Combining Multiple Views and Temporal Associations for 3-D Object Recognition

Amin Massad, Bärbel Mertsching, and Steffen Schmalz

University of Hamburg, Dep. of Computer Science, AG IMA
Vogt-Kölln-Str. 30, D-22527 Hamburg, Germany
mertsching@informatik.uni-hamburg.de

Abstract. This article describes an architecture for the recognition of three-dimensional objects on the basis of viewer centred representations and temporal associations. Considering evidence from psychophysics, neurophysiology, as well as computer science we have decided to use a viewer centred approach for the representation of three-dimensional objects. Even though this concept quite naturally suggests utilizing the temporal order of the views for learning and recognition, this aspect is often neglected. Therefore we will pay special attention to the evaluation of the temporal information and embed it into the conceptual framework of biological findings and computational advantages. The proposed recognition system consists of four stages and includes different kinds of artificial neural networks: Preprocessing is done by a Gabor-based wavelet transform. A Dynamic Link Matching algorithm, extended by several modifications, forms the second stage. It implements recognition and learning of the view classes. The temporal order of the views is recorded by a STORE network which transforms the output for a presented sequence of views into an item-and-order coding. A subsequent Gaussian-ARTMAP architecture is used for the classification of the sequences and for their mapping onto object classes by means of supervised learning. The results achieved with this system show its capability to autonomously learn and to recognize considerably similar objects. Furthermore the given examples illustrate the benefits for object recognition stemming from the utilization of the temporal context. Ambiguous views become manageable and a higher degree of robustness against misclassifications can be accomplished.

1 Introduction

The utilization of image processing seems promising for the development of machines capable of circumspect, flexible, or even autonomous interactions with their environment. As scenes are usually composed of objects, artificial image understanding needs the implementation of a functioning object recognition system. The latter is the main topic of the presented work which aims at the recognition of real three-dimensional objects. Special attention is given to the evaluation of the temporal dimension for the representation as well as for the recognition. Hereby, ambiguous views become manageable and a higher degree of robustness can be achieved for the recognition process.

Recent studies in psychophysics and neurophysiology stress the advantages and plausibility of viewer centred representations in contrast to object centred

models. These biological findings together with computational aspects give reason for the usage of a view-based representation scheme as presented here.

Another important concern is the realization of a system equipped with learning abilities to ensure adaptability to newly perceived views of known objects and to allow the addition of previously unknown objects. With respect to the desired properties of the solution, we focus on a universality regarding the kind of objects, features, and restrictions.

Facing the complexity of the subject we only can propose one possible way based on a sufficient number of arguments. Accordingly, the selection, interpretation, evaluation, and connection of various approaches are regarded as main components of this study.

2 Representing 3-D Objects

2.1 Viewer Centred vs. Object Centred Representations

The implementation of an object recognition necessarily requires the definition of a representation scheme for the objects, which does not only determine the type of storage but even the manner of the recognition process. Special attention shall be paid to the representation schemes because of considerable contradictions between the concepts of object centred and viewer centred descriptions. Object centred models store a single representation for each object (e. g. Marr and Nishihara 1978; Biederman 1985; Thompson and Mundy 1987; Lowe 1986). Accordingly, such a representation has to be three-dimensional and often resembles models used in computer graphics, esp. CAD. In contrast, viewer centred models contain collections of representations for the different perspectives under which the object appears to a (virtual) viewer. Thus a self-contained description of an object is embodied in the combination of its views, which can be three-dimensional as well as two-dimensional. Concerning invariance it has been argued for the usage of object centred models. Likewise the anticipation of a combinatorial explosion, caused by the multitude of imaginable views, lead to the rejection of viewer centred representations. Object centred models have the one purpose in common: the description of objects by high level features which provide stability over all perspectives. These features, however, involve a high degree of complexity and computational efforts. As a matter of fact, a great number of representation schemes is merely designed for theoretical treatments without being implemented, not even for testing purposes.

Recent studies in computer science as well as in psychology and neurophysiology support the notion of viewer centred models by not only refuting the fear of a combinatorial explosion but even stating several advantages over object centred representations. Therefore a short review of the relevant work will be given together with the corresponding implications.

Ullman and Basri (1991) propose a viewer centred model based on two-dimensional views: It is not restricted to rigid transformations, does not involve the explicit reconstruction and representation of the three-dimensional structure for storing the objects. Another noteworthy aspect is the proof that under certain assumptions all the views of a three-dimensional object, which may arise by affine transformations, can be derived from the linear combination of a few 2D views. On the other hand this method presupposes the visibility of all object points from every perspective. Even though these assumptions must be regarded as

hardly realizable for real scenes or automated model acquisition, the scheme gives impressive hints of the potential information content of two-dimensional views.

An early implementation of a view based recognition system by means of an artificial neural network is presented by Poggio and Edelman (1990). They postulate that for every object an appropriate function can be found which is capable of transforming all possible views into a single standard view. The approximations of these functions are expected to be evolved by *RBF* networks (*Radial Basis Functions*) after separately training them on different views of their corresponding object. Recognition involves the application of the transformation functions to the input view and the comparison of the resulting outputs with the stored standard views. Because of the necessity of a constant number of feature points together with an exact correspondence relation between image and model, the previously mentioned drawbacks also hold for this approach.

In contrast, the *CLF* network (*Conjunctions of Localized Features*) suggested by Edelman and Weinshall (1991) does not need the computation of an explicit correspondence but uses topological feature maps. While CLF networks have the capability to simulate error rates and recognition measures found in psychological tests, they seem to be inadequate for general image processing purposes because of their generalization method, which is mainly based on gaussian blurring of the stored model representations. Therefore, preliminary experiments with real images instead of artificially designed objects showed the tendency towards improper matches with models containing a higher number of feature points. This systematic fault is caused by single feature points in the input image overlapping with several widened model points simultaneously and can only be circumvented by a constant number of feature points over all views and all objects as it was the case in the study of Edelman and Weinshall.

More realistic input images are processed by the *VIEWNET* architecture (*View Information Encoded with NETworks*) described by Grossberg and Bradski (1995). The view based model includes a biologically motivated preprocessing chain to convert the input images into a representation invariant under illumination changes, translation, plane rotation, and scaling. The classification of the resulting patterns is done by a Fuzzy-ARTMAP network (cf. Carpenter et al. 1992), an artificial neural network stemming from the adaptive resonance theory (ART). While this study hints at the advantages resulting from the consideration of view sequences instead of single images, it still neglects the order in which the views appear.

Such an evaluation of the serial information can be found in Seibert and Waxman (1992). Their system is able to create transition matrices from view sequences and thus offers a method for the automated construction of aspect graphs as defined by Koenderink and van Doorn (1979). However, the edges of an aspect graph indicate the transition between merely two views. The assembly of longer sequences requires the implicit assumption of transitivity along edges. Although the connections are augmented by the relative frequency of the according transition, detailed information about longer sequences is not provided because of the missing serial relations.

Viewer centred representations not only give rise to these successful technical implementations but even have a close relationship to biological findings. Yet, the existence of so-called *canonical* and *accidental views* has diverging interpretations. Supporters of object centred representations might argue that accidental

views coincide with mathematical singularities which impair the reconstruction of a three-dimensional model. However, this hypothesis cannot account for the continuously decreasing recognition rates reported by Edelman and Weinshall (1991), which show obvious dependencies from the distance to previously trained views. These results were obtained for monocular as well as for stereoscopic presentation.

The outcome of the psychological experiments may be explained by Perrett et al. (1991) on a neurophysical level. Single cell recordings in the macaque *superior temporal sulcus (STS)* showed some cells with clear viewer centred behaviour. They were only stimulated by certain perspectives of familiar faces. Another kind of cells reacted nearly uniformly over all the views of a person and can therefore be considered as an object centred representation. The discovery of two cell types does not unveil a contradiction but rather demonstrates that object centred outputs can be interpreted as the simple supposition of several viewer centred cells. Tanaka (1996) gives an account of face sensitive cells in another brain area, the *anterior inferotemporal cortex* (AIT), which not only react in a viewer centred fashion but even show a systematic arrangement with locations for neighbouring views being structured in columns. Additionally, the neurophysiological evidence seems to reveal a universal principle: The search for a highly invariant recognition mechanism is solved in a way which represents a trade-off between memory and computation.

2.2 Utilizing Temporal Associations for Viewer Centred Represent

Although viewer centred representations are frequently implemented as the simple accumulation of the different aspects, a structured description showing the interrelations between the stored views is expected to be more beneficial. Time has proven to be a good guideline because the knowledge of temporal neighbourhood naturally facilitates the deduction of causal and spatial relationships. Therefore, the assumption of continuity leads to the conclusion that views appearing adjacently will probably stem from the same object.

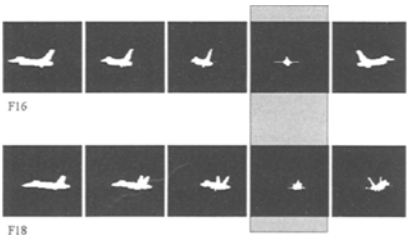


Fig. 1: The consideration of the temporal context allows to resolve ambiguities as illustrated for the fourth view of the sequence. (The depicted planes are kindly provided by the authors of Seibert and Waxman 1992).

Especially within the framework of viewer centred representations information about temporal relations is important to resolve ambiguities of single views since the models of similar objects may contain indistinguishable aspects as depicted in Fig. 1. Furthermore recognition of single views is always subject to inevitable errors which become remediable by the evaluation of view sequences: A higher number of aspects assigned to the same object indicates a higher propability for the correctness of the classification. Moreover, recording the occurrence of the sequences makes it possible to define a measure which denotes how typical is a certain motion of an object. In addition the deduction of exceptional situations is facilitated, hence allowing the detection of possible dangers or required

interventions. Given the example of a car, a rotation about the horizontal axis (fortunately) occurs less often than a turning about its vertical axis and may therefore be regarded as an exception.

There are several neurophysical and psychological experiments hinting at the utilization of temporal associations in biology. Miyashita (1988) reports the training of macaque monkeys with a set of 97 randomly generated fractal images. Conducting *delayed matching to sample (DMS)* tasks the animals had to decide whether two consecutively presented patterns were the same or not. Training was performed over quite a long period of time by presenting a circular repetition of the images and thus maintaining the order of the training set. Subsequent single cell recordings in the inferotemporal cortex exhibited neurons with effective stimuli formed by clusters of consecutive patterns of the training set. Accordingly these randomly generated patterns established associations not for their geometric similarity but for their temporal connections. This conclusion is affirmed by further experiments of Sakai and Miyashita (1991) who managed to create associations between any pairs of shapes on the basis of a consecutive presentation.

Psychophysical evidence for the existence of temporal associations can be found in Wallis (in press). His study makes use of training sequences which are artificially created by the combination of consecutive faces belonging to different persons. Assuming the development of temporal associations, one would suppose interconnections between the views of a sequence possibly leading to the fusion of a single virtual face. In fact, recognition errors showed more often a confusion between faces belonging to a common artificial sequence than between faces of different sequences. In other words the subjects formed associations between views because of their coincident appearance and not because of their similarity.

Summarizing the findings about temporal associations, Stryker (1991) deduces a powerful scheme which seems to supersede the need for mechanisms of geometric transformations or hierarchical connections of so-called trigger features. As previously mentioned for viewer centred representations, he considers temporal associations to be a trade-off between memory and computation providing a means for the brain to accomplish perceptual constancy.

3 Recognizing 3-D Objects from View Sequences

3.1 Overview of the Recognition System

The design of our object recognition system is based on the combination of viewer centred representations and temporal associations. According to the notion of viewer centred representations discussed in Sect. 2.1, it is possible to store three-dimensional objects as collections of two-dimensional views and to achieve an invariant object recognition by the connection of several variant view classes with broadly tuned outputs. Hence the need arises for mechanisms capable of generalizing across a range of vantage points, which is a task suggesting the use of artificial neural networks. However, objects do not come into sight as single views, instead they are naturally embedded into a temporal context. The stated advantages of a hypothesized continuity for a view based recognition propose the utilization of temporal information by the integration of view sequences into the representation scheme. Therefore the model will be augmented by mechanisms with the capacity to record the temporal order of successive views and to divide

these sequences into classes. Thus object classes can be built upon collections of sequence classes. By defining a single view to be a special case of a (very short) sequence, one can drop the additional shortcut connections between views and object classes as depicted in Fig. 2.

Obviously the use of long sequences yields a high number of possible sequence classes. In anticipation of an objection to a combinatorial explosion, it is important to notice that in practice only a very limited subset of the theoretically possible sequences will appear. Constraints originate from natural laws or conditions: gravity, for example, affects the viewer as well as the objects and considerably restricts the motions under which objects can be perceived; living beings show characteristic movements determinded by morphology or learnt behaviour. Matsakis et al. (1990) report investigations of cosmonauts of the space station MIR obtaining improved capabilities in mental rotation experiments under weightlessness. They presume the improvements to be effected by learning and to be partly caused by physiological processes in conjunction with the lost sense of gravity which make it easier to mentally relate the positions of object and viewer. Evidence for the relevancy of certain kinds of movements can be found in Perrett et al. (1990) where neurophysiological experiments exhibited cells of the macaque temporal cortex selectively reacting on complex movements or actions. Furthermore the number of cells coding compatible movements (e. g. the left profile of a person walking to left) was noticeably higher than the number of cells representing unusual movements (e. g. the left profile of a person moving to the right, i. e. walking backwards). Similarly Sumi (1984) reports higher recognition rates for normally presented biological motion stimuli than for inverted sequences.

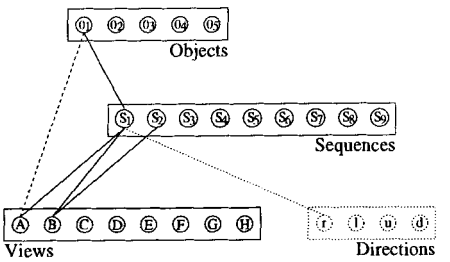


Fig. 2: Relations between the elements of the representation scheme, which comprises classes of views, sequences and objects. The system can be extended by direction classes indicating whether a rotation is right-, left-, up-, or downwards.

Moreover, the number of sequence classes can be reduced by the inclusion of mechanisms for invariance against modifications of the temporal patterns. The proposed system does explicitly not aim at storing all possible view sequences into the object model, instead it is designed to employ its learning capabilities in order to extract characteristics of the objects from the presented training sequences, to weigh them, and to use them for future recognition tasks.

Although spatial information is not necessarily required for the object recognition intended here, it should be included later to complete the concept of the object model: We suggest to seek the useful spatial information of an object in the relationship of its views. For this the sequence classes could be augmented by a small number of detectors indicating some basic directions of rotation about the main axes. Assuming mutually exclusive direction classes, such a modelling will be in agreement with the results of mental rotation experiments (cf. Metzler and Shepard 1974) in which the subjects have more difficulties in the simultaneous rotation of three-dimensional objects about more than one axis than in rotations about a single main axis.

Based on the proposed representation scheme the structure of the recognition system is implemented as depicted in Fig. 3. The process starts with the extraction of local features resulting from a Gaborjet transform applied to the input images. The inclusion of a visual attention algorithm to select a subset of these features for further processing would be beyond the scope of this article and is therefore preliminarily integrated as relatively simple method for the selection of a window around the interesting object within the input scene.

Usually the following steps employ a kind of coordinate transform and the computation of global features in order to yield an invariant representation as the foundation of a subsequent view classification. In Sect. 3.3 we will discuss why such a technique has serious drawbacks. Therefore we have chosen another approach which combines several of the processing steps and profits from the consideration of the topological information coded in feature maps. This so-called *Dynamic Link Matching (DLM)* has its origin in von der Malsburg (1981), however several extensions had to be applied to this basic concept before it could be used in the given context for the foundation of view classes.

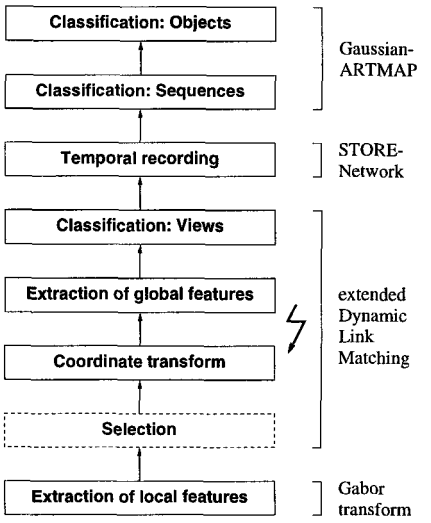


Fig. 3: Simplified block diagram of the recognition system depicting the relations of the processing steps to the implemented algorithms.

The temporal recording of the view classes is done by a STORE network according to Bradski et al. (1992) which transforms the patterns into an *item-and-order coding* as described in Sect. 3.4. A Gaussian-ARTMAP architecture, which was introduced by Williamson (1996), divides the outputs from the STORE network into sequence classes and maps them onto object classes by means of a supervised training method. Section 3.5 deals with the details of this stage.

3.2 Feature Extraction: Gaborjet Transform

In principle, the generality of the DLM algorithm allows to use every kind of input that is organized in feature maps. Even though DLM is designed to tolerate variations of the feature positions, it is necessary to find features which are stable under a large range of varying conditions and on the other hand not too widespread in order to reduce the number of initial ambiguities. Würtz (1994) suggests the usage of a Gabor-wavelet transform because Gabor filters are well known to have computational advantages and to give a theoretical account for findings about biological visual systems. Accordingly, preprocessing is implemented as the convolution of the input images with a number of n_{dir} different oriented kernels on n_{lev} resolution levels defined as

$$(\mathcal{F}\psi_{\boldsymbol{k}})(\boldsymbol{\omega}) = \exp\left(-\frac{\sigma^2\,(\boldsymbol{\omega}-\boldsymbol{k})^2}{2\boldsymbol{k}^2}\right) - \exp\left(-\frac{\sigma^2\,(\boldsymbol{\omega}^2+\boldsymbol{k}^2)}{2\boldsymbol{k}^2}\right) \text{ with } \boldsymbol{k} = \begin{pmatrix} k_{lev}\cos\phi_{dir} \\ k_{lev}\sin\phi_{dir} \end{pmatrix}$$

where $k_{lev} = n_{lev}/\sqrt{2}^{lev}$, $\phi_{dir} = \pi \cdot dir/n_{dir}$, and $\mathcal{F}$ denotes the Fourier transform. The resulting amplitudes of the filter outputs form a feature vector f, called Gaborjet, of the dimension $n_{dir} \cdot n_{lev} = 12 \cdot 2$. An additional subsampling yields patterns of the size 64 × 64 pixels from the input images of 256 × 256 pixels.

3.3 Classifying Views: Extended Dynamic Link Matching

It is common practice to transform the input patterns into a special representation invariant against certain kinds of mathematically defined operations (e. g. the magnitudes of the Fourier transform or a centred log-polar transform) or to predetermine a transforming function (e. g. an affine transform) and compute its parameters for the mapping of the current input image onto the stored patterns. However, the success and the applicability of these approaches are limited because an appropriate formal description for real scenes is still to be found. Serious problems arise for example from noise, occlusions, distortions, non-rigid objects, changes of illumination or background. In its most general formulation matching consists in mapping local features of the input pattern onto corresponding features of a trained pattern and thus yielding a global transformation between both patterns. In this process ambiguities of local features have to be solved by the consideration of neighbouring features. Therefore, special attention is paid to the information contained by the topological structure of the patterns.

Konen et al. (1994) present a neural formulation of such a match process, called *Dynamic Link Matching*, which uses topography as a guideline on the assumption that a transformation needed to map a local feature of the stored pattern onto its counterpart in the input pattern is very likely to be applicable to match the neighbours onto their counterparts. The system consists of two layers, the image layer X representing the current input pattern and the model layer Y storing the pattern against which the input is to be matched. The neurons of each layer are labeled by local features. Both layers are connected by interlayer links J_{ba} carrying information about the correspondence between each cell b of layer Y and each cell a of layer X. These connections, called *dynamic links*, model a kind of working memory with rapid weight changes and can be understood as a measure for the probability that a cell a is the correct correspondence for b. Additionally, the wiring contains static intralayer weights in both layers which couple each cell with its local neighbours by excitatory connections and with more distant cells by inhibitory connections. The process for self-organization of the dynamic links has two formulations, a description of the neural activity by dynamic equations and a simplified algorithm based on the blob equilibrium solution, which proves the emergence of a connected active region in each layer. Our discussion will be focussed on this *Fast DLM (FDLM)* algorithm outlined in Fig. 4.

This basic DLM concept was usually applied to sets of quite dissimilar objects whereas our application has additionally to cope with similar views depicting the same objects under different perspectives. By separately applying the DLM to every model, we frequently obtained high correspondences between a given input image and several models. Neither the correlation value of the blob regions

1	Init dynamic links according to the similarity $S(f_b, f_a)$ between the features f_b and f_a with $J_{ba} = T_{ba} / \sum_{a' \in X} T_{ba'}$ where $T_{ba} = S(f_b, f_a)$.
2	Choose a random centre $a_c \in X$ and place the blob there: $x_a = B(a - a_c)$. Compute the resulting input of each cell $b \in Y$: $I_b = \sum_a J_{ba} T_{ba} B(a - a_c)$
3	Use I_b to compute the maximum $b_c \in Y$ of the potential V: $V(b_c) = \sum_{b \in Y} B(b - b_c) I_b$ and place the blob there: $y_b = B(b - b_c)$.
4	Update the links between the active blob regions and renormalize: $J_{ba} := \frac{J_{ba} + \Delta J_{ba}}{\sum_{a' \in X}(J_{ba'} + \Delta J_{ba'})}$ with $\Delta J_{ba} = \epsilon J_{ba} T_{ba} Y_b X_a$.
5	Proceed with step 2 until stop criterion is true (e. g. $t > t_{max} = 2000$).

Fig. 4: Fast Dynamic Link Matching according to Konen et al. (1994). The blob solution B can be of arbitrary and simple shape, e. g. a rectangular window function.

nor the structure of the dynamic links allowed the definition of a measure to discern between them. Therefore, the algorithm has to be modified to competitively match all stored models against the current input. In contrast to Wiskott and von der Malsburg (1996) we will maintain the principle of the FDLM algorithm. The scheme of Fig. 4 is extended to consider m layers Y_m simultaneously, where m denotes the number of stored models. Step 3 now has to compute the maximum of V over all models m. Only the layer $Y_{m'}$ containing the region of highest activation will form a blob and update its weights to the current blob in X (step 4).

The modification makes it possible to define a useful recognition measure $r^m(t)$ for a model m and an iteration step t by

$$r^m(t_0) = 1 \qquad \dot{r}^m(t) = \lambda_r r^m \left(F^m - \max_{m'} \left(r^{m'} F^{m'} \right) \right)$$

where F^m denotes the „fitness" of the layer m, computed as total activity of all its cells, and λ_r is a time constant. The right equation expresses a competition between the model with the highest fitness and all the weaker models, which will be further suppressed by the last term becoming negative. During the recognition process the recognition value of the best fitting model approaches one while the remaining values decrease to zero. A speed up of recognition is achieved by ignoring models which have dropped under a threshold $r^m(t) < \theta_r$. Additionally, we demand a certain significance of the maximum found in step 3 before admitting the weight updates of step 4, i. e. there has to be a sufficient distance to the second best value $V(b'_c)$. A precondition defined as $V(b_c)/V(b'_c) \geq k$ (e. g. $k = 1.1$) intensifies the competition and ensures that connections are only made between characteristic regions and not between areas belonging to several models (e. g. some background pixels which are inaccurately included in the models).

A full connectivity between the model layers Y^m and the input layer X yields large matrices $J_{b^m a}$ for the links and $T_{b^m a}$ for the similarity values. To save memory and to reduce computation Wiskott and von der Malsburg (1996) propose to restrict the receptive field of each cell b to patches of the size $w_{patch} \cdot h_{patch}$, which are evenly distributed over the image layer. In addition to the primarily intended increase of performance, this restriction has great advantages for the arrangement of the generated correspondences: The feature extraction can roughly be interpreted as a kind of edge detector, therefore the DLM process sometimes

finds implicit symmetries (e. g. by matching opponent parts of a lengthy edge and thus creating a intersection of the links) which are then gradually extended over the complete image. The local bounding of the weights prevents such effects by initially excluding connections between too distant features.

The size of the patches must be chosen very carefully: On the one hand performance depends on small sizes, on the other hand the patches have to be large enough to allow displacements of the stored model in larger input scenes. For that reason the size of the patches is individually computed for each model according to its size (note: smaller models need larger patches).

Generally the model layers Y^m will be much smaller than the image layer X, for example when the system has to search an object within a scene. In this case the results of the DLM can be further enhanced by the introduction of an attention window which limits the positions of blobs in layer X to the region of a presumed object. After a predefined number of iterations (e. g. $t_w = 800$) we set the region of interest by considering the model m' with the highest recognition value and computing the centre $(\bar{a}_x, \bar{a}_y)$ of the window on the basis of its link matrix $J_{b^{m'}a}$ as $\bar{a}_{x/y} = \sum'_{b^m} \sum_{a_{x/y}} a_{x/y} \cdot J_{b^{m'}a_{x/y}}$, height and width of the window are deduced from the standard deviations $(2\sigma_{a_x}, 2\sigma_{a_y})$.

The extended DLM algorithm described above forms the first stage of the matching process. For a given input image it returns the best fitting model from the set of known views. In a second stage we have to test whether the selected model is sufficiently similar to the input. If this is true we will update the model else we will create a new model to hold the current view. The definition of an appropriate similarity measure is facilitated by the DLM algorithm and can be derived from the smoothness of the correspondence grids as depicted in Fig. 5. We have implemented two different distortion measures yielding comparable results, one based on the normalized standard deviation of the node distances and another indicating the perpendicularity of the grid by summing up the deviations of the angles from 90 degrees.

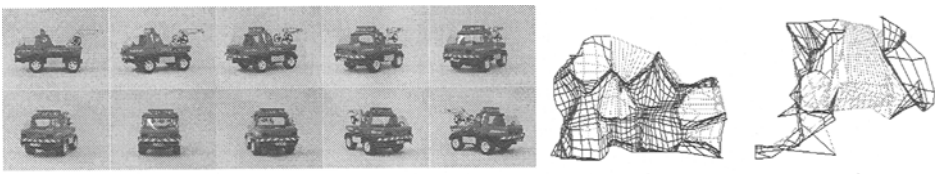


Fig. 5: Two different kinds of correspondence grids as exhibited while training of view sequences. Each node corresponds to a cell of the model layer Y, edges connect neighbouring neurons. The positions of the nodes indicate the best corresponding location within the image layer X. This correspondence is based on the correlation values C_{ba} between a model neuron b and an image neuron a. In order to reduce small irregularities an additional smoothing has been applied to the grid. Nodes b with no sufficiently correlated cell within X are positioned by interpolation with their matching neighbours and connected by light-grey edges. **a)** Example of a view sequence taken from object 1. **b)** Regular correspondence grid stemming from the match of view 4 onto the previously learnt model view 2. **c)** Irregular grid originating from the match of view 9 onto view 2 for the lack of a more similar candidate within the model database.

If the second stage requires the update of an existing model an adaptation of this model to the current input will be performed by placing the feature vector

located at (b_x, b_y) to $(b_x, b_y)' := (b_x, b_y) + \lambda[(c_x, c_y) - (b_x, b_y)]$where λ denotes the learn rate and (c_x, c_y) is the position of the cell b within the correspondence grid.

3.4 Recording View Sequences: Sustained Temporal Order Network

We have chosen to represent the temporal order of the view sequences by a *STORE* (Sustained Temporal Order REcurrent) model according to Bradski et al. (1992). The network is capable of encoding the invariant temporal order of sequential events (e. g. regardless of their durations and interstimulus intervals). Basically it transforms the input sequences into an item-and-order coding which encodes the events that have occurred and the temporal order in which they have occurred. This coding ensures that the presentation of new events will not invalidate the previously learnt patterns. We are especially interested in the possibility to use a network of the ART architecture for the following classification because this design facilitates fast learning and provides a solution for the stability-plasticity-dilemma of artificial neural networks. Both aspects can be regarded as advantages over other kinds of temporal representations realized by time shifter networks or the frequently applied JORDAN and ELMAN nets.

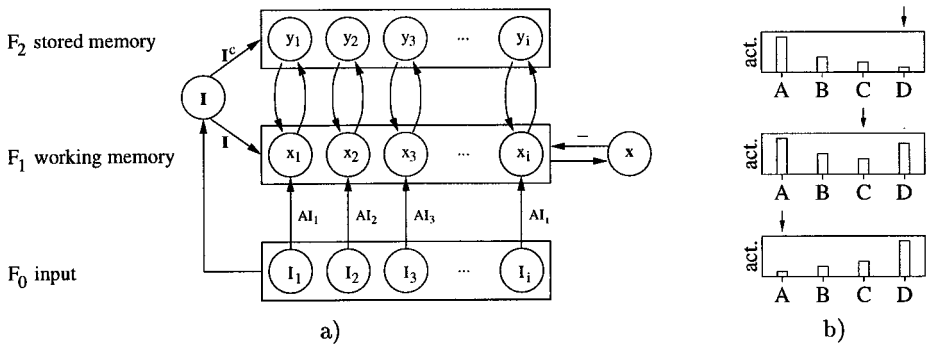


Fig. 6: a) STORE architecture according to Bradski et al. (1992). The bold-faced letters denote the sums $\mathbf{I} = \sum_k I_k$, $\mathbf{I^c} = 1 - \mathbf{I}$, and $\mathbf{x} = \sum_k x_k$; A is an arbitrary factor controlling the shape of the generated gradient. **b)** Representing temporal information by means of an *item-and-order coding*. Top: *Primacy* gradient for small A. Middle: *Bowing* gradient for $0 < A < 1$. Bottom: *Recency* gradient for $A \geq 1$ of the sequence $A \rightarrow B \rightarrow C \rightarrow D$. The arrow indicates the position at which the bowing starts.

Figure 6a shows the structure of a STORE network. The layer F_1 implements a short term memory (STM) which represents the temporal order of the input sequence by its activities x_i and is regarded as the output of the STORE architecture. The negative feedback of the total layer activity $\mathbf{x}$ ensures a partial normalization in order to model psychological findings of a limited STM capacity. The second layer F_2 consists of neurons which track the activities of the F_1-cells and function as a kind of memory. Changes in both layers occur mutually exclusive in dependence of the gain control $\mathbf{I}$ which indicates whether an input is present or not. The factor A applied to the inputs controls the shape of STM gradients as depicted in Fig. 6b. The following stage of our system will receive recency gradients as inputs by setting $A = 1.1$.

3.5 Classifying Sequences and Objects: Gaussian ARTMAP

A Gaussian ARTMAP network as introduced by Williamson (1996) is used for incremental learning of the sequences and their mapping onto object classes. This architecture implements a combination of a Gaussian classifier and an ART neural network which remedies known deficiencies of other ARTMAP architectures, especially Fuzzy ARTMAP (Sarle 1995). It is more resistant to noise, prevents the proliferation of the generated classes, achieves independence of the order in which the patterns are trained, and pays attention to the statistical distribution of the patterns stored within the classes.

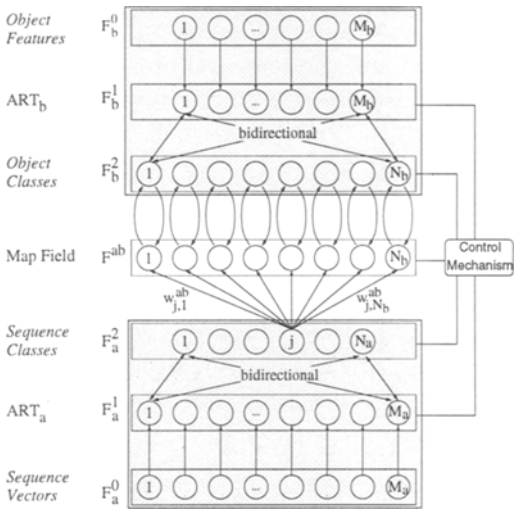


Fig. 7: Gaussian ARTMAP architecture consisting of two Gaussian ART nets and a map field. ART$_a$ receives at F_a^0 the STORE patterns as input and divides them into classes represented by the cells of F_a^2. Similarly ART$_b$ yields object classes at F_b^2 from object features.

Gaussian ARTMAP (Fig. 7) is based on the concept of the Gaussian ART architecture, which defines its classes by Gaussian distributions. Such a class j is represented by trainable M-dimensional vectors storing the mean $\boldsymbol{\mu}_j$ and standard deviation $\boldsymbol{\sigma}_j$ together with a scalar n_j counting the number of coded samples. The ART choice function and the match function are computed by means of the a priori probabilities and their normalization to unit height, respectively.

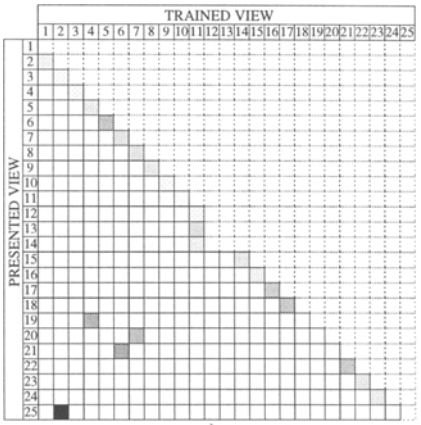
The map field is trained to map a class J of F_a^2 onto its predicted class K in F_b^2, a function which is generally a many-to-one assignment. If a class J is chosen during training that maps onto an incorrect prediction $K' \neq K$ a mechanism called match tracking will be invoked. As a consequence of match tracking, the class J will be reset in order to let ART$_a$ bring up another class which obtains a correct mapping or is an untrained class that will be recruited for the storage of the current input pattern.

4 Results and Discussion

4.1 Demonstration of the Learn- and Recognition-Process

In the first example we illustrate the learn-process. Learning of the first object has already been mentioned by Fig. 5; therefore we present the sequence of Fig. 9a now. The system is supposed to perform a gradual update of the set of model views by means of the extended DLM algorithm. To make the interpretation of the view classes easier, the combination of more than one input image into a view class is prevented by demanding very low distortion measures.

Figure 8a shows the results. The greyed entries mark the stored view which is returned by the extended DLM algorithm as the view best matching the



a)

class	views	class	views	class	views
1	1	7	9, 10	13	18
2	2, 3	8	11, 12	14	19, 20
3	4, 5	9	13, 14	15	21
4	6	10	15	16	22, 23
5	7	11	16	17	24, 25
6	8	12	17		

b)

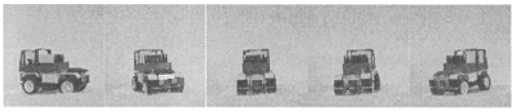
Fig. 8: a) Distortion measures of the first stage matches occurring during the successive learning of object 2. **b)** Assignments of views to classes resulting from the learn-process.

a) b)

Fig. 9: Objects used for training. Each object was presented as a sequence of 25 views which shows a complete rotation of the object about the vertical axis. Object 1 is depicted in Fig. 5a. **a)** Object 2. **b)** Object 3.

current input (first matching stage). The grey level corresponds to the distortion measure assigned to the grid of this match. Dotted fields in the upper right part denote views which have not been included into the set of model views yet. As expected most matches lie on the main diagonal of the table and indicate highest similarity with the directly preceding view which has just been stored into the model database.

The mean of the distortion measures on the diagonal is 0.98, whereas especially aberrant matches show higher values. If one sets the admissible distortion to a value less than the mean, the system will be allowed to combine similar views into a common view class as shown by Fig. 8b. It is noteworthy that slightly changed distortion measures can result from the adaptation of the stored views to one another dependent on the chosen learn rate.



view 26 obj. 2 | view 28 obj. 2 | **view 31 obj. 2** | **view 31 obj. 2** | view 33 obj. 2

Fig. 10: Example of the classification of a view sequence belonging to object 2. The resulting view classes and their assignment to objects are shown below.

The following example will demonstrate the processes of learning and recognition for three similar objects. Note that the resemblance between the objects

intentionally sets the system a difficult task. Therefore, we have to expect that some misclassifications will occur on the view level. They elucidate why it is necessary to consider sequences instead of single views.

The system was trained for the objects depicted in Fig. 5a and Fig. 9. Learning created a total of 55 view classes for the storage of 75 training views. The classes are almost evenly spread over the three objects (17, 18, 20 classes respectively). The presentation of the test sequence illustrated in Fig. 10 demonstrates the advantages of the temporal context. Despite of an ambiguous view (view 31) appearing twice in the sequence, the knowledge about the preceding views still allows a reliable recognition of the object.

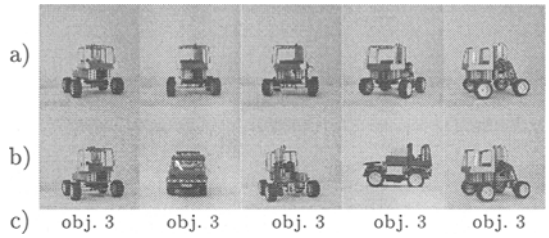


Fig. 11: Example of the classification of view sequence belonging to object 3. **a)** Presented test sequence of object 3. **b)** Representatives of the view classes matching each view of a). The second and the fourth view have been erroneously mapped onto similar views of other objects. The third image yields a wrong view class of the corresponding object. **c)** Classification output, which is stable despite of the misclassified views.

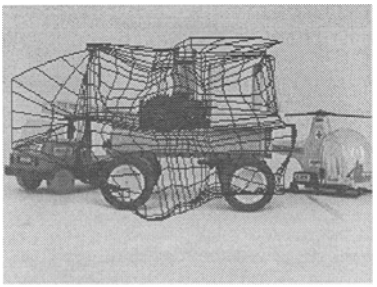


Fig. 12: Example for the recognition of object 3 within a cluttered scene. The correspondence grid is overlayed over the scene.

In addition to the manageability of uncertain views, an object recognition based on sequences facilitates an improved tolerance against misclassifications of single views: Figure 11 shows a test sequence of object 3 together with the views matching each input image. Although two of five views (2^{nd} and 4^{th} view) are mapped onto an incorrect object, the outputs resulting from the sequence classification yield the actual object class over all inputs.

A test with cluttered scenes is depicted in Fig. 12. After approx. 2000 iteration steps the correspondence grid is unfolded and positioned over the known object. Some distortions can be noted. However, they are restricted to the boundary nodes and still allow to localize the object within the scene. The effects of occlusion can be handled similarily. Further investigations of these aspects are currently in progress.

4.2 Comparison with Related Work

We have presented an image processing system for the recognition of three-dimensional objects which is based on the approach of viewer centred representations and on the utilization of temporal associations between the views.

To our knowledge the outlined architecture is unique with respect to the combination of its modules while particular elements were adapted from different known approaches. However, most implementations of viewer centred representations include a direct mapping of view classes onto objects and neglect the temporal order of the image sequences. As an example of such a system we have already mentioned the VIEWNET architecture of Grossberg and Bradski (1995). Figure 13 depicts the trained weights of some view classes generated by the Fuzzy-ARTMAP classifier of the VIEWNET system.

Due to an insufficient position invariance, training creates many similar classes (e. g. class 1, 2, 4, 7, and 9) and causes a proliferation of the number of classes. Moreover, the usage of the Fuzzy-AND operation produces serious „deletion effects": The rotation of an object shows several similar images in succession matching the same class. This class is then trained to store the intersection of the subsequent views. Hence, the rotation of the plane yields classes which represent a virtual image formed by the central region of the object. However, the implied instability of the view classes will obviously impede a reliable definition of sequence classes.

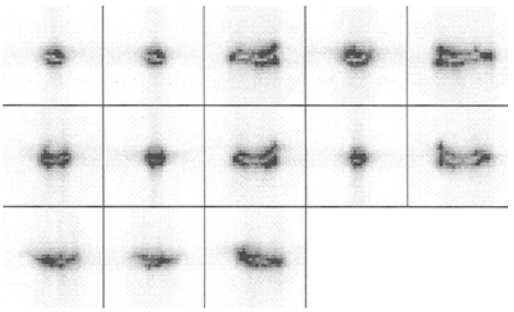


Fig. 13: View classes created by a VIEWNET architecture. The system has been trained for the views of an airplane as depicted in Fig. 1.

The Dynamic Link Matching developed by von der Malsburg et al. deals with the achievement of position invariance and robustness against distortion but requires that all model views are of the same size and manually aligned (e. g. the application to face recognition needed the eyes in all the stored faces to be placed at corresponding positions). Furthermore, recent publications focus on the computationally more expensive formulation of the DLM in terms of dynamical equations whereas our system still relies on the faster FDLM algorithm. Above all, the DLM architectures contain no mechanism for an automated learning of the models. The only report of an automatic model acquisition is given by von der Malsburg and Reiser (1995) where the attempt is made to map all the views of an object onto a single model view. However, the applicability of this approach seems questionable and is not in agreement with the concept of viewer centred representations proposed here.

With respect to the classification of sequences a comparable system can be found in Bradski and Grossberg (1993), which is based on a Fuzzy-ART network. This implementation has the same drawback of a deletion effect analogous with the one explained in the context of view classes: If a sequence misses out a view even once, this view will be deleted from the stored sequence class and will never be added to this class again.

Darrell and Pentland (1993) present a system for the processing of view sequences. The employed image processing algorithms, however, are quite simple. Moreover, training and recognition require an alignment, called dynamic time warping, of the current input sequence and all the known sequences.

4.3 Further Development

We are planning to increase the performance of the system with respect to speed as well as concerning its recognition capabilities. Another important aspect is the development of a benchmark test which provides an appropriate image database for a quantitative comparison of efficiency and accuracy. Currently none of the mentioned references offers such a set of input patterns applicable for testing our recognition system. The test sets either do not contain sequences or images without interior structure of the objects.

The results, presented here, indicate that some misclassifications originate from the subsampling rates. As the usage of an increased resolution comes along with considerably longer response times, certain mechanisms for a speed-up must be implemented. The replacement of the fixed stop criterion of the DLM by a measure which dynamically ends the recognition process seems promising. Furthermore, the use of additional features is possible without changing the architecture and should be investigated in the future. Another extension will be the embedding of the recognition system into an active stereo vision environment. Extracting information about motion and depth will enable the segmentation of the scene into regions of interest and thus speed up the recognition process. Moreover, by tracking the objects the system will be allowed to observe moving objects on its own instead of being dependent on artificially created training sequences.

References

Biederman, I. (1985): Human image understanding: Recent research and a theory. *Comput. Vision Graphics Image Processing* 32: 29–73.

Bradski, G.; Carpenter, G. and Grossberg, S. (1992): Working memory networks for learning temporal order with application to three-dimensional visual object recognition. *Neural Comput.* 4: 270–286.

Bradski, G. and Grossberg, S. (1993): Fast learning VIEWNET architectures for recognizing 3-D objects from multiple 2-D views. Technical Report CAS/CNS-TR-93-053, Boston Univ., Boston, MA.

Carpenter, G.; Grossberg, S.; Markuzon, N.; Reynolds, J. and Rosen, D. (1992): Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Trans. Neural Networks* 3(5): 698–713.

Darrell, T. and Pentland, A. (1993): Recognition of space-time gestures using a distributed representation. In R. Mammone (ed.), *Artificial neural networks for speech and vision*. London: Chapman & Hall, pp. 502–519.

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

Grossberg, S. and Bradski, G. (1995): VIEWNET architectures for invariant 3-D object recognition from multiple views. In B. Bouchon-meunier; R. Yager and L. Zadeh (eds.), *Fuzzy logic and soft computing*. Singapore: World Scientific Publishing.

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

Konen, W.; Maurer, T. and von der Malsburg, C. (1994): A fast link matching algorithm for invariant pattern recognition. *Neural Networks* 7: 1019–1030.

Lowe, D. G. (1986): *Perceptual organization and visual recognition*. Boston: Kluwer.

Marr, D. and Nishihara, H. K. (1978): Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond.* 200: 269–294.

Matsakis, Y.; Berthoz, A.; Lipschits, M. and Gurfinkel, V. (1990): Mental rotation of three-dimensional shapes in microgravity. In *Proc. Fourth European Symposium on Life Sciences in Space*. ESA, pp. 625–629.

Metzler, J. and Shepard, R. (1974): Transformational studies of the internal representaion of three-dimensional objects. In R. Solso (ed.), *Theories in cognitive psychology: The Loyola Symposium*. Erlbaum.

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

Perrett, D.; Harries, M.; Benson, P.; Chitty, A. and Mistlin, A. (1990): Retrieval of structure from rigid and biological motion: An analysis of the visual responses of neurones in the Macaque temporal cortex. In A. Blake and T. Troscianko (eds.), *AI and the Eye*, chap. 8. Wiley & Sons, pp. 181–200.

Perrett, D.; Oram, M.; Harries, M.; Bevan, R.; Hietanen, J.; Benson, P. and Thomas, S. (1991): Viewer-centred and object-centred coding of heads in the macaque tempral cortex. *Experimental Brain Research* 86: 159–173.

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

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

Sarle, W. (1995): Why statisticians should not FART. ftp://ftp.sas.com/pub/ neural/fart.doc.

Seibert, M. and Waxman, A. (1992): Adaptive 3-D object recognition from multiple views. *IEEE Trans. Pattern Anal. and Machine Intel.* 14(2): 107–124.

Stryker, M. P. (1991): Temporal associations. *Nature* 354: 108–109.

Sumi, S. (1984): Upside-down presentation of the Johansson moving light-spot pattern. *Perception* 13: 283–286.

Tanaka, K. (1996): Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19: 109–139.

Thompson, D. W. and Mundy, J. L. (1987): Three dimensional model matching from an unconstrained viewpoint. In *Proc. IEEE Int. Conf. Robotics and Automation*. Raleigh, NC: IEEE, pp. 208–220.

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

von der Malsburg, C. (1981): The correlation theory of brain function. Internal report, Max-Planck-Institut für Biophysikalische Chemie, Göttingen.

von der Malsburg, C. and Reiser, K. (1995): Pose invariant object recognition in a neural system. In *Proc. Int. Conf. on Artificial Neural Networks ICANN*. pp. 127–132.

Wallis, G. (in press): Temporal order in human object recognition learning. *To appear in Journal of Biological Systems* .

Williamson, J. (1996): Gaussian ARTMAP: A neural network for fast incremental learning of noisy multidimensional maps. *Neural Networks* 9(5): 881–897.

Wiskott, L. and von der Malsburg, C. (1996): Face recognition by Dynamic Link Matching. Internal Report IR-INI 96-05, Inst. f. Neuroinformatik, Ruhr-Univ. Bochum.

Würtz, R. (1994): *Multilayer dynamic link networks for establishing image point correspondences and visual object recognition*. Ph.D. thesis, Ruhr-Univ. Bochum.