Where is the Hole Punch? Object Localization Capabilities on a Specific Bureau Task

E. Michaelsen[*], U. Ahlrichs[◊+], U. Stilla[*], D. Paulus[◊], H. Niemann[◊+]

[*]FGAN-FOM Research Institute for Optronics and Pattern Recognition
Gutleuthausstr. 1, 76275 Ettlingen, Germany
mich@fom.fgan.de, usti@fom.fgan.de (www.fom.fgan.de)
[◊]Lehrstuhl für Mustererkennung (LME, Informatik 5), Universität Erlangen-Nürnberg
Martensstr. 3, 91058 Erlangen, Germany
paulus@cs.fau.de (www5.informatik.uni-erlangen.de)

Abstract. In this paper, knowledge-based recognition of objects in a bureau scene is studied and compared using two different systems on a common data set: In the first system active scene exploration is based on semantic networks and an A*-control algorithm which uses color cues and 2-d image segmentation into regions. The other system is based on production nets and uses line extraction and views of 3-d polyhedral models. For the latter a new probabilistic foundation is given. In the experiments, wide-angle overviews are used to generate hypotheses. The active component then takes close-up views which are verified exploiting the knowledge bases, i.e. either the semantic network or the production net.

1 Introduction

Object localization from intensity images has a long history of research, but has not led to a general solution yet. Approaches proposed differ in objectives, exploited features, constraints, precision, reliability, processing time etc. Although surveys exist on knowledge-based object recognition [8, 6, 1], little has been published on experiments by different groups on a common task. Comparisons mainly exist on data-driven or appearance-based approaches, e.g. on the COIL-data base [7]. We compare two different approaches developed by different groups to solve one common task. We chose the localization of a hole punch from oblique views on an office desk. Fig. 3a,f,g,h below show such frames taken with different focal lengths by a video camera.

In the experiments camera parameters (focal length, pan, tilt) are adjustable and camera actions are controlled by the recognition process. The 3-d position of the hole punch is constrained by a desk. The rotation is restricted to the axis perpendicular to the ground plate (azimuth). The overviews are used to generate hypotheses of the hole

[+] This work was partially supported under grand number NI 191/12 by Deutsche Forschungsgemeinschaft.

Michaelsen E, Ahlrichs U, Stilla U, Paulus D, Niemann H (2001) Where is the hole Punch? Object localization capabilities on a specific bureau task. In: Radig B, Florczyk S (eds) Pattern Recognition, 23rd DAGM Symposium, Berlin: Springer, 337-344

punch´s position which result in camera actions to take close-up views. These are the input for final recognition or verification.

In Sect. 2 we outline the structure and interaction of the two systems and present a new probabilistic foundation of the production net system. Results of experiments on a common data-base are given in Sect. 3. In Sect. 4 a discussion of pros and cons of both approaches is given.

2 Architectures of the Two Systems

Initially we describe how the two systems interact on common data.

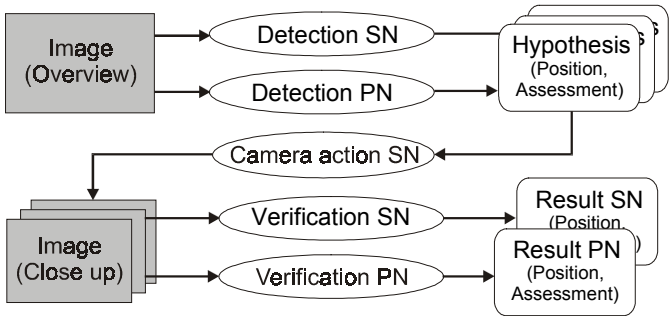

Fig. 1. Overview of the experimental localization setup with semantic network system (SN) and production net system (PN)

Fig. 1 shows the different components of the two localization systems and the data flow between them. Starting with an overview color-image (like the one presented in Fig. 3a) two different algorithms are applied that generate hypotheses for the hole punch´s location. The first system uses a pixel-based color classifier resulting in an interest map (Sect. 2.1), whereas the second system determines the hypotheses with a knowledge-based approach (Sect. 2.2). Both detection systems provide hypotheses as 2-d-coordinates in the image and an assessment value.

Based on the hypotheses, close-up views are then generated by adjusting pan and tilt and increasing the focal length of the active camera. Since close-up views contain objects in more detail, recognition is expected to be more reliable. Results of region segmentation are interpreted by the SN-system providing the center of gravity for each hypothesis. Lines constitute the input for the PN-system yielding a 3-d pose estimate for each hypothesis. This verification utilizes a different parameter setting and finer model compared to the detection phase. Both systems give an assessment value for the results.

2.1 Localization with the Semantic Network System (SN)

The semantic network approach uses the object's color as a cue to hypothesize possible positions of the hole punch in a scene. An interest operator based on histogram back-projection is applied [13], which learns the color distribution of the hole punch and applies this distribution to find pixels of similar color in the overview images. We calculate the color histograms in the normalized *rg* color space to be insensitive to illumination changes. Since the hole punch is red, the interest operator yields hypotheses for red objects.

The verification of the hypotheses is done by matching the hole punch's model to color regions. These regions are determined by segmenting the close-up views using a split and merge approach. The semantic network represents the 2-d object model by a concept which is linked to a color region concept [2]. The latter concept contains attributes for the region's height, width, and color as well as the allowed value range for each of these attributes. During analysis the expected values for the object are compared to the corresponding feature values calculated for each color region of the close up views. A match is judged according to a probability based optimality criterion [3]. The search for the best matching region is embedded into an A*-search.

2.2 Localization with the Production Net System (PN)

Production nets [12] have been described for different tasks like recognition of roads in maps and aerial images, 3D reconstruction of buildings, and vehicle detection. A syntactic foundation using coordinate grammars is given in [9]. Initially contours are extracted from gray-value images and approximated by straight line segments. The production system works on the set of lines reconstructing according to the production net the model structure of the hole punch. This search process is performed with a bottom-up strategy. Accumulating irrevocable control and associative access is used to reduce the computational load [11, 9].

The view-based localization utilized here for the hole punch search implements accumulation of evidence by means of cycles in the net with recursive productions [10]. The accumulation resembles generalized Hough transform [4]. The hole punch is modeled by a 3-d polyhedron. The 3-d pose space is equidistantly sampled rotating the object in azimuth α in steps of 10^o and varying the distance d in 5 steps of 10cm. For each of these 180 poses a 2-d model is automatically generated off-line by a hidden line projection assuming perspective projection and internal camera parameters estimated by previous calibration (see Fig. 2). The recognition relies on matching structures formed of straight line segments. Below only L-shaped structures are used, that are 4-d attributed by the location of the vertex and the two orientations.
If a L-structure in the image is constructed from two lines, then similar L-structures in each 2-d model are searched, where the two orientations account for the similarity. Matches are inserted as cue instances into an 4-d accumulator of position in the image *(x,y)*, azimuth α, and distance d. Accumulation is performed by recursive productions operating on the associative memory which is discretized in Pixel, 1^o and 1cm. Values found in the accumulator highly depend on structures and parameters. High values indicate the presence of the object for a corresponding pose.

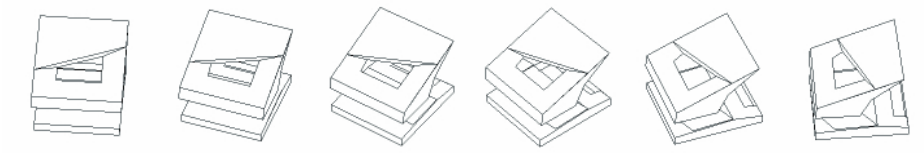

Fig. 2. Selected set of 2-d models projected from a 3-d polyhedron model ($\Delta\alpha=15^o$)

We now replace the accumulator values by an objective function based on probabilistic assessment. For this purpose we modified the theory derived by Wells [14]. But while he uses contour primitives attributed by their location, orientation and curvature our approach matches L-structures; while he operates in the image domain we operate in the accumulator.

Wells´ Theory of Feature-Based Object Recognition

Wells uses a linear pose vector β of dimension 4 (for similarity transform) or 8 (for limited 3-d rotations according to the linear combination of views method), and a correspondence function Γ, mapping the image features to a model feature or to the background. A scaled likelihood

$$\mathbf{L}(\Gamma,\beta)=-\frac{1}{2}(\beta-\beta_0)^T\psi_\beta^{-1}(\beta-\beta_0)+\sum_{i,j:\Gamma_i=M_j}\left[\lambda-\frac{1}{2}(Y_i-M_j\beta)^T\psi_{ij}^{-1}(Y_i-M_j\beta)\right] \qquad (1)$$

of an image-to-model correspondence and a pose is derived from independence and distribution assumptions. The first term in Eq. 1 results from a normal prior distribution on the pose, where ψ_β is the corresponding covariance matrix and β_0 the center. The second term is due to the conditional probability that a set of correspondences between the image features Y_i and model features M_j may be true, given a pose β. Wells gives the model features M_j in a matrix format, that enables linear transformation to the image feature domain. Inside the sum there appears a trade-off rewarding each image-to-model correspondence by a constant λ and punishing the match errors. The punishing term for each correspondence results from the assumption of linear projection and normal distributed error in the mapping of object to image features with covariance ψ. To reduce the complexity of the estimation process, this matrix is independent of the indices *i* and *j*. The reward term λ is to be calculated from a representative training-set according to

$$\lambda=\ln\left(\frac{1}{(2\pi)^{\nu/2}m}\frac{(1-B)}{B}\frac{W_1\cdots W_\nu}{\sqrt{|\psi|}}\right). \qquad (2)$$

The middle factor in this product is calculated from the ratio between the probability *B* that a feature is due to the background, and the probability *(1-B)/m* that it corresponds to a certain model feature, where *m* is the number of features in the model. The rightmost factor in the product is given by the ratio between the volume of the whole feature domain W_1 ... W_ν and the volume of a standard deviation ellipsoid of ψ.

Modification for Accumulator-Productions in the PN-System

To apply the theory of Wells to our problem we set $\beta^T=(x,y,\alpha,d)$. The objective funktion L is calculated for each cluster of cues. The pose β is estimated as mean $\hat{\beta}^T = (\hat{x},\hat{y},\hat{\alpha},\hat{d})$ of the poses of the member cues of the cluster. The correspondence Γ is coded as an attribute of the cues. For each model feature j put into correspondence in the cluster the closest cue i to the mean is taken as representative of the set of all cues i corresponding to j. This is done, because we regard multiple cues to the same model feature as not being mutual independent. The attribute values $(x_i, y_i, \alpha_i, d_i)$ directly serve as Y_i for formula (1). There is no need for coding model features in a matrix format, because the projection has been treated off-line in the generation of the 2-d models. We just determine the deviation for each such cue

$$L = \sum_j \left[\lambda - \underset{\Gamma_i=j}{\mathrm{Min}} \left[\frac{1}{2} (Y_i - \hat{\beta})^T \psi^{-1} (Y_i - \hat{\beta}) \right] \right]. \qquad (3)$$

The covariance matrix ψ of the cues and the background probability B are estimated from the training-set. These differ in the present bureau application significantly between overviews and close-ups. For the overviews the reward λ is close to the critical value zero indicating that recognition in these data is difficult and not very stable. Recall that the maximization must not take those Γ into account, that include negative terms into the sum. This condition gives a new way to infer the threshold parameters for adjacency in the cluster productions from a training set. In the verification step parameters are set different compared to the detection step, e. g. the accumulator is now sampled in $\Delta\alpha=5^o$ and Δd=5cm. Fig. 2 shows 2-d models used for close-up views, whereas Fig. 3d,e show two coarser 2-d models used for the overviews.

The theory of Wells rejects scenes as non recognizable, if λ turns out to be negative according to Eq. 2. In such situation we still may use a positive reward λ' instead indicating that cues with high values for this objective function will contain more false matches than correct ones with high probability. Still among the set of all cues exciding a threshold, there will be the correct hypothesis with a probability that may be calculated from the difference λ-λ'.

For the close-ups a ML-decision is needed and we have to use the correct reward term λ. For these data the estimation for λ is much bigger. Compared to the Hough-accumulator value used as decision criterion in [10] the likelihood function includes an error measurement on the structures set in correspondence with the model and evaluates the match based on an estimated background probability.

3 Experiments

Each system used its own set of training images for hypothesis generation and object recognition. The training of the SN-system is based on 40 close-up images for model parameter estimation and a histogram for red objects that is calculated using one close-up image of the hole punch. For the PN-system 7 desk scene overview and 7 close-up images were used as training set. The evaluation was done on 11 common

scenes disjoint from the training sets. One of these overviews is depicted in Fig. 3a. For each test scene both systems generated their hypotheses (Fig. 3b,c), and corresponding close-up sets were taken by the scene exploration system.

Success and failure was judged manually. In Fig. the highest objective function value L is detected by the PN-system in the correct location. Fig. 3d presents the 2-d model corresponding to this result. The pose is incorrectly determined on this image. Fig. 3e shows 2-d model of a cue cluster with correct pose and location but having a smaller likelihood. On the 11 overview images only two localization results are successful where one gives the correct pose, too. This shows that in this case pure ML is not sufficient. Therefore clusters are sorted according to L and the 1‰-highest-L scoring clusters were taken as hypotheses (see white crosses in Fig. 3b). A successful localization according to the definition is contained in 5 of the 11 hypotheses sets.

The color-based detection of the SN-system does not determine pose. It gives 8 correct ML-localization results in the overview images. In 10 results the hypotheses set contains the correct cue. Fig. 3c shows an interest-map of the overview image. Dark regions correspond to high likelihood of the hole punch´s position. Note that hypotheses sets of the two systems differ substantially, as can be seen comparing Fig. 3b and Fig. 3c. Where the SN-system finds red objects like the glue stick and the adhesive tape dispenser, the PN-system finds rectilinear structures like books.

Fig. 3f,g,h show the three close-up views taken according to the PN-system detection. In the verification step the ML-decision includes all cues from a close-up set resulting from one overview. Fig. 3i,j,k display the result, where the third scores correctly the highest. A successful verification with the PN-system additionally requires the correct pose. This is performed correctly on 3 of the 11 examples. The SN-system succeeds on 9 close-up sets without giving a pose. The PN-system succeeds on one of the two failure examples of the SN-system.

4 Discussion

In this contribution we demonstrated how the difficult problem of object recognition can be solved for a specific task. An office tool is found by two model-based active vision systems. In both cases the object was modeled manually in a knowledge base. A 3-d polyhedral model was used in the PN-system requiring line segmentation for the recognition. 2-d object views were modeled in SN-system using a region based approach. The experiments revealed that color interest maps of the SN-system outperform the line-based hypothesis generation of the PN-system on the considered scenery. We conjecture that this is due to high color saturation and small size of the object in the overview images. Close-up views captured by the active camera increase the recognition stability of both systems; in some cases overview images already yielded the correct result.

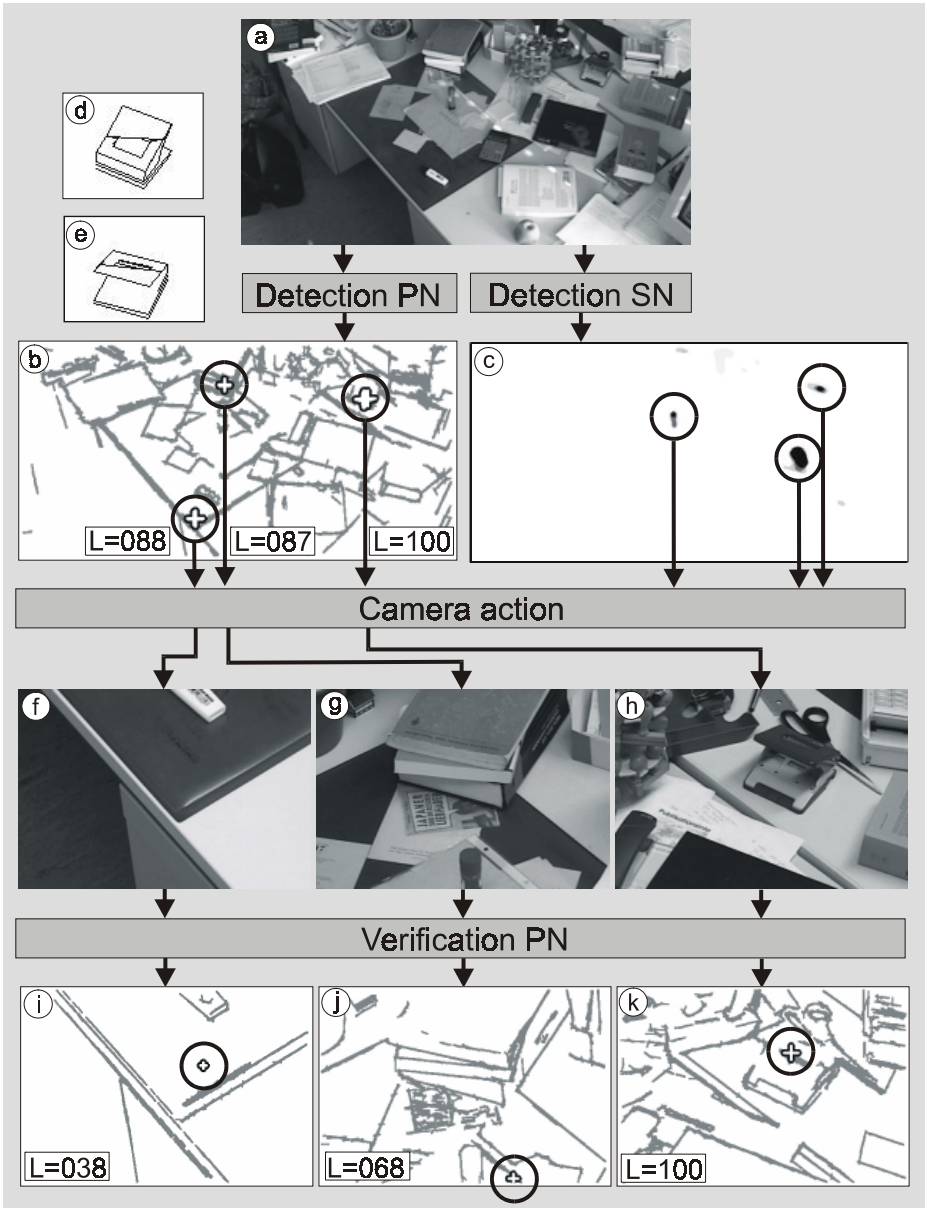

Fig. 3. Localization of the hole punch in a bureau scene; close-up views and verification of the SN-system omitted

For the line-based recognition the process had to be parameterized differently for overview and close-up images. A new probabilistic objective function for the PN-system allows parameter inference from a training set, and opens the way for a better interpretation of the results. Both systems achieved recognition rates that – with respect to the complexity of the task - were satisfactory. It is expected that the

combination of line and color segmentation will eventually outperform either approach. This is subject to future work.

The PN-system is designed to work on T-shaped structures as well. Other possibilities like U-shaped structures would be a straight forward extension. Further investigations will include an EM-type optimization of pose and correspondence in the final verification step also following [14].

We proved that one common experimental set-up can be used by two different working groups to generate competitive hypotheses and to verify these hypotheses, even in an active vision system. The image data is publicly available to other groups to allow further comparisons under the web site of the authors.

References

1. Ade F.: The Role of Artificial Intelligence in the Reconstruction of Man-made Objects from Aerial Images. In: Gruen A., Baltsavias E. P., Henricsson O.(eds.): Automatic Extraction of Man-Made Objects from Aerial and Space Images (II), Birkhäuser, Basel (1997). 23-32
2. Ahlrichs U., Paulus D., Niemann H.: Integrating Aspects of Active Vision into a Knowledge-Based System. In: A. Sanfeliu et al. (eds.): 15th International Conference on Pattern Recognition, IEEE, Barcelona, Vol. IV (2000) 579-582
3. Ahlrichs U.: Wissensbasierte Szenenexploration auf der Basis erlernter Analysestrategien. Diss., Uni. Erlangen-Nürnberg, to appear (2001)
4. Ballard D. H., Brown C. M.: Computer Vision. Prentice Hall, Englewood Cliffs, New Jersey, (1982)
5. Binfort T. O., Levitt T. S.: Model-based Recognition of Objects in Complex Scenes. In: ARPA (ed.). Image Understanding Workshop 1994. Morgan Kaufman, San Francisco (1994) 149-155
6. Crevier D., Lepage R.: Knowledge-Based Image Understanding Systems: A Survey. CVIU, 6 (1997) 161-185
7. COIL-100. http://www.cs.columbia.edu/CAVE (confirmed 06.07.2001).
8. Lowe D. G.: Three-dimensional Object Recognition from Single Two-dimensional Images. AI, 31 (1987) 355-395
9. Michaelsen E.: Über Koordinaten Grammatiken zur Bildverarbeitung und Szenenanalyse. Diss., Uni. Erlangen-Nürnberg, (1998)
10. Michaelsen E., Stilla U.: Ansichtenbasierte Erkennung von Fahrzeugen. In: Sommer G., Krüger N., Perwas C. (eds.).: Mustererkennung 2000, Springer, Berlin (2000) 245-252
11. Nilsson N. J.: Principles of Artificial Intelligence. Springer, Berlin (1982)
12. Stilla U.: Map-aided structural analysis of aerial images. ISPRS Journal of Photogrammetry and Remote Sensing, 50, 4 (1995) 3-10
13. Swain M. J., Ballard D. H.: Color indexing. IJCV, 7, 1 (1995) 11-32
14. Wells III W. M.. Statistical Approaches to Feature-Based Object Recognition. IJCV, 21 (1997) 63-98