Lucas Paletta
John K. Tsotsos
Erich Rome
Glyn Humphreys (Eds.)

LNCS 3368

Attention and Performance in Computational Vision

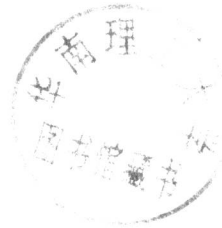
Second International Workshop, WAPCV 2004
Prague, Czech Republic, May 2004
Revised Selected Papers

Springer

Lucas Paletta John K. Tsotsos
Erich Rome Glyn Humphreys (Eds.)

Attention and Performance in Computational Vision

Second International Workshop, WAPCV 2004
Prague, Czech Republic, May 15, 2004
Revised Selected Papers



Springer

Volume Editors

Lucas Paletta
Joanneum Research, Institute of Digital Image Processing
Wastiangasse 6, 8010 Graz, Austria
E-mail: lucas.paletta@joanneum.at

John K. Tsotsos
York University, Department of Computer Science and Center for Vision Research
4700 Keele Street, Ontario, M3J 1P3, Toronto, Canada
E-mail: tsotsos@cs.yorku.ca

Erich Rome
Fraunhofer Insitute for Autonomous Intelligent Systems
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
E-mail: erich.rome@ais.fraunhofer.de

Glyn Humphreys
University of Birmingham, Behavioural Brain Sciences Centre
B15 2TT, Birmingham, UK
E-mail: g.w.humphreys@bham.ac.uk

Library of Congress Control Number: 2004117729

CR Subject Classification (1998): I.4, I.2, I.5, I.3

ISSN 0302-9743
ISBN 3-540-24421-2 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11378754 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3368

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Preface

In recent research on computer vision systems, attention has been playing a crucial role in mediating bottom-up and top-down paths of information processing. In applied research, the development of enabling technologies such as miniaturized mobile sensors, video surveillance systems, and ambient intelligence systems involves the real-time analysis of enormous quantities of data. Knowledge has to be applied about what needs to be attended to, and when, and what to do in a meaningful sequence, in correspondence with visual feedback. Methods on attention and control are mandatory to render computer vision systems more robust.

The 2nd International Workshop on Attention and Performance in Computational Vision (WAPCV 2004) was held in the Czech Technical University of Prague, Czech Republic, as an associated workshop of the 8th European Conference on Computer Vision (ECCV 2004). The goal of this workshop was to provide an interdisciplinary forum to communicate computational models of visual attention from various viewpoints, such as from computer vision, psychology, robotics and neuroscience. The motivation for interdisciplinarity was communication and inspiration beyond the individual community, to focus discussion on computational modelling, to outline relevant objectives for performance comparison, to explore promising application domains, and to discuss these with reference to all related aspects of cognitive vision. The workshop was held as a single-day, single-track event, consisting of high-quality podium and poster presentations. Invited talks were given by John K. Tsotsos about attention and feature binding in biologically motivated computer vision and by Gustavo Deco about the context of attention, memory and reward from the perspective of computational neuroscience.

The interdisciplinary program committee was composed of 21 internationally recognized researchers. We received 20 manuscripts responding to the workshop call for papers; each of the papers was assigned at least 3 double-blind reviews; 16 of the papers were accepted, as they corresponded to the requested quality standards and suited the workshop topic; 10 were attributed to 4 thematic oral sessions, and 6 were appropriate for representation as posters. The low rejection rate was commonly agreed to be due to the high quality of the submitted papers.

WAPCV 2004 was made possible by the support and engagement of the European Research Network for Cognitive Computer Vision Systems (ECVision). We are very thankful to David Vernon (Coordinator of ECVision) and Colette Maloney of the European Commission's IST Program on Cognition for their financial and moral support. We are grateful to Radim Sara, for the perfect local organization of the workshop and the registration management. We also wish to thank Christin Seifert, for doing the difficult task of assembling these proceedings.

October 2004

Lucas Paletta
John K. Tsotsos
Erich Rome
Glyn W. Humphreys

Organization

Organizing Committee

Chair

Lucas Paletta (Joanneum Res., Austria)
John K. Tsotsos (York Univ., Canada)
Erich Rome (Fraunhofer AIS, Germany)
Glyn W. Humphreys (Birmingham, UK)

Program Committee

Minoru Asada (Osaka Univ., Japan)
Gerriet Backer (Krauss SW, Germany)
Marlene Behrmann (CMU, USA)
Leonardo Chelazzi (Univ. Verona, Italy)
James J. Clark (McGill Univ., Canada)
Bruce A. Draper (Univ. Colorado, USA)
Jan-Olof Eklundh (KTH, Sweden)
Robert B. Fisher (Univ. Edinburgh, UK)
Horst-M. Gross (TU Ilmenau, Germany)
Fred Hamker (Univ. Münster, Germany)
John M. Henderson (MSU, USA)
Laurent Itti (USC, USA)
Christof Koch (Caltech, USA)
Bastian Leibe (ETH Zurich, Switzerland)
Michael Lindenbaum (Technion, Israel)
Nikos Paragios (ENPC Paris, France)
Satyajit Rao (Univ. Genoa, Italy)
Ronald A. Rensink (UBC, Canada)
Antonio Torralba (MIT, USA)
Jeremy Wolfe (Harvard Univ., USA)
Hezy Yeshurun (Tel Aviv Univ., Israel)

Sponsoring Institutions

ECVision — European Research Network for Cognitive Computer Vision Systems
Joanneum Research, Austria

Table of Contents

Attention in Object and Scene Recognition

Distributed Control of Attention
Ola Ramström, Henrik I Christensen 1

Inherent Limitations of Visual Search and the Role of Inner-Scene Similarity
Tamar Avraham, Michael Lindenbaum 16

Attentive Object Detection Using an Information Theoretic Saliency Measure
Gerald Fritz, Christin Seifert, Lucas Paletta, Horst Bischof 29

Architectures for Sequential Attention

A Model of Object-Based Attention That Guides Active Visual Search to Behaviourally Relevant Locations
Linda Lanyon, Susan Denham 42

Learning of Position-Invariant Object Representation Across Attention Shifts
Muhua Li, James J. Clark 57

Combining Conspicuity Maps for hROIs Prediction
Claudio M. Privitera, Orazio Gallo, Giorgio Grimoldi, Toyomi Fujita, Lawrence W. Stark 71

Human Gaze Control in Real World Search
Daniel A. Gajewski, Aaron M. Pearson, Michael L. Mack, Francis N. Bartlett III, John M. Henderson 83

Biologically Plausible Models for Attention

The Computational Neuroscience of Visual Cognition: Attention, Memory and Reward
Gustavo Deco 100

Modeling Attention: From Computational Neuroscience to Computer Vision
Fred H. Hamker 118

Towards a Biologically Plausible Active Visual Search Model
Andrei Zaharescu, Albert L. Rothenstein, John K. Tsotsos 133

Modeling Grouping Through Interactions Between Top-Down and Bottom-Up Processes: The Grouping and Selective Attention for Identification Model (G-SAIM)
Dietmar Heinke, Yaoru Sun, Glyn W. Humphreys 148

TarzaNN: A General Purpose Neural Network Simulator for Visual Attention Modeling
Albert L. Rothenstein, Andrei Zaharescu, John K. Tsotsos 159

Applications of Attentive Vision

Visual Attention for Object Recognition in Spatial 3D Data
Simone Frintrop, Andreas Nüchter, Hartmut Surmann 168

A Visual Attention-Based Approach for Automatic Landmark Selection and Recognition
Nabil Ouerhani, Heinz Hügli, Gabriel Gruener, Alain Codourey 183

Biologically Motivated Visual Selective Attention for Face Localization
Sang-Woo Ban, Minho Lee 196

Accumulative Computation Method for Motion Features Extraction in Active Selective Visual Attention
Antonio Fernández-Caballero, María T. López, Miguel A. Fernández, José Mira, Ana E. Delgado, José M. López-Valles 206

Fast Detection of Frequent Change in Focus of Human Attention
Nan Hu, Weimin Huang, Surendra Ranganath 216

Author Index 231

Distributed Control of Attention

Ola Ramström and Henrik I Christensen

KTH, 10044 Stockholm, Sweden
{olar, hic}@nada.kth.se
http://www.nada.kth.se/cvap

Abstract. Detection of objects is in general a computationally demanding task. To simplify the problem it is of interest to focus the attention to a set of regions of interest. Indoor environments often have large homogeneous textured objects, such as walls and furniture. In this paper we present a model which detects large homogeneous regions and uses this information to search for targets that are smaller in size. Homogeneity is detected by a number of different descriptors and a coalition technique is used to achieve robustness. Expectations about size allow for constraint object search. The presented model is evaluated in the context of a table top scenario.

1 Introduction

In everyday life we have the impression to constantly perceive everything in the visual field coherently and in great detail. One would normally expect to notice a gorilla walking across the scene while watching a basketball game. However, we often fail to notice salient events that are not expected [SM01]. Indeed, only a small fraction of the visual properties of a scene is attended and consciously perceived. Tsotsos' complexity analysis [Tso90] concludes that an attentional mechanism, that selects relevant visual features and regions for higher level processes, is required to handle the vast amount of visual information in a scene.

Garner [Gar74] found that similarity between objects is measured differently depending on whether they differ in integral or separable features. From these findings Treisman and Gelade [TG80] developed the "Feature-Integration Theory of Attention", which states that integral features (denoted dimensions) are processed pre-attentively across the visual field. Consequently a target will appear to pop-out if it is unique in one dimension, such as a red target among green distractors. However, in conjunction search, when the target is not uniquely described by any dimension, such as in search for a red vertical target among red horizontal and green vertical distractors, we must inspect each object in turn and hence the search time will be proportional to the number of distractors. The theory furthermore predicts perceptual grouping to be processed pre-attentively across the visual field. In the conjunction search example, a red vertical target among red horizontal and green vertical distractors, the target can appear to pop-out if e.g. all green objects are on the left side and the red are on the right side of a display. The two groups need to be inspected in turn and the red vertical target will pop-out as the only vertical object in the red group.

L. Paletta et al. (Eds.): WAPCV 2004, LNCS 3368, pp. 1–15, 2005.

Wolfe et. al. [WCF89] and others have found many cases where conjunction search is much faster than the Feature-Integration Theory predicts, clearly different search strategies are used depending on the scene-properties. Treisman and Sato revised the theory [TS90] and confirmed the use of multiple strategies. One of these strategies is to inhibit parts of the background; they found that search performance depends on the homogeneity of the background. Apparently, the background context is processed to ease the search for foreground objects. Moreover, many experiments have demonstrated that we detect and implicitly learn unattended background context [KTG92] [DT96] [HK02]. This implicit memory affects our visual search performance but cannot be accessed by our conscious mind. The Inattentional Amnesia Hypothesis [Wol99] explains this as: Although we perceive and process the whole visual field, only attended locations are consciously remembered.

Clearly, the processing of unattended background information plays an important role in object detection.

The Coherence Theory [Ren00] defines the concept of volatile proto-objects that are formed pre-attentively across the visual field. Proto-objects are described as "relatively complex assemblies of fragments that correspond to localized structures in the world"; for example occluded objects are processed to estimated complete objects [ER92]. Attention is needed for proto-objects to become stable and for conscious processes to access its information. When attention is released the proto-objects become volatile again. This implies that there is little short-term memory apart from what is being attended; this is consistent with the Inattentional Amnesia Hypothesis. Recent biological findings [MvE02] confirm pre-attentive processes corresponding to proto-objects in the Coherence Theory.

We propose a model that is inspired by the Coherence Theory in that it models a way for pre-attentive proto-objects to become stable when attended. The assemblies of proto-objects are used as background context and their statistics are used to efficiently search for objects that are defined only by their size.

1.1 Related Work

Most models of visual attention are space based [Mil93] [Wol94] [TSW+95] [IK00]. Some have modeled different aspects of object based visual attention: Li [Li00] has developed a model for pre-attentive segmentation, Sun and Fisher for computing saliency of hierarchical segments and the attentional shift among these [SF03]. However, none of the above models how pre-attentive segments become stable when attended as predicted in the Coherence Theory.

2 Conceptual Model

A model has been developed which searches for target objects of an expected size. The model is designed to be implemented on a distributed system and uses concepts from game theory to minimize the need for inter-process communication.

The strategy is to use knowledge of the environment; e.g. in a living-room we might expect large items with homogeneous surfaces such as table and cupboard. The large

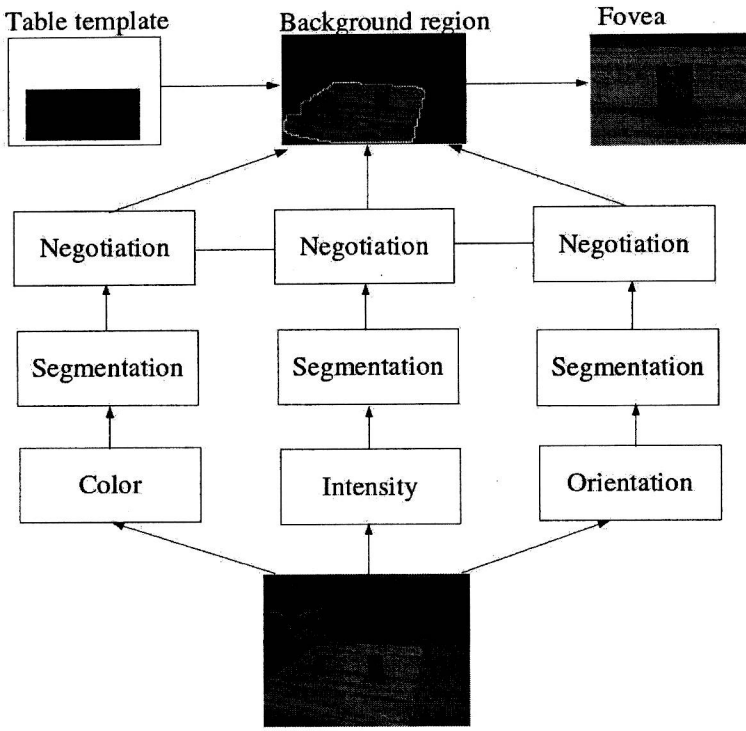


Fig. 1. A raw image is processed by a set of distributed nodes resulting in a set of background regions, which often corresponds to large objects

homogeneous regions provide layout and contextual information of the scene, which can be used to guide the attention.

Figure 1 illustrates the system: A raw image from a camera is decomposed into a set of feature maps at separate nodes, namely color, intensity, and orientation (see section 3). A segmentation algorithm searches for large homogeneous regions locally at each node. The resulting segments are sensitive to variations in the intrinsic parameters and the camera pose, similar to proto objects discussed in [Ren00]. A negotiation scheme forms coalitions of segments, which are more stable than the individual segments, similar to the nexus discussed in [Ren00]. The coalitions of segments are formed by only sharing real valued coalition values across the nodes and the final winning segmentation mask (see section 4). The coalitions are denoted background regions and provide layout information; a spatial template selects interesting background regions. The feature statistics of each interesting background region is computed as local context and saliency can thereafter be computed at each node with respect to the local context. Only a sparse set of the saliency data need to be integrated across the nodes to achieve accurate object detection (see section 5). As reference a center-surround saliency algorithm has been developed (see section 6). The performance of the model is evaluated in section 7.

3 Image Processing

The system processes a raw image into a set of feature maps which are subsequently segmented into a set of homogeneous background regions. The raw image has 300 $\times$ 224 pixels in resolution using the YUV color space and the decomposed feature maps have 75 $\times$ 56 pixels in resolution with different dimensionality. The format of the raw images enable accurate processing of image features and the four times down sampling of the feature maps reduce noise and thus improves extraction of homogeneous regions.

From experimental psychology [ER92]and biology [MvE02] it is clear that the visual cortex performs segmentation preattentively using several separate visual features, Julez denoted such features textons [Jul81]. Treisman [TG80] found that segments cannot be formed by conjunctions of separate features. We will in this model restrict us to three separate feature dimensions: Color, intensity, and orientation. These are suitable to the environment we will use for evaluation. Note that it is not claimed that these are better than any other feature dimensions nor that three separate dimensions is an optimal number of dimensions. However, these feature dimensions allow us to compare the results to [IK00], where similar although not identical features are used.

Feature map identity is denoted $d \in \{color, intensity, orientation\}$ and the feature maps are denoted f^d. In order to make output from the different feature maps comparable when searching for homogeneous regions all feature maps are normalized to have zero mean and unit variance.

4 Background Regions

The processing of feature maps is distributed over a set of processing nodes. Distributed computing increases the processing power if the communication across nodes is limited. It is of interest to enable distributed control where the nodes processes a majority of information locally and only integrates a small subset of the full data set.

This is enabled using background regions, which are created using a game theoretic negotiation scheme.

Knowing the context of a homogeneous background region we can efficiently search for target objects, e.g. knowledge of the appearance of a tablecloth can efficiently guide search for a cup on a table.

Having a rough estimate of the pose of a table relative to the camera, we need a mechanism to find background regions, which might represent large homogeneous objects and are stable with respect small variations in camera pose.

The mean-shift algorithm [CM99] is a fairly well established segmentation algorithm based on local similarities. It has two intrinsic parameters: spatial scale s and feature range r. Small changes in these two intrinsic parameters or in the camera pose can result in different segmentation results.

However, by using redundant mean-shift segments, corresponding to different parameters, we can increase the stability. We will in this section define a negotiation scheme to form coalitions of similar mean-shift segments. Such coalitions are used to extract a segmentation mask which is more stable than the ingoing members.

4.1 Clustering of Redundant Mean-Shift Segments

The mean-shift segmentation depend on two intrinsic parameters (r, s). Different values of these parameters might result in different segmentation result. Since they relate to distances in the spatial and feature domain, such variations are related to variations in pose and illumination. To increase the stability we compute the mean-shift segmentation varying $r \in M_r$ and $s \in M_s$. We will in this work restrict $M_r = \{2, 3, 4\}$ and $M_s = \{0.15, 0.22, 0.33\}$. Furthermore, since we are interested in background regions which are much larger than the expected target size, we select only the $N = 4$ largest segments and discard all segments smaller than 400 pixels (10% of the feature map size) for each selection of $(r, s) \in M_r \times M_s$.

The mean-shift segmentation algorithm is processed locally at each node for each $(r, s) \in M_r \times M_s$. Let P^d represent the resulting set of mean-shift segments at node d; hence the size of $|P^d| \leq N|M_s \times M_r|$. Each mean-shift segment $p_i \in P^d$ is associated with a segmentation mask S_i^d and a histogram of the feature values inside the segmentation mask h_i^d. The similarity between two mean-shift segments p_i and p_j is defined as the normalized intersection of their segmentation masks and histograms:

$$Sim(i, j, d) = \frac{S_i^d \cap S_j^d}{|S_i^d| + |S_j^d|} \cdot \frac{h_i^d \cap h_j^d}{|h_i^d| + |h_j^d|} \tag{1}$$

The second factor is fairly standard in histogram matching, the first factor borrows the same normalization technique and gives a penalty when they differ spatially in size, location, and shape.

To find the optimal background regions we need to evaluate all possible cluster combinations of mean-shift segments, which has exponential complexity $O(2^{|P^d|})$. This complexity is reduced using a modified version of the coalition formation process proposed by [SK98], which only have square complexity with respect to the number of mean-shift segments $O(|P^d|^2)$.

4.2 Negotiation

Coalitions of mean-shift segments are formed by an iterative negotiation process. At iteration $t = 0$ each mean-shift segment $p_i \in P^d$ broadcast its description, (S_i^d, h_i^d), and selects a set of possible coalition members $C_i^d(0) \subseteq P^d$, including all other mean-shift segments with a spatial similarity larger than th:

$$C_i^d(0) = \{\forall p_j \in P^d | Sim(i, j, d) > th\} \tag{2}$$

We define the value of a coalition $C_i^d(t)$ at iteration t as:

$$V_i^d(t) = \sum_{j \in C_i^d(t)} Sim(i, j, d) \tag{3}$$

mean-shift segments are valued relative to their contribution, hence the value of p_j in coalition $C_i^d(t)$ is:

$$V_i^d(j, t) = Sim(i, j, d) \tag{4}$$

Thus, each segment forms a coalition including similar segments at the same node. From this set of coalitions, background regions are iteratively extracted by a distributed negotiation scheme. In each iteration of the negotiation the strongest coalition across all nodes is chosen and used to form a background region and to inhibit segments in succeeding negotiation iterations.

In more detail, stable coalitions are formed when each mean-shift segment p_i at each node d iteratively perform the following:

1. Compute and announce $V_i^d(t) = \sum_{j \in C_i^d(t)} V_i^d(j,t)$ to all other segments at all nodes.
2. Choose the highest among all announced coalition values, $V_{max}(t)$.
3. If no other coalition at any node has a stronger coalition value, $V_i^d(t) = V_{max}(t)$, then compute the weighted segmentation mask $W^d = q \sum_{j \in C_i^d} V_i^d(j,t) S_j^d$, where $q \in \mathbf{R^1}$ is a constant which normalize W^d to have maximal value one. Remove all $p_i \in C_i^d$ from further negotiation.
4. Update $V_i^d(j,t+1) = (1 - 2\frac{S_i^d \cap S_{max}^d}{|S_i^d| + |S_{max}^d|}) V_i^d(j,t)$;
5. Start over from 1.

At each iteration a weighted segmentation mask, W^d, is computed and $|C_i^d|$ is decreased for at least one mean-shift segment. The process will be repeated until all C_i^d are empty or for a fixed number of iterations.

Note that the set of nodes only compares values of maximal coalitions, all other computation of image data is processed locally at each node. The weighted segmentation masks resulting from a raw image illustrated in figure 2.

4.3 Segmentation Mask

A segmentation mask can be extracted by thresholding each weighted segmentation mask W^d. However, we do not have any analytic way to extract such threshold value. Instead we calculate a Gaussian-mixture model (GMM) of the complement region of W^d in the associated feature map f^d, using 5 Gaussian models.

The resulting probability maps are suitable competition to W^d. We compute the E_{ij} maps, which is the probability of pixel i to belong to the Gaussian model j, for each $j = 1,2,3,4,5$. Furthermore, we denote the Gaussian model at W^d with $j = 0$, hence E_{i0} is the probability map for pixel i to belong to the Gaussian model at W^d. Finally, the probability for pixel i to belong to background region C_{max}^d is the joint probability $W_i^d E_{i0}$.

The segmentation mask, S^d, associated with C_{max}^d is defined as the pixels i where $W_i^d E_{i0} > E_{ij}$ for all $j = 1,2,3,4,5$.

The segmentation mask from a raw image is illustrated in figure 3.

4.4 Region Completion

[ER92] demonstrated in a series of experiments that the visual cortex preattentively completes homogeneous regions to object hypothesis. Inspired by this result we perform region completion of the segmentation masks S^d. The object completion is not as advanced as the completion process demonstrated by [ER92], however it enables

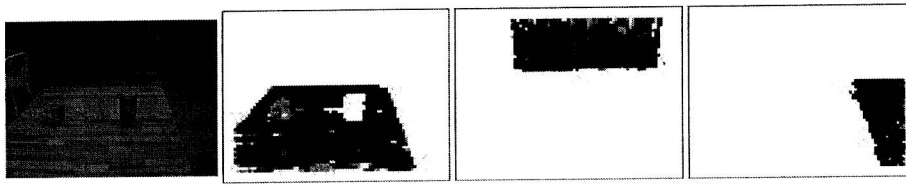


Fig. 2. Raw image and weighted summation mask for first, second, and third background region

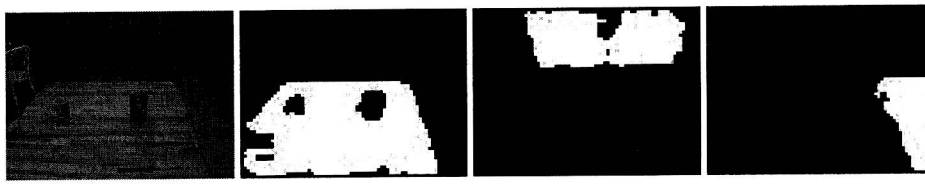


Fig. 3. Raw image and segmentation mask for first, second, and third background region

detection of objects within homogeneous regions. One obvious completion process is to fill in holes. Furthermore, objects e.g. at the border of a table often pop-out from the table leaving a notch on the border of the segmentation mask. Regions corresponding to artificial large indoor-objects, in the set of evaluation scenes, are often square or have some vertical or horizontal straight lines. Following this discussion, we define the region completion as filling in gaps where a vertical or horizontal straight line can be attached to the original segment. Note that a notch in the corner will not be completed by this process as predicted by [ER92] .

Moreover, we do not want to detect other overlapping background regions as salient. Therefore we restrict object completion to regions not occupied by other original segmentation masks.

Figure 4 illustrates a completed segmentation mask. We observe that the region is more compound.

This process is solely based on intuition from the [ER92] experiments and can obviously be improved. However, the completion process enables accurate objects detection and is sufficient at this point in the development process.

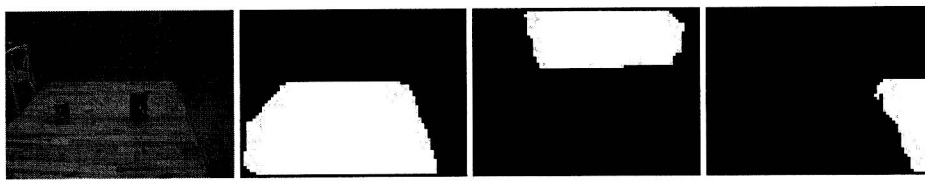


Fig. 4. Raw image and completed segmentation mask for first, second, and third background region

5 Target Objects

All background regions that are spatially similar to a template are attended, e.g. all background regions in the lower part of the visual field. The feature statistics of each background regions is extracted and used to search for outliers which are likely to be a target objects.

This process involves sending a sparse set of conspicuous data to all other nodes. After this integration of sparse data, salient locations can be attended by foveated vision, and the target hypothesis can be figure-ground segmented.

5.1 Background Region of Interest

The task is to find objects on a table which are expected to occupy a large fraction of the lower half of the scene. We use a template to represent this knowledge. A background region which overlaps more than 25% with the template it is considered a background region of interest. Figure 5 illustrates the template used here.

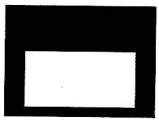


Fig. 5. Template for background region

5.2 Foreground Region of Interest

To find foreground regions of interest (ROI) we calculate the feature statistics of each background region of interest. The weighted summation mask W^d of each such background region is used to calculate (m^d, Σ^d) at each node $d \in \{color, intensity, orientation\}$:

$$m^d = \frac{1}{|W^d|} \sum_{x=0}^{X} \sum_{y=0}^{Y} f^d(x,y) W^d \tag{5}$$

$$\Sigma^d = \frac{1}{|W^d|} \sum_{x=0}^{X} \sum_{y=0}^{Y} (f^d(x,y) - m^d)(f^d(x,y) - m^d)^T W^d \tag{6}$$

where W^d is the sum of all pixels in W^d.

We calculate the set of pixels p_S^d which can be excluded from background region S^d with confidence γ with respect to a Normal distribution (m^d, Σ^d). If this set is larger that q, the confidence value is increased and a new set p_S^d is computed. This process is repeated until the size $|p_S^d| < q$.

In the current implementation we have chosen the set $\gamma \in 0.5, 0.6, 0.7, 0.8, 0.9, 1$ and $q = 0.25|S|$ without further investigation.

The sparse set of conspicuous pixels p_S^d is distributed to all other nodes. At each node an integrated saliency map is constructed from the sum of all p_S^d. The integrated

saliency map in convolved with a Gaussian kernel with a standard deviation equal to the expected target size. Each peak of the saliency map, which is larger than one, is extracted as ROI. Hence, we only consider regions, of expected target size, that at least one node has found conspicuous.

Each peak, which is selected in the saliency map, is attended with a foveated camera with sixteen times higher resolution, right-most image in figure 6.

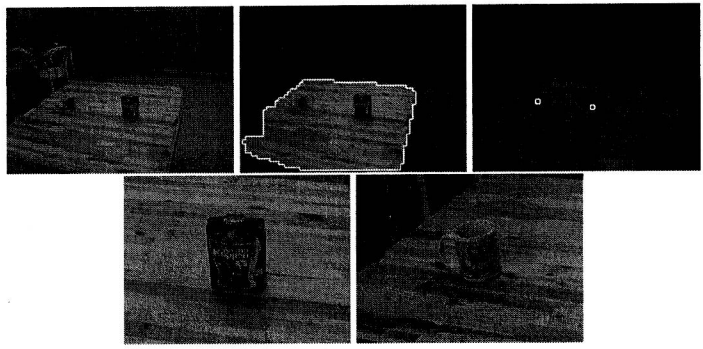


Fig. 6. Top row: Raw image, background region, and interest points at background region Bottom row: foveated view at interest points

6 Center-Surround Saliency

A well-established attention model is the center-surround saliency model developed by [IK00]. The source code is available at `http://ilab.usc.edu/toolkit/`. However, to suite our choices of feature map definitions, an own implementation has been developed based on [IK00]. Our implementation uses fewer scales and integral feature maps (described in section 3), and is hence not equally good as the original. However, it has similar properties as the original and is used here as a comparison of typical behavior. It will be denoted CS-search.

It should be pointed out that CS-search does not have the same focus on distributed processing as the proposed model; at the final step complete pixel maps are integrated.

6.1 CS-Search

Using the integral images we define center-surround saliency as the Euclidean distance between the mean feature vector inside a center rectangle (rc) and the mean feature vector inside a surrounding larger rectangle (rs). Let (rc, rs) denote a center rectangle with width rc and a surrounding rectangle with width rs, as illustrated in figure 7.

Six different center-surround saliency maps $CS^{d}_{(rc,rs)}$ are computed at each node with:

$$(rc, rs) \in \{(20, 50); (20, 60); (30, 60); (30, 70); (40, 70); (40, 80)\}$$

and

$$d \in \{color, intensity, orientation\}$$