Proc. 20th International Conference on Pattern Recognition (ICPR 2010), Istanbul, Turkey, August 2010.

A Statistical Learning Approach to Spatial Context Exploitation for Semantic Image Analysis*

G. Th. Papadopoulos[1,2], V. Mezaris[2], I. Kompatsiaris[2] and M. G. Strintzis[1,2]
[1]*Electrical & Computer Eng. Dep., Aristotle Univ. of Thessaloniki, Greece*
[2]*Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece*
mail: {papad,bmezaris,ikom}@iti.gr, strintzi@eng.auth.gr

Abstract

In this paper, a statistical learning approach to spatial context exploitation for semantic image analysis is presented. The proposed method constitutes an extension of the key parts of the authors' previous work on spatial context utilization, where a Genetic Algorithm (GA) was introduced for exploiting fuzzy directional relations after performing an initial classification of image regions to semantic concepts using solely visual information. In the extensions reported in this work, a more elaborate approach is followed during the spatial knowledge acquisition and modeling process. Additionally, the impact of every resulting spatial constraint on the final outcome is adaptively adjusted. Experimental results as well as comparative evaluation on three datasets of varying complexity in terms of the total number of supported semantic concepts demonstrate the efficiency of the proposed method.

1. Introduction

Semantic image analysis techniques aim among others to localize and recognize the actual objects that are depicted in the image. They have been utilized during the development of sophisticated systems for the efficient manipulation of the image content and have so far exhibited promising results [8]. However, their efficiency is significantly hindered by the inherent visual information ambiguity. For overcoming this limitation, the use of contextual information has been proposed.

Among the available contextual information types, spatial context is of increased importance in semantic image analysis. The latter represents and models the spatial attributes of the real-world objects and consequently is used for discriminating between objects that exhibit similar visual characteristics. In [2], a Conditional Random Field (CRF)-based approach is presented that incorporates co-occurrence and spatial contextual information. Additionally, spatial constraints specific to individual scene types are modeled in the form of factor graphs in [1]. Singhal et al. [7] propose a Bayesian Network (BN)-based framework for learning and consequently utilizing probabilistic spatial context models. Moreover, individual spatial context techniques are comparatively evaluated with different combinations of classifiers and low-level features in [6].

In this paper, a statistical learning approach to spatial context exploitation for semantic image analysis is presented. The proposed method constitutes an extension of the authors' previous work on spatial context [5] in the following crucial parts: a) the spatial knowledge acquisition process and subsequent enforcement of the resulting spatial constraints, and b) the determination of the impact of every estimated constraint on the final outcome.

The paper is organized as follows: Section 2 discusses the low-level visual information processing and the extraction of the spatial relations. Section 3 briefly describes the visual classification algorithm. The authors' previous work on spatial context exploitation is outlined in Section 4. The proposed spatial constraint acquisition and application approach is detailed in Section 5. Experimental results are presented in Section 6 and conclusions are drawn in Section 7.

2. Segmentation and Feature Extraction

In order to perform an initial region-concept association based solely on visual information, the examined image is segmented using the algorithm of [3] and the created spatial regions are denoted by s_n, $n =$

*The work presented in this paper was supported by the European Commission under contracts FP6-045547 VIDI-Video, FP7-214306 JUMAS and FP7-248984 GLOCAL.

$1, ...N$. For every generated image segment, the following MPEG-7 descriptors are extracted and concatenated to form a region feature vector: Scalable Color, Homogeneous Texture, Region Shape and Edge Histogram.

In parallel to visual descriptor extraction, a set of fuzzy directional spatial relations are estimated for every ordered pair of image regions $(s_n, s_m),\ n \neq m$. The set of directional relations utilized in this work, denoted by $R = \{r_\gamma\ :\ \gamma \in [1, \Gamma]\}$, comprises the following relations: Above, Right, Below, Left, Below-Right, Below-Left, Above-Right and Above-Left. Relation r_γ estimated for the region pair (s_n, s_m) is denoted by $r_\gamma(s_n, s_m) \in [0, 1]$. A detailed description of the extraction procedure can be found in [5].

3. Visual Classification

In this work, Support Vector Machines (SVMs) are employed for performing an initial association of the computed image regions with one of a set of pre-defined semantic concepts based on the estimated low-level features. An individual SVM is introduced for every concept c_k, $k = 1, ...K$, to detect the corresponding instances, and is trained under the *'one-against-all'* approach. Each SVM, which receives as input the region feature vector described in Section 2, returns at the evaluation stage a posterior probability h_{nk}, which denotes the degree to which concept c_k is assigned to region s_n. For each region, $\arg\max_k(h_{nk})$ indicates its concept assignment based solely on visual information. A detailed description of this procedure can be found in [5].

4. Previous Work on GA

Spatial contextual information is used in this work for improving the initial analysis results that have been computed based solely on visual cues. Two important aspects of this problem are: a) for an image with N regions and for an application case where the detection of K concepts is required, the image solution space (i.e. the number of possible image interpretations that need to be examined) comprises K^N instances. The latter number is huge for most typical applications. b) for each candidate image solution, all possible pairs of region to concept mappings need to be examined and evaluated with respect to the spatial arrangement of the corresponding semantic concepts.

In order to incorporate spatial contextual information in the semantic image analysis process, while addressing the aforementioned challenges, a GA is employed in [5] on top of the initial region-concept association results for deciding upon the optimal semantic image interpretation by treating image analysis as a global optimization problem. The latter choice is justified by the fact that GAs have been extensively used in a wide variety of optimization problems [4], where they were shown to outperform other traditional methods.

Spatial context is obtained following a simple learning process. For that purpose, a set of annotated image content, denoted by D_{tr} and for which the fuzzy directional relations are computed as described in Section 2, is assembled. Then, for every ordered concept pair (c_k, c_l) the mean values $\bar{r}_\gamma^{kl}$ of the relations r_γ, which have been computed for all region pairs $(s_n, s_m),\ n \neq m$, assigned to the concepts (c_k, c_l), respectively, are estimated. The set of values $\bar{r}_\gamma^{kl}$ obtained for concept pair (c_k, c_l) define a spatial constraint, denoted by d^{kl}, which represents the 'allowed' spatial topology of concepts c_k and c_l.

At the evaluation stage, 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_k to an image region s_n; this assignment is denoted g_{nk} and therefore $V = \{g_{nk}\ :\ n \in [1, N]\}$. After the population initialization, new generations are iteratively produced, where each new generation comes from the current one after the application of evolutionary operators like selection, crossover and mutation, until the optimal solution is reached. The GA is provided with a fitness function for evaluating the plausibility of every possible image interpretation, which has the form:

$$f(V) \ = \ \lambda \cdot FS \ + \ (1 - \lambda) \cdot SC \ , \qquad (1)$$

where FS refers to the degree of visual features similarity, SC stands for the degree of spatial relations consistency, and variable $\lambda \in [0, 1]$ is introduced to adjust the degree to which FS and SC should affect the final outcome; the value of the latter is estimated according to a separate optimization procedure [5], where a sub-set of D_{tr} serves as a validation set. FS is defined as:

$$FS = \frac{\sum_n h_{nk} - I_{min}}{I_{max} - I_{min}}$$
$$I_{min} = \sum_n \min_k(h_{nk}),\ I_{max} = \sum_n \max_k(h_{nk}) \qquad (2)$$

On the other hand SC is calculated as follows:

$$SC = \frac{\sum_{n,m} B_{d^{kl}}(g_{nk}, g_{ml})}{N(N-1)} \ , \qquad (3)$$

where $B_{d^{kl}}(g_{nk}, g_{ml})$ denotes the corresponding spatial constraint's verification factor for the region pair (s_n, s_m) and $N(N-1)$ denotes the number of permutations of the N image regions taken 2 at a time (i.e. the number of ordered region pairs that are present in the

examined image and which contribute to the summation in the numerator). $B_{d^{kl}}(g_{nk}, g_{ml}) \in [0,1]$ is estimated based on a normalized Euclidian distance calculation:

$$B_{d^{kl}}(g_{nk}, g_{ml}) = 1 - \frac{\sqrt{\sum_{\gamma}(\bar{r}_{\gamma}^{kl} - r_{\gamma}(s_n, s_m))^2}}{\sqrt{\Gamma}} \quad (4)$$

Output of this procedure is a final region-concept association, which corresponds to the solution with the highest fitness value.

5. Proposed Approach

The proposed approach constitutes an extension of the GA-based method outlined in Section 4 with respect to the spatial knowledge acquisition process and the determination of the impact of every estimated constraint on the final outcome, i.e. in the way that the term SC is calculated in Eq. (1).

Regarding the spatial knowledge acquisition process, a statistical learning approach is followed for attaining and efficiently modeling the complex spatial relationships between the supported concepts. For that purpose, the set of annotated image content D_{tr} (Section 4), for which the fuzzy directional relations are computed, is utilized. Then, for every ordered concept pair (c_k, c_l) the mean vector $\bar{\mathbf{r}}^{kl}$ and the corresponding covariance matrix $cov(\mathbf{r}^{kl})$, with respect to the relations r_{γ}, are calculated as follows:

$$\begin{aligned} \mathbf{r}_{n,m} &= [r_1(s_n, s_m),\ r_2(s_n, s_m) \ldots r_{\Gamma}(s_n, s_m)]^T \\ \bar{\mathbf{r}}^{kl} &= [\bar{r}_1^{kl},\ \bar{r}_2^{kl} \ldots \bar{r}_{\Gamma}^{kl}]^T = E[\mathbf{r}_{n,m}] \\ cov(\mathbf{r}^{kl}) &= E[(\mathbf{r}_{n,m} - \bar{\mathbf{r}}^{kl})(\mathbf{r}_{n,m} - \bar{\mathbf{r}}^{kl})^T], \end{aligned} \quad (5)$$

where for the calculations the spatial relations $r_{\gamma}(s_n, s_m)$ which have been computed for all region pairs (s_n, s_m), $n \neq m$, that are assigned to the concepts (c_k, c_l), respectively, are taken into account. Similarly to the case of d^{kl} in Section 4, the set of values $\bar{\mathbf{r}}^{kl}$ and $cov(\mathbf{r}^{kl})$ are used to define a spatial constraint, denoted by u^{kl}, with respect to the concept pair (c_k, c_l). The aforementioned statistical values do not encompass only the expected values $\bar{\mathbf{r}}^{kl}$ of the spatial relations for every possible concept pair (c_k, c_l), but also encode the correlations between their respective values $cov(\mathbf{r}^{kl})$, hence resulting in a more complete representation of the supported concepts' spatial configuration.

At the evaluation stage, the GA follows the same evolutionary procedure described in Section 4 and the respective fitness function is again given by Eq. (1). However, the value of SC, which indicates the consistency of each solution with respect to the acquired spatial constraints and is evaluated for all ordered region pairs that are present in the image, is now calculated as:

$$SC = \frac{\sum_{n,m} w_u^{kl} \cdot Y_{u^{kl}}(g_{nk}, g_{ml})}{\sum_{n,m} w_u^{kl}}, \quad (6)$$

where $Y_{u^{kl}}(g_{nk}, g_{ml})$ denotes the verification factor of u^{kl} for the region pair (s_n, s_m) and w_u^{kl} its weight. $Y_{u^{kl}}(g_{nk}, g_{ml}) \in [0,1]$ is defined according to the following mahalanobis distance-based expression:

$$Y_{u^{kl}}(g_{nk}, g_{ml}) = \frac{1}{1 + \sqrt{\mathbf{q}_{n,m}^T cov^{-1}(\mathbf{r}^{kl}) \mathbf{q}_{n,m}}}, \quad (7)$$

where $\mathbf{q}_{n,m} = \mathbf{r}_{n,m} - \bar{\mathbf{r}}^{kl}$. Greater values of $Y_{u^{kl}}(g_{nk}, g_{ml})$ indicate more plausible spatial arrangements. The weight of every constraint is set equal to:

$$w_u^{kl} = \frac{co(c_k, c_l)}{\sum_{\gamma} \sigma_{r_{\gamma}}^{kl}}, \quad (8)$$

where the standard deviations $\sigma_{r_{\gamma}}^{kl}$ are calculated from the corresponding $cov(\mathbf{r}^{kl})$ and $co(c_k, c_l)$ is the frequency of co-occurrence of concepts c_k, c_l in the images of D_{tr}. According to the above definition, it can be seen that the lower the values of $\sigma_{r_{\gamma}}^{kl}$ are, the higher the corresponding constraint weight gets. The motivation for this is that pairs of concepts with clear spatial topology should have increased impact on the final outcome compared to pairs of concepts with not so well-defined spatial arrangement. Additionally, from the definition of w_u^{kl} it can be seen that constraints that correspond to concepts with high co-occurrence frequency are also favored. This mainly accounts for application cases where a large number of concepts is defined and every individual concept tends to co-exist with only a subset of them. Output of the procedure is a final region-concept association, which again corresponds to the solution with the highest fitness value.

6. Experimental Results

In this section, experimental results as well as comparative evaluation from the application of the proposed approach to three datasets, denoted by D_1, D_2, and D_3, are presented. Regarding the creation of D_1, a set of 535 images depicting only coastal scenes was assembled and the following set of concepts c_k, $k = 1, \ldots 7$, which represent meaningful real-world objects that can be present in images of the formed set, was defined: *Sand*, *Sea*, *Boat*, *Vegetation*, *Rock*, *Person* and *Sky*. Then, every image was manually annotated. The aforementioned image set was divided into two sub-sets, namely D_{tr} (263 images) and D_{te} (272 images). The

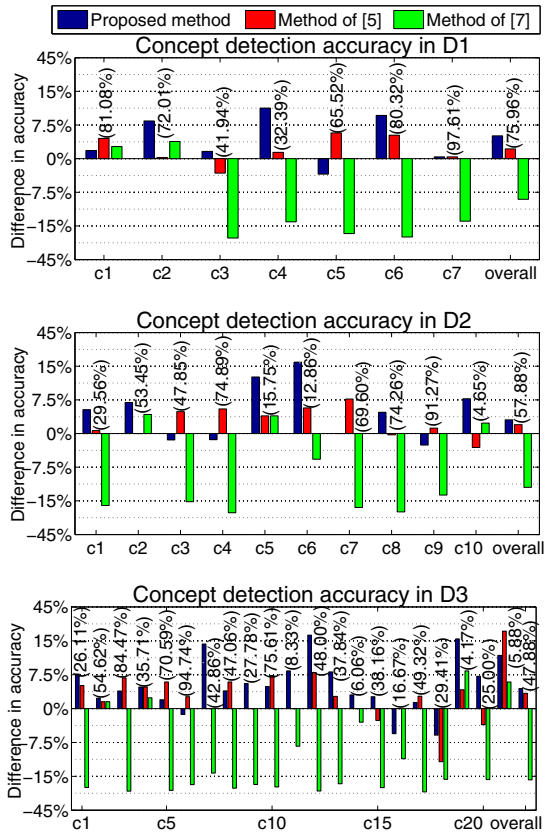

Figure 1. Concept detection results.

first one, D_{tr}, was used for training the SVM-based classification algorithm and acquiring the appropriate spatial-related knowledge, while the second, D_{te}, was used for evaluation. The SCEF[1] dataset, which is denoted by D_2 (923 images) and was introduced in [6], is also used for experimentation. For this, the following 10 concepts are defined: *Building*, *Foliage*, *Mountain*, *Person*, *Road*, *Sailing-boat*, *Sand*, *Sea*, *Sky* and *Snow*. The corresponding sets D_{tr} and D_{te} comprise 400 and 523 images, respectively. Regarding the dataset D_3, the MSRC[2] v2 dataset was utilized. For the latter dataset, the following 21 semantic concepts are supported: *Building*, *Grass*, *Tree*, *Cow*, *Sheep*, *Sky*, *Aeroplane*, *Water*, *Face*, *Car*, *Bicycle*, *Flower*, *Sign*, *Bird*, *Book*, *Chair*, *Road*, *Cat*, *Dog*, *Body* and *Boat*. The corresponding sets D_{tr} and D_{te}, comprise 295 and 296 images, respectively.

In Fig. 1, quantitative performance measures from the application of the proposed spatial context exploitation technique to the utilized datasets are presented in terms of the difference in concept detection accuracy. The latter is calculated by subtracting the detection accuracy accomplished based solely on visual features from the corresponding one obtained after the application of the proposed spatial context exploitation approach. The initial classification results computed based on visual information are depicted in parentheses. Accuracy is defined as the percentage of the image regions that are assigned to the correct concept.

From the presented results, it can be seen that the proposed approach achieves an overall performance improvement of 5.09%, 3.06% and 4.44% in the D_1, D_2 and D_3 datasets, respectively, compared to the initial classification results. Additionally, the detection rates for most of the supported concepts are significantly increased in all datasets. In particular, it is shown that concepts with more well-defined spatial context present the highest in percentage improvement (D_1: Sea, Vegetation, Person; D_2: Road, Sailing-boat, Snow, D_3: Aeroplane, Flower, Dog). Concept c_k is considered to have well-defined spatial context if the sum $\sum_l tr(cov(\mathbf{r}^{kl}))$ receives relatively low values (where $tr(.)$ denotes the trace of a matrix), i.e. the spatial relations of concept c_k with all other concepts c_l of the respective dataset do not present significant variations in their values. These results demonstrate the efficiency of spatial context exploitation in improving the region classification results that have been computed based solely on visual information.

Comparing the performance of the proposed approach for the selected datasets, it can be observed that the overall concept detection improvement tends to decrease when the corresponding number of supported concepts increases. The reason for this is twofold: a) For dataset D_1, a smaller number of concepts has been defined. This results in reduced problem complexity, which in turn accommodates the proposed approach in efficiently discriminating between the supported concepts. b) D_2 and D_3 constitute significantly broader datasets, including concepts from multiple and different semantic categories. As a consequence, the concepts' spatial configuration becomes less well-defined, while it is more likely for many individual concept pairs to share very similar spatial characteristics. The latter ambiguity is partially overcome by the incorporation of the concepts' co-occurrence frequency ($co(c_k, c_l)$) in the optimization procedure. It must be noted that a higher overall performance improvement is observed in D_3 than D_2, contrary to the fact that a greater number of concepts is defined in D_3. This is mainly due to the images of D_3 including very few different kinds of objects (usually no more than two or three), and only particular concept pairs tend to co-exist. As a result, many co-occurrence frequencies $co(c_k, c_l)$ are equal to zero, which in turn facilitates the discrimination between the concepts.

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

[2] http://research.microsoft.com/en-us/projects/objectclassrecognition/

In Fig. 1, the performance of the proposed method is also compared with the spatial context exploitation approach presented in [5]. From the presented results, it can be seen that the proposed method outperforms the method of [5] for most of the supported concepts, as well as in overall detection accuracy, for all datasets. Specifically, the proposed method exhibits 2.92%, 1.09% and 1.07% higher overall accuracy in the D_1, D_2 and D_3 datasets, respectively. These observations indicate that the more sophisticated statistical learning-based approach followed for the spatial constraints acquisition led to increased detection performance, compared to the simpler learning process followed in [5]. Additionally, the proposed method is shown to be advantageous for concepts with more well-defined spatial context, as described earlier in this section. Moreover, it can be seen that when the initial classification rate is significantly high for a particular concept (e.g. concept Sky in all datasets), both methods present marginal changes in its detection rate. On the other hand, significant performance improvements can be obtained by the application of the proposed approach for concepts with low recognition rates (e.g. concepts Sailing-boat and Snow in D_2).

The performance of the proposed method is also compared with the spatial context-aware concept detection system presented in [7]. Specifically, Singhal et al. utilize a series of N BNs (N being the number of regions in the examined image), which are gradually constructed and solved in an iterative manner, for learning and consequently utilizing probabilistic spatial context models. It must be noted that despite the good recognition results reported in [7], their framework was evaluated with a small number of supported concepts, namely the concepts Sky, Grass, Foliage, Water and Snow. According to the results presented in Fig. 1, the proposed method outperforms the method of [7] for most of the supported concepts, as well as in overall classification accuracy, for all datasets. This is mainly due to the following limitations of the method of [7]: a) every pair of image regions can only be connected with a single binary spatial relation, and b) the image region s_n, which is imported to the BN structure at each of the N steps, is selected according to the descending sorted set of the values $\arg\max_k(h_{nk})$. Although this selection is intuitive, it is in principle heuristic and as a consequence the BN inference can be significantly misguided if image regions are associated with an incorrect concept with a high degree of confidence h_{nk}. The obtained results are also consistent with the ones reported in the more recent work of the same authors [1], where their improved algorithm was shown to lead to a decrease in the overall classification performance of approximately 8.5% in a dataset including ten concepts, when only spatial-related contextual information was used, compared to the initial classification results based solely on visual information. On the contrary, the proposed approach performs significantly better than the method of [7], since it handles the aforementioned limitations in the following ways: a) every pair of image regions is connected through a set of fuzzy directional spatial relations. As a consequence, a more complete representation of the spatial properties of the depicted objects is achieved, while a significantly more elaborate statistical learning approach is followed during the spatial constraints acquisition process, as opposed to the simple frequency counting approach adopted in the work of [7]. b) the proposed formulation of semantic image analysis as a global optimization problem and the subsequent use of a GA for solving it, render the proposed method less likely to converge to local maxima in the solution space, compared to the method of [7] that follows a heuristic approach for estimating the optimal solution that can be easily misguided, as discussed above.

7. Conclusions

In this paper, a statistical learning approach to spatial context exploitation for semantic image analysis, which makes use of fuzzy directional relations, was presented and comparatively evaluated on three datasets. Future work includes the investigation of additional contextual information sources and their integration in the developed framework.

References

[1] M. Boutell et al. Improved Semantic Region Labeling Based on Scene Context. In *Proc. IEEE ICME*, 2005.

[2] C. Galleguillos et al. Object categorization using co-occurrence, location and appearance. In *Proc. IEEE CVPR*, 2008.

[3] V. Mezaris et al. Still Image Segmentation Tools for Object-Based Multimedia Applications. *Int. J. of Patt. Recog. and Artif. Intell.*, 18, 2004.

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

[5] G. T. Papadopoulos et al. Combining Global and Local Information for Knowledge-Assisted Image Analysis and Classification. *EURASIP J. on Adv. in Signal Proc.*, 2007.

[6] G. T. Papadopoulos et al. Comparative evaluation of spatial context techniques for semantic image analysis. *In Proc. WIAMIS '09*, pages 161–164, 2009.

[7] A. Singhal et al. Probabilistic spatial context models for scene content understanding. In *Proc. IEEE CVPR*, 2003.

[8] A. Smeulders et al. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. on PAMI*, pages 1349–1380, 2000.