Adaptive scene-dependent filters in online learning environments

M. Götting[1], J. J. Steil[1], H. Wersing[2], E. Körner[2] and H. Ritter[1]

1- Bielefeld University - Neuroinformatics Group, Faculty of Technology
P.O.-Box 10 01 31, D-33501 Bielefeld - Germany

2- Honda Research Institute GmbH
Carl-Legien-Str. 30, 63073 Offenbach - Germany

Abstract. In this paper we propose the Adaptive Scene Dependent Filters (ASDF) to enhance the online learning capabilities of an object recognition system in real-world scenes. The ASDF method proposed extends the idea of unsupervised segmentation to a flexible, highly dynamic image segmentation architecture. We combine unsupervised segmentation to define coherent groups of pixels with a recombination step using top-down information to determine which segments belong together to the object. We show the successful application of this approach to online learning in cluttered environments.

1 Introduction

In the field of intelligent man-machine interaction it is widely recognized that attention control and object recognition are two important and connected issues. At the lower level, attention control based on topographic feature maps like color, entropy or orientation is often used to guide fixations to salient image regions, to which object recognition then can be restricted. But due to the complexity of real-world scenes, most work in this area has been concentrated on explicitely or implicitly constrained scenarios like e.g. uncluttered background, homogenous coloring of foreground objects, or predefined object classes. However, we are aiming at a more complex scenario where humans manipulate and present objects to be learned to the machine in an unconstrained environment. Only a few vision systems meet the computational challenges to enable real-time online learning in this context. One approach recently presented is based on a holistic object recognition system [1]. In this system image acquisition is triggered by pointing gestures indicating objects on a table, and is followed by a training phase taking some minutes. A neural approach to recognition is suggested in [2], which supports online learning capabilities by using precomputed and hierarchically generated robust sparse feature sets. Here object-specific learning is directed to the highest levels of a visual hierarchy. This recognition architecture is a part of a stereo vision framework for visual object recognition as described in [3]. For this architecture it has explicitly been shown that good segmentation of objects increases the performance for online object learning as well as for the object recognition [4]. This motivates to search for fast online approaches to unsupervised segmentation, because a priori knowledge about the object to segment is not available.

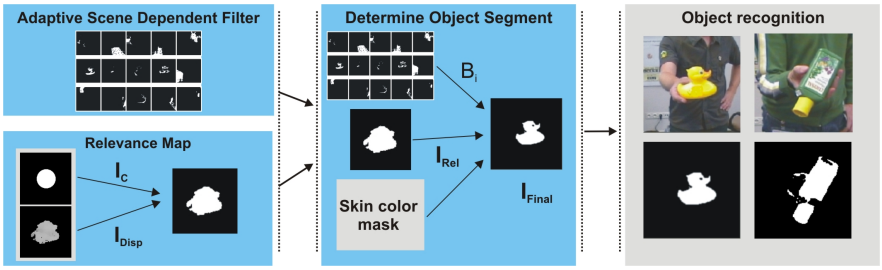

Fig. 1: The multi-path ASDF processing scheme for image segmentation. While the upper path generates the ASDF maps, the second path computes the Relevance Map used as a prediction for the object position in the image. The third path adds a skin color segmentation to subtract hands and forearm from the object. All three paths project to the Determine Object Segment module and are integrated to a single object segmentation mask providing unsupervised figure ground segmentation for the object recognition architecture.

A number of unsupervised segmentation schemes have been proposed [5, 6], mostly based on different color spaces and sometimes on the pixel coordinates as feature space. Though such quantization methods can potentially be fast, they assume that objects have to be homogeneously colored, can be covered by one segment, and are isolated from each other. A combination of unsupervised segmentation and top-down Bayesian classification approaches into a single figure ground segmentation process is proposed in [7]. The unsupervised step of this approach consists of generating a hierarchy of segments [8] ordered in a tree and a successive optimization procedure to label the segments as belonging to the object with respect to a cost function based on the top-level information. The ASDF method proposes a multi-stage, flexible, highly dynamic, online capable and robust image segmentation architecture with real-time performance in the stereo vision framework [3]. It uses feature maps typically available from an attentive system like orientation, intensity, difference images, velocity fields, disparity, image position or different color spaces for forming a combined feature space. Clustering in the combined space results in binarized disjunct segment masks through winner-take-all competition between the cluster prototypes. Because we recombine segments in a later processing step we do not rely on a homogeneity assumption on the object. Additionally we limit the number of segments for avoiding over-segmentation and saving computation time. In this paper, we focus on the application of the ASDF architecture in online-learning of "objects-in-hand" presented by a human partner.

2 The Adaptive Scene Dependent Filter architecture

The ASDF architecture is a multistage and multi-path segmentation architecture and is used as preprocessing step for the feature based object recognition

architecture described in [2], see Fig. 1. There are three paths: skin color detection, computation of the relevance mask, and unsupervised learning of the multi-feature segmentation. Those project to the "Determine Object Segment" module, which computes the final segmentation mask.

We assume that in earlier stages of the vision architecture low level filter operations on an input image provide topographic feature maps, which usually are combined by an additive weighting in a fixation-guiding saliency map. Thus in layer 1 of the ADSF architecture (Fig. 2) we rely on M such basic filter maps F_i with features $m^i_{(x,y)}, i = 1..M$ at pixel positions (x, y). In layer 2, a vector quantization network (VQ) is used to obtain N prototypic codebook vectors $\vec{c}^j, j = 1..N$ representing the most frequent and salient feature combinations. These are used to generate new adaptive topographic activation maps in layer 3, which are binarized by a winner-take-all mechanism in layer 4 accordingly.

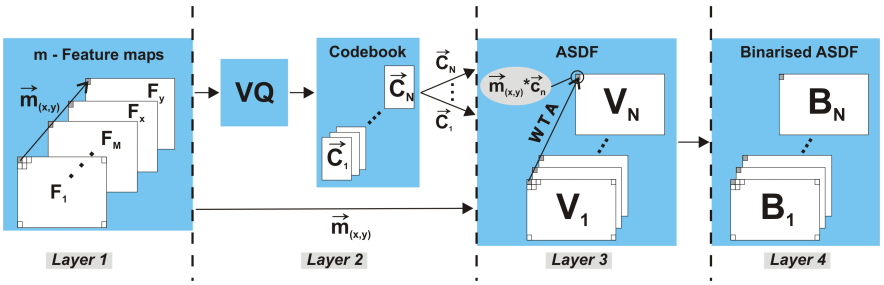

Fig. 2: The multistage ASDF architecture.

The generation of the activation maps employs a standard VQ with a fixed number of training steps (to speed up computation) and training data $\vec{m}_{(x,y)}$:

$$\vec{m}_{(x,y)} = \left(\xi^1 \frac{m^1_{(x,y)}}{\sigma(m^1)^2}, ..., \xi^n \frac{m^M_{(x,y)}}{\sigma(m^n)^2}, \xi^x m^x_{(x,y)}, \xi^y m^y_{(x,y)} \right)^T , \tag{1}$$

where (x, y) is the respective pixel index and $m^x(x, y) = x, m^y(x, y) = y$ include the pixel position as feature. We normalize each component by its variance $\sigma(m_i)^2$. ξ^i is an additional heuristically determined weighting factor, which can be used to weight the relative importance of different map. In each step, the minimal distance $d_{min} = \min_j \parallel \vec{m}_{(x,y)} - \vec{c}^j \parallel^2, \vec{c}^j \in C$ is calculated and the winning codebook vector is adapted through the standard VQ rules. The initialization of the VQ codebook C starts with an empty codebook and incrementally assigns new codebook vectors by the following procedure:
Draw a random (x, y)-position from the image, generate the feature vector $\vec{m}_{(x,y)}$ at this position and compute the minimal distance d_{min} of $\vec{m}_{(x,y)}$ to all $\vec{c}_j$ in the current codebook. A new codebook vector $\vec{c}^{j'}$ is assigned dependent on d_{min}

$$\vec{c}^{j'} = \begin{cases} \vec{m}_{(x,y)} & \text{if } d_{min} > \bar{d} \\ \text{else} & \text{draw a new } \vec{m}_{(x,y)} \end{cases} , \tag{2}$$

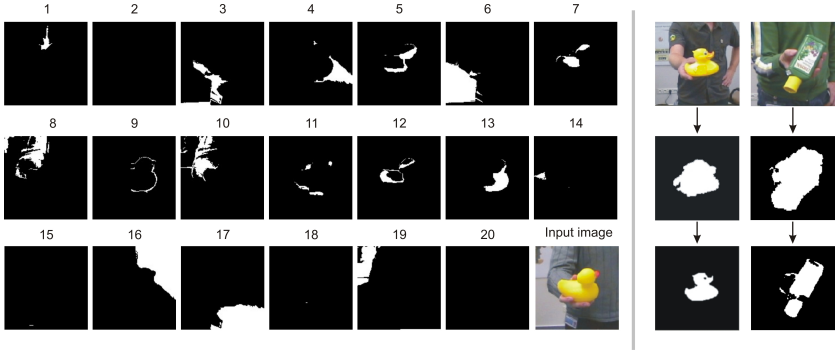

Fig. 3: Binarized ASDF segments B_i (shown on the left). A combination of 5,7,9,11,12,13 constitutes the object mask for the shown object. The selection of the appropriated ASDF segments is done by the Determine Object Segment module. The segmentation results of the architecture, are shown on the right.

where $\bar{d}$ is a threshold value to ensure a good distribution of the codebook vectors. This procedure is done before each adaption step of the VQ until the maximum number of codebook vectors is reached.

The input for the third layer consists of the adaptive codebook C and the basic filter maps F_i. Based on the codebook, N scene dependent activation maps V^j are computed as $V^j_{(x,y)} = \| \vec{m}_{(x,y)} - \vec{c}^j \|^2$ and binarized as

$$B^j_{(x,y)} = \begin{cases} 1 & \text{if } \| \vec{m}_{(x,y)} - \vec{c}^k \|^2 < \| \vec{m}_{(x,y)} - \vec{c}^j \|^2 \ ,\forall k \neq j \\ 0 & \text{else} \end{cases} . \tag{3}$$

The *relevance map* I_{Rel} is used as a prediction mask for a rough region around the focused object in which segments are likely to belong to the object. Currently it is an additive superposition of a circular center map I_C defining an interest range around the current fixation and a disparity map I_{Disp} consisting of pixels with disparities within a defined range only. This flexible integration of different inputs for the relevance map allows for handling objects in different orientations and perspectives. Furthermore, it is not limited to utilizing disparity or interest and could be enhanced by any topographic map predicting the object position.

The *Determine Object Map* module combines the ASDF and the relevance map with an adaptive and robust skin color segmentation [9] in order to subtract hand and forearm from the object. We compute the number of pixels P_{in} of the intersection I_{Rel} and B_i ($P_{in} = \#(B_i \cap I_{Rel})$) and the number of pixels P_{out}, B_i without I_{Rel} ($P_{out} = \#(B_i \setminus I_{Rel})$). The probability of mask B_i belonging to the object is estimated by the relative frequency P_{out}/P_{in} and included in the final segment mask I_{final} if $P_{out}/P_{in} < 0.2$. Thus the final mask can contain pixels from segments which are not completely inside the original relevance map. The final mask I_{Final} is then computed as the additive superposition of the selected B_i and the skin color pixels are removed from this mask ($I_{Final} = \cup_i B_i \setminus I_{Skin}$).

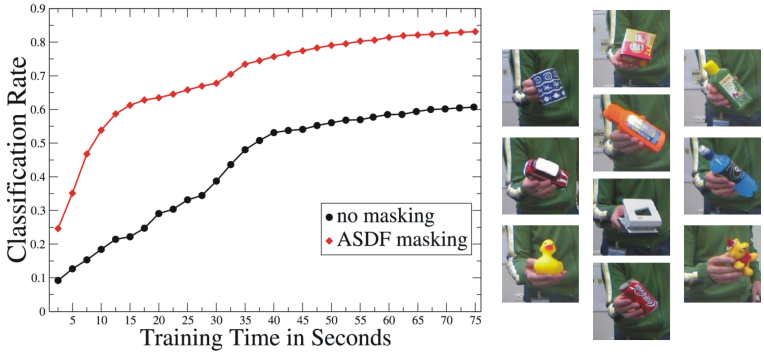

Fig. 4: Comparison of the classification results with ASDF segmentation and without segmentation and the images used for test and training. Training time results from the frame rate of 0.25 sec per image.

3 Experimental results

For evaluating the performance of the ASDF in combination with the recognition architecture an image database of 10 every-day objects of various, but partially equal, colors and shapes (Fig. 4) are shown in front of a cluttered background. We record 300 training images from a first training person and 300 test images for each object from a different person. The overall systems runs currently at 4 Hz including image recording, control of stereo vision, ASDF and object recognition, resulting in 75 sec presentation time for a full sequence of 300 training images. Training within the ASDF (Fig. 2) is performed with 3000 training steps per image with a constant learning rate to ensure the controlled online performance. As input for the ASDF we use the following feature maps: red, green, blue and disparity. Control experiments for performance evaluation have shown that the influence of the edge filters (with orientation at 0, 45, 90 and 135 degree) with respect to the object recognition performance in the current setting of the recognition architecture is insignificant. A high weighting parameter ξ^i for the edge filters may even decrease the object recognition performance.

To evaluate the incremental learning and online performance of the overall architecture, first we train nine objects from the training database with their complete 300 training images. Then the tenth object is trained in steps of 10 images (2.5 sec in Fig. 4) and a validation step is performed. Test performance is measured over all 300 test images of the currently trained object giving the classification rate as percentage of correctly recognized objects at this point of online learning. Then training proceeds until all 300 training images are used and test results are shown in Fig. 4 as average of classification rates for all ten objects. The plot shows the average learning curve for adding each of the ten objects as the final object to nine previously learned objects. The results demonstrate that training with the ASDF speeds up the recognition process and leads to a significantly higher recognition rate.

4 Conclusion

In this paper we have presented the ASDF architecture to enhance the performance of visual online object recognition in cluttered environments and "object-in-hand" presentation. The main advantage of our approach is the flexible generation and usage of a relevance map to group adaptively generated segements to object specific masks for figure ground seperation. We have shown that the architecture is reliable and fast enough to work online in real-time in a realistic human-machine interaction loop. The ability of changing the input stimuli for the relevance map as well as for the ASDF input F_i makes it easy to tailor the architecture for a wide range of applications. The current setting mainly relys on the combination of disparity, robust skin color detection, and color features and has proven to be robust to different clothing of the presenting persons. Other applications like e.g. scene interpretation may also include texture filters, motion maps, or other specifically useful topographic information while keeping the overall processing structure. This is an important step towards a flexible vision architecture to perform fast object recognition in arbitrary environments. Future work will be concentrated on optimizing the recombination step of the binarized ASDF filters B_i in the *Determine Object Segment* module and on evaluation the impact of different input maps F_i for the ASDF.

References

[1] G. Heidemann, H. Bekel, I. Bax, and H. Ritter. Interactive online learning. *Pattern Recognition and Image Analysis*, 15(1):55–58, 2005.

[2] H. Wersing and E. Körner. Learning optimized features for hierarchical models of invariant object recognition. *Neural Computation*, 15:1559–1588, 2003.

[3] C. Goerick, H. Wersing, I. Mikhailova, and M. Dunn. Peripersonal space and object recognition for humanoids. In *Proc. IEEE Humanoids, Tsukuba, Japan*, 2005.

[4] H. Wersing S. Kirstein and E. Körner. Online learning for object recognition with a hierarchical visual cortex model. In *Proc. Int. Conf. on Artif. Neur. Netw., Warsaw*, pages 487–492, 2005.

[5] Guo Dong and Ming Xie. Color clustering and learning for image segmentation based on neural networks. *IEEE Trans. on Neural Networks*, 16(14):925–936, 2005.

[6] Y. Jiang and Z.-H. Zhou. Som ensemble-based image segmentation. *Neural Processing Letters*, 20(3):171–178, 2004.

[7] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. *CVPRW, Washington D. C.*, 4:46, 2004.

[8] R. Basri E. Sharon, A. Brandt. Segmentation and boundary detection using multiscale intensity measurements. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition, Hawaii*, 1:469–476, 2001.

[9] J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink, and G. Sagerer. Improving adaptive skin color segmentation by incorporating results from face detection. In *Proc. IEEE ROMAN*, pages 337–343. IEEE, September 2002.