International Journal of Computer Vision

DOI: 10.1007/s11263-006-9965-3

Three-Dimensional Shape Knowledge for Joint Image Segmentation and Pose Tracking

BODO ROSENHAHN
Max-Planck-Center Saarbrücken, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany
rosenhahn@mpi-sb.mpg.de

THOMAS BROX
Computer Vision and Pattern Recognition Group, University of Bonn, Römerstr. 164, 53117 Bonn, Germany
brox@cs.uni-bonn.de

JOACHIM WEICKERT
Mathematical Image Analysis Group, Saarland University, Building 27, 66041 Saarbrücken, Germany
weickert@mia.uni-saarland.de

Received August 10, 2005; Revised July 18, 2006; Accepted July 18, 2006

First online version published in September, 2006

Abstract. In this article we present the integration of 3-D shape knowledge into a variational model for level set based image segmentation and contour based 3-D pose tracking. Given the surface model of an object that is visible in the image of one or multiple cameras calibrated to the same world coordinate system, the object contour extracted by the segmentation method is applied to estimate the 3-D pose parameters of the object. Vice-versa, the surface model projected to the image plane helps in a top-down manner to improve the extraction of the contour. While common alternative segmentation approaches, which integrate 2-D shape knowledge, face the problem that an object can look very differently from various viewpoints, a 3-D free form model ensures that for each view the model can fit the data in the image very well. Moreover, one additionally solves the problem of determining the object's pose in 3-D space. The performance is demonstrated by numerous experiments with a monocular and a stereo camera system.

Keywords: pose estimation, segmentation, variational methods, shape priors

1. Introduction

Image segmentation and pose estimation are two principal problems in computer vision. Segmentation determines the location and shape of objects in the image plane, thereby performing a significant abstraction step from the raw pixel data to object regions. Pose estimation, on the other hand, determines the pose of objects in 3-D space given some sensor data, e.g., images from one or multiple cameras.

Both tasks have been intensively investigated and a lot of progress has been made in recent years, as shown by many seminal papers and textbooks on segmentation (Geman and Geman, 1984; Blake and Zisserman, 1987; Kass et al., 1988; Mumford and Shah, 1989; Zhu and Yuille, 1996; Shi and Malik, 2000; Leventon et al.,

2000; Chan and Vese, 2001; Paragios and Deriche, 2002; Cremers et al., 2002) and pose estimation (Lowe, 1980, 1987; Grimson, 1990; Li, 1995; Araújo et al., 1998; Ma et al., 2003; Murray et al., 1994; Gallier, 2001; Faugeras, 1993; Blaschke, 1960; Vacchetti et al., 2004; Lepetit and Fua, 2005). See also the theses (Goddard, 1997; Rosenhahn, 2003) on pose estimation. Nevertheless, both segmentation and pose estimation still face severe difficulties, in particular in natural scenes. The reason for such difficulties is in most cases a violation of model assumptions. In image segmentation, for instance, the model usually assumes homogeneous (e.g. constant (Chan and Vese, 2001) or smooth (Mumford and Shah, 1989)) object regions. Due to noise, texture, shading, or occlusion, however, this model is often not appropriate to delineate object regions. A successful remedy is the statistical modeling of regions and the supplement of additional information, such as texture and motion, which greatly extends the number of situations where image segmentation can succeed (Zhu and Yuille, 1996; Malik et al., 2001; Paragios and Deriche, 2002; Rousson et al., 2003). Another strategy is to impose additional constraints like the restriction to a certain object shape. This introduction of shape priors into segmentation models has been proposed in Leventon et al. (2000) and has been extended and modified in a large number of successive works (Cremers et al., 2001; Rousson and Paragios, 2002; Cremers et al., 2002; Riklin-Raviv et al., 2004; Cremers et al., 2004).

Also pose estimation, or the related task of pose tracking, work very reliably and often even in real-time when applied to controlled situations. In this paper, we focus on 2D-3D pose tracking of a rigid body, i.e., we seek a 3-D rigid motion that fits the model to some 2-D image data.[1] The difficult part in this task is to reliably match some 2-D features to their 3-D counterparts on the model. Numerous different types of features have been used in the past, e.g., lines (Beveridge, 1993), viewpoint dependent point features, such as vertices, *t*-junctions, cusps, three-tangent junctions, edge inflections, etc. (Kriegman et al., 1992), multi-part curve segments (Zerroug and Nevatia, 1996), and complete contours (Drummond and Cipolla, 2000; Rosenhahn and Sommer, 2004).

In this paper, we follow the concept of matching the modeled object surface to the object contour(s) extracted from one or multiple images by level set segmentation. While segmentation and 2D-3D pose estimation have so far been investigated more or less independently from each other, the main contribution of the present paper, is to formulate a joint energy functional and a corresponding optimization scheme that solves both tasks simultaneously.

From the segmentation perspective, our approach extends the above mentioned segmentation methods that integrate 2-D prior knowledge. Compared to these existing methods, the 3-D model in our approach ensures a good description of the model contour from arbitrary viewpoints by directly taking the 3-D nature of most real objects into account.

Moreover, from the perspective of pose tracking, our method integrates the feature extraction step into the pose estimation process, i.e., there is a feedback of the pose result that helps to improve the features used for matching, in our case the object contour.

While the coupling of segmentation and pose estimation (registration) has already been investigated in case of 2-D segmentation and 2-D shape models as well as in case of volumetric segmentation and 3-D shape models, e.g. (Yezzi et al., 2001; Rousson et al., 2004), to the best of our knowledge, we present here the first approach that integrates segmentation in the image plane and pose estimation in 3-D. Compared to the previous cases, this comes along with additional difficulties, as the approach implies a projection as well as an inverse projection to match the 3-D shape to 2-D image data and vice-versa.[2] Related works on 2D-3D pose estimation, on the other hand, either work on pre-computed contours that are independent from the pose result (Drummond and Cipolla, 2000; Rosenhahn, 2003), or rely on simpler features such as edge maps and line segments (Haag and Nagel, 1999; Marchand, 2001).

This paper comprises and extends an earlier work presented on a conference (Brox et al., 2005). In comparison to this introduction of the basic idea, the present paper contains a much more detailed description of the approach, suggests a confidence measure in the coupling of segmentation and pose estimation, and demonstrates the generality of the method by means of additional experiments that rule out many alternative techniques for solving the task.

Paper Organization. We start in the next section with a review of the level set based image segmentation model that provides the basis for our variational approach. The section further includes a region model based on local statistics that aims at the handling of inhomogeneous objects and backgrounds. Section 3 extends the variational segmentation model by an additional term that integrates the 3-D surface and its

pose parameters. The section describes step by step the optimization procedure that yields the object contour and the sought pose parameters. Experiments in Section 4 demonstrate the performance of the proposed technique and illustrate the conceptual difference to other methods. The paper is concluded by a short summary in Section 5.

2. Image Segmentation

2.1. Level Set Formulation

Our method is based on variational image segmentation with level sets (Dervieux and Thomasset, 1979; Osher and Sethian, 1988; Caselles et al., 1993; Malladi et al., 1995; Paragios and Deriche, 1999; Chan and Vese, 1999, 2001; Tsai et al., 2001; Paragios and Deriche, 2002). Level set formulations of the image segmentation problem have several advantages. One is the convenient embedding of a 1-D curve into a 2-D, image-like structure. This allows for a convenient and sound interaction between constraints that are imposed on the contour itself and constraints that act on the regions separated by the contour. Moreover, the level set representation yields the inherent capability to model topological changes. This can be an important issue, for instance, when the object is partially occluded by another object and is hence split into two parts.

In the prominent case of a two-phase segmentation, a level set function $\Phi \in \Omega \mapsto \mathbb{R}$ splits the image domain Ω into two regions Ω_1 and Ω_2, with $\Phi(x) > 0$ if $x \in \Omega_1$ and $\Phi(x) < 0$ if $x \in \Omega_2$. The zero-level line thus marks the boundary between both regions, i.e., it represents the object contour that is sought to be extracted.

Most works on level set segmentation focus on this special case with two regions. It automatically rules out overlapping or vacuum regions and therefore eases implementation. Since the present paper is concerned with the extraction of exactly one known object and its pose, we will also restrict to two regions: the object and the background. However, the reader can find numerous works that extend the level set framework to multiple regions in a more or less simple and efficient manner (Zhao et al., 1996; Vese and Chan, 2002; Mansouri et al., 2004; Brox and Weickert, 2005).

As an optimality criterion for contour extraction, three constraints are imposed:

1. the data within each region should be similar
2. the data between regions should be dissimilar
3. the contour dividing the regions should be minimal

These model assumptions can be expressed by the following energy functional (Zhu and Yuille, 1996; Chan and Vese, 2001; Paragios and Deriche, 2002):

$$E(\Phi) = -\int_\Omega \big(H(\Phi)\log p_1 + (1-H(\Phi))\log p_2\big)\,dx + \nu \int_\Omega |\nabla H(\Phi)|\,dx \tag{2.1}$$

where $\nu > 0$ is a weighting parameter between the third and the two other constraints, and $H(s)$ is a regularized Heaviside function with $\lim_{s\to-\infty} H(s) = 0$, $\lim_{s\to\infty} H(s) = 1$, and $H(0) = 0.5$ (e.g. the error function). It indicates to which region a pixel belongs. Minimizing the first two terms maximizes the total a-posteriori probability given the probability densities p_1 and p_2 of Ω_1 and Ω_2, i.e., pixels are assigned to the most probable region according to the Bayes rule. The third term minimizes the length of the contour.

Energy minimization can be performed according to the gradient descent equation

$$\partial_t \Phi = H'(\Phi)\left(\log\frac{p_1}{p_2} + \nu\,\mathrm{div}\left(\frac{\nabla\Phi}{|\nabla\Phi|}\right)\right) \tag{2.2}$$

where $H'(s)$ is the derivative of $H(s)$ with respect to its argument. Applying this evolution equation to some initialization Φ^0, and given the probability densities p_i, the contour converges to the next local minimum for the numerical evolution parameter $t \to \infty$.

2.2. Region Statistics

A very important factor for the quality of the contour extraction process is the way how the probability densities p_1 and p_2 are modeled. This model decides on what is considered as similar or dissimilar. There are several possibilities of image cues to use for the density model, for instance, gray value, color, texture (Sifakis et al., 2002; Paragios and Deriche, 2002; Rousson et al., 2003), or motion (Cremers and Soatto, 2005). Moreover, there are various possibilities how to model the probability densities given these image cues, e.g., a Gaussian density with fixed standard deviation (Chan and Vese, 2001), a full Gaussian density (Zhu and Yuille, 1996; Rousson and Deriche, 2002), a generalized Laplacian (Heiler and Schnörr, 2005), or nonparametric Parzen estimates (Kim et al., 2002; Rousson et al., 2003; Kadir and Brady, 2003; Kim et al., 2005).

For the segmentation here, we use the texture feature space proposed in Brox and Weickert (2006), which yields $M = 5$ feature channels I_j for gray scale images, and $M = 7$ channels if color is available. The color channels are considered in the CIELAB color space. The texture features described in Brox and Weickert (2006) contain basically the same information as the frequently used responses of Gabor filters, yet the representation of this information is less redundant, so 4 feature channels substitute 12-64 Gabor responses.

The probability densities of the M feature channels are assumed to be independent, thus the total probability density comes down to

$$p_i = \prod_{j=1}^{M} p_{ij}(I_j) \quad i = 1, 2. \tag{2.3}$$

Though assuming independence of the probability densities is only an approximation of the true densities, it keeps the density model tractable. This has to be seen particularly with regard to the fact that the densities have to be estimated by means of a limited amount of image data given.

Estimating both the probability densities p_{ij} and the region contour works according to the *expectation-maximization principle* (Dempster et al., 1997; McLachlan and Krishnan, 1997). Having the level set function initialized with some partitioning Φ^0, the probability densities in these regions can be approximated. With the probability densities, on the other hand, one can compute an update on the contour according to (2.2), leading to a further update of the probability densities, and so on. Since the process converges to a local minimum, the initialization matters. In order to attenuate the dependency on the initialization, one can apply a continuation method in a coarse-to-fine manner (Blake and Zisserman, 1987).

It has been shown in Rousson and Deriche (2002) that in case of Gaussian densities, the expectation-maximization procedure is equivalent to a gradient descent in both the contour Φ and the densities p_i. For other density models, the gradient descent contains additional terms that are neglected by the expectation-maximization procedure. In Heiler and Schnörr (2005) it has been shown empirically for the Laplacian density model that the influence of these higher order terms is very small. In the appendix we show that for the local region statistics described in the next section, the influence of the additional term is restricted, too.

2.3. Local Region Statistics

While most image segmentation methods assume a global model for the probability density of each region, i.e., the probability density function only depends on the region but does not change within one region, it has been suggested in Kadir and Brady (2003) to consider density functions that may vary within regions. This can be advantageous in scenes with complex objects, shadows, and highlights, where differences between object and background are often only locally visible. A global statistical model looses this local information and can thus loose the capability to separate the regions.

We model the regions by the following local Gaussian probability density that varies with the position x in the image:

$$p_{ij}(s, x) = \frac{1}{\sqrt{2\pi}\sigma_{ij}(x)} \exp\left(\frac{(s - \mu_{ij}(x))^2}{2\sigma_{ij}(x)^2}\right). \tag{2.4}$$

The parameters $\mu_{ij}(x)$ and $\sigma_{ij}(x)$ are computed in a local Gaussian neighborhood K_ρ around x by:

$$\begin{aligned} \mu_{ij}(x) &= \frac{\int_{\Omega_i} K_\rho(\zeta - x) I_j(\zeta)\, d\zeta}{\int_{\Omega_i} K_\rho(\zeta - x)\, d\zeta} \\ \sigma_{ij}(x) &= \frac{\int_{\Omega_i} K_\rho(\zeta - x)(I_j(\zeta) - \mu_{ij}(x))^2\, d\zeta}{\int_{\Omega_i} K_\rho(\zeta - x)\, d\zeta} \end{aligned} \tag{2.5}$$

where ρ denotes the standard deviation of the Gaussian window. In order to obtain reliable estimates for the parameters $\mu_{ij}(x)$ and $\sigma_{ij}(x)$, it is recommended to choose $\rho \geq 6$.

Two major drawbacks come along with these local statistics. Firstly, they demand a considerably larger amount of computation time than global estimates (about one order of magnitude). Secondly, they induce more local minima in the energy functional. It is mainly due to the latter reason that we choose Gaussian densities and not a non-parametric model as in Kadir and Brady (2003). With global statistics, non-parametric region models are often advantageous, since they can better adapt to multimodal densities. Within a local window, however, such multimodal situations are rare. Consequently, one can expect the less complex model to be more reliable.

3. Integration of a 3-D Surface Model and Estimation of its Pose

Segmentation approaches from the type described in the last section can perform very well, if all model assumptions are satisfied. In many images, however, the model will prefer to separate other regions than the object regions. Additional constraints have to be introduced in order to restrict the sought contour to stay close to a certain shape. This concept is already well-established in the segmentation of 2-D images using 2-D shape knowledge, (Leventon et al., 2000; Cremers et al., 2001; Rousson and Paragios, 2002; Cremers et al., 2002) or the segmentation of 3-D volume data and 3-D shape knowledge (Yezzi et al., 2001; Rousson et al., 2004).

In this section, we will implement this concept when 2-D images and a 3-D shape model are given. In comparison to the above-mentioned situations, this comes along with additional difficulties. The integration of prior shape knowledge always includes an estimation of the shape's pose in the image. This is necessary since usually one wants the method to be invariant to a certain class of pose transformations, e.g. rigid transformations. In contrast to matching 2-D images to a 2-D shape or 3-D volumes to a 3-D shape, the case investigated here implies a projection as well as an inverse projection to match the 3-D shape to 2-D image data and vice-versa. In the following, we introduce a way to integrate a joint evolution of the contour and the pose in the variational setting from the last section.

3.1. Extending the Energy by a Shape Term

To this end, the energy functional in (2.1) is extended by an additional term. This term implements the new model assumption that the shape in the image should be close to the projection of a given object model that can be obtained, e.g., as the mean surface of a set of 3-D training shapes. The extended energy reads:

$$\begin{aligned} E(\Phi, \theta\xi) = & -\int_\Omega (H(\Phi)\log p_1 \\ & + (1 - H(\Phi))\log p_2)\, dx \\ & + \nu \int_\Omega |\nabla H(\Phi)|\, dx \\ & + \underbrace{\lambda \int_\Omega (\Phi - \Phi_0(\theta\xi))^2\, dx}_{\text{Shape}}. \end{aligned} \quad (3.6)$$

The parameter $\lambda \geq 0$ determines the variability of the estimated contour from the modeled one. This variability could as well be estimated in a more sophisticated manner from a set of training shapes. Since this work, however, does not focus on the statistic modeling of shapes but on the general integration of 3-D shapes into 2-D image segmentation, we keep the shape model simple.

The quadratic error measure in the shape term of (3.6) has been proposed in the context of 2-D shape priors, e.g. in Rousson and Paragios (2002), and is not new. However, the prior $\Phi_0 \in \Omega \to \mathbb{R}$ is now derived from a 3-D model and depends on a 3-D rigid transformation $\theta\xi$.

3.1.1. Surface Representation. There exist different ways to represent 3-D shapes (Besl, 1990; Campbell and Flynn, 2001): a common way is to use local representations, e.g. point sets, line segments or curve segments. Global representations can roughly be divided into implicit representations, e.g., by using superquadrics, generalized cylinders, or 3-D level set functions, and explicit ones, e.g. two-parametric meshes, triangulated surfaces, or by using Fourier descriptors. An overview of free-form representations can, e.g., be found in Campbell and Flynn (2001), though the focus of their work is on object recognition and not on pose estimation.

In this work we use a parametric model in terms of a two-parametric surface. This means, a surface F is represented by two sampling parameters ϕ_1 and ϕ_2 and points on the surface are given as

$$F(\phi_1, \phi_2) = (f^1(\phi_1, \phi_2), f^2(\phi_1, \phi_2), f^3(\phi_1, \phi_2))^T.$$

Thus, the surface is represented by three 2-D functions $f^i(\phi_1, \phi_2) : \mathbb{R}^2 \to \mathbb{R}$ acting on the three Euclidean basis vectors. Although other free form models of the surface are applicable as well, we have chosen this representation, since it provides instant access to surface points and allows for quickly computing image silhouettes and projections of the surface mesh. To project the surface mesh to an image plane, just the sample points need to be projected. These points are then connected by line segments. Moreover, one can derive a low-pass object description from this representation by using Fourier descriptors. This has proven beneficial to avoid local minima in the pose estimation. For details we refer to Rosenhahn and Sommer (2004).

3.1.2. Representation of Rigid Transformations using Screw Transformations. Every 3-D rigid body motion (RBM) can be represented as a 4×4 matrix

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{R}_{3\times3} & \boldsymbol{t}_{3\times1} \\ \boldsymbol{0}_{1\times3} & 1 \end{pmatrix} \tag{3.7}$$

for a given rotation matrix $\boldsymbol{R}_{3\times3} \in SO(3)$, with $SO(n) := \{\boldsymbol{R} \in \mathbb{R}^{n\times n} : \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}, \quad det(\boldsymbol{R}) = +1\}$, and a translation vector $\boldsymbol{t}_{3\times1}$. By using homogeneous coordinates, a point $\boldsymbol{x}$ can be transformed by matrix-vector multiplication $\boldsymbol{x}' = \boldsymbol{M}\boldsymbol{x}$. The 3-D rigid body motion has six degrees of freedom, three for the rotation and three for the translation (Gallier, 2001). A common way to represent rigid body motions is by using Euler angles and a translation vector (Murray et al., 1994), thus resulting in a consecutive evaluation of the RBM.

In fact, $\boldsymbol{M}$ is an element of the one-parametric Lie group $SE(3)$, known as the group of direct affine isometries. Elements of a Lie-Group can be represented in an exponential form, thus resulting in a continuous representation of the RBM (i.e. the rotation and translation is evaluated simultaneously for a velocity parameter θ). A main result of Lie theory is that to each Lie group there exists a Lie algebra which can be found in its tangential space by derivation and evaluation at its origin; see (Gallier, 2001; Murray et al., 1994; Sommer, 2001) for more details. The corresponding Lie algebra to $SE(3)$ is $se(3) = \{(\boldsymbol{v}, \omega) \,|\, \boldsymbol{v} \in \mathbb{R}^3, \omega \in so(3)\}$, with $so(3) = \{\boldsymbol{A} \in \mathbb{R}^{3\times3} \,|\, \boldsymbol{A} = -\boldsymbol{A}^T\}$. The elements in $se(3)$ are called *twists*, which can be denoted as

$$\theta\hat{\xi} = \theta\begin{pmatrix} \hat{\omega} & \boldsymbol{v} \\ 0_{3\times1} & 0 \end{pmatrix}, \text{ with}$$
$$\hat{\omega} = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}. \tag{3.8}$$

A twist is sometimes written as vector

$$\theta\xi = \theta(\omega_1, \omega_2, \omega_3, v_1, v_2, v_3)^T, \text{ with}$$
$$\|\omega\|_2 = \|(\omega_1, \omega_2, \omega_3)^T\|_2 = 1. \tag{3.9}$$

It contains six parameters, namely θ, v_1, v_2, v_3 and ω with $\|\omega\|_2 = 1$. To reconstruct a group action $\boldsymbol{M} \in SE(3)$ from a given twist, the exponential function $\exp(\theta\hat{\xi}) = \boldsymbol{M} \in SE(3)$ can be used. The parameter $\theta \in \mathbb{R}$ corresponds to the motion velocity, i.e., the rotation velocity and pitch. For varying θ, the motion can be identified as screw motion around an axis in space. This is proven by Chasles Theorem (Murray et al., 1994) from 1830. Representing a rigid body motion as a screw transformation means to evaluate the rotation and translation simultaneously. This is an important aspect for the later used gradient descent approach for minimizing the constraint equations for pose estimation. Indeed, computing the exponential of a matrix is not trivial, but the RBM from a given twist can be calculated efficiently by using the Rodriguez formula (Murray et al., 1994),

$$\exp(\hat{\xi}\theta) = \begin{pmatrix} \exp(\theta\hat{\omega}) & (I - \exp(\hat{\omega}\theta))(\omega \times \boldsymbol{v}) + \omega\omega^T\boldsymbol{v}\theta \\ 0_{1\times3} & 1 \end{pmatrix}, \quad \text{for } \omega \neq 0 \tag{3.10}$$

with $\exp(\theta\hat{\omega})$ computed by calculating

$$\exp(\theta\hat{\omega}) = I + \hat{\omega}\sin(\theta) + \hat{\omega}^2(1 - \cos(\theta)), \tag{3.11}$$

i.e., only sine and cosine functions of real numbers need to be computed.

3.1.3. Projection of the 3-D Surface to Yield a 2-D Prior. To interact with the segmentation in the image, the surface has to be projected to the image plane. Moreover, the projected shape Φ_0 in the energy (3.6) is assumed to be represented by the signed Euclidean distance function, i.e., $\Phi_0(x)$ yields the Euclidean distance of x to the silhouette of the projected object surface.

For each pose configuration $\theta\hat{\xi}$ one can derive $\Phi_0(\theta\hat{\xi}, x)$ as follows: let X_S denote the set of points $\boldsymbol{X}$ on the object surface. Projection of the transformed points $\exp(\theta\hat{\xi})X_S$ into the image plane yields the set x_S of all 2-D points x on the image plane that correspond to a 3-D point on the surface

$$x = P\exp(\theta\hat{\xi})\boldsymbol{X}, \qquad \forall \boldsymbol{X} \in X_S \tag{3.12}$$

where P denotes a projection with known camera parameters.[3] This set of points yields the binary function $\tilde{\Phi}_0$ representing the projected surface by setting

$$\tilde{\Phi}_0(x) = \begin{cases} 1 & \text{if } x \in x_S \\ -1 & \text{otherwise.} \end{cases} \tag{3.13}$$

By applying the signed distance transform to $\tilde{\Phi}_0$, one

obtains

$$\Phi_0(x) = \begin{cases} \text{dist}(x, C) & \text{if } \tilde{\Phi}_0(x) > 0 \\ -\text{dist}(x, C) & \text{otherwise} \end{cases} \tag{3.14}$$

where $\text{dist}(x, C)$ denotes the Euclidean distance of x to the zero-level line C of $\tilde{\Phi}_0$. An efficient implementation of the Euclidean distance transform can be found in Felzenszwalb and Huttenlocher (2004).

3.1.4. Interpretation of the Energy. The shape term in (3.6) penalizes deviations of the contour Φ from the contour of the projected object model Φ_0. This ensures that Φ cannot deviate too much from the modeled shape. The weighting parameter $\lambda \geq 0$ thereby determines just how far the contour can deviate from the prior. If the correct pose parameters were known, a large value of λ would ensure that the contour fully converges to the shape of the projected object model.

However, the pose parameters are *not* known but are free variables and supposed to be optimized together with the contour. Thus the shape term in (3.6) not only draws the contour towards the projected object model, but also makes the object model to change its pose such that the projection Φ_0 resembles the contour Φ. While Φ_0 thereby has to respect the constraint of a 3-D rigid motion, Φ has to respect the data in the image. In order to minimize the total energy, we suggest an explicit iterative scheme where one optimization variable is kept constant while the other is optimized, and vice-versa.

3.2. Optimization with Respect to the Contour

Since the shape term is modeled in the image domain, minimization of (3.6) with respect to Φ is straightforward and equal to the approach in Rousson and Paragios (2002). It yields the gradient descent equation

$$\begin{aligned} \partial_t \Phi = H'(\Phi) \left(\log \frac{p_1}{p_2} + \nu \,\text{div}\left(\frac{\nabla \Phi}{|\nabla \Phi|} \right) \right) \\ + 2\lambda \left(\Phi_0(\theta \xi) - \Phi \right). \end{aligned} \tag{3.15}$$

One can see again from this evolution equation that the shape term pushes Φ towards the projected surface model, while on the other hand, Φ is still influenced by the image data trying to ensure homogeneous regions according to the maximum a-posteriori criterion.

3.3. Optimization with Respect to the Pose Parameters

Optimization with respect to the pose parameters needs considerably more care, since the way how the projection of the shape varies with a certain 3-D rigid transformation is quite complex. Thus, it is far from straightforward to solve the inverse problem, i.e., to find the 3-D rigid transformation that minimizes the distance of the contours in 2-D. In the following we describe step by step how such an optimization scheme can be derived.

3.3.1. Point Correspondences Between the Contours. Due to Φ and Φ_0 being signed distance functions, the error measure in (3.6) integrates for each point on one contour the distance to the closest point on the other contour. Therefore, we collect the point correspondences from all points on the zero-level of Φ_0 to their closest point on the zero-level of Φ and vice-versa. For each point on the zero-level of Φ_0 we know its 3-D coordinates (Φ_0 was obtained by projecting these 3-D points to the image). Consequently, we obtain a set of correspondences between 2-D points stemming from Φ and 3-D points from the surface model.[4]

Since a rigid transformation changes Φ_0, it may also change the points that correspond to each other. Thus, each iteration has to update the point correspondences. One may note the strong similarities to iterated closest point (ICP) algorithms (Besl and McKay, 1992; Zhang, 1994).

3.3.2. Inverse Projection and Plücker Lines. In order to estimate a 3-D transformation from the correspondences, we change the 2-D points into 3-D entities, i.e., their projection rays need to be constructed. A projection ray contains all 3-D points that, when projected to the image plane, yield a zero distance to the contour point there. Hence, for minimizing the distance in the image plane, one can as well minimize the distance between the model points and the rays reconstructed from the corresponding points.

There exist different ways to represent projection rays. As we have to minimize distances between correspondences, it is advantageous to use an implicit representation for a 3-D line. It allows instantaneously to determine the distance between a point and a line.

One implicit representation of projection rays is by means of so-called *Plücker lines* (Shevlin, 1998; Sommer, 2001). A Plücker line $\boldsymbol{L} = (\boldsymbol{n}, \boldsymbol{m})$ is given as

a unit vector $\boldsymbol{n}$ and a moment $\boldsymbol{m}$ with $\boldsymbol{m} = \boldsymbol{x} \times \boldsymbol{n}$ for a given point $\boldsymbol{x}$ on the line. An advantage of this representation is its uniqueness (apart from possible sign changes). Moreover, the incidence of a point $\boldsymbol{x}$ on a line $\boldsymbol{L} = (\boldsymbol{n}, \boldsymbol{m})$ can be expressed as

$$\boldsymbol{x} \in \boldsymbol{L} \Leftrightarrow \boldsymbol{x} \times \boldsymbol{n} - \boldsymbol{m} = 0. \tag{3.16}$$

This equation provides us with an error vector. Let $\boldsymbol{L} = (\boldsymbol{n}, \boldsymbol{m})$, with $\boldsymbol{m} = \boldsymbol{v} \times \boldsymbol{n}$ as shown in Fig. 2, and $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, with $\boldsymbol{x} \notin \boldsymbol{L}$ and $\boldsymbol{x}_2 \perp \boldsymbol{n}$.

Since $\boldsymbol{x}_1 \times \boldsymbol{n} = \boldsymbol{m}$, $\boldsymbol{x}_2 \perp \boldsymbol{n}$, and $\|\boldsymbol{n}\| = 1$, we have

$$\begin{aligned} \|\boldsymbol{x} \times \boldsymbol{n} - \boldsymbol{m}\| &= \|\boldsymbol{x}_1 \times \boldsymbol{n} + \boldsymbol{x}_2 \times \boldsymbol{n} - \boldsymbol{m}\| = \|\boldsymbol{x}_2 \times \boldsymbol{n}\| \\ &= \|\boldsymbol{x}_2\| \end{aligned} \tag{3.17}$$

where $\|\cdot\|$ denotes the Euclidean norm. This means that $\boldsymbol{x} \times \boldsymbol{n} - \boldsymbol{m}$ in (3.16) results in the (rotated) perpendicular error vector to line $\boldsymbol{L}$.

3.3.3. Pose Estimation. With the result from the last section, the pose can now be determined as the rigid transformation that minimizes the total error over all correspondences i:

$$\sum_i \|(\exp(\theta\hat{\xi})\boldsymbol{x}_i)_{3\times 1} \times \boldsymbol{n}_i - \boldsymbol{m}_i\|_2^2 \to \min. \tag{3.18}$$

Indeed, $\boldsymbol{x}_i$ is a homogeneous 4-D vector, and after multiplication with the 4×4 transformation matrix $\exp(\theta\hat{\xi})$ we neglect the homogeneous component (which is 1) to evaluate the cross product with $\boldsymbol{n}_i$.

While minimizing this total 3-D error is not exactly equivalent to the minimization of the sum of errors in the image plane as stated in the energy, it has been shown in Rosenhahn (2003) that one can provide equivalence between these measures by appropriately rescaling each error vector with a suited η_i:

$$\sum_i \eta_i \|(\exp(\theta\hat{\xi})\boldsymbol{x}_i)_{3\times 1} \times \boldsymbol{n}_i - \boldsymbol{m}_i\|_2^2 \to \min. \tag{3.19}$$

The scalar η_i can be used to rescale the 3-D error vector to gain a different error metric. Let $\boldsymbol{x}_i'$ be the image point of the projection ray $(\boldsymbol{n}_i, \boldsymbol{m}_i)$, $P\boldsymbol{x}_i$ be the projection of $\boldsymbol{x}_i$ and v be the distance of the 3-D point $\boldsymbol{x}_i$ to the projection ray. Then the scaling factor

$$\eta_i = \frac{\|P\boldsymbol{x}_i - \boldsymbol{x}_i'\|_2}{v} \tag{3.20}$$

leads to the desired error in the image plane. For most objects and camera configurations, the rescaling has only very little influence on the estimation result. Thus, if desired, in our implementation the local rescalings can be switched on, but we usually skip it for efficiency reasons, in particular since the 3-D error is not wrong but only inconsistent with the energy functional.

The minimization problem in (3.18) or (3.19) is a least squares problem. Unfortunately, however, the equations are non-quadratic due to the exponential form of the RBM. For this reason, the RBM is linearized, and the pose estimation procedure is iterated, i.e., the nonlinear problem is decomposed into a sequence of linear problems.

By using $\exp(\theta\hat{\xi}) = \sum_{k=0}^{\infty} \frac{(\theta\hat{\xi})^k}{k!} \approx \boldsymbol{I} + \theta\hat{\xi}$, with $\boldsymbol{I}$ as identity matrix, linearization of

$$(\exp(\theta\hat{\xi})\boldsymbol{x})_{3\times 1} \times \boldsymbol{n} - \boldsymbol{m} = 0$$

results in

$$((\boldsymbol{I} + \theta\hat{\xi})\boldsymbol{x})_{3\times 1} \times \boldsymbol{n} - \boldsymbol{m} = 0. \tag{3.21}$$

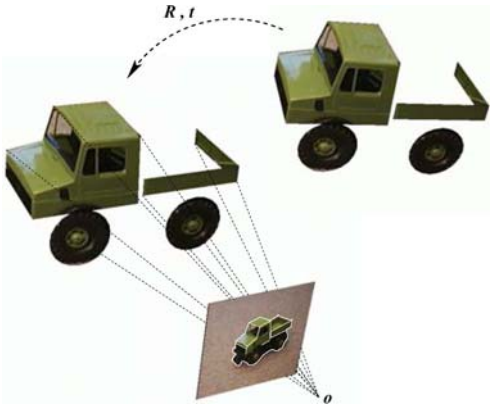

Figure 1. The pose estimation scenario: the aim is to estimate the rigid transformation $\exp(\theta\hat{\xi}) = \boldsymbol{R}, \boldsymbol{t}$ that produces the object contour seen in the image.

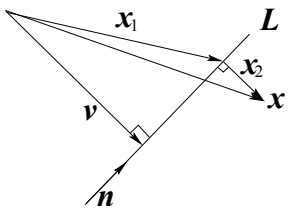

Figure 2. Comparison of a 3-D point $\boldsymbol{x}$ with a 3-D line $\boldsymbol{L}$.

This equation can be reordered into the form $\boldsymbol{A}\xi = \boldsymbol{b}$. Collecting a set of such equations (each is of rank two) leads to an overdetermined linear system of equations, which can be solved using, for example, the Householder algorithm. The Rodriguez formula can be applied to reconstruct the group action $\boldsymbol{M}$ from the estimated twist ξ. Then, the 3-D points can be transformed and the process is iterated until it converges.

Every (linearized) constraint equation yields three rows in the system of equations with respect to the unknown pose parameters. If a confidence measure of the extracted contour Φ is available, it is further possible to scale the equations with respect to the confidence measure, similar to Eq. (3.19). This has the effect that correspondences with a higher confidence are reinforced, whereas correspondences with lower confidence are alleviated. In Rosenhahn (2003) this property to manipulate local correspondences is called *adaptive pose estimation*. Note, however, that the scaling of equations is no longer consistent with the energy in (3.6).

In case of minimizing the 3-D error measure, the projection rays only need to be reconstructed once, and can be reconstructed from orthographic, projective or even catadioptric cameras. The algorithm is very fast (e.g., it needs 2 ms on a standard (2 GHz) Linux PC for 100 point correspondences). In Rosenhahn (2003) extensions to point-plane, line-plane constraint equations and kinematic chains are presented using Clifford algebra (Sommer, 2001).

3.3.4. A Confidence Measure for Contour Points. Although we are currently not able to state an energy that is minimized by the scaled pose estimation equations, we introduce here a possibility to derive a measure of confidence for the contour. It takes into account that the separability of the object and the background region can be considerably reduced in some areas. This happens in particular at locations where the shape prior contradicts the local region statistics, e.g., due to occlusions.

The sought confidence at a certain point x on the contour can be expressed by the a-posteriori probability of the region the point has been assigned to. This reads

$$\tilde{c}(x) = \frac{p_1(x)\int_{\Omega_1} K_\rho(x)\,d\xi}{p(x)} H(\Phi(x)) + \frac{p_2(x)\int_{\Omega_2} K_\rho(x)\,d\xi}{p(x)}(1 - H(\Phi(x))) \tag{3.22}$$

where K_ρ is the Gaussian kernel from Section 2.3. If a pixel assigned to region Ω_1 also fits well to region Ω_2, i.e., $p_1 \approx p_2$, the precise location of the contour will be ambiguous and the confidence will be around 0.5. Obversely, if a pixel assigned to Ω_1 does not fit to region Ω_2, i.e., $p_1 \gg p_2$, the contour location will be definite and the confidence will be close to 1. If a pixel is assigned to the wrong region according to the statistics—this can happen due to contradictions with the object prior or the length constraint—the confidence will be even smaller than 0.5.

Due to slightly blurred edges, pixels directly on the contour often have a quite low confidence, although the separability of the regions in the surrounding area is high. Therefore, it is reasonable to take also pixels from the neighborhood into account. This can be achieved by a simple convolution with a Gaussian kernel K_σ

$$c(x) = (K_\sigma * \tilde{c})(x) \tag{3.23}$$

where we set $\sigma = 1.5$. When scaling the equations in Section 3.3.3 by $c(x)$ one can obtain slightly improved results as demonstrated by one experiment in Section 4.

3.4. Summary of the Optimization Procedure

The complete optimization procedure can be summarized as follows:

1. Initialize the pose parameters by some values that are sufficiently close to the true pose.
2. Project the surface model with the current pose parameters to the image plane and construct Φ_0 as described in Section 3.1.3.
3. Initialize Φ with Φ_0.
4. Compute the probability densities p_i for the current segmentation Φ.
5. Update Φ according to (3.15).
6. Compute the set of point correspondences as described in Section 3.3.1.
7. Reconstruct projection rays from the 2-D points as described in Section 3.3.2.
8. Find the pose parameters that minimize the total distance between the 3-D points and the projection rays as described in Section 3.3.3.
9. Update Φ_0 as in 2.
10. Iterate 4–9.
11. Repeat 3–10 for each frame in the image sequence.

An illustration can be found in Fig. 3.

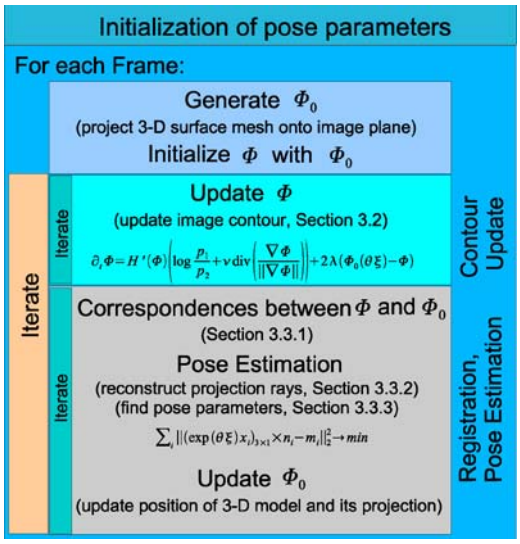

Figure 3. Summary of the optimization scheme.

4. Experiments

We investigated the performance of our joint contour extraction and pose estimation method in a couple of experiments. Figure 4 first demonstrates the general advantage of integrating object knowledge into the segmentation process. Without object knowledge, parts of the tea box are missing as they better fit to the background. The object prior can constrain the contour to the vicinity of the projected object model derived from those parts of the contour that can be extracted reliably. This concept is also the key issue of approaches that use 2-D shape knowledge. With 3-D shape knowledge, however, it is no longer necessary to model several views. Moreover, the object model can perfectly fit the data, while in 2-D approaches there remain discrepancies if the current view does not coincide perfectly with one of the modeled views.

In Fig. 5 we show the robustness of the method in the case of a changing background. One can see that the estimated pose of the tea box is not distracted by any of the objects moved in the background, though the CDs even reflect the tea box surface. Later on in the sequence, also the tea box itself is moved, which shows that the method is not tuned for static objects.

In the experiment shown in Fig. 6, we tested the influence of artifacts like reflections, shadows, and noise. The motion of the object causes partially severe reflections on the metallic surface of the tea box. Moreover, the tea box throws a shadow as it is tilted. Additionally, Gaussian noise with standard deviation 30 has been added to the sequence. These difficulties partially lead to small errors, yet the overall results remain stable. Also the slight occlusion due to the fingers does not harm the pose estimation. The presence of noise in this sequence clearly rules out methods that are based on background subtraction. Also simple thresholding methods for contour extraction would fail due to the cluttered background and the reflections.

Figure 7 compares the results obtained with and without the suggested confidence measure, respectively. The confidence measure prevents the result from being deteriorated by the shadow and the occluding fingers. With a homogeneous weighting, the correct contour points have not enough weight to ensure the correct pose estimate.

In the experiment depicted in Fig. 8, the monocular camera has been extended to a stereo system. In this case, another significant advantage of using 3-D shape knowledge becomes apparent. In contrast to 2-D approaches, our method can fuse the information from two images. If the information in one image is not reliable, e.g due to occlusions, the information from the other image can still determine the pose. Even if there are occlusions in both images, the combined information from both images can be still sufficient for a reliable pose estimation. The object model with the correct pose, on the other hand, constrains the contour and keeps it from breaking away.

Figure 4. From left to right: (a) Initialization. (b) Segmentation result with object knowledge. (c) Pose result. (d) Segmentation result without object knowledge.

Figure 5. *Top row:* Input images for frames 51, 189 and 450 of an image sequence containing 560 frames. *Bottom row:* Pose results. The algorithm is able to deal with a cluttered and changing background.

Figure 6. *Top row:* Initialization at the first frame. Contour at frames 49, 50, and 116 of the sequence. *Bottom row:* Pose results at frames 0, 49, 50, and 116. The tea box is moved, causing partially severe reflections on the box. Furthermore, Gaussian noise with standard deviation 30 has been added.

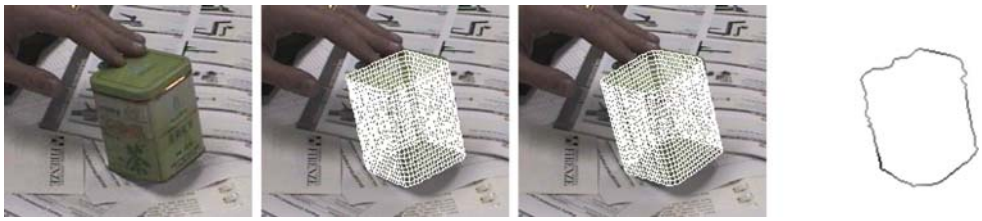

Figure 7. From left to right: (a) Frame 13. (b) Pose estimation result without the proposed confidence measure. (c) Result when exploiting the confidence. (d) Confidence along the contour. Dark values represent a high confidence.

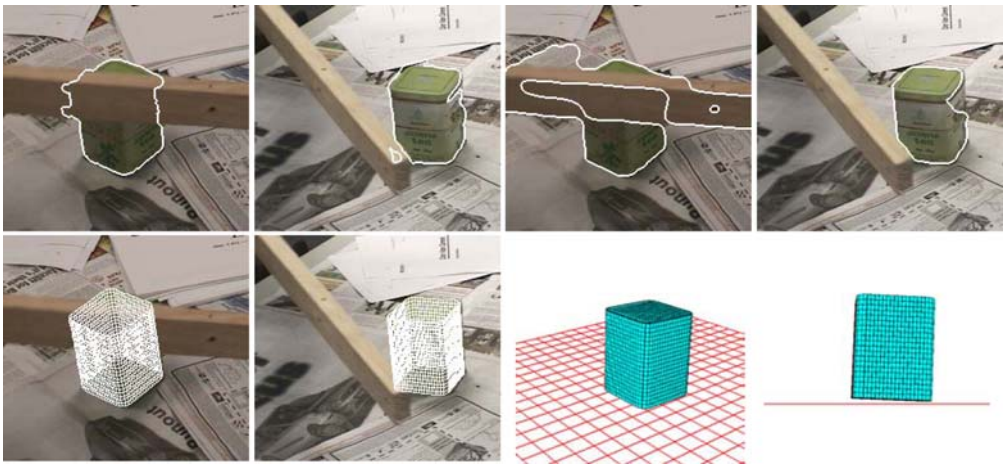

Figure 8. Frame 97 from a stereo sequence with 400 frames. In both views the object is partially occluded. *Top left:* Due to the shape prior, the contour is kept close to object. *Top right:* Here, the contour has been initialized at this frame with the correct contour, but the shape constraint has been neglected ($\lambda = 0$). Consequently, the contour breaks away. *Bottom left:* Pose result. *Bottom right:* Visualization of the object pose from two different views. Each square of the floor represents two centimeters in the world coordinate system. The pose is recovered with a deviation of only a few millimeter.

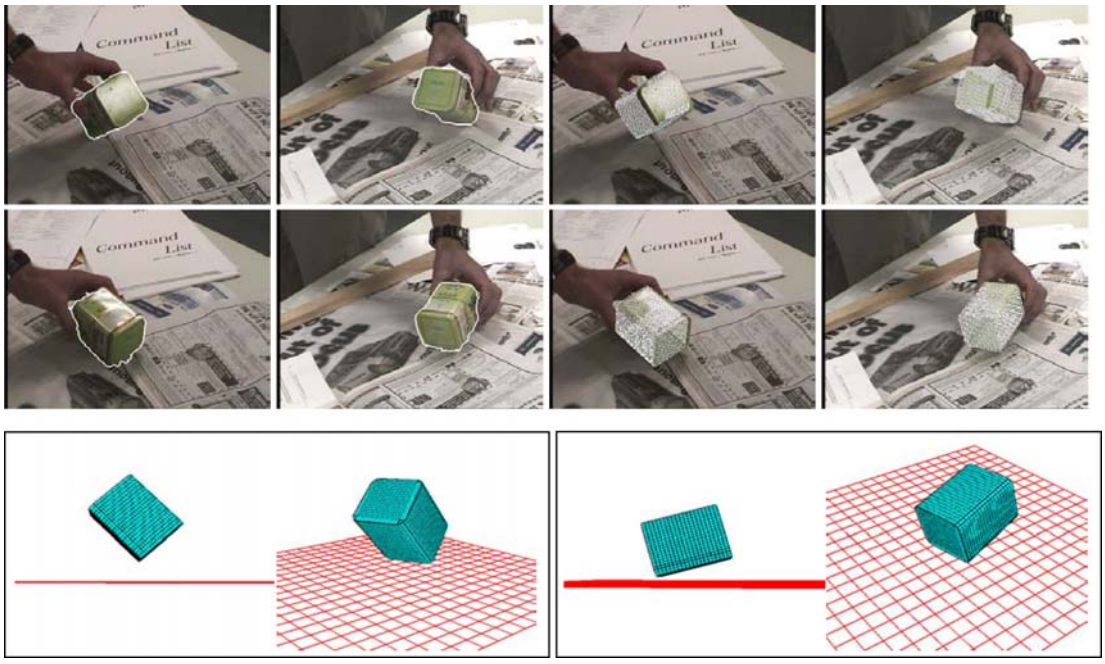

Figure 9. *Top:* Contour and pose results at frame 190 and 212 for the stereo sequence from Fig. 8. *Bottom:* Pose results at frame 190 and 212 from two different perspective views. Each floor square represents two centimeter in the world coordinate system.

In the sequel of this stereo sequence, the tea box is moved. Two further frames are depicted in Fig. 9. Again there appear reflections on the surface of the box, and there are further partial occlusions due to the hand.

In order to demonstrate that the approach is not restricted to a certain type of object, Fig. 10 shows an experiment with a teapot model. This object is non-convex and even contains a hole. Dealing with such a kind of object, it is particularly beneficial to represent the contour by means of a level set function. In the level set framework, the more complex topology does not change anything. Thus, the region encircled by the

Figure 10. *From left to right:* (a) Stereo image with a teapot. The initialization is quite far away from the object. (b) Resulting contour when performing only one iteration. The contour is restricted to the initial, bad pose and cannot fully capture the object. (c) Consequently, the pose remains close to the initialization. (d) Pose result after 20 iterations.

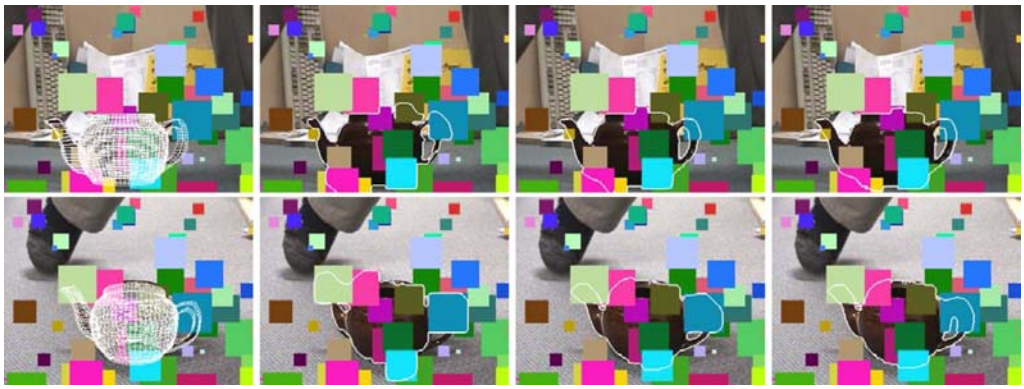

Figure 11. *Most left:* Stereo image of the teapot. The image pair is disturbed with rectangles of random size, position, and color. The initialization is close to the correct pose. *From left to Right:* Segmentation results for varying weights of the shape constraint $\lambda = 0.001, 0.06$, and 0.1.

handle of the teapot can correctly be assigned to the background region. The roundish teapot immediately rules out line based methods for this task. Also methods based on feature matching may have difficulties due to the homogeneous surface of the object. Further note the rather bad initialization. A decoupled concatenation of the segmentation technique and the pose estimation method cannot succeed in finding the right contour and pose. Only the mutual improvement of both the contour and the pose allows for a good result in the steady state.

Figure 11 illustrates the role of the weighting parameter λ for the shape prior. In this example we use one frame of the stereo sequence with the tea pot disturbed by colored rectangles of random size and position. The initialization is shown on the left. The remaining images show the contour for $\lambda = 0.001$, 0.06 and 0.1, respectively. Small λ give the contour much freedom to evolve, enabling the pose to follow. If the object region can be clearly separated from the background, choosing λ small is therefore beneficial. On the other hand, if the contour is distracted by background clutter, the shape information keeps the contour from running too far away from the object. In our other experiments we have therefore chosen λ in the area of 0.06.

Figure 12 depicts a sequence where object and camera are static to allow a quantitative error measurement. The diagrams on the left show the translational and angular errors along the three axes, respectively. Despite the change of the lighting conditions and partial occlusions, the error has a standard deviation of less than 2 mm and 6.0 degrees. The main rotational errors occur for rotations around the x-axis of the calibrated system. This is due to the fact that such a rotation causes smaller changes of the silhouette than rotations around the other axes. Therefore, this degree of freedom is more sensitive to inaccuracies or errors in the

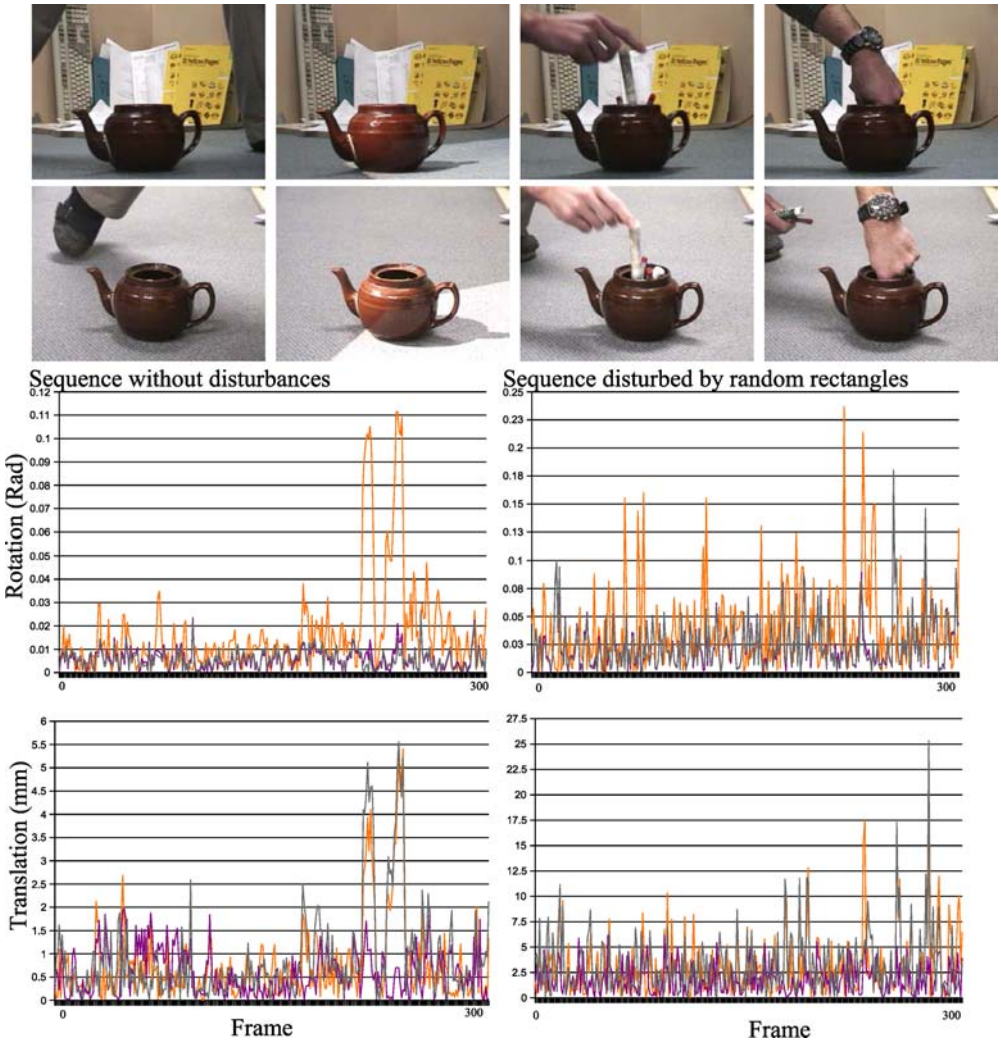

Figure 12. *Top row:* Some frames from a static stereo sequence with illumination changes and partial occlusions (left and right view in the top and bottom row, respectively). *Diagrams left:* Rotational and translational errors in radians and millimeters for the undisturbed sequence. *Diagrams right:* The sequence has been disturbed with random rectangles as shown in Fig. 13. The estimation errors increase to up to 2.5 cm in space.

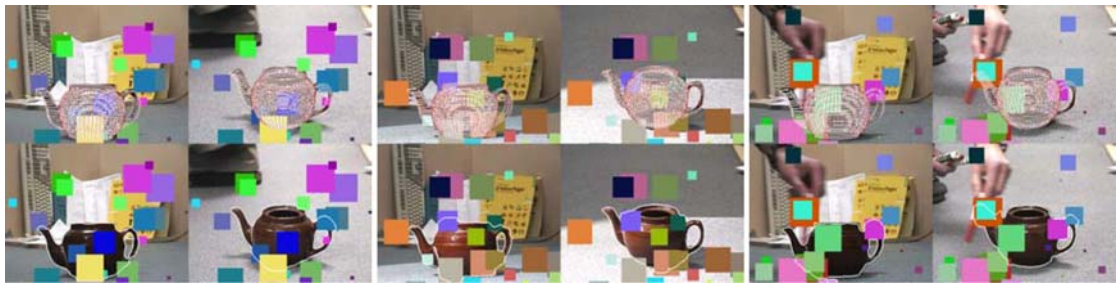

Figure 13. The sequence from Fig. 12 has been disturbed with rectangles of random size, position, and color, which leads to occlusions of the object. *Top row:* Pose results for different frames. *Bottom row:* Segmentation results. The occlusions can be compensated due to the object model. The result on the right shows the worst pose according to the diagram in Fig. 12.

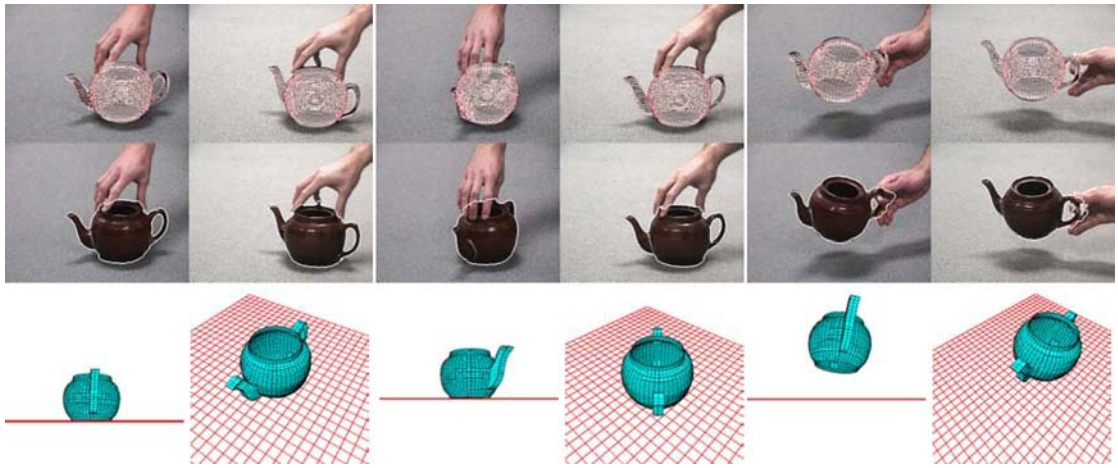

Figure 14. A second stereo sequence with the tea pot, where the handle of the tea pot vanishes behind the container and reappears. Finally the tea pot is moved around (345 frames). *Top row:* Pose results for different frames. *Middle row:* Segmentation results. Due to the shape model, the occluding hand only slightly disturbs the contour extraction. *Last row:* Pose results from different perspective views. The rotation on the ground floor is accurately estimated.

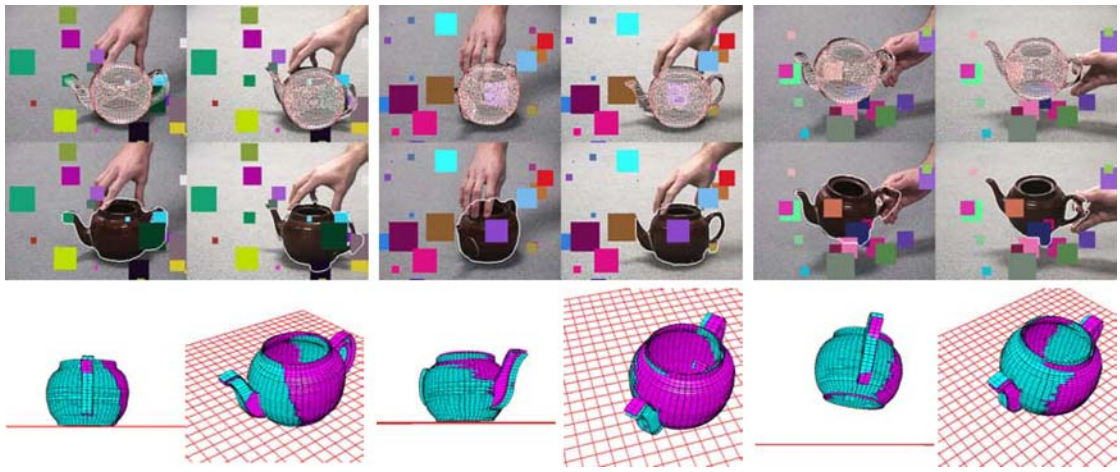

Figure 15. The stereo sequence from Fig. 14 is now randomly disturbed with rectangles. *Top row:* Pose results. *Middle row:* Segmentation results. *Last row:* Pose results from different perspective views. The pose result of the non-disturbed images is blended with the scene. The deviation is in the area of one centimeter.

extracted contour. The x-axis is located horizontally along the teapot, crossing the center of the teapot, and pointing from the handle to the spout.

The diagrams on the right hand side of the same figure show the translational and angular errors for the same sequence disturbed by random rectangles, as shown in Fig. 13. The occlusions partially lead to bad segmentations, which can be compensated due to the object model. Here, the error is up to 25 mm and 12 degrees.

In order to demonstrate the ability of the approach to deal with topological changes in the contour, Fig. 14 shows pose and segmentation results of a second stereo sequence. At the beginning, the teapot is rotated on the ground floor (along the y-axis), such that the handle and the opening of the handle vanish behind the container and reappear later. Thanks to the shape prior, the level set segmentation can recapture the hole when it reappears.

In the sequel, the handle is grabbed and the tea pot is moved around. The bottom row in Fig. 14 visualizes the 3-D pose in a virtual environment. During rotation on the ground floor, the tea pot is nearly perfectly on the floor as it should be (see also Fig. 16).

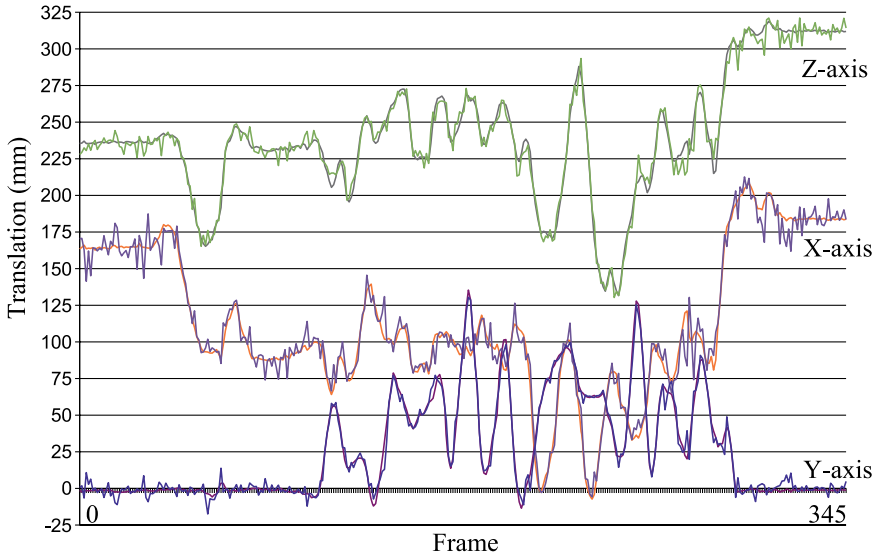

Figure 16. Comparison of the x-, y- and z-axis of the estimated pose for the stereo sequence in Figs. 14 and 15. In the first part of the sequence, the tea pot is just rotated on the floor. The nearly constant values of the y-axis (with a deviation of 2–5 mm) indicate a stable result. Then the tea pot is grabbed and moved around. The curves for the disturbed images show larger errors (up to 2 cm), but it is still possible to track the tea pot.

Again, we disturbed the sequence by occluding rectangles. The results are depicted in Fig. 15. For analyzing the impact of the disturbances, in the last row, pose results from the disturbed sequence are blended with the results from the non-disturbed data. The deviation is in the area of one centimeter, which demonstrates the stability of the method in case of occlusions. The diagram in Fig. 16 further quantifies this outcome. It shows the tracking curves for the disturbed and undisturbed sequences in Figs. 14 and 15, respectively. In the first part of the sequence, the tea pot is just rotated on the floor. The nearly constant values of the y-axis (with a deviation of 2–5 mm) indicate a stable result. The values of the disturbed sequence have a higher deviation (up to 2 cm), but it is still possible to reliably track the tea pot.

The overall computation time depends on the number of iterations necessary for the method to converge. For the last (and hardest) sequence that includes the disturbances by random rectangles, the computation time per stereo pair was approximately 2 min (1 min and 50 sec to 2 min and 2 sec) on a 2.4 GHz opteron Linux machine. The computation time is significantly larger than with other pose tracking models that often achieve real-time performance. However, in contrast to these approaches, our model includes a sophisticated interlocking of region based segmentation and pose estimation as well as statistical region models that allow for good results in situations where current real-time approaches may fail. Even recent pose trackers based on local descriptors, which yield very good results,[5] generally do not work well with homogeneous objects like the teapot. We have shown that such objects can be tracked reliably by our approach despite background clutter and occlusions, which disturb other contour based techniques.

5. Conclusion

In this work, variational and statistical methodologies have been combined with geometric techniques previously developed in the language of Clifford algebras. We introduced a method that integrates 3-D shape knowledge into a variational model for level set based image segmentation. While the utilization of 2-D shape knowledge has been investigated intensively in recent time, the presented approach takes the three-dimensional nature of the world into account. The method relies on a powerful image-driven segmentation model on one side, and an elaborated technique for contour based 2D-3D pose estimation on the other side. The combination of both techniques in a joint energy minimization problem improves the quality of contour extraction and, consequently, also the robustness of pose estimation, which relies on the contour. This allows for the tracking of three-dimensional objects in cluttered scenes with inconvenient illumination effects and partial occlusions.

Appendix A: Euler-Lagrange Equations for Local Statistics

A popular way to minimize energies like the one in (2.1) is the EM-algorithm. EM keeps the probability densities p_i fixed for computing an update of the contour Φ and then, vice-versa, updates the densities while retaining the contour. Although it is known that the EM-algorithm converges, the dependency between the contour and the densities is neglected. Alternatively, one can write (2.1) as a functional that only depends on Φ and compute the corresponding Euler-Lagrange equation. In the following, we derive this Euler-Lagrange equation for local Gaussian region statistics as introduced in Section 2.3.

First we introduce the characteristic functions $\chi_1 := H(\Phi)$ and $\chi_2 := (1 - H(\Phi))$. Then the energy in (2.1) can be written as:

$$-\sum_{i=1}^{2} \int_\Omega \chi_i(\Phi) \log p_i \, dx + \nu \int_\Omega |\nabla H(\Phi)| \, dx.$$

local mean μ and variance σ^2 yields

$$\int_\Omega \chi(\Phi) \left[\frac{\left(I - \frac{\int K I \chi(\Phi)\, d\zeta}{\int K \chi(\Phi)\, d\zeta}\right)^2}{\frac{\int K I^2 \chi(\Phi)\, d\zeta}{\int K \chi(\Phi)\, d\zeta} - \left(\frac{\int K I \chi(\Phi)\, d\zeta}{\int K \chi(\Phi)\, d\zeta}\right)^2} + \log\left(\frac{\int K I^2 \chi(\Phi)\, d\zeta}{\int K \chi(\Phi)\, d\zeta} - \left(\frac{\int K I \chi(\Phi)\, d\zeta}{\int K \chi(\Phi)\, d\zeta}\right)^2\right)\right] dx$$

where K denotes a local Gaussian window centered at x. For the Euler-Lagrange equations we compute

$$\frac{\partial}{\partial \epsilon} \int_\Omega \mathcal{L}(\Phi + \epsilon h)\, dx \Big|_{\epsilon=0}.$$

With the abbreviations

$$A := \int_\Omega K \chi(\Phi)\, d\zeta \qquad s := \int_\Omega K \chi'(\Phi)\, d\zeta$$

$$\mu_C := \frac{\int K I \chi'(\Phi)\, d\zeta}{s} \qquad \theta_C := \frac{\int K I^2 \chi'(\Phi)\, d\zeta}{s}$$

we obtain

$$\underbrace{\chi'(\Phi)\left(\frac{(I-\mu)^2}{\sigma^2} + \log\sigma^2\right)}_{\text{usual term}} \underbrace{-\chi(\Phi)\frac{s}{A}\left[\frac{2(I-\mu)}{\sigma^2}(\mu_C - \mu) + \left(\frac{(I-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}\right)(\theta_C - 2\mu\mu_C - \sigma^2 + \mu)\right]}_{\text{additional term}} = 0.$$

In the following we neglect the last term, since it is independent from the region statistics. We further see that the terms for both regions are symmetric. We can thus concentrate on computing the Euler-Lagrange equation for one of the regions. The additional term for the second region can be derived the same way. So we compute the Euler-Lagrange equation of

$$-\int_\Omega \chi(\Phi) \log p \, dx = \frac{1}{2}\int_\Omega \chi(\Phi)\left(\frac{(I-\mu)^2}{\sigma^2} + \log(2\pi) + \log\sigma^2\right) dx$$

As the pre-factor $1/2$ has no influence on the minimizer, it can be neglected. Also the term $\log(2\pi)$ can be neglected, as the same term reappears for the other region with opposite sign and thus cancels out. Expanding the

Besides the usual term known from the EM-algorithm, there is a further term that is zero if the window K covers the whole image domain Ω (global density estimate). Then the integral of $I - \mu$ is zero and the integral of $\frac{(I-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}$, too. In general the term is not zero, yet one can see that it is small, as it depends on the ratio between the contour length and the region area. For a window K with reasonable size, this ratio is considerably smaller than 1. Moreover, the importance of the term depends on the difference between the statistics along the contour and within the region. This difference is usually very small, as well, especially when the contour is still far from the state of convergence. For this reason, the additional term can be neglected without loosing the advantages of the variational framework.

Acknowledgments

We gratefully acknowledge funding by the German Research Foundation (DFG) under the projects Ro2497/1, We2602/1, and Cr250/1.

Notes

1. Extensions from rigid bodies to kinematic chains have been suggested in Bregler and Malik (1998), and Bregler et al. (2004).
2. A related problem appears in certain works on shape reconstruction, e.g. in Faugeras and Keriven (1998), and Yezzi and Soatto (2003a,b).
3. In case of a discrete mesh representation of the surface, as assumed above, one has to fill the gaps in the projected mesh to obtain a continuous representation of the projected surface.
4. For being fully consistent with the energy in (3.6), one had to match not only the points on the zero-level lines of Φ and Φ_0, but additionally all points where $\Phi_0 > 0$. This could be done, for instance, with a variation of the framework in Paragios et al. (2003). For efficiency reasons, however, we take only point correspondences for points on the contours into account.
5. for instance, the method in Vacchetti et al. (2004) can deal with larger displacements than our method.

References

Araújo, H., Carceroni, R.L., and Brown, C.M. 1998. A fully projective formulation to improve the accuracy of Lowe's pose-estimation algorithm. *Computer Vision and Image Understanding*, 70(2):227–238.

Besl, P. and McKay, N. 1992. A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:239–256.

Besl, P.J. 1990. The free-form surface matching problem. In *Machine Vision for Three-Dimensional Scenes*, H. Freemann (Ed.), Academic, Press: San Diego, pp. 25–71.

Beveridge, J.R. 1993. Local search algorithms for geometric object recognition: Optimal correspondence and pose. Technical Report Technical Report CS 93–5, University of Massachusetts, Amherst.

Blake, A. and Zisserman, A. 1987. *Visual Reconstruction*. MIT Press: Cambridge, MA.

Blaschke, W. 1960. *Kinematik und Quaternionen, Mathematische Monographien*. 4. Deutscher Verlag der Wissenschaften.

Bregler, C. and Malik, J. 1998. Tracking people with twists and exponential maps. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Santa Barbara, California, pp. 8–15.

Bregler, C., Malik, J., and Pullen, K. 2004. Twist based acquisition and tracking of animal and human kinetics. *International Journal of Computer Vision*, 56(3):179–194.

Brox, T., Rosenhahn, B., and Weickert, J. 2005. Three-dimensional shape knowledge for joint image segmentation and pose estimation. In *Pattern Recognition*, W. Kropatsch, R. Sablatnig, and A. Hanbury (Eds.), volume 3663 of *LNCS*, Springer, pp. 109–116.

Brox, T. and Weickert, J. 2005. Level set segmentation with multiple regions. Technical Report 145, Dept. of Mathematics, Saarland University, Saarbrücken, Germany.

Brox, T. and Weickert, J. 2006. A TV flow based local scale estimate and its application to texture discrimination. *Journal of Visual Communication and Image Representation*, To appear.

Campbell, R. and Flynn, P. 2001. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, (81):166–210.

Caselles, V., Catté, F., Coll, T., and Dibos, F. 1993. A geometric model for active contours in image processing. *Numerische Mathematik*, 66:1–31.

Chan, T. and Vese, L. 1999. An active contour model without edges. In *Scale-Space Theories in Computer Vision*, M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert (Eds.), volume 1682 of LNCS, Springer, pp. 141–151.

Chan, T. and Vese, L. 2001. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277.

Cremers, D., Osher, S., and Soatto, S. 2004. A multi-modal translation-invariant shape prior for level set segmentation. In *Pattern Recognition*, C.-E. Rasmussen, H. Bülthoff, M. Giese, and B. Schölkopf (Eds.), volume 3175 of LNCS, Springer, Berlin, pp. 36–44.

Cremers, D., Schnörr, C., and Weickert, J. 2001. Diffusion-snakes: Combining statistical shape knowledge and image information in a variational framework. In *Proc. First IEEE Workshop on Variational and Level Set Methods in Computer Vision*, Vancouver, Canada, IEEE Computer Society Press, pp. 137–144.

Cremers, D. and Soatto, S. 2005. Motion competition: A variational framework for piecewise parametric motion segmentation. *International Journal of Computer Vision*, 62(3):249–265.

Cremers, D., Tischhäuser, F., Weickert, J., and Schnörr, C. 2002. Diffusion snakes: Introducing statistical shape knowledge into the mumford-shah functional. *International Journal of Computer Vision*, 50(3):295–313.

Dempster, A., Laird, N., and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society series B*, 39:1–38.

Dervieux, A. and Thomasset, F. 1979. A finite element method for the simulation of Rayleigh–Taylor instability. In *Approximation Methods for Navier–Stokes Problems*, R. Rautman (Ed.), volume 771 of *Lecture Notes in Mathematics*, Springer pp. 145–158.

Drummond, T. and Cipolla, R. 2000. Real-time tracking of multiple articulated structures in multiple views. In *Proc. 6th European Conference on Computer Vision, ECCV*, Dublin, Ireland, Springer, pp. 20–36.

Faugeras, O. 1993. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press: Cambridge, MA.

Faugeras, O. and Keriven, R. 1998. Variational principles, surface evolution, PDE's, level set methods, and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):336–344.

Felzenszwalb, P.F. and Huttenlocher, D.P. 2004. Distance transforms of sampled functions. Technical Report TR2004-1963, Computer Science Department, Cornell University.

Gallier, J. 2001. *Geometric Methods and Applications For Computer Science and Engineering.* Springer-Verlag: New York Inc.

Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Goddard, J. 1997. *Pose and Motion Estimation From Vision Using Dual Quaternion-Based Extended Kalman Filtering*. PhD thesis, Knoxville.

Grimson, W.E.L. 1990. *Object Recognition by Computer.* MIT Press: Cambridge, MA.

Haag, M. and Nagel, H.-H. 1999. 3d-model-based vehicle tracking in traffic image sequences. *International Journal of Computer Vision*, 35(3):295–319.

Heiler, M. and Schnörr, C. 2005. Natural image statistics for natural image segmentation. *International Journal of Computer Vision*, 63(1):5–19.

Kadir, T. and Brady, M. 2003. Unsupervised non-parametric region segmentation using level sets. In *Proc. Ninth IEEE International Conference on Computer Vision*, vol. 2, pp. 1267–1274.

Kass, M., Witkin, A., and Terzopoulos, D. 1988. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331.

Kim, J., Fisher, J., Yezzi, A., Cetin, M., and Willsky, A. 2002. Nonparametric methods for image segmentation using information theory and curve evolution. In *IEEE International Conference on Image Processing*, Rochester, NY vol. 3, pp. 797–800, .

Kim, J., Fisher, J., Yezzi, A., Cetin, M., and Willsky, A. 2005. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Transactions on Image Processing*, 14(10):1486–1502.

Kriegman, D., Vijayakumar, B., and Ponce, J. 1992. Constraints for recognizing and locating curved 3D objects from monocular image features. In *Proc. 2nd European Conference on Computer Vision (ECCV '92)*, G. Sandini (Ed.), volume 588 of *Lecture Notes in Computer Science*, Springer, pp. 829–833.

Lepetit, V. and Fua, P. 2005. Monocular model-based 3D tracking of rigid objects: A survey. *Computer Graphics and Vision*, 1(1):1–89.

Leventon, M.E., Grimson, W.E.L., and Faugeras, O. 2000. Statistical shape influence in geodesic active contours. In *Proc. 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Hilton Head, SC, vol. 1, pp. 316–323.

Li, S.Z. 1995. *Markov Random Field Modeling in Computer Vision*. Springer Verlag: New York.

Lowe, D. 1980. Solving for the parameters of object models from image descriptions. In *Proc. ARPA Image Understanding Workshop*, pp. 121–127.

Lowe, D. 1987. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395.

Ma, Y., Soatto, S., Kosecka, J., Sastry, S.S., and Soatta, S. 2003. *An Invitation to 3-D Vision*. Springer Verlag: New York.

Malik, J., Belongie, S., Leung, T., and Shi, J. 2001. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27.

Malladi, R., Sethian, J.A., and Vemuri, B.C. 1995. Shape modeling with front propagation: A level set approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2):158–175.

Mansouri, A., Mitiche, A., and Vázquez, C. 2004. Image partioning by level set multiregion competition. In *Proc. International Conference on Image Processing*, vol. 4, pp. 2721–2724.

Marchand, E., Bouthemy, P., and Chaumette, F. 2001. A 2D-3D model-based approach to real-time visual tracking. *Image and Vision Computing*, 19(13):941–955.

McLachlan, G. and Krishnan, T. 1997. *The EM Algorithm and Extensions*. Wiley series in probability and statistics. John Wiley & Sons.

Mumford, D. and Shah, J. 1989. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42:577–685.

Murray, R., Li, Z., and Sastry, S. 1994. *Mathematical Introduction to Robotic Manipulation.* CRC Press: Boca Raton, FL.

Osher, S. and Sethian, J.A. 1988. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *Journal of Computational Physics*, 79:12–49.

Paragios, N. and Deriche, R. 1999. Unifying boundary and region-based information for geodesic active tracking. In *Proc. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Forth Collins, Colorado, vol. 2, pp. 300–305.

Paragios, N. and Deriche, R. 2002. Geodesic active regions: A new paradigm to deal with frame partition problems in computer vision. *Journal of Visual Communication and Image Representation*, 13(1/2):249–268.

Paragios, N. and Deriche, R. 2002. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal of Computer Vision*, 46(3):223–247.

Paragios, N., Rousson, M., and Ramesh, V. 2003. Distance transforms for non-rigid registration. *Computer Vision and Image Understanding*, 23:142–165.

Riklin-Raviv, T., Kiryati, N., and Sochen, N. 2004. Unlevel-sets: Geometry and prior-based segmentation. In *Proc. 8th European Conference on Computer Vision*, T. Pajdla and J. Matas (Eds.), volume 3024 of *LNCS*, Springer, Berlin, pp. 50–61.

Rosenhahn, B. 2003. *Pose Estimation Revisited*. PhD thesis, University of Kiel, Germany.

Rosenhahn, B. and Sommer, G. 2004. Pose estimation of free-form objects. In *Computer Vision - Proc. 8th European Conference on Computer Vision*, T. Pajdla and J. Matas (Eds.), vol. 3021 of *LNCS*, Springer, pp. 414–427.

Rousson, M. Brox, T., and Deriche, R. 2003. Active unsupervised texture segmentation on a diffusion based feature space. In *Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, pp. 699–704.

Rousson, M. and Deriche, R. 2002. A variational framework for active and adaptive segmentation of vector-valued images. In *Proc. IEEE Workshop on Motion and Video Computing*, Orlando, Florida, pp. 56–62.

Rousson, M. and Paragios, N. 2002. Shape priors for level set representations. In *Computer Vision – ECCV 2002*, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen (Eds.), vol. 2351 of *LNCS*, Springer, Berlin pp. 78–92.

Rousson, M., Paragios, N., and Deriche, R. 2004. Implicit active shape models for 3D segmentation in MR imaging. In *7th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 3216 of *LNCS*, Springer, Berlin, pp. 209–216.

Shevlin, F. 1998. Analysis of orientation problems using Plücker lines. In *International Conference on Pattern Recognition (ICPR)*, Brisbane vol. 1, pp. 685–689.

Shi, J. and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Sifakis, E., Garcia, C., and Tziritas, G. 2002. Bayesian level sets for image segmentation. *Journal of Visual Communication and Image Representation*, 13(1/2):44–64.

Sommer, G. (Ed) 2001. *Geometric Computing with Clifford Algebra*. Springer Verlag: Berlin.

Tsai, A., Yezzi, A., and Willsky, A. 2001. Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE Transactions on Image Processing*, 10(8):1169–1186.

Vacchetti, L., Lepetit, V., and Fua, P. 2004. Stable real-time 3D tracking using online and offline information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1391–1391.

Vese, L. and Chan, T. 2002. A multiphase level set framework for image segmentation using the Mumford and Shah model. *International Journal of Computer Vision*, 50(3):271–293.

Yezzi, A. and Soatto, S. 2003a. Stereoscopic segmentation. *International Journal of Computer Vision*, 53(1):31–43.

Yezzi, A. and Soatto, S. 2003b. Structure from motion for scenes without features. In *Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, vol. 1, pp. 171–178.

Yezzi, A., Zollei, L., and Kapur, T. 2001. A variational framework for joint segmentation and registration. In *Proc. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pp. 44–51.

Zerroug, M. and Nevatia, R. 1996. Pose estimation of multi-part curved objects. In *Proc. Image Understanding Workshop*, pp. 831–835.

Zhang, Z. 1994. Iterative points matching for registration of free form curves and surfaces. *International Journal of Computer Vision*, 13(2):119–152.

Zhao, H.K., Chan, T., Merriman, B., and Osher, S. 1996. A variational level set approach to multiphase motion. *Journal of Computational Physics*, 127:179–195.

Zhu, S.-C. and Yuille, A. 1996. Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884–900.