A Hybrid Object Recognition Architecture

Gunther Heidemann, Franz Kummert, Helge Ritter, Gerhard Sagerer

University of Bielefeld,
33501 Bielefeld,
Germany

Abstract. We present an architecture for 3D-object recognition based on the integration of neural and semantic networks. The architecture consists of mainly two components. A neural object recognition system generates object hypotheses, which are verified or rejected by a semantic network. Thus the advantages of both paradigms are combined: in the low level field adaptivity and the ability to learn from examples is realized by a neural network, whereas the high level analysis is performed by representing structured knowledge in a semantic network.

1 Introduction

One of the main reasons that make 3D-object recognition an extremely hard task in computer vision is that knowledge acquisition and representation has to cover the wide range from very low level (pixel) data up to a high level symbolic representation. On the one hand there are sensor data which are hard to describe by explicit models, but can be classified holistically by an artificial neural network (ANN), on the other hand the structure of objects often is too complex for a pure holistic recognition, but can be modeled explicitly in a semantic net. Therefore, it seems reasonable to combine the benefits of ANNs and semantic networks in a hybrid approach. Knowledge about the objects that can be well structured such as (in our case) shape is represented by a semantic net whereas the bridge to the pixel data is realized using a neural object recognition system that can be trained from examples.

2 The hybrid object recognition architecture

In our approach, the hybrid system performs object recognition in mainly three steps: 1. a low level preprocessing and search for regions of interest, 2. generation of object hypotheses by the neural recognition system, 3. knowledge based analysis and verification or rejection of the hypotheses by the semantic network.

In the low level part, first a segmentation for colors of the domain of the objects is performed. From this we get regions of interest which are the basis for both the neural and the semantic analysis. The semantic net operates on features of the regions such as eccentricity and compactness. Moreover, in the low level preprocessing the monochrome intensity image is transformed to an edge

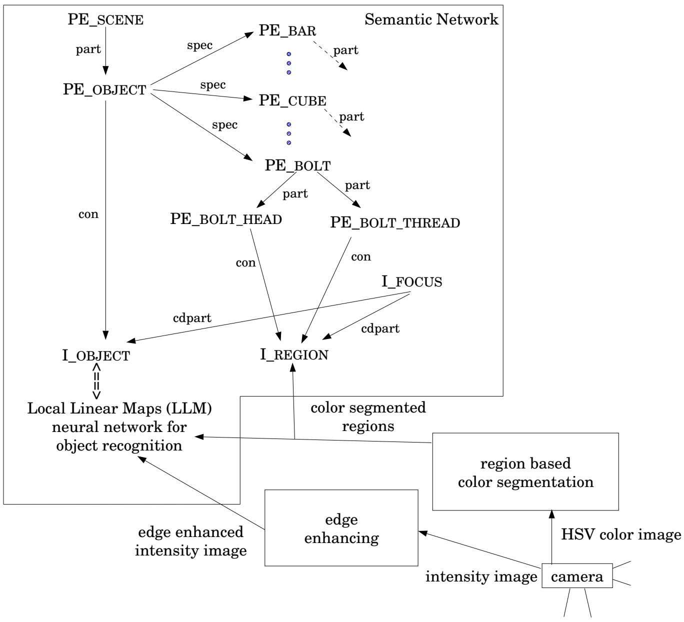

Fig. 1. The hybrid object recognition architecture. Only a part of the knowledge base is shown.

enhanced image by laplace filtering and a subsequent logarithmic transformation. From this image the features for the neural classification are extracted by Gabor filter kernels within the regions of interest.

The neural system then tries to classify the features extracted and generates up to three competing object hypotheses, combined with a judgment. By this means the search space of the semantic net can be directed, and search is started with the hypotheses with highest probability. The semantic net then tries to verify or reject the hypotheses by decomposition of the objects according to the knowledge base. In other words, it is the task of the ANN to have a "first look" at the scene and give an overview quickly, which is the starting point for a closer inspection by the semantic net.

2.1 The neural object recognition system

From the low level color segmentation the neural system gets the blob centers as "focus points". At each focus point, a feature vector is extracted by currently 16 Gabor filter kernels. The parameters of the Gabor filter kernels (location with respect to the focus point, width, wavevector and phase) were optimized to the classification task by a method outlined in [5, 1]. In short, the proposed algorithm optimizes the parameters of the filter kernels by (local) minimization of an energy function on the parameters, which is constructed such that the extracted feature vectors belonging to one type of object tend to cluster in feature space, whereas clusters belonging to different object types are separated as far as possible.

Classification of the feature vectors is performed by an ANN of the Local Linear Map (LLM) – type. The LLM network is related to the self-organizing map [2] and the GRBF approach. It can be trained to approximate a nonlinear function by a set of locally valid linear mappings, for details see e.g. [6]. For the classification task we use a "winner takes all" network. In this case, for a given input $\mathbf{x}$ only one node, the best match or "winner" node k, contributes to the output vector $\mathbf{y}$:

$$\mathbf{y} = \mathbf{w}_k^{out} + \mathbf{A}_k(\mathbf{x} - \mathbf{w}_k^{in}), \tag{1}$$

where $\mathbf{w}_k^{in}$ and $\mathbf{w}_k^{out}$ are the input and output weight vectors of node k, respectively. The input space has in our case as dimension the number of Gabor kernels n_G, the output space is (n_O+1) dimensional, this is the number of object classes n_O plus one as a rejection class. Therefore, $\mathbf{A}_k$ is a $(n_O+1) \times n_G$ matrix associated with node k.

For the training, the output vector of training example α has the form

$$y_i^{(\alpha)}(l) = \delta_{il}, \quad \text{with} \quad i, l = 1 \dots n_O + 1, \tag{2}$$

where l is the class of the object to be trained. When applying the network to an unknown input vector, the resulting class o_{res} is determined by

$$o_{res} = \arg \max_{i=1 \dots n_O+1} (y_i). \tag{3}$$

The main limitation of the neural recognition system is the lack of a universal rejection class. Up to now, only objects trained to be rejected will be classified correctly. However, the semantic analysis is able to reject completely unknown objects in most cases.

2.2 Knowledge based object recognition

The semantic network language ERNEST [3] builds the basis for knowledge representation and utilization. In contrast to other approaches like KL-ONE or PSN in ERNEST only three different types of nodes and three different types of links exist. *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. In an intermediate state of processing instances 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 link *part* decomposes a concept into its natural components. Another link type is the *specialization* with a related inheritance mechanism by which a 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. Additionally, a concept is described by attributes representing numerical features and restrictions on these values according to the modeled term. Furthermore, relations defining constraints for 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. 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.

In the following the declarative knowledge base (see Fig. 1) and the processing strategy is described in some detail. Actually, the network consists of two levels of abstraction namely the image level (indicated by the prefix I_) and the level of perception (indicated by the prefix PE_). The concept I_FOCUS was motivated by the works of Moratz, see e.g. [4]. It mainly allows to focus on certain areas in the image to restrict the object recognition task. This focus can be established by an utterance or a gesture during the construction dialogue (actually not yet considered) or by the objects detected so far. This concept has two context–dependent parts namely I_REGION representing a color segmented region and I_OBJECT representing an object hypothesis calculated by the underlying LLM network. For every competing LLM hypothesis an instance I_OBJECT$^{(I)}$ is created which are stored in competing search tree nodes. Dependent on the object type detected by the LLM network the corresponding concept in the perceptual level is selected to verify the object hypothesis due to the structural knowledge stored in the semantic network. That means if an instance I_OBJECT$^{(I)}$ with type 'bolt head' exists then a modified concept PE_BOLT$^{(M)}$ is created with the concretization I_OBJECT$^{(I)}$. This link is inherited by the concept PE_OBJECT. In the next step, the control algorithm tries to detect the parts of a modified perceptual object as they are modeled in the semantic network. For our bolt example this yields in instances for 'bolt head' and 'bolt thread' which are concretized by one instance of I_REGION respectively. Currently, these instances are based on the regions detected by the preprocessing. But we are working on an object-dependant segmentation relying on inter-object comparisons. During the instantiation process restrictions for position, color and shape are propagated in a model–driven way. Additionally, the restrictions of the actual focus are taken into account. If a successful instance of a perceptual object is created then it is added as part of PE_SCENE which refers to all objects in the scene detected so far. After this step, the focus is adapted due to the newly detected object and the next object hypotheses are processed.

3 Results

The proposed architecture has been investigated so far in the scenario of SFB 360. The task is the recognition of a set of wooden toy pieces ("Baufix"), which are a "bar" with three, five or seven holes, a "felly", a "cube", a "rhomb-nut" a "tyre", a "socket", a "ring", and "bolts" with round or hexagon head. The bolts have four different lengths. The objects may be freely arranged within the

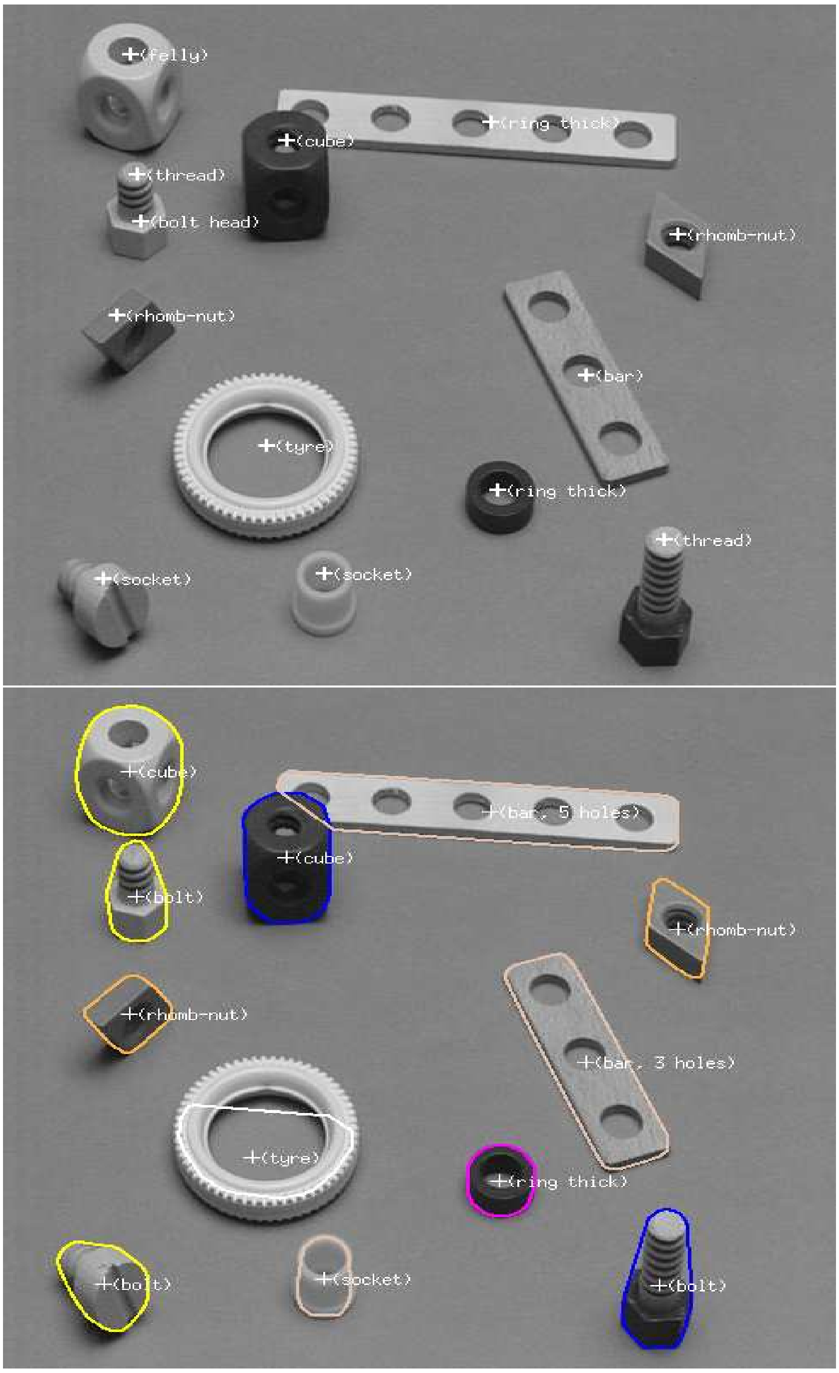

Fig. 2. Above: Wooden toys with best judged object hypotheses from the neural system, below: region boundaries and correct classification by the semantic analysis

range of a table from where the training images were taken as long as there is no occlusion, see Fig. 2. As a training set for the neural recognition system, 50 images of each part were used, for the bolts 200 images were used. On the training images the parts are arranged in different views and distances from the camera. By this way rotational invariance and scaling up to 30% were trained. For the LLM, 40 nodes approved to be the optimum.

The misclassifcation rate of the neural system is about 20%, it is reduced by the semantic analysis to about 10%.

4 Conclusion, outlook, and acknowledgement

We have presented a hybrid architecture for 3D-object recognition. Due to the hybrid architecture, knowledge acquistion becomes simple because using a semantic net we have the possibility of structuring, but avoid the difficulty of modeling knowledge about the sensor data explicitly by use of a neural network. By this means robustness and computational efficiency can be achieved.

Up to now the system is bound to a special geometric situation, because the ANN is trained only to a limited range of camera distance and angles. Adding an initialization phase, in which the semantic analyzer checks for camera distance and angle without help of the ANN, we want to get the parameters needed to choose a specialized ANN for the specific situation. After this initialization phase, the system could run as described here. This will be the aim of further investigation.

This work has been supported by the German Research Foundation (DFG) under SFB 360.

References

1. G. Heidemann and H. Ritter. A Neural 3-D Object Recognition Architecture Using Optimized Gabor Filters. In *Proceedings of 13th International Conference on Pattern Recognition, Vienna*, volume IV, pages 70–74. IEEE Computer Society Press, 1996.
2. T. Kohonen. Self-organization and associative memory. In *Springer Series in Information Sciences 8.* Springer Verlag Heidelberg, 1984.
3. F. Kummert, H. Niemann, R. Prechtel, and G. Sagerer. Control and Explanation in a Signal Understanding Environment. *Signal Processing, special issue on 'Intelligent Systems for Signal and Image Understanding'*, 32:111–145, 1993.
4. R. Moratz, H.J. Eikmeyer, B. Hildebrandt, A. Knoll, F. Kummert, G. Rickheit, and G. Sagerer. Selective visual perception driven by cues from speech processing. In *7th Portuguese Conference on AI, EPIA95, Workshop on Applications of AI to Robotics and Vision Systems*, pages 63–72, Portugal, 1995. Trans Tech Publications. Ltd.
5. H. Ritter, G. Sagerer, G. Heidemann, and R. Moratz. Hybride Wissensrepräsentation: neuronale und semantische Netzwerke für die Bildanalyse. In *Arbeits- und Ergebnisbericht*, pages 27–65. Universität Bielefeld, SFB 360, 1995.
6. H.J. Ritter, T.M. Martinetz, and K.J. Schulten. *Neuronale Netze.* Addison-Wesley, München, 1992.