Pattern Recognition in Practice IV
E.S. Gelsema and L.N. Kanal

A hybrid system to detect hand orientation in stereo images

A.Drees, F. Kummert, E. Littmann, S. Posch, H. Ritter, G. Sagerer[a]*

[a] Technische Fakultät, Universität Bielefeld, Postfach 100131, D-33501 Bielefeld, FR Germany

The recognition of hands and the 3-dimensional characterization of hand orientation is a difficult and practically important vision problem. In this work we propose a hybrid system attaching artificial neural networks to concepts of a semantic network to solve this problem. The neural networks are attached as a holistic representation at varying levels of the abstraction hierarchy modeled by the semantic network. In this way we are able to combine the advantages of both techniques. However, it can also be applied to other problems in computer vision and tasks of signal interpretation in general.

1. INTRODUCTION

Recognition and description of objects is one of the main problems in computer vision. Furthermore, this task is one of the key capabilities required for the successful operation of both biological organisms and artificial robots. Two different paradigms to solve this problem are often discussed as competing approaches: On the one hand artificial neural networks (ANNs) based on trained parameters, and on the other hand semantic networks with knowledge based techniques. In this work we do not intend to argue which one of these approaches is better suited to the image interpretation task. Rather we aim at utilizing the advantages of both techniques by combining them in a hybrid system for knowledge representation and utilization. The starting points for this hybrid system are the semantic network ERNEST and Local-Linear-Map networks. The latter ones are attached to concepts of the semantic net modeling objects.

As an application we chose the visual recognition of a hand and the characterization of its orientation. This poses a challenging vision problem the solution of which is of great practical interest, for instance to facilitate the control of multifingered anthropomorphic manipulators. However, the techniques developed here can also be applied to other problems in computer vision and tasks of signal interpretation in general like speech understanding. The work described in this paper is an extension of previous research ([1,2]) where monocular images are used as input. This system is further developed to utilize depth information acquired from stereo images.

In the next section, the advantages and disadvantages of both techniques and our approach for their combination in a hybrid system are discussed. Section 3 gives a brief overview of the neural and semantic networks used as the starting point for this work.

*This work has been partly supported by the German Research Foundation (DFG) in the project SFB 360 and by the German Ministry of Research and Technology (BMFT), Grant No. ITN9104AO.

Then the application of the hybrid system to recognize 3D-hand orientations is presented in detail and in section 5 results are discussed. We conclude with a discussion and plans for future work.

2. OUTLINE OF THE APPROACH

As in all semantic network approaches, the domain knowledge in ERNEST is explicitly structured using a decomposition and specialization hierarchy of concepts. Likewise, knowledge about attributes and relations between parts of the concepts is modeled in an explicit way. The analysis process is strongly influenced by this decompositional view of the world. Based on intensity information at pixel level, the sensor data are transformed into increasingly abstract representational levels. This process may proceed in both a data or model driven way, as well as in a mixed strategy. While the decomposition of objects and the explicit description of their attributes is one of the main advantages of semantic networks, it also may cause drawbacks. The inherent ambiguity of signal interpretation, especially for primitive parts of the segmentation hierarchy may lead to many competing interpretations for more complex objects of the knowledge base. Additionally, the acquisition and adaptation of the knowledge base imposes a considerable effort on semantic network approaches.

In contrast to semantic networks, the neural net approach does not attempt a decomposition into symbolic object parts. Instead, the properties of objects are modeled "holistically" in the weight parameters of an artificial neural network. Thus, the use of heuristics and world knowledge in the first approach is replaced by learning from examples. The trained network represents implicit knowledge about the attributes required for the identification of the object. This allows for a fast recognition of the learned objects that is also robust with regard to noise and variations in the signal. However, being a holistic system it is not feasible to build a single ANN that can cope with all possible configurations of many simultaneous objects in a complex scene. Rather, multiple ANNs have to be applied to certain regions of interest, but their coordination is not (yet) well understood in the neural paradigm. Another possible problem is the necessary size of the training set for ANNs. For complex images, the number of required training samples may be too large for realistic applications.

To overcome the disadvantages of both approaches we propose a hybrid system combining neural and semantic network techniques. The main idea is to associate or attach ANNs as holistic models to concepts of the semantic net, with both components modeling the same object.[1] That is, the interface between the different network types is not defined at one fixed level of the segmentation hierarchy, rather it is determined as appropriate for the given task, knowledge base, or the current state of the analysis process. Given such a hybrid knowledge base, different options are available to recognize a modeled object in a model-driven strategy. If a concept node is to be instantiated the associated ANN can be activated and the object is recognized in a fast and robust way without the necessity to detect the parts of the object as modeled by the semantic network. If no ANN has been attached to a concept node the analysis works in the usual manner pursuing the de-

[1] The same applies to other concepts modeled in the semantic network, like events or abstract conceptions. For simplicity, however, we only refer to objects in the following.

composition hierarchy. In this mode of operation the semantic network is mainly utilized to control the analysis process and to focus the various ANNs attached to the semantic network on different image regions. If in a later phase of the analysis process information about parts and attributes of parts is required, the knowledge about the structure of objects modeled in the semantic network can still be exploited. In a data driven analysis strategy the interaction is done in a similar way. After an object has been recognized by an ANN the corresponding concept can be instantiated even if its parts are not (yet) detected. In a mixed strategy the instantiated objects recognized by ANNs can be used to select appropriate goal concepts of more abstract levels of the semantic network. In this way the number of competing interpretations is drastically reduced and the analysis process can be focused by propagating the contraints from the estimated goal concepts and the instantiated objects.

As indicated above, it is not necessary to attach an ANN to each concept of the semantic network. Rather, one might choose to first train and associate ANNs for objects that occur frequently or that are difficult to recognize by a semantic network. In cases when sufficient training data are not available for a successful training of an ANN, no ANN is bound to the corresponding concept. On the other hand, the hybrid approach gives the option not to fully decompose some of the objects alleviating the effort to acquire and adapt the knowledge base of the semantic network.

3. FORMALISMS FOR NEURAL AND SEMANTIC NETWORKS

The neural network used in the following is the *Local-Linear-Map network (LLM)* [10, 4,5]. This sort of network consists of units that are significantly more complex than the usually employed sigmoid neurons. Therefore, a moderate number of units is sufficient for many tasks.

Each of these units processes the same input vector $\mathbf{x}$ of dimensionality L and computes a node response $\mathbf{y}$ of dimensionality M. Each LLM-unit is characterized by three components: an input weight vector $\mathbf{w}_r^{(in)} \in \mathbb{R}^L$, an output weight vector $\mathbf{w}_r^{(out)} \in \mathbb{R}^M$ and a MxL-matrix $\mathbf{A}_r$. The matrix $\mathbf{A}_r$ implements a locally valid linear mapping. The node response $\mathbf{y}$ of a unit r is determined by

$$\mathbf{y}_r = \mathbf{w}_r^{(out)} + \mathbf{A}_r(\mathbf{x} - \mathbf{w}_r^{(in)}).$$

For computing the final net output two different variants can be distinguished. If the LLM network acts like a "winner-takes-all" network the node response of one single unit is used as final output. Otherwise, a weighted superposition of several node responses is used. The contribution of each node to the superposition can depend e.g. on the distance between the input vector and the input weight vector of the node and can also be influenced by the previous node response. In this way, a short-term memory-effect might be produced.

In our approach the "winner-takes-all" variant is used. The "winner" unit s is determined in this case by comparing the distances $d_r = \|\mathbf{x} - \mathbf{w}_r^{(in)}\|$ between the input vector $\mathbf{x}$ and the input weight vectors $\mathbf{w}_r^{(in)}$, $r = 1, 2, 3 \ldots$, and selecting the node with the minimal distance.

To perform the required transformation between input and output space the necessary values of the network parameters are learned during a training phase. For this purpose correct input-output pairs $(\mathbf{x}^{(\alpha)}, \mathbf{y}^{(\alpha)})$, $\alpha = 1, 2, \ldots T$, of a set of T training samples are presented repeatedly and in a random sequence. The input and output weight vectors as well as the matrix coefficients are adapted according to the following simple error-correction rules:

$$\begin{aligned}
\Delta \mathbf{w}_s^{(in)} &= \epsilon_1(\mathbf{x}^{(\alpha)} - \mathbf{w}_s^{(in)}), \\
\Delta \mathbf{w}_s^{(out)} &= \epsilon_2(\mathbf{y}^{(\alpha)} - \mathbf{w}_s^{(out)}) + \mathbf{A}_s \Delta \mathbf{w}_s^{(in)}, \\
\Delta \mathbf{A}_s &= \epsilon_3(d_s^2)^{-1}(\mathbf{y}^{(\alpha)} - \mathbf{y}^{(net)})(\mathbf{x}^{(\alpha)} - \mathbf{w}_s^{(in)})^T.
\end{aligned}$$

During this adaptation process each training step parameter ϵ_i decays exponentially from a large initial value (typically $\epsilon_i^{initial} = 0.9$) to a small final value (typically $\epsilon_i^{final} = 0.01$).

In the following the semantic network is described. In contrast to other approaches like KL-ONE or PSN, in the ERNEST semantic network language only three different types of nodes and three different types of links exist. They have well defined semantics and we believe that these structures are adequate to represent the knowledge for different pattern understanding tasks. *Concepts* represent classes of objects, events, or abstract conceptions having some common properties. In the context of image understanding an important step is the interpretation of the sensor signal in terms modeled in the knowledge base. The second node type, called *instance*, represents these extensions of a concept. It associates certain areas of the image with concepts of the knowledge base. It is a copy of the related concept where common property descriptions of a class are substituted by values derived from the signal. In an intermediate state of processing instances of some concepts may not be computable because certain prerequisites are missing. Nevertheless, the available information can be used to constrain an uninstantiated concept. This is done via the node type *modified concept*. As in all approaches to semantic networks the *part* link decomposes a concept into its natural components. Another well-known link type is the *specialization* with a related inheritance mechanism by which a special concept inherits all properties of the general one. For a clear distinction of knowledge of different levels of abstraction the link type *concrete* is introduced. In addition to its links, a concept is described by attributes representing mainly numerical features and restrictions on these values according to the modeled term. Furthermore, relations defining constraints for the attributes can be specified and must be satisfied for valid instances.

The creation of modified concepts and instances constitutes the knowledge utilization in the semantic network. For the creation of instances, this process is based on the fact that the recognition of a complex object needs the detection of all its parts as a prerequisite. For concepts which model terms only defined within a certain context the instantiation process must proceed in the opposite direction. In this case the context must exist before an instance of the context-dependent concept can be created. In the network language, these ideas are expressed by six problem-independent inference rules. Context-independent parts, contexts, and concretes are the prerequisites for the creation of instances and modified concepts in a data-driven strategy. The opposite link directions are used for model driven inferences. Since the results of an initial segmentation are not perfect, the definition of a concept is completed by a judgment function estimating the degree of correspondence of an image area to the term defined by the related concept. On the basis of these estimates and the inference rules an A*-like control algorithm is applied. For a detailed description of the network language see [7,3].

4. APPLICATION

The complexity and variety of real scenes poses challenging problems for computer vision systems aiming at a complete description of objects and their relations. Therefore, we choose an incremental design of our system and concentrate on the location and orientation of multifingered anthropomorphic hands in scenes as shown in figure 1. This description is essential information for the control of such manipulators.

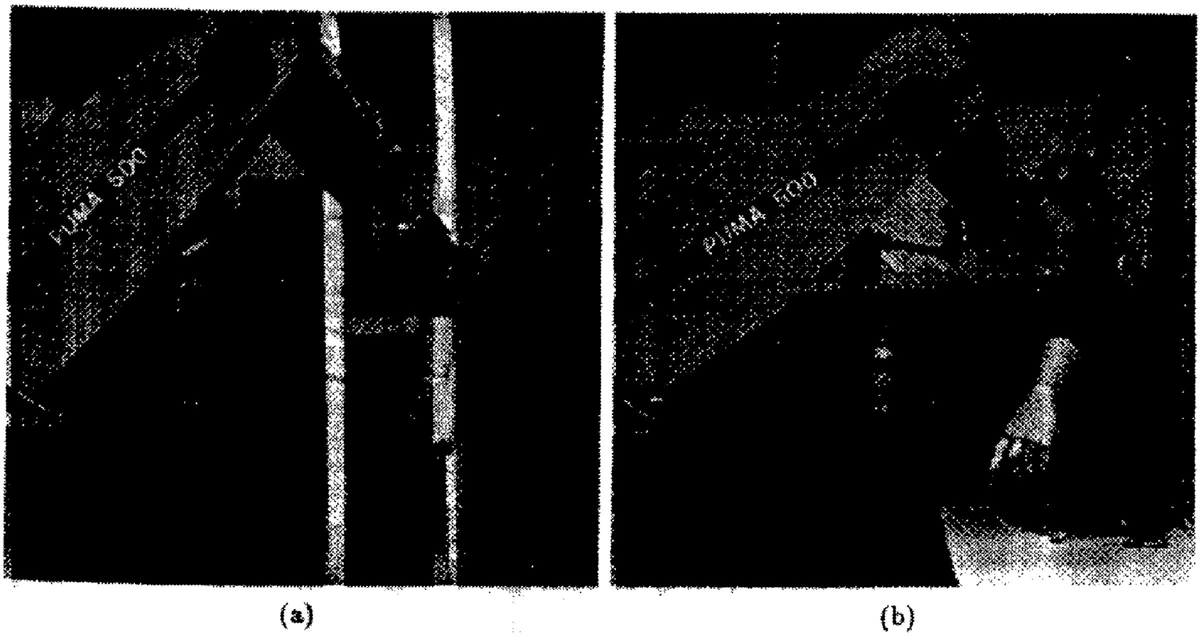

Figure 1. Two examples for left input images shown as grey value images.

The system uses color images as input, either monocular or stereo images. As the result, one or possibly more hypotheses for the location and orientation of a hand are computed. The additional information provided by stereo images can be exploited to enhance performance of the system as will be demonstrated later, imposes however stronger computational requirements. In the rest of this section we describe the declarative knowledge base as shown in figure 2, the attached neural networks, and the procedural knowledge in some detail.

The lowest level of the hybrid network is the concept *INPUT_IMAGE* forming the interface to the input data. The concept *SKIN_COLOR_ACTIVITY_MAP* is realized by an LLM network. The network was trained to map the local color information of each pixel onto a real-valued "skin-color activity" value (see Fig. 3 (a) for an example). This training was based on an additional calibration image that had been segmented manually in hand and non-hand pixels.

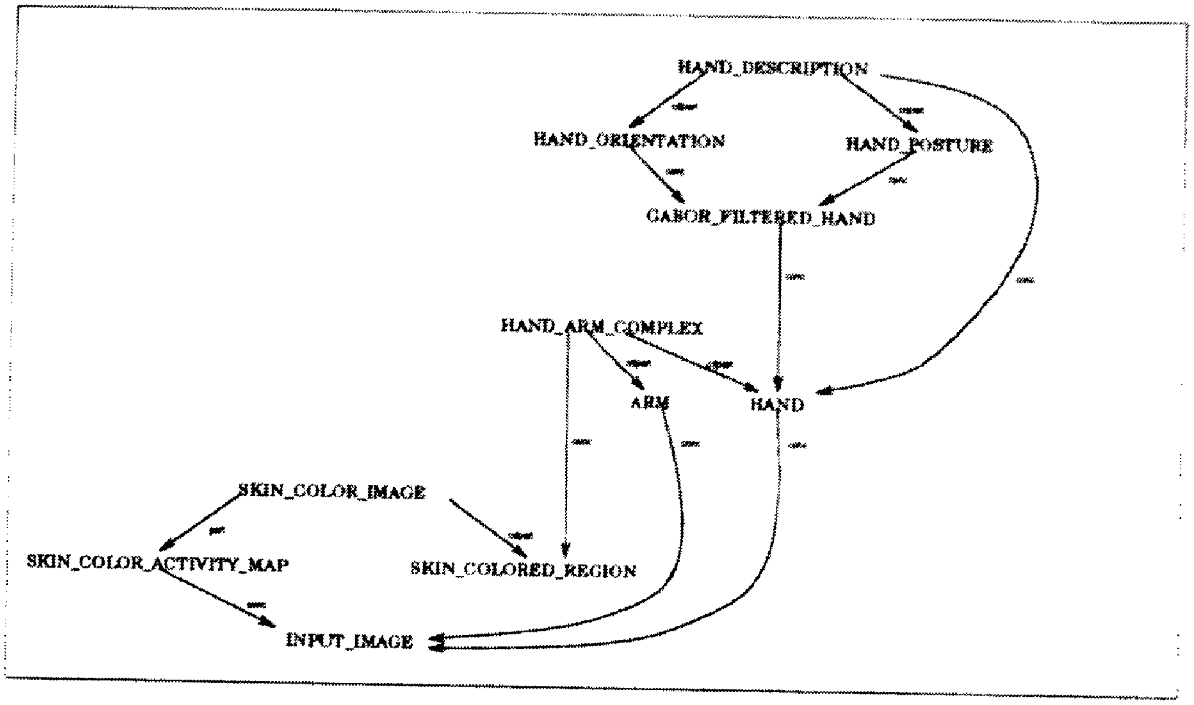

Figure 2. The declarative knowledge base of the hybrid network

The resulting skin-color activity values for all pixels form the *SKIN_COLOR_ACTIVITY_MAP* and is the basis for the detection of hand and arm regions: Applying discriminant analysis to the histogram of this activity map the concept *SKIN_COLOR_IMAGE* calculates a threshold. This threshold is applied to the low pass filtered activity map yielding a binary image with foreground pixels corresponding to skin colored areas. Connected regions of skin colored pixels are candidates for a hand-arm-complex and are represented as instances of *SKIN_COLORED_REGION*. (see figure 3 (b) for an example). However, different skin colored objects overlapping in the image plane cannot be separated using only color information. Yet in most cases overlapping objects are separated in 3D space and depth information may be exploited to discriminate the objects. To acquire three-dimensional information we employ a contour based stereo algorithm (see [8,9]). Straight lines or polygons are extracted only for foreground pixels of the *SKIN_COLOR_IMAGE* and matched subsequently. To confine matching to pairs of regions is not feasible, since a region in one image may correspond to more then one region in the other image due to different viewpoints. Nevertheless, focusing on a subset of all contours in the image alleviated the correspondence problem for the stereo algorithm. The resulting sparse disparity map has to be expanded to all foreground pixels. If only few contours exist within skin colored objects, this disparity map may not be reliable for subsequent splitting of regions since disparity can only be computed for contour pixels. Assuming that two objects are projected into the region and that the depth (and hence disparity) of each object does not vary significantly, discriminant analysis is applied to the histogram of disparity values within each region. The resulting threshold is used to split the regions. At this point of the analysis no attempt is made to decide whether overlapping objects indeed occured

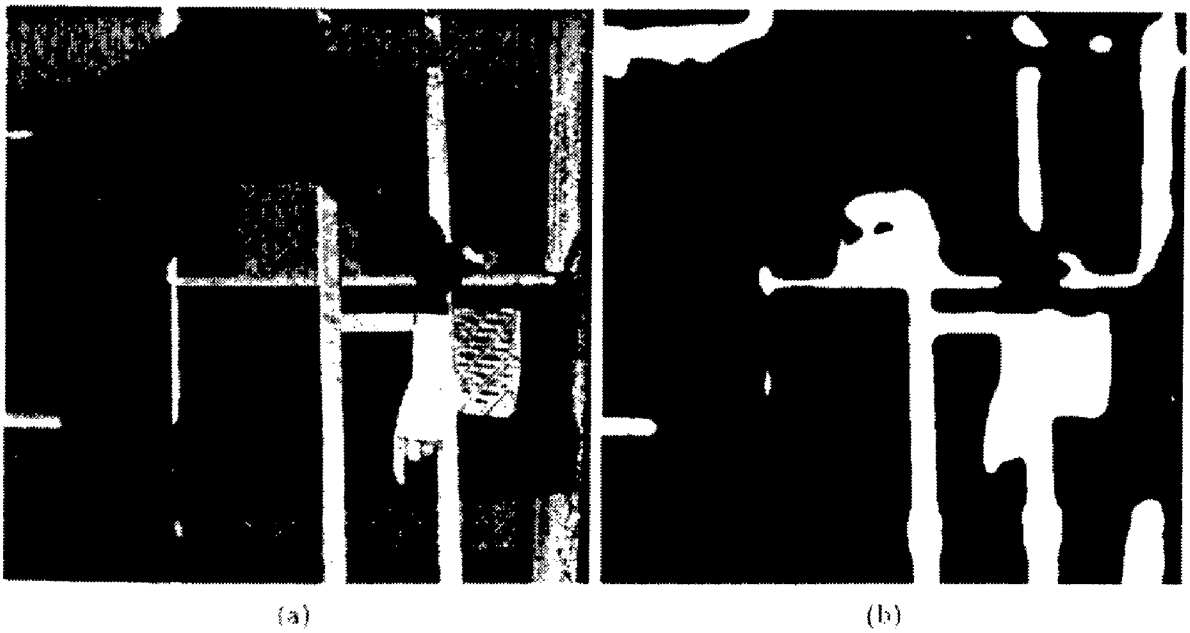

Figure 3. (a) Skin-color activity map for the image shown in figure 1 (a); (b) connected regions of skin colored pixels as candidates for a hand-arm-complex for the same scene

or not. Hence, for each connected region derived from the binary skin color image, the region itself and all subregions computed from the disparity information yield instances of *SKIN_COLORED_REGION*. All theses hypotheses are further processed and judged later on. Since *SKIN_COLORED_REGION* is a concrete of *HAND_ARM_COMPLEX*, each instance of the former gives rise to an instance of the latter. As an example, in figure 4 a subwindow of the image in figure 1 (a) is displayed and the two competing regions for a hand-arm-complex detected in this subwindow using disparities are shown. The union of both regions forms the third hypothesis corresponding to the non overlapping case.

The concepts *HAND* and *ARM* are modeled as context-dependent parts of *HAND_ARM_COMPLEX*. To separate the arm from the hand the hand-arm-complex is transformed into the normalized orientation by a Karhunen-Loeve transformation. Based on general knowledge about the shape and proportions of hands and arms the hand is extracted from the complex. For each of the competing hand hypotheses a judgment has to be computed in order to control the interpretation process. This is done by projecting each hand candidate into a subspace spanned by "eigenhands" derived from a test set of hands in an approach adopted from the "eigenface approach" (see [11]). The concept *ARM* is not considered further for our application.

For our task the *HAND_DESCRIPTION* is composed of the *HAND_ORIENTATION* (including location) and *HAND_POSTURE*, the latter one not being considered for the current state of the system. The concept *HAND_ORIENTATION* is again modeled by

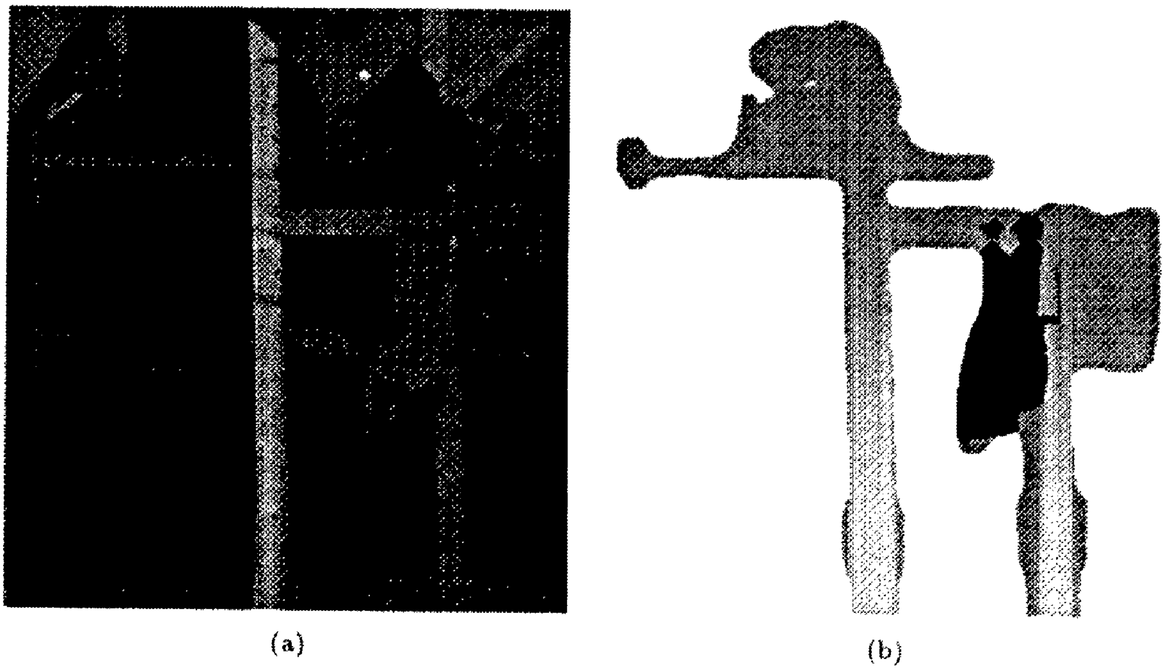

Figure 4. (a) A subwindow of the image in figure 1 (a); (b) the competing regions for a hand-arm-complex

an LLM network. This network operates on a feature space represented in the concept *GABOR_FILTERED_HAND*: A Laplace operator is applied to the hand and negative filter values are clipped. Then a Gabor filter with four different orientations operates on a 3×3 "filter grid", resulting in a 36-dimensional feature vector. From this feature representation the LLM network attached to the concept *HAND_ORIENTATION* calculates the two angles characterizing the orientation in our setup.

5. RESULTS

To evaluate our approach we use a set of 300 stereo images with 512×512 color pixels each. Examples are shown in figure 1. The images were taken in 15 groups of 20 images each. For each group the basic setup of the scene is constant, while the end effector of the robot was positioned in each image in a random orientation. This orientation was generated by two subsequent rotations around a local coordinate system of the hand. First the z-axis is rotated in the range of $[-40°, 40°]$, followed by a rotation of the resulting y-axis in the range of $[10°, 70°]$. The z-axis is approximatly aligned with the middle finger of a stretched hand, while the y-axis is roughly defined by the finger tips.

An additional calibration image was used for the training of an LLM network to detect skin-colored pixels. The current LLM network for this task consists of one layer containing 50 nodes. It is trained to provide an output value of 1 for pixels belonging to the hand-arm-complex and -1 otherwise. 200.000 adaptation steps were carried out by randomly drawing pixels from the calibration image. In contrast to [1], the chosen network and input representation is rather simple to guarantee robust recognition of hand pixels at the expense of the specificity of the mapping, thus yielding large hand activity values for wooden objects.

The basic setup of the 15 groups of scenes was chosen to challenge the system with frequent overlaps of the robot hand with other skin colored objects. As a result, 176 of the 299 scenes[2] show such an overlap for the left image.[3] For these images the correct hand description cannot be derived using only monocular images as in [1,2], but can potentially be handled using stereo images. The evaluation of the segmentation results show that for a total of 230 scenes the system was able to correctly segment the hand region.[4] Different types of problems can be identified analysing the remaining 69 scenes: About 75% of theses scenes are from one of four groups. In two groups the overlapping objects have no contours in the vicinity of the border between hand and object. Therefore, no reliable disparity values can be derived near these borders with the contour based stereo algorithm. In another group there were two skin colored objects located in the background, overlapping the hand at both sides. In addition, these objects have repetitive patterns and positional reversal of scene points along a scanline occurs. This results in difficult conditions for stereo algorithms. In the fourth group, the overlapping objects have about the same depth, and hence cannot be separated using disparity information. The remaining errors are due to wrong stereo matching or problems in correctly thresholding the disparity histogram. For the remaining discussion we consider only the 230 correctly segmented scenes.

As mentioned in the previous section, we use the "eigenhand" approach to judge all competing instances of the concept *HAND* in a given images. To train these eigenhands, a given set of training images is scaled to size mxn and the eigenvectors are computed, considering the scaled images as vectors of length mn. The k eigenvectors with largest eigenvalues constitute the set of eigenhands spanning the k-dimensional subspace of hands. In order to judge for a given region the similarity to a hand, the region is again scaled and then projected into this subspace of hands. The error of the projection, i.e. the distance to the hand space yields the required judgment. The training set was derived in two steps: First, for each of three scenes ten images were chosen and a tentative set of 15 eigenhands derived. In the second step for each of the remaining 200 scenes with correct hand segmentation, the distance of the hand was compared to the distance of the competing regions. For 165 scenes the correct hand region had the smallest distance, thus the best judgment. The inital training set was augmented with the other 35 images and the final set of eigenhands computed. Figure 5 shows the mean of all eigenhands and the eigenhand with largest eigenvector of the eigenhands we used. For 3, 22, and 58 eigenhands corresponding to the largest eigenvalues, table 1 shows the number of scenes where the hand region yields the smallest distance and therefore is returned as the best hypothesis for the hand location. In most cases the correct region is found, while the number of eigenhands has no significant influence on the result, indicating that the hands are located in a low dimensional subspace.

The subsequent orientation detection works on these areas of interest. The data set is randomly split into two subsets: 150 images are used for training, the remaining 80

[2] One of the originally 300 images was unusable due to technical problems

[3] The left image is used to locate the hand and determine its orientation

[4] The correctness was judged visually and is therefore subjectivly to a certain degree.

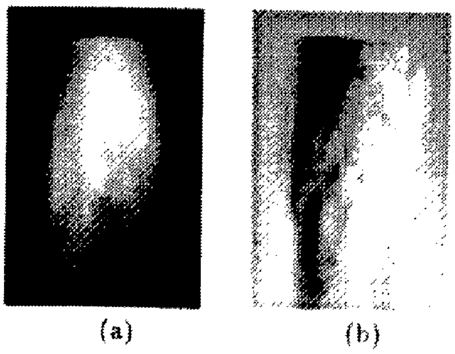

Figure 5. (a) Mean of all eigenhands and (b) the eigenhand with largest eigenvalue of the eigenhands from the final training set

Table 1
Absolut number and percentage of the correctly segmented 230 scenes, where the eigenhand projection yields the correct hand region as the best hypothesis. The results are given for 15 eigenhands of the tentativ set of eigenhands and 3, 22, respectivly 58 eigenhands of the final set.

	15 tentative eigenhands	3 eigenhands	22 eigenhands	58 eigenhands
absolut	195	215	220	222
relativ	0.85	0.93	0.96	0.97

Table 2
Performance of the LLM network for the recognition of orientation on the training/test set presenting the mean square root error as an absolute value as well as normalized by the standard deviation of the data sets.

	MSRE	NMSRE
training	3.76°	0.139
test	7.54°	0.270

images form the test set. The task is accomplished by an LLM-network consisting of five units. For the 36-dimensional input space the training set contains very few examples. Therefore, the network is trained with the relatively small number of 40.000 adaptation steps only, since more training iterations would lead to overfitting.

The performance of the orientation LLM network is represented in Table 2. Evaluated on the independent test set, the mean square root error (MSRE) for the Euclidean distance in both dimensions is 7.54°. Normalizing the MSRE with the corresponding standard deviations of the data sets, we obtain a total NMSRE of 0.270. These results are presented graphically in figure 6.

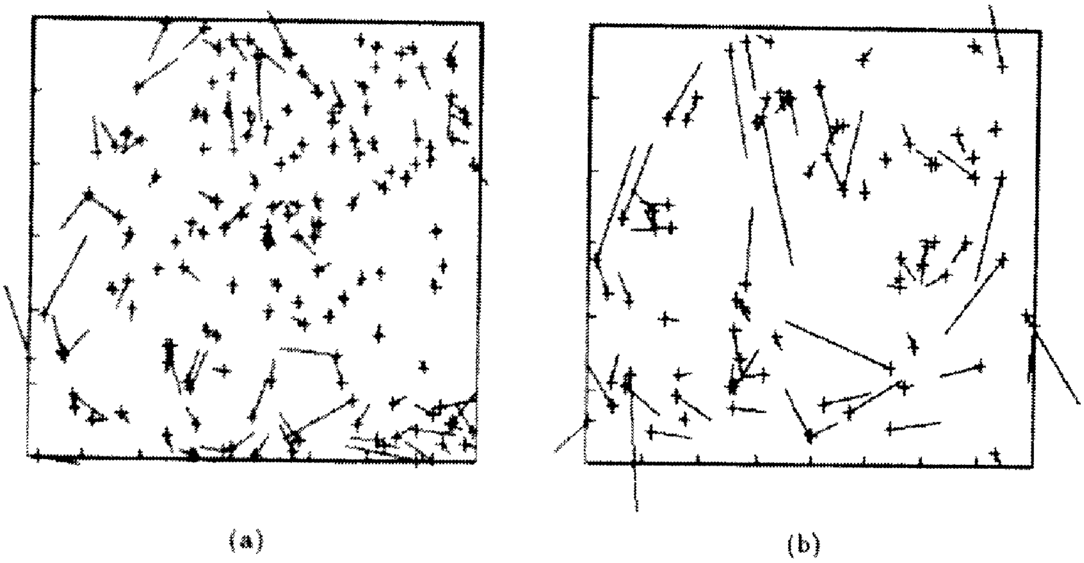

Figure 6. Representation of the orientation error on the training set (a) and the test set (b). The center of a cross is the target point whereas the length of the tail represents the size of the total error for rotation about z- and y-axis in degree. The rotation about z-axis in the range of $[-40°, 40°]$ is shown on the abscissa, the rotation about the y-axis in the range of $[10°, 70°]$ on the ordinate.

6. CONCLUSION

In this work we describe a hybrid system for the analysis of complex scenes combining semantic and artificial neural networks. The ANNs are attached to concepts of the semantic network adding a holistic model of a given object to the decompositional view of the semantic network. Given such a hybrid knowledge base, different analysis strategies can be realized. In this way, the robust holistic recognition by ANNs is combined with explicit structuring of domain knowledge resulting in a flexible and powerful analysis system. The same technique can also be applied to other problems in computer vision and tasks of signal interpretation in general.

The system is successfully applied to the problem of recognition of a hand and the determination of its 3D-orientation in complex real world scenes. Using stereo images as input, the system is able to correctly process many scenes with overlapping skin colored objects which cannot be handled by a monocular system as realized previously in ([1,2]).

In further work we will explore issues of control strategies to exploit the potentials of holistic and decompositional modeling in more detail. Additionally, the knowledge base of the system is incrementally expanded. A complementing direction of research aims at realizing procedural knowledge of the semantic network by neural networks, like computation of attributes or judgments ([6]). The resulting tighter coupling of both network types will be combined with the approach described in this paper.

REFERENCES

1. F. Kummert, E. Littmann, A. Meyering, S. Posch, H. Ritter, and G. Sagerer. A hybrid approach to signal interpretation using neural and semantic networks. *Proceedings 15. DAGM-Symposium*, pages 245–252, 1993.
2. F. Kummert, E. Littmann, A. Meyering, S. Posch, H. Ritter, and G. Sagerer. Recognition of 3d-hand orientation from monocular color images by neural semantic networks. *Pattern Recognition and Image Analysis*, 3(3):312–316, 1993.
3. F. Kummert, G. Sagerer, and H. Niemann. A Problem-Independent Control Algorithm for Image Understanding. In *11th International Conference on Pattern Recognition*, volume I, pages 297–301, The Hague, 1992.
4. Enno Littmann, Andrea Meyering, and Helge Ritter. Cascaded and parallel neural network architectures for machine vision – a case study. *Proceedings 14. DAGM-Symposium*, pages 81–87, 1992.
5. Andrea Meyering and Helge Ritter. Learning 3d-hand postures from perspective pixel images. In I. Aleksander and J. Taylor, editors, *Artificial Neural Networks II*, pages 821–824. Elsevier Science Publishers (North Holland), 1992.
6. R. Moratz, S. Posch, and G. Sagerer. Controlling multiple neural nets with semantic networks. *Proceedings 16. DAGM-Symposium*, page to appear, 1994.
7. H. Niemann, G. Sagerer, S. Schröder, and F. Kummert. ERNEST: A Semantic Network System for Pattern Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9):883–905, 1990.
8. Stefan Posch. *Automatische Tiefenbestimmung aus Grauwertstereobildern*. Deutscher Universitäts Verlag, Wiesbaden, 1990.
9. Stefan Posch. Stereozuordnung mit geraden Liniensegmenten und Polygonen. *Proceedings 14. DAGM-Symposium*, pages 385–391, 1992.
10. Helge Ritter. Learning with the self-organizing map. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks*, pages 379–384. Elsevier Science Publishers (North Holland), 1991.
11. Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.