Active 3D scene segmentation and detection of unknown objects

Mårten Björkman and Danica Kragic

Abstract— **We present an active vision system for segmentation of visual scenes based on integration of several cues. The system serves as a visual front end for generation of object hypotheses for new, previously unseen objects in natural scenes. The system combines a set of foveal and peripheral cameras where, through a stereo based fixation process, object hypotheses are generated. In addition to considering the segmentation process in 3D, the main contribution of the paper is integration of different cues in a temporal framework and improvement of initial hypotheses over time.**

I. Introduction

The next important milestone for embodied machine vision systems is to make them flexible and robust in a variety of environments and tasks. Recent examples of machine vision systems for humanoid robots [1] demonstrate the necessity for active aspects of the system, both in terms of actively changing the parameters of the vision system and interacting with the environment. Visual attention serves as a core process for generating hypotheses about the structure of the scene and allows the system to deal with the complexity of natural scenes. The requirements on machine vision systems are highly dependent on the task, and have historically been developed with this in mind. To deal with the complexity of the environment, prior task and context information have commonly been integrated with low level processing structures, the former being denoted as top-down and latter bottom-up principle. This has many times been motivated by human visual processing. Humans build a representation of a visual scene using a temporal process of integration of several scene 'glances', [2]. A cumulative memory allows them to detect and recall objects seen during several short, separate presentations even when these are several minutes apart. Likewise, in machine vision systems, generating hypotheses about objects in the scene is a necessary prerequisite for interaction. Although generation of hypotheses may be solved through a classical process of object recognition, our main interest is to generate hypotheses of *previously unseen* objects. This process may also help the recognition and classification processes by reducing the search space.

The main contribution of the work presented here is 3D scene segmentation based on the integration of several visual cues. However, this work should not be viewed as a typical work on image segmentation, since the hypotheses of objects are generated in 3D, thus facilitating shape attribution and pose estimation. We also show how segmentation can evolve over time and gradually produces better hypotheses. This is another important difference from the classical segmentation approaches that are typically demonstrated on a single image. We also evaluate the presented method using an active humanoid head in realistic scenarios. As said, this work relates to classical approaches to segmentation, however, most of these have been demonstrated only in the image space. Segmentation in 3D offers not only the possibility to attribute 3D regions based on their shape properties, [3], but also gives direct input to an object grasping and manipulation system, [4].

The work presented here is related to image segmentation methods such as GrabCut, [5] in that it models segmentation as a hypotheses generation and verification process. However, in the GrabCut approach only two hypotheses are used: one for the foreground and one for the background. We will show that in a 3D segmentation process, additional hypotheses increase the quality of the results. In addition, we employ belief propagation for verification of hypotheses, that differs from the energy minimization approaches of [5] and [6]. The most important difference and also a contribution is that our method uses a temporal framework and verifies the hypotheses over time, whereas methods of [5] and [6] work on a single image.

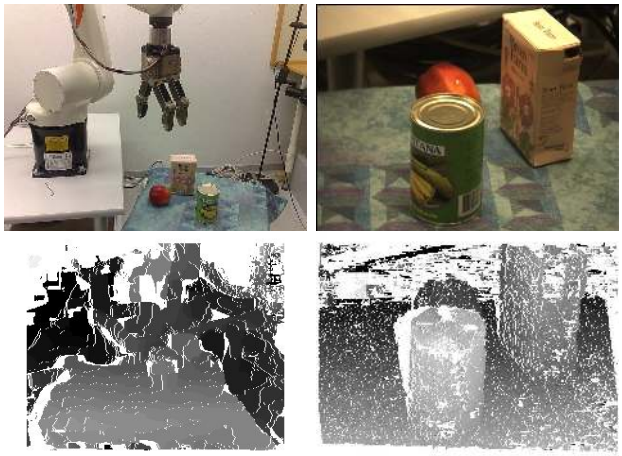

Fig. 1. Left: A peripheral view of a typical experimental scene (upper), with a corresponding disparity map (lower). Right: A foveal view of the same scene (upper) with a disparity map (lower).

The goal behind the presented work is to enable a vision guided robotic system to learn about its environment through interaction with the objects therein. First, the hypotheses of possible scene objects need to be generated within reasonable time. This means that an attention system that directs the vision system towards the most conspicuous parts of the

This work was supported by EU through the project PACO-PLUS, IST-FP6-IP-027657, and GRASP, IST-FP7-IP-215821 and Swedish Foundation for Strategic Research. The authors are with the Centre for Autonomous Systems and Computational Vision and active Perception Lab, CSC-KTH, Stockholm, Sweden. celle,dani@kth.se

scene is needed. Second, extraction of attributes related to an observed object often requires it first to be segregated from its background. With the attention system already presented elsewhere, [4] here we concentrate on the second problem, figure-ground segmentation of objects in typical indoor scenes.

A. Experimental platform

Our experimental platform includes the 7-joint Armar III robotic head, [7]. The stereo head carries four Point Grey Dragonfly cameras grouped in two pairs, a peripheral and a foveal one, see Fig. 1. These are parts of an existing vision system [4] that uses attention in the peripheral view to direct cameras towards nearby regions of interest. After gaze direction such regions are placed in fixation in the foveal view. Binocular disparities are exploited in both views, for gaze control in the peripheral view and for object analysis and manipulation in the foveal view.

Visual attention, gaze control and manipulation are beyond the scope of this paper, yet they serve as the context in which the presented segmentation approach is to be used. The disparity maps shown in Fig. 1 are computed using Stable Matching [8], a method that is able to cope with wide disparity ranges. The range we typically use for the foveal views, 64 pixels, is more than what most disparity methods are able to handle within reasonable time. Stable Matching is suitable for our needs, since instead of aiming for the highest possible density, it tries to minimize the number of false positive matches.

B. Assumptions

In typical indoor environments most physical objects are placed on flat surfaces. However, based on our previous work [9], an object may be impossible to separate from the surface: they may be similar in appearance[1]. In this paper we thus expand a typical framework for figure-ground segmentation with an additional model, that of a flat surface. A foreground object is defined as the object fixated on by the stereo system. Thus it is expected to be placed in the center of view at about zero disparity. In GrabCut [5], a foreground object is similarly defined by a given bounding box. We also assume that models change only slightly while the object is in fixation and that the system knows when the gaze is shifted and segmentation has to be reinitialized. Finally, the system should be able to operate autonomously through sequences of gaze shifts and tolerate disparity data that arises through non-perfect calibration and limited disparity search ranges.

II. Prerequisites

The segmentation method presented in this paper is based on measurements of colors and binocular disparities. Given these measurements the scene is divided into 3 parts; a foreground object, a flat surface and a background. We later describe a scheme with which model parameters can be estimated and images segmented on a per-pixel basis.

[1]See http://www.csc.kth.se/∼danik/HeadArmDemo-centering.avi for an example of using the system for object grasping.

A. Measurements and model parameters

An image, here assumed to be part of a stereo pair, contains image points that are characterized by their positions (x_i, y_i) and measured colors $c_i = (h_i, s_i, v_i)$ given in HSV space, with h_i being the hue, s_i the saturation and v_i the luminance value. Also associated to each such point is a measured binocular disparity d_i, that can either be a value within a given disparity range or be undefined. There are primarily two reasons for the disparity to be undefined; either a point lacks sufficient texture to be matched in stereo or it is occluded in one of the two images. We denote the total set of image measurements by $\mathbf{m} = \{m_i\}$, with each point characterized by $m_i = (p_i, c_i)$, where $p_i = (x_i, y_i, d_i)$ are the three spatial measurements and c_i is the color.

We assume each image point to originate from one of three possible scene parts; a foreground object $\mathbf{F}$, a planar surface $\mathbf{S}$ and a background $\mathbf{B}$, each of which a characterized by a corresponding model. The foreground $\mathbf{F}$ is assumed to be a connected set of 3D points representing some physical object in the center of the image and close to the fixation point. It is further assumed that the scene contains a large planar surface $\mathbf{S}$, upon which objects could be placed. The background $\mathbf{B}$ is defined as all points that neither belong to the foreground nor the planar surface. The scene part that a particular point p_i belongs to is given by a label $l_i \in L$, where $L = \{l_f, l_s, l_b\}$ is the set of values that corresponds to each scene part respectively.

The three different parts of the scene are modeled by a set of parameters $\theta = \theta_f \cup \theta_s \cup \theta_b$. These will be defined later in Section II-B. Given the measurements $\mathbf{m}$ our goal is to find the most likely parameter set θ and distribution of labels $\mathbf{l} = \{l_i\}$. The joint probability of $\mathbf{m}$ and $\mathbf{l}$ given θ can be written as

$$p(\mathbf{m}, \mathbf{l}|\theta) = p(\mathbf{m}|\mathbf{l}, \theta)p(\mathbf{l}|\theta) \tag{1}$$

with the measurement distribution given by

$$p(\mathbf{m}|\mathbf{l}, \theta) = \prod_i p(m_i|\theta_f)^{I_i^f} p(m_i|\theta_b)^{I_i^b} p(m_i|\theta_s)^{I_i^s} \tag{2}$$

and the prior label probabilities

$$p(\mathbf{l}|\theta) = \prod_k p(l_k) \prod_i \prod_{j \in N_i} p(l_i, l_j). \tag{3}$$

In the equations above, I_i^x equals 1 if $l_i = l_x$ and 0 otherwise, and N_i is the set of neighbors to point i. The priors in (3) will be defined later in Section III-A.

B. Scene part models

For all three scene parts we model the distributions of image point positions, disparities and colors. The spatial distributions of the background and surface parts are assumed to be uniform across the image space $\mathbf{X}$, i.e. $p(x_i, y_i|\theta_b) = p(x_i, y_i|\theta_s) = 1/N$, where $N = |\mathbf{X}|$ is the number of image points. Their counterparts in disparity space are modeled as Gaussians with $p(d_i|\theta_b) = n(d_i; d_b, \Delta_b)$ and $p(d_i|\theta_s) = n(d_i; \alpha_s x_i + \beta_s y_i + \delta_s, \Delta_s)$, where $d_s = (\alpha_s, \beta_s, \delta_s)$ are disparity parameters that belong to the surface model. Here

we denote by $n(x; \bar{x}, \Delta)$ a Gaussian distribution of a d-dimensional variable x, with mean $\bar{x}$ and covariance Δ,

$$n(x; \bar{x}, \Delta) = \frac{1}{\sqrt{(2\pi)^d |\Delta|}} \exp^{-\frac{1}{2}(x-\bar{x})^\top \Delta^{-1}(x-\bar{x})}$$

While the conditional probability of the background is the same for all image points, it varies for the flat surface. Note that $d = \alpha_s x + \beta_s y + \delta_s$ represents a plane in (x, y, d) space that, assuming a projective camera, corresponds to a plane also in the 3D metric space. The spatial positions of the foreground object are modeled using a single 3D Gaussian that includes both image point positions and disparities, with conditional probabilities given by $p(x_i, y_i, d_i|\theta_f) = n(p_i; p_f, \Delta_f)$. The disparity dimension is ignored for points with undefined disparities and for these points Δ_f is replaced by its projection in (x, y)-space.

The distributions of colors within a given scene part are assumed to be the same for all image points. We represent such distributions as 2D histograms, based on hue and saturation; $p(h_i, s_i|\theta_b) = H_b(h_i, s_i)$, $p(h_i, s_i|\theta_s) = H_s(h_i, s_i)$ and $p(h_i, s_i|\theta_f) = H_f(h_i, s_i)$. With color histograms included in the set of model parameters, the complete set is given by

$$\begin{aligned}\theta_f &= \{p_f, \Delta_f, c_f\},\\ \theta_b &= \{d_b, \Delta_b, c_b\},\\ \theta_s &= \{d_s, \Delta_s, c_s\},\end{aligned}$$

where c_f, c_b and c_s denote the color histogram bins stacked into vectors. The other parameters are the means and variances of the Gaussians mentioned above. The joint measurement conditionals can finally be summarized as

$$\begin{aligned}p(m_i|\theta_f) &= n(p_i; p_f, \Delta_f)\, H_f(h_i, s_i),\\ p(m_i|\theta_b) &= N^{-1} n(d_i; d_b, \Delta_b)\, H_b(h_i, s_i),\\ p(m_i|\theta_s) &= N^{-1} n(d_i; \alpha_s x_i + \beta_s y_i + \delta_s, \Delta_s)\, H_s(h_i, s_i).\end{aligned}$$

III. Estimating the model parameters

One way of estimating the model parameters θ would be to determine a maximum likelihood estimate for $p(\mathbf{m}|\theta)$ using the Expectation-Maximization (EM) algorithm, with all labels $\mathbf{l}$ treated as hidden variables. Given $p(\mathbf{m}, \mathbf{l}|\theta)$, that was defined in (1), the hidden variables can be eliminated through marginalization,

$$p(\mathbf{m}|\theta) = \sum_{\mathbf{l}} p(\mathbf{m}, \mathbf{l}|\theta).$$

The EM algorithm is based on maximization of an objective function $Q(\theta|\theta')$ that given a previous estimate θ' is guaranteed to increase $p(\mathbf{m}|\theta)$. In the first step of the algorithm, the Expectation step, $Q(\theta|\theta')$ is expressed as the expected value of $\log p(\mathbf{m}, \mathbf{l}|\theta)$ with respect to the conditional distribution $w(\mathbf{l}) = p(\mathbf{l}|\mathbf{m}, \theta')$ under the previous estimate θ', that is

$$Q(\theta|\theta') = \sum_{\mathbf{l}} w(\mathbf{l}) \log p(\mathbf{m}, \mathbf{l}|\theta). \tag{4}$$

The model parameters θ are updated in the second step, the Maximization step, through maximization of $Q(\theta|\theta')$. This two-step procedure is then repeated until convergence.

As can be seen in (4), the algorithm essentially performs a summation over the conditional distribution $w(\mathbf{l})$. Unfortunately, this fact makes the EM algorithm intractable for our purpose. In our case labels from neighboring image points are assumed to be dependent. This means that the summation has to be done across all 3^N possible combinations of labels, where N is the number of image points, rather than $3N$ combinations that would otherwise have been the case.

To make summation computationally tractable, we introduce an approximation that treats labels as if they are in fact independent. We do this by replacing the conditional distribution $w(\mathbf{l})$ with the product of the marginal distributions for each unobserved label, that is

$$\hat{w}(\mathbf{l}) = \prod_i w(l_i) = \prod_i p(l_i|\mathbf{m}, \theta').$$

Since a measurement m_i at a given point only depends on the label l_i at that point, not on neighboring labels, the summation in (4) becomes

$$Q_1(\theta|\theta') = \sum_i \sum_{l_i \in L} w(l_i) \log p(m_i, l_i|\theta). \tag{5}$$

With dependencies ignored the joint probability for a single point (see (1) and (2)) can be written as

$$p(m_i, l_i|\theta) = p(m_i|l_i, \theta) p(l_i),$$

where

$$p(m_i|l_i, \theta) = p(m_i|\theta_f)^{I_i^f} p(m_i|\theta_b)^{I_i^b} p(m_i|\theta_s)^{I_i^s}.$$

Note that it is only when marginal distributions are summed up to produce an estimate of θ that dependencies between labels are ignored. The marginals $w(l_i)$ themselves determine the final segmentation and are computed with dependencies taken into consideration.

A. An iterative two-stage approach

Our optimization approach consists of two stages, that are iterated until either convergence or the number of iterations reaches a given maximum. Given an initial estimation of the conditional marginals for all individual labels, or the marginals from the previous iteration, the model parameters are estimated by maximizing $Q_1(\theta|\theta')$ in (5), where θ' are the parameters from which the marginals were computed. The corresponding update functions for all foreground parameters can be found in the appendix.

In the second stage the conditional marginals $w(l_i) = p(l_i|\mathbf{m}, \theta)$ are recomputed for each label. This is done using loopy belief propagation [10]. First, however, we have to rewrite the equations into energy functions suitable for belief propagation. From Bayes' rule and using the fact that m_i only depends on l_i, we have that

$$p(\mathbf{l}|\mathbf{m}, \theta) = \frac{p(\mathbf{m}|\mathbf{l}, \theta) p(\mathbf{l}|\theta)}{p(\mathbf{m}|\theta)} = \frac{\prod_i p(m_i|l_i, \theta)}{\prod_i p(m_i|\theta)} p(\mathbf{l}|\theta)$$

and from the label priors in (3)

$$p(\mathbf{l}|\mathbf{m}, \theta) = \frac{\prod_k p(m_k|l_k, \theta) p(l_k)}{\prod_k \sum_{l \in L} p(m_k|l_k = l, \theta)} \cdot \prod_i \prod_{j \in N_i} p(l_i, l_j).$$

The network of image points can be considered a Markov Random Field (MRF), with the first factor in the equation above representing cliques of one point each and the second involving pairs of points. The corresponding energy functions are given by the negative logarithms of these factors. Note that the second factor represents a smoothing term that is intended to capture the spatial continuity in typical scenes, and penalizes solutions that include discontinuities.

With no penalty if two neighboring points are labeled the same and a constant penalty when labeled differently, the joint probabilities of two neighboring points can be modeled using the Potts model [11], [12]

$$p(l_i, l_j) = \exp^{-V_{i,j}[l_i \neq l_j]}$$

where $[C]$ denotes an indicator function that takes a value 1 if C is true and 0 otherwise. Similar to [13] and [5] we use a pair-wise penalty based on the difference in luminance between image points;

$$V_{i,j} = 50 \, \exp^{-\beta(v_i - v_j)^2},$$

where

$$\beta = (2\langle (v_i - v_j)^2 \rangle)^{-1}.$$

and $\langle \cdot \rangle$ denotes the expectation over an image.

An alternative solution to the problem above could have been based on maximum a posteri (MAP) estimates, instead of the conditional marginals of each label. A local maximum of $p(\mathbf{m}, \mathbf{l}|\theta)$ is searched, while alternating between keeping $\mathbf{l}$ or θ fixed. This is what is done in GrabCut [5]. It is known that if there are only two possible labels per point, an exact MAP solution can be found using graph-cuts [14], and even if the problem becomes NP-hard with more than two labels, there are efficient approximate solutions at hand [6]. While the EM algorithm estimates model parameters by an enumeration over all possible configuration of labels, a MAP based approach would use only one such configuration.

Since we have an interest in the model parameters themselves, in particular those of the foreground, a MAP approach can become problematic. What frequently occurs in figure-ground segmentation are cases where the interpretation of a particular non-textured background region alternates between foreground and background. This leads to model parameters radically change from frame to frame. EM takes such uncertainties into consideration and their respective probabilities are weighted in when parameters are estimated.

B. Initialization

The iterative scheme described above is initialized through a rough segmentation of the image into the three scene parts, using the assumptions mentioned in Section I-B. At this stage only pixels for which disparities exist are considered. Occluded or non-textured areas are ignored until after initialization. From the assumption that the foreground object is in fixation, image points located within a 3D ball are sought and assigned to the foreground model $\mathbf{F}$. The size of the ball is set so that its projective size is equals to half the image height.

Among the remaining image points a flat surface is sought using random sampling with 1000 trials. For each such trial three points are randomly selected and the parameters of a plane $d = \alpha_s x + \beta_s y + \delta_s$ are determined. Since the robot head knows its approximate orientation, planes that are not horizontal enough can immediately be discarded. Among the non-discarded planes, the plane with the highest number of matching image points across the whole image is then selected. A point is considered as matching if its disparity is within 2 pixel values from that of the plane. Points that match the selected plane equation are finally assigned to the surface model $\mathbf{S}$, while the rest are assigned to the background $\mathbf{B}$. Once image points have been assigned, the iterative scheme in section III-A can get started.

IV. Adding dependency over time

In an active vision system image point positions, disparities and colors can be expected to change only slightly from one frame to the next, at least as long as there are no rapid gaze shifts. This consistency over time can be exploited in the estimation of model parameters. In our system we do this by regarding the estimated parameters from the previous frame, θ^t, as measurements when considering the current. Instead of searching the maximum likelihood estimate for $p(\mathbf{m}|\theta)$, we do it for $p(\mathbf{m}, \theta^t|\theta)$.

With labels and point measurements independent of θ^t, the objective function $Q_1(\theta|\theta')$ in (5) is replaced by

$$Q_2(\theta|\theta') = \sum_i \sum_{l_i \in L} w(l_i) \log p(m_i, l_i|\theta) + \log p(\theta^t|\theta) \quad (6)$$

The transition probabilities $p(\theta^t|\theta)$ have three factors, one for each scene part, that is

$$p(\theta^t|\theta) = p(\theta_f^t|\theta_f) p(\theta_b^t|\theta_b) p(\theta_s^t|\theta_s),$$

where

$$\begin{aligned} p(\theta_f^t|\theta_f) &= n(p_f^t; p_f, \Lambda_f)\, n(c_f^t; c_f, \sigma_c^2 I)\, g(\Delta_f^t; \Delta_f, S_f), \\ p(\theta_b^t|\theta_b) &= n(d_b^t; d_b, \Lambda_b)\, n(c_b^t; c_b, \sigma_c^2 I)\, g(\Delta_b^t; \Delta_b, S_b), \\ p(\theta_s^t|\theta_s) &= n(d_s^t; d_s, \Lambda_s)\, n(c_s^t; c_s, \sigma_c^2 I)\, g(\Delta_s^t; \Delta_s, S_s). \end{aligned} \quad (7)$$

Here Λ_f is the expected variance over time for the positional parameters of the foreground, while Λ_b and Λ_s are corresponding variances for the disparity parameters of the background and surface models. The expected variance of the color histogram bins is denoted σ_c^2. The remaining functions $g(\Delta^t; \Delta, S)$ capture the assumed consistency of covariance matrices over time and are defined as follows.

A. Time consistency of covariance matrices

Assume we would like to estimate a covariance matrix Δ given some measurements $\{x_i\}$, and a previously estimated covariance matrix Δ^t at time t. If we assume the underlying distribution changes gradually from one instance in time to the next, we need some way to express its consistency over

Fig. 2. Segmentation results for every fourth frame of a sequence generated by the attention system. Segmentation is re-initiated after each saccade.

Fig. 3. Segmentation results for various scenes. The 9th frame in a sequence is shown in each case.

Fig. 4. Segmentation results with foreground, surface and background models. The images show the 1st, 3rd, 5th and 7th frames of a sequence.

time. In this study we assume the consistency between Δ and Δ^t to be given by

$$g(\Delta^t; \Delta, S) = \left(\frac{1}{2\pi|\Delta|}\right)^{S/2} \exp\left(-\frac{S}{2}\sum_i \lambda_i \mu_i^\top \Delta^{-1} \mu_i\right),$$

where μ_i and λ_i are the eigenvectors and eigenvalues of Δ^t, and S is the strength of the dependency. The equation can be interpreted as $\prod_j p(y_j|\Delta^t)$, where S samples $\{y_j\}$ are drawn from a Gaussian distribution with zero mean and variance Δ^t. If we assume there are no measurements $\{x_i\}$ at time t and Δ only depends on Δ^t, then an estimate Δ^* can be determined from $\arg\max_\Delta g(\Delta^t; \Delta, S)$. We first compute the logarithm of the consistency function

$$\log g(\Delta^t; \Delta, S) = -\frac{S}{2}(\log(2\pi|\Delta|) - \sum_i \lambda_i \mu_i^\top \Delta^{-1} \mu_i,$$

and its derivative with respect to Δ^{-1}

$$\frac{\delta}{\delta \Delta^{-1}} \log g(\Delta^t; \Delta, S) = \frac{S}{2}\left(\Delta - \sum_i \lambda_i \mu_i \mu_i^\top\right).$$

Setting the derivative to 0 results in

$$\Delta^* = \sum_i \lambda_i \mu_i \mu_i^\top = \Delta^t.$$

Hence, if there are no measurements, then Δ will be directly given by Δ^t. In this case the consistency strength factor S has no influence on the result. It will become important, however, when consistency over time is combined with the image point measurements.

V. Experimental evaluation

We performed a series of realistic experiments with objects scattered on a table. A short sequence[2] of foveal views from such an experiment can be seen in Fig. 2. This sequence illustrates how the system is able to rapidly segment an object in its foveated view. For each view the attention system has controlled the cameras and placed an object hypothesis in the center of view.

Using a typical Core 2 processor, the segmentation, including disparity extraction, requires about a second per update

[2]Available as a movie at http://www.csc.kth.se/∼danik/ICRA2010_AVI.avi

with 640×480 pixel images and five iterations per update. For all these experiments we set the expected variances over time of the position parameters (defined in (7)) to Λ_f = diag{1000, 1000, 4}, Λ_b = 25 and Λ_s = diag{0.0001, 0.0004, 1}. We used normalized color histograms with 10×10 bins each, with an expected variance of σ_c^2 = 0.00001 for each bin. The time consistency values for the covariance matrices were set to $S_f = S_b = S_s = N$, i.e. the number of image points. Finally, the prior label probabilities were assumed to be $p(l_f)$ = 20%, $p(l_b)$ = 40% and $p(l_s)$ = 40%. All remaining model parameters were estimated from image and disparity measurement, using the procedure described in Section III.

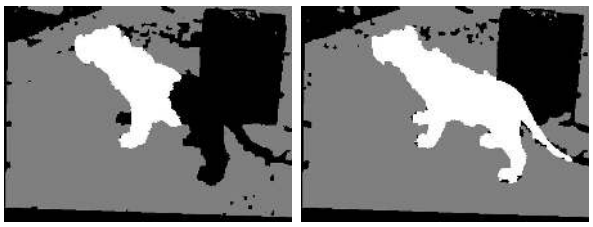

Fig. 5. Point labels of the first and last images of Fig. 3. Pixels labeled as surface points are shown in gray, while white pixels indicate foreground.

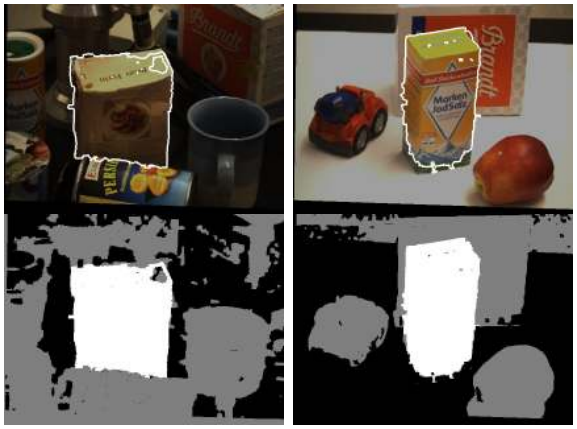

Fig. 6. Segmentation results without an obvious surface plane. The lower images show pixels labeled as surface points in gray.

Fig. 7. Segmentation results without a surface model. The images show the 1st and 7th frames of a sequence.

A. Segmentation results

Using the above mentioned method, segmentation results can be seen in Fig. 3 for a selection of scenes, some more challenging than others. Since the inner part of the cup in the third image lacks reliable disparities and its shade resembles that of a background object, a fragment is still labeled as background after the 9th update. The last image shows an case where the assumption that the foreground object can be described as an ellipsoid fails. The tail of the giraffe will eventually be included, but never the legs. Fig. 4 shows how segmentation evolves over time. With the initial assumption that the foreground can be represented by a ball around zero disparity, it takes a few updates for the model to extend to include the whole cat. Labeling results for the first and last updates can be seen in Fig. 5. As shown by the gray pixels, the table top is captured by the surface model already from the first update.

Fig. 8. Segmentation results without disparity measurements. The images show the 1st and 7th frames of a sequence.

Fig. 9. Segmentation results without color measurements. The images show the 1st and 7th frames of a sequence.

We also consider how the method behaves if no distinct flat surface exists in the scene. Two such examples are shown in Fig. 6. From the gray pixels we observe that the background and surface models have essentially changed order, while the foreground segmentation is unaffected. The surface model finds some non-physical plane across the background objects. The thickness of the plane is gradually extended to include large parts of the scene. The background model is unable to compete, since image points are assumed to be uniformly distributed, even though scene points are typically not.

B. Benefits of multiple cues and models

The method presented here differs from the traditional figure-ground segmentation: it exploits multiple cues for segmentation (colors, positions and disparities) and together with the foreground and background hypotheses it also includes a third, that of a flat surface. Fig. 7-9 show how important these additions are by showing what happens when they are removed. If no flat surface hypothesis were added, one would get results similar to those of Fig. 7. Since the initial ball around the cat includes parts of the table and these parts are located on about the same depth, the foreground segment cannot differentiate between cat and table. The

foreground segment will grow from frame to frame and eventually the whole table will be included.

The behavior could become even worse when disparity measurements are not taken into consideration. Fig. 8 shows an example of that. Without disparities the surface model loses its function and becomes just another background model. Cues that would otherwise have prevented the table top from being included in the foreground become even weaker. Similar behaviors can sometimes be observed in GrabCut, [5], when the initial selected region contains too much of a similarly colored background. Samples from such a false background may result in a distinct peak in the foreground color histogram, which strengthens the hypothesis that these samples do in fact belong to the foreground in next update. With high-quality disparities and a flat surface hypothesis, segmentation often becomes trivial, even without color measurements. However, for regions with unreliable or undefined disparities, color measurements can still be beneficial, as can be seen in Fig. 9.

VI. Discussion and Conclusions

Generating hypotheses about objects in natural scene is a prerequisite for enabling robots to interact with the environment. In this paper, we have presented an active vision system consisting of a two sets of stereo cameras: one for foveal and one for peripheral vision. The system is used for 3D segmentation of visual scenes based on integration of several cues. The main application of the system is to serve as a visual front end and generate object hypotheses for objects not known *a-priori*. The active part of the system is the use of a stereo based fixation process, where objects hypotheses are generated and improved over time. The main contributions of the work is i) that the process of segmentation is considered in 3D thus also providing the input for direct interaction with the environment; ii) the process of temporal segmentation is modeled, showing how the quality of object hypotheses improves over time.

Experimental evaluation demonstrates segmentation of objects in natural scenes with some of the underlying assumptions being violated. Still, the presented method performs well and provides several good object hypotheses. We believe that this is an important result towards equipping robots with the capability of detecting novel objects in the environments and use metric information for direct grasping and manipulation of objects. Our current work explores the use of the system for generation of 3D shape attributes of objects. In addition, we will extend the method for automatic 3D object model generation using several different views of the same object and thus improve the quality of generated grasps.

Appendix

For conciseness we denote the foreground marginal probability of point i by $w_f^i = w(l_i{=}l_f)$. With the color histogram bin corresponding to the same point denoted by b_i, the value of this bin is $c_{f,b_i} = H_f(h_i, s_i)$, where c_f is the foreground color histogram vector. Given the objective function

$$Q_f(\theta) = \sum_i w_f^i \log p(m_i, l_f|\theta_f) + \log p(\theta_f^t|\theta_f)$$

the following update functions of the foreground model can be derived:

$$\frac{\delta Q_f(\theta)}{\delta p_f} = \sum_i w_f^i \Delta_f^{-1}(p_f - p_i) + \Lambda_f^{-1}(p_f - p_f^t) = 0 \Rightarrow$$

$$p_f \leftarrow (\sum_i w_f^i + \Delta_f \Lambda_f^{-1})^{-1}(\sum_i w_f^i p_i + \Delta_f \Lambda_f^{-1} p_f^t)$$

$$\frac{\delta Q_f(\theta)}{\delta \Delta_f^{-1}} = \frac{1}{2}(\sum_i w_f^i \Delta_f + S_f(\Delta_f - \Delta_f^t) - \sum_i w_f^i (p_f - p_i)(p_f - p_i)^\top) = 0 \Rightarrow$$

$$\Delta_f \leftarrow \frac{\sum_i w_f^i (p_f - p_i)(p_f - p_i)^\top + S_f \Delta_f^t}{\sum_i w_f^i + S_f}$$

$$\frac{\delta Q_f(\theta)}{\delta c_{f,j}} = \frac{1}{c_{f,j}} \sum_{b_i = j} w_f^i + \frac{1}{\sigma_c^2}(c_{f,j}^t - c_{f,j}) = 0 \Rightarrow$$

$$c_{f,j} \leftarrow \hat{c}_{f,j} / \sum_i \hat{c}_{f,i}, \text{ where}$$

$$\hat{c}_{f,j} = \frac{1}{2} c_{f,j}^t + \frac{1}{2} \sqrt{{c_{f,j}^t}^2 + 4\,\sigma_c^2 \sum_{b_i = j} w_f^i}$$

Update functions for the background and surface models can be derived similarly.

References

[1] A. Ude, D. Omrcen, and G. Cheng, "Making object learning and recognition an active process," *International Journal of Humanoid Robotics*, vol. 5, no. 2, pp. 267–286, 2008.

[2] D. Melcher, "Persistance of visual memory for scenes," *Nature*, vol. 412, no. 6845, p. 401, 2001.

[3] K. Huebner, M. Björkman, B. Rasolzadeh, M. Schmidt, and D. Kragic, "Integration of Visual and Shape Attributes for Object Action Complexes," in *6th International Conference on Computer Vision Systems*, ser. LNAI, vol. 5008. Springer-Verlag, 2008, pp. 13–22.

[4] B. Rasolzadeh, M. Björkman, K. Huebner, and D. Kragic, "An Active Vision System for Detecting, Fixating and Manipulating Objects in Real World," *International Journal of Robotics Research*, 2009, to appear, available from http://ijr.sagepub.com/pap.dtl.

[5] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut – interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (SIGGRAPH)*, August 2004.

[6] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[7] T. Asfour, K. Regenstein, P. Azad, J. Schröder, and R. Dillmann, "Armar-iii: A humanoid platform for perception-action integration," in *Proceedings of the International Workshop on Human-Centered Robotic Systems (HCRS)*, 2006.

[8] R. Sara, "Finding the largest unambiguous component of stereo matching," in *Proceedings 7th European Conference on Computer Vision (ECCV)*, vol. 2, May 2002, pp. 900–914.

[9] D. Kragic, M. Bjorkman, H. I. Christensen, and J.-O. Eklundh, "Vision for robotic object manipulation is domestic settings," *Robotics and Autonomous Systems*, pp. 85–100, Jun 2005.

[10] Y. Weiss, "Correctness of local probability propagation in graphical models with loops," *Neural Computation*, vol. 12, pp. 1–41, 2000.

[11] R. Potts, "Some generalized order-disorder transformation," *Proceedings of the Cambridge Philosophical Society*, vol. 48, pp. 106–109, 1952.

[12] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 609–628, July 1990.

[13] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," in *Proceedings of International Conference on Computer Vision (ICCV)*, vol. I, 2001, pp. 105–112.

[14] D. Greig, B. Porteous, and A. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 51, no. 2, pp. 271–279, 1989.