Kernel Fusion for Image Classification Using Fuzzy Structural Information

Emanuel Aldea[1], Geoffroy Fouquier[1], Jamal Atif[2], and Isabelle Bloch[1]

[1] GET - Télécom Paris (ENST), Dept. TSI, CNRS UMR 5141 LTCI
46 rue Barrault, 75634 Paris Cedex 13, France Geoffroy.Fouquier@enst.fr,
[2] Unité ESPACE S140, IRD-Cayenne/UAG, Guyanne Française

Abstract. Various kernel functions on graphs have been defined recently. In this article, our purpose is to assess the efficiency of a marginalized kernel for image classification using structural information. Graphs are built from image segmentations, and various types of information concerning the underlying image regions as well as the spatial relationships between them are incorporated as attributes in the graph labeling. The main contribution of this paper consists in studying the impact of fusioning kernels for different attributes on the classification decision, while proposing the use of fuzzy attributes for estimating spatial relationships.

1 Introduction

Most of traditional machine learning techniques are not designed to cope with structured data. Instead of changing these algorithms, an alternative approach is to go in the opposite direction and to adapt the input for classification purposes so as to decrement the structural complexity and at the same time to preserve the attributes that allow assigning data to distinct classes.

In the particular case of images, fundamentally different strategies have been outlined in recent years. One of them copes with images as single indivisible objects [1] and tends to use global image features, like the color histogram. Other strategies treat them as bags [2] of objects, thus taking into account primarily the vectorization of the image content. Finally, a third strategy considers images as organized sets of objects [3, 4], making use of components and also of the relationships among them; our approach falls into this category. The interest of this latter model in retrieving complex structures from images is that it handles view variations and complex inference of non-rigid objects, taking into account their intrinsic variability in a spatial context.

In [5], an image classification method using marginalized kernels for graphs was presented. In a preprocessing step, images are automatically segmented and an adjacency graph is built upon the resulting neighboring regions. Intrinsic region attributes are computed. The only structural information retrieved from the image is the neighborhood relationship between regions that is implicitly stored in the graph structure by the presence of an edge between two vertices. Once the graph is built, a marginalized kernel extension relying on the attributes mentioned above is used to assess the similarity between two graphs and to build a classifier.

In this paper, we extend this image classification method. We propose to automatically create a kernel based on more than one attribute. The presence of multiple attributes emphasizes the importance of a generic, reliable method that combines data sources in building the discriminant function of the classifier [6]. We also enrich the graph by adding more edges and more complex structural information retrieved from the image, such as topological relations or metric spatial relations [7] (distance, relative orientation). This raises specific methodological problems, that are addressed in this paper, in particular by using different kernels for each type of relation and combining them under a global optimization constraint. The framework is open to the introduction of any other features that describe image regions or relationships between them. However, we stress the importance of selecting relevant features and of finding positive definite kernels that give an intuitive similarity measure between them. The general scheme of the proposed method is illustrated in Figure 1.

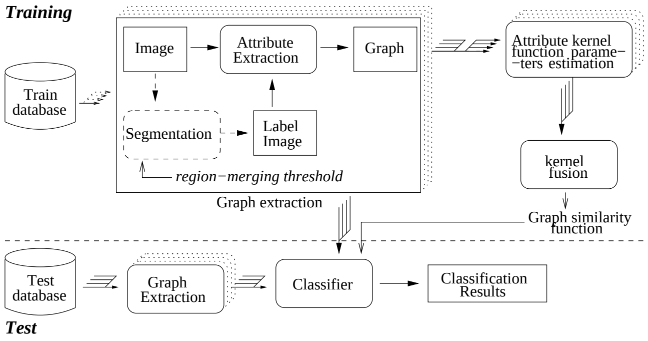

Fig. 1. Block diagram. Training step: If needed, images are segmented. A graph is extracted from each image of the training database, using the corresponding label image. Then, for each graph attribute, the corresponding kernel function parameters are estimated. Finally, the kernel functions are merged. Test step: A graph is extracted from each image of the test database. The resulting graphs are compared with the graphs of the training database and classified using the learned similarity function

The structure of this paper is as follows. First the original method is summarized in Section 2. Section 3 presents the graph structure and edge attributes. Section 4 presents how kernel fusion is used to merge different attribute kernels. Experimental results are outlined in Section 5.

2 Classification based on kernels for graphs

This section briefly presents the general principle of our classification technique based on random walk kernels for graphs [5].

The image is first over-segmented using an unsupervised hierarchical process [8, 9]. Then neighboring regions with close average gray levels are merged. The stopping criterion is a function of a dynamic threshold based on the differences between neighboring regions, updated at each step of the process[3]. An adjacency graph is constructed with all regions as vertices. In [5], only the adjacency between regions is considered as an implicit edge attribute. The following real value attributes are then computed for each region: the surface in pixels, the ratio between the surface of the region and the surface of the image (relative surface), average gray level, relative (to the dynamic range of the image) gray level, perimeter, compacity and neighboring degree.

The kernel between two graphs G and G' measures the similarity according to an attribute a of all the possible random walk labels [10, 11], weighted by their probabilities of apparition. As compared to previous frameworks that use this type of method [12], the region neighborhood has a lower importance in image than it has in a chemical structure between its constituents, for example. Variable space used in labeling becomes continuous and multi-dimensional, and a significant part of the information migrates from the graph structure to the labeling of its constituent parts. Therefore, the similarity function for a continuous-valued attribute such as the gray level must be less discriminative than a Dirac function. For this purpose, a Gaussian kernel $K_a^{RBF}(a_1, a_2) = \exp[-\|a_1 - a_2\|^2/(2\sigma^2)]$ or a triangular kernel $K_a^{\varDelta}(a_1, a_2) = \max(1 - \|a_1 - a_2\|/\varGamma, 0)$ is used for assessing the similarity between two numeric values a_1 and a_2 of an attribute a.

For two graphs G and G' to compare, these basic kernels allow us to evaluate the similarity $k_a(h, h')$ between two random walks $h \in G$ and $h' \in G'$, by aggregating the similarity of attribute a of all vertices (resp. edges) along h and h'. In [5], an extension of the base kernel $k_a(h, h')$ is proposed to better cope with specific image attributes. Under this framework, continuous similarity values between graph constituents (vertices, edges) are interpreted as transition probability penalties that will influence the random walks, without terminating them prematurely. Finally, the kernel between G and G' sums the similarity of all the possible random walks, weighted by their probabilities of apparition: $K_a(G, G') = \sum_h \sum_{h'} k(h, h')p(h|G)p(h'|G')$. This function is subsequently used in a 1-norm soft margin SVM [6] for creating the image classifier.

3 Graph representation of images including spatial relations

In addition to the region-based attributes from the original method, we propose to improve the structure of the graph (by adding some edges) and to add structural information on these edges.

The original method [5] uses an adjacency graph. One way to enrich the graph is by adding structural information on the adjacency graph, i.e. no edges

[3] Any other segmentation method achieving the same goal could be used as well (e.g. Markov Random Fields)

are added or removed. On the other hand, the adjacency graph from the original method is too restrictive since adjacency is a relation that is highly sensitive to the segmentation of the objects and whether it is satisfied or not may depend on one point only.

Therefore, using edges carrying more than adjacency and corresponding attributes better reflects the structural information and improves the robustness of the representation. Thus, the resulting graph is not an adjacency graph anymore, it may even become complete if this is not a performance drawback.

In [5], only region-based features are computed. We propose some new features based on structural information, more precisely spatial relations. They are traditionally divided into topological relations and metric relations [13]. Among all the spatial relation, we choose here the most usual examples of the latter: distance and directional relative position (but the method applies to any other relation). As a topological relation, instead of the adjacency, we compute an estimation of the adjacency length between two regions. We now present each of these features.

Distance between regions. The distance between two regions R_1 and R_2 is computed as the minimal Euclidean distance between two points $p_i \in R_1$ and $q_j \in R_2$: $\min_{p_i \in R_1, q_j \in R_2}(d_{euclidian}(p_i, q_j))$.

Directional relative position. Several methods have been proposed to define the directional relative position between two objects, which is an intrinsically vague notion. In particularly, fuzzy methods are appropriate [14], and we choose here to represent this information using histograms of angles [15].
This allows representing all possible directional relations between two regions. If R_1 and R_2 are two sets of points $R_1 = p_1, ..., p_n$ and $R_2 = q_1, ..., q_n$, the relative position between regions R_1 and R_2 is estimated from the relative position of each point q_j of R_2 with respect to each point p_i of R_1. The histogram of angles $H_{R_1R_2}$ is defined as a function of the angle θ and $H_{R_1R_2}(\theta)$ is the frequency of the angle θ:

$$H_{R_1R_2}(\theta) = \left| \{(p_i, q_j) \in R_1 \times R_2 / \angle\ (\overrightarrow{i}, \overrightarrow{p_i q_j}) = \theta\} \right|$$

where $\angle\ (\overrightarrow{i}, \overrightarrow{p_i q_j})$ denote the angle between a reference vector $\overrightarrow{i}$ and $\overrightarrow{p_i q_j}$. In order to derive a real value, we compute the center of gravity of the histogram.

Adjacency measure based on fuzzy satisfiability. Distance and orientation may not be always relevant, for instance the distance between two regions is the same if those two regions are adjacent by only one pixel, or if a region is surrounded by another region. In the latter case, the center of gravity of the histogram of angles has no meaning. Therefore we propose to include a third feature which is a topological feature that measures the adjacency length between two regions.

One way to estimate this measure is to compute the matching between the portion of space “near” a reference region and the other region. This measure is maximal in the case where the reference region is embedded into the other one, and is minimal if the two regions are far away from each other.

Fuzzy representations are appropriate to model the intrinsic imprecision of several relations (such as "near") and the necessary flexibility for spatial reasoning [7]. We define the region of space in which a relation to a given object is satisfied. The membership degree of each point to this fuzzy set corresponds to the satisfaction degree of the relation at this point [7]. Note that this representation is in the image space and thus may be more easily merged with an image of a region.

The spatial relation "near" is defined as a distance relation. A distance relation can be defined as a fuzzy interval f of trapezoidal shape on $\mathbb{R}^+$. A fuzzy subset μ_d of the image space $\mathcal{S}$ can then be derived by combining f with a distance map d_R to the reference object R: $\forall x \in \mathcal{S},\ \mu_d(x) = f(d_R(x))$, where $d_R(x) = \inf_{y \in R} d(x, y)$. Figure 3 presents a region (a) and the fuzzy subset corresponding to "Near region 1" (d). In our experiments, the fuzzy interval f is defined with the following fixed values: 0, 0, 10, 30.

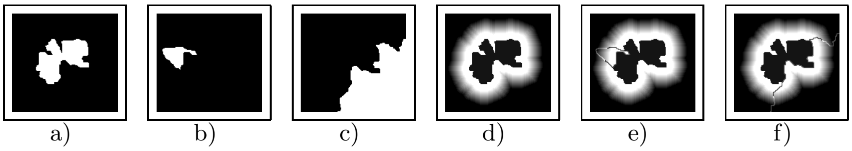

Fig. 2. (a) Region 1. (b) Region 2. (c) Region 3. (d) Fuzzy subset corresponding to "Near region 1". (e) The same with boundary of region 2 added. (f) The same with boundary of region 3 added

So far we have defined the portion of space in which the relation "near" a reference object is defined. The next step consists in estimating the matching between this fuzzy representation and the other region. Among all possible fuzzy measures, we choose as a criterion a *M-measure of satisfiability* [16] defined as:

$$Sat(near(R_1), R_2) = \frac{\sum_{x \in \mathcal{S}} \min(\mu_{near(R_1)}(x), \mu_{R_2}(x))}{\sum_{x \in \mathcal{S}} \mu_{near(R_1)}(x)}$$

where $\mathcal{S}$ denotes the spatial domain. It measures the precision of the position of the object in the region where the relation is satisfied. It is maximal if the whole object is included in the kernel of $\mu_{near(R_1)}$. Note that the size of the region where the relation is satisfied is not restricted and could be the whole image space. If object R_2 is crisp, this measure reduces to $\frac{\sum_{x \in R_2} \mu_{near(R_1)}(x)}{\sum_{x \in \mathcal{S}} \mu_{near(R_1)}(x)}$, i.e. the portion of $\mu_{near(R_1)}$ that is covered by the object.

Figure 3 presents three regions: the reference region (a), a small region adjacent to the first one (b) and a bigger region which is partially represented (c). The fuzzy subset corresponding to "Near region 1" is illustrated in (d) and the border of the others regions have been added in (e) and (f). The value of the satisfiability measure between the fuzzy subset "Near region 1" and region 2 is 0.06, and for region 3, 0.29.

We also choose a symmetric measure, on the contrary of the satisfiability measure, the *M-measure of resemblance* [16] defined as :

$$Res(near(R_1), R_2) = \frac{\sum_{x \in \mathcal{S}} \min(\mu_{near(R_1)}(x), \mu_{R_2}(x))}{\sum_{x \in \mathcal{S}} \max(\mu_{near(R_1)}(x), \mu_{R_2}(x))}$$

This measure is maximal if the object and the relation are identical: this resemblance measure accounts for the positioning of the object and for the precision of the fuzzy set as well.

4 Attribute fusion

We have presented three features corresponding to the principal spatial relations. All these features are normalized in the following. We present now how those features are incorporated in the kernel.

The interest of fusion is to provide a single kernel representation for heterogeneous data, here different types of attributes. For a given graph training set, the first step of the classification task is to build the base kernel matrices $\{K_{a_1}, \ldots, K_{a_n}\}$ corresponding to each attribute take into account. These matrices are basic in the way that each of them represents a narrow view of the data. For a difficult set of images, classification in such basic feature spaces might not be efficient, because a reliable discrimination is not performed using only one attribute. In these cases, fusion of the information brought by each kernel is necessary. The most straightforward solution to this problem is to build a linear combination of the base kernels $K = \sum_{i=1}^{n} \lambda_i K_{a_i}$.

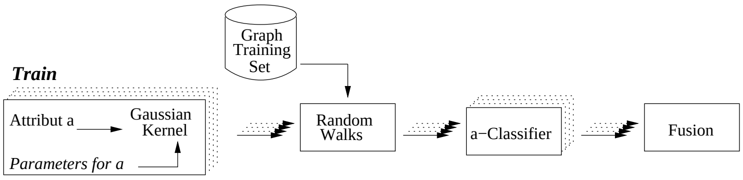

Fig. 3. Fusion of attribute kernels at learning step. For each attribute, a Gaussian kernel is computed with the corresponding parameter. For each of these attribute kernels, the random walk function creates a different classifier using the graphs extracted from the training database. Finally, classifiers are merged using a linear combination

This type of linear combination represents a compromise that allows mutual compensation among different views of the data, thus ameliorating the classification flexibility. The problem of optimally retrieving the weight vector λ has been addressed in [6] and consists in globally optimizing over the convex cone P of symmetric, positive definite matrices: $P = \{X \in \mathbb{R}^{p \times p} \mid X = X^T, X \succeq 0\}$ the following SVM-like dual problem

$$\min_{\lambda \in \mathbb{R}^n, K \in P} \max_{\alpha \in \mathbb{R}^m} 2\alpha^T e - \alpha^T D(y)\ K\ D(y)\alpha\ ,\ \textit{subject to} \qquad (1)$$

$$C \geq \alpha \geq 0,\ trace(K) = c,\ K = \sum_{i=1}^{n} \lambda_i K_{a_i},\ \alpha^T y = 0$$

where m is the size of the training database, $e \in \mathbb{R}^m$ is the vector whose elements are equal to 1 and $D(y) \in \mathbb{R}^m \times \mathbb{R}^m$ is the matrix whose elements are null except those on diagonal which are the labels (+1 or -1) of the training examples, $D(y)_{ii} = y_i$. In the problem specified above, C represents the soft margin parameter, while $c \geq 0$ fixes the trace of the resulting matrix. The interest of this program is that it minimizes the cost function of the classifier with respect to both the discriminant boundary and the parameters λ_i. The output is a set of weights and a discriminant function that combines information from multiple kernel spaces.

The problem can be transposed into the following quadratically constrained quadratic program [6], whose primal-dual solution indicates the optimal weights λ_i:

$$\min_{\alpha,t} 2\alpha^T e - ct\ ,\ subject\ to \tag{2}$$

$$t \geq \frac{1}{trace(K_i)} \alpha^T\ D(y)\ K_{a_i}\ D(y)\ \alpha \quad i = 1, \ldots, n$$

$$C \geq \alpha \geq 0, \alpha^T y = 0$$

We define a kernel function for each attribute, using one of the basic types mentioned above (Gaussian or triangular). Kernel parameters are selected according to their feature variability in the data. More precisely, the threshold for the discrimination function should roughly indicate the smallest distance between two feature values that would trigger a 0-similarity decision for an observer. This threshold is closely correlated to the type of the attribute and equally to the data being analyzed.

For each of the attribute kernels above, we build a graph kernel that will provide us with a graph similarity estimate based on a single feature of the data. Some features are more discriminative than others for a specific data set and therefore generate a better classifier. The fusion method presented above allows us to build a heterogenous decision function that weighs each feature based on its relative relevance in the feature set through its weight μ_i, thus providing optimal performance with the given feature kernels as inputs.

5 Experiments and results

The IBSR database[4] contains real clinical data and is a widely used 3D healthy brain magnetic resonance image (MRI) database. It provides 18 manually-guided expert brain segmentations, each of them being available for three different views: axial, sagittal and coronal. Each element of the database is a set of slices that cover the whole brain.

The main purpose of the database is to provide a tool for evaluating the performance of segmentation algorithms. However, the fact that it is freely available

[4] Internet Brain Segmentation Repository,
available at http://www.cma.mgh.harvard.edu/ibsr/

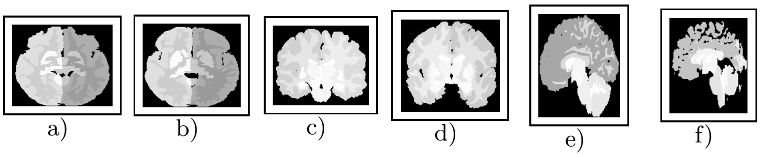

Fig. 4. Samples from IBSR database. Gray levels represent labels. (a) (b) Two slices of the axial view of the same 3D MRI volume representing both classes. (c) (d) Coronal view. (e) (f) Sagittal view

Table 1. Identification of the slices composing the database in each view of the 3D volume, for the three possible views: axial (A), sagittal (S) and coronal (C)

View	# slices	Slices class 1	Slices class 2
A	255	121, 122, 123	126, 127, 128
S	255	121, 122, 123	126, 127, 128
C	128	58, 59, 60	64, 65, 66

and that it offers high quality segmentations makes it also useful for our experiments. Image classification between two different views is performed with a 100% success rate for many of the attributes that we take into account; as a result, we had to build a more challenging classification problem. We try to perform classification on images belonging to the same view; each element of the database belonging to the view will provide three slices in a row for the first class, and other three for the second one. In each set of 54 images that define a class, we choose fifteen images for training (randomly), and the rest of them are used for testing the classifier. Table 1 references the index of slices that are used for defining each class, and for each of the three views.

For assessing attribute similarity, we use Gaussian kernels with relatively small thresholds that render them sensitive to the differences in the labeling. Each attribute kernel is injected in a graph marginalized kernel that we use in the SVM algorithm. For the regularization parameter C of the SVM that controls the trade-off between maximizing the margin and minimizing the L_1 norm of the slack vector, we perform a grid-search with uniform resolution in $\log_2$ space: $\log_2 C \in \{-5, \ldots, 15\}$. For each classification task we use $N = 30$ training graphs and $T = 78$ test graphs, both evenly divided for the two classes.

Further, fusion is performed for k multiple attributes (spatial relations and region descriptors), based on their corresponding marginalized kernels. We fix the trace constraint parameter of the fusion algorithm $c = kN$ and we compute the weights $\lambda_1, \ldots, \lambda_k$ for the input kernels in the fusion function, by solving the system 2 with `cvx`[5]. Finally, the performance of the resulting kernel is tested in an SVM classifier.

In most cases, preliminary results show an amelioration of the performance compared to the initial classification rates, thus proving the interest of the fusion approach for these image kernels. In lines 1, 3, 4, 6 and 16, attributes seem to

[5] Matlab Software for Disciplined Convex Programming,
available at http://www.stanford.edu/ boyd/cvx/

Table 2. Classification performance for different attributes (sa: satisfiability; re: resemblance; su: relative surface; co: compacity; gr: gray level). Columns 2, 4 and 6 list the kernel parameters, and columns 3, 5 et 7 outline the individual classification performance for each attribute applied to each view

	sa		re		su		co		gr	
	Par.	%	Par.	%	Par.	%	Par.	%	Par.	%
axial	0.01	0.79			0.01	0.69	0.01	0.87	0.10	0.65
coronal	0.05	0.74	0.01	0.82			0.01	0.81	0.10	0.86
sagittal	0.05	0.85	0.01	0.95	0.01	0.96	0.01	0.91	0.10	0.81

Table 3. Classification using fusion kernels. Columns 2, 5 and 8 present the attributes used for fusion, and columns 3, 6 and 9 present the performance of the fusion kernel

Axial			Coronal			Sagittal		
No.	**Att.**	**Fusion**	**No.**	**Att.**	**Fusion**	**No.**	**Att.**	**Fusion**
1	sa,su	0.92	6	re,ng	0.99	11	re,su	0.96
2	sa,co	0.90	7	sa,co	0.83	12	re,ng	0.83
3	sa,ng	0.94	8	sa,ng	0.90	13	sa,su	0.96
4	su,ng	0.97	9	ng,co	0.87	14	sa,ng	0.83
5	sa,su,ng	0.96	10	sa,ng,co	0.87	15	sa,co	0.91
						16	ng,co	0.95
						17	sa,ng,co	0.95

provide overall complementary views of the data and therefore their individual performances are greatly topped by that of the fusion. In lines 5 and 17, triple fusion performs as well as the best possible double fusion for the given attributes, thus indicating a saturation effect, based on previous high classification scores. There are also cases (lines 10, 12 or 14) where the fusion weighs more the kernel with a lower performance, thus creating an average performance interpolator. Indeed, optimizing the global convex problem does not directly guarantee a better performance on any testing sample, but gives a better statistical bound on the proportion of errors. Another important aspect that has to be taken into account is the fact that fusion increases the dimensionality of the kernel feature space, and overlearning may occur for small size training sets.

The heaviest step of the algorithm is the computation of the kernel K_{a_i} between two graphs G and G'. The computational complexity associated with this operation is $O((|G||G'|)^3)$, corresponding to a few milliseconds for the images of the IBSR database and one minute for more complex graphs with 60-70 nodes.

6 Conclusion

A method for image classification based on marginalized kernels has been proposed. In particular, we show that a graph representation of the image, enriched with numerical attributes characterizing both the image regions and the spatial relations between them, associated with a fusion of the attributes, leads to improved performances. A kernel is derived for each attribute and fusion of the kernels is performed using a weighted average, in which weights are automatically estimated so as to give more importance to the most relevant attributes.

Preliminary results on medical images illustrate the interest of the proposed approach.

Future work aims at extending the experimental study on other and larger image databases and for more meaningful problems. From a methodological point of view, it could be interesting to investigate different types of fusion.

References

1. Chapelle, O., Haffner, P., Vapnik, V.: Svms for histogram-based image classification. In: IEEE Transactions on Neural Networks, special issue on Support Vectors. (1999)
2. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: Proc. IEEE Int. Conf. on Computer Vision. (2005)
3. Neuhaus, M., Bunke, H.: Edit distance based kernel functions for attributed graph matching. In: 5th IAPR TC-15 Workshop on Graph-based Representations in Pattern Recognition, Poitier, France (2005) 352–361
4. Neuhaus, M., Bunke, H.: A random walk kernel derived from graph edit distance. In: 11th International Workshop on Structural and Syntactic Pattern Recognition. Volume 4109., Hong-Kong, China, springer (2006) 191–199
5. Aldea, E., Atif, J., Bloch, I.: Image Classification using Marginalized Kernels for Graphs. In: 6th IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, GbR'07. Volume 1., Alicante, Spain (2007) 103–113
6. Lanckriet, G., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.: Learning the Kernel Matrix with Semidefinite Programming. Journal of Machine Learning Research **5** (2004) 27–72
7. Bloch, I.: Fuzzy Spatial Relationships for Image Processing and Interpretation: A Review. Image and Vision Computing **23** (2005) 89–110
8. Brun, L., Mokhtari, M., , Meyer, F.: Hierarchical watersheds within the combinatorial pyramid framework. In: 12th International Conference on Discrete Geometry for Computer Imagery. Volume 3429., Poitiers, France, Springer (2005) 34–44
9. Haris, K., Estradiadis, S.N., Maglaveras, N., Katsaggelos, A.K.: Hybrid image segmentation using watersheds and fast region merging. IEEE Transactions on Image Processing **7** (1998) 1684–1699
10. Gaertner, T., Flach, P., Wrobel, S.: On graph kernels: Hardness results and efficient alternatives. In: 16th Annual Conference on Computational Learning Theory, Washington, DC, USA (2003) 129–143
11. Kashima, H., Tsuda, K., Inokuchi, A.: Marginalized kernels between labeled graphs. In: Proc. 20st Int. Conf. on Machine Learning. (2003) 321–328
12. Mahé, P., Ueda, N., Akutsu, T., Perret, J.L., Vert, J.P.: Extensions of marginalized graph kernels. In: ICML '04: Proc. 21st Int. Conf. on Machine Learning. (2004)
13. Kuipers, B.: Modeling spatial knowledge. Cognitive Science **2** (1978) 129–153
14. Bloch, I., Ralescu, A.: Directional Relative Position between Objects in Image Processing: A Comparison between Fuzzy Approaches. Pattern Recognition **36** (2003) 1563–1582
15. Miyajima, K., Ralescu, A.: Spatial organization in 2d segmented images: representation and recognition of primitive spatial relations. Fuzzy Sets and Systems **65** (1994) 225–236
16. Bouchon-Meunier, B., Rifqi, M., Bothorel, S.: Towards general measures of comparison of objects. Fuzzy sets and Systems **84(2)** (1996) 143–153