Ontology Based Object Learning and Recognition : Application to Image Retrieval

Nicolas Maillot Monique Thonnat Céline Hudelot
INRIA Sophia Antipolis - Orion Team,
2004 Route des lucioles - B.P. 93
06902 Sophia Antipolis, France.
email: firstname.name@sophia.inria.fr

Abstract

This paper presents a new object categorization method and shows how it can be used for image retrieval. Our approach involves machine learning and knowledge representation techniques. A major element of our approach is a visual concept ontology composed of several types of concepts (spatial concepts and relations, color concepts and texture concepts). Visual concepts contained in this ontology can be seen as an intermediate layer between domain knowledge and image processing procedures. Our approach is composed of three phases: (1) a knowledge acquisition phase, (2) a learning phase and (3) a categorization phase. This paper is mainly focused on phases (2) and (3). A major issue is the symbol grounding problem which consists of linking meaningfully symbols to sensory information. We propose a solution to this difficult issue by showing how learning techniques can map numerical features to visual concepts.

1. Introduction

This paper presents an object categorization method based on a visual concept ontology. The paper also shows how we have applied the proposed approach to image retrieval. Both knowledge representation and machine learning techniques are involved in the categorization process. The proposed approach is designed for semantic interpretation of isolated objects of interest. Related work on scene analysis issues (i.e. involving non isolated objects) can be found in [2]. Our approach is composed of three phases: (1) a knowledge acquisition phase, (2) a learning phase and (3) a categorization phase.

A long experience in complex object categorization [3] has shown that experts often use a well defined and shared vocabulary for describing the objects of their domain. In [4], we have explained how ontological engineering can be applied to acquire expert knowledge. Our goal is now to show how to use this expert knowledge in order to guide object categorization.

Section 2 gives an overview of key issues and existing approaches in object categorization. Section 3 gives an overview of the proposed knowledge acquisition process. Section 4 explains how visual concepts are learned by machine learning techniques. Section 5 presents an object categorization algorithm. Section 6 details results based on the proposed methodology applied to image indexing and retrieval. We finally conclude in section 7.

2. Related Work

An overview of object categorization techniques can be found in [1]. When object recognition is performed using the whole image, the image is often considered as an arrangement of colored pixels rather than a picture of objects. This abstraction is often called appearance. Object recognition is often achieved by computing global statistics on the whole image (e.g. histograms, color coherence vectors [7]). This kind of approach does not take into account object level semantics.

Considering parts of an image produced by a segmentation process is useful to object-level semantics. In [5], it is explained how image parts can be handled by high-level bayesian image interpretation techniques. The author explains that bayesian analysis techniques are more widely applicable and reliable than ad hoc algorithms. Such statistical models are explicit and allow to evaluate confidence about conclusions.

Knowledge based vision systems have proven to be effective for complex object recognition [3] and for scene understanding [6]. These systems give access to a high semantic level. They offer a great capacity of reusability and extendability and better tractability of the different sub-problems (i.e. image processing, symbol grounding and image interpretation) encountered in image understanding. The major negative point of these systems is that they rely on knowledge bases which are difficult to produce.

To achieve object recognition, we propose an intermediate approach: to use expert knowledge to structure prior distributions of relevant visual features (i.e. texture, color, shape). This means that we aim at using expert knowledge to perform a focused learning of semantically meaningful visual features.

3. Knowledge Acquisition Phase

First come knowledge acquisition issues which have been discussed in [4]. This phase is driven by a visual concept ontology. As seen in fig. 1, knowledge acquisition process consists of achieving the following tasks: domain taxonomy acquisition (i.e. hierarchy of domain classes); ontology driven visual description of domain classes which leads to a more complete domain knowledge base; image sample management (i.e. annotation and manual segmentation of samples of object classes of interest). Sample annotation consists of labeling each manually segmented region of interest by a domain class name.

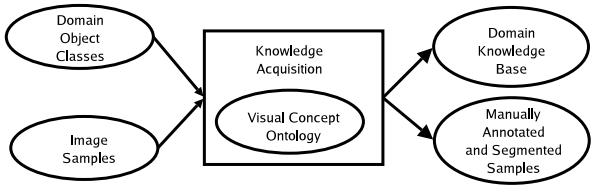

Figure 1. Knowledge acquisition phase overview

Definition 1 *Let Θ be the set of visual concepts. $\preceq_\Theta$ is a partial order between visual concepts. $\forall (C_i, C_j) \in \Theta^2, C_i \preceq_\Theta C_j$ means that C_i is a sub-concept of C_j.*

Definition 2 *Let Φ be the set of domain classes. $\mathcal{A} \in \Theta$ is the set of domain class attributes. $\mathcal{A}$ is a predefined set of visual concepts. $\mathcal{A} = \{geometry, size, orientation, position, hue, brightness, saturation, repartition, contrast, pattern\}$. For a class $\alpha \in \Phi$, $\mathcal{A}_\alpha \subseteq \mathcal{A}$ is the set of attributes of α. $\preceq_\Phi$ is a partial order between domain classes (i.e. superclass attribute). We define $\mathcal{S} : \Phi \rightarrow \Phi$ so that $\mathcal{S}(\alpha)$ is the set of subparts of α (i.e. subparts attribute).*

Definition 3 *Let $a \in \mathcal{A}_\alpha$ be an attribute of $\alpha \in \Phi$. We define $\mathcal{V} : \mathcal{A}_\alpha \rightarrow \Theta$ so that $\mathcal{V}(a)$ is the set of values of a and so that $\forall C_i \in \mathcal{V}(a), C_i \preceq_\Theta a$.*

Definition 4 *Let Γ be the set of manually segmented regions of interest. We define $\mathcal{L}_\Phi : \Phi \rightarrow \Gamma$ so that $\mathcal{L}_\Phi(\alpha)$ is the set of representative regions of interest of a domain class α.*

Knowledge acquisition phase consists of defining Φ, $\preceq_\Phi$, $\mathcal{S}(\alpha)$ and $\mathcal{V}(a)$. Φ, $\preceq_\Phi$ and $\mathcal{S}(\alpha)$ belong to domain knowledge. This knowledge is shared by the specialists of the domain (e.g. biologists, astronomers). It is also independant of any vision layer and can be reused for other purposes. Defining $\mathcal{V}(a)$ allows to reduce the semantic gap between expert knowledge and image level. As explained in the next section, semantic gap is completely filled during a learning phase.

The complete ontology is composed of 103 visual concepts (e.g. Granulated Texture, Coarse Texture, Circular Surface, Dark) . The depth of the ontological tree is 8. This ontology is an extendible basis that can be specialized depending on the application domain. Numerical features are the low-level definitions of visual concepts. 16 numerical features are associated with spatial visual concepts (e.g. compacity, surface). 127 features are used to characterize texture concepts (e.g. cooccurence matrices). 512 features (e.g. color coherence vectors [7]) are associated with color concepts. Association between features and visual concepts defines how visual concepts are computed on image data. For instance, extraction of the visual concept *GranulatedTexture* is performed by computing cooccurence matrices on a region of interest.

An example of a pollen grain class formalized with the frame formalism is shown in table 1. This example results from a knowledge acquisition phase involving palynologists. In this case $\alpha = \{Poaceae\}$, $\mathcal{S}(\alpha) = \{PollenWithPori\}$, $\mathcal{A}_\alpha = \{geometry, size, hue, brightness, pattern, contrast\}$. Attribute value *hue* is defined as $\mathcal{V}(hue) = \{Pink\}$.

4. Visual Concept Learning Phase

The role of visual concept learning is to learn representative samples of visual concepts used during knowledge acquisition phase. Visual concept learning fills the gap between ontological concepts and image level. An example of a visual concept learning task is to learn how to detect the visual concept $Pink$ in any image.

Visual concept learning consists of training a set of detectors $D = \{d_i\}$ to recognize visual concepts contained in the ontology. This learning is done thanks to a set of training vectors computed during feature extraction on manually segmented and annotated regions of interest. The visual concept ontology is used because the learning process is done in a hierarchical way by using ontological tree structure. Visual concept learning is composed of three steps : training set building, feature selection and training (fig. 2).

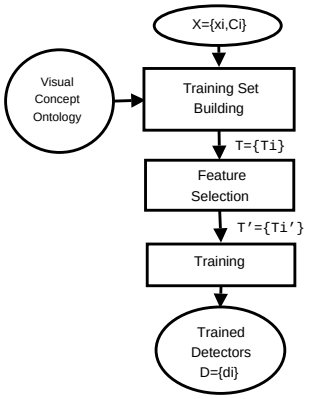

Figure 2. Visual concept learning

Domain Class	POACEAE
SuperClass:	POLLENWITHPORI
SubParts:	
PORI pori1	{PORIWITHANULUS}
SpatialAttributes :	
geometry :	{*CircularSurface, EllipticalSurface*}
size :	{*ImportantSize*}
ColorAttributes :	
hue:	{*Pink* }
brightness:	{*Dark*}
TextureAttributes :	
pattern:	{*GranulatedTexture*}
contrast:	{*Slight*}

Table 1. High level description of domain class Poaceae. Attributes names are in bold face. Attribute values are in italic. Expert terminology is in small caps.

The proposed architecture is designed to learn the set of visual concepts used during knowledge acquisition. A training set T_i is associated with each visual concept $C_i \in \Theta$. A training set is a set of N labeled vectors $\mathbf{x_i} \in \mathbf{R}^n$. Vectors are labeled by $y_i \in \{-1, 1\}$. $y_i = 1$ means that $\mathbf{x_i}$ is a representative sample of C_i. $y_i = -1$ means that $\mathbf{x_i}$ is a negative sample of C_i. We define d_i as an estimation of the probability distribution $p(C_i|\mathbf{x})$. Each d_i is built thanks to the training vectors $\mathbf{x_i}$ labeled by y_i such that $y_i = 1$ if $\mathbf{x_i}$ is a representative of C_i. A training vectors $\mathbf{x_j}$ is labeled by $y_j = -1$ if $\mathbf{x_j}$ is a negative sample of C_j.

Training Set Building uses the set of training vectors $X = \{\mathbf{x_i}, \mathbf{C_i}\}$. Let us consider an image sample $\gamma \in \Gamma$ so that $\mathbf{x}$ has been computed on γ. $\mathbf{x}$ is labeled by C_i if $\exists a \in \mathcal{A}_\alpha \mid (C_i \in \mathcal{V}(a) \wedge \gamma \in \mathcal{L}_\Phi(\alpha))$. For instance, if a region of interest is a representative sample of a domain class in which attribute *pattern* has for value *GranulatedTexture*, then the feature vector $\mathbf{x}$ computed on this region of interest by applying a gabor filter will be labeled by the visual concept *GranulatedTexture*. Training set building consists of obtaining $T = \{T_i\}$ from X so that:

$$\begin{cases} P_i = \bigcup_j \{(\mathbf{x_j}, +1) \mid C_j \preceq_\Theta C_i\} \\ N_i = \bigcup_j \{(\mathbf{x_j}, -1) \mid C_j \preceq_\Theta \\ \qquad\qquad (C_k \in brothers(C_i)) \wedge (\mathbf{x_j}, +1) \notin P_i\} \\ T_i = P_i \cup N_i \end{cases}$$

P_i is the set of representative training vectors of a visual concept C_i. N_i is the set of training vectors computed on negative samples of a visual concept C_i. The training set associated with C_i is noted T_i. Training set building result is T_i (feature vectors labeled by $+1$ or -1) for each C_i. The hierarchical structure of the ontology is used to compute each T_i.

Feature selection chooses the most characterizing features for better visual concept detection. We currently use a Sequential Forward Floating Selection (SFFS) algorithm [8] to perform feature selection. Feature selection consists of computing $T' = \{T'_i\}$ so that $T'_i = SFFS(T_i)$. This method iteratively adds or removes features until some termination criterion is met. Bhattacharyya distance between classes is used as a separability criterion.

Training consists of computing $D = \{d_i\}$ so that d_i is an estimation of the probability distribution of $p(C_i|\mathbf{x})$. Estimation uses T'_i to build each d_i. We currently use multi layer perceptrons to perform estimation. The next section shows how $D = \{d_i\}$ is used to perform object categorization. The combination of the domain knowledge base (fig. 1) and visual concept detectors is called an augmented knowledge base.

5. Object Categorization Phase

Figure 3 presents an overview of the categorization phase. Figure 4 describes the recognition algorithm in a simplified way.

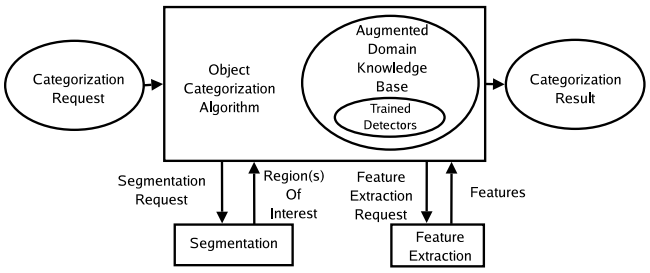

Figure 3. Object categorization phase overview

This algorithm is divided into four steps. It matches an unknown object to categorize with one or several classes of the domain.

(1) The categorization process is initiated by a categorization request which contains an image of the object to categorize.

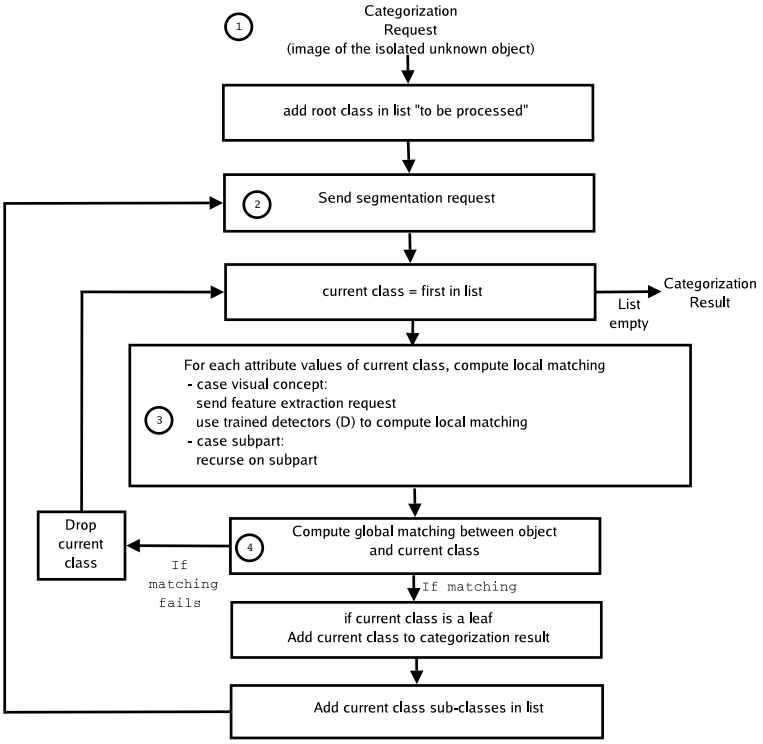

Figure 4. Simplified Version of object categorization algorithm

(2) The object of interest has to be segmented from background. To achieve object extraction, we use a region growing segmentation algorithm. Initial seeds are placed at the corners of the image. If the algorithm tries to classify a subpart, the segmentation task consists of extracting the subpart from the main object. In both cases, a segmentation request has to be sent to initiate segmentation.

(3) Then comes local matching between current class attribute values (e.g. *CircularSurface* for attribute *geometry*) and visual concepts recognized by the detectors trained during the learning process. Local matching function $m_{\mathcal{A}}$ is defined as :

$$\begin{cases} m_{\mathcal{A}} : \mathcal{A} \rightarrow [0,1] \\ \forall \mathbf{x} \in \mathbf{R}^n, \forall \alpha \in \Phi, \forall a \in \mathcal{A}_\alpha, \forall C_i \in \mathcal{V}(a) \\ m_{\mathcal{A}}(a) = max\{d_i(\mathbf{x})\} \end{cases}$$

Feature vector $\mathbf{x}$ used to compute local matching is the result of feature extraction. The result of local matching is a set of probabilities associated with each attribute value. For the subpart attribute, a recursive call has to be made so as to compute its global matching value.

(4) Global matching consists of evaluating if current class matches the object to be recognized. This matching is done by combining probabilities computed during local matching. Global matching function m_Φ is defined as :

$$\begin{cases} m_\Phi : \Phi \rightarrow [0,1] \\ \forall \alpha \in \Phi, \forall a \in \mathcal{A}_\alpha, \forall \beta \in \mathcal{S}(\alpha) \\ m_\Phi(\alpha) = \sum_a m_{\mathcal{A}}(a)/Card(\mathcal{A}_\alpha) + \\ \qquad\qquad \sum_\beta m_\Phi(\beta)/Card(\mathcal{S}(\alpha)) \end{cases}$$

If m_Φ is greater than a predefined threshold $th_{compatibility} \in [0,1]$ then matching between current class and unknown object is validated. If object matches current class, the classification algorithm tries to go deeper in the domain class hierarchy defined by the partial order $\preceq_\Phi$. If matching fails, current class is dropped.

6. Application to Image Indexing and Retrieval

This section shows how the methology presented in this paper is used to perform image linguistic indexing. For a given image, indexing is done by using the class name of the main object contained in this image. This approach has to be opposed to feature-based indexing. Once the indexing process is finished, retrieval is straightforward.

As shown in fig. 5, we define the set of domain classes as $\Phi = \{OutdoorScene, AerialScene, MaritimeScene, Aircraft, Ship, Sea, Sky, Landscape\}$. We also define $\mathcal{S}(AerialScene) = \{Aircraft, Landscape, Sky\}$ and $\mathcal{S}(MaritimeScene) = \{Ship, Sea, Sky\}$. Table 2 shows how domain classes are visually described by visual concepts. Three attributes are used : *geometry*, *hue* and *repartition*. Attribute *repartition* value is defined as $\mathcal{V}(repartition) = \{Uniform\}$ for class Sky.

We have used OntoVis (a knowledge acquisition tool described in [4]) to perform manual segmentation (fig. 6) and annotation of 120 samples images (fig. 7) (60 images for class

Domain Class	**geometry**	**hue**	**repartition**
SKY	-	*Blue* or *White*	*Uniform*
SEA	-	*Blue* or *Green*	*Random*
AIRCRAFT	*Aircraft Shape*	*Grey* or *White* or *Green*	-
SHIP	*Ship Shape*	*Grey* or *White*	-

Table 2. Examples of visual concept description of domain classes.

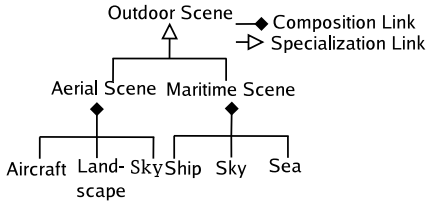

Figure 5. Hierarchical ($\preceq_\Phi$) and composition ($\mathcal{S}(\alpha)$) relations between classes.

Aircraft, 20 images for class Ship, 20 for class Sky and 20 for class Sea). Due to lack of expertise, we have described the geometry of aircrafts and ships respectively as Aircraft Shape and Ship Shape. This is an example of a specialization of the visual concept *PolygonalSurface* provided by the visual concept ontology. This shows that our approach is still valid even if little knowledge is provided.

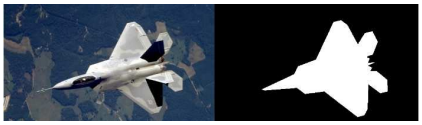

Figure 6. A sample image and its associated manual segmentation.

Figure 7. Typical Images of interest : aircrafts and ships in their environment.

During the learning phase, 120 annotated and manually segmented samples obtained during knowledge acquisition are used to compute the set of visual concept detectors (D). Multi layer perceptrons are used to perform visual concept detection. The input layer of a given perceptron associated with a visual concept depends on the number of features defining this visual concept. For instance, *hue* attribute values are characterized by histograms quantified on 255 levels. This implies that the perceptron associated with concepts used as hue attribute values has an input layer of size 255. The intermediate layer of each perceptron contains 20 neurons. The output layer of a perceptron trained to recognize a visual concept C_i is of size one. Once the learning phase is over, the object categorization is able to use the resulting augmented knowledge base to perform categorization.

Our object categorization algorithm is used to perform image retrieval in both video streams and still image databases. Frames are acquired from video streams one by one. The algorithm is entirely programmed in C++. This categorization method has been tested on an image database structured as following: French TV news (7000 frames); Aircraft and Ship images found on Google^{TM}image search engine (http://images.google.com) (169 frames) and images taken with a personal camera (1000 frames). A categorization example is given in fig. 8. The output of the categorization process is a symbolic explanation of the result. Categorization time (on a P4 3.06Ghz with 1.5Gb of RAM) for a 360x288 image is of 500 ms. Figure 9 shows the precision/recall tradeoff. Precision is defined as the ratio between the number of relevant retrieved images and the number of retrieved images. Recall is defined as the ratio between the number of relevant retrieved images and the number of relevant images in the image database. This curve has been obtained by a variation (step=0.01) of global matching threshold $th_{compatibility}$ (see section 5). It can be seen that for a precision rate of 73%, a recall rate of 19% is obtained. This means that 73% of retrieved images are relevant and that 19% of the relevant images contained in the image database have been retrieved. For our end-users, a good precision is more important than a good recall : the system satisfies their needs. From our point of view, results are promising. The results depend on the quality of automatic segmentation : when segmentation of the object is bad, categorization often fails. Segmentation process improvement should increase system performances.

7. Conclusion

This paper presents an original approach to complex object recognition. This approach takes advantage of explicit aspects of knowledge based approaches. Moreover, machine learning techniques allow to reduce the knowledge acquisi-

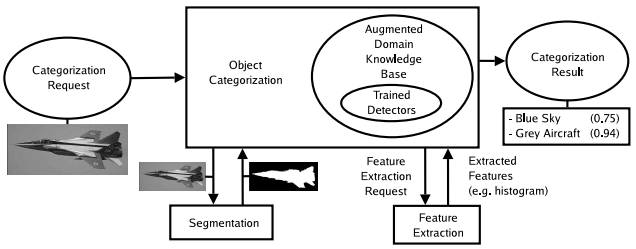

Figure 8. Categorization result is a symbolic explanation of the input image.

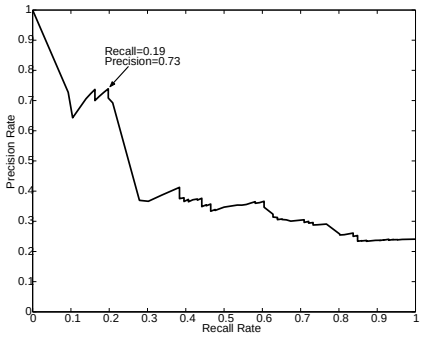

Figure 9. Plot of relative percentage for precision vs. recall.

tion effort. This approach is structured in three main phases. A knowledge acquisition phase which consists of describing a set of domain classes with visual concepts provided by a visual concept ontology. This ontology is composed of the following types of visual concepts : spatial concepts and relations, color concepts and texture concepts. The result is a domain knowledge base. An object learning phase follows the knowledge acquisition process in order to obtain a knowledge base augmented by a set of detectors trained to the recognition of the visual concepts used for the description of each class. The categorization phase tries to match an unknown object with one or several domain classes. The matching is done between visual concepts computed on the unknown object and visual concepts used for the description of domain classes. The proposed approach allows semantic and explicit object categorization. The global architecture does not act as a black box and is able to explain categorization results, unlike categorization techniques which consider the image at the appearance level and not at the object level. One strong point is the modularity of the approach. New algorithms can be easily integrated in order to obtain a better segmentation or a better feature extraction. Changes at the low-level part have no consequence on the high-level part. An important contribution to past knowledge based vision systems is that learning techniques simplify knowledge acquisition : the expert provides well shared knowledge (i.e domain knowledge) and not image processing knowledge. In systems like [6] or [3], symbol grounding was performed by inference rules which were difficult to define. In our approach, a learning phase achieves this task.

We have applied the proposed approach to image indexing which enables straightforward retrieval. Promising results have been obtained. Since querying is based on expert knowledge, no query image is needed. As explained in [10], such a knowledge oriented approach allows one to gain access to a meaningful conceptual level that is difficult to reach with classic query-by-example paradigm. Indeed, in the query-by-example paradigm, it is difficult for the system to determine which semantical aspects make a given image relevant.

There are several remaining issues. At the segmentation level, a major remaining challenge is to define precisely the feedback to the segmentation level when object categorization fails. A good object segmentation as well as good subpart segmentations are needed in this approach. For some specific applications, this hypothesis is reasonable. In general, segmentation remains a major issue. We plan to use visual concepts, program supervision [9] and learning techniques to deal with this problem.

References

[1] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2002.

[2] C. Hudelot and M. Thonnat. A cognitive vision platform for automatic recognition of natural complex objects. In *ICTAI*, Sacramento, USA, November 2003.

[3] S. Liu, M. Thonnat, and M. Berthod. Automatic classification of planktonic foraminifera by a knowledge-based system. In *The Tenth Conference on Artificial Intelligence for Applications*, pages 358–364, San Antonio, Texas, March 1994. IEEE Computer Society Press.

[4] N. Maillot, M. Thonnat, and A. Boucher. Towards ontology based cognitive vision. In *ICVS, Graz, Austria*, volume 2626 of *LNCS*. Springer, 2003.

[5] K. Mardia. Shape in images. In S. Pal and A. Pal, editors, *Pattern Recognition – From Classical to Modern Approaches*, pages 147–167. World Scientific, 2002.

[6] T. Matsuyama and V.-S. Hwang. *SIGMA - A Knowledge-Based Aerial Image Understanding System*. Plenum Press New York USA, 1990.

[7] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *ACM Multimedia*, pages 65–73, 1996.

[8] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature-selection. *Pattern Recognition Letters*, 15(11):1119–1125, November 1994.

[9] C. Shekhar, S. Moisan, R. Vincent, P. Burlina, and R. Chellappa. Knowledge-based control of vision systems. *Image and Vision Computing*, 17:667–683, 1998.

[10] C. Town and D. Sinclair. Language-based querying of image collections on the basis of an extensible ontology. *IVC*, 22(3):251–267, March 2004.