ORIGINAL PAPER

Computational object recognition: a biologically motivated approach

Tim C. Kietzmann · Sascha Lange · Martin Riedmiller

Abstract We propose a conceptual framework for artificial object recognition systems based on findings from neurophysiological and neuropsychological research on the visual system in primate cortex. We identify some essential questions, which have to be addressed in the course of designing object recognition systems. As answers, we review some major aspects of biological object recognition, which are then translated into the technical field of computer vision. The key suggestions are the use of incremental and view-based approaches together with the ability of online feature selection and the interconnection of object-views to form an overall object representation. The effectiveness of the computational approach is estimated by testing a possible realization in various tasks and conditions explicitly designed to allow for a direct comparison with the biological counterpart. The results exhibit excellent performance with regard to recognition accuracy, the creation of sparse models and the selection of appropriate features.

Keywords Biologically inspired computer vision ·
Object recognition · View-based object representations ·
Feature selection · Incremental learning

T. C. Kietzmann (✉)
Institute of Cognitive Science, University of Osnabrück,
Osnabrück, Germany
e-mail: tkietzma@uos.de

S. Lange · M. Riedmiller
Institute of Computer Science, University of Osnabrück,
Osnabrück, Germany

1 Introduction

In recent years, the area of computer vision has made great advances coming up with a variety of different approaches and solutions. Especially in the area of robotics, the analysis of visual information is very important. Here, the task of object recognition (also known as object identification) is of major significance because it forms the basis for further computations such as reasoning and decision making. Since robots directly interact in real world environments, creating highly adaptive and reliable systems capable of real time recognition is a crucial issue.

Although considerable progress has been made, performance of most of the current systems is still limited to special tasks and areas of application. As a result, automatically finding solutions capable of dealing with a great variety of tasks while still preserving the effectiveness of task-specificity still remains a difficult challenge. For humans, however, object recognition forms a very basic capability, working seamlessly in most diverse situations and tasks. Based on this superior performance, it is reasonable to exploit our biological and psychological knowledge to guide and inspire the creation of artificial vision systems.

The idea is not new. Among the first authors to describe this kind of approach were Poggio and Edelman (1990) who employed neural networks in order to learn and represent object models for visual object recognition. However, their approach dealt only with artificial data in the form of wire-frame 3D objects. Task complexity was thus comparatively low and far from real world applications. SEEMORE, a neurally inspired architecture capable of dealing with more realistic images was later introduced by Mel (1997). Here, the biologically relevant notion of view-based object representation was applied together with a great variety of visual features. Object recognition was achieved by a nearest neighbor

decision strategy, i.e. comparing each input vector to all training patterns and selecting the nearest one to assign an object label. A more recent approach was provided by Wallraven and Bülthoff (2001a). In addition to explicitly using only some of the training vectors as stored object views, temporal information was successfully used to fully represent object models.

The selection of a good feature space is a very important issue because it forms the basis of object representations, learning and recognition. A considerable amount of work has been put into the design of biologically motivated features and feature hierarchies (Mutch and Lowe 2006; Mel 1997; Serre et al. 2005; Wersing and Korner 2002), exhibiting very good performance. Taken together, there is an increasing body of research pointing into the direction that biologically motivated systems are able to achieve very promising results while being able to deal with more generic and thus less specialized object recognition tasks.

In this context, we will address some important and yet unsolved issues of computer vision systems such as the selection of an appropriate object representation scheme, automatic extractions of object model connections, how to achieve increased robustness and abstractness, the selection of an appropriate feature space and the choice for an adaptive learning mechanism. To tackle these problems, we identify and discuss important neurophysiological and neuropsychological evidence and argue for the necessity in this particular case. As a consequence, we conclude that a consistent model of computational object recognition must respect the following five aspects: (1) Object representations should be view-based. (2) The resulting view-prototypes should be associated by Hebbian connections to form aspect graphs, thereby representing the 3D structure of the object. (3) Visual computations should be layered to achieve increasing stability and level of abstraction. (4) Visual features should be selected automatically and task-specific during the learning process and not a priori. (5) Any applied learning mechanism has to account for variable view-based object representations and should be able to increase the amount of resources on an object-specific basis. Starting from these aspects, we develop a conceptual framework and describe a possible realization. For the latter, we show that the required characteristics can be realized by three central building blocks: the learning mechanism incremental Generalized Relevance Learning Vector Quantization (iGRLVQ) (Kietzmann et al. 2008), Hebbian connections of view-prototypes (Tarr and Bülthoff 1998) and object cells. Taken together, these three main components cover the five required aspects. To verify the effectiveness of the approach, the system's recognition performance was tested together with a variety of related effects such as the semantics of feature selection, rotation invariance, Hebbian connections of prototypes and the automatic creation of view- and object cells.

2 Human object recognition

Problems of artificial vision systems are known to arise from a variety of sources, including changing illumination, occlusion and object variability. Because of the human superiority in this task, it is sensible to let the creation of artificial systems be inspired and guided by biological findings. In the following, we will highlight some central questions, that arise when investigating biological and computational object recognition systems. The first issue which has to be addressed is the question of how to store knowledge about objects. Given the present understanding, this question comes down to following either the view-based approach or the 3D model-based approach. Given that the decision is to follow the view-based approach the second decisive question is how to integrate 3D information from a representation based on only 2D views. The third question arises from the need to assign the same identifying label to all the various perspectives and views of an object. How can consistent and view-invariant object recognition be achieved from lower-level representations? Moreover, there is a great diversity of possible visual features which can be used by humans as well as computational systems. Do we need all possible features or is a small set of decisive measurements sufficient? This issue is covered in question four. Another question is whether each object should have a model of constant or variable representations and complexity. An answer requires a closer look at object models. The following list shortly summarizes these questions.

Question 1: *How should knowledge about objects be stored? View- or object-centered?*
Question 2: *How can temporal and 3D information be integrated with the view-based approach?*
Question 3: *How can view-invariant representations be built from simple visual features and object views?*
Question 4: *Should all possible features be used?*
Question 5: *Should the complexity and representation of each object model be variable or fixed?*

In order to find possible answers, we review aspects of neurophysiological and neuropsychological object recognition strategies in humans which we find to be essential for successful object recognition. The described evidence forms the basis of the proposed framework.

Aspect 1: Object representations are stored view-based

There are two major streams in the literature dealing with mental representation schemes of objects, i.e. view- and object-centered. The former assumes objects to be represented and recognized by referring to their 2D views under

which they can be seen by an external viewer. The latter expects objects to be stored in the form of a single and more abstract 3D model, which can mentally be turned in order to be mapped to the perspective currently seen. Although not totally uncontroversial, there is increasing evidence speaking in favor of the view-based approach. Single cell recordings of Perrett et al. (1987) revealed some neurons in superior temporal sulcus (STS) responsive to certain perspectives of faces. More view-centered cells selective to faces were found by Tanaka (1996). Located in anterior inferotemporal cortex (AIT), these cells are arranged in overlapping columns such that neighboring views are encoded by adjacent cell structures. Today, neurophysiologists believe that object views are represented by groups of neurons, each responsive to a collection of visual features (Abbott et al. 1996; Young and Yamane 1992).

On a more abstract level, objects are represented as view-invariant. In a study by Logothetis et al. (1995) investigating inferotemporal cortex (IT), view-selective cells were found together with a smaller amount of neurons, responsive to an object being present independent from the current viewing-perspective. Further support for these object-encoding cells was provided by Perrett et al. (1991) and Booth and Rolls (1998). In addition to supporting a view-centered theory of object-representations, these findings can be seen as evidence for the assumption that object centered behavior is formed by integrating information from view-dependent cells (Massad et al. 1998).

Psychophysical evidence speaking in favor of view-centered behavior in humans was put forward in (Tarr and Pinker 1989; Bülthoff and Edelman 1992; Tarr and Bülthoff 1995; Wallis and Bülthoff 1999). In a prominent study, Bülthoff and Edelman (1992) used artificial, computer-generated 3D objects shown from two oscillating perspectives. Recognition was tested on static views which lay either inside (intro condition) or outside (extra condition) the trained ones or orthogonal to the trained meridian (ortho condition). The resulting performance pattern of the subjects clearly supports view-based representations and proved incompatible with object-centered schemes. For more detailed reviews on object representation schemes, see Tarr and Bülthoff (1998) and Riesenhuber and Poggio (2000). For the described reasons, we follow the view-based theory of human object recognition.

Aspect 2: Co-occurring object views are interconnected to form the overall object representation and to improve efficiency

The apparent question of how to integrate 3D information arises from the assumption that object representations are based on 2D views. As an answer, theorists propose the
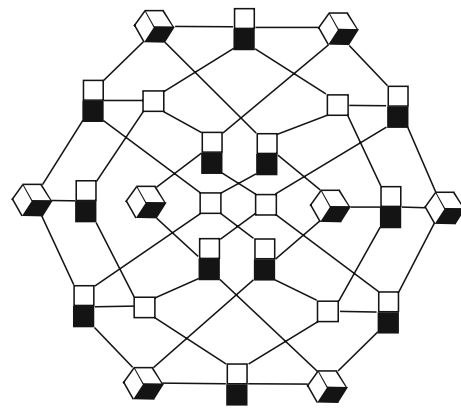

Fig. 1 An aspect graph is created by interconnecting adjacent views of an object. This way, the graph represents the object's three-dimensional structure (adapted from Luong Chi)

use of temporal associations, which connect co-occurring object views. In the case of object recognition, this is sensible because object views are often seen in rapid succession (Edelman and Weinshall 1991; Wallis and Bülthoff 1999). The resulting interconnections of adjacent views can be interpreted as a graph being able to resemble the 3D structure of the object (Tarr and Bülthoff 1998). As put by Poggio and Edelman (1990), "having enough 2-D views is equivalent to having its 3-D structure specified" (p. 263). This representation scheme is also known as aspect graph (Fig. 1).

There are several psychophysical and neurophysiological experiments providing scientific evidence for temporal correlations of object views. Single cell recordings in IT of macaque monkeys, as performed by Miyashita (1988); Miyashita (1993) and Sakai and Miyashita (1991), showed that neuronal associations could be built on the basis of temporal connectedness regardless of geometric similarity. Psychophysical evidence was put forward by Wallis (1996, 1998) and Wallis and Bülthoff (2001). The latter performed an experiment in which subjects were shown sequences of faces in which the identity shown changed during rotation. As predicted by the temporal associations theory, subjects exhibited the tendency to treat the different views as a single person.

In addition to being able to represent the 3D structure of objects, there is an additional advantage of connecting object views. As was shown in various experiments (Erickson and Desimone 1999; Vuilleumier et al. 2002), the use of visual associations of object views enhances efficiency. As an effect, response times decreased when subjects were primed by different viewpoints of the same object and also by associated, but different stimuli. In both experiments, expectancy was shown to be able to enhance performance.

The positive effects on efficiency were also described by Bar (2003), who suggested that the connections could be used to reduce the number of prototypes that need to be compared

with the current input. The prior expectations would thus reduce search-space by offering an 'educated guess' (Bar 2003). Finally, some out of all possible views occur more often and thus allow for easier and faster recognition than others. These are known as 'canonical' views (Palmer et al. 1981). In order to incorporate temporal and 3D information, we thus identify the second important aspect of human object recognition as the presence of interconnections of object views.

Aspect 3: Computations and representations become more abstract in subsequent levels

An important characteristic of human visual processing is the fact that visual representations become more and more abstract in subsequent areas. Thus, higher level neuronal activity corresponds to more complex features and feature clusters (Mareschal et al. 1999). Most neuroscientists agree on the initial visual processing in the first milliseconds of feedforward hierarchical processing. The resulting picture is also known as the standard model of object recognition (Riesenhuber and Poggio 1999), which has been subject to intensive computational modeling. Starting with the retina and Lateral Geniculate Nucleus (LGN), visual information enters V1 where simple and complex cells are selective to bars of light and blobs to colors. Afterwards, information is passed through the extrastriate areas V2 and V4. Later, it reaches higher cortical areas such as posterior inferotemporal cortex (PIT), anterior inferotemporal cortex (AIT), and central inferotemporal cortex (CIT). Here, the neurons respond to complex shapes, object views and faces. Object-selective neurons are thought to pool activity of view-selective cells. This way, the overall object is represented through the collection of views. Figure 2 shows the corresponding schematic layers together with their biological counterparts. Among

many others, biological evidence for this hierarchical model was put forward by Perrett et al. (1992), where it was shown that response latencies of view-dependent cells are faster than view-independent cells.

Aspect 4: Features are selected according to the task and objects presented

The human visual system is capable of using a great variety of features for object recognition. In addition to color and brightness information and bars of light, more complex and finer grained features exist in higher brain areas, encoding external form and shape information (Milner and Goodale 1993; Tanaka 1992) as well as texture-shape combinations (Kobatake and Tanaka 1994). Even more importantly, neuronal selectivity to visual features ranges from highly localized information to object-centered information. As described in the standard model, this difference in selectivity is a result of increasing receptive field sizes. The existence of different levels of position specificity is especially advantageous with regard to object occlusion. In this case, localized information is highly beneficial because global object features are not extractable. Out of this wide diversity of features and feature combinations, humans are known to select important visual features in a highly task-specific manner (Walther and Fei-Fei 2007; Murray and Wojciulik 2004). In the case of attentional weighting, known from the area of perceptual learning, people rapidly adapt to the respective task by varying attention paid to different features. In addition to emphasizing important features through attentive selection, unimportant ones are decreased or completely disregarded (Goldstone 1998). Further evidence for this highly adaptive way of selecting features comes from the area of object categorization, where shifts towards features that are useful for specific tasks can be observed (Nosofsky 1984).
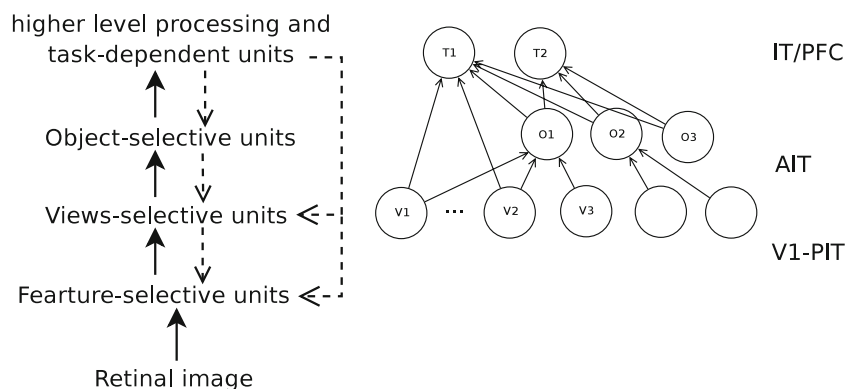

Fig. 2 Representations of objects can be seen on different levels of abstraction in artificial as well as biological systems. Shown are the connections between the different layers and the corresponding brain-structures. Higher level elements and rely on information present in a greater number of lower ones (poolings). The topmost and thus most

abstract layer represents a controlling element, which can influence a priori information used for the decision process and guide possible feature extractions and learning on lower levels. Moreover, it is needed in order to integrate other modalities as well as task-specific information. (The figure was partially adapted from Riesenhuber and Poggio (2000))

In addition to the online selection of task-relevant dimensions, feature selectivity is known to be subject to learning. Neurophysiological evidence was provided by Bichot et al. (1996) who tested the effects of learning on neuronal selectivity in the frontal eye field (FEF). After macaque monkeys were trained exclusively on targets of a certain color the recorded neurons exhibited selectivity corresponding to the trained features. A related effect was found for neurons in AIT, which show increased activity upon behaviorally significant stimuli after training (Jagadeesh et al. 2001). These effects are also known as long-term visual priming (Chun and Marois 2002).

The positive effect of feature selection is clearly an increase in efficiency. As all incoming information competes for limited neural resources, a selection procedure is able to facilitate performance by suppressing irrelevant information. As an effect, task-relevant information is separated from task-redundant information. For instance, Haider and Frensch (1996) performed an experiment in which subjects were able to improve their performance in an alphabetic string verification task by limiting their processing to relevant aspects and dimensions of the task. For the reasons described above, we identify the fourth aspect of human vision: Task-specific feature selectivity.

Aspect 5: Neural selectivity is increased and views are created with learning and task complexity

The neuronal representations of object views allow for the observations that the number of views and the selectivity of the corresponding neurons increase with experience. Evidence for the creation of views comes from Kobatake et al. (1998), who trained monkeys to discriminate 28 complex shapes. After training, there were significantly more neurons selective to views of the trained than untrained objects. Support for the changing selectivity was given by Perrett et al. (1998). In this experiment, selectivity of neurons in temporal cortex adapted to match the features present in the most commonly seen view of an object. As reviewed above, the number and selectivity of neurons is thus assumed to be variable.

3 An artificial object recognition system

Having described the underlying neurophysiological and psychophysical assumptions of the human visual system, we now turn to artificial object recognition. First, we propose a conceptual framework integrating the aforementioned biological aspects, which is, at this level of description, mainly independent of realizations of its different parts. For this, we translate aspects 1–5 to possible technical realizations which are described in the following notions. Note however, that the framework should be seen as a suggestion on what properties artificial systems should include and not as a complete biological equivalent or simulation. A description of a possible realization will be provided afterwards.

3.1 General description

Notion 1: Views are represented through prototype-based systems

As specified in Aspect 1, there is increasing evidence for the view-based representations of 3D objects in humans. Although many of the early computational approaches relied on single 3D and thus object-centered representations (Biederman 1986; Lowe 1985; Marr and Nishihara 1978; Thompson and Mundy 1987), there is an increasing number of view-centered approaches. Among the first authors to employ object views for object recognition were Poggio and Edelman (1990), who used Generalized Radial Basis Function Networks (GRBFs) in order to represent their object models. One of the main assumptions underlying this approach is that the 3D structure of an object can be specified by interpolating between stored 2D views of the respective item. Ullman and Basri showed that any 3D projection of an object can be obtained by a linear combination of 2D views (Ullman and Basri 1991). Thus, matching and recognition of a novel view can be accomplished by interpolating between the stored views. This basic assumption led to a great variety of approaches for 3D object recognition (Roobaert and Van Hulle 1999; Shokoufandeh et al. 1999; Wallraven and Bülthoff 2001b), including the current work.

The view-centered representation scheme forms one of the basic notions of the current description. One of its most important implications is that it is not necessary to store all possible image instances of an object, but only its most descriptive views. As mentioned before, neurophysiologic findings see object views represented by collections of neurons responsive to combinations of features present in each view. Transferred to the computational domain, this corresponds to the prototype-based approaches known from the machine learning literature, where each prototype resembles a specific combination of input features. In this view, interpolating between views is realized by neighborhood relations in feature space (compared to the mathematical transition procedure used by Ullman and Basri (1991)). In the following, the terms *view* and *prototype* will thus be used interchangeably.

In the recent past, various artificial object recognition systems based on prototypical views have shown very promising performance (Jugessur and Dudek 2000; Lowe 2000; Tuytelaars et al. 1999), and thereby provided empirical evidence for the efficiency of this representation scheme.

Notion 2: Prototypes are interconnected by Hebbian connections to form aspect graphs

Because of the potential increase in efficiency and the increasing evidence for the existence of temporal associations between object views (Aspect 2), the use of aspect graphs in artificial systems is clearly reasonable. Since such a structure cannot be known a priori without providing a large amount of external knowledge, an artificial system should be able to automatically extract it from its training data. The temporal order of information in form of view sequences naturally exhibits the underlying structure of the object, i.e. seeing sequences of moving objects over time reveal their 3D structure.

Waxman and Seibert were among the first authors to apply the notion of aspects in artificial object recognition (Seibert and Waxman 1992). Their system automatically constructs aspect transition matrices, similar to aspect graphs as defined by Koenderink and Doorn (1979). Recognition is then performed on the basis of accumulated evidence to find a best-match hypothesis. A different solution to the problem of learning aspect graphs from data comes from the theory of temporal associations. The main idea is to enhance connections of subsequently winning prototypes. Being presented with different image sequences of the same object, this naturally results in a graph in which often co-occurring representations will be strongly connected.

In addition to providing a straightforward method for obtaining an object's aspect graph, only characteristic view sequences are extracted. Thus, the resulting models represent typical ways of motion (Massad et al. 1998). Canonical views can be obtained by storing the total number of activations of the prototypes together with the connections. This way, the views with the most activations can be used before less frequent ones are considered.

Beyond providing a computationally less complex alternative to the object-centered representations, there are further positive effects speaking in favor of the use of aspect graphs. For instance, they can be used in decision making by influencing the a priori probability of a hypothesis. Having seen a prototype A and having a strong connection $H(A, B)$ in the graph implies that the prototype B often co-occurred with A. Thus, the a priori probability or prototype activity for the next classification can be altered to prefer hypothesis B. This is particularly important in the case of 'ambiguity in appearance' (Paletta and Pinz 2000; Massad et al. 1998). This term describes the phenomenon that two different objects can look similar when viewed from a certain perspective. In this case, the resulting classification of a single image is clearly ambiguous. However, integrating previous information can disambiguate the decision and thus improve efficiency and reliability (Bradski and Grossberg 1995; Rao 1997). This effect broadly corresponds to the behavior of cells in the ventral stream, whose responsiveness can be modulated by prior occurrences of stimuli (Goodale 1993); an effect which can be interpreted as a feature memory trace (Mareschal et al. 1999).

Finally, this approach easily integrates with further information and modalities in a sensor fusion process, which is clearly reasonable in the light of real-time robotics. For instance, it would be reasonable to integrate auditory information or data from other 'visual' devices such as laser scanners or additional cameras (Voigtländer et al. 2007) in the recognition process.

Notion 3: Information becomes more abstract in subsequent stages of processing

In most computer vision systems, increasing abstractness is an implicit result of the process of feature extraction. After mapping from the raw pixel space into feature space, features resemble more complex and and mostly more abstract input configurations. Visual features can be manifold. Among many others, present approaches reach from highly localized image features (Lowe 1999), to computational models based on our knowledge of visual cortex (Serre et al. 2005; Mutch and Lowe 2007), which additionally incorporate different layers of processing and abstraction in a layered, feedforward hierarchy, to directly object based features such as area, diameter and color. After features are extracted, which can potentially include several stages of hierarchical processing already, prototypes represent collections of feature values in input space. Their activity can be interpreted as being caused by an object viewed from a certain perspective. By pooling over all view-selective units for one object, the overall object representation is formed (Fig. 2). In this final and object-selective level, objects are encoded invariant from the perspective, position and scale currently seen. In the following, these cells will be referred to as object cells (OCs), whereas prototypes form the underlying view cells (VCs). A computationally related approach worth noting are convolutional neural nets (Lecun et al. 1998). This technique employs subsequent layers of processing in order to yield higher-order features. Starting from selectivity for localized patterns in the input image, representations become increasingly position invariant through subsampling. A final, fully connected layer of the network can then be trained to recognize high-level, position invariant patterns, such as handwritten digits or objects.

Notion 4: Features are selected online according to the underlying tasks and requirements

Extractions of features are computationally very expensive. Moreover, problems arise from the presence of great variability in visual data. If a set of calculations is useful in

a certain task or situation, it is often the case that it turns out to be useless in others. For these reasons, the system should be able to automatically select relevant from irrelevant features in a highly task- and situation-specific fashion. Even more important, a selection should be performed during the learning process. With this capability, it is possible to significantly increase efficiency by rejecting unnecessary features and thus computations. Furthermore, a task-specific and online selection of useful features allows for a broader area of application. This is because an a priori specification of a fixed input space/feature set is specifically done for a certain task. However, starting with a more general feature set and then iteratively selecting the right dimensions gives greater variability and a greater range of feasible situations.

Notion 5: Incremental methods are able to account for varying task and object complexity

As described in Aspect 5, the number of views and neuronal selectivity change with increasing expertise and task complexity. For two reasons, this aspect is especially important for artificial systems. First, efficiency can be increased by assigning only the minimum number of required resources to each object. This is achieved by starting with a very small number of prototypes. By adding prototypes on demand, very sparse models are created. Applying this principle to the use of object views leads to the observation that simple objects are often able to be represented by a single view, whereas more complex ones often need considerably more views in order to allow for a complete representation and consequently good recognizability. Second, the underlying input space naturally varies when rejecting parts of it. As was shown by Kietzmann et al. (2008), the number of clusters formed by the object instances varies together with input space. Consequently, when considering online feature selection, as in the currently proposed framework, the required number of prototypes can vary for one and the same object. The use of incremental methods enables the system to adapt to these changes in a straightforward manner by automatically adding prototypes if objects turn out to be misclassified repeatedly.

Moreover, changing the feature space demands variable rather than stable prototypes. This way, the existing prototypes are able to change their selectivity in order to account for new evidence or changing input space after having been put into the system. In short, moving prototypes in feature space accounts for changing selectivity whereas additions of prototypes are expected to lead to increased selectivity.

Dealing with image sequences instead of single object instances is of great importance to artificial vision systems especially in the case of real-time robotics. Temporal associations can only be extracted from temporally connected visual information and an integration of subsequent evidence can support a more reliable recognition. We propose the

following procedure: The currently winning prototype is always stored in the form of a 'keyframe', which was also the procedure in (Wallraven and Bülthoff 2001b). If the active keyframe of a previous image is also the most appropriate prototype for the next image in the sequence, it is adjusted according to a learning rule, its winning count is increased and a feature-selection procedure is run. If a different prototype is more appropriate for the new image, it is set to be the current keyframe and is adapted. Whenever the keyframe changes, the Hebbian connection of the underlying prototypes is increased. In order to be able to add prototypes to complex classes, the best matching prototype of a different class (negative prototype) is retrieved together with the best one of the same class (positive prototype). The case that the negative prototype is performing better than the positive one can also be interpreted as a case of misclassification and an error-term is increased. Whenever the total number of errors exceeds a selected threshold, signaling that still too many errors occur and that the current resources are not sufficient, new representations for problematic objects are added to the system (Fig. 3). With this procedure, the system is able to dynamically select the appropriate number of prototypes for each object.

Summing up, the currently proposed framework requires object representations to be altered by updating existing and adding new prototypes. Hebbian connections are used to form aspect graphs and OCs provide a view-invariant object representation. Features should be selected as part of its learning procedure. Taken together, the system is expected to be able to provide directly task dependent solutions for both, prototypes and feature set. Successive updates of view-connections further improve stability and produce structural representations of the underlying objects. So far, the actual learning and feature selection mechanism were left unspecified. As an overview, Table 1 shows a direct comparison of biological aspects and computational notions. The current description is meant as a guideline on what properties are expected to improve performance when included in artificial object recognition systems. Still, the actual realization of these elements can be chosen independently.

3.2 A possible realization

To allow for a detailed evaluation of the approach, a possible realization of the described notions was implemented. The first central component of the system is the learning mechanism. Here, we chose to use iGRLVQ, which was recently introduced and algorithmically evaluated (Kietzmann et al. 2008). Additionally, we use a Hebbian learning procedure to automatically extract aspect graphs (second component) and introduce a more abstract object representation in form of object cells (third component). In the following, we will give a more detailed description of the realization including
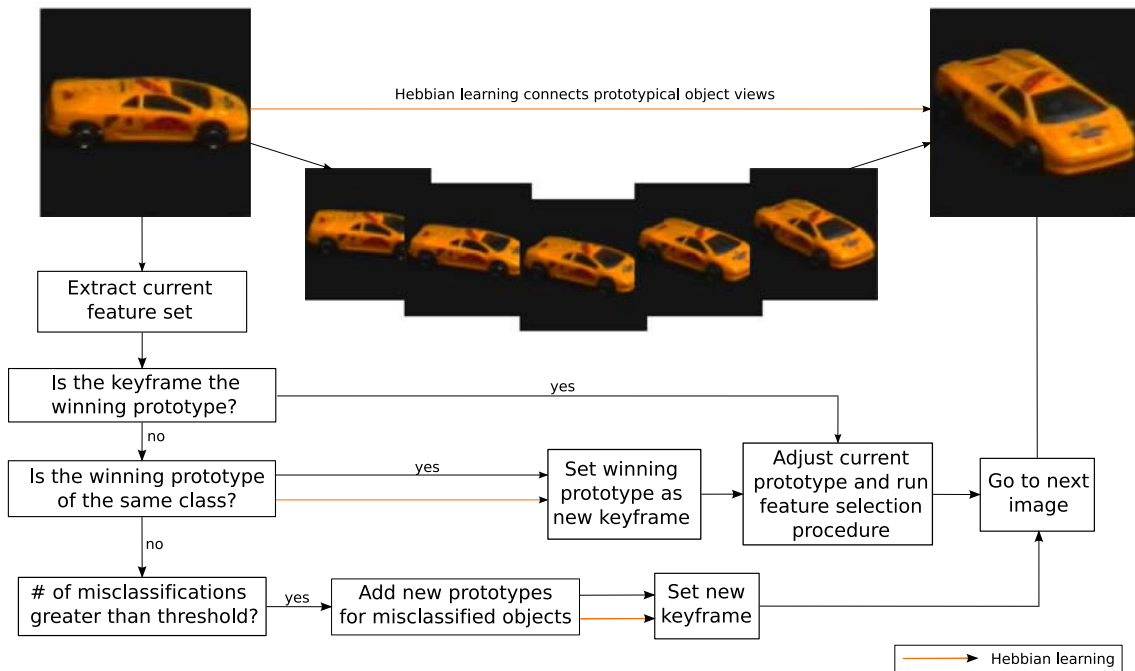
Fig. 3 A modular scheme of the framework showing the main learning process and handling of image sequences. At this level of description, it is still independent from a possible realization

Table 1 A direct comparison of the reviewed biological aspects and suggested computational notions

Biological aspect	Computational suggestion
View-selective cells represent objects	Prototype based approaches resemble object-views
Co-occurring views are interconnected	Hebbian connections create aspect graphs
Representations become more abstract in subsequent levels of processing	Prototypes react on combinations of lower level features
Perceptual learning and feature attention strengthen the influence of relevant features	Features are selected during the learning process to find task- and situation-dependent solutions
The number of views increases and neuronal selectivity changes with learning and task complexity	Incremental learning methods changing the number and selectivity of prototypes

a description of iGRLVQ, OCs, the underlying feature space, the feature selection method and the Hebbian learning procedure.

3.2.1 Prototype updates via iGRLVQ

As specified in Notion 1, prototype-based methods form an excellent combination with view-based approaches. In order to be able to deal with visual data, the highly redundant image data presented in form of raw pixels has to be transformed into points in input space. The system's prototypes, which can be seen as (labeled) points in the same feature space, exist for every learned object and are stored as vectors in a codebook. When being presented with an object to be recog-

nized, the best matching prototype is used to assign the label. Thus, the goal of a learning procedure is to adjust the prototypes of each class such that they represent it as accurately as possible.

In particular, learning vector quantization (LVQ) is very appealing because of its straightforward update rule and good generalization properties. Recently, we proposed the learning method iGRLVQ. In addition to incrementally adding prototypes to the codebook, it also allows for an elegant way of feature selection, as illustrated below. Prototype learning is achieved by using each training vector to attract the closest prototype of the same class (w^J) and push away the closest one of a different class (w^K) according to $w_{new}^J = w_{old}^J + \Delta w^J$ and $w_{new}^K = w_{old}^K - \Delta w^K$ respectively. The

update elements are defined as:

$$\Delta w^J := \epsilon \times \frac{d_K}{(d_J + d_K)^2}(x^i - w^J)$$
$$\Delta w^K := \epsilon \times \frac{d_J}{(d_J + d_K)^2}(x^i - w^K) \qquad (1)$$

d_J and d_K describe distances between the codebook and the training vector. With this rule, the prototypes are lead into parts of feature space where they best resemble their corresponding class instances (Notion 5).

For the update calculations, LVQ relies on standard Euclidean distance. However, it is not necessarily the case that this particular metric is suitable for the respective learning task. Because every dimension is regarded equally, the data has to be preprocessed such that the input dimensions have approximately the same magnitude w.r.t. the classification. In order to avoid these problems, relevance learning applies adjustable weights $\lambda = (\lambda_1, \ldots, \lambda_n)$ to the input dimensions, which are altered during learning together with the prototypes. When calculating distances between prototypes and input patterns, the differences in the corresponding dimensions are scaled by these relevances (Eq. 2). This way, dimensions with higher weights are emphasized whereas others become less pronounced. Following the main update principle for prototypes, relevances are adjusted online according to the learning rule provided in Eq. 3. This rule can be interpreted as follows: On correct classification, the update increases the weight terms of dimensions in which the training data is close to the positive prototype, whereas relevances for terms with greater distance are decreased. In contrast, on false classification the more distant dimensions are increased and the closer ones weakened. As a result, the mechanism facilitates dimensions which contribute to the right classification and which do not contribute to a false one. As will become evident later, an important advantage is that the resulting relevance terms can be used for feature selection.

$$d_J = \| x - w^J \|_\lambda^2 = \sum_{i=1}^n \lambda_i (x_i - w_i)^2 \qquad (2)$$

$$\lambda_m := \lambda_{m-1} - \epsilon_1 \times \left(\frac{d_K}{(d_J + d_K)^2}(x_m^i - w_m^J)^2 \right.$$
$$\left. - \frac{d_J}{(d_J + d_K)^2}(x_m^i - w_m^K)^2 \right) \qquad (3)$$

Activity of prototypes i belonging to class j, a_{ij} is computed according to the Winner-Takes-All (WTA) principle. Thus, only the most appropriate prototype becomes active, whereas all others remain silent. When being presented with an input pattern x, the distance to each prototype is calculated and the closest one becomes active (Eq. 4).

$$a_{ij} = \begin{cases} 1 \text{ if } & i = \text{argmin}_i \| x, w_i \|_\lambda^2 \\ 0 \text{ else} \end{cases} \qquad (4)$$

Standard methods use the same fixed amount of prototypes for every class. As described in Notion 5, however, the required number naturally varies with the complexity of the object and with the input space, as it is in the case of online feature selection. A solution to these issues was provided by using incremental methods. Having already proven to be quite effective for monitoring technical systems (Bojer et al. 2003) and for object recognition in a biologically motivated approach (Kirstein et al. 2005), prototypes are successively added for classes, which get misclassified repeatedly after starting with only one prototype for every class. As a result, every object receives the optimal number of prototypes needed for successful recognition.

3.2.2 Object cells

VCs encode object information implicitly by being associated with a defined object. To achieve an explicitly view-invariant object model, a more abstract representation is needed (Notion 3). Object cells respond to the presence of objects independent of the current view and define their activity by pooling from the underlying VCs' activity. Thus, whereas VCs are embedded in feature space, OCs are superimposed (Fig. 4).

In detail, each OC gets input from all VCs belonging to the same object. Since we are dealing with a WTA network of view-cells in the subjacent layer, an obvious way of modeling such behavior is to activate an OC whenever one of its associated VCs becomes active. This is achieved by taking the maximum VC activity as activation for the OC (Eq. 5). This procedure was also biologically motivated and applied in the HMAX approach (Riesenhuber and Poggio 1999, 2003).

$$OC_k = \max_i a_{ik} \qquad (5)$$

3.2.3 Feature set

As we apply a prototype-based algorithm, the only constraint lies in the need for a fixed dimensionality of input space. As shortly noted above, visual features can be extracted from complete, segmented object images (global) or in a more localized manner dealing with smaller image patches (local). Local image features, such as SIFT (Lowe 1999), which transform small image parts into a high dimensional feature space, have become very popular in the computer vision community. The clear benefit of these approaches lies in potentially high recognition rates and invariance to changes in scale and rotation. Nevertheless, the number of features found in an image is not fixed and can vary from image to image such that a classification with procedures being dependent on a fixed dimensionality of input space, such as the current one, are not possible in a straight forward way. Moreover, highly localized features are expected to be less efficient in
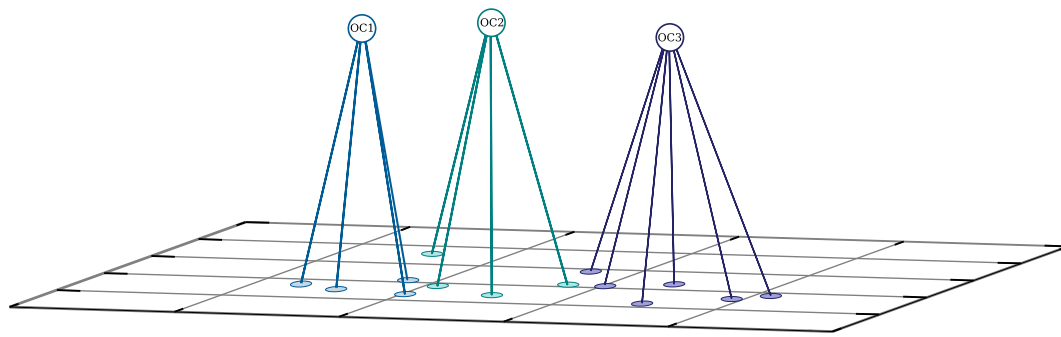
Fig. 4 VCs are embedded in feature space whereas OCs are superimposed. Each OC receives input from the underlying VCs belonging to the same class. This way, a more robust and abstract representation is formed

tasks requiring a high level of generalization. On a more abstract level are intermediate image features such as the ones put forward by Ullman et al. (2002). In their approach, the most informative image fragments are automatically selected from a training set, each fragment-template then formed one dimension of the resulting feature space. The object recognition system proposed by Serre et al. (2005), which is based on the standard model mentioned earlier, uses a different approach. Their hierarchy of visual features is based on localized image properties, which are integrated in subsequent levels of processing to form intermediated feature levels with more complex and less local features. In addition to the biological relevance and robust recognition rates, the feature hierarchy also gives a fixed dimensionality. Finally, global or object features are on the most abstract level.

Out of the great variety of possible feature sets, two possibilities were selected for the experiments. First, a generic set of object features is used which deals with global object features. Despite their great performance and expressibility, a disadvantage of these features lies in the need of image segmentation. Although promising approaches exist, segmentation is still a highly discussed and yet greatly unsolved issue in computer vision and neurosciences. We greatly acknowledge the importance of this problem. However, the work with this first feature set explicitly concentrates on object recognition from already segmented images. For reasons of simplicity, a region-growing algorithm, as proposed in (Adams and Bischof 1994), was applied as a preprocessing step to separate the object from the image background. To solve the problem of the dependence on image segmentation, a second type of feature set was used, which is based on the extended feature hierarchy of Serre et al. (2005), which was proposed by Mutch and Lowe (2007). Among many other advantages of this approach is the fact that the features trained and extracted without the need for a segmentation of the image.

The generic feature set was composed of the following parts. Color information of the pixels was encoded in the YUV color space. In order to include color information, a 32-dimensional Y-luminance histogram, and an UV-color histogram (8×8 bins $= 64$ dimensions) were included. Moreover, the area of the object, its centroid, perimeter, eccentricity, circularity, compactness, and maximum and orthogonal diameter were extracted together with the Hu set of image moments (Hu 1962). The latter are calculated as particular weighted averages of the object's pixel intensities and have the special property of being invariant under translation, scale and rotation, a characteristic which is also thought to support recognition memory in humans (Milner and Goodale 1996). Finally, Gabor wavelets with three scales and four orientations were extracted, which can be seen as detecting specially oriented and scaled bars of intensity in the image. Their usage was suggested by Würtz (1995) because of computational advantages and biological plausibility.

In the second feature set, the visual hierarchy, information is processed through subsequent layers of increasing complexity. For this, localized information is integrated to learn intermediate image features. The hierarchy consists of four layers. S1, C1, S2 and C2. S1. Starting from the image layer, subsequent layers pool their activity from each previous level, which leads to increasing location invariance and more complex features. Activity in S1 corresponds to simple cells in V1, C1 corresponds to their complex counterpart. S2 is calculating information which is expected to be present in area V4 or posterior IT. In the final layer, C2, all position and scale information is removed, this level is representing a "bag of features". In its original version, the output of the hierarchy was used as input for a classical support vector machine. Instead, the current approach extended the hierarchy by adding two layers of VCs and OCs. This way, VCs resemble collections of intermediate features and their activity corresponds to spread activity in the underlying layer C2. As can be seen in Fig. 5, VCs and OCs integrate very well with the standard model of object recognition and the aspect of increasing complexity in subsequent layers. This feature set was used as additional input space for experiments designed to assess recognition performance, as explained in more detail below.

Fig. 5 The extension of the used feed forward hierarchy with our notion of VCs and OCs (partially adapted from Riesenhuber and Poggio 2003). Layer C2 represents a set of scale and position invariant features, which acts as input for the prototypes of the learning procedure.
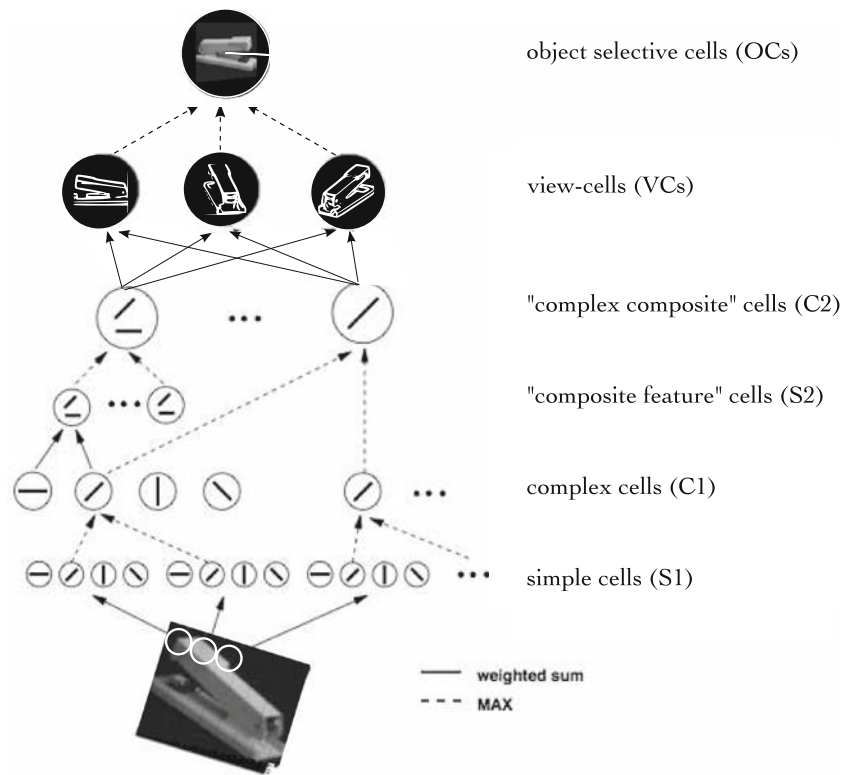
3.2.4 Feature selection

Extractions of visual features are known to be computationally very expensive, which is why the dimensionality of input space should be kept as small as possible. Consequently, selecting the right amount and types of features is decisive for the later performance of the system which is why this decision is highly task-specific and normally done by a human expert. With iGRLVQ, however, this problem can be solved differently. Initially offering a whole variety of standard measurements, the system automatically selects relevant features and prunes irrelevant ones in order to decrease dimensionality. The features initially used can broadly be put into two categories, appearance and shape selective.

The required capability of selecting relevant features and pruning irrelevant ones (Notion 4) is applied for efficiency reasons and is done during learning to achieve better task-dependency. Due to the high variability in visual data, it is, even for humans, very hard to assess which measurements will be useful for a particular data set and in a particular condition. As mentioned before, the assigned relevance terms are updated during learning and enhance important dimensions. As also described in (Strickert et al. 2001), in addition to only diminishing the influence of unimportant dimensions, the system iteratively prunes the weakest dimensions, i.e. the ones with the smallest λ-values, in order to completely exclude them from the following learning process and, even more importantly, from future calculations. In order to stop

pruning automatically, we suggest the usage of a validation set. If recognition performance on this set drops, pruning is stopped. One of the main advantages of using relevance terms as a basis for pruning is that it can be done online, making it an integrated part of the overall learning process. Hence, feature selection and model learning cannot be seen as two distinct processes in this setting. Both work interactively in order to solve the current task and to adapt to the individual situation.

3.2.5 Hebbian learning

In order to account for temporal associations of prototypes and even more importantly to represent the inherent 3D structure of the objects, we use a mechanism comparable to Hebbian learning (Notion 2). Because learning is based on image sequences, it is possible to keep track of the currently active prototype. If it changes, the connection-weight $H(w^r, w^s)$ between the old w^r and the new prototype w^s is increased according to:

$$H_{\text{new}}(w^r, w^s) = \tau H_{\text{old}}(w^r, w^s) = \tau H_{\text{old}}(w^s, w^r) \qquad (6)$$

τ is the learning rate. This simple and computationally very efficient approach is able to successively extract aspect graphs during learning.

The current implementation is a prototype- or view-based approach. It uses a generic set of visual features together with an automatic pruning algorithm. Because feature selec-

tion is based upon relevance learning of iGRLVQ, the system can reduce the input space as part of its learning process in order to speed up future computations. Prototypes are connected through Hebbian connections and OCs receive their input from associated VCs to form a coherent object representation. An incremental mechanism is applied in order be able to deal with varying numbers of clusters and to create sparse object models. Considering the number of parameters, the approach is able to find the correct number of prototypes and input dimensions and selects the most important features. This significantly reduces the amount of external knowledge, which has to be put into the system.

4 Empirical Evaluation

In order to test the effectiveness and behavior of the current approach, multiple experiments have been carried out. In addition to recognition performance, which is clearly most interesting from a technical point of view, the current work also includes an empirical evaluation of the system's behavior in comparison to the characterized biological findings. The first section examines the system's performance, which is compared to other state of the art methods. Afterwards, an elaborated investigation of the system's properties with regard to its biological counterparts is provided. In a different setup, the automatically chosen number and the selectivity of the prototypes are analyzed and compared to non-incremental methods. Moreover, some instances of the learned aspect graphs are presented. The final experimental setup explicitly deals with the feature selection mechanism, which is tested in a standard and a one-vs-all setting. Finally, the rotation invariance of the VCs and OCs is tested and compared, providing evidence for the increasing level of abstraction and the effectiveness of the view-based prototype learning. An overview of the performed experiments together with the system's behavior and corresponding biological effects is provided in Table 4.

4.1 Recognition performance and generalization capabilities

Because the goal of the current work is to suggest a way of creating high-performance object recognition systems inspired by biological equivalences, the resulting recognition performance is clearly one of the most important components in the evaluation of the overall approach. In detail, the system's object identification performance is tested on the COIL100 database (Nene et al. 1996) and the CSC-LAB image database Murphy-Chutorian et al. (2005). The first provided image sequences needed for the extractions of aspect-graphs. The second data set introduces object images with heavy occlusion and varying illumination, scale and viewpoint. Occlusion is one of the hardest problems with which artificial recognition systems have to cope.

4.1.1 COIL100

The data set includes 360° image sequences of 100 objects. In each subsequent picture of an object, it is turned by 5° such that there are 72 images for every object in total. The experimental setup was the same as in (Obdrzalek and Matas 2002; Schneider et al. 2004), 4 (8, 18 and 36) images of each object were used for training, and performance was tested on the remaining ones forming the test set. Thus, in the case of 18 views per object, the training set included 1,800 data points and recognition performance was tested on the remaining 5,400 vectors. Some of the stimuli of the database are shown in Fig. 6, resulting accuracies, adapted from Kietzmann et al. (2008), are shown in Table 2. Our biologically motivated approach is clearly able to compete with other state of the art recognition systems even when drastically reducing the dimensionality of the underlying feature space. The generic feature set performs slightly better than the localized feature hierarchy. The latter does not include any color information. A possible explanation for the differ-

Fig. 6 **a** Some examples of the images in the COIL100 database. **b** Labelled example scene from the CSCLAB database
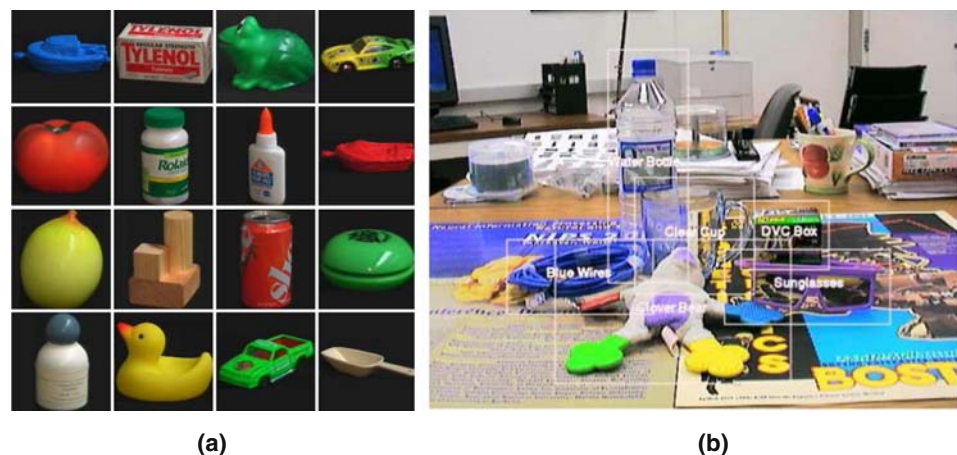
(a)　　　　　　**(b)**

Table 2 Recognition performance of various approaches based on the COIL100 database

Method	# Training views			
	36 (%)	18 (%)	8 (%)	4 (%)
iGRLVQ Generic (137)	99.3	97.9	92.6	85.4
iGRLVQ Generic (50)	99.2	97.7	92.0	82.9
iGRLVQ Hierarchy (500)		96.5		
iGRLVQ Generic (20)	98.5	96.3	88.9	76.3
PCA & NN[a]	98.2	96.5	–	–
Spin-Glass MRF	–	96.8	88.2	69.4
iGRLVQ Hierarchy (250)		95.9		
iGRLVQ Hierarchy (100)		94.8		
iGRLVQ (10)	97.0	93.8	85.3	69.2
Linear SVM	–	91.3	84.8	78.5
Nearest Neighbor	–	87.5	79.5	74.6

The numbers in parentheses of the generic feature set correspond to the resulting dimensionality of feature space after pruning
[a] Results on 40 of the 100 objects using the first 12 PCAs

ence in performance is that color information is very beneficial in this data set. This interpretation is also supported by our feature-selection experiments with the generic feature set. which showed that dimensions dealing with color information are among the most relevant ones.

4.1.2 CSCLAB image database

Because the COIL100 database does not contain cases of occlusion or varying illumination and scale, the approach was tested on a second, more difficult database. The CSC-LAB image database consists of images of 50 objects in ten highly cluttered scenes with heavy occlusion and changes in illumination, scale and perspective. Exemplary stimuli are shown in Fig. 6. Together with the object scenes, the database provides one binary mask per image separating the scene's foreground from its background. This mask, however, does not include information about the separation of the individual objects in a scene. As proposed by Murphy-Chutorian et al. (2005), the images with only one object present and half of the object images from the cluttered scenes were used for training while the other half was used for testing. The resulting test set thus only included images of occluded objects. Scenes for which binary masks were present were used in the experiments (967 for training and 490 for testing purposes, 1,457 in total). Although the database is comparably difficult, the system was able to achieve 80.53% accuracy on the test set (Table 3). Despite the promising results, the degraded performance, as compared to the COIL100 results, shows the limitations of the system when dealing with occlusion. An explanation for this is the used generic feature set. Global features fail to capture

Table 3 Recognition performance of the approach using different types of features

Feature set used	Test accuracy (%)
Generic feature set	80.53
Feature hierarchy 100	92.04
Feature hierarchy 250	93.88
Feature hierarchy 500	94.90

Whereas the generic feature set is based on global features, the feature hierarchy includes localized information. In the case of occlusion, as in the CSCLAB database, localized information greatly improves recognition performance

the localized information needed to successfully deal with only partly visible objects. This assumption, which implies that localized information is needed to cope with occluded objects, was verified by experiments in which an extension of the feature hierarchy following the standard model (Mutch and Lowe 2007). As expected, running experiments on the same training and test sets increased performance dramatically up to 94.9% (Table 3).

4.2 Amount and selectivity of prototypes

The number of object views is known to increase with learning progress and supports recognition with increasing task complexity. In the same manner, the algorithm is able to recruit different numbers of prototypes depending on the given task during the learning procedure. When dealing with easy tasks, no further recruitment of additional prototypes is needed in order to achieve good results. However, with increasing difficulty, the system automatically adds more prototypes for complex objects in order to compensate for the increased demands. The resulting possibility of the incremental method to create sparse object models can very well be seen when taking a closer look at the resulting numbers of prototypes. Whereas the most simple object in the data set (onion) was able to be represented with only one prototype, more complex objects (e.g. hook) needed considerably more resources. Exemplary selectivity of prototypes can also be seen in Fig. 9. Generally speaking, because iGRLVQ starts with only one representation per object and successively adds resources on demand, the average number of prototypes is significantly smaller as compared to standard methods. The latter was explicitly tested in a prior experiment (based on the generic feature set) in which iGRLVQ was tested against the non-incremental variant GRLVQ with different numbers of prototypes (3, 6, 9, 12 and 15) assigned to each object. Both approaches were tested on 18 different training sets, which were chosen such that every image was only once included in a training set. Although iGRLVQ uses on average only 3.17 prototypes, it performs significantly better than GRLVQ
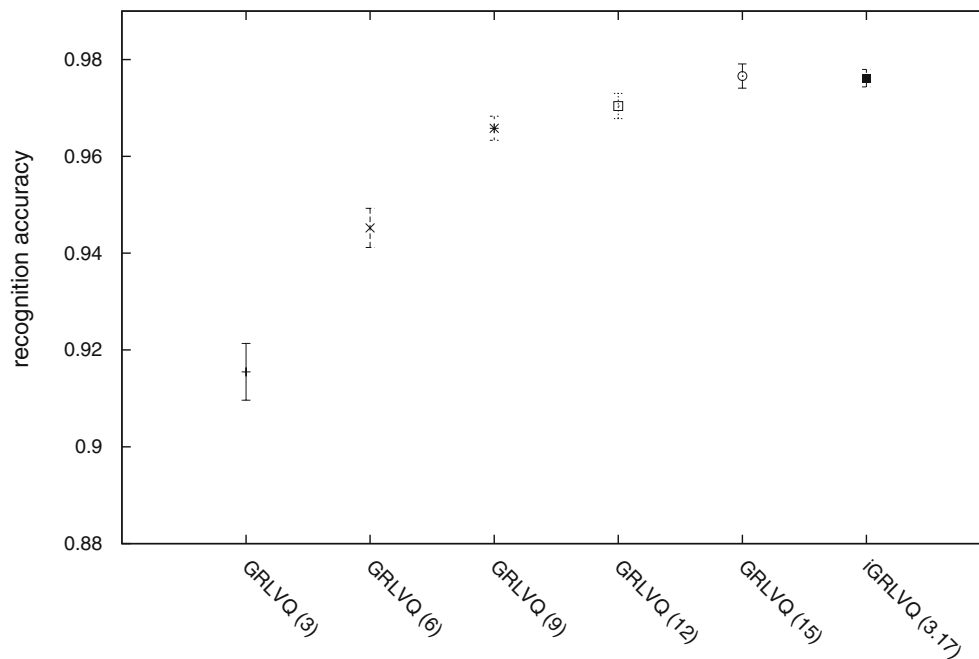
Fig. 7 The average recognition accuracy and standard deviation of GRLVQ with 3, 6, 9, 12 and 15 prototypes per object compared to iGRLVQ (which used only 3.17 prototypes on average)

when equipped with up to 12 prototypes per object (Wilcoxon signed rank test with $p < 0.001$). Only when being equipped with 15 prototypes per object, which is close to the number of training patterns, was GRLVQ able to compete with iGRLVQ leading to a non-significant difference (Wilcoxon signed rank test with $p > 0.49$). As important as the average recognition performance is the resulting variance. Figure 7 shows the recognition accuracy together with the corresponding standard deviation. Results of iGRLVQ proved to be least variant. The degree of improvement becomes especially clear when comparing iGRLVQ to GRLVQ (3). Although nearly the same amount of resources was used, results of accuracy and standard deviation differ remarkably.

An explanation for this remarkable difference can be given by the fact that, with regard to the amount of prototypes, GRLVQ does not differentiate between objects. Because of this, every object has to receive the same amount as the most complex one.

In the current system, the change in responsiveness, which can also be seen in neurons in IT, was accomplished by changing the position of prototypes in feature space. Together with the additions of new prototypes, the system is thus able to find the optimal views needed for successful recognition. Figure 8 shows the training patterns of a complex object and a simple one together with the resulting prototypes. As can readily be seen, the prototypes moved to parts of input space where they could maximize the coverage of instances of their corresponding class. In other words, the prototypes changed their selectivity with increasing experience.

4.3 Learning of aspect graphs

Hebbian connections of prototypes were extracted during the learning process. As a result, frequently co-occurring object views are strongly interconnected. As can nicely be seen in Fig. 9, the resulting aspect graph represents the 3D structure of the object very well. Thus, the computational simulation of creating coherent object views based on temporal associations proved to lead to quite effective solutions, which can be used to infer the 3D structure of the object and improve recognition stability.

4.4 Feature selection performance and semantics

Because it is among the most important capabilities of the current system, the effects of feature selection were tested in more detail. A good feature selection procedure should be able to reduce input space as much as possible, while keeping recognition accuracy comparably stable. The positive effects of online feature selection are twofold. First, efficiency is increased considerably because the pruned dimensions can be completely excluded from the prototypes and therefore from computations and future extractions. Second, the selection of relevant features during learning leads to solutions particularly fitted for the current setting. With regard to efficiency, results are provided in Table 2. As can readily be seen, the feature space could be reduced to about 1/6 of the original dimensionality while still keeping performance very
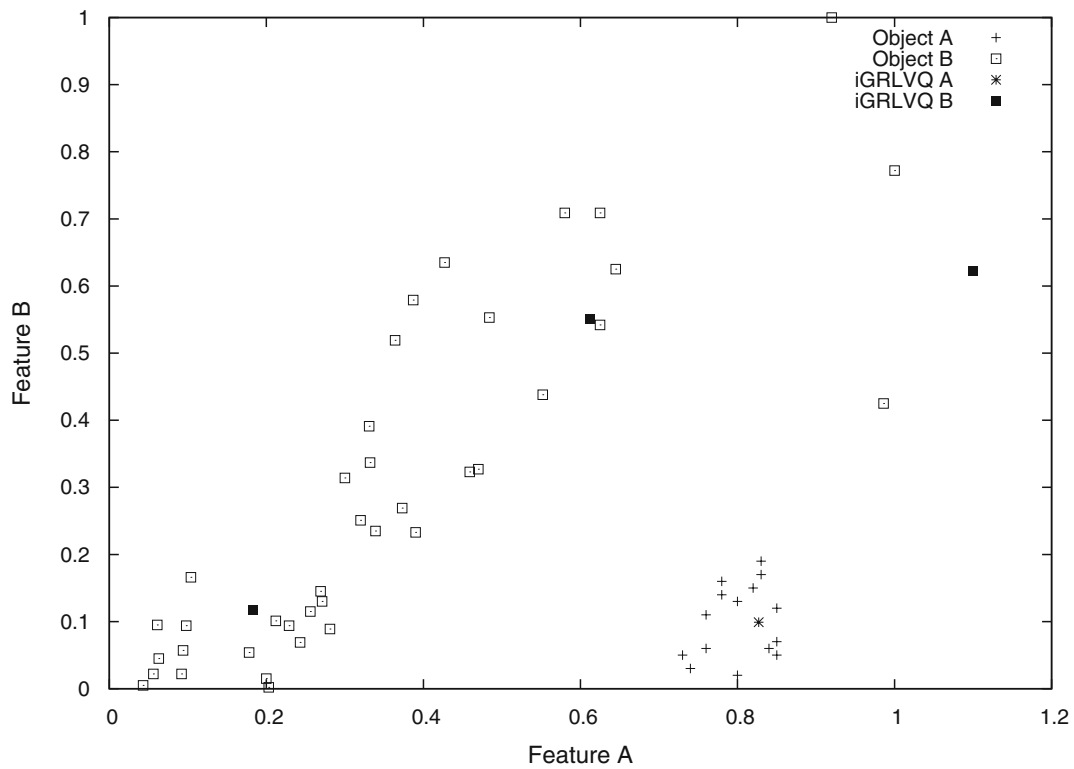
Fig. 8 The *x*- and *y*-axis were selected to be the two most important dimensions of input space. The training instances of a simple and a complex object are drawn together with the resulting prototypes of iGRLVQ. If the currently present amount is not able to cover the complexity of an object, new prototypes are introduced (Object B). However, if the object forms a clear cluster, as in the case of the easy object, a comparably small amount of prototypes suffices (Object A)
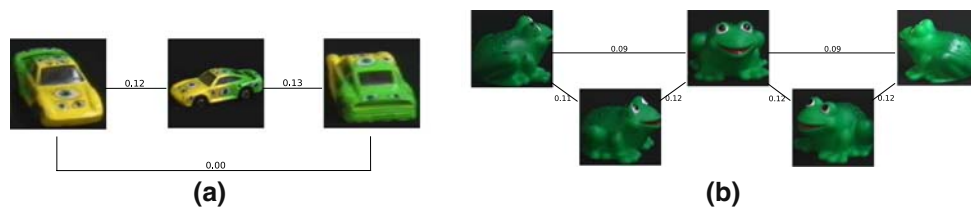

(a)

(b)

Fig. 9 a The three interconnected prototypes of a car. While it often happened that a car was seen from the side followed by either the front or back, it never occurred that a car was seen from front and back subsequently. This is integrated in the graph by a zero-connection. Notice that the canonical view of the car is put to the middle of the graph. **b** The resulting aspect-graph of a more complex representation using five prototypes

high. This can also be seen as proof for the effectiveness of the selection algorithm.

Still, there is considerably more to be examined. In order to test the feature selection semantics, two additional experiments were conducted. In the first, subsets of objects were formed according to two conditions. In the same shape condition, the sets contained only objects of similar shape (boats), whereas the same color condition examined objects with comparable color (wood, blue and red). The different training sets are shown in Fig. 10. Naturally, relevant features vary with the training set and task. Thus, the feature selector was expected to give emphasis on color in the first condi-

tion and to diminish relevance of color information in the second. Because these feature expectancies were able to be determined in a straightforward manner, this was used to verify the semantics behind the selection as well as the system's capability of finding task-dependent solutions. The system worked as expected in both conditions. In the first, all shape information was pruned and only color information corresponding to the colors in the training data was kept. In the second condition, in which the color of all objects was similar, color information was greatly diminished and shape-selective features were emphasized (Fig. 11). Notice that in the latter case, the color information present corre-
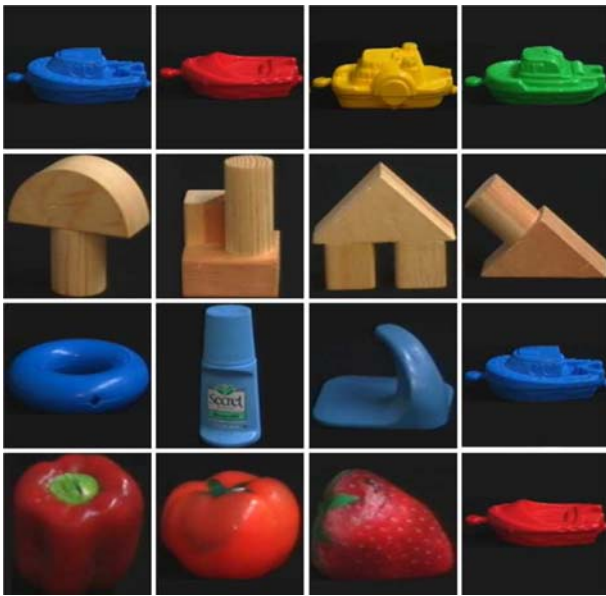
Fig. 10 Instances of the training sets. Objects in the same shape condition are shown at the *top row*, the *three other rows* show object sets belonging to the same color/different shape condition

sponds to the colors in the training data. If any, only differences in these colors could differentiate the objects.

In a second experiment, the system was trained in a one-versus-all setting. For this setup, all but one object were assigned to the same class whereas the left object formed a second one. The difference between this method and the procedure used in previous experiments is that it results in directly object specific features and not in a feature set best fitting the whole recognition task. Results of this experiment are given in Fig. 12. Regarding color histograms, it is especially remarkable that the importance of the selected colors directly matches the distribution in the object. While the object's main colors were chosen to be highly descriptive, minor nuances and variations were disregarded.
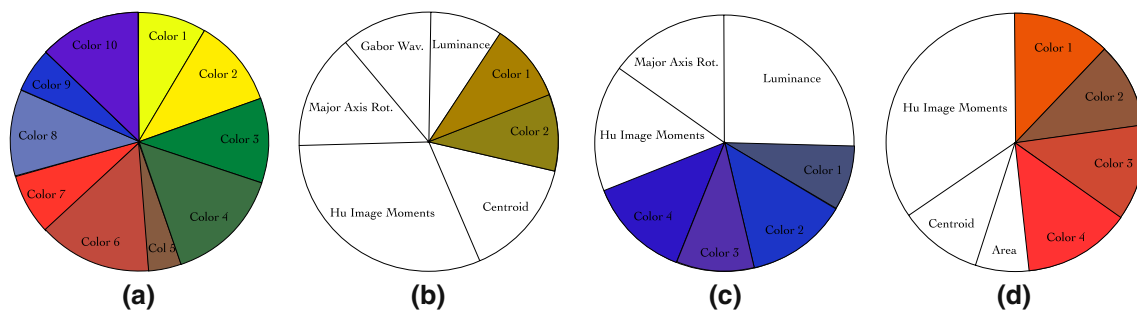
The advantage of having object-specific feature knowledge is that the amount of extractions needed is reduced, which is sensible in a great variety of situations. For instance, when searching for a particular object in an image, as in the case of object detection, it is especially valuable to have a small but highly expressive set of features for which the image can be scanned (Lange and Riedmiller 2006). In a different but related setting, namely the case of feature-based attention, the knowledge of object-specific features is used. As reviewed in (Maunsell and Treue 2006), neuronal activity in visual cortex can be modulated by attention paid to certain features. With this in mind, the selected features can be integrated into saliency maps in order to simulate attention paid to particular features and therefore objects. In yet another but still related case, the system could extract only some very general features and let further calculations be guided by a first guess. Thus, the system first comes up with a broad expectation with further extractions relying on the object-specific features extracted to verify the first assumption. This broadly corresponds to the reverse hierarchy theory (Ahissar and Hochstein 2004; Oliva 2005), where the gist of a scene is processed first in order to guide a later and more detailed analysis.

4.5 Rotation invariance of VCs and OCs

We argued for the importance of invariant object recognition, i.e. an object should be recognized independently of its currently seen perspective. In this regard, rotation- or view-invariance is an interesting feature of VCs and OCs. Testing a biologically motivated object recognition system, this attribute was also used by Einhäuser et al. (2005). Generally, this measurement describes the sensitivity of the cell's activity with regard to the object's perspective, i.e. cells being only active on a certain view of an object are less rotation-invariant and thus more view-dependent than cells with high activity for a greater amount of views. Because we deal

Fig. 11 The pie charts show the resulting feature space after training objects with **a** different color but similar shape, **b** *wooden color*, **c** *blue color*, **d** *red color*. The sizes of the parts are proportional to the relevance values assigned to the corresponding features. Although shape information was completely pruned in the same shape/different color

condition, the relevance values for shape selective features were greatly enhanced in the cases where the colors were selected to be similar. In the latter case, the remaining colors directly resembled the ones present in the training set (e.g *blue colors* in **c**)
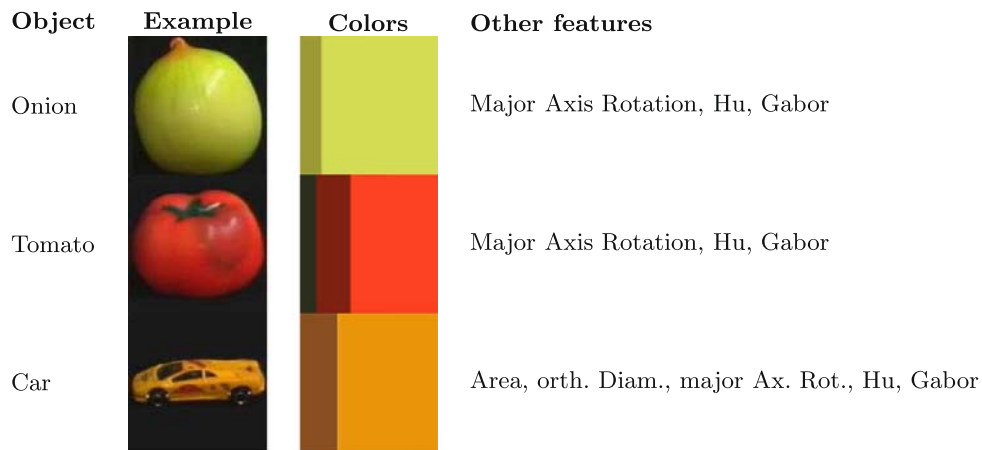
Object	Example	Colors	Other features
Onion			Major Axis Rotation, Hu, Gabor
Tomato			Major Axis Rotation, Hu, Gabor
Car			Area, orth. Diam., major Ax. Rot., Hu, Gabor

Fig. 12 Some exemplary objects together with the resulting features when trained in a one-vs-all condition. In this case, feature selectivity was set to 10, such that the most relevant 10 features were iteratively selected. Still, it is possible to apply a stopping criterion to decide when to stop pruning based on the current performance. This way, each object would also yield its required number of features. Together with an example of the training images, the selected color features are shown. Here, the proportion of the colors in the images illustrates the weight of the corresponding input dimension. Finally, the remaining feature types selected together with the color information are provided, sorted by relevance
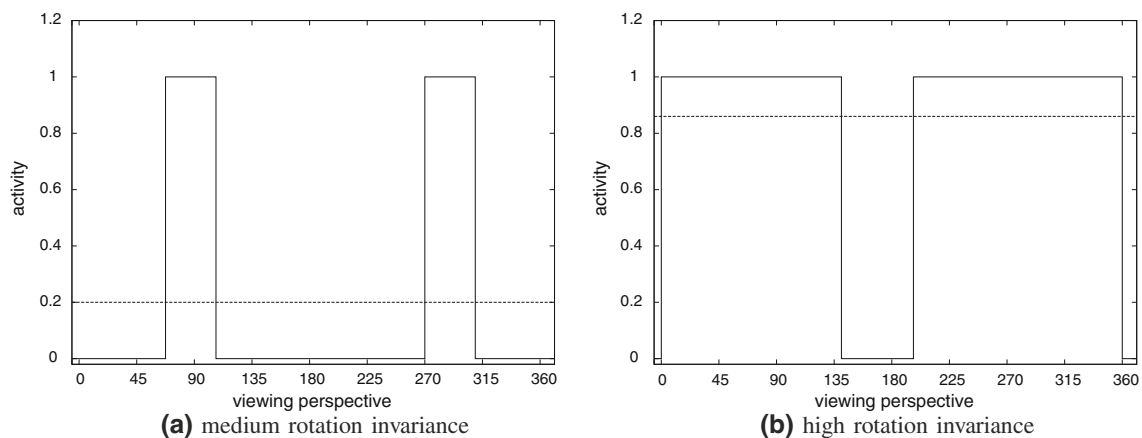

(a) medium rotation invariance

(b) high rotation invariance

Fig. 13 This figure shows examples of the rotation invariance measure used. The more specific a cell reacts on certain views compared to others, the smaller is the value of rotation invariance

with WTA networks and binary prototypical activity, a good measure of rotation invariance is the area below the activation graph (Fig. 13) or simply the number of angles for which the cell is active.

As in (Einhäuser et al. 2005), we train the system with 12 views of 50 objects and probe each VC and OC of our system. The activity of OCs is calculated according to Eq. 5. Because these cells represent objects on a more abstract level, using the maximal activity of the connected VCs guarantees achieving higher invariance than the VCs.

The trained system contained 127 VCs and 50 OCs. Because the number of VCs was selected automatically, there is a variable amount for each object whereas each OC represents exactly one. We compared the average rotation invariance of all VCs versus OCs. For VCs, the average activity and thus rotation invariance is 38% of the whole 360°. As expected, the OC rotation invariance was confirmed to be significantly larger with a value of 98%. This means that about 7° of the possible object views are not covered by the object cells. Obviously, this is equivalent to the achieved recognition accuracy.

These findings clearly demonstrate the system's ability of finding prototypes, which are explicitly responsive to views of objects. Moreover, the proposed definition of VCs and OCs proved to be able to realize different layers of abstraction. Cells in the lower layer correspond to specialized views, whereas cells on a higher level resemble view-invariant object representations.

Table 4 A summary of the performed experiments, observed effects and their biological equivalents

Experiment	Computational effect	Biological equivalent
Number of prototypes	Increasing numbers of prototypes are recruited on demand for complex objects and in complex tasks	The number of views increases with learning and task complexity (Aspect 5)
Selectivity of prototypes	Adjustments of prototypes result in different positions in feature space and thus changing selectivity	Neurons change their selectivity while learning (Aspect 5)
Feature selection semantics	Features were selected to best suit the task at hand. Irrelevant information was pruned in order to increase performance and stability	Perceptual learning, popout effects and visual attention provide evidence for selective feature relevance (Aspect 4)
Feature selection: one-vs-all	The selected features are directly dependent on the underlying object and resemble the input dimensions, which are suited best to distinguish the object from all others	Experiments show that relevant features are selected for object detection (Aspect 4)
Learning of aspect graphs	The resulting aspect graphs directly resembled the 3D structure and typical way of motion of the learned objects	Neurons are known to connect co-occurring visual stimuli such as object views (Aspect 2)
Rotation invariance	VCs exhibit increased rotation variance, whereas OCs form a more abstract and invariant object representation.	Object cells in IT are known to be rotation invariant. However, view-selective cells respond to particular views of objects (Aspects 1, 3)

5 Discussion

The current work proposes a conceptual framework for biologically motivated object recognition systems. Integrating the suggested notions for artificial approaches and being conform to the described aspects of human vision, the implemented system proved the effectiveness and biological plausibility of the approach in various settings. The aim of the current work was not to give a biological simulation of visual processing in primate cortex, but to create a highly variable and automatic object recognition system by relying on some major biological notions and findings. This was accomplished by applying an automatic feature selection procedure together with an incremental learning method, Hebbian connections of VCs and more abstract level of representation in form of OCs. With this combination, the necessary amount of human expertise and domain knowledge could significantly be reduced.

The incremental learning procedure proved to be very effective especially when applied together with feature selection because the optimal number of prototypes cannot be known a priori when dealing with a varying input space. As a result, the system is able to recruit new prototypes on demand when the task at hand turns out to be too complex to be solved with the resources currently available. iGRLVQ is designed

to generalize from the presented data and thus to use the smallest possible amount of prototypes. As indicated, each object is represented by only one prototype in the system's starting state. However, the task of concurrently optimizing recognition accuracy clearly forced the system to create more specialized representations on a lower level, resulting in view-selective cells. Although some objects could successfully be represented by only one view, more complex objects required multiple but specialized cells. This can be seen as computational evidence for the necessity of view-cells in object recognition.

The system's ability to select relevant visual features out of all available ones is clearly one of the most important benefits. Since the feature selection is part of the learning procedure itself, there is no necessity for relearning as in other standard methods such as artificial neural nets (ANN) (Bishop 1995). In addition to the clear computational advantages resulting from the reduction of needed feature extractions, the system is capable of creating highly task- and situation-dependent solutions without the need for additional external knowledge.

Most of the current approaches to object recognition and the underlying feature spaces are handcrafted to suit particular tasks. However, when a more general approach is needed, thus being able to deal with a greater variety of tasks and situations, many systems use very general setups and give solu-

tions suitable for many tasks with the same configuration. This has the clear disadvantage of decreased performance due to lacking task specificity. Our approach is capable of finding task specific solutions while still being able to deal with a great variety of situations and tasks. This substantial advantage is due to its ability to automatically adjust the size of the codebook as well as the feature space.

Limitations of the approach using the generic set of global features were found in cases of heavy occlusion. Object based image features only insufficiently capture local information which is needed to successfully recognize partly visible objects. A second limitation of the generic set of features is the need for reliable image segmentation. As a possible solution to both problems, the output of a feed forward hierarchy of visual features, which integrates localized image features to form intermediate representation schemes, was used as input for the described framework. As can readily be seen in Fig. 5, the feature hierarchy integrates very well with our notion of VCs and OCs and is in line with the standard model of object recognition. In more detail, the prototypes of iGRLVQ react on collections of features from the output layer of the hierarchy. As expected, including localized information in the feature set greatly enhanced performance on occluded objects.

As shown, object identification in cases of occlusion requires highly specific and localized image features. On the other extreme lies the task of object categorization where multiple different class instances have to be classified as belonging to one more abstract class. Although our current description and implementation is explicitly designed to deal with object recognition and not categorization, the system was tested on the ETH80 database (Leibe and Schiele 2003) which consists of ten object categories with ten instances each. Based on the generic feature set, our approach reached 82% in a leave-one-out cross validation procedure, which is comparable to results based on PCA (Leibe and Schiele 2003). Superior, but more task-adapted methods have been published by Leibe and Schiele (2003) and Suard et al. (2006). More interestingly, the localized feature hierarchy, which clearly outperformed the global features in the CSC-LAB recognition experiment, performs significantly worse on this categorization tasks. Testing accuracy of the localized feature hierarchy with 250 features was found to be 72.44%, which is comparable to results of a patch-based approach using SIFT Teynor et al. (2006). This shows the relatively poor generalization capabilities of localized image features. Taken together with the benefits of localized features in cases of occlusion, the results clearly illustrate advantages and disadvantages of the described feature sets. While no feature type exists, which performs well in all tasks, the different approaches were shown to complement each other. Thus, a combined use of localized and more global feature types in a hybrid approach with the capability of selecting the cor-

rect features based on the current task is clearly sensible. Although a more detailed analysis of these approaches is beyond the scope of the current work, the proposed framework and the described learning mechanism for features and prototypes are well equipped to deal with this integration in future work. For instance, an integration of color information and local features could be achieved by adding a color histogram to the input space. Input to the histogram could be provided by color information from image patches around the extracted local features. If global image features are available, which always includes the need for image segmentation, local and global features could both form the overall input space. After training, prototype positions then correspond to collections of local and global features and task-irrelevant dimensions are pruned. However, this approach does not change the relative impact of the different feature types on a case-to-case basis during recognition. A solution which does not imply retraining could be to train two different modules, one based on local features and one on global ones. When a decision is required, each system "votes" for the object identity. The votes are weighted by the certainty of the decision, calculated through the distance of the object currently seen to the winning prototype and the overall reliability of the module.

Additional future work includes a more detailed analysis of the use of aspect graphs and visual context for changing a priori probabilities of prototypes and its application in a fusion process, which are expected to further improve performance of the overall system. Moreover, the system is currently trained offline in batch mode. Nevertheless, an implementation of the described learning procedure in an online system is especially reasonable in the light of dynamics of visual information and a possible application in the area of robotics. Finally, integrating color information in possible feed-forward hierarchies of visual processing is expected to further improve performance and stability of the approach.

References

Abbott L, Rolls E, Tovee M (1996) Representational capacity of face coding in monkeys. Cereb Cortex 6(3):498–505

Adams R, Bischof L (1994) Seeded region growing. IEEE Trans Pattern Anal Mach Intell 16(6):641–647

Ahissar M, Hochstein S (2004) The reverse hierarchy theory of visual perceptual learning. Trends Cogn Sci 8(10):457–464

Bar M (2003) A cortical mechanism for triggering top-down facilitation in visual object recognition. J Cogn Neurosci 15(4):600–609

Bichot N, Schall J, Thompson K (1996) Visual feature selectivity in frontal eye fields induced by experience in mature macaques. Nature 381(6584):697–699

Biederman I (1986) Human image understanding: recent research and a theory. Papers from the second workshop, vol 13 on Human and Machine Vision II table of contents, pp 13–57

Bishop C (1995) Neural networks for pattern recognition. Oxford University Press, Inc., New York

Bojer T, Hammer B, Koers C (2003) Monitoring technical systems with prototype based clustering. European Symposium on Artificial Neural Networks, pp 433–439

Booth M, Rolls E (1998) View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb Cortex 8(6):510–523

Bradski G, Grossberg S (1995) Fast-learning VIEWNET architectures for recognizing three-dimensional objects from multiple two-dimensional views. Neural Netw 8(7):1053–1080

Bülthoff H, Edelman S (1992) Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc Natl Acad Sci USA 89(1):60–64

Chun M, Marois R (2002) The dark side of visual attention. Curr Opin Neurobiol 12(2):184–189

Edelman S, Weinshall D (1991) A self-organizing multiple-view representation of 3d objects. Biol Cybern 64(3):209–219

Einhäuser W, Hipp J, Eggert J, Körner E, König P (2005) Learning viewpoint invariant object representations using a temporal coherence principle. Biol Cybern 93(1):79–90

Erickson C, Desimone R (1999) Responses of macaque perirhinal neurons during and after visual stimulus association learning. J Neurosci 19(23):10404

Goldstone R (1998) Perceptual learning. Ann Rev Psychol 49

Goodale M (1993) Visual pathways supporting perception and action in the primate cerebral cortex. Curr Opin Neurobiol 3(4):578–585

Haider H, Frensch P (1996) The role of information reduction in skill acquisition. Cognit Psychol 30(3):304–337

Hu M (1962) Visual pattern recognition by moment invariants. IEEE Trans Inf Theory 8(2):179–187

Jagadeesh B, Chelazzi L, Mishkin M, Desimone R (2001) Learning increases stimulus salience in anterior inferior temporal cortex of the macaque. J Neurophysiol 86(1):290–303

Jugessur D, Dudek G (2000) Local appearance for robust object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol 1, 2000

Kietzmann TC, Lange S, Riedmiller M (2008) Incremental GRLVQ: Learning relevant features for 3D object recognition. Neurocomputing 71:2868–2879

Kirstein S, Wersing H, Korner E (2005) Rapid online learning of objects in a biologically motivated recognition architecture. 27th Pattern Recognition Symposium DAGM, pp 301–308

Kobatake E, Tanaka K (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol 71(3):856–867

Kobatake E, Wang G, Tanaka K (1998) Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. J Neurophysiol 80(1):324–330

Koenderink J, Doorn A (1979) The internal representation of solid shape with respect to vision. Biol Cybern 32(4):211–216

Lange S, Riedmiller M (2006) Appearance based robot discrimination using eigenimages. In: Nardi D, Riedmiller M, Sammut C, Santos-Victor J (eds) RoboCup-2004: Robot Soccer World Cup VIII. Springer, LCNS, Berlin, 2005

Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

Leibe B, Schiele B (2003) Analyzing appearance and contour based methods for object categorization. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR'03)

Logothetis N, Pauls J, Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. Curr Biol 5(5):552–563

Lowe D (1985) Perceptual Organization and Visual Recognition. Kluwer Academic Publishers, Norwell

Lowe D (1999) Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision,vol 2

Lowe D (2000) Towards a computational model for object recognition in it cortex. Biol Motiv Comput Vis 1811:20–31

Luong Chi M (2006) Introduction To Computer Vision and Computer Graphics. Institute of Information Technology, Hanoi, Vietnam

Mareschal D, Plunkett K, Harris P (1999) A computational and neuropsychological account of object-oriented behaviours in infancy. Dev Sci 2(3):306–317

Marr D, Nishihara H (1978) Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. Proc R Soc Lond Ser B Biol Sci 200(1140):269–294

Massad A, Mertsching B, Schmalz S (1998) Combining multiple views and temporal associations for 3-d object recognition. Proc ECCV 98:699–715

Maunsell J, Treue S (2006) Feature-based attention in visual cortex. Trends Neurosci 29(6):317–322

Mel B (1997) SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. Neural Comput 9(4):777–804

Milner A, Goodale M (1993) Visual pathways to perception and action. Prog Brain Res 95:317–337

Milner A, Goodale M (1996) The visual brain in action. Oxford University Press, NY

Miyashita Y (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature 335(6193):817–820

Miyashita Y (1993) Inferior temporal cortex: Where visual perception meets memory. Ann Rev Neurosci 16(1):245–263

Murphy-Chutorian E, Aboutalib S, Triesch J (2005) Analysis of a biologically-inspired system for real-time object recognition. Cognit Sci Online 3(2):1–14

Murray S, Wojciulik E (2004) Attention increases neural selectivity in the human lateral occipital complex. Nature Neurosci 7: 70–74

Mutch J, Lowe D (2006) Multiclass object recognition with sparse, localized features. In: Proceedings of the 2006 IEEE computer society conference on computer vision and pattern recognition, vol 1, pp 11–18

Mutch J, Lowe D (2007) Object class recognition and localization using sparse features with limited receptive fields. In IJCV

Nene S, Nayar S, Murase H (1996) Columbia object image library (COIL-100). Techn. Rep. No. CUCS-006-96, dept. Comp. Science, Columbia University

Nosofsky R (1984) Attention, Similarity, and the Identification-Categorization Relationship. Dissertation, Harvard University

Obdrzalek S, Matas J (2002) Object recognition using local affine frames on distinguished regions. BMVC 2002, pp 113–122

Oliva A (2005) Gist of a scene. Neurobiology of attention, pp 251–256

Paletta L, Pinz A (2000) Active object recognition by view integration and reinforcement learning. Rob Auton Syst 31(1-2): 71–86

Palmer S, Rosch E, Chase P (1981) Canonical perspective and the perception of objects. Attention and performance IX, pp 135–151

Perrett D, Hietanen J, Oram M, Benson P, Rolls E (1992) Organization and functions of cells responsive to faces in the temporal cortex. Philos Trans Biol Sci 335(1273):23–30

Perrett D, Mistlin A, Chitty A (1987) Visual cells responsive to faces. Trends Neurosci 10:358–364

Perrett D, Oram M, Ashbridge E (1998) Evidence accumulation in cell populations responsive to faces: An account of generalization of recognition without mental transformations. Cognition 67: 111–145

Perrett D, Oram M, Harries M, Bevan R, Benson P, Thomas S (1991) Viewer centered and object centered coding of heads in the macaque temporal cortex. Exp Brain Res 86:159–173

Poggio T, Edelman S (1990) A network that learns to recognize three-dimensional objects. Nature 343:263–266, 34

Rao R (1997) Dynamic appearance-based recognition. In: Proceedings of computer vision and pattern recognition

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025

Riesenhuber M, Poggio T (2000) Models of object recognition. Nat Neurosci 3:1199–1204

Riesenhuber M, Poggio T (2003) How visual cortex recognizes objects: The tale of the standard model. Vis Neurosci 2:1640–1653

Roobaert D, Van Hulle M (1999) View-based 3d object recognition with support vector machines. Neural Networks for Signal Processing IX, 1999. In: Proceedings of the 1999 IEEE Signal Processing Society Workshop, pp 77–84

Sakai K, Miyashita Y (1991) Neural organization for the long-term memory of paired associates. Nature 354(6349):152–155

Schneider G, Wersing H, Sendhoff B, Korner E, Schneider G, Wersing H (2004) Evolution of hierarchical features for visual object recognition. Third Workshop on SelfOrganization of Adap-tiVE Behavior (SOAVE 2004) Ilmenau, pp 104–113

Seibert M, Waxman A (1992) Adaptive 3-d object recognition from multiple views. IEEE Trans Pattern Anal Mach Intell 14(2):107–124

Serre T, Wolf L, Poggio T (2005) Object recognition with features inspired by visual cortex. In: IEEE Computer society conference on Computer vision and pattern recognition, 2005. CVPR 2005, vol 2

Shokoufandeh A, Marsic I, Dickinson S (1999) View-based object recognition using saliency maps. Image Vis Comput 17(5):445–460

Strickert M, Bojer T, Hammer B (2001) Generalized relevance LVQ for time series. Springer, Berlin, pp 677–683

Suard F, Rakotomamonjy A, Bensrhair A (2006) Object categorization using kernels combining graphs and histograms of gradients. In: International conference on image analysis and recognition, vol 2, pp 23–34

Tanaka K (1992) Inferotemporal cortex and higher visual functions. Curr Opin Neurobiol 2(4):502–505

Tanaka K (1996) Inferotemporal cortex and object vision. Ann Rev Neurosci 19(1):109–139

Tarr M, Bülthoff H (1995) Is human object recognition better described by geon-structural-descriptions or by multiple-views. J Exp Psychol Human Percept Perform 21(6):1494–1505

Tarr M, Bülthoff H (1998) Image-based object recognition in man, monkey and machine. Cognition 67(1):1–20

Tarr M, Pinker S (1989) Mental rotation and orientation-dependence in shape recognition. Cognit Psychol 21(2):233–282

Teynor A, Rahtu E, Setia L, Burkhardt H, Teynor A, Rahtu E, Setia L, Burkhardt H (2006) Properties of patch based approaches for the recognition of visual object classes. In: Pattern recognition, DAGM 2006 Proceedings, lecture notes in computer science, vol 4174, pp 284–293

Thompson D, Mundy J (1987) Three-dimensional model matching from an unconstrained viewpoint. In: Proceedings of the IEEE International Conference on Robotics and Automation, vol 4

Tuytelaars T, Van Gool L et al (1999) Content-based image retrieval based on local affinely invariant regions. International conference on visual information systems, pp 493–500

Ullman S, Basri R (1991) Recognition by linear combinations of models. IEEE Trans Pattern Anal Mach Intell 13(10):992–1006

Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. Nat Neurosci 5:682–687

Voigtländer A, Lange S, Lauer M, Riedmiller M (2007) Real-time 3d ball recognition using perspective and catadioptric cameras. In ECMR 2007

Vuilleumier P, Henson R, Driver J, Dolan R (2002) Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. Nat Neurosci 5(5):491–499

Wallis G (1996) How neurons learn to associate 2d-views in invariant object recognition. Technical report, Technical Report No

Wallis G (1998) Temporal order in human object recognition learning. J Biol Syst 6(3):299–313

Wallis G, Bülthoff H (1999) Learning to recognize objects. Trends Cognit Sci 3(1):22–31

Wallis G, Bülthoff H (2001) Effects of temporal association on recognition memory. Proc Natl Acad Sci 98(8):4800–4804

Wallraven C, Bülthoff H (2001a) Automatic acquisition of exemplar-based representations for recognition from image sequences. In: Proceedings of the CVPR'01-workshop models versus exemplars, vol 28

Wallraven C, Bülthoff H (2001b) View-based recognition under illumination changes using local features. In: Proceedings of the CVPR'01-workshop on identifying objects across variations in lighting: psychophysics and computation, vol 3

Walther D, Fei-Fei L (2007) Task-set switching with natural scenes: Measuring the cost of deploying top-down attention. J Vis 7(11):9

Wersing H, Korner E (2002) Unsupervised learning of combination features for hierarchical recognition models. International conference on artificial neural network, ICANN. 11

Würtz R (1995) Multilayer dynamic link networks for establishing image point correspondences and visual object recognition. Verlag Harri Deutsch

Young M, Yamane S (1992) Sparse population coding of faces in the inferotemporal cortex. Science 256(5061):1327