Recognition of Gestural Object Reference with Auditory Feedback

Ingo Bax, Holger Bekel, and Gunther Heidemann

Neuroinformatics Group, Faculty of Technology, Bielefeld University,
Postfach 10 01 31, D-33501 Bielefeld, Germany
{ibax,hbekel,gheidema}@techfak.uni-bielefeld.de
http://www.TechFak.Uni-Bielefeld.DE/ags/ni/index_d.html

Abstract. We present a cognitively motivated vision architecture for the evaluation of pointing gestures. The system views a scene of several structured objects and a pointing human hand. A neural classifier gives an estimation of the pointing direction, then the object correspondence is established using a sub-symbolic representation of both the scene and the pointing direction. The system achieves high robustness because the result (the indicated location) does not primarily depend on the accuracy of the pointing direction classification. Instead, the scene is analysed for low level saliency features to restrict the set of all *possible* pointing locations to a subset of highly *likely* locations. This transformation of the "continuous" to a "discrete" pointing problem simultaneously facilitates an auditory feedback whenever the object reference changes, which leads to a significantly improved human-machine interaction.

1 Introduction

Establishing a common *focus of attention* (FOA) is a major task of communication. To influence the spatial FOA humans often use hand gestures. Therefore, pointing direction evaluation is a key topic in the field of human-machine interaction, in particular, gestural reference to objects increasingly attracts interest. However, humans use several modalities at once to establish a common FOA like gesture, speech and gaze direction. Moreover, feedback from the partner is constantly evaluated. In contrast, gesture recognition used in machine vision to direct the FOA is still mostly *stand alone* and *unidirectional*, i.e. without feedback. To compensate for these shortcomings, much effort is spent to increase the *accuracy* of pointing direction recognition. However, we argue that it is not primarily pointing accuracy which leads to good results but interaction with a system. Therefore, a system for gestural object reference recognition should offer three features: (*i*) A basic "understanding" of the scene to limit the set of objects that can be pointed at, (*ii*) feedback to the user to indicate how a gesture was understood, and (*iii*) the possibility to include other modalities like speech.

In this paper, we present a human-machine interaction system that addresses these three points. It allows the user to refer to objects or structures of objects

(like buttons of a technical device) by pointing gestures. It relies on the interaction of a neural classifier for pointing directions and a saliency map S generated from context-free attentional mechanisms. The attentional system is related to the approach of Backer et al. [1]; an earlier version was presented in [2, 7]. Similarly motivated architectures for FOA were proposed by Itti et al. [9] and Walther et al. [17]. The latter approach is based on the Neocognitron introduced by [3].

The maxima of S are used to select conspicuous image structures. These results are combined with the down-propagated output of the classifier (pointing angle) on a sub-symbolic level using an "attention map" (ATM). The representation of the common FOA as the maximum of the ATM facilitates (*i*) the stabilisation even in the presence of inaccurate or noisy pointing results, (*ii*) an auditory feedback when the indicated point "hops" to a new object and (*iii*) the future integration of spatial anticipations from other modalities.

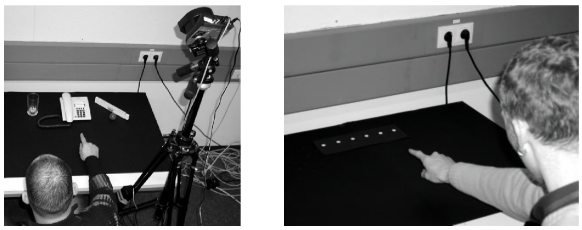

Fig. 1. Left: In a test scenario viewed by a stationary camera a user is pointing to objects on the table. Right: Setting used for system performance evaluation as described in Sect. 3.

2 System description

The system relies on two data driven processing branches (Fig. 2): From the camera image first three feature maps are calculated (Sect. 2.1) in which different image features stand out (Fig. 5). The saliency map S is calculated as a weighted sum of the feature maps by an adaptive weighting (Sect. 2.2). Maxima of S are considered as "interesting" areas and serve as a possible pointing targets.

In the second branch, a skin colour segmentation module yields candidate regions which might contain a hand. These regions are input to the VPL classifier (Sect. 2.3) which (a) decides if the region is a pointing hand and if so (b) determines the pointing direction. The symbolic output angle is translated back to the sub-symbolic level by calculating a *manipulator map* (Sect. 2.4) which is multiplied to S to obtain the attention map, the maximum of which is the FOA. The *focus shift detection* module (FSD) outputs an auditory feedback to the user, reconsulting intermediate processing results from the ATM module. Next, the single components are described in more detail.

2.1 Generation of context-free feature maps

We currently use three different methods to generate saliency maps which complement each other: Grey value entropy, local symmetry and edge-corner detection. A *local entropy* map M_1 yields high saliency value for image windows which

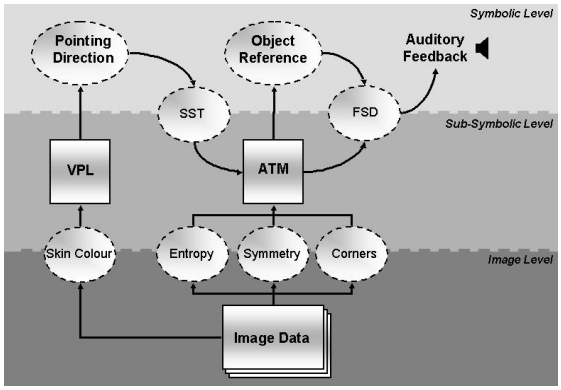

Fig. 2. System architecture. In contrast to approaches which integrate pointing gesture information and object locations on the symbolic level, the pointing angle is down-propagated to the sub-symbolic level using a "symbol-signal-transformer" (SST) and integrated as a spatial weighting of the feature maps.

have a high informational content in the sense of information theory [10]. The window size determines the scale on which structures are evaluated. Here, we use windows large enough to detect large objects (see Fig. 5).

A *symmetry map* M_2 after Reisfeld et al. [13] attracts attention to objects or details which are locally symmetric. The use of symmetry is cognitively motivated e.g. by [12]. The third feature map M_3 concentrates on *edges and corners* as small salient details of objects. Here we use the detector proposed by Harris and Stephens [4], which proved to be superior to other detectors in [15].

2.2 Adaptive integration algorithm

From the $N = 3$ feature maps $M_i(x, y)$ the saliency map S is calculated as a weighted sum. S is spatially weighted by the *manipulator map* $L(x, y)$ which codes the pointing direction (Sect. 2.4) to obtain the attention map C (Fig. 3):

$$C(x,y) = S(x,y) \cdot L(x,y) \qquad \text{with} \qquad S(x,y) = \sum_{i=1}^{N} \theta(w_i \cdot M_i(x,y)), \tag{1}$$

with $\theta(\cdot)$ as a threshold function. The maximum of $C(\cdot,\cdot)$ determines the common FOA of user and machine, which can be used for further processing.

To equalise contributions of the maps M_i, we calculate the contributions $\bar{M}_i$ as a sum over all pixels of each M_i. To reach approximate equalisation of the $\bar{M}_i$, the map weights w_i are adapted by iterating

$$w_i(t+1) = w_i(t) + \epsilon(w_i^s(t) - w_i(t)), \quad 0 < \epsilon \leq 1, \tag{2}$$

with the following target weights w_i^s:

$$w_i^s = \frac{1}{N^2} \cdot \frac{\sum_{k=1}^{N} \bar{M}_k}{\bar{M}_i} \qquad \text{with} \qquad \bar{M}_i = \frac{\sum_{(x,y)} (M_i(x,y) + \gamma)}{\xi_i}. \tag{3}$$

γ enforces a limit for weight growing. The parameters ξ_i can be used if certain saliency features should a priori be weighted higher. In Sect. 2.4 we make use of this possibility to give entropy higher weight for large-scale selection of objects and low weight when object details are pointed at.

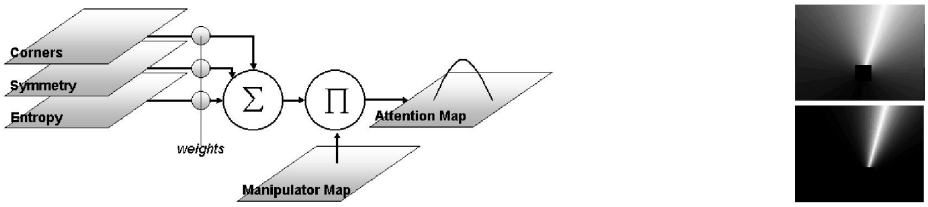

Fig. 3. Left: Processing flow of the ATM-module (central box in Fig. 2). The saliency map S is generated from an (adaptively) weighted superposition of the feature maps. The manipulator map, which allows the coupling of information from other modules like the pointing direction recognition, is multiplicatively overlaid on S. Right: Examples of manipulator maps L ("spotlight of attention"). A wide cone is used as long as the user wants to indicate large objects, a narrow one for precise pointing to details.

2.3 The neural VPL classification system

We use the VPL system [5] for visual classification, which was previously applied to several computer vision tasks [6]. "VPL" stands for three processing stages: **V**ector quantisation, **P**CA and **L**LM-network. The VPL classifier combines visual feature extraction and classification by means of a local principal component analysis (PCA) for dimension reduction followed by a classification stage using neural networks, see Fig. 4. Local PCA can be viewed as a nonlinear extension of simple, global PCA [16].

The vector quantisation is carried out on the raw image windows to provide a first data partitioning with N_V reference vectors $\boldsymbol{r}_i \in \mathbb{R}^D, i = 1 \ldots N_V$, using the *Activity Equalisation Algorithm* proposed in [8]. To each reference vector $\boldsymbol{r}_i$ a single layer feed forward network for the successive calculation of the principal components (PCs) as proposed by Sanger [14] is attached. It projects the input $\boldsymbol{x} \in \mathbb{R}^D$ to the $N_P < D$ PCs with the largest eigenvalues: $\boldsymbol{x} \rightarrow \boldsymbol{p}_i(\boldsymbol{x}) \in \mathbb{R}^{N_P}, i = 1 \ldots N_V$. In the third stage, to each PCA-net one "expert" neural classifier of the Local Linear Map – type (LLM network) is attached. It performs the final mapping $\boldsymbol{p}_l(\boldsymbol{x}) \rightarrow \boldsymbol{y}$. The LLM network is related to the self-organising map [11], see e.g. [5] for details. It can be trained to approximate a nonlinear function by a set of locally valid linear mappings.

The output vector $\boldsymbol{y}$ codes both the decision as to whether the input $\boldsymbol{x}$ is a pointing hand, and, if so, its pointing angle. The three VPL processing stages are trained successively with labelled sample windows of the cropped pointing

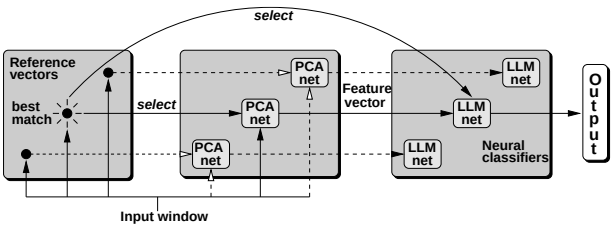

Fig. 4. The VPL classifier performs a local PCA for feature extraction and a subsequent neural classification.

hand plus objects assigned to a rejection class. The rejection class contains other objects which are part of the scenario, e.g. the objects the user points at or parts of the background. In addition, hand postures other than pointing gestures are part of the rejection class, e.g. a fist.

The major advantage of the VPL classifier is its ability to form many highly specific feature detectors (the $N_V \cdot N_P$ local PCs). It could be shown that classification performance and generalisation properties are well-behaved when the main parameters are changed, which are N_V, N_P and the number of nodes in the LLM nets N_L [6].

2.4 Translation from symbolic to sub-symbolic level

Skin colour segmentation and the VPL classifier yield the position of the hand (x_H, y_H) and the pointing direction α, respectively. Both these (symbolic) informations are translated to a manipulator map L and thus back to the sub-symbolic level. The manipulator map shows a "Gaussian cone" of width σ_c which determines the effective angle of beam spread

$$L(x,y) \;=\; \frac{1}{\sqrt{2\pi}\sigma_c} \;\exp\left(-\frac{(\arctan(\frac{y-y_H}{x-x_H}) - \alpha)^2}{\sigma_c^2}\right), \tag{4}$$

here in the form for the first quadrant for simplicity, see Fig. 3. The cone gives higher weight in the attention map to image regions in the pointing direction and thus "strengthens" salient points in this area.

To facilitate selection of objects on differing scales, σ_c is adjusted online according to the user behaviour. The pointing angles α and hand positions (x_H, y_H) are recorded over the last six frames. If they show large variance, it is assumed that the user has moved the hand on a large scale to select a big object, so also a large σ_c is chosen. In contrast, σ_c is reduced for small variance to establish a "virtual laser pointer" since it is assumed that the user tries to point to detail.

As an additional assistance for coarse / fine selection, the a priori weights ξ_i of (3) are changed such that the large scale entropy map M_1 dominates for large pointing variance whereas the symmetry map M_2 and the corner saliency M_3 are weighted higher for detail selection.

2.5 Auditory feedback

The FSD module in Fig. 2 detects spatial shifts of the FOA for auditory user feedback. A short "bop" sound is produced when the current FOA shifts to a different maximum $\hat{\boldsymbol{s}}_j$ of the saliency map S. Such an event is detected when

$$\left|\left(\frac{1}{\Delta t}\sum_{i=1}^{\Delta t}\hat{\boldsymbol{s}}^*(t-i)\right) - \hat{\boldsymbol{s}}^*(t)\right| > d, \tag{5}$$

where $\hat{\boldsymbol{s}}^*(t)$ denotes the maximum of S closest to the FOA (i.e. the maximum of the ATM) in frame t. The parameter Δt has to be adjusted according to the processing frame rate of the system and the threshold d can be estimated by analysing the distance matrix of the maxima $\hat{\boldsymbol{s}}_i$ of the map.

3 Results

Figure 5 shows the results of intermediate processing stages. The user has just slowed down the pointing movement, so the manipulator map shows a cone of medium width in the pointing direction, starting at the approximate position of the hand centre. The parameter ξ_i of the entropy map is decreased while the symmetry map is weighted higher. So the edges of the phone are suppressed while the key pad is highlighted most through the combination of high weighted symmetry map and the manipulator cone.

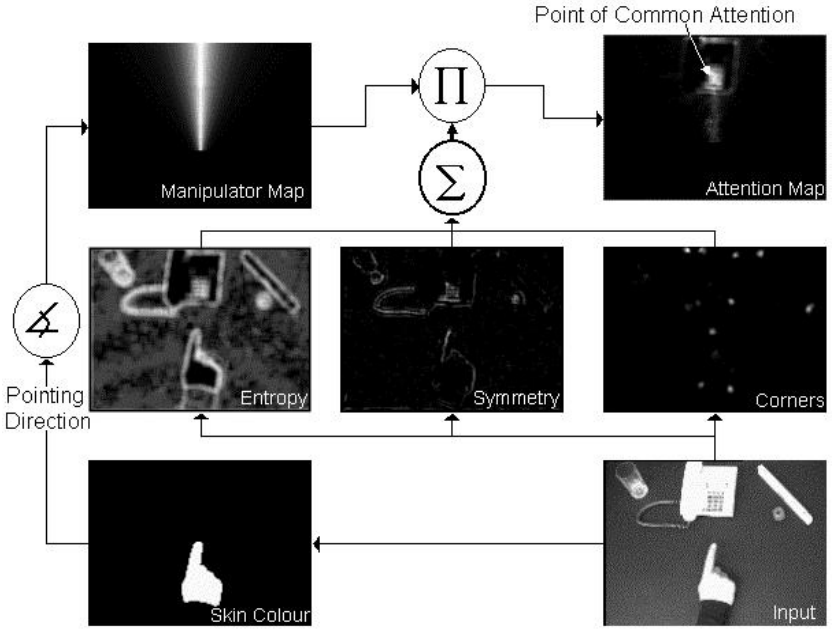

Fig. 5. Processing results for a pointing gesture towards an object. From the input image (bottom right) skin colour is segmented (bottom left); the VPL classifier calculates the angle which is transformed to a manipulator map (top left). The manipulator cone "illuminates" the object; the maxima of the feature maps stand out.

To evaluate the system performance we choose a setup which is based on a "generic" pointing task that can be easily reproduced: A proband points at a row of six white circles on a black table (Fig. 1, right). The distance between the hand and the targets is approximately 40 cm. So the diameter of each circle is of an angular range of 1.7° and the distances between the circle centres vary from an angular resolutions of 4° to 28°. To test performance for pointing to details, circles of a diameter of 0.9° with a distance angle of 2° were used in an additional experiment. A supervisor gives the command to point at one of the circles by reading a randomly generated circle number. We use only the inner

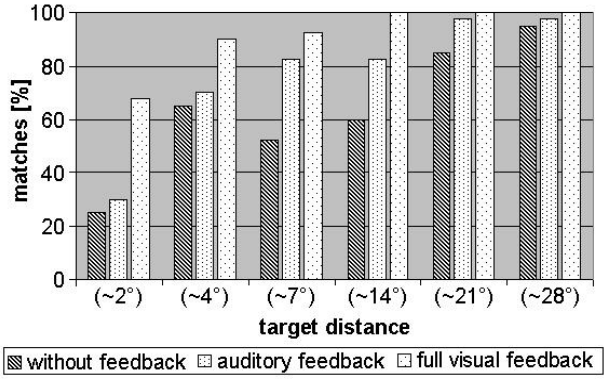

Fig. 6. The chart shows the results of the evaluation experiment. The values are averaged for three probands with 20 items each. On the x-axis the the corresponding angles for a pointing distance of about 40 cm are shown.

four circles to avoid border effects. A match is counted if the system outputs a focus point on the correct circle within three seconds. The experiment was repeated under three conditions. As the results in Fig. 6 show, the best match percentages are reached under the *full visual feedback* condition (proband sees system output on a computer screen), whereas the values substantially decrease under the *without feedback* condition at small distances. Under the *auditory feedback* condition (proband hears a "bop" sound, if the focus point shifts from on maximum to another) a better percentage could be reached.

The major result achieved in this test scenario is that system performance can be significantly increased by giving feedback because (*i*) the user is enabled to adjust single pointing gestures to a target and (*ii*) the user can adapt himself or herself to the system behaviour. This way the achievable *effective resolution* can be improved, because it does not solely rely on the accuracy of the pointing gesture recognition anymore. It could be shown that the rather simple means of giving auditory feedback already leads to a better performance.

A limitation is that the hand has to be completely visible, otherwise the VPL classifier gets an unknown input. A "beep" is used as an error signal if the hand is too close to the border. Another restriction is that the saliency operators do not yield maxima on all of the objects or not on the desired locations.

4 Conclusion and Acknowledgement

We have presented a human-machine interface for visual detection of gestural object reference. It could be shown that using auditory feedback to indicate shifts of the FOA increases performance significantly.

The functionality of the presented system is not limited to the current scenario. Since arbitrary other saliency features like colour or movement can be integrated, the bottom-up focus of attention can be directed to a wide variety of objects. Even more important is the possibility to transform cues from other modules top-down to the sub-symbolic level using further manipulator maps. One of the first steps will be the integration of speech-driven cues to generate

spatial large scale anticipations. As precision and user independence are still problems in the field of gesture recognition, a major advantage of the new approach is that it does not require high recognition accuracy. This is achieved by the system's anticipation that only salient image points will be selected.

This work was conducted within the scope of the project VAMPIRE (Visual Active Memory Processes and Interactive REtrieval) which is part of the IST programme (IST-2001-34401).

References

1. G. Backer, B. Mertsching, and M. Bollmann. Data- and Model-Driven Gaze Control for an Active-Vision System. *IEEE Trans. PAMI*, 23(12):1415–1429, 2001.
2. M. Fislage, R. Rae, and H. Ritter. Using visual attention to recognize human pointing gestures in assembly tasks. In *7th IEEE Int'l Conf. Comp. Vision*, 1999.
3. K. Fukushima. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition unaffected by Shift in Position. *Biol. Cybern.*, 36:193–202, 1980.
4. C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proc. 4th Alvey Vision Conf.*, pages 147–151, 1988.
5. G. Heidemann. *Ein flexibel einsetzbares Objekterkennungssystem auf der Basis neuronaler Netze.* PhD thesis, Univ. Bielefeld, 1998. Infix, DISKI 190.
6. G. Heidemann, D. Lücke, and H. Ritter. A System for Various Visual Classification Tasks Based on Neural Networks. In A. Sanfeliu et al., editor, *Proc. 15th Int'l Conf. on Pattern Recognition ICPR 2000, Barcelona*, volume I, pages 9–12, 2000.
7. G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter. Integrating Context-Free and Context-Dependent Attentional Mechanisms for Gestural Object Reference. In *Proc. Int'l Conf. Cognitive Vision Systems*, Graz, Austria, 2003.
8. G. Heidemann and H. Ritter. Efficient Vector Quantization Using the WTA-rule with Activity Equalization. *Neural Processing Letters*, 13(1):17–30, 2001.
9. L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. PAMI*, 20(11):1254–1259, 1998.
10. T. Kalinke and W. v. Seelen. Entropie als Maß des lokalen Informationsgehalts in Bildern zur Realisierung einer Aufmerksamkeitssteuerung. In B. Jähne et al., editor, *Mustererkennung 1996.* Springer, Heidelberg, 1996.
11. T. Kohonen. Self-organization and associative memory. In *Springer Series in Information Sciences 8.* Springer-Verlag Heidelberg, 1984.
12. P. J. Locher and C. F. Nodine. Symmetry Catches the Eye. In A. Levy-Schoen and J. K. O'Reagan, editors, *Eye Movements: From Physiology to Cognition*, pages 353–361. Elsevier Science Publishers B. V. (North Holland), 1987.
13. D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-Free Attentional Operators: The Generalized Symmetry Transform. *Int'l J. Comp. Vision*, 14, 1995.
14. T. D. Sanger. Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network. *Neural Networks*, 2:459–473, 1989.
15. C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of Interest Point Detectors. *Int'l J. of Computer Vision*, 37(2):151–172, 2000.
16. M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.
17. D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional Selection for Object Recognition – a Gentle Way. In *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02)*, Tübingen, Germany, 2002.