Scale-Aware Object Tracking with Convex Shape Constraints on RGB-D Images

Maria Klodt, Jürgen Sturm, and Daniel Cremers

TU München, Germany

Abstract. Convex relaxation techniques have become a popular approach to a variety of image segmentation problems as they allow to compute solutions independent of the initialization. In this paper, we propose a novel technique for the segmentation of RGB-D images using convex function optimization. The function that we propose to minimize considers both the color image and the depth map for finding the optimal segmentation. We extend the objective function by moment constraints, which allow to include prior knowledge on the 3D center, surface area or volume of the object in a principled way. As we show in this paper, the relaxed optimization problem is convex, and thus can be minimized in a globally optimal way leading to high-quality solutions independent of the initialization. We validated our approach experimentally on four different datasets, and show that using both color and depth substantially improves segmentation compared to color or depth only. Further, 3D moment constraints significantly robustify segmentation which proves in particular useful for object tracking.

1 Introduction

Image segmentation and tracking are of central importance in image analysis. Many successful approaches to image segmentation from monochrome or color images have been proposed in the past [1,18]. Unfortunately, in many real-world applications object and background share similar colors such that purely 2D color-based segmentation methods invariably fail – see Figure 1.

With the rise of novel RGB-D cameras like the Microsoft Kinect, inexpensive sensors became available that provide both color images and depth maps synchronized and at high resolution. While depth alone is usually not sufficient to achieve good segmentation results (different objects may share the same depth), it is well-known that the combination of depth and color information outperforms purely color-based segmentation [10] and allows for significant speed-ups of the segmentation process [17]. Moreover, as we will see in this paper, when prior knowledge about the object is available – like for example, its surface area, centroid, or shape covariance matrix – this knowledge can be exploited during object segmentation.

In this paper, we show how a recently introduced convex framework for color image segmentation [13] can be extended to RGB-D image data. In particular, the contributions of this work are three-fold:

- We show that the data term of respective segmentation energies can be extended to incorporate the local depth information. As a consequence, the respective algorithm favors a separation of object and background based on both color and depth

J. Weickert, M. Hein, and B. Schiele (Eds.): GCPR 2013, LNCS 8142, pp. 111–120, 2013.

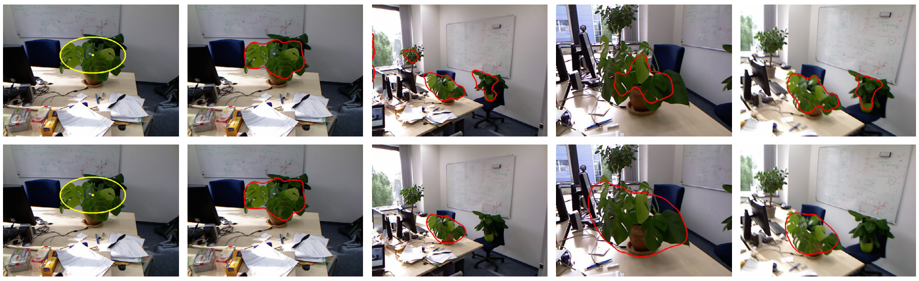


Fig. 1. Tracking with area constraints: RGB area constraints (first row) cannot deal with camera motion, whereas the RGB-D area constraints (second row) are scale-invariant.

information: It will therefore distinguish structures of the same color but different depth. As a consequence we can segment objects that would be difficult to separate by color or depth alone – see Figure 1.

- We show that the moment constraints introduced in [13] can be made invariant using the depth information to the object's distance from the camera. More specifically, the depth maps enable us to impose constraints on the object's *absolute* shape in 3D, whereas purely color based tracking methods can only impose constraints on the object's *projected* shape. These constraints can either be specified manually by user input, or automatically extracted from an initial segmentation for example for object tracking. In several experiments, we demonstrate that our approach allows us to reliably segment and track humans and plants in RGB-D images. Further, we show that respective moment constraints can be generalized to the RGB-D setting thereby assuring that – for example – the surface area in 3D space is preserved. In tracking experiments beyond constraining the object's sideways motion we can thus also constrain the motion of the object along the camera axis.

2 Related Work

Image segmentation is among the most studied problems in image analysis. Popular algorithms to solve the arising shape optimization problems include level set methods [15], graph cuts [11] or convex relaxation [5], with respective extensions to the multi-region case [6,2,20,14,3].

While it was shown that segmentation results can be substantially improved by imposing shape priors [12,7,9], existing approaches have several limitations: Firstly, apart from a few exceptions such as [19,16], computable solutions are only *locally* optimal thus requiring appropriate initializations and leading to often suboptimal solutions. Secondly, many shape priors require an entire training set of familiar shapes [7,8], making them unpractical for generic interactive image segmentation where the user may have a good idea of what he/she wants but will be hard pressed to construct an entire training set of shapes.

As a remedy it was recently proposed [13] to interactively impose constraints on the lower-order moments of the shape in a convex relaxation framework for image segmentation. The aim of this paper is to generalize these concepts to the problem of RGB-D image segmentation.

3 Tracking in RGB-D Sequences with Shape Constraints

We structured the description of our approach into three parts. First, we introduce in Sec. 3.1 how image segmentation can be formulated as a convex relaxation problem. Second, we describe in Sec. 3.2 how moment constraints can be incorporated during image segmentation. Third, we show in Sec. 3.4 how a user can intuitively provide these constraints with a minimum of effort.

3.1 Segmentation with Convex Relaxation

We formulate the problem of image segmentation as a minimization of functionals of the following form:

$$E(u) = \int_\Omega f(x)\, u(x)\, dx \; + \int_\Omega |Du(x)|, \tag{1}$$

Here, $u \in BV(\mathbb{R}^d; \{0,1\})$ is an indicator function on the space of binary functions of bounded variation, where $u = 1$ and $u = 0$ denote the interior and exterior of a hyper surface in $\mathbb{R}^d$, i.e. a set of closed boundaries in the case of 2D image segmentation or a set of closed surfaces in the case of 3D segmentation.

The second term in (1) is the total variation. Here Du denotes the distributional derivative which for differentiable functions u boils down to $Du(x) = \nabla u(x) dx$. By relaxing the binary constraint and allowing the function u to take on values in the interval between 0 and 1, the optimization problem becomes that of minimizing the convex functional (1) over the convex set $BV(\mathbb{R}^d; [0,1])$.

Functionals of this form can be globally optimized in a spatially continuous setting by means of convex relaxation and thresholding. The thresholding theorem [4] assures that thresholding the solution u^* of the relaxed problem preserves global optimality for the original binary labeling problem. We can therefore compute global minimizers for functional (1) in a spatially continuous setting as follows: Compute a global minimizer u^* of (1) on the convex set $BV(\mathbb{R}^d; [0,1])$ and threshold the minimizer u^* at any value $\theta \in (0,1)$.

With additional depth information from RGB-D images, the boundary length can be measured in absolute values instead of the image domain. Functional (1) can be generalized to

$$E(u) = \int_\Omega f(x)\, u(x)\, dx \; + \int_\Omega d(x) |Du(x)|, \tag{2}$$

with depth values $d : \Omega \to \mathbb{R}$. This formulation compensates the fact that objects that are far away to the camera appear smaller in the image due to perspective projection. Weighting with $d(x)$ allows regularization on the absolute size of the boundary – in contrast to assuming a uniform pixel size as in (1).

3.2 Moment Constraints for RGB-D Images

In the following, we will successively constrain the moments of the segmentation with depth information and show how all of these constraints give rise to nested convex sets. We will denote by $\mathcal{B} = BV(\Omega; [0,1])$ the convex hull of the set of binary indicator functions $u \in BV(\Omega; \{0,1\})$ of bounded variation on the domain $\Omega \subset \mathrm{I\!R}^d$.

Area Constraint. The 0-th order moment corresponds to the area of the shape u and can be computed by

$$\mathrm{Area}(u) := \int_{\Omega} d^2(x) u(x) \,\mathrm{dx}, \tag{3}$$

where $d(x)$ gives the depth of pixel x. Here, we assume that $d(x) = KD(x)$, with K being the focal length of the camera and $D(x)$ being the depth of the pixel measured in meters. Note that $d^2(x)$ corresponds to the size of a back-projected pixel in 3D space, and thus the integral measures the absolute surface area (scaled by K^2) instead of the projected area in the image. This is in contrast to [13], where all pixels are treated equally.

We can impose that the absolute area of the shape u to be bounded by constants $c_1 \leq c_2$ by constraining u to lie in the set:

$$\mathcal{C}_0 = \left\{ u \in \mathcal{B} \;\middle|\; c_1 \leq \mathrm{Area}(u) \leq c_2 \right\}. \tag{4}$$

The set $\mathcal{C}_0$ is linearly dependent on u and therefore convex for any constants $c_2 \geq c_1 \geq 0$.

In practice, we can either impose an exact area by setting $c_1 = c_2$, or we can impose upper and lower bounds on the area. Alternatively, we can impose a soft area constraint by enhancing the functional (1) as follows:

$$E_{total}(u) = E(u) + \lambda \left(\int d^2 u \,\mathrm{dx} - c \right)^2, \tag{5}$$

which imposes a soft constraint with a weight $\lambda > 0$ favoring the area of the estimated shape to be near $c \geq 0$. Note that the functional (5) is also convex.

Centroid Constraint. The 1-st order moment corresponds to the center of gravity (or *centroid*) of the shape. It can be computed by integrating over all 3D positions of the shape, i.e.,

$$\mu(u) := \begin{pmatrix} \overline{x} \\ \overline{d} \end{pmatrix} = \frac{\int_{\Omega} \binom{x}{d} u \,\mathrm{dx}}{\int_{\Omega} d^2 u \,\mathrm{dx}}, \tag{6}$$

where $\overline{x} \in \mathbb{R}^2$ is the centroid in pixel coordinates and $\overline{d} \in \mathbb{R}$ is the centroid in depth. Together, $\mu \in \mathbb{R}^3$ corresponds to the centroid of the shape in 3D.

We can now impose bounds on the centroid for the object we want to segment by constraining the solution u to the set $\mathcal{C}_1$:

$$\mathcal{C}_1 = \left\{ u \in \mathcal{B} \;\middle|\; \mu_1 \leq \mu(u) \leq \mu_2 \right\}, \tag{7}$$

where all inequalities are to be taken point-wise and $\mu_1, \mu_2 \in \mathrm{I\!R}^3$. This imposes the centroid to lie between the two constants $\mu_1 \leq \mu_2$. In particular, for $\mu_1 = \mu_2$, the centroid is fixed.

Proposition 1. *For any constants $\mu_2 \geq \mu_1 \geq 0$, the set $\mathcal{C}_1$ is convex. (The proof is analogous to proof 2 in [13].)*

Alternatively, we can impose the centroid as a soft constraint by minimizing the energy:

$$E_{total}(u) = E(u) + \lambda \left| \int_{\Omega} \left(\mu d^2 - \left(\begin{smallmatrix} x \\ d \end{smallmatrix} \right) \right) u \, \mathrm{dx} \right|^2 ,$$

which is also convex in u.

Covariance Constraint. The proposed concept can be generalized to moments of second order. In the following, we focus on central moments (i.e. moments with respect to a specified centroid μ). The 3D covariance of a shape u is given by

$$\mathrm{Cov}(u) := \frac{\int_{\Omega} \left(\left(\begin{smallmatrix} x \\ d \end{smallmatrix} \right) - \mu \right) \left(\left(\begin{smallmatrix} x \\ d \end{smallmatrix} \right) - \mu \right)^{\top} u \, \mathrm{dx}}{\int_{\Omega} d^2 u \, \mathrm{dx}}. \tag{8}$$

The covariance structure can be considered by the following convex set:

$$\mathcal{C}_2 = \left\{ u \in \mathcal{B} \;\middle|\; A_1 \leq \mathrm{Cov}(u) \leq A_2 \right\} \tag{9}$$

where the inequality constraint should be taken element wise. Here $\mu \in \mathbb{R}^3$ denotes the centroid and $A_1, A_2 \in \mathbb{R}^{3\times3}$ denote symmetric matrices such that $A_1 \leq A_2$ element wise. This constraint is particularly meaningful if one additionally constrains the centroid to be μ, i.e. considers the intersection of the set (9) with a set of the form (7).

Optimization with Moment Constraints. Shape optimization and image segmentation with respective moment constraints can now be done by minimizing convex energies under respective convex constraints. The optimization was implemented using the projection approach as described in [13].

3.3 Tracking with 3D Constraints

The 3D moments of a shape can be used for tracking objects in a sequence of images. Given the moments of the shape in the first frame, constraints can be imposed on segmentations in all subsequent frames. Here, the moments of a shape are computed directly in the 3D space, not in the projection to the image plane. This makes the method independent of the projected size of the object in the image. Without the need of defining a window in which subsequent shapes should be found, the proposed method simply applies the moment constraints of the current frame to the subsequent. We allow the centroid to change inside a small range to handle motion of the camera and/or the object. The area and covariance are supposed to stay constant in the 3D space over all time frames.

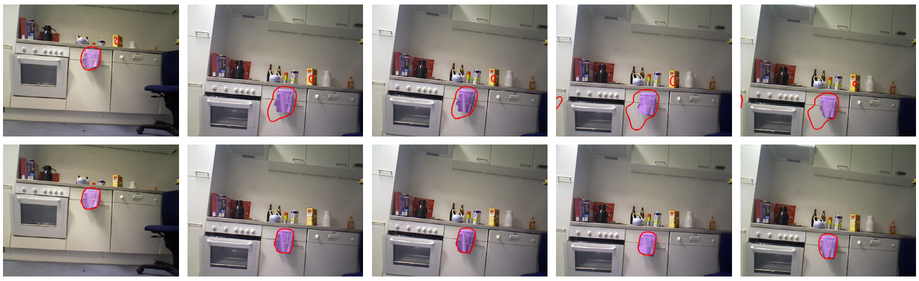


Fig. 2. Comparison of tracking an object with and without area constraint. **Top row:** Color-only tracking. **Bottom row:** RGB-D tracking: The surface area is constrained on the absolute dimension via additional information from the depth images.

3.4 Segmentation Priors from User Input

The data term used throughout our experiments has the following form:

$$f(x) = \log \frac{p_{\mathrm{bg}}(I(x))}{p_{\mathrm{obj}}(I(x))}. \tag{10}$$

Here, $I : \Omega \to \mathbb{R}^n$ refers to an image with n channels. For example, $n = 1$ for depth or gray-scale images, $n = 3$ for color images and $n = 4$ for RGB-D images. The data priors p_{obj} and p_{bg} assign probabilities to each pixel belonging to the object or the background, respectively, and satisfy $p_{\mathrm{obj}} + p_{\mathrm{bg}} = 1$. We compute them from histograms for foreground and background. The moment constraints that we consider in our experiments include the centroid, area and covariance of the shape.

Both the data prior as well as the moment constraints can be specified by the user. We found that an intuitive interface is to ask the user to mark the object of interest with an ellipse (see Fig. 1). From the pixels within and outside the ellipse, we train the n-dimensional color/depth/RGB-D histograms corresponding to the probability distributions p_{obj} and p_{bg}, respectively. Further, we extract the surface area, 3D centroid and 3D covariance matrix that we use as moment constraints during segmentation from the projection of the ellipse into 3D space, with the information of the depth image.

4 Experimental Results

In this section we present an evaluation of our approach for RGB-D image segmentation with moment constraints. The goal of our experiments was to verify that (1) segmentation on RGB-D data is more reliable than segmentation of color or depth images alone, and that (2) object tracking with 3D moment constraints is more robust than 2D moment constraints.

All images and videos shown in this paper were captured using the Microsoft Kinect sensor. Run-times on a GPU implementation are less than 1 second per image, making the method useful for interactive applications.

4.1 Tracking with Moment Constraints

Figure 1 shows results on moment-consistent tracking in 2D and 3D with large camera motion. In the top row we see the results for color-only tracking: The area constraint is imposed on the projected shape of the object. The method cannot cope with increasing and decreasing appearance in the image domain, although the absolute size of the object stays the same. The bottom row shows RGB-D tracking: The area constraint is imposed on the absolute dimension via additional information from the depth images. The method enables area-consistent tracking with arbitrary camera motion. In the Figure, we took images of a plant in an office scene with a hand-held Kinect sensor from different view points. Of course, the basic properties of the 3D shape – and thus the surface area and covariance structure – of the selected object remains the same during the sequence. However, the projection of the object's shape in 2D changes its size due to object and/or camera motion. As a result, simple 2D moment tracking fails, as it tries to keep the area in image space constant. In contrast, 3D moment constraints are scale-invariant and are thus more robust against camera and/or object motion. From these examples, we conclude that in the case of arbitrary camera motion 3D moment constraints are better suited for object tracking than 2D moment constraints.

The image sequence in Fig. 2 was captured by a flying quadrocopter with a Kinect camera mounted on top of it. The towel's shape and color distribution vary over time due to camera motion and wind caused by the quadrocopter's rotors. The figure shows that color-only segmentation (first row) is not sufficient to track the object, whereas additional information from the depth images allow 3D moment constraints to track the exact surface area (second row).

4.2 Segmentation with Color, Depth, and RGB-D

We tested our segmentation method with moment constraints in several scenes to demonstrate that RGB-D segmentation can outperform segmentation based on color or depth alone. To demonstrate this, we segmented different objects in the color, depth, and the (combined) RGB-D image.

Our first example is shown in Fig. 3 where we aimed at segmenting individual persons from the crowd. We found that neither color nor depth information are sufficient to uniquely separate a single person in the image, see Fig. 3(b+c). In more detail, the person in the first row is hard to segment in the color image because of the blue jeans in front of the blue door. The person in the second row wears a black shirt and is partially occluded by the wardrobe, and the person in the third row overlaps with the person in the background, having similar histograms which makes the segmentation task hard. Depth segmentation alone has shortcomings in other regions of the image. We found that there are often pixels in an image with similar depth values as the foreground object – with the exception of the person sitting on the chair, where no other pixels had the same depth values. In the first two rows of Fig. 3, the segmentation problems are resolved when RGB and depth information is jointly considered. To conclude, all persons could be separated well in the RGB-D case.

Another interesting example is depicted in Fig. 4 where we found that even the absence of information in the depth image can be exploited to successfully segment an

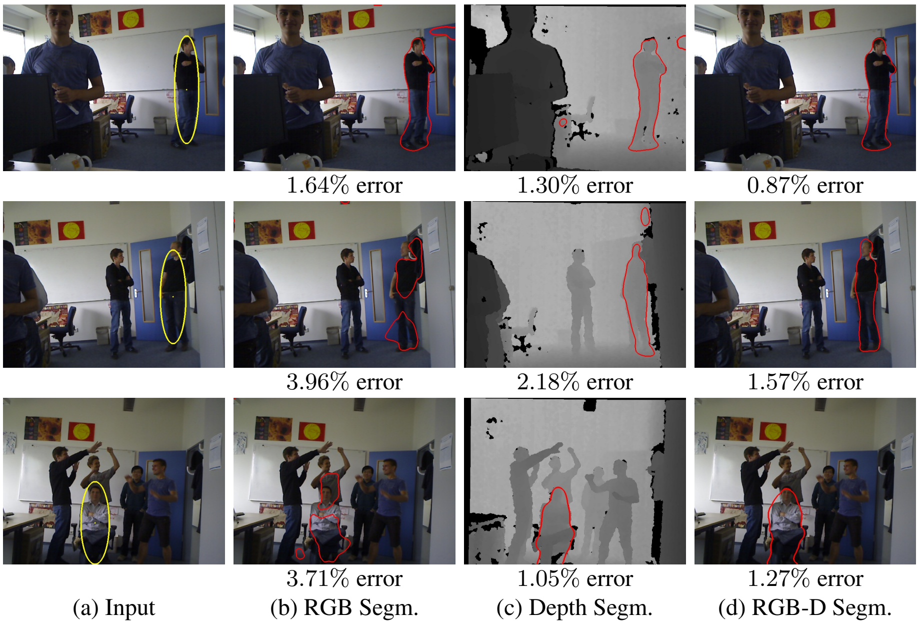


Fig. 3. Segmentation of images with ambiguous color and depth information. Moment constraint parameters are derived from user input (a). Purely color (b) and depth (c) images alone do not provide enough information to uniquely segment one person. The combination (d) allows for segmentation of one single person in all three examples. Segmentation errors can be reduced by combining depth and color information.

image. Here, we consider a water glass located on a table. In the color image, the glass is difficult to see because of its transparency. Moreover, the depth of the glass pixels cannot be estimated due to the material's reflective property. By considering both the color and the depth image, we found that the glass is well separable.

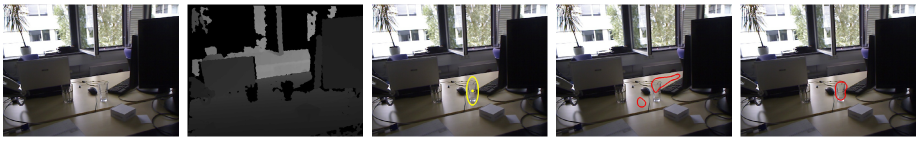
4.3 Quantitative Analysis

For a quantitative analysis of the presented method, we measured the amount of pixels that differ from a manually segmented ground truth for segmentation with and without constraints, as well as segmentations using color, depth, and their combination. Segmentation errors were computed for the images in Fig. 3.

Table 1 shows average segmentation errors compared to the ground truth. Here we also compared to segmentations without moment constraints, where segmentations were computed using only the color information of the histograms inside and outside the ellipse drawn by the user. The table clearly shows that the amount of misclassified pixels can be reduced by combining depth and color information for segmentation with moment constraints. Interestingly, segmentation with depth only yields significantly better results than color only.

Table 1. Average segmentation errors with and without moment constraints, compared to ground truth. The combination of color and depth leads to better results, even more improvement is achieved by additionally constraining the moments of the segmentation.

		Average Segmentation Error
Without Constraints:	Color only	29.25%
	Depth only	16.99%
	RGB-D	17.93%
With Constraints:	Color only	3.10%
	Depth only	1.51%
	RGB-D	1.24%



(a) Input Color Image (b) Input Depth Image (c) User Input (d) Color-only Segmentation (e) RGB-D Segmentation

Fig. 4. Segmentation of reflective material. (a+b) Input image and (c) user input. (d) When only the color image is considered, the glass is indistinguishable from the background due to its transparency. (e) When the depth image is taken into account, the glass becomes separable.

5 Conclusion

We introduced a convex framework for interactive RGB-D image segmentation and tracking. Building up on state-of-the-art approaches for color segmentation, we showed that depth information can be integrated in the data terms for image segmentation so as to favor segmentations of coherent depth. In particular objects of similar color but different depth can be discriminated. Moreover, we show that the availability of depth allow to impose constraints on the *absolute* shape rather than the *projected* shape. And lastly we show that one can impose moment constraints in 3D space – thereby we exploit the fact that the 3D motion of a tracked object is constrained over time. Our studies demonstrate that combining color and depth drastically enhances the possibilities of variational segmentation methods. In particular, it allows to generalize respective constraints from the image plane to the physical 3D space. Experiments show that with a minimal amount of user input we can obtain fast interactive segmentations of good quality in a variety of challenging real-world scenarios.

References

1. Boykov, Y., Jolly, M.-P.: Interactive organ segmentation using graph cuts. In: Delp, S.L., DiGoia, A.M., Jaramaz, B. (eds.) MICCAI 2000. LNCS, vol. 1935, pp. 276–286. Springer, Heidelberg (2000)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. on Patt. Anal. and Mach. Intell. 23(11), 1222–1239 (2001)

3. Chambolle, A., Cremers, D., Pock, T.: A convex approach for computing minimal partitions. Communications on Pure and Applied Mathematics (2008)
4. Chan, T., Esedoglu, S., Nikolova, M.: Algorithms for finding global minimizers of denoising and segmentation models. SIAM Journal on Applied Mathematics 66(5), 1632–1648 (2006)
5. Chan, T., Esedoḡlu, S., Nikolova, M.: Algorithms for finding global minimizers of image segmentation and denoising models. SIAM Journal on Applied Mathematics 66(5), 1632–1648 (2006)
6. Chan, T., Vese, L.: A level set algorithm for minimizing the Mumford–Shah functional in image processing. In: IEEE Workshop on Variational and Level Set Methods, Vancouver, CA, pp. 161–168 (2001)
7. Cremers, D., Osher, S.J., Soatto, S.: Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. Int. J. of Computer Vision 69(3), 335–351 (2006)
8. Etyngier, P., Segonne, F., Keriven, R.: Shape priors using manifold learning techniques. In: IEEE Int. Conf. on Computer Vision. Rio de Janeiro (October 2007)
9. Foulonneau, A., Charbonnier, P., Heitz, F.: Affine-invariant geometric shape priors for region-based active contours. IEEE Trans. on Patt. Anal. and Mach. Intell. 28(8), 1352–1357 (2006)
10. Gordon, G., Darrell, T., Harville, M., Woodfill, J.: Background estimation and removal based on range and color. In: Int. Conf. on Computer Vision and Pattern Recognition, pp. 459–464 (1999)
11. Greig, D.M., Porteous, B.T., Seheult, A.H.: Exact maximum *a posteriori* estimation for binary images. J. Roy. Statist. Soc., Ser. B. 51(2), 271–279 (1989)
12. Grenander, U., Chow, Y., Keenan, D.M.: Hands: A Pattern Theoretic Study of Biological Shapes. Springer, New York (1991)
13. Klodt, M., Cremers, D.: A convex framework for image segmentation with moment constraints. In: IEEE Int. Conf. on Computer Vision (2011)
14. Lellmann, J., Kappes, J., Yuan, J., Becker, F., Schnörr, C.: Convex multi-class image labeling by simplex-constrained total variation. In: Tai, X.-C., Mørken, K., Lysaker, M., Lie, K.-A. (eds.) SSVM 2009. LNCS, vol. 5567, pp. 150–162. Springer, Heidelberg (2009)
15. Osher, S.J., Sethian, J.A.: Fronts propagation with curvature dependent speed: Algorithms based on Hamilton–Jacobi formulations. J. of Comp. Phys. 79, 12–49 (1988)
16. Schoenemann, T., Cremers, D.: A combinatorial solution for model-based image segmentation and real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2009)
17. Taylor, C., Cowley, A.: Fast scene analysis using image and range data. In: Proc. of the Intl. Conf. on Robotics and Automation, ICRA (2011)
18. Unger, M., Pock, T., Cremers, D., Bischof, H.: TVSeg - interactive total variation based image segmentation. In: British Machine Vision Conference (BMVC), Leeds, UK (September 2008)
19. Veksler, O.: Star shape prior for graph-cut image segmentation. In: Europ. Conf. on Computer Vision., pp. 454–467 (2008)
20. Zach, C., Gallup, D., Frahm, J.M., Niethammer, M.: Fast global labeling for real-time stereo using multiple plane sweeps. In: Vision, Modeling and Visualization Workshop VMV 2008(October 2008)