Electronic Letters on Computer Vision and Image Analysis 7(2):35-61, 2008

Representations for Cognitive Vision: A Review of Appearance-Based, Spatio-Temporal, and Graph-Based Approaches

Axel Pinz[a], Horst Bischof[b], Walter Kropatsch[c], Gerald Schweighofer[a],
Yll Haxhimusa[d], Andreas Opelt[a], Adrian Ion[c]

[a] *EMT - Inst. of El. Measurement and Measurement Signal Processing, TU Graz, Austria*
[b] *ICG - Institute of Computer Graphics and Vision, TU Graz, Austria*
[c] *PRIP - Pattern Recognition and Image Processing Group, TU Vienna, Austria*
[d] *Department of Psychological Sciences, Purdue University, USA*

Received 17th October 2007; revised 20th March 2008; accepted 22nd June 2008

Abstract

The emerging discipline of cognitive vision requires a proper representation of visual information including spatial and temporal relationships, scenes, events, semantics and context. This review article summarizes existing representational schemes in computer vision which might be useful for cognitive vision, and discusses promising future research directions. The various approaches are categorized according to appearance-based, spatio-temporal, and graph-based representations for cognitive vision. While the representation of objects has been covered extensively in computer vision research, both from a reconstruction as well as from a recognition point of view, cognitive vision will also require new ideas how to represent scenes. We introduce new concepts for scene representations and discuss how these might be efficiently implemented in future cognitive vision systems.

1 Introduction

Cognitive vision brings together such diverse fields of research as digital image analysis, computer vision and cognitive sciences. The *Research Roadmap of Cognitive Vision* [160] presents this emerging discipline as ‘a point on a spectrum of theories, models, and techniques with computer vision on one end and cognitive systems at the other’. Potential definitions of a cognitive vision system range from ‘visually enabled cognitive system’ to ‘cognitively enabled vision system’. Typical results that we would expect from a cognitive vision system are for instance to be able to correctly answer queries regarding the relative position of occluded objects or to recognize previously unseen objects of a learned category (‘object categorization’ [136]).

From a computer scientists point of view, several major issues have to be solved in engineering a cognitive vision system: embodiment, learning, recognition, and reasoning. At the basis of all these efforts, we require a proper *representation* of visual information, spatial and temporal relationships, scenes, events, semantics, and context. This paper reviews existing approaches to representations that should suit some of the requirements of a cognitive vision system and outlines promising research directions. Special emphasis is put on appearance-based, spatio-temporal, and graph-based representations, including a comparison of these rather diverse approaches.

Correspondence to: <axel.pinz@tugraz.at>

Recommended for acceptance by João Manuel R. S. Tavares and Renato Natal Jorge
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

1.1 From Computer Vision to 'Cognitive Vision'

The first complete theory of a *computational* approach to vision has been presented by the late David Marr [106]. This seminal work has not only significantly influenced a decade of Computer Vision research, but is still used by researchers from related fields as the reference framework in setting up new theories (see e.g. Palmer [133]). Marr distinguished between a *reconstruction* and a *recognition* approach, which has been further detailed by Aloimonos and Shulman [5]. He also initiated work on representations when he introduced what he called a 'primal sketch' (local 2D saliency), a '2-1/2–D sketch' (visible surface depth and orientation), and a '3D object model' (volumetric object representation, typically Marr's 'generalized cones'). Primal sketch and 2-1/2–D sketch are 'viewer centered' in image coordinates, and the 3D model is 'object centered' in a specific object coordinate system. This allows to build a scene model (in scene coordinates), which is composed of individual objects, and their poses (position and orientation) and scales. The idea also supports nicely the decomposition of an object into (volumetric) parts, and Marr's book has triggered a 'Recognition by Components' (RBC) school, which has been advocated by Biederman [19] from a cognitive psychology viewpoint, and was supported by Dickinson and others in the computer vision community (e.g. [42]).

Some computer vision researchers have reviewed Marr's theory from a more critical point of view. One interesting aspect is provided by Medioni et al. [111]. They concentrate on the limitations of a 2-1/2–D sketch, which may even complicate the problem of reconstruction when many different views are used. Instead, they propose to use *layers*, a layered representation of visible curves and surfaces, and they present tensor voting as the appropriate computational framework. Other researchers, including Ullman [157] and Edelman [45], advocate view-based approaches, which avoid computationally expensive and sometimes ill-posed 3D reconstruction. View-based recognition also has strong support from cognitive scientists (e.g. [153]) and biologists [150].

Computer Vision has seen a rapid and fruitful development of 3D reconstruction from multiple images and image sequences (stereo, structure from motion), and also of 2-1/2–D (shape from X), with an especially concise treatment of algebraic projective geometry (see [48, 67, 105], but also [148]). While it seems now possible, to reconstruct a 3D scene in terms of visual features and their positions in scene coordinates, the automated assembly of 3D object models has turned out to be more difficult. RBC, especially geon-based recognition suffered from the problem of insufficient low-level image analysis - while higher level algorithms worked nicely, the necessary segmentation had to be circumvented by line-drawings [41]. Some success was reported for very narrowly limited cases, for instance, by Zerroug and Nevatia [169] for a few special types of generalized cones under orthographic projection.

On the other hand, appearance-based object recognition (without requiring 3D reconstruction, and working purely in the 2D image domain) has been very successful, and still has not reached its limitations, over the past 10 years (e.g. early work by Murase and Nayar [120] or robust PCA by Leonardis and Bischof [93]). Recently, Nistér and Stewénius [124] presented a vocabulary tree, and claim that they can recognize a specific object out of 110 million candidates in less than 6 seconds. While global PCA and related subspace approaches (e.g. LDA, ICA, etc.) work on the 2D images themselves, i.e. on well defined pixel arrays (including certain size and brightness normalization), recent developments are relating back to Marr's primal sketch and try to reduce the complexity of the problem by looking only at salient points. The theoretical foundation for this work is scale space theory [98] and some of the saliency detectors are invariant to scale and/or affine distortions [78, 116, 101, 109]. They have been successfully used to represent, detect, and recognize individual objects and even object categories by a collection of object specific local features (e.g. [163, 50, 127, 1, 53, 91]).

Perceptual grouping approaches may be considered somewhere in the middle between purely appearance-based and 3D reconstructionist approaches. Starting from basic primitives (points, lines, curves), these approaches work towards grouping these primitives into higher-level entities (closed contours, surfaces, volumes). Seminal work in this area has been contributed by Lowe [103], Dickinson [42], and Sarkar [142]. Grouping may be either data- (bottom-up) or model-driven (top-down). Recently good results for object recognition were achieved with object representations using curves as primitives (e.g. [17, 87, 51, 149, 129]).

Up to this point, we did not explicitly mention time in our discussion. Early work in computer vision was mainly based on the interpretation of individual images or on stereo pairs, which were captured at a certain instance in time. A reconstructionist approach for dealing with time will include a history, object trajectories, and prediction of future motion. The straightforward way is to extend 3D reconstruction techniques towards 4D (3D space + time), which has partly been covered for static scenes (see e.g. work by Pollefeys, Nistér and others [119, 123, 2]). There are many applications of tracking in videos which are recorded by a stationary camera and deal with the 3D space-time domain (2D image coordinates + time, for instance person tracking and traffic monitoring). A recognition point of view is to extend scale-space theory towards scale in space and time, and to detect salient space-time events. This formal extension towards space-time scale space has been presented by Laptev and Lindeberg [88].

A suitable representation for cognitive vision has to bridge the representational gap between raw images and high level interpretations of scenes [81, 82]. Computer Vision over the past 30 years has followed a path from generic (prototypical) models (generalized cylinders, superquadrics and geons) to individual (exemplar-based) models (starting with 3D CAD based models, appearance based models, generative probabilistic models). There is evidence that some mechanisms in human visual cognition are view-based and do not require the reconstruction of a 3D object model, or a 3D scene model [153]. However, to a certain extent, when we have to reason with objects, their motion, and their relations to each other in space and time, we will require to explicitly represent these entities. Some cognitive psychologists also advocate that relational structural representations are required in object category recognition (see e.g. [71]). A survey on human representation of visual perception of objects can be found in [24].

1.2 Terminology

A common terminology is required when several fields are related to each other under a new perspective*. We look at the main concepts involved in representation (and reasoning), discuss the related terms, and indicate their dimensionality in space and time (boldface numbers). We deal with *images* (dimensionality **2:** 2D space), *image sequences* (**3:** 2D space + time), *scenes* (**3:** 3D space), and *objects* (**3:** 3D or **2:** 2D projections) Objects can be represented by one or more *view(s)* (**2:** 2D space), or by a *3D model* (**3:** 3D space), and their motion can be described by *trajectories* (**3:** 2D space + time, or **4:** 3D space + time). Depending on the type of the objects and their motion, these trajectories can be simpler (e.g. rigid objects) or more complex (e.g. articulated motion of a human). Furthermore, we might want to represent *events*, which are characterized by spatial extent (2D or 3D), and occur at a certain instance in time. Events have a *duration* (time interval).

It turns out, that one case is not well grounded in this terminology: while an image sequence (video **3:** 2D space + time) is a common term, we do not have anything comparable for a sequence of scenes. How should the successive states of a scene be called? In Computer Graphics, there is the term of an *animation sequence*. In this paper we will use the term *scene sequence* for a 4D development of a scene (**4:** 3D space + time). A scene sequence is certainly more than an object trajectory. A scene sequence could be represented by a number of independently moving objects (thus resembling a number of simultaneous, related or unrelated trajectories), or, depending on the scene representation, it might be represented by a sequence of occupancy grids or graphs (as proposed by [74]).

Further terms that will be used include *history* (knowledge about the past states of an object/scene), *prediction*, *topology* (adjacency, containment and decomposition/part relations), and *behavior* of an object, parts of an object, or a group of objects (e.g. hiding, seeking, a certain motion pattern, etc.).

*Cognitive vision is related to such diverse research areas as computer vision, cognitive psychology, computer graphics, visualization, human-computer interaction, and augmented reality.

1.3 Outline of the paper

The previous introductory sections already showed that the topic of representation in vision is quite broad so that it deserves to be reviewed from various viewpoints. We start with brief and rather general aspects on visual abstraction and representational levels (section 2), and focus then on object representations in section 3. The extension from objects to scene representations is covered by section 4, and we conclude with a dedicated section on promising future research directions (section 5).

2 Visual Abstraction and Representational Levels

Recognition, manipulation and *representation* of visual objects can be simplified significantly by "abstraction". By definition abstraction extracts essential features and properties while it neglects unnecessary details. Two types of unnecessary details can be distinguished: redundancies and data of minor importance.

Details may not be necessary in different contexts and under different objectives which reflect different types of abstraction. In general, four different types of abstraction are distinguished [86]:

isolating abstraction: important aspects of one or more objects are extracted from their original context.

generalizing abstraction: typical properties of a collection of objects are emphasized and summarized.

idealizing abstraction: data are classified into a (finite) set of ideal models, with parameters approximating the data and with (symbolic) names/notions determining their semantic meaning.

discriminative abstraction: only aspects discriminating one object from the other are considered.

These four types of abstraction have strong associations with well known tasks in computer vision: recognition and object detection tries to *isolate* the object from the background; perceptual grouping needs a high degree of *generalization*; and classification assigns data to *"ideal"* classes, *discriminating* between them, disregarding noise and measurement inaccuracies. Such generalization allows to treat all the elements of a general class in the same way. When applied successively, the four types of abstraction imply a hierarchical structure with different levels

- of concepts for representing knowledge about the world, e.g. the conceptual hierarchy in [8],
- of representation,
- of processing stages, e.g. hierarchies of invariance in cognition [10], and
- in the complexity of processing images.

In all cases abstraction drops certain data items which are considered less relevant. Hence the *importance* of the data needs to be computed to decide which items to drop during abstraction. The importance or the relevance of an entity of a (discrete) description must be evaluated with respect to the purpose or the goal of processing. The system may also change its focus according to changing goals after knowing certain facts about the actual environment, other aspects that were not relevant at the first glance may gain importance. Representational schemes must be flexible enough to accommodate such attentional shifts in the objectives. With respect to cognitive vision, abstraction can help in obtaining compact but still very descriptive representations. It is one of the known ways to connect low level data with high level processes such as high level reasoning (Where is the cup?), and is needed to comunicate with humans in their natural language.

Multiple abstraction levels have been identified, spanning from the low, image-based (pixels) to the high, object, model, and topology based. Table 1 shows the main abstraction categories and some of their properties.

To build a scene representation one needs at the basis techniques for the semantic interpretation of images, in particular to localize and name objects contained in a scene and to assess their mutual relationships. A general

Table 1: Abstraction Levels

	addressing and axis	entities	neighborhood
image-based	2D (row, column)	pixel	4,8-neighborhood
appearance (view) -based	$n \times m$-D subspaces	points in subspace	distance to subspace
part-based	part-whole relation	properties, parts	semantics
object/model-based	name, location	sub-objects/models	part/whole
scene-based	(x,y,z,t,...)	objects	spatio-temporal semantics
topology-based	relational paths	topology domain	explicitly encoded

topic within this interpretation is the question of the representation levels. Different aims (tasks) and different scenes might need more or less detail.

Some of the pictorial entities, their information content, and the operations that can be performed at different processing levels are summarized in Table 2.

Table 2: Pictorial entities at different levels of processing

Entity	information content	examples for operations
Picture	imaging conditions, geometry	sampling, rectification
Pixel	gray value / color vector	enhancement, classification
Neighborhood	spatial locality	shrink, expand
(Step) edge	magnitude, orientation	edge detection and linking
Region	homogeneity, connectivity	segmentation
Boundary	shape	connecting continuous curve segments
Image Part	specific image properties	property measurement
Object Part	specific object properties	property matching
Object	functionality	relational matching
Situation	specific configuration of objects	interpretation
Scene	visible situations of the world	description

In [155], the *connection table* allows the transition between the different levels of abstraction. Five different representation levels are identified from the real to the cognitive world: 2D image (image-based), 3D skeleton (feature-based), connection table (part-based), object description language (model-based), natural language (language-based) (see Table 1). Basic descriptive notions are: objects - parts - primitive parts. The *connection table* describes the way in which parts form an object.

While visualization generates an image from a computer stored description, digital image analysis is supposed to produce descriptions of a digital image. Still, both fields have at the basis descriptions at different levels of abstraction. The following levels are identified:

1. 2D digital image with pixels;

2. image segments such as region, edge, or texton [77];

3. image segments with specific properties such as generalized cylinders;

4. fragments, parts of objects, 'GEON' [20];

5. objects, models;
6. functional areas [110];
7. natural language like in [155].

3 Object Representations

One of the most important decisions that have to be taken when designing a vision system is how objects and their properties are represented. This determines the classes of features that could be used, how they are grouped, and how they are matched. Current object representations, depending mostly on the task, span from prototypical (high abstraction level, used mainly for generic object recognition) to exemplar-based (low abstraction level, used mainly for recognizing particular instances). We give here a summary of existing object representation frameworks and discuss their advantages and disadvantages. The section is structured into three major subsections on appearance-based, spatio-temporal, and graph-based representations.

3.1 Appearance-Based Object representation

Appearance-based representation of objects has probably been one of the most researched areas in Computer Vision over the past decade. More recently, we have seen a distinction between global appearance-based methods (such as PCA), and local (e.g. saliency) detectors and descriptors.

3.1.1 Global Subspace Methods

The basic idea underlying subspace methods for visual learning and recognition is that an image can be represented as a point in a high-dimensional space (the space spanned by its pixels), a change of the object in the image (e.g., object rotation) has not an arbitrary effect on the point in the high-dimensional space. Therefore, an object (or even an object class, e.g. faces) can be characterized by the set points (the subspace) they occupy in the high-dimensional space. Since this subspace is usually of much lower dimensionality than the original space a considerable amount of compression can be achieved by characterizing this low dimensional subspace. The difficulty is to find a compact representation of this usually highly non-linear space. A common approach is to choose a linear approximation:

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

where, $\mathbf{x} \in \mathbb{R}^n$ is the original image, $\mathbf{y} \in \mathbb{R}^m$ its low dimensional subspace representation and $\mathbf{A} \in \mathbb{R}^{n \times m}$ the linear subspace. Depending on the required properties of the subspace we can obtain different (linear) subspace representations. Among the most commonly used representations are:

Principal Component Analysis (PCA): The most commonly used technique for compression of training images is based on *principal component analysis* (PCA) [70]. PCA requires that the reconstruction error (the error obtained when reconstruction $\mathbf{x}$ from its low dimensional representation $\mathbf{y}$) over all training images is minimal. To achieve this goal, the directions with the largest variance of input data are found in the high-dimensional input space. The dimension of the space can be reduced by discarding the directions with small variance of the input data. By projecting the input data into this subspace, which has the principal directions for the basis vectors, we obtain an approximation with an error, which is minimal (in the least squares sense) among all linear transformations to a subspace of the same dimension. It turns out that the correlation between two images can be approximated by the distance between their projections in the principal subspace. Thus, the recognition can be carried out by projecting an image of an unknown object into the principal subspace and finding the nearest projected training image [120].

Independent Component Analysis (ICA): [72, 7] is a powerful technique from signal processing known also as *blind source separation*. Contrary to PCA it does not only find uncorrelated components, but it delivers a linear transformation $\mathbf{A}$ such that the projections are as statistically independent as possible. This can be seen as an extension of PCA, where the projections of the input data into the subspace are not only uncorrelated but also independent. Independent representations lead to sparse codes which is considered as one goal of sensory coding in the brain (cf. [11]).

Linear Discriminant Analysis (LDA): PCA and ICA are *unsupervised* methods, which means that no additional information about the training images is necessary to build the representation. If, for instance, PCA is used for classification, no information on classes is used, thus the discriminant information might be lost. In this case, rather than maximizing the variance of all projections, one would prefer to maximize the distance between the projected class means, which increases the discriminant power of the transformation. This is the goal of *linear discriminant analysis* (LDA) [107]. Furthermore, next to maximizing the distance between the classes, *Fisher's linear discriminant* [15] minimizes the distances within classes by minimizing within-class variance of the projections. It has been a popular tool in the field of pattern recognition, where it is frequently used to reduce the dimensionality of the input signal to alleviate the subsequent classification step.

Canonical Correlation Analysis (CCA): If the task is regression (not classification), *canonical correlation analysis* (CCA) [112] is the method of choice. It relates two sets of observations by determining pairs of directions (canonical factors) that yield maximum correlation between the projections of these sets. Thus, it is suitable, for example, for estimation of orientation, where one set of observations consists of observed images, while the observations in the second set are object orientations from which the corresponding images were acquired.

Non-negative Matrix Factorization (NMF): Another subspace technique is *non-negative matrix factorization* (NMF) [90]. It is similar to PCA (finds the representation with the minimal error) with the constraint that the factors consist of non-negative elements only. Due to this non-negativity constraint it tends to decompose the input images into parts (e.g., learn from a set of faces the parts a face consists of, i.e., eyes, nose, mouth, etc.), leading to a part based representation.

Kernel methods: As explained above the subspace of images is usually not linear, and all the methods discussed so far provide only a linear approximation. They can, however, be extended to *nonlinear* feature extractors [34, 145, 113, 112]. This can be done by first mapping input vectors using a nonlinear mapping into a high-dimensional feature space and then performing a linear method on the obtained high-dimensional points. This procedure is equal to the employment of a non-linear method in the original space. To avoid computing a nonlinear mapping into a space of a very high (possibly infinite) dimension, the so called *kernel trick* can be applied [34]. This method was originally proposed in the context of Support Vector Machines (SVM) [159]. It can be applied whenever it is possible to formulate the algorithm in such a way that it uses only dot products of the transformed input data. The dot products in feature space are then expressed in terms of kernel functions in input space, thus all operations can be performed in the original lower-dimensional input space.

The major advantage of the subspace approach is that both learning as well as recognition are performed using just two-dimensional brightness images without any low- or mid-level processing. However, due to the direct use of two-dimensional images there are various problems associated with the direct application of the methods, in particular, robustness against occlusion, scaling, varying background, illumination changes etc. Recently some new methods that can cope with these problems have been proposed (e.g., [93] has demonstrated how to handle occlusion, varying background and other kinds of non-Gaussian noise in PCA, [23] has demonstrated how to handle severe illumination variations). Also the problem of learning these representations in a robust manner has been addressed recently [151].

Besides using these methods on whole images they can also be applied locally to image patches [126] or local descriptors, a recent example is the so called PCA-SIFT [80] descriptor.

Another characterization of the subspace approaches is to distinguish, whether a generative model (i.e., the ability of reconstruction and generation of samples), or a discriminative model is employed [122]. Each of them offers distinctive advantages. Generative models such as PCA, ICA, etc. enable robustness in model construction due to their ability to reconstruct the input from partial data Their representation is also not task dependent and can be used for different purposes. On the negative side, generative representations are usually wasteful in the resources and do not scale well. On the other hand discriminative methods such as LDA, SVM, etc. achieve in general higher recognition rates. Their representation is tailored for the specific task and they are usually faster. On the negative side, discriminative reresentations do not enable reconstruction, therefore, robust methods cannot be easily used. Furthermore, the representation is less flexible and cannot be adapted to new tasks. Since these two representations have quite complementary properties it makes sense to combine them, and recently people have started to work on such combinations, e.g., [55].

3.1.2 Local Detectors and Descriptors

As there are many problems with global representations (sub-space methods, aspect graphs) of an object (e.g. outliers, occlusion, varying background) recent research focuses on a local description of the object. The basic idea is to first extract distinguished regions in an image (such that the regions can be re-detected with a high probability), then describe the region and/or its local neighborhood with a possibly invariant photometric descriptor and use the descriptor for matching with new images. The advantage of these approaches is that they do not require a segmentation and can deal with occlusions and clutter. Using photometric descriptors the approaches are discriminative (see [117] for a comprehensive review and evaluation of different approaches). There is a wide variety of different distinguished region detectors.

Simple detectors: A large class of detectors is based on measures of cornerness, among those the well known Harris corner detector [66]. The idea is reformulated using the structure tensor [21] and the second moment matrix respectively, leading to different variants of corner detectors [56, 154, 140, 84]. Other approaches use the second derivatives (Hessian matrix) instead of the first derivatives. All these approaches can be considered belonging to one class of simple interest point detectors. They all detect only a location. Therefore, for a subsequent task like image matching via cross-correlation the size and orientation of the necessary matching window has to be chosen independently. This is a severe limitation when dealing with differently scaled or affinely transformed regions.

Scale and Affine invariant detectors: This limitation was addressed by estimating a proper scale for every detected interest point. The first work going into this direction was presented by Tony Lindeberg [99] in 1998. Other approaches followed shortly by David Lowe [104] or Krystian Mikolajczyk [115]. This class of interest operators is usually called scale-invariant interest operators.

However, research again went one step further. According to the success of interest operators which are invariant to scale changes, methods were sought to create interest operators invariant to a larger class of image transformations. This was driven mostly by developments in wide baseline image matching were significant perspective distortions occur. Research therein led to a new class of interest detectors, affine-invariant detectors. In most cases such a detection consists of a point location and an elliptical delineation of the detection. The ellipse representation captures the affine transformation of the detection. By normalizing the ellipse to an unit-circle the affine transformation can be removed. This method was first suggested in 2000 by Baumberg et al. [12]. This has lead to a wide variety of affine-invariant detectors [116, 109, 79, 156]. The common property of these approaches is that they provide information how the region around the detection can be normalized to allow image matching. The detections themselves, however, may not be simple point locations anymore. In the case of the MSER detector [109] a detection is a whole image region showing similar gray-values. Approaches like that are usually referred to as interest region detectors, moreover as every affine detector defines its own support region too.

Besides the detection of local regions we need also a description of local photometric content of the regions, which is then in turn used for matching. The descriptors can be roughly divided into three classes:

Distribution-based: Distribution based methods represent certain region properties by (sometimes multi-dimensional) histograms. Very often geometric properties (e.g. location, distance) of interest points in the region (corners, edgels) and local orientation information (gradients) are used. In this class falls the well-known SIFT descriptor [102] that uses histograms of local edge orientations. Ke and Sukthankar [80] modified the SIFT-key approach by reducing the dimensionality of the descriptor by applying principal component analysis to the scale-normalized patches. A rotation invariant version of SIFT is obtained by the Gradient location-orientation histogram (GLOH) descriptor, which divides the patch into a radial and angular grid [114]. Other well known distribution based descriptors are the spin image [76, 89] and the shape context [16] that uses the distribution of relative point positions and corresponding orientations collected in a histogram as descriptor.

Filter-based: The basic idea of filter-based methods is to use the response of a set of filters as a description of the region. Properties of local derivatives (local jets) are well investigated and can be combined to sets of differential operators in order to obtain rotational invariance. Such a set is called "differential invariant descriptor" [144]. "Complex filters" is an umbrella term used for all filter types with complex valued coefficients. In this context, all filters working in the frequency domain (e.g. Fourier - transformation) are also called complex filters, examples are [13, 143, 31].

Other methods: The simplest method that can be used as a descriptor is to take the gray-value patch as it is and use cross-correlation for matching. To obtain invariance, moments can be used [158].

Local descriptors are also used for object categorization. The idea is to learn local descriptors which are category-specific. Various methods are used to learn the features which are best for classification: Boosting [127], Naive Bayes [166], SVM or PCA on local features ([166]), and others.

A problem with all these approaches is the fact, that the learning algorithms do not know where the objects are in the image, so that they also learn features on the background which are related to the object (e.g. [128]).

We have shown in [146] that such learned classifiers give good classification rates on images which are similar to the images used for learning, but they give poor recognition rates on ground truth data (just the object without any contextual information). We have also shown that object localization based on spatio-temporal reasoning is one method, which can improve the learning procedure to give also good recognition rates on ground truth data.

In recent work we have studied the problem of learning local descriptors from image sequences for specific objects [60] and object categories [132]. In particular, local features are tracked in image sequences leading to local trajectories containing dynamic information. Based on these trajectories the quality and robustness of the local feature can be evaluated (and only those that are stable enter the representation). In addition the most representative local description can be selected based on the information obtained from the trajectory. This approach shows that by using dynamic information compact and distinctive local object descriptions can be obtained.

3.1.3 Models Based on Local Descriptions

The biggest problem in using collections of local features for object categorization is the fact that the features can be located anywhere in the image. We know for instance that there has to be a nose between two eyes to be a face, but the algorithms listed above do not take this information of spatial relation between features into account. Many papers exist in face detection which have a pre-learned representation of the model of the face, which consists of eyes, mouth, nose and also ears [68, 168]. These are only useful in limited cases where we know the model of suited features (namely the parts of the face). Nowadays people try to learn the model also from the training images.

In [163, 50], for extensions see [49, 51], the model is learned as a flexible constellation of features, where the variability within a class is represented by a probability density function. The main problem with this approach is that the images used for learning must look similar, which means that the objects in the images must be roughly aligned, resulting in similar position, orientation and scale of the objects. This approach can only learn 2D models from rather aligned images.

An enhancement to this approach was made by [69] to be translation and scale invariant in both learning and recognition of the objects. Results are only reported for face images, where the learned parts look like eyes, chin and eyebrows.

The constellation model is probably the most prominent out of several similar representations. The constellation model proposes a fully connected graph of all model parts. Crandall et al. [37] presented their k-fan model, where k denotes the number of parts that are fully connected to all other parts in the model. A 1-fan can be regarded a star-shaped model, as was also presented by Fergus et al. [52].

Constellations and similar representations model object shape explicitly by a limited number of salient parts and their spatial constellation. Leibe et al. [91] presented a codebook of local appearance (local salient features and their descriptors) that is used together with an implicit shape model (the location of each salient point is mapped relative to the object center).

3.2 Spatial and Temporal Representations

In comparison to appearance, the methods discussed here can be used to represent object *shape*, 3D scene structure, volumetric object models, and temporal characteristics like typical motion patterns.

3.2.1 Curves, Boundaries, Fragments

Although there has been significantly more recent work on object representation by local, salient patches and their descriptors, 2D object shape can often be efficiently represented by an object's internal and external contour. When shape is a dominant cue (e.g. in distinguishing cows from horses), such models may be better suited than patch-based methods. On the other hand, patch-based representation can emphasize texture (e.g. to distinguish horses from zebras, which is impossible based just on the external contour). It is slightly more difficult to represent boundaries at varying scales, orientations and other spatial transformations, but together with the idea of a codebook with object centroid votes, there is recent success in representing codebooks of contour fragments [149] or boundary fragments [130]. An obvious, promising direction of future research will be to combine patch and boundary representation into a unified model (as we report in [131]). Yet another approach models object categories as a highly connected graph of pairwise relationships between boundary fragments [94].

3.2.2 3D Object Representations

There is a vast amount of literature on shape from X methods that recover 2-1/2-D representations and on the recovery of 3D object models either from images, or from 2-1/2-D. Photogrammetric methods include calibrated stereo and block bundle adjustment methods, while the Computer Vision approach is rather directed towards the recovery of scene structure from uncalibrated video [119, 123], or from potentially very disparate views [109]. The typical object representation that emerges from such approaches is a 3D point cloud of salient points. It is not only necessary that a certain saliency detector responds above threshold, but it is also required that point correspondences between views can be established. One way to obtain high quality point clouds is to texture the objects (when this is possible, e.g. by spraying them with a random pattern), another one to mount them on a rotating table. Tracking of feature points while the object is rotated, or while the camera is moved around the object can substantially ease correspondence search.

There are cases, when 3D point clouds, either in scene coordinates, or in object-centered coordinates, are a sufficient 3D model. For instance, in computer graphics, point based rendering attaches a grey-, color- or texture-value to each point and obtains very realistic rendering results when the point cloud is sufficiently dense. In computer vision, however, the necessary next step is to aggregate metric 3D point clouds into more abstract models. One obvious way is to try to fit parametric models to the 3D data. This can be the fitting of dominant planes, or of higher order parametric surfaces such as superquadrics (e.g. [152], obtained from dense 3D range images). Another idea is to model 3D objects and to try to recover their 2D projections in the images, e.g. by geometric hashing [167].

The above approaches all obtain metric 3D models from metric 3D reconstruction. While this is an important research goal on its own, cognitive vision will probably require other kinds of 3D object representations, which are more qualitative, but at the same time may generalize well to represent object categories as well as individual objects. However, the research landscape in qualitative 3D object representation is far more sparse than for 3D reconstruction.

Marr proposed to recover generalized cones and cylinders from single intensity images. This has been achieved for a limited number of specific types of generalized cylinders, based on clues like curvilinearity, symmetry, and low- and mid-level geometric reasoning [169].

Geons have been introduced based on a cognitive theory [19] and have been recovered from intensity images [42], but many open problems remain [41]. However, geons would be attractive, because they constitute *qualitative* models, which eases their use for generalization and categorization, and a certain amount of success has been reported in using geons in content-based image retrieval [22]. The graphs produced in [42], however, lack a proper representation of spatial configuration. A workaround has been presented in [137], which might also work for representing spatial and temporal relations.

A very different approach is advocated by Medioni, who argues that the Marr paradigm has been very influential. It triggered numerous reconstructionist research in shape from X and in 2-1/2-D. Medioni prefers a more direct, layered representation, which circumvents the necessity of reconstructing 2-1/2-D (tensor voting [111]).

3.3 Graph-Based Representations

Handling "structured geometric objects" is important for many applications related to Geometric Modeling, Computational Geometry, Image Analysis, etc.; one often has to distinguish between different parts of an object, according to properties which are relevant for the application (e.g. mechanical, photometric, geometric properties). For instance for geological modeling, the sub-ground is made of different layers, maybe split by faults, so layers are sets of (maybe not connected) geological blocks. For image analysis, a region is a (structured) set of pixels or voxels, or more generally a (structured) set of lower-level regions. At the lowest level, such an object is a subdivision[†], i.e. a partition of the object into cells of dimensions 0, 1, 2, 3 ... (i.e. vertices, edges, faces, volumes, ...).

The structure, or the topology, of the object is related to the decomposition of the object into sub-objects, and to the relations between these sub-objects: basically, topological information is related to the cells and their adjacency or incidence relations. Other information (embedding information) is associated to these sub-objects, and describes for instance their shapes (e.g. a point, a curve, a part of a surface, is associated with each vertex, each edge, each face), their textures or colors, or other information depending on the application.

[†]For instance, a Voronoi diagram in the plane defines a subdivision of the plane

3.3.1 Aspect Graphs

The use of Aspect Graphs (see e.g. [57], chapter 20) as an object representation is a generalized method, representing the view space. The main idea is to combine different viewing directions, where the object looks alike, to one aspect. The object is represented by a number of aspects, a representation of these aspects and a graph which describes the possible transitions between them. Aspect Graphs were used in the early 90's to recognize simple polyhedral objects, or objects which could be decomposed into generalized cones.

[137] sketches an extension of Aspect Graphs using CAD prototypes and a view-sphere for generic object recognition. This work is also based on simple geometric objects where the generality of the method deals only with small variations of the parameters of the used models (generalized cones).

An aspect graph is defined only for polyhedral objects, but the concept can be generalized for arbitrary objects. In this case one is interested in partitioning the view sphere of on object, such that the view of the object changes only slightly within the partition. An example how this can be done using a PCA based representation is the multiple eigenspace algorithm [92].

One advantage in using aspect graphs and related representations is that with the recognition of an object we not only know the object, but also its corresponding aspect, and the possible next aspects. This aspect gives us an idea of the viewing direction i.e. the pose of the camera.

3.3.2 Characteristic View

The concept of a characteristic view (CV) is useful in appearance-based object recognition [161]. Characteristic views are intended to help obtain a representative and adequate grouping of views, such that a given level of recognition accuracy may be achieved using a minimum number of stored views [44]. Clearly, this has important implications for the storage space needed to represent each object, and the number of matches which must be performed at run-time for the purpose of recognition. View grouping has been addressed using CVs and aspect graphs. An extension to the original idea of Koenderink et al. [83], the so called appearance graph uses the appearance of the object under consideration as well as information like illumination, texture etc. Problems in building aspect graphs occur when the object under consideration has curved surface, non-uniform illumination, etc, since it is hard to find stable views. Instead of building a complete aspect graph one can build an approximate of the object's appearance [27].

3.3.3 Generic Models based on Graphs

In learning a prototype from a set of noisy examples of the same object the goal is to find a representative model. If the examples are given as graphs, Jiang et.al. [75] introduced a concept of set median and generalized median graphs and a genetic algorithm to obtain the prototype graph. The generalized median concept is more powerful since it does not constrain the resulting graph as being one of the example graphs. Spectral methods were utilized to cluster graphs of different views [100].

Recently, Keselman and Dickison [82] introduced a novel approach based on graph shortest paths approximation to close the representation gap in the domain of automatic acquisition of 2D view-based models. The harder task of recognition is not tackled.

Cyr and Kimia [39] introduced a 3D object recognition algorithm based on 2D views. The aspects are based on a notion of shape similarity between views.

3.3.4 Dual Graphs and Combinatorial Maps

Dual graphs [85] can be seen as an extension of the well known region adjacency graph (RAG) representation. In 2D space a dual graph representation consists of a pair of the primary planar graph and its dual (called also geometric dual [65]). This representation is able to encode any subdivision of the 2D topological space. Encoding higher dimensions with graphs is a difficult problem. Combinatorial maps or generalized maps are well-suited representations to overcome this problem. In 2D space simple dual graphs are equivalent to 2D combinatorial maps.

N-dimensional combinatorial maps [95] may be seen as a graph with an embedding in an N-dimensional space, i.e., in the case of 2D [29], combinatorial maps are planar graphs encoding the orientation of edges around vertices. The base elements of an N-dimensional combinatorial map are the darts, also called half edges, which are connected (sewed) together by the orbits of 1 permutation and $N-1$ involutions. In the case of 2D [29], the permutation is called σ and forms vertices, and the involution is called α and specifies edges (other attributions for the permutations exist). One of the advantages of combinatorial maps is that in the 2D case, unlike dual-graphs, they explicitly encode the orientation of the plane, correctly handling all the complicated cases with self-loops and parallel edges.

Like combinatorial maps, n-dimensional generalized maps [95, 96] are defined in any dimension and correctly represent all topological configurations of the n-dimensional space (including 2D). Their base elements are darts and use only involutions to represent the connections between them. With these relations they describe cells in any dimension.

3.4 Geometry and Topology

Many topological models have been conceived for representing the topology of subdivided objects, since different types of subdivisions have to be handled: general complexes [32, 38, 47, 165] or particular manifolds [6, 14, 164], subdivided into any cells [61, 43] or into regular ones (e.g. simplices, cubes, etc.) [54, 134]. Few models are defined for any dimensions [18, 141, 28, 97]. Some of them are (extensions of) incidence graphs or adjacency graphs. So, their principle is often simple, but:

- they cannot deal with any subdivision without loss of information, since it is not possible to describe the relations between two cells precisely if they are incident in several locations;
- operations for handling such graphs are often complex, since they have to handle simultaneously different cells of different dimensions.

Other structures are "ordered" [28, 97, 47], and they do not have the drawbacks of incidence or adjacency graphs. A comparison between some of these structures is presented in [96]. A subdivided object can be described at different levels. For instance, a building is subdivided into floors, each floor is subdivided into wings, each wing is subdivided into rooms, etc. Thus, several contributions deal with hierarchical topological models and topological pyramids [40, 18, 85]. For geometric modeling, there are often only few levels. For image analysis, more levels are needed since the goal is to derive information which is not known a priori.

Since a geometric object is represented by a topological structure and its embedding in a geometric space we distinguish: (i) topological operations which modify the structure; (ii) embedding operations which modify the embedding; and (iii) geometric operations which modify both topology and embedding. For the animation of articulated objects, the object structure is not modified. Therefore, animation can be performed by applying embedding operations. Local operations can be easily defined and performed (e.g. chamfering, contraction, removal, extrusion, split, etc.), and this plays an important role when wanting to simultaneously (in parallel) apply them when an image is analyzed.

Moreover, topological features can be computed from the topological structure: orientability for pseudo-manifolds [30], genus for surfaces, and homology groups which provide information about the "holes" in the object for any dimension [3]. Such information can be used to control the construction of the object. For instance, when simplifying an image and constructing a pyramid, one often wants to keep some properties like connectedness invariant. When an object is made of many parts, one requires tools in order to check it. Topology and shape are complementary, and it is very useful to compute both types of information.

The use of geometry and topology for a generally valid representation should also incorporate the local object appearances. This third cue increases the use for a real world scene representation. Different scenarios can be conceived: for instance, a box with a picture of a lion on top might be represented by two layers in topological means (top and bottom of the box). Each of these layers is again represented by a topological description combined with its geometric appearance. One can imagine that the use of the surface appearance (here the picture of the lion) increases the discriminative power of such a representation.

4 Scene Representations

We can regard the scene representation as the internal state of a cognitive vision system. This representation should correspond as accurately as possible to the real scene which is observed by the system.

Some of the previously discussed methods for object representation in 2D extend quite naturally towards appearance-based scene representation. These can be global methods modeling brightness, contrast, color, texture, or integral features, which have for instance been applied in image retrieval, but also local methods. Bags of keypoints (and their descriptors) can be similarly used for object recognition (and, thus, object representation) [58], and for image retrieval (and, thus, scene representation) [115].

There is also a straightforward extension of the above explanations on 3D object representations to a type of scene representation, which works either in the 2D ('viewer centered') image coordinates, or in a 3D scene coordinate system, where each individual object in the scene is represented by a 3D vector that points at the origin of an 'object centered' coordinate system, which in turn is used to properly represent the individual object.

A scan of literature on scene representations has led to several results, which either do not originate from computer vision (artificial intelligence, linguistics, topology, geometry) and are hard to adapt to cognitive vision, or they have been tailored for a very specific vision application (navigation, tracking, surveillance, brain atlas, geographic information systems) and are not sufficiently general. Obviously, there is still plenty of room for future research in this area, as will be seen from the subsequent sections.

In general, a scene representation for cognitive vision should address the following topics:

1. The scene representation is not independent of the object representation, therefore it is important how objects are represented.

2. Time is an essential factor of the scene representation, especially the time resolution defines what types of events need to be represented (i.e., what type of object interactions can we resolve?).

3. The question of a purposive and qualitative representation should be addressed. It is clear that we do not need a full representation of every detail (reconstructive), but we must address the question of the purpose of the representation (cf. purposive and qualitative vision).

4.1 4D-Coordinates

Masunaga and Ukai [108] propose a database of 3D, moving objects, which is a 4D (Euclidean space+time) representation. The representation is at the object level and consists of:

- A list of all objects.

- For every object and every time instance: 6 DoF of object pose (3D position + orientation).

- A set of primary relations like velocity of an object, distance between objects and a set of topological relations: Disjoint, Contains, Inside, Overlaps, Touches, Equals, Covers and CoveredBy.

As a difference to the topology-based representations this (4D) scene representation has the advantage that one can use history information. This could be useful as discussed in section 1.2.

4.2 Appearance-based scene representations

Similar to object representations we could also use the appearance of the scene as a representation of it. In the simplest case we just store some snapshots of the scene. In more sophisticated approaches (mainly used in a robotics context) more complex appearance-based representations are used. Similar to object recognition we have global appearance based scene representations. A typical example is a robot equipped with an omni-directional camera and global PCA as a representation, e.g. [118]. Another class of appearance-based representations are local ones. Similar to the object recognition case, only distinguished regions in the scene are represented. A recent approach for this type of representation can be found in Se et al. [147]. In this paper, a vision-based mobile robot self localization and mapping (SLAM) algorithm is presented that uses SIFT descriptors as a scene representation of visual landmarks. It is interesting to note that visual landmarks are also used by a variety of animals for navigation purposes.

4.3 Occupancy Grids

Occupancy grids [46] are different from the scene representations discussed above. The world is divided into fixed grid cells. In every cell there is a value stored, which stands for the probability that this cell is empty / occupied. Thus, this representation does not deal with different objects and their motion in the scene, but with the space they occupy. From the nature of this representation, it will also be difficult to distinguish between objects and static background.

Occupancy grids are mostly used in robotic applications. Here the grid is a 2D floor plan of the scene, which describes where the robot can move around in the world. In most applications the grid is estimated with sonar sensors or laser-range finders, but there exist also applications where stereo vision is used.

There have also been attempts to model imperfect knowledge about the scene (ignorance, imprecision, and ambiguity), e.g. within a complete representational framework for fuzzy mathematical morphology by Isabelle Bloch [25]. Probabilistic, possibilistic, and fuzzy occupancy grids have been proposed [139, 26]. One problem with these approaches is their static description (description of one frame) of the scene.

4.4 Topology-based

Topology-based approaches often relate to linguistics or artificial intelligence and try to simplify the representation based on relations. Such a representation may consist of a list of objects plus a list of relations between these objects. Interval calculus [4] is used in systems that require some form of temporal reasoning capabilities. In [4] 13 interval-interval relations are defined: ‘before’, ‘after’, ‘meets’, ‘met-by’, ‘overlaps’, ‘overlapped-by’, ‘started-by’, ‘starts’, ‘contains’, ‘during’, ‘ended-by’, ‘ends’ and ‘equals’. In [138], motivated by the work in [4, 35, 36], an interval calculus-like formalism for the spatial domain, the so called region connection calculus (RCC) was presented. In (RCC-8) [138], the set of 8 possible relations between two regions are: ‘is disconnected from’, ‘is externally connected with’, ‘partially overlaps’, ‘is a tangential proper part of’, ‘is non-tangential proper part of’, ‘has a tangential proper part’, ‘has non-tangential proper part’, and ‘equals’. A more expressive calculus can be produced with additional relations to describe regions that are either inside, partially inside, or outside other regions (RCC-15). There exist also extensions of RCC to 3D space and time ([162]).

Different graph based representations have been used to describe the changes / events in a dynamic space. In [33] graphs are used to describe actions (vertices represent actions). Graphs are also used in [9], but here vertices represent objects. Balder [9] argues that arbitrary changes can be best described by a state approach: the state of the world before and after the change characterizes the change completely. The Unified Modeling Language (UML), in its state diagram, also defines a graph based representation for tracking temporal changes.

4.5 Event Representation

Representing objects and their spatio-temporal behavior in a scene can be done in different ways on different levels. For a cognitive vision system it might be interesting to detect events in a video sequence instead of just behavior. For instance, lifting a cup, moving it and finally putting it over a smaller item might be described by the events: 'lift', 'move', 'put down'. It could also be represented by the event 'hide'. For example, for a short video with a hand using 2 cups to hide a ball, such a description would be: 'hand from left', 'grasps left cup', 'moves it over ball', 'releases cup', 'shifts it to the left', 'releases cup', etc.

Event representation from video sequences has broad interest (e.g. video surveillance). There is a vast amount of literature and we give only a very brief overview to state-of-the-art research. Recently a consortium of researchers developed a formal language to describe the ontology of events in videos which they called 'VERL' (Video Event Representation Language) [121]. In [73] a proposal on spatio-temporal graphs is presented. The authors proposed to use new relations (like grasp, move and release) to describe events as the changes in the scene and to build a hierarchical graph-based representation to keep track of actions, events, and relations. Hakeem et al. [63] presented an extension of the 'CASE' description method which bridges the representational gap between low level vision and human scene description. Such representations can be learned by e.g. the method of Hakeem and Shah [62] which uses a video event graph and a video correlation graph on a set of training videos.

5 Promising Research Directions

Finally, we conclude with the hard task of presenting promising directions of research for a field as active and developing as cognitive vision is today. We identify a number of evident research goals, and we also present our more specific ideas. Some of them are reflected by our own current research activities, others will probably be addressed in the near future.

Representational concert: It is quite clear that for a versatile cognitive system we need a multitude of representations to have to work together. Probably all (and maybe more) that have been discussed in this paper. But this requires to answer several hard questions, like how we can keep these representations consistent? What happens if they are inconsistent? Should we jointly update them or keep them separate? How do we decide which representation is the most appropriate to solve a task?

Recognition versus reconstruction: To which extent is an explicit metric representation of space and time required? What are the limitations of purely appearance-based approaches? It seems that a hybrid representation might be most appropriate. This could cover the approximate reconstruction of objects and their trajectories (spatio-temporal reconstruction, including camera and object pose), a qualitative (graph-based) representation of spatial and temporal relations and events, and purely appearance-based recognition.

Object versus background: What is a relevant object for the current state of representation and reasoning in the cognitive system? What is 'background'? Attention may switch, and an object may gain importance and be separated from the surrounding background, for instance by grasping it. On the other hand, some object may loose importance. Having not been used for an extended period of time, it might be merged with the background. This kind of approach has for instance been followed by the EU-IST project VAMPIRE, where a 'visual active memory' served the purpose of storing and retrieving relevant visual information at various levels of abstraction [64].

Spatio-temporal representation: A 4D scene representation similar to the one described in section 4.1 may be well-suited to represent spatio-temporal relationship. Using such a scene representation one could answer the following important questions:

- Where are the objects? This information is directly available from the representation.
- Object trajectories? This information is directly available from the history of the representation.
- Which object hides which object? This information can be inferred from the topological relations of the representation.

At the object level, a certain amount of metric representation will be required to represent camera-to-object pose, object and camera trajectories in scene coordinates etc. At a base representational level, objects could be represented as a 3D point cloud, and motion by the correspondence of moving points between subsequent frames. This representation might serve as the basis for an increasingly complex representational framework. One can think about attaching local descriptors to points of interest, combining point and contour information, thus representing an object as a collection of loosely related pieces of information. Even the boundary of an object might be represented in a fuzzy manner. First experimental results in this direction show that a representation based on a point cloud is sufficient to improve learning and recognition of object categories [146], but for a proper object representation and for spatio-temporal analysis, this can only be considered a starting point.

Appearance-based representation: Currently appearance-based representations are quite popular in the object recognition community, but there are some obstacles that need to be overcome. First of all we need to address the issues of scaling (i.e., how can we build a system that can recognize thousands of objects). It is clear that one-to-one matching is not useful, we need proper ways to hierarchically structure the object representation. A recent approach in that direction can be found in [125], but this approach is not the solution because it needs too much storage space. Very recently Nistér and Stewénius [124] presented a hierarchical system with that scales to many specific objects. It remains to be shown how the generalization and accuracy is influenced by the approximations introduced. Another interesting issue is how we can combine local and global appearance-based representations in a coherent framework in order to use the best of both worlds. Ideally we would like to have a seamless integration between these two representations so that we can always select the most appropriate one. A question closely related to learning is how we can build a hierarchy of appearance-based representations, starting with simple patches based on interest points to more complex constellations and expressive representations. It is clear that such a representation needs to be learned (in an un-supervised manner).

Graph-based representation: Existing graph based representations support any number of dimensions but the question of minimum required complexity is still open. Do we really need more than $2D + time$ and represent everything in maximum detail or should we focus more on the advantages of representations which keep embedding information, structure, and topology? Preliminary experimental results show that in most of the cases humans do very well using simple descriptions enhanced by the power of relations. RCC is definitely something that should be considered in the future, along with n-dimensional graph based representation like combinatorial maps and generalized maps, for which many properties in $3D$ and $4D$ still need to be studied. Holes of different dimension result from the local and global connectivity of parts. Homology generators characterize holes and can be efficiently computed using a graph pyramid [135]. Other topological invariants like cohomology rings [59] could also be considered. Staying with qualitative measures, landmark based addressing, mosaicking graph based patches, and, graph and shape matching should enforce the way to human like generic object recognition capabilities and, thus, will certainly be part of our future research.

Acknowledgements

This work was supported by the Austrian Science Funds FWF under contracts S9103-N04 and S9103-N13.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *Proc. European Conference on Computer Vision*, pages 113–130, 2002.

[2] A. Akbarzadeh, J.-M. Frahm, P. Mordohai, B. Clipp, C. Engels, D. Gallup, P. Merrell, M. Phelps, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, H. Towles, D. Nistér, and M. Pollefeys. Towards urban 3D reconstruction from video. In *Proc. 3DPVT*, 2006.

[3] S. Alayrangues, X. Daragon, J.-O. Lachaud, and P. Lienhardt. Computation of homology groups and generators. In *Proceedings of DGCI 2005*, pages 195–205. Springer, 2005.

[4] J.F. Allen. An Interval-based Representation of Temporal Knowledge. In *Proc. 7th Inter. Joint Conf. on AI*, pages 221–226, 1981.

[5] Y. Aloimonos and D. Shulman. *Integration of visual modules: an extension of the Marr paradigm*. Academic Press, 1989.

[6] S. Ansaldi, L. de Floriani, and B. Falcidieno. Geometric modeling of solid objects by using a face adjacency graph representation. *Computer Graphics*, 19(3):131–139, 1985.

[7] K. Baek, B. Draper, J. R. Beveridge, and K. She. PCA vs. ICA: A comparison on the FERET data set. In *The 6th Joint Conference on Information Sciences*, pages 824–827, Durham, North Carolina, March 8-14 2002.

[8] R. Bajcsy and D. A. Rosenthal. Visual and conceptual focus of attention. In S. Tanimoto and A. Klinger, editors, *Structured Computer Vision*, pages 133–149. Academic Press, 1980.

[9] Norman I. Balder. *Temporal Scene Analysis: Conceptual Descriptions of Object Movements*. PhD thesis, University of Toronto, Canada, 1975.

[10] Dana H. Ballard. Interpolation coding: A representation for numbers in neural models. Technical Report TR–175, Dept. of CS, Univ. of Rochester, September 1986.

[11] H.B. Barlow. The coding of sensory messages. In W. H. Thorpe and O. L. Zangwil, editors, *Current Problems in Animal Behavior*, pages 331–360. Cambridge University Press, 1961.

[12] A. Baumberg. Reliable feature matching across widely separated views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, pages 774–781, 2000.

[13] Adam Baumberg. Reliable feature matching across widely separated views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, volume 1, pages 774–781, June 2000.

[14] B. Baumgart. A Polyhedron Representation for Computer Vision. In *AFIPS Nat. Conf. Proc.*, volume 44, pages 589–596, 1975.

[15] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenspaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[16] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. In *PAMI*, volume 24, pages 509–522, April 2002.

[17] Elliot Joel Bernstein and Yali Amit. Part-based statistical models for object classification and detection. In *CVPR*, volume 2, pages 734–740, 2005.

[18] Yves Bertrand, Guillaume Damiand, and Christophe Fiorio. Topological Encoding of 3D Segmented Images. In Gunilla Borgefors, Ingela Nyström, and Gabriella Sanniti di Baja, editors, *Proceedings DGCI'00, Discrete Geometry for Computer Imagery*, volume 1953 of *Lecture Notes in Computer Science*, pages 311–324, Uppsala, Sweden, 2000. Springer, Berlin Heidelberg, New York.

[19] I. Biederman. Human image understanding: Recent research and a theory. *CVGIP*, 32:29–73, 1985.

[20] Irving Biederman. Matching image edges to object memory. In *Proceedings of the First International Conference on Computer Vision*, pages 384–392, London, England, 1987.

[21] J. Bigün and G. H. Granlund. Optimal orientation detection of linear symmetry. In *Proceedings of the IEEE First International Conference on Computer Vision*, pages 433–438, London, Great Britain, June 1987.

[22] G.-A. Bilodeau and R. Bergevin. Qualitative part-based models in content-based image retrieval. *Machine Vision and Applications*, 2007. In press, published online: 10 January.

[23] H. Bischof, H. Wildenauer, and A. Leonardis. Illumination insensitive recognition using eigenspaces. *Computer Vision and Image Understanding*, 95(1):86–104, 2004.

[24] Randolph Blake and Robert Sekuler, editors. *Perception*. McGraw-Hill, 2006.

[25] I. Bloch. Fuzzy relative position between objects in image processing: A morphological approach. *PAMI*, 21(7):657–664, 1999.

[26] I. Bloch and A. Saffiotti. On the representation of fuzzy spatial relations in robot maps. In B. Bouchon-Meunier, L. Foulloy, and R.R. Yager, editors, *Intelligent Systems for Information Processing*, pages 47–57. Elsevier, NL, 2003. Online at http://www.aass.oru.se/˜asaffio/.

[27] Peter Boros and Richard E. Blake. Appearance graph generation using ray tracing and graph matching. In *Proceedings of the 2nd Asian Conference on Computer Vision*, Singapore, 1995.

[28] E. Brisson. Representing geometric structures in d dimensions: Topology and order. *Discrete and Computational Geometry*, 9:387–426, 1993.

[29] Luc Brun and Walter G. Kropatsch. Dual Contraction of Combinatorial Maps. Technical Report PRIP-TR-54, Institute f. Computer Aided Automation 183/2, Pattern Recognition and Image Processing Group, TU Wien, Austria, 1999. Also available through http://www.prip.tuwien.ac.at/ftp/pub/publications/trs/tr54.ps.gz.

[30] Luc Brun and Walter G. Kropatsch. Inside and Outside within Combinatorial Pyramids. *Pattern Recognition, accepted*, 2005.

[31] Gustavo Carneiro and Allan D. Jepson. Phase-based local features. In *Proc. 7th European Conference on Computer Vision, Copenhagen, Denmark*, 2002.

[32] P.R. Cavalcanti, P.C.P. Carvalho, and L. Martha. Non-manifold modeling: an approach based on spatial subdivision. *Computer-Aided Design*, 29(3):299–220, 1997.

[33] A. Chella, M. Frixione, and S. Gaglio. Understanding Dynamic Scenes. *Artificial intelligence*, 123:89–132, 2000.

[34] N. Christianini and J. S. Taylor. *Support vector machines and other kernel-based methods*. Cambridge university press, 2000.

[35] B.L. Clarke. A Calculus of Individuals Based on Connection. *Notre Dame Journal of Formal Logic*, 23(3):204–218, 1981.

[36] B.L. Clarke. Individuals and Points. *Notre Dame Journal of Formal Logic*, 26(1):61–75, 1985.

[37] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2005.

[38] G. Crocker and W. Reinke. An editable non-manifold boundary representation. *Computer Graphics and Applications 11,2 (1991)*, 11(2), 1991.

[39] C. M. Cyr and B. B. Kimia. A Similarity-based Aspect-graph Approach to 3D Object Recognition. *International Journal of Computer Vision*, 57(1):5–22, April 2004.

[40] Leila De Floriani, Enrico Puppo, and Paolo Magillo. A formal approach to multiresolution hypersurface modeling. in w. strasser, r. klein and r. rau eds. geometric modeling :. In Gilles Bertrand, Michel Couprie, and Laurent Perroton, editors, *Geometric Modelling Theory and Practice*, volume 1568 of *Lecture Notes in Computer Science*, pages 3–18, Marne-la-Vallée, France, 1999. Springer, Berlin Heidelberg, New York.

[41] S. Dickinson, R. Bergevin, I. Biederman, J.-O. Eklundh, R. Munck-Fairwood, A.K. Jain, and A. Pentland. Panel report: the potential of geons for generic 3-d object recognition. *Image and Vision Computing*, 15:277–192, 1997.

[42] S. Dickinson, A. Pentland, and A. Rosenfeld. 3-d shape recovery using distributed aspect matching. *PAMI*, 14(2):174–198, 1992.

[43] D. Dobkin and M. Laszlo. Primitives for the manipulation of three-dimensional subdivisions. In *Proc. 3rd Symposium on Computational Geomtry*, pages 86–99, Waterloo, Canada, 1987.

[44] C. Dorai and A. K. Jain. Shape Spectrum Based View Grouping and Matching of 3D Free-form Object. *Pattern Analysis and Machine Intelligence*, 19(10):1139–1145, 1997.

[45] S. Edelman. *Representation and Recognition in Vision*. MIT Press, 1999.

[46] A. Elfes. *Occupancy grids: A probabilistic framework for robot perception and navigation*. PhD thesis, Carnegie Mellon University, 1989.

[47] H. Elter and P. Lienhardt. Cellular complexes as structured semi-simplicial sets. *Int. Journal of Shape Modeling*, 1(2):191–217, 1994.

[48] O.D. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993.

[49] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *CVPR GMBV Workshop on Generative-Model Based Vision*, 2004.

[50] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, 2003.

[51] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proc. European Conference of Computer Vision*, pages 242–256, 2004.

[52] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *Proc. CVPR*, 2005.

[53] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneos object recognition and segmentation by image exploration. In *Proc. European Conference on Computer Vision*, pages 40–54, 2004.

[54] V. Ferruci and A. Paoluzzi. Extrusion and boundary evaluation for multidimensional polyhedra. *Computer-Aided Design*, 23(1):40–50, 1991.

[55] S. Fidler and A. Leonardis. Robust lda classification. In C. Beleznai and T. Schlögl, editors, *Vision in a Dynamic World, Proc. of 27th ÖAGM/AAPR 2003*, pages 119–126. Austrian Computer Society, 2003.

[56] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *ISPRS Intercommission Workshop, Interlaken*, June 1987.

[57] D. Forsyth and J. Ponce. *Computer Vision, a modern approach*. Prentice Hall, 2003.

[58] G.Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, pages 1–22, 2004.

[59] Rocío González-Díaz and Pedro Real. On the cohomology of 3d digital images. *Discrete Appl. Math.*, 147(2-3):245–263, 2005.

[60] M. Grabner and H. Bischof. Extracting object representations from local feature trajectories. In D. Chetverikov, L. Czuni, and M. Vincze, editors, *Proc. Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition*, volume 192, pages 265–272. Austrian Computer Society, 2005.

[61] L. Guibas and J. Stolfi. Primitives for the Manipulation of General Subdivisds and the Computation of Voronoi Diagrams. *ACM. Transactions on Graphics*, 4(2):74–123, 1985.

[62] A. Hakeem and M. Shah. Multiple agent event detection and representation in videos. In *The Twentieth National Conference on Artificial Intelligence (AAAI)*, 2005.

[63] A. Hakeem, Y. Sheikh, and M. Shah. CaseE: A hierarchical event representation for the analysis of videos. In *The Ninteenth National Conference on Artificial Intelligence (AAAI)*, 2005.

[64] M. Hanheide, Ch. Bauckhage, and G. Sagerer. Memory consistency validation in a cognitive vision system. In *Proc. ICPR*, volume 2, pages 459–462, 2004.

[65] F. Harary. *Graph Theory*. Addison-Wesley, 1994.

[66] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988.

[67] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2003.

[68] Bernd Heisele, Thomas Serre, Massimiliano Pontil, and Tomaso Poggio. Component-based face detection. In *CVPR*, 2001.

[69] Scott Helmer and David G. Lowe. Object class recognition with many local features. In *CVPR GMBV Workshop on Generative-Model Based Vision*, 2004.

[70] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

[71] J. E. Hummel. Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich and A. Markman, editors, *Cognitive Dynamics: Conceptual Change in Humans and Machines*, pages 157–185. Erlbaum, 2000.

[72] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, 2001.

[73] Adrian Ion, Yll Haxhimusa, and Walter G. Kropatsch. A Graph-Based Concept for Spatiotemporal Information in Cognitive Vision. In L. Brun and M. Vento, editors, *5th IAPR-TC15 Workshop on Graph-based Representation in Pattern Recognition*, volume 3434 of *Lecture Notes in Computer Science*, pages 223–232, Poitiers, France, April 2005. Springer, Berlin Heidelberg, New York.

[74] Adrian Ion, Yll Haxhimusa, and Walter G. Kropatsch. A graph-based concept for spatiotemporal information in cognitive vision. Technical Report PRIP-TR-98, Institute f. Computer Aided Automation 183/2, Pattern Recognition and Image Processing Group, TU Wien, Austria, 2005. Also available through http://www.prip.tuwien.ac.at/ftp/pub/publications/trs/tr98.pdf.

[75] Xiaoyi Jiang, Adreas Muenger, and Horst Bunke. On median graphs: Properties, algorithms and applications. *Transaction on Pattern Recognition and Machine Intelligence*, 23(10):1144–1151, October 2001.

[76] Andrew E. Johnson and Martial Hebert. Using spin-images for efficient multiple model recognition in cluttered 3-d scenes. *Trans PAMI*, 21(5):433–449, 1999.

[77] B. Julesz and J. R. Bergen. Textons, the fundamental elements in preattentive vision and perception of textures. *The Bell System Technical Journal*, 62(6):1619–1645, July-August 1983.

[78] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[79] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Proc. 8th European Conference on Computer Vision, Prague, Czech Republic*, pages Vol I: 228–241, 2004.

[80] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proc. CVPR 2004*, pages 506–513. IEEE CS Press, 2004.

[81] Y. Keselman and S. Dickinson. Generic model abstraction from examples. In *CVPR 2001*. IEEE CS Press, 2001.

[82] Y. Keselman and S. Dickinson. Generic model abstraction from examples. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(7):1141–1156, 2005.

[83] J. J. Koenderink and A. J. Doorn. The Internal Representation of Solid Shape with Respect to Vision. *BioCyber*, 32:211–216, 1979.

[84] Ullrich Köthe. Edge and junction detection with a improved structure tensor. *Lecture Notes in Computer Science - 25th Pattern Recognition Symposium DAGM*, pages 25–32, 10.-12. September 2003.

[85] Walter G. Kropatsch. Building Irregular Pyramids by Dual Graph Contraction. *IEE-Proc. Vision, Image and Signal Processing*, 142(6):366–374, December 1995.

[86] Walter G. Kropatsch. Abstract pyramid on discrete represtations. In In J. O. Lachaud, A. Braquelaire, and A. Vialard, editors, *DGCI 2002 Lecture Notes in Computer Science, 2301*, pages 1–21, France, 2002. Springer Verlag.

[87] M.P. Kumar, P.H.S. Torr, and A. Zisserman. Extending pictural structures for object recognition. In *In Proc. of British Machine Vision Conference*, 2004.

[88] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. ICCV*, pages 432–439, 2003.

[89] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using affine-invariant regions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin*, volume 2, pages 319–324, June 2003.

[90] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[91] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV'04 Workshop on Statistical Learning in Computer Vision, Prague*, May 2004.

[92] A. Leonardis, H. Bischof, and J. Maver. Multiple eigenspaces. *Pattern Recognition*, 35(11):2613–2627, 2002.

[93] Ales Leonardis and Horst Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding: CVIU*, 78(1):99–118, 2000.

[94] M. leordeanu, M. Hebert, and R. Sukthankar. Beyond local appearance: Category recognition from pairwise interactions of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, Minnesota*, June 2007.

[95] P. Lienhardt. Subdivisions of n-dimensional spaces and n-dimensional generalized maps. In Kurt Mehlhorn, editor, *Proceedings of the 5th Annual Symposium on Computational Geometry (SCG '89)*, pages 228–236, Saarbrücken, FRG, June 1989. ACM Press.

[96] P. Lienhardt. Topological models for boundary representation: a comparison with n-dimensional generalized maps. *Computer-Aided Design*, 23(1):59–82, 1991.

[97] P. Lienhardt. N-dimensional generalized combinatorial maps and cellular quasi-manifolds. *Int. Journal of Computational Geometry and Applications*, 4(3):275–324, 1994.

[98] T. Lindeberg. *Scale space theora in computer vision*. Kluwer, 1994.

[99] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[100] Bin Lou, Richard C. Willson, and Edwin R. Hancock. Spectral embedding of graphs. *Pattern Recognition*, 36:2213–2230, 2004.

[101] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[102] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[103] D.G. Lowe. *Perceptual Organization and Visual Cognition*. Kluwer, 1985.

[104] D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV99*, pages 1150–1157, 1999.

[105] Y. Ma, S. Soatto, J. Košecá, and S.S. Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer, 2004.

[106] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information.* W.H. Freeman, 1982.

[107] A. M. Martinez and A. C. Kak. PCA versus LDA. *PAMI*, 23(2):228–233, February 2001.

[108] Yoshifumi Masunaga and Noriko Ukai. Towards a 3d moving object data model - a preliminary consideration -. In *Proc. IEEE Int.Symp. on Database Applications in Non-Traditional Environments, DANTE'99*, page 302ff, 1999.

[109] Jiri Matas, Ondrej Chum, Martin Urban, and Tomas Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 384–393, 2002.

[110] David M. McKeown and John McDermott. Toward expert systems for photo interpretation. In *Proc. of Trends and Applications*, pages 33–39. IEEE Comp.Soc., 1983.

[111] G. Medioni, M.-S. Lee, and C.-K. Tang. *A computational framework for segmentation and grouping*. Elsevier, 2000.

[112] T. Melzer, M. Reiter, and H.Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition, Special Issue on Kernel and Subspace Methods for Computer Vision*, 36(9):1961–1971, 2003.

[113] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müller. Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing*, 9:41–48, 1999.

[114] Krystian Mikolajczk and Cordelia Schmid. A performance evaluation of local descriptors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin*, volume 2, pages 257–263, June 2003.

[115] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. ICCV*, pages 525–531, 2001.

[116] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, volume 1, pages 128–142, 2002.

[117] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, November 2005.

[118] M.Jogan and A. Leonardis. Robust localization using an omnidirectional appearance-based subspace model of environment. *Robotics and Autonomous Systems*, 45(1):51–72, 2003.

[119] M.Pollefeys, R. Koch, and L. Van Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. *Int.J. Computer Vision*, 32(1):7–25, 1999.

[120] Hiroshi Murase and Shree K. Nayar. Visual learning and recognition of 3-d objects from appearance. *Int.J. Computer Vision*, 14(1):5–24, 1993.

[121] R. Nevatia, J. Hobbs, and B. Bolles. An ontology for video event representation. In *IEEE Workshop on Event Detection and Recognition*, 2004.

[122] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, 2002.

[123] D. Nister. Preemptive RANSAC for live structure and motion estimation. In *Proc. ICCV*, pages 199–206, 2003.

[124] D. Nistér and H. Stewénius. Stable recognition with a vocabulary tree. In *Proc. CVPR*, 2006.

[125] S. Obdrzalek and J. Matas. Sub-linear indexing for large scale object recognition. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 1–10, 2005.

[126] K. Ohba and K. Ikeuchi. Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *PAMI*, 9:1043–1047, 1997.

[127] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In T. Pajdla and J. Matas, editors, *ECCV'04*, volume 3022 of *LNCS*, pages 71–84. Springer, 2004.

[128] A. Opelt and A. Pinz. Object localization with boosting and weak supervision for generic object recognition. In *Proceedings of the 14th Scandinavian Conference on Image Analysis*, 2005.

[129] A. Opelt, A. Pinz, and A. Zisserman. A Boundary-Fragment-Model for object detection. In *Proc. ECCV*, volume II, pages 575–588, 2006.

[130] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. In *Proc. European Conference of Computer Vision*, volume II, pages 575–588, 2006.

[131] A. Opelt, A. Pinz, and A. Zisserman. Fusing shape and appearance information for object category detection. In *Proc. BMVC*, 2006.

[132] A. Opelt, J. Sivic, and A. Pinz. Generic object recognition from video data. In *Proceedings of the Austrian Cognitive Vision Workshop*, 2005.

[133] S.E. Palmer. *Vision Science - Photons to Phenomenology*. MIT Press, 1999.

[134] A. Paoluzzi, F. Bernardini, C. Cattani, and V. Ferrucci. Dimension independent modeling with simplicial complexes. *ACM Transactions on Graphics*, 12(1), 1993.

[135] Samuel Peltier, Adrian Ion, Yll Haxhimusa, Walter G. Kropatsch, and Guillaume Damiand. Computing homology group generators of images using irregular graph pyramids. In F. Escolano and M. Vento, editors, *Proceedings of the 15th International Workshop on Graph-based Representation for Pattern Recognition.*, volume 4538 of *Lecture Notes in Computer Science*, pages 283–294, Alicante, Spain, June 2007. Springer, Berlin Heidelberg, New York.

[136] A. Pinz. Object categorization. *Foundations and Trends in Computer Graphics and Vision*, 1(4):255–353, 2006.

[137] A. Pinz and J-Ph. Andreu. Qualitative spatial reasoning to infer the camera position in generic object recognition. In *Proceedings ICPR'98*, volume I, pages 770–773, 1998.

[138] D.A. Randell, Z. Cui, and A.C. Cohn. A Spatial Logic Based on Regions and Connection. In *Proc. 3rd Intern. Conf. on Knowledge Representation and Reasoning*, pages 165–176. Morgan Kaufmann, 1992.

[139] M. Ribo and A. Pinz. A comparison of three uncertainty calculi for building sonar-based occupancy grids. *Int.J. Robotics and Autonomous Systems*, 35:201–209, 2001.

[140] K. Rohr. Localization properties of direct corner detectors. *Journal of Mathematical Imaging and Vision*, 4:139–150, 1994.

[141] J. Rossignac and M. O'Connor. A dimension-independent model for pointsets with internal structures and incomplete boundaries. In J. Turner M. Wozny and K. Preiss, editors, *Proc. Geometric Modeling for Product Engineering*, pages 145–180. North-Holland, 1989.

[142] S. Sarkar and K. Bowyer. *Computing Perceptual Organization in Computer Vision*. World Scientific, 1994.

[143] Frederik Schaffalitzky and Andrew Zisserman. Multi-view matching for unordered image sets, or how do i organize my holiday snaps? In *Proc. 7th European Conference on Computer Vision, Copenhagen, Denmark*, volume 1, pages 414–431, 2002.

[144] Cordelia Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19:530–535, 1997.

[145] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[146] G. Schweighofer, A. Opelt, and A. Pinz. Improved object categorization by unsupervised object localization. In *Proc. Int. Workshop on Learning for Adaptable Visual Systems LAVS'04*, August 2004.

[147] S. Se, D. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research*, 21(8):735–758, 2002.

[148] J.G. Semple and G.T. Kneebone. *Algebraic Projective Geometry*. Oxford University Press, 1952.

[149] J. Shotton, A. Blake, and R. Cipolla. Contour-based learning for object detection. In *Proc. ICCV*, 2005.

[150] N. Sigala. *Visual object categorization and representation in primates: psychophysics and physiology*. Logos Verlag, 2002.

[151] D. Skocaj, H. Bischof, and A. Leonardis. A robust PCA algorithm for building representations from panoramic images. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proc. ECCV02*, volume IV, pages 761–775. Springer, 2002.

[152] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12(2):131–147, 1990.

[153] M. Tarr and H. Bülthoff, editors. *Object Recognition in Man, Monkey, And Machine*. MIT Press, 1998.

[154] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, 1991.

[155] Staffan Truvé and Whiteman Richards. From Waltz to Winston (via the connection table). In *Proceedings of the First International Conference on Computer Vision*, pages 393–404, London, England, 1987.

[156] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 1(59):61–85, 2004.

[157] S. Ullman. *High-level Vision: Object Recognition and Visual Cognition*. MIT Press, 1997.

[158] Luc Van Gool, T. Moons, and D. Ungureanu. Affine/ photometric invariants for planar intensity patterns. In *Proc. 4th European Conference on Computer Vision, Cambridge, UK*, volume 1, pages 642–651, 1996.

[159] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[160] D. Vernon, editor. *A research Roadmap of Cognitive Vision*. Aug 2005. ECVision Report, Version 5.0 available at http://www.ecvision.org/research_planning/Research_Roadmap.htm.

[161] R. Wang and H. Freeman. Object Recognition based on Characteristic View Classes. In *Internation Conference on Pattern Recognition, ICPR90*, volume I, pages 8–12, 1990.

[162] Sheng-Sheng Wang, DA-You Liu, Xiao-Dong Liu, and Bo Yang. Spatio-temporal representation for multi-dimensional occlusion relation. In *Procedings of the Second International Conference on Machine Learning and Cybernetics*, pages 1677–1681, 2003.

[163] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. ECCV*, volume 1, pages 18–32, 2000.

[164] K. Weiler. Edge-based data structures for solid modeling in curved-surface environments. *Computer Graphics and Applications*, 5(1):21–40, 1985.

[165] K. Weiler. The radial-edge data structure: A topological representation for non-manifold geometry boundary modeling. In *Geometric Modeling for CAD Applications*, pages 3–36, Rensselaerville, USA, 1988.

[166] Jutta Willamowski, Damian Arregui, Gabriella Csurka, Christopher R. Dance, and Lixin Fan. Categorizing nine visual classes using local appearance descriptors. In *Proc. Int. Workshop on Learning for Adaptable Visual Systems LAVS'04*, 2004.

[167] H.J. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Computational Science & Engineering*, 4(4):10–21, 1997.

[168] Binglong Xie, Dorin Comaniciu, Visvanathan Ramesh, Markus Simon, and Terrance Boult. Component fusion for face detection in the presence of heteroscedastic noise. In *Proc. DAGM*, pages 434–441, 2003.

[169] M. Zerroug and R. Nevatia. Segmentation and 3-d recovery of SHGCs from a single intensity image. In *Proc. ECCV*, pages 319–330, 1994.