ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 301-306

Combining Gestural and Contact Information for Visual Guidance of Multi-Finger Grasps

Gunther Heidemann,* Helge Ritter
University of Bielefeld, AG Neuroinformatik,

D-33501 Bielefeld, Germany

Abstract. A computer vision system for a three-fingered robot hand is presented which can solve two entirely different tasks at a time: First, to guide the robot hand, hand gestures of a human instructor are classified using the hand camera. Second, when an object has been grasped the success or failure of the grasping action can be judged qualitatively by the same system. Both tasks are solved using a view based approach which classifies a set of prototypical situations instead of exact geometric reconstruction.

1 Introduction

A challenging subject in robotics is the "intelligent" grasping of objects with anthropomorphic artificial hands. Computer vision (CV) plays an important role in this field because it serves at least two different purposes: (1) CV has to solve tasks for which humans use their eyes, e.g. the overall localization of objects or the recognition of guiding hand gestures carried out by another person. (2) Apart from this, CV often has to replace human haptic capabilities.

The second point must be explained in more detail: As haptics are essential for the human ability to grasp arbitrarily shaped objects, equivalent technical solutions have to be found. However, though it is no problem to provide high quality sensors in general, sensors the size and shape of a human fingertip are extremely difficult to produce (and apply). In our experience such fingertip sensors are still unreliable and coarse in measurement. Consequently, other sensory sources have to be exploited. A suitable means is visual control as miniature cameras and image processing hardware are easily available. Our approach is therefore to supplement fingertip sensing by a visual check of the points where the fingers touch the object.

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) within research project "Hybrid Knowledge Representation" of SFB 360, Bielefeld.

*email: gheidema@techfak.uni-bielefeld.de

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 301-306

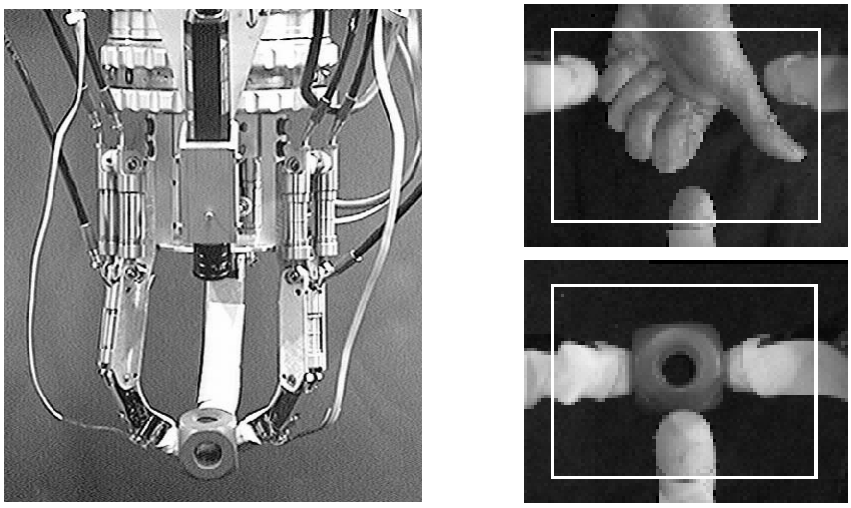

Figure 1: Left: The three-fingered hydraulic hand with camera mounted in front, holding an object. Right: View of the hand camera, above gesture of the instructor (robot hand almost open), below grip on an object (hand closed). The input window for the VPL-classifier is marked.

We address both "types" of robot vision in this contribution: the recognition of human hand gestures for guidance of the robot manipulator towards an object and the visual judgment of grasping positions. In an anthropomorphic scenario different cameras would be used as "eyes" (overview) for gesture recognition and "haptics" (hand camera) for grasping tasks. In contrast, both tasks are carried out by the hand camera in our scenario. This setup is technically simpler and enables a direct interaction close to the object. Moreover, coordinate transformations between two different camera systems can be avoided. However, this approach requires a CV system that can solve two entirely different recognition tasks at a time.

In the following, we will first describe shortly the robot setup (which itself is not the topic of this paper), then, the CV system will be outlined. It relies on a view based approach which uses local principal component analysis for feature extraction in combination with neural expert classifiers. The focus of this contribution is on how visual "classes" or "categories" can be defined to describe the hand gestures as well as the grasping positions.

2 The three-fingered robot hand

The robot system consists of a standard industrial robot arm (PUMA 560) with six degrees of freedom and an oil-hydraulic robot hand developed by Pfeiffer et al. [6]. The hand has three equal fingers mounted in an equilateral triangle, pointing in parallel directions (Fig. 1). Each finger has three degrees of freedom:

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 301-306

bending the first joint sideways and inward at the wrist, and bending the coupled second and third joint. The oil pressure in the hydraulic system serves as sensory feedback. A detailed description of the setup can be found in [4]. The force/torque wrist sensor which is also shown in Fig. 1 is not used in these experiments. The sensed hydraulic oil pressure can be used as feedback on the success of a grasping movement. However, these data allow only to detect if the object is in the grasp at all or if it was lost. A reliable estimation on the stability of the object position cannot be achieved, since this requires information about the fingertip positions. Though there are additional position sensors on the hydraulic motor pistons, the fingertip positions can not be estimated with the required accuracy due to mechanical hysteresis. Hence, the camera mounted in front (Fig. 1) has to evaluate the position of the object.

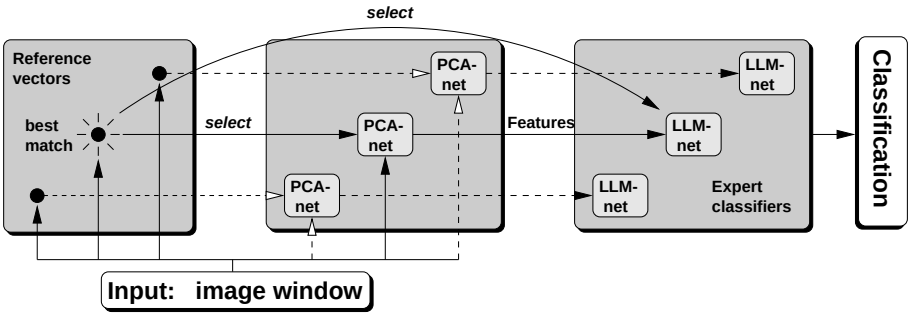

Figure 2: The neural VPL-classifier: To classify an input window, first the best match reference vector is determined. Then the input window is projected to the local principal components which have been calculated from all training data within the Voronoi tessellation cell of this reference vector. The projection is subsequently classified by the neural network (LLM-net) associated with the reference vector.

3 The visual recognition system

The grey value camera output of 768×576 pixels is first sub-sampled by a factor 8 to 96×72 which is sufficient for both recognition tasks. Of the sub-sampled image a slightly smaller (fixed) window of 81×59 pixels serves as input to the classifier (marked in Fig. 1).

3.1 VPL-classifier

For the classification of the image windows, a neural classifier is applied which performs a mapping $\vec{x} \rightarrow \vec{y}, \vec{x} \in \mathbb{R}^M, \vec{y} \in \mathbb{R}^N$. The input dimension M is the number of pixels of the window. The window vector $\vec{x}$ has to be mapped to a discrete valued output k that denotes the class. There is one separate output

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 301-306

channel for each of the N output classes. Training is performed with hand-labeled sample windows $\vec{x}_i^{Tr}$ and binary output vectors $y_i^{Tr} = \delta_{ij}, i = 1 \ldots N$, coding the class $1 \leq j \leq N$. Classification of unknown windows $\vec{x}$ is done by taking the class k of the channel with maximal output: $k = \arg\max_i(\vec{y}_i(\vec{x}))$.

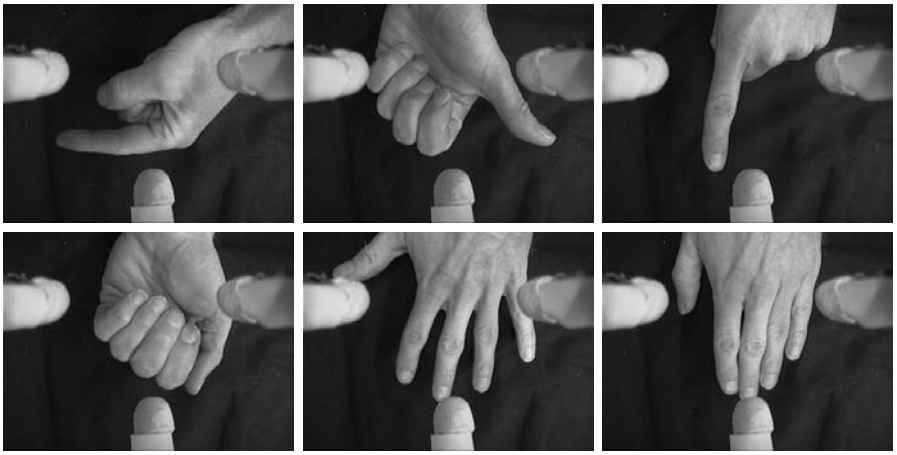

Figure 3: Hand gestures used to guide the robot in the three translatory DOG, as seen from the hand camera (from upper left to lower right): "Move: left, right, away, towards instructor, up, down".

To reduce and exploit redundancy of the input data, the VPL-classifier (Fig. 2) combines visual feature extraction and -classification. It consists of three processing stages which perform a local principal component analysis (PCA) as feature extraction followed by a classification by neural expert networks. Local PCA can be viewed as a nonlinear extension of simple, global PCA [8]. "VPL" stands for the three stages: **V**ector quantization, **P**CA and **L**LM-network. The vector quantization is carried out on the raw image windows to provide a first data partitioning with N_V reference vectors $\vec{r}_i \in \mathbb{R}^M, i = 1 \ldots N_V$. We apply for vector quantization the Activity Equalization algorithm, which is an extension of the well-known "winner takes all" method [3].

To each reference vector $\vec{r}_i$ a single layer feed forward network for the successive calculation of the principal components (PCs) as proposed by Sanger [7] is attached which projects the input $\vec{x}$ to the $N_P < M$ PCs with the largest eigenvalues: $\vec{x} \rightarrow \vec{p}_l(\vec{x}) \in \mathbb{R}^{N_P}, l = 1 \ldots N_V$. To each of the N_V different PCA-nets one expert neural classifier of the Local Linear Map – type (LLM-network) is attached, see e.g. [1] for details. It performs the final mapping $\vec{p}_l(\vec{x}) \rightarrow \vec{y} \in \mathbb{R}^N$. The LLM-network is related to the self-organizing map [5]. It can be trained to approximate a nonlinear function by a set of locally valid linear mappings.

The three processing stages are trained successively, first vector quantization and PCA-nets (unsupervised), finally the LLM-nets (supervised). Classi-

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 301-306

fication of input $\vec{x}$ is carried out by computation of the best match reference vector $\vec{r}_{n(\vec{x})}$, mapping $\vec{x}$ to $\vec{p}_{n(\vec{x})}(\vec{x})$ by the attached PCA-net and finally mapping $\vec{p}_{n(\vec{x})}(\vec{x}) \to \vec{y}$.

The VPL-classifier is applied in this scenario because of its ability to deal efficiently with high dimensional data: The local PCA leads to many highly specific feature detectors (the $N_V \cdot N_P$ local PCs), but for a single classification only $N_V + N_P - 1$ filter operations are needed. Performance and generalization properties of the VPL-classifiers proved to be well behaved for several vision tasks, for details see [1].

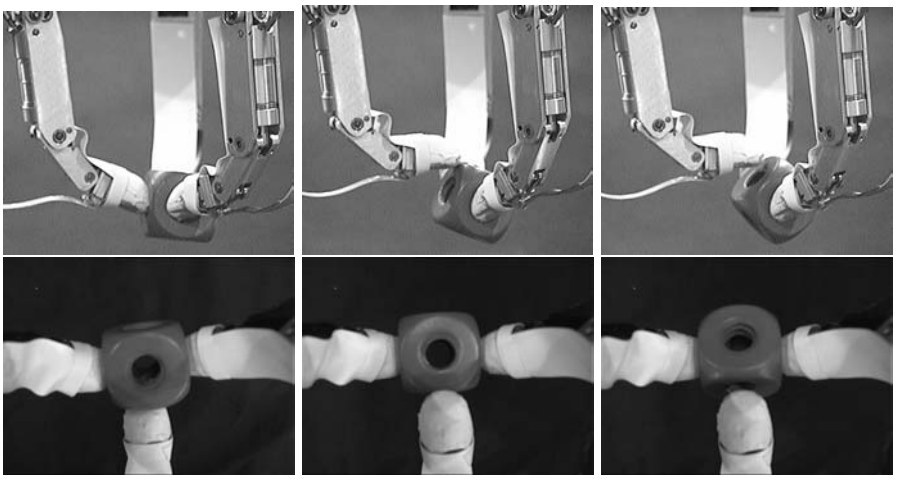

Figure 4: Grasping positions of different quality, above view from the side, below view of the hand camera. Left: Object held correctly by all fingers. Middle: Left finger has lost its grip, but object still stable. Right: Like (Middle), but object in a different position. (From [2].)

3.2 Choice of visual classes

One of the major benefits of the VPL-classifier is that it can be used for entirely different visual classification tasks. Hence, there is no need for a prior decision between either hand gestures or object poses. Instead, all hand gestures and object poses are treated equally. There are six different classes for hand postures (Fig. 3), which allow to give visual commands for translatory movements (left–right, away–come, up–down).

We chose five different objects to be grasped. For each object two classes were defined: grasp stable or unstable. In this distinction human expert knowledge about the physics of grasping can be brought in, see Fig. 4 for examples of grasp stability. In total, there are $6 + 5 \cdot 2 = 16$ classes which give information about the identity of the hand gesture or object, and, in case of an object, the stability of the grasp.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 301-306

4 Results and outlook

The system was tested on a series of 130 images of a human hand giving instructions and a series of 60 images of the grasped objects. We used more hand images due to the greater variability of hand postures. Moreover, the human hand is in some cases partially occluded by the robot fingertips. The 81×59 input window was classified by a VPL for which $N_V = 9$ reference vectors, $N_P = 6$ local PCs and $N_L = 16$ LLM-nodes proved to be the optimal, in particular significantly better than $N_V = 1$ (the "normal" – PCA case).

The system was tested using the "leaving one out" – method on the 190 images and achieved 86.3% correct classifications. The relatively high error was mostly due to misclassifications between stable/unstable object poses which are in some cases very difficult to distinguish from the viewpoint of the hand camera. This problem can be overcome using another VPL for a separate, more detailed classification of the robot fingertip position on a higher resolution [2].

For the future, the following directions of development are promising: (1) We hope to improve the classification results by an additional classification step judging the positions of the single fingertips relative to the object as described in [2]. (2) Still, the hand gestures are "symbolic" and non-continuous. Therefore, a more natural interface involving intermediate postures will be developed.

References

[1] G. Heidemann. *Ein flexibel einsetzbares Objekterkennungssystem auf der Basis neuronaler Netze.* PhD thesis, Univ. Bielefeld, Techn. Fak., 1998. Infix, DISKI 190.

[2] G. Heidemann and H. Ritter. Visual checking of grasping positions of a three-fingered robot hand. In *Proc. ICANN 2001.* Springer, 2001.

[3] G. Heidemann and H.J. Ritter. Efficient Vector Quantization using the WTA-rule with Activity Equalization. *Neural Processing Letters*, 13(1):17–30, 2001.

[4] Ján Jockusch. *Exploration based on Neural Networks with Applications in Manipulator Control.* PhD thesis, Universität Bielefeld, 2000.

[5] T. Kohonen. Self-organization and associative memory. In *Springer Series in Information Sciences 8.* Springer-Verlag Heidelberg, 1984.

[6] R. Menzel, K. Woelfl, and F. Pfeiffer. The development of a hydraulic hand. In *2nd Conf. on Mechatronics and Robotics*, pages 225–238, 1993.

[7] T.D. Sanger. Optimal unsupervised learning in a single-layer linear feed-forward neural network. *Neural Networks*, 2:459–473, 1989.

[8] Michael E. Tipping and Christopher M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.