Integrating Image Segmentation and Classification for Fuzzy Knowledge-based Multimedia Indexing (*)

Thanos Athanasiadis[1], Nikolaos Simou[1], Georgios Papadopoulos[2], Rachid Benmokhtar[3], Krishna Chandramouli[4], Vassilis Tzouvaras[1], Vasileios Mezaris[2], Marios Phiniketos[1], Yannis Avrithis[1], Yiannis Kompatsiaris[2], Benoit Huet[3], and Ebroul Izquierdo[4]

[1] Image, Video and Multimedia Systems Laboratory, National Technical University of Athens, Greece, (thanos,nsimou,tzouvaras,finik,iavr)@image.ntua.gr
[2] Informatics and Telematics Institute, Centre for Research and Technology Hellas (CERTH), Greece, (papad,bmezaris,ikom)@iti.gr
[3] Institut Eurécom, Département Multimédia, France, (rachid.benmokhtar,benoit.huet)@eurecom.fr
[4] Department of Electronic Engineering, Queen Mary University of London, UK, (krishna.chandramouli,ebroul.izquierdo)@elec.qmul.ac.uk

Abstract. In this paper we propose a methodology for semantic indexing of images, based on techniques of image segmentation, classification and fuzzy reasoning. The proposed knowledge-assisted analysis architecture integrates algorithms applied on three overlapping levels of semantic information: i) no semantics, i.e. segmentation based on low-level features such as color and shape, ii) mid-level semantics, such as concurrent image segmentation and object detection, region-based classification and, iii) rich semantics, i.e. fuzzy reasoning for extraction of implicit knowledge. In that way, we extract semantic description of raw multimedia content and use it for indexing and retrieval purposes, backed up by a fuzzy knowledge repository. We conducted several experiments to evaluate each technique, as well as the whole methodology in overall and, results show the potential of our approach.

1 Introduction

Production of digital content has become daily routine for almost every person, leading to an immense size of accessible multimedia data. Consequently, public and research interest has partly shifted from the production of multimedia content to its efficient management, making apparent the need of mechanisms for automatic indexing and retrieval, thematic categorization and content-based

(*) This research was supported by the European Commission under contract FP6-027026 K-SPACE.

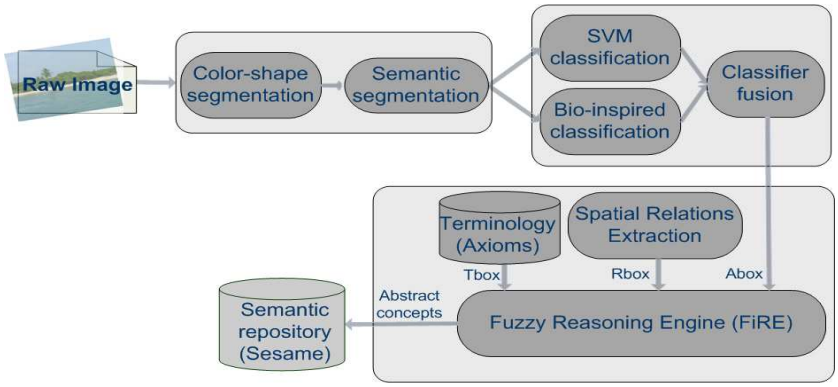


Fig. 1. Overview of the proposed architecture.

search (among many others). Efficient multimedia content management and usability requires focus on the semantic information level, with which most users desire to interact; other than that would render any results ineffective.

The importance of semantic indexing and retrieval of multimedia has brought out several benchmarking activities, such as TRECVID [11] with increasing participation every year. Most approaches in semantic-based analysis and indexing are grounded on multimedia segmentation and object recognition techniques. The majority of classification techniques employ statistical modeling, associating low-level visual features with mid-level concepts [8]. There have been proposed techniques for region-based classification using machine learning techniques such as Self Organizing Maps (SOMs) [5], Genetic Algorithms [10], Support Vector Machines (SVMs) [16, 10] and biologically inspired optimization techniques. To achieve better recognition rates, it has been found that, it is better to fuse multiple simple classifiers than to build a single sophisticated classifier [4].

During the late years, various attempts were made in order to extract complicated concepts using multimedia analysis results combined with taxonomies and ontologies. In [12] WordNet is used to include lexical relationships between abstract and detected mid-level concepts. Ontologies based on Description Logics (DLs) [3] are a family of knowledge representation languages; however, despite the rich expressiveness of DLs, they lack the ability to deal with vague and uncertain information which is commonly found in multimedia content. This was the reason that a variety of DLs capable of handling imprecise information, like probabilistic and fuzzy [14, 13] have been proposed.

Within this context, our paper presents a knowledge assisted image analysis and semantic annotation methodology consisting of several novel and state-of-the-art techniques. As depicted in Figure 1, we discuss methods for semantic-aware segmentation, object detection and recognition, as well as detection of abstract concepts that cannot be detected directly, but can only be inferred using higher level knowledge. We follow a bottom-up approach and therefore we initially segment the image based on color and shape criteria, followed by a

novel semantic region growing methodology which incorporates object detection simultaneously with region merging that improves extraction of semantic objects.

Next, two different classification approaches are employed and used for recognition of several concepts: i) Support Vector Machines and ii) a biologically inspired classifier. Combination of multiple classifier decisions is a powerful method for increasing classification rates in recognition problems. We fuse the two sets of classification results, using a neural network based on evidence theory method, obtaining a single list of concepts with degrees of confidence for all regions.

So far a list of concepts (together with degrees of confidence for each one) have been linked to the image. Our goal lies beyond this and we want to extract additional, implicit knowledge, improve region-based classification by incorporating spatial relations and neighborhood information and finally infer abstract concepts on a global image basis. Towards this aim, a fuzzy reasoning engine is employed. The final results are stored in an online semantic repository, in a strictly structured format, allowing query mechanisms for semantic retrieval.

The manuscript is structured as follows: Section 2 details the mechanism of each algorithm used towards a bottom-up image classification. Section 3 presents the role of fuzzy multimedia reasoning and its application in fuzzy semantic indexing, storing in appropriate knowledge bases and querying mechanisms for retrieval purposes. We provide extended experimental results of the overall approach in section 4 and we draw our conclusions in section 5.

2 Bottom-up Image Classification

In this section we describe a series of image analysis techniques, whose integration leads to the detection and recognition of a set of concepts used as the basis for the semantic handling of the content. As a bottom-up technique, it starts from the pixel level jumping to the region level using a color and shape image segmentation, further refined with a semantic region growing technique (subsection 2.1). Two classifiers are used in parallel, described in subsection 2.2, which assign concepts in a fuzzy manner (i.e. with a degree of confidence) for each region of the image. The last subsection presents a fusion mechanism, based on a neural network, which fuses the results of the two classifiers and produces a single set of concepts detected in the image. This set of concepts provides the initial vocabulary for the semantic description of the image.

2.1 Semantic Image Segmentation

Initially, a segmentation algorithm, based on low-level features such as color and shape [1], is applied in order to divide the given image into a set of non overlapping regions. In previous work ([2]) we have shown how extracted visual descriptors can be matched to visual models of concepts resulting to an initial fuzzy labeling of the regions with concepts from the knowledge base, i.e. for region a we have the fuzzy set (following the sum notation [7]) $L_a = \sum_k C_k/w_k$, where $k = 1, \ldots, K$, K is the cardinality of the crisp set of all concepts $\mathbf{C} = \{C_k\}$

in the knowledge base and $w_k = \mu_a(C_k)$ is the degree of membership of element C_k in the fuzzy set L_a.

Segmentation based only on syntactic features usually creates more regions than the actual number of objects. We examine how a variation of a traditional segmentation technique, the Recursive Shortest Spanning Tree (RSST) can be used to create more semantically coherent regions in an image. The idea is that neighbor regions, sharing the same concepts, as expressed by the labels assigned to them, should be merged, since they define a single object. To this aim, we modify the RSST algorithm to operate on the fuzzy sets of labels $\mathcal{L}$ of the volumes in a similar way as if it worked on low-level features (such as color, texture) [2]. The modification of the traditional algorithm to its semantic equivalent lies on the re-definition of the two criteria: (i) The dissimilarity between two neighbor regions a and b (vertices v_a and v_b in the graph), based on which graph's edges are sorted and (ii) the termination criterion. For the calculation of the similarity between two regions we defined a metric between two fuzzy sets, those that correspond to the candidate concepts of the two regions. This dissimilarity value is computed according to the following formula and is assigned as the weight of the respective graph's edge e_{ab}:

$$w(e_{ab}) = 1 - \sup_{C_k \in \mathbf{C}} (\top(\mu_a(C_k), \mu_b(C_k))) \tag{1}$$

where $\top$ is a t-norm, a and b are two neighbor regions and $\mu_a(C_k)$ is the degree of membership of concept $C_k \in \mathbf{C}$ in the fuzzy set L_a.

Let us now examine one iteration of the S-RSST algorithm. Firstly, the edge e_{ab} with the least weight is selected, then regions a and b are merged. Vertex v_b is removed completely from the ARG, whereas v_a is updated appropriately. This update procedure consists of the following two actions:

1. Re-evaluation of the degrees of membership of the labels fuzzy set in a weighted average (w.r.t. the regions' size) fashion.
2. Re-adjustment of the ARG edges by removing edge e_{ab} and re-evaluating the weight of the affected edges.

This procedure continues until the edge e^* with the least weight in the ARG is bigger than a threshold: $w(e^*) > T_w$. This threshold is calculated in the beginning of the algorithm, based on the histogram of all weights of the set of all edges.

2.2 Region Classification

SVM-based Classification. SVMs have been widely used in semantic image analysis tasks due to their reported generalization ability and their efficiency in solving high-dimensionality pattern recognition problems [15]. Under the proposed approach, SVMs are employed for performing the association of the computed image regions to one of the defined high-level semantic concepts based on the estimated region feature vector. In particular, a SVM structure is utilized, where an individual SVM is introduced for every defined concept $C_k \in \mathbf{C}$,

to detect the corresponding instances. Every SVM is trained under the '*one-against-all*' approach. The region feature vector, consisting of seven MPEG-7 visual descriptors, constitutes the input to each SVM, which at the evaluation stage returns for every image segment a numerical value in the range $[0, 1]$. This value denotes the degree of confidence, $\mu_a(C_k)$, to which the corresponding region is assigned to the concept associated with the particular SVM [10]. For each region, the maximum of the K calculated degrees of confidence, $argmax(\mu_a(C_k))$, indicates its concept assignment, whereas the pairs of all supported concepts and their respective degree of confidence $\mu_a(C_k)$ computed for segment a comprise the region's concept hypothesis set $H_a^{\mathbf{C}} = \{\mu_a(C_k)\}$.

Bio-inspired Classifier. Neural network based clustering and classification has been dominated by Self Organizing Maps (SOMs) and Adaptive Resonance Theory (ART). In competitive neural networks, active neurons reinforce their neighbourhood within certain regions, while suppressing the activities of other neurons. This is called on-center/off-surround competition. The objective of SOM is to represent high-dimensional input patterns with prototype vectors that can be visualized in a usually two-dimensional lattice structure. Input patterns are fully connected to all neurons via adaptable weights. During the training process, neighbouring input patterns are projected into the lattice, corresponding to adjacent neurons. An individual SOM network is employed to detect instances of the defined high-level semantic concepts. Each SOM is trained under the one against all approach. In the basic training algorithm are the prototype vectors trained according to $m_d(t+1) = m_d(t)+g_{cd}(t)[x-m_d(t)]$ where m_d is the weight of the neurons in the SOM network, $g_{cd}(t)$ is the neighbourhood function and d is the dimension of the input feature vector. Each SOM network corresponding to defined high-level concept returns for every segment a numerical value in the range of $[0, 1]$, denoting the degree of confidence to which the corresponding region is assigned to the concept associated with the particular SOM.

2.3 Classifier Fusion

In this section, we describe how the evidence theory can be applied to fusion problems and outline our recently proposed neural network based on evidence theory (NNET) to address classifier fusion [4]. The objective is to associate for each object x (image region), one class from the set of classes $\Omega = \{w_1, .., w_K\}$. In our case, the set of classes is equivalent to the set of concepts $\mathbf{C}$, defined previously. This association is given via a training set of N samples, where each sample can be considered as a part of belief for one class of Ω. This belief degree can be assimilated to evidence function m^i, with 2 focal elements: The class of x^i noted w_q, and Ω. So, if we consider that the object x^i is near to x, then a part of belief can be affected to w_q and the rest to Ω. The mass function is obtained by decreasing function of distance as follows:

$$\begin{cases} m^i(\{w_q\}) = \alpha^i \phi_q(d^i) \\ m^i(\Omega) = 1 - \alpha^i \phi_q(d^i) \end{cases} \tag{2}$$

Where $\phi(.)$ is a monotonically decreasing function such as an exponential function $\phi_q(d^i) = \exp(-\gamma_q(d^i)^2)$, and d^i is an Euclidean distance between the vector x and the i^{th} vector of training base. $0 < \alpha < 1$ is a constant which prevents a total affectation of mass to the class w_q when x and i^{th} samples are equal. γ_q is a positive parameter defining the decreasing speed of mass function. A method for optimizing parameters (α, γ_q) has been described in [6]. We obtain N mass functions, which can be combined into a single one using (3):

$$m(A) = (m^1 \oplus ... \oplus m^N) = \sum_{(B_1 \bigcap ... \bigcap B_N)=A} \prod_{i=1}^{N} m^i(B_i) \tag{3}$$

We propose to resume work already made with the evidence theory in the connectionist implementation [4, 6], and to adapt it to classifier fusion. For this aim, an improved version of RBF neural network based on evidence theory which we call NNET, with one input layer L_{input}, two hidden layers L_2 and L_3 and one output layer L_{output} has been devised.

Layer L_{input}. It contains N units and is identical to an RBF network input layer with an exponential activation function ϕ. d is a distance computed using training data and dictionary created (clustering method). K-means is applied on the training data in order to create a "visual" dictionary of the regions.

Layer L_2. Computes the belief masses m^i (2) associated to each prototype. It is composed of N modules of $K+1$ units each $m^i = (m^i(\{w_1\}), ..., m^i(\{w_{K+1}\})) = (u_1^i s^i, ..., u_K^i s^i, 1 - s^i)$ where u_q^i is the membership degree to each class w_q, q class index $q = \{1, ..., K\}$. The units of module i are connected to neuron i of the previous layer. Note that each region in the image can belong to only one class.

Layer L_3. The Dempster-Shafer combination rule combines N different mass functions in one single mass, given by the conjunctive combination (3). For this aim, the activation vector $\vec{\mu^i}$ can be recursively computed by $\mu^1 = m^1$, $\mu_j^i = \mu_j^{i-1} m_j^i + \mu_j^{i-1} m_{K+1}^i + \mu_{K+1}^{i-1} m_j^i$ and $\mu_{K+1}^i = \mu_{K+1}^{i-1} m_{K+1}^i$

Layer L_{output}. In [6], the output is directly obtained by $O_j = \mu_j^N$. The experiments show that this output is very sensitive to the number of prototype, where for each iteration, the output is purely an addition of ignorance. Also, we notice that a small change in the number of prototype can change the classifier fusion behavior. To resolve this problem, we use normalized output: $O_j = \frac{\sum_{i=1}^{N} \mu_j^i}{\sum_{i=1}^{N} \sum_{j=1}^{K+1} \mu_j^i}$. Here, the output is computed taking into account the activation vectors of all prototypes to decrease the effect of an eventual bad behavior of prototype in the mass computation.

The different parameters (Δu, $\Delta \gamma$, $\Delta \alpha$, ΔP, Δs) can be determined by gradient descent of output error for an input pattern x. Finally, the maximum of plausibility P_q of each class w_q is computed: $P_q = O_q + O_{K+1}$.

3 Fuzzy Reasoning and Indexing

Image classification algorithms can provide reliable results on the recognition of specific concepts, however, it is very difficult to recognize higher-level concepts

that do not have specific low-level features. That kind of concepts can be effectively represented by an ontology capable of handling the imprecise information provided by image segmentation and classification algorithms. A DL that fullfills these requirements is f-SHIN [13]. Using fuzzy reasoning engine FiRE[5], which supports f-$\mathcal{SHIN}$ and its reasoning services, we improve region-based classification results and extract additional implicit concepts that categorize an image. The extracted information is stored in a semantic repository permitting fuzzy conjunctive queries for semantic image and region retrieval.

3.1 Fuzzy Reasoning Services and Querying

A f-$\mathcal{SHIN}$ knowledge base Σ is a triple $\langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$, where $\mathcal{T}$ is a fuzzy $TBox$, $\mathcal{R}$ is a fuzzy $RBox$ and $\mathcal{A}$ is a fuzzy $ABox$. $TBox$ is a finite set of fuzzy concept axioms which are of the form $C \equiv D$ called fuzzy concept inclusion axioms and $C \sqsubseteq D$ called fuzzy concept equivalence axioms, where C, D are concepts, saying that C is equivalent or C is a sub-concept of D, respectively. Similarly, $RBox$ is a finite set of fuzzy role axioms of the form $\mathsf{Trans}(R)$ called fuzzy transitive role axioms and $R \sqsubseteq S$ called fuzzy role inclusion axioms saying that R is transitive and R is a sub-role of S respectively. Ending, $ABox$ is as finite set of fuzzy assertions of the form $\langle a : C \bowtie n \rangle$, $\langle (a, b) : R \bowtie n \rangle$, where $\bowtie$ stands for $\geq, >, \leq$ or $<$ or $a \neq b$. Intuitively, a fuzzy assertion of the form $\langle a : C \geq n \rangle$ means that the membership degree of a to the concept C is at least equal to n. Finally, assertions defined by $\geq, >$ are called *positive* assertions, while those defined by $\leq, <$ *negative* assertions.

The main reasoning services of crisp reasoners are deciding satisfiability, subsumption and entailment of concepts and axioms w.r.t. an Σ. In other words, these tools are capable of answering queries like "Can the concept C have any instances in models of the ontology T?" (satisfiability of C), "Is the concept D more general than the concept C in models of the ontology T ?" (subsumption $C \sqsubseteq D$) of does axiom Ψ logically follows from the ontology (entailment of Ψ). These reasoning services are also available by FiRE together with *greatest lower bound queries* which are specific to fuzzy assertions. Since in fuzzy DLs individuals participate in concepts and are connected with a degree, satisfiability queries are of the form "Can the concept C have any instances with degree of participation $\bowtie n$ in models of the ontology T ?". Furthermore, it is in our interest to compute the best lower and upper truth-value bounds of a fuzzy assertion. The term of *greatest lower bound* of a fuzzy assertion w.r.t. Σ was defined in [14]. Greatest lower bound are queries like "What is the greatest degree n that our ontology entails an individual a to participate in a concept C?".

In order to store the fuzzy knowledge base produced by FiRE in a Sesame RDF Repository we serialize it into RDF triples. For this purpose, we use blank nodes in order to represent fuzzy information by defining three new entities: `frdf:membership`, `frdf:degree` and `frdf:ineqType` as types of `rdf:Property` while properties are defined for each role assertion. In that way, Sesame is used

[5] FiRE can be found at `http://www.image.ece.ntua.gr/~nsimou/FiRE/`

as a back end for storing and querying RDF triples while FiRE is the front end by which the user can store and query a fuzzy knowledge base.

Since in our case we extend classical assertions to fuzzy assertions, new methods of querying such fuzzy information are possible. More precisely, in [9] the authors extend ordinary conjunctive queries to a family of significantly more expressive query languages, which are borrowed from the fields of fuzzy information retrieval. These languages exploit the membership degrees of fuzzy assertions by introducing weights or thresholds in query atoms. Similarly using FiRE and Sesame permits conjunctive fuzzy queries. Queries are converted from the FiRE syntax to the SeRQL query language supported by Sesame. Sesame engine evaluates the results which are then visualized by FiRE.

Queries consist of two parts: the first one specifies the individual(s) that will be evaluated while the second one states the condition that has to be fulfilled for the individuals. This query asks for individuals x and y, x has to participate in concept Beach to at least 0.7, it also has to be the subject of a contains assertion with participation greater than 1, having as a role-filler individual y which has to participate in concept Person to at least 0.8.

3.2 The fuzzy knowledge base

In order to effectively categorize images and also improve the semantic segmentation process we have implemented an expressive terminology. The terminology defines new concepts that characterize an image and also refines concepts extracted by the classification modules considering regions' spatial configuration.

First, we present the input used as the assertional part of the fuzzy knowledge base provided by the analysis modules. After an initial segmentation, an image is divided into a number of segments. Their spatial relations extracted by the semantic RSST comprise the $RBox$ of the fuzzy knowledge base. The classification algorithms evaluate a participation degree in a set of concepts for every segment. The obtained results are then fuzzed and used as positive assertions to represent the $ABox$ of the fuzzy knowledge base. Hence, the alphabet of concepts $\mathbf{C}$ and roles $\mathbf{R}$ is: $\mathbf{C} = \{$*Sky Building Person Rock Tree Vegetation Sea Grass Ground Sand Trunk Dried-plant Pavement Boat Wave*$\}$ and $\mathbf{R} = \{$*above-of below-of left-of right-of contains*$\}$. The set of individuals consist of the amount of segments obtained for each image together with the whole image. The $TBox$ can be found in Table 1. As can be observed, concepts like Sky that are extracted by the classification modules have been re-defined using spatial relations.(Those concepts are shown in capitals.) Hence, SKY has been defined as a segment that was classified as Sky and has a $\mathsf{above-of}$ neighbor that is either Sea or Building or Vegetation. Additionally, higher concepts that refer to a segment have been defined like WavySea and SandyBeach also concepts like Beach that refer to the whole image and categorize it. Within our knowledge base Beach has been defined as an image that contains segments labeled as Sky and Sea. According to the defined terminology implicit knowledge is extracted. For every image *greatest lower bound* (glb) reasoning service is used for the defined concepts of the

Table 1. The terminology $TBox$.

$\mathcal{T} = \{$SKY	$\equiv$ Sky $\sqcap$ ($\exists$above $-$ of.Sea $\sqcup$ $\exists$above $-$ of.Building $\sqcup$ $\exists$above $-$ of.Vegetation),
SAND	$\equiv$ Sand $\sqcap$ $\exists$below $-$ of.Sea,
PAVEMENT	$\equiv$ Pavement $\sqcap$ $\exists$below $-$ of.Building,
TRUNK	$\equiv$ Trunk $\sqcap$ ($\exists$above $-$ of.Ground $\sqcup$ $\exists$above $-$ of.Grass),
VEGETATION	$\equiv$ Grass $\sqcup$ Tree $\sqcup$ Vegetation,
WavySea	$\equiv$ Sea $\sqcap$ Wave,
SandyBeach	$\equiv$ Sea $\sqcap$ Sand,
PartOfComplexBuilding	$\equiv$ Building $\sqcap$ ($\exists$left $-$ of.Building $\sqcup$ $\exists$right $-$ of.Building),
Beach	$\equiv$ $\exists$contains.Sea $\sqcap$ $\exists$contains.SKY,
Landscape	$\equiv$ $\exists$contains.VEGETATION,
City	$\equiv$ $\exists$contains.Building $\sqcup$ $\exists$contains.Pavement$\}$
$\mathcal{R} = \{$contains, left $-$ of$^-$ $=$ right $-$ of, above $-$ of$^-$ $=$ below $-$ of$\}$	

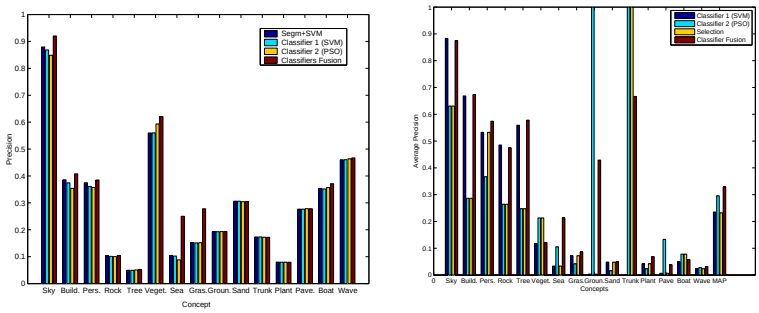
terminology. The obtained implicit results together with the explicit information provided by classifiers(i.e. $ABox$) are stored to a Sesame repository.

4 Experimental Results

In this section we present a series of experiments conducted to demonstrate the gain achieved using the proposed approach in comparison to other techniques. We have set up two datasets of images: One consisting of 500 images, which is accompanied by ground truth at the image level, i.e. we know that concepts either exist or not in the whole image, without any information to which region correspond. The second dataset consists of 350 images, for which we have a finer grained ground truth at a region level, i.e. annotation for every region (2185 regions in total).

In order to evaluate the performance of our integrated approach we compare the recognition rates to those of each individual classifier, as well as to a basic classification method. In the case of the first dataset we had to align the available image level ground truth to the region level classification results. We assumed that when a region has been classified to a concept with a certain degree of confidence, then the maximum degree (over all regions) can be propagated to the image itself. Following this procedure for every concept we end up with a list of concepts and confidence values detected *somewhere* in the image.

First Experiment: Evaluation at the image level. For this first experiment, we calculated the performance of a simple classification approach which is based on a simple color RSST segmentation, descriptor extraction and SVM classification. We examined the performance of the semantic segmentation, of the SVM classifier (2.2), of the bio-inspired classifier (2.2), as well as that of the fusion mechanism (2.3). Figure 2a illustrates the precision rate of the above four algorithms for all 15 concepts. Additionally we calculated the overall performance of



(a) (b)

Fig. 2. Precision for the different classification results for every concept.

the above four techniques, irrespectively to the concept, using a weighted average of precision and recall values of each concept. Each concept's weight depends on the frequency of appearance of that concept in the dataset according to the ground truth. Moreover, the weighted harmonic mean (F-measure) of precision and recall was calculated to measure the effectiveness of the classification. In the application of multimedia indexing, we consider precision more important measure than recall, since it is the user's preference to retrieve relevant content with little noise, rather than all the available relevant dataset (which is usually of immense size) and therefore in the computation of the harmonic mean the precision to recall rate is 2:1. The first three columns of Table 2 provide those figures. It is apparent that the NNET fusion provides the best precision for every single concept and also is the most effective (according to the F-measure), while the bio-inspired classifier tops in the recall figures.

Moreover, we calculated the precision and recall of a selected subset of concepts, the most frequent in the dataset. We observe (Table 2 last three columns) a significant increase of all figures, which can be explained by the fact that classifiers were better trained since more example samples were available. We have selected the 6 most frequent concepts, which correspond approximately to the two thirds of detected concepts in the whole dataset.

Table 2. Average classification results for all concepts (image level granularity).

	All Concepts			Frequent Concepts		
Technique	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Segm+SVM	0.45	0.58	0.48	0.57	0.63	0.58
Clasif.1 (SVM)	0.45	0.57	0.48	0.56	0.64	0.58
Clasif.2 (PSO)	0.44	0.82	0.52	0.56	0.85	0.63
NNET Fusion	0.48	0.71	0.54	0.60	0.72	0.64

Second Experiment: Evaluation at the region level. In order to demonstrate the significance of our region-based approach to semantically index images, we set up another experiment, based on the second dataset, for which we have ground truth on the region level. The classification task consists of retrieving regions expressing one of the considered semantic concepts. The performance has been measured using the standard precision and recall metrics. We are interested by the average precision to have the measure of the ability of a system to present only relevant regions.

Figure 2b shows the average precision for the four systems (PSO, SVM for classification, and our NNET in fusion, along with a simple selection approach based on the best classifier output over a validation set). We observe that our fusion model achieves respectable performance with respect to the number of concepts detected, in particular for certain semantic concepts (Building, Person, Tree, Sea, Grass, Sand and Dried plant). Here, NNET fusion combines the converging classifier outputs (PSO and SVM) to obtain an effective decision which improves upon the individual classifier outputs. In comparison to the classifier chosen by the selection, which due to low data representativity of the validation set has not allowed the best detection in the test set, the fusion mechanism is more robust. Interesting findings are obtained for the concepts (Vegetation, Pavement and Boat). The performance of fusion is lower than the result given by one of the two classifiers. This is due to both numerous conflicting classification and limited training data. This also explains the extreme cases obtained for concepts Ground and Trunk.

In order to measure the overall performance for the region-based image classification, we calculate the Mean Average Precision (MAP). The PSO classifier detects concepts with more precision than the SVM classifier, $MAP_{PSO} = 0.30$ and $MAP_{SVM} = 0.23$, while NNET fusion combines the two classifiers and allows an overall improvement $MAP_{NNET} = 0.33$. We observe that these figures are pretty lower than those of Table 2 $(0.48 - 0.54)$, but this should be expected since this evaluation metric has region level granularity. For instance, when searching for images of people in a beach (see also example in the following subsection) the evaluation metric for the image as a whole will consider the maximum degree of confidence for the concept Person, while the region-level approach will also detect the exact position of it in the image. This, we think, is a reasonable trade-off between spatial accuracy and global precision rate.

5 Conclusions

This paper contributes to the semantic indexing of images based on algorithms of varying granularity of semantic information, each one targeting to solve partially the problem of bridging the semantic gap. The integrated framework consists of a novel semantic image segmentation technique, a bottom-up image classification using two classifiers of different philosophy and a neural network to fuse the results of the classifiers. This intermediate outcome is further refined and enriched using fuzzy reasoning based on domain-specific knowledge in the for-

mal representation of fuzzy description logics. Finally, the semantic description of the image is stored in a knowledge base which facilitates querying and retrieving of images. Future work of the authors includes implementation of more robust classifiers, integration of richer semantics and broader knowledge, as well as extension to video sequences.

References

1. T. Adamek, N.O'Connor, and N.Murphy. Region-based segmentation of images using syntactic visual features. In *Proc. Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2005*, Switzerland, April 2005.
2. T. Athanasiadis, P. Mylonas, Y. Avrithis, and S. Kollias. Semantic image segmentation and object labeling. *IEEE Trans. on Circuits and Systems for Video Technology*, 17(3):298–312, March 2007.
3. F. Baader, D. McGuinness, D. Nardi, and P. Patel-Schneider. *The Description Logic Handbook: Theory, implementation and applications*. Cambridge University Press, 2002.
4. R. Benmokhtar and B. Huet. Neural network combining classifier based on dempster-shafer theory for semantic indexing in video content. *International MultiMedia Modeling Conference*, 4351:196–205, 2007.
5. K. Chandramouli and E. Izquierdo. Image classification using self organizing feature maps and particle swarm optimization. *in Proc. 7th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2006.
6. T. Denoeux. An evidence-theoretic neural network classifier. *International Conference on Systems, Man and Cybernetics*, 3:712–717, 1995.
7. G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice-Hall, 1995.
8. M. Naphade and T. S. Huang. A probabilistic framework for semantic video indexing, filtering and retrieval. *IEEE Trans. on Multimedia*, 3(1):144–151, March 2001.
9. J. Pan, G. Stamou, G. Stoilos, and E. Thomas. Expressive querying over fuzzy DL-Lite ontologies. In *Proceedings of the International Workshop on Description Logics (DL 2007)*, 2007.
10. G. T. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Combining global and local information for knowledge-assisted image analysis and classification. *EURASIP J. Adv. Signal Process*, 2007(2):18–18, 2007.
11. A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.
12. C. Snoek, B. Huurninkm, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Trans. on Multimedia*, 9(5):144–151, August 2007.
13. G. Stoilos, G. Stamou, V. Tzouvaras, J. Z. Pan, and I. Horrocks. Reasoning with very expressive fuzzy description logics. *Journal of Artificial Intelligence Research*, 30(5):273–320, 2007.
14. U. Straccia. Reasoning within fuzzy description logics. *Journal of Artificial Intelligence Research*, 14:137–166, 2001.
15. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.
16. L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. *International Conference on Image Processing*, 2, 2001.