ratings on a carefully chosen subset of picture implemented as a grouping task. Instead of using a predefined number of categories, participants could make their own choice on a) how many groups they thought were appropriate and b) which icons/images belong into these groups. The results show that the overall underlying conceptual structure created by the participants corresponds well to the classification provided through the algorithm.

Key words: Qualitative spatial representation, qualitative similarity, image categorisation and retrieval, psychophysical evaluation

1 Introduction

Content based image classification and retrieval systems have gained more importance and have been an active research area in recent years with an increasing requirement of robust and flexible techniques to handle dynamic and complex visual content in large volumes of digital data at a higher semantic level. Most of the research in this area is primarily based on the use of low level image features like colour, texture, shape etc [1–3]. Although low level image processing algorithms and methodologies are quite mature, such systems are hard to be used effectively by a novice due to the semantic gap between user perception and understanding, and system requirements. Furthermore, humans tend to describe scenes using

* corresponding author

natural language semantic keywords/concepts like sky, water etc and specify retrievals "an image with water next to fields and having sky at the top ..." or ".... has a small lake with high peaks of mountains behind and fields on left....". This suggests that the use of underlying semantic knowledge in a qualitative representation language may provide a way to model the human context and a natural way to bridge this semantic gap for better image understanding, categorization and retrieval capabilities. In an earlier work, we had, therefore proposed a qualitative knowledge driven framework for image categorisation and retrieval using local semantic content of images [4, 5].

There are number of methods discussed in the literature to assess the similarity of objects, events and spatial relations [6–8]. Keeping in the characteristics of the proposed categorisation and retrieval framework [4, 5], a psychophysical evaluation is an obvious choice and is therefore used as an alternative evaluation approach in order to perform:

Categorization Evaluation: The participants were provided with an appropriately chosen set of images and asked to group images into as many number of images as they deemed fit, and to assign a label/keyword to each group.
This provides an evaluation of whether the number of classes from data set matches those selected by the study participants. Moreover, it provides a validation of the *'ground truth'* as participants will place images in a group based on their notion of 'similarity'; so comparing the images in a group created by a participant vs the pre-assigned class labels provides a cognitive evaluation of the *'ground truth'*. Moreover, it can reveal level of confusion between certain existing classes of images in ground truth.

Retrieval Evaluation: The above task provides an evaluation of the image retrieval approach as well, because if the participants *'mostly'* place the images of the same pre-assigned class in one group - it implies that humans also choose certain images to belong to particular group if they are 'visually similar (in qualitative terms)'. Furthermore, the participants are requested to give a natural language description of sampled images in each group they create to get an idea about their notion of 'relative similarity' in images of each group. This provides an assessment of the QSD-based approach for categorization and *'qualitative similarity measure'*-based retrieval tasks.

A metric based evaluation of qualitative approaches is always regarded a difficult task because of the nature of qualitative representations. In order to evaluate the overall performance of our qualitative knowledge driven approach to image categorization and retrieval [4, 5], manually assigned categories for the image data were used. However, to evaluate the efficiency and effectiveness of the qualitative representations based categorisation and retrieval framework , a psychophysical evaluation approach is proposed. Qualitative representations of image content is a cognitively more plausible way to describe images; psychophysical evaluation of such approaches, therefore is an obvious choice and is used as an alternative approach.

Before discussing these psychophysical evaluation experiments, the baseline approach used in our previous work on *qualitative semantic image description* based categorisation and retrieval is briefly described to make this paper self-contained.

1.1 Baseline Approach for Qualitative Categorisation and Retrieval

There has been substantial work done in areas like computer vision, pattern recognition etc on detecting, recognising and categorizing objects of interest in images and other application domains. Most of the techniques in image description and categorisation has been based on describing the image using low level features such as colour and texture [9–11, 1, 2, 12], whereas semantic scene description is arguably a natural way to describe image features and it may bridge the gap between a human's description and that of a computer. Although, there has been more recent work on the use of semantic content in such systems [13–15, 4, 16, 17] and using spatial context and spatial relations in image analysis/automatic annotation [18–20], the need to bridge the semantic gap between low level synthetic features and high level semantic meanings has been regarded as an open problem [21].
In an earlier work, we had proposed a qualitative knowledge-driven semantic modelling approach for image categorisation and retrieval [4, 5]. It was demonstrated that how category descriptions for a set of images can be learned using qualitative spatial representations over a set of local semantic concepts (LSC) such as sky, grass and categorizing the images into one of six global categories (e.g. Coasts, Landscapes with Mountains etc). Four kinds of qualitative spatial representations, namely *'RSizeRep'* based on relative size of each of concept occurences in each image, *'AllenRep'*based on Allen relations [22] and are measured on the vertical axis between the intervals representing the maximum vertical extent of each concept occurrence, *'ChordRep'* based on Morchen and Ultsch's work [23] in which each row in grid like image is a *chord* which is labelled by concatenating the corresponding patch labels in that row and the *'TouchRep'* to model whether one patch type is spatially in contact with another in the image, were used in these experiments.

Qualitative semantic image descriptions were obtained by applying the above QSRs and their variants to learn class descriptions and categorise images into one of a fixed number of semantic classes (such as sky_clouds, coasts, landscapes_with_mountains (lwm), fields, forests and waterscapes) [4]. It demonstrated that supervised learning of a pure qualitative and spatially expressive representation of semantic image concepts can rival a non-qualitative one for image categorization [24]. In order to evaluate the performance of this approach to image categorization, manually assigned categories for the images were used.
Image categorisation is one of the critical steps for retrieving images. We used the same semantic descriptions as in the categorization work summarized above to evaluate the validity of our hypothesis that the qualitative representations which were able to effectively support categorization may also provide an effective and natural way to support content-oriented querying approach [5]. Further details of this approach can be viewed in our previously published work [4, 5].

A collection of seven hundred natural scenes images was used in this work. The labelled data set was provided by Julia Vogel who has developed a quantitative semantic modelling framework for image categorisation and retrieval [16].

2 Psychophysical Experimental Setup

As already argued previously, a qualitative representation of image content is cognitively plausible way to describe images and perform categorization and retrieval based on such descriptions. Moreover, it was argued that the *qualitative similarity measure* used to accomplish image retrieval tasks is a natural way to arrange the images in order of their respective *typicality* with respect to a query image. Psychophysical evaluation of such approaches, therefore, is an obvious choice and is used as an alternative approach to evaluate our qualitative knowledge driven approach for categorisation and retrieval of images. The details of these experiments and corresponding results are presented in the following sections.

2.1 Experiments:

This sections details the behavioural validation of the algorithm discussed in the above section. The goal is to provide validation, or the ground truth for the algorithm that has been advised on the basis of a qualitative spatial representation.

Participants: Twenty students including two females were used to conduct these experiments. They were selected from different subject areas like chemistry, medevial studies, mathematics, geography, computing etc with average age of about thirty two years (age of participants ranges between 24 to 52 years). The reason to choose participants with such diverse academic backgrounds was to achieve a varied understanding and perception of the participants with regard to the type of experiment and the nature of the data.

Materials: We used 72 icons to depict landscapes. The 72 icons were chosen from a data set that contains 700 pictures of natural landscapes. The following procedure was applied to make this selection and to obtain a representative example. A representative example should be first of all be related to the six groups: landscape with mountains, forest, field, coast, waterscape, and sky_ clouds; corresponding to the hypothesis that the algorithm classifying images into these six categories is operating on cognitively adequate principles. Hence, we adopted the following strategy: Consider the average penalty weights for all images in a class in the corpus based on all four base qualitative representations used in actual categorization and retrieval experiments. It is worth reporting here again that the sequence of penalty weights for each image in each of the six classes is generated by taking each image of the corpus as a query image iteratively in a leave-one-out fashion and k such sequences per class are generated, where k is the total number of images in each class. The penalty weights for each image are then averaged to get a single sequence of weights for each class. This approach rules out bias due to choice of query image in these experiments. The respective average penalty weights sequences for each of the pre-assigned category of images are then sorted in ascending order to arrange images in decreasing order of qualitative similarity. For each image category, twelve images are selected by taking six, four and three images from top, middle and bottom of the sorted sequence of k images for each category which means that the

selected images have varied typicality level with respect to a class. The sampled data set thus contained seventy two images from all six categories in the corpus.

Procedure: The icons were integrated into the grouping tool. Figure 1 shows a screenshot of a mimicked ongoing experiment. The grouping tool divides the screen into two parts. On the left side, participants find the stimulus material, consisting of all icons 72 depicting different landscapes, which are placed in a different random order for each participant. The number of icons required scrolling to access all items. (Scrolling is a common procedure in interacting with computer interfaces; no problems were expected nor found during the experiments.) The right side of the screen is empty at the start; participants move icons to this side in order to group them during the experiment. The interface was kept simple so that participants could perform only the following three actions: Create a new group, Delete an existing group, Finish. The experiment took place in the participants' workplace

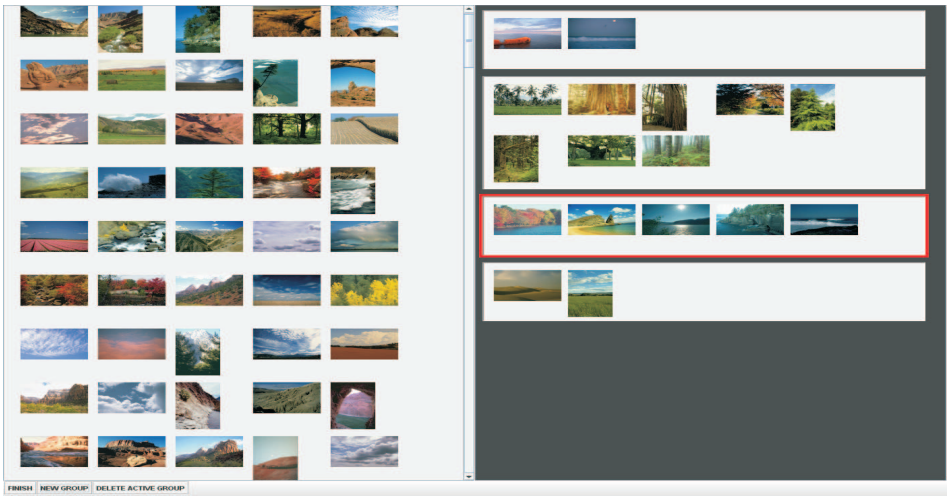

Fig. 1. The grouping tool (snapshot from an ongoing experiment). On the left side, icons representing landscapes are presented in random order. On the right side, a participant has started to group icons according to her categories of landscapes.

and were tested individually. After arriving and obtaining some basic information, we explained to them the general procedure of the experiments and demonstrated the functionality of the grouping tool in form of a mock-up grouping task using different animal. Participants were advised that there is no right or wrong number of groups. In contrast to other grouping tasks [24] there was no predefined set of groups that participant had to create. The algorithm, though, assigns six different labels for groups. We therefore tested two aspects of validity at the same time: on the one hand, whether the categorization of landscapes into 6 distinct groups is something that is reflected in the behavioral data created by the participants,

and, on the other hand, whether the category boundaries are comparable to the judgment of human cognitive agent.

After the grouping task, participants were presented with the groups that they had created and were asked to provide two verbal labels for each group, i.e., a linguistic description for the kind of landscape a particular group represents. They were first asked to provide a global or general label and additionally for a slightly more elaborate description of the content within each group. Again, the participants were free in their choice of labels (compared to other experiments in this a predefined set of labels had to be assigned). The participants had a varied linguistic background and we refrain here from an in depth discussion and just use example of the labels participants provided.

2.2 Results

The groupings of each participant result in a 72 x 72 similarity matrix, the number of the icons used in this experiment on each axis. This matrix allows us to code all possible similarities between two icons simultaneously for all icons in the data set; it is a symmetric similarity matrix with 5184 cells. Similarity is coded in a binary way; any pair of icons is coded as '0' if its two items are not placed in the same group and '1' if its two items are placed in the same group. The overall similarity of two items is obtained by summing over all the similarity matrices of individual participants. For example, if two icons (called A and B) were placed into the same group by all 20 participants, we add 20 individual '1's to obtain an overall score of 20 in the respective cells for matrix position AB and BA.

This data was subjected to several agglomerative hierarchical cluster analysis. Cluster analysis identifies "natural" groupings within data that minimize within-group and maximize between-group variation. Agglomerative cluster analysis initially treats each case as a separate cluster, recursively combining the most similar clusters until all clusters are combined. After Aldenderfer and Blashfield [25], the cluster analysis used can be summarized according to the following five criteria:

1. Software: CLUSTAN software were used to perform the cluster analysis.
2. Similarity measure: Squared Euclidean distance was used as the analysis similarity measure.
3. Cluster method: Four common cluster methods were used and compared: complete linkage (minimizes the maximum distance between clusters), average linkage (minimizes the average distance between clusters), and Ward's methods (minimizes the distance to the center mean).
4. Number of clusters: In many studies, deciding on the appropriate number of clusters is a critical step in the analysis. Our results confirm a clear 6 cluster solution (see below).
5. Validation: The results were initially validated through the comparison of the results of several cluster methods (item 3 above) suggested by Kos and Psenicka [26]. Further validation was performed by repeating the analysis with two randomly selected sub-groups (a procedure suggested by Clatworthy et al. [27]).

Figure 2 shows a cluster analysis using Ward's method. The dendrogram shows a clear 6 cluster solution. The labels on the left side of the dendrogram provide the picture (running) number and the ground truth that was calculated through the algorithm, i.e. the category in which they would have been placed based on the algorithm. This classification shows the successful categorization of pictures by the algorithm compared to the results of this user study, i.e. the dendrogram on the left side. From the 72 pictures all picture in the category sky-clouds were categorized correctly, all pictures in the category forest, all pictures in the category field, all but two pictures in the category landscape-with-mountain (lwm), all but two pictures in the category waterscape, and all but one picture in the category coast. Hence, we have a total of 5 out of 72 pictures placed in other than expected groups.

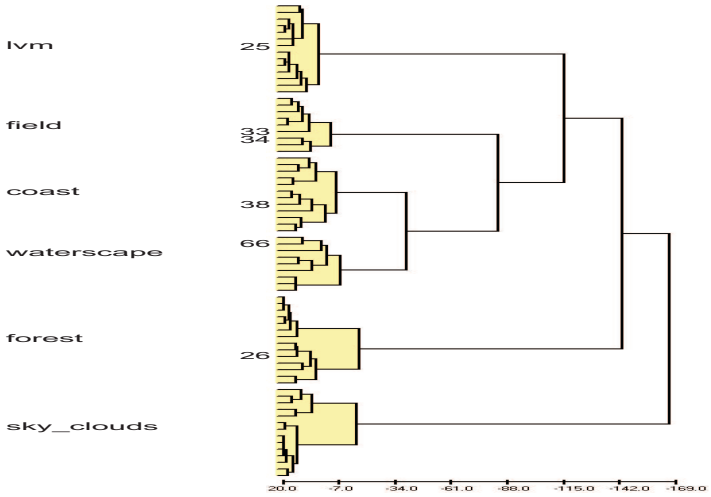

Fig. 2. The figure shows a dendrogram as the result of a cluster analysis using Wards method. Indicated is a six cluster solution. On the left side the predicted categories based on the algorithm are indicated: landscape with mountain (lwm), field, coast, waterscape, forest, and sky-clouds. The derived clustering structure indicates strong agreement between the human participant test and the results of the algorithm, i.e. only 6 pictures are not grouped according to the expected grouping provided by the algorithm. The numbers of the misclassified pictures are written next to the dendrogram.

Validation As indicated above, the validation was performed by comparing different clustering methods as suggested by Kos and Psenicka [26]. Figure 3 compared the results of Ward's, average linkage, and complete linkage. This comparison shows nearly identical cluster structure. The cluster structure at the cut-off point for the six-cluster solution (shaded areas) is shown (the number of misplaced pictures is: 6 for Ward's method, 7 for average linkage, and 8 for complete linkage). A further validation technique was applied (which also can be seen as a means

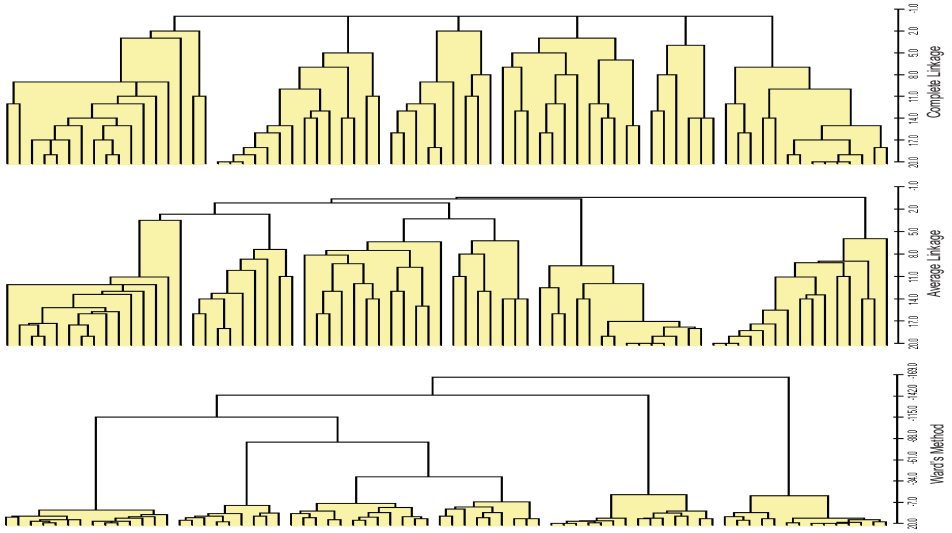

Fig. 3. Depicted are the results from three different cluster methods: Wards, average linkage, and complete linkage. While average and complete linkage to not exhibit the same clear 6 cluster solution, the assignment of pictures into 6 clusters is nearly identical across all three methods.

to test whether the number of participants was sufficient), i.e. the participants were split into two groups using numbers from the website random.org. We performed Ward's method on both groups and the results show that both sub-groups exhibited nearly identical cluster structures. In both groups some icons were 'incorrectly' (compared to the algorithm) placed into groups (7 and 10) (see Figure 4).

3 Conclusion

The results of the above experiments provide an evaluation of our approach of categorization and retrieval in natural scene images using qualitative semantic image descriptions - QSID. It was argued that an expressive image description is key to success in categorization and retrieval approaches and that humans tend to describe different categories using qualitative terms and relations. Qualitative representations provide a natural way to model human cognition and therefore, QSID's based on an image's local semantic content provide a more plausible way to represent/describe the images. (The verbal descriptions of each group in these experiments has not been formally analysed, but participants used qualitative keywords like *more, above, below, meet etc* and local semantic features like *sky, grass, water etc* in their descriptions which is similar to QSIDs used to model the image content in our categorization and retrieval approach.) The proposed framework also helps to reduce the semantic gap between humans and most of content based image retrieval systems. Moreover, the results of these experiments

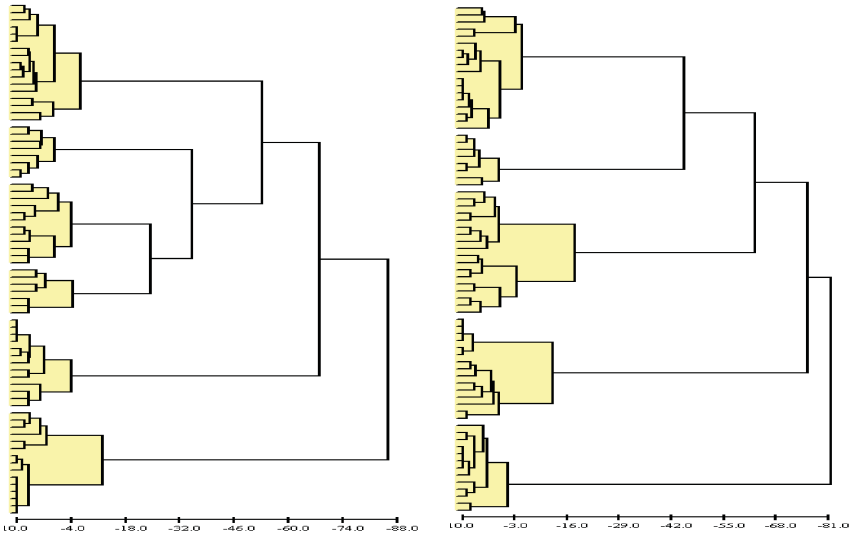

Fig. 4. Participants were randomly assigned to two groups. Both groups were subjected to Wards method. Depicted are dendograms for both groups (i.e. 10 participants each).

tend to validate the *ground truth* of the preassigned category label in terms of number of classes and indicate , at the same time, the confusion in pre-assigned image labelling.

References

1. Rui, Y., Huang, T.S., Chang, S.F.: "Image Retrieval: Past, Present and Future". Journal of Visual Communication and Image Representation **10** (1999) 1–23
2. Veltkamp, R.C., Tanase, M.: "Content-Based Image Retrieval Systems: A Survey". Tech. Rep. UU-CS-2000-34, Univ. Utrecht, Utrecht, The Netherlands (2000)
3. Deb, S., Zhang, Y.: "An Overview of Content-based Image Retrieval Techniques". Proceedings of 18th Int Conf on Advanced Information Networking and Application (AINA04) **1, issue 2004** (2004) 59–64
4. Qayyum, Z.U., Cohn, A.G.: "Qualitative Approaches to Semantic Scene Modelling and Retrieval". Proceedings of 26th SGAI Int. Conference on Innovative Techniques and Applications of Artificial Intelligence, Research and Development in Itelligent Systems, XXIII, Springer-Verlag (2006) 346–359
5. Qayyum, Z.U., Cohn, A.G.: "Image Retrieval through Qualitative Representation over Semantic Features". Proceedings of 18th British Machine Vision Conference (BMVC'07) (2007) 610–619
6. Rogowitz, B.E., Frese, T., Smith, J.R., Bouaman, C.A., Kalin, E.: "Perceptual Image Similarity Experiments". In B.E. Rogowitz and N.P. Thrasyvoulos (Eds.), Proceedings of SPIE: Human Vision and Electronic Images III (1998) 576–590
7. Tversky, A.: "Features of Similarity". Psychological Review **84** (1977) 327–352
8. Knauff, M., Rauh, R., Renz, J.: "A Cognitive Assessment of Topological Spatial Relations: Results from an Empirical Investigation". In S.C. Hirtle and A.U. Frank (Eds.); Spatial Information Theory: A Theoretical Basis for GIS, LNCS 1329, Springer (1997) 193–206

9. Szummer, M., Picard, R.: "Indoor Outdoor Image Classification". Proceedings of IEEE Int. Workshop on Content-based Access of Image and Video Databases (CAIVD), India (1998) 42–51
10. Vailaya, A., Jain, A., Zhang, H.: "On Image Classification: City vs Landscape". Pattern Recognition **31(12)** (1998) 1921–1935
11. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.J.: "Image Classification for Content-based Indexing". IEEE Transactions on Image Processing **10(1)** (2001) 117–130
12. Enser, P., Sandom, C.: "Towards a Comprehensive Survey of the Semantic Gap in Visual Image Retrieval". Proceedings of Int. Conference on Image and Video Retrieval, LNCS, Springer **27-28** (2003) 279–287
13. Serrano, N., Savakis, A., Luo, J.: "Improved Scene Classification using Efficient Low Level Features and Semantic Cues". Pattern Recognition, **37(9)** (2004) 1773–1784
14. Pal, N.R., Pal, S.K.: "A Review of Image Segmentation Techniques". Pattern Recognition **26,** (1993) 1277–1294
15. Picard, R., Minka, T.: "Vision Texture for Annotation". ACM Journal of Multimedia Systems **3(1)** (1995) 3–14
16. Vogel, J., Schiele, B.: "Semantic Modelling of Natural Scenes for Content-Based Image Retrieval". International Journal of Computer Vision, Springer, Netherlands **72(2)** (2006) 133–157
17. Wang, W., Song, Y., Zhang, A.: "Semantics-Based Image Retrieval by Region Saliency". Proceedings of Int. Conference on Image and Video Retrieval **LNCS-2383** (2002) 29–37
18. Papadopoulos, G.T., Mezaris, V., Dasiopoulou, S., Kompatsiaris, I.: "Semantic Image Analysis using a Learning Approach and Spatial Context". Y. Avrithis et al. (Eds.): First Int. Conference on Semantics and Digital Media Technology (SAMT 2006) **LNCS 4306** (2006) 199–211
19. Millet, C., Bloch, I., Hede, P., Moellic, P.A.: "Using Relative Spatial Relationships to Improve Individual Region Recognition". European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies, EWIMT'05 (2005) 119–126
20. Hollink, L., Nguyen, G., Schreiber, G., Wielemaker, J., Wielinga, B., Worring, M.: "Adding Spatial Semnatics to Image Annotatins". Proceedings of 4th Int. Workshop on Knowledge Markup and Semantic Annotation at ISWC'04 (2004)
21. Aghbari, Z., Makinouchi, A.: "Semantic Approach to Image Database Classification and Retrieval". NII J (Natl Inst Inform) **7** (2003) 1–8
22. Allen, J.F.: "Maintaining Knowledge About Temporal Intervals". Commun. ACM **26(11)** (1983) 832–843
23. Morchen, F., Ultsch, A.: "Mining Hierarchical Temporal Patterns in Multivariate Time Series". Proceedings of the 27th German Conference on Artificial Intelligence (KI), Germany, Springer **LNCS-3238** (2004) 127–140
24. Vogel, J., Schiele, B.: "A Semantic Typicality Measure for Natural Scene Categorization". In Rasmussen,C.E., H.H. Blthoff, B. Schlkopf and M.A. Giese Eds.): German Symposium on Pattern Recognition DAGM 2004, Tuebingen, Germany **LNCS 3175** (2004)
25. Aldenderfer, M.S., Blashfield, R.K.: "Cluster Analysis". Newbury Park, CA: Sage. (1984)
26. Kos, A.J., Psenicka, C.: Measuring cluster similarity across methods. Psychological Reports **86** (2000) 856–862
27. Clatworthy, J., Buick, D., Hankins, M., Weinman, J., Horne, R.: "The Use and Reporting of Cluster Analysis in Health Psychology: A Review". British Journal of Health Psychology **10** (2005) 329–358