COMPARATIVE EVALUATION OF SPATIAL CONTEXT TECHNIQUES FOR SEMANTIC IMAGE ANALYSIS

G. Th. Papadopoulos[1,2], C. Saathoff[3], M. Grzegorzek[3], V. Mezaris[2], I. Kompatsiaris[2], S. Staab[3] and M. G. Strintzis[1,2]

[1]Information Processing Lab. Electrical & Computer Eng. Dep. Aristotle University of Thessaloniki Thessaloniki, GR-54124, Greece

[2]Informatics & Telematics Institute Centre for Research & Technology Hellas 6th Klm. Charilaou - Thermi Road Thessaloniki, GR-57001, Greece

[3]Information Systems & Semantic Web Research Group Institute for Computer Science University of Koblenz - Landau Koblenz, Germany

ABSTRACT

In this paper, two approaches to utilizing contextual information in semantic image analysis are presented and comparatively evaluated. Both approaches make use of spatial context in the form of fuzzy directional relations. The first one is based on a Genetic Algorithm (GA), which is employed in order to decide upon the optimal semantic image interpretation by treating semantic image analysis as a global optimization problem. On the other hand, the second method follows a Binary Integer Programming (BIP) technique for estimating the optimal solution. Both spatial context techniques are evaluated with several different combinations of classifiers and low-level features, in order to demonstrate the improvements attained using spatial context in a number of different image analysis schemes.

1. INTRODUCTION

The continuously increasing amount of image content generated every day has led to the development of vast multimedia collections, which are stored in individuals' personal archives or are made available over the internet. Tasks like image search and retrieval in such collections have become common practice for the average user. As a consequence, new needs have emerged regarding the efficient manipulation of images. To this end, several approaches have been proposed in the literature regarding the skilful indexing, search and retrieval of images based on semantic criteria; hence, trying to bridge the so called *semantic gap* [1].

Among these approaches, semantic image analysis techniques, i.e. techniques that among others aim to detect and recognize the actual objects that are depicted in the image, have received particular attention. Region-level semantic image annotation can provide a good foundation for boosting global image classification, enabling the realization of complex queries or facilitating further inference [2]. However, the information ambiguity that is inherent in the visual medium reduces significantly the efficiency of the aforementioned techniques. In order to overcome this limitation, the use of contextual information has been proposed.

Among the available contextual information types, spatial context is of increased importance in image analysis. Boutell et al. [3] propose generative models for scene configurations, consisting of regions' identities and their spatial relations. In [4], Conditional Random Fields (CRFs) are used to maximize object label agreement according to both semantic and spatial relevance. Additional approaches to spatial context exploitation include the works of [5, 6, 7].

In this paper, two approaches making use of spatial knowledge for semantic image analysis are presented and comparatively evaluated. Initially, the examined image is segmented and two individual sets of low-level descriptors are extracted for every resulting segment. In parallel to this procedure, for every pair of image regions a corresponding set of fuzzy directional spatial relations are estimated. Then, the computed descriptors are provided as input to two different classifiers, namely a Support Vector Machine (SVM)- and a Maximum Likelihood (ML)-based one, in order to associate every region with a predefined high-level semantic concept based solely on visual information. The latter is used for denoting a real-world object that can be present in the examined image. Consequently, the two aforementioned spatial context techniques, which perform on top of the initial classification results and which make use of the available spatial knowledge, are applied in order to estimate an optimal region-concept assignment. Extensive experiments have been conducted for investigating the influence of the different combinations of the employed low-level features and classification algorithms on the performance of the presented spatial context exploitation techniques.

The paper is organized as follows: Section 2 discusses the low-level visual information processing and the extraction of the spatial relations. Section 3 outlines the employed classification algorithms. The developed spatial context exploitation techniques are described in Section 4. Experimental results are presented in Section 5 and conclusions are drawn in Section 6.

2. SEGMENTATION AND FEATURE EXTRACTION

In order to perform the initial region-concept association, the examined image has to be segmented into regions and suitable low-level descriptions have to be extracted for every resulting segment. In this work, a modified K-Means-with-connectivity-constraint pixel classification algorithm has been used for segmenting the image [8]. Output of this segmentation algorithm is a segmentation mask, where the created spatial regions $s_n,\ n = 1, ...N$, are likely to represent meaningful semantic objects.

For every generated image segment, two individual sets of low-level features are estimated. Regarding the first set, the following MPEG-7 descriptors are extracted and concatenated to form a region feature vector: Scalable Color, Homogeneous Texture, Region Shape and Edge Histogram. This results in a 433-dimensional low-level feature vector. The method followed for the extraction of

The work presented in this paper was supported by the European Commission under contract FP6-027026 K-Space.

the second set of features is based on the wavelet transform. For that purpose, a two-dimensional discrete signal decomposition technique is applied to local image neighborhoods, while the Johnston $8-TAB$ wavelet is used as the basis function [9]. Then, a four-dimensional feature vector is estimated for every RGB image region: $\mathbf{d}_n = (d_{n1}, d_{n2}, d_{n3}, d_{n4})^T$, where d_{n1} is calculated while taking into account all channels of the input image, while $d_{n2},\ d_{n3},\ d_{n4}$ result from the red, green, and blue channel processing, respectively. Detailed description of this feature extraction method can be found in [7]. The aforementioned descriptors are in turn utilized by the classification algorithms, i.e. they constitute a common data set, for performing the region-concept assignment.

In parallel to the descriptor extraction procedure, a set of fuzzy directional spatial relations are estimated for every pair of image regions $(s_n, s_m),\ n \neq m$. The set of directional relations utilized in this work, denoted by $R = \{r_k,\ k = 1, ...K\}$, comprises the following relations: Above (A), Right (R), Below (B), Left (L), Below-Right (BR), Below-Left (BL), Above-Right (AR) and Above-Left (AL). Relation r_k estimated for the region pair (s_n, s_m) is denoted by $r_k(s_n, s_m)$ and receives values in the interval $[0, 1]$. The aforementioned relations are utilized by the spatial context exploitation techniques, in order to refine the initial region-classification results. A detailed description of the fuzzy directional relations extraction procedure can be found in [5].

3. VISUAL CLASSIFICATION

3.1. Support Vector Machines

SVMs have been widely used in semantic image analysis tasks due to their reported generalization ability and their suitability for handling high-dimensional data [10]. Under the proposed approach, SVMs are employed for performing an initial association of the computed image regions to one of the defined high-level semantic concepts, denoted by $C = \{c_l,\ l = 1, ...L\}$, based on the estimated low-level features. An individual SVM is introduced for every defined concept c_l to detect the corresponding instances and is trained under the '*one-against-all'* approach. Each SVM, which receives as input either one of the two region feature vectors described in Section 2, returns at the evaluation stage for every segment a numerical value in the range $[0, 1]$ denoting the degree of confidence, h_{nl}, to which the corresponding region is assigned to the concept associated with the particular SVM. The degree of confidence is calculated as follows: $h_{nl} = \frac{1}{1+e^{-p \cdot z_{nl}}}$, where z_{nl} is the distance of the particular input feature vector from the corresponding SVM's separating hyperplane and *p* is a slope parameter set experimentally. For each region, $argmax_l(h_{nl})$ indicates its concept assignment. A detailed description of this procedure can be found in [11].

3.2. Maximum Likelihood Classifier

A classifier that is based on ML estimation [12] is also employed for performing an initial region-concept association, similarly to the SVM-based one. During the training procedure, the elements of the feature vector, either the MPEG-7- or the wavelet-based one (Section 2), are considered as random variables and modeled as normal-distribution probability density functions (pdfs), i.e. the mean value and standard deviation are computed for each element. At the evaluation stage, the degree of confidence h_{nl} is set equal to the probability value estimated using the aforementioned pdfs. $argmax_l(h_{nl})$ indicates again the concept that is eventually associated with region s_n. More details about the developed classifier can be found in [7].

4. SPATIAL CONTEXT

4.1. Genetic Algorithm

GAs have been extensively used in a wide variety of optimization problems [13], where they have shown to outperform other traditional methods. In the present analysis framework, a GA, which constitutes an extension of the approach presented in [11], is employed on top of the initial region-concept association results (Section 3) for deciding upon the optimal semantic image interpretation by treating semantic image analysis as a global optimization problem. More specifically, the GA receives as input the estimated degrees of confidence, h_{nl}, for all region-concept pairs (Section 3), the spatial relations among the image segments and spatial knowledge.

Spatial knowledge is obtained following a statistical learning approach. For that purpose, a set of annotated image content, denoted by Q_{tr} and for which the fuzzy directional relations described in Section 2 are computed, is assembled. Annotation is performed by applying the segmentation algorithm described in Section 2 to every image and consequently manually associating every resulting image region with a single concept. Then, for every concept pair $(c_l, c_{l'})$ the covariance matrix $\sum_{ll'}$ and the mean values $r^{ll'}_{k_{mean}}$ of the relations r_k are estimated for all region pairs $(s_n, s_m),\ n \neq m$, that are assigned to the aforementioned concept pair. The aforementioned statistical values serve as constraints denoting the concepts 'allowed' spatial topology.

Under the proposed approach, the GA employs an initial population of randomly generated chromosomes. Every chromosome V represents a possible solution, i.e. each gene assigns one of the defined concepts c_l to every image region s_n; this assignment is denoted by g_{nl}. After the population initialization, new generations are iteratively produced, where each generation results from the current one through the application of evolutionary operators like selection, crossover and mutation, until the optimal solution is reached. The GA is provided with an appropriate fitness function for denoting the plausibility of every possible image interpretation and has the form:

$$f(V) \ = \ \lambda \cdot FS_{norm} \ + \ (1 - \lambda) \cdot SC_{norm} \ , \tag{1}$$

where FS_{norm} refers to the degree of visual features similarity, SC_{norm} stands for the degree of spatial relations consistency, and variable $\lambda \epsilon [0, 1]$ is introduced to adjust the degree to which FS_{norm} and SC_{norm} should affect the final outcome; the value of the latter is estimated according to a separate optimization procedure [11]. FS_{norm} is defined as: $FS_{norm} = \frac{\sum_n h_{nl} - I_{min}}{I_{max} - I_{min}}$, where $I_{min} = \sum_n \min_l h_{nl}$ and $I_{max} = \sum_n max_l h_{nl}$. On the other hand SC_{norm} is defined as:

$$SC_{norm} \ = \ \frac{\sum_u w_u^{ll'} \cdot Y_u(g_{nl}, g_{ml'})}{\sum_u w_u^{ll'}} \ , \tag{2}$$

where u denotes a particular constraint, $Y_u(g_{nl}, g_{ml'})$ the spatial constraint verification factor and $w_u^{ll'}$ its weight. The spatial constraint verification factor is defined as follows:

$$Y_u(g_{nl}, g_{ml'}) = \frac{1}{1 + mahalanobis(g_{nl}, g_{ml'})} \ , \tag{3}$$

where for calculating the mahalanobis distance $\sum_{ll'}$, $r^{ll'}_{k_{mean}}$ and $r_k(s_n, s_m)$ are taken into account. Additionally, the weight of every constraint is set equal to $w_u^{ll'} \ = \frac{co(c_l, c_{l'})}{\sum_k \sigma^{ll'}_{r_k}}$, where $co(c_l, c_{l'})$ is the frequency of co-occurrence of concepts $c_l,\ c_{l'}$ in the images

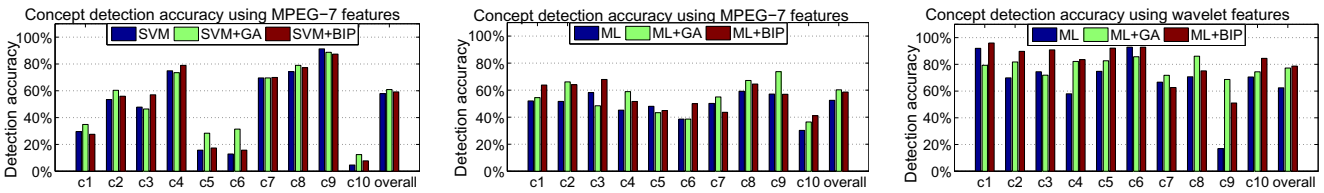

Fig. 1. Detailed concept association results (c1: *Building*, c2: *Foliage*, c3: *Mountain*, c4: *Person*, c5: *Road*, c6: *Sailing-boat*, c7: *Sand*, c8: *Sea*, c9: *Sky* and c10: *Snow*)

of Q_{tr}. According to the above definition, it can be seen that the lower the values of $\sigma_{r_k}^{ll'}$ are the higher the corresponding constraint weight gets. The motivation behind this is that pairs of concepts with clear spatial topology exhibit very similar spatial relations r_k; hence, they should have increased impact on the final outcome, compared to pairs of concepts with not so well-defined spatial arrangement. Moreover, constraints that correspond to concepts with high co-occurrence frequency are also favored.

Output of the procedure is a final region-concept association, which corresponds to the solution with the highest fitness value. A detailed description of the GA implementation can be found in [11].

4.2. Binary Integer Programming

The second approach for exploiting spatial context is based on binary integer programming. Similarly to the GA-based approach (Section 4.1), the set of annotated image content Q_{tr} is used again for extracting spatial-related constraints. These constraints are computed using support and confidence as selection criteria. Output of this procedure is an individual constraint for every pair of concepts c_l, $c_{l'}$ and every spatial relation r_k, denoted by $\mathcal{T}_k(c_l, c_{l'})$, which is defined as equal to 1 for an 'allowed' spatial arrangement and equal to 0 otherwise.

In order to represent the problem at hand as a binary integer program [14], a set of linear constraints for each spatial relation is created. In particular, let $O_n \subseteq R$ be the set of outgoing relations for region s_n, i.e. $O_n = \{r_k(s_n, s_m)\}$, and $E_n \subseteq R$ the set of incoming spatial relations, i.e. $E_n = \{r_k(s_m, s_n)\}$. Then, for every supported spatial relation r_k a respective binary integer variable $b_{nkm}^{ll'}$ is created, which represents the possible assignments g_{nl}, $g_{ml'}$ with respect to relation $r_k(s_n, s_m)$. If $b_{nkm}^{ll'} = 1$, the assignments g_{nl}, $g_{ml'}$ are made, while if $b_{nkm}^{ll'} = 0$ they are not. Since every such variable represents exactly one assignment of concepts to the involved regions, and only one concept might be assigned to a region in the final solution, this restriction has to be added as a set of linear constraints: $\forall r_k(s_n, s_m) : \sum_{c_l \in C} \sum_{c_{l'} \in C} b_{nkm}^{ll'} = 1$. These constraints assure that there is only one pair of concepts $(c_l, c_{l'})$ assigned to a pair of regions (s_n, s_m) per spatial relation r_k.

While the final solution requires that there is only one concept assigned to every region, appropriate constraints that "link" the variables accordingly have to be added, since the created variables from two relations involving the same region might possibly assign different concepts. This can be accomplished by linking pairs of relations. The basic mechanism for doing that for the case of outgoing relations is as follows: A base relation $r_k \in O_n$ is arbitrarily chosen and consequently constraints for all $r_{k'} \in O_n$, $k \neq k'$, are defined. Let $r_k(s_n, s_m)$ and $r_{k'}(s_n, s_{m'})$ be the two relations to be linked. Then, the constraints are defined as: $\forall c_l \in C$: $\sum_{c_{l'} \in C} b_{nkm}^{ll'} - \sum_{c_{l'} \in C} b_{nk'm'}^{ll'} = 0$. The first sum takes the value 1 if c_l is assigned to s_n by relation r_k. The second sum has to take the same value, since both are subtracted and the whole expression has to be equal to 0. Therefore, if one of the relations assigns c_l to s_n, the other has to do the same. Following the same approach, the incoming relations can be linked, as well as the incoming to the outgoing ones.

Eventually, the objective function is defined as:

$$\sum_{r_k(s_n,s_m)} \sum_{c_l} \sum_{c_{l'}} \min(h_{nl}, h_{ml'}) \cdot r_k(s_n, s_m) \cdot \mathcal{T}_k(c_l, c_{l'}) \cdot b_{nkm}^{ll'}. \quad (4)$$

This function rewards concept assignments that satisfy both the background knowledge and involve concepts with high degree of confidence. A detailed description of this approach can be found in [7].

5. EXPERIMENTAL RESULTS

In this section, experimental results regarding the evaluation process of the presented spatial context exploitation techniques are presented. For that purpose a set of 922 images[1], Q, belonging to the general category of outdoor images, was assembled. Then, the following set of 10 concepts, C, which represent meaningful real-world objects that can be present in images of the formed set, was defined: *Building*, *Foliage*, *Mountain*, *Person*, *Road*, *Sailing-boat*, *Sand*, *Sea*, *Sky* and *Snow*. Every image was manually annotated, i.e. after the segmentation algorithm described in Section 2 was applied, a single concept was associated with each resulting image region. The aforementioned image set Q was divided into two sub-sets, namely Q_{tr} (400 images) and Q_{te} (522 images). The first one, Q_{tr}, was used for training the classification algorithms and acquiring the appropriate spatial-related knowledge, while the second, Q_{te}, was used for the evaluation.

After every image belonging to the formulated set Q was segmented, two sets of low-level features, namely MPEG-7 descriptors and wavelet-based features, were extracted for every image region as described in Section 2. Additionally, for every possible pair of image regions a corresponding set of fuzzy directional spatial relations was also estimated. Subsequently, each of the classification algorithms (i.e. SVM- and ML-based one), which receives as input each of the aforementioned sets of low-level features separately, was applied to perform an initial association of every image region of the test set with one of the predefined semantic concepts (Section 3). Then, the spatial context exploitation techniques, i.e. GA and BIP, presented in Section 4 were used for deciding on the final region-concept assignment, while taking into account the computed spatial relations.

In Table 1, quantitative performance measures from the application of the spatial context exploitation techniques are presented in terms of concept detection accuracy for all possible combinations of

[1] http://mklab.iti.gr/project/scef

low-level features and classification algorithms. Accuracy is defined as the percentage of the image regions that are assigned to the correct semantic concept. It must be noted that classification results for the combination of the SVM-based classifier and the wavelet-based features are not given, since these features were experimentally found to be less suitable for SVM-based classification, possibly due to the very low dimensionality of the corresponding region feature vector $\mathbf{d}_n$. In Fig. 1, the respective detailed concept association results for all the aforementioned combinations are presented.

From the above results, it can be seen that the application of both GA and BIP techniques leads to a significant improvement of the overall concept detection accuracy for all combinations of features and classification algorithms, thus highlighting the effectiveness of spatial context exploitation in improving the region-concept association results that have been generated based solely on visual information. Additionally, no significant difference in the overall concept detection performance is observed between the two spatial context exploitation techniques for all possible features-classifier combinations, since the overall concept detection performance differences are lower than 1.80%. Moreover, it is shown that the combination of the wavelet-based features and the ML classifier leads to the highest overall performance improvement for both the GA and BIP techniques. This fact demonstrates that when the initial classification results are good for most supported concepts and not only for a relatively small subset of them, including concepts with not so well-defined spatial context like *Building*, *Road*, *Sailing-Boat* and *Snow*, then the detection performance improvement attained by the application of the spatial context exploitation techniques increases significantly for concepts with more well-defined spatial context, such as *Sea* and *Foliage*, as well as overall.

Examining the performance of the different approaches for each concept individually, it can be observed that the detection rate of certain concepts is generally boosted by the incorporation of spatial knowledge in the analysis process, since they present significant detection performance improvement after the application of both GA and BIP for any combination of low-level features and classifiers. These concepts are: *Person*, *Foliage*, *Sea* and *Snow*. Additionally, it is experimentally shown that the detection of some concepts is particularly favored by the GA or the BIP algorithm, regardless of the employed features-classifier combination. Specifically, concepts with more well-defined spatial context, like *Sea*, *Sand* and *Sky*, exhibit increased recognition rates when the GA approach is applied, instead of the BIP. This suggests that the statistical learning approach followed for obtaining the GA's fuzzy spatial constraints, i.e. calculation of $\sum_{ll'}$ and $r^{ll'}_{k_{mean}}$, is more suitable for modeling the spatial relations of these concepts. On the other hand, the BIP technique is shown to be more appropriate than the GA for localizing concepts with not so well-defined spatial context, like *Building*, *Mountain* and *Snow*. This indicates that the set of binary spatial constraints $\mathcal{T}_k(c_l, c_{l'})$, which are computed using support and confidence as selection criteria and are utilized by the BIP technique, is advantageous for detecting the aforementioned concepts.

Table 1. Concept detection accuracy

Features	Classifier	Spatial context technique
MPEG-7	SVM: 57.88%	GA: 60.94%
		BIP: 59.15%
	ML: 52.45%	GA: 60.24%
		BIP: 58.58%
Wavelet	ML: 62.45%	GA: 77.24%
		BIP: 78.73%

6. CONCLUSIONS

In this paper, two approaches to spatial context exploitation, which make use of fuzzy directional relations, were presented and comparatively evaluated. Extensive experiments demonstrated the influence of different features and classification algorithms on their region-concept association performance. Future work includes further investigation of the impact of fuzzy and binary spatial constraints and the possible integration of them into a single model.

7. REFERENCES

[1] A.W.M. Smeulders et al., "Content-Based Image Retrieval at the End of the Early Years," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, pp. 1349–1380, 2000.

[2] S. Dasiopoulou, J. Heinecke, C. Saathoff, and M.G. Strintzis, "Multimedia reasoning with natural language support," *Int. Conf. on Semantic Computing*, pp. 413–420, 2007.

[3] M.R. Boutell, J. Luo, and C.M. Brown, "Factor Graphs for Region-based Whole-scene Classification," in *Proc. of Conf. on Computer Vision and Pattern Recognition Workshop*, 2006.

[4] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," *Comp. Vision and Pat. Recog., IEEE Conf. on*, pp. 1–8, 2008.

[5] G.T. Papadopoulos, V. Mezaris, S. Dasiopoulou, and I. Kompatsiaris, "Semantic Image Analysis Using a Learning Approach and Spatial Context," *Int. Conf. on Semantics and Digital Media Technologies*, vol. 4306, pp. 199, 2006.

[6] J. Yuan, J. Li, and B. Zhang, "Exploiting spatial context constraints for automatic image region annotation," in *Proc. of ACM Multimedia*, 2007, pp. 595–604.

[7] C. Saathoff, M. Grzegorzek, and S. Staab, "Labelling image regions using wavelet features and spatial prototypes," in *Int. Conf. on Semantics and Digital Media Technologies*, 2008.

[8] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis, "Still Image Segmentation Tools for Object-Based Multimedia Applications," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, pp. 701–726, 2004.

[9] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[10] K. Kim, K. Jung, S. Park, and H. Kim, "Support vector machines for texture classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1542–1550, 2002.

[11] G. Th. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Combining Global and Local Information for Knowledge-Assisted Image Analysis and Classification," *EURASIP Journal on Adv. in Signal Proc.*, vol. 2007, pp. 1–15, 2007.

[12] A. R. Webb, *Statistical Pattern Recognition*, John Wiley & Sons Ltd, Chichester, UK, 2002.

[13] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, 1996.

[14] D. Luenberger, *Linear and Non-Linear Programming*, 1989.