In Pattern Recognition, Springer LNCS 3663, W. Kropatsch, R. Sablatnig, and A. Hanbury (Eds.), pp. 109-116, Vienna, Austria, Aug. 2005

Three-Dimensional Shape Knowledge for Joint Image Segmentation and Pose Estimation ⋆

Thomas Brox[1], Bodo Rosenhahn[2], and Joachim Weickert[1]

[1] Mathematical Image Analysis Group, Faculty of Mathematics and Computer Science, Saarland University, Building 27, 66041 Saarbrücken, Germany
{brox,weickert}@mia.uni-saarland.de

[2] Centre for Imaging Technology and Robotics (CITR), University of Auckland, New Zealand
bros028@cs.auckland.ac.nz

Abstract. This paper presents the integration of 3D shape knowledge into a variational model for level set based image segmentation and tracking. Having a 3D surface model of an object that is visible in the image of a calibrated camera, the object contour stemming from the segmentation is applied to estimate the 3D pose parameters, whereas the object model projected to the image plane helps in a top-down manner to improve the extraction of the contour and the region statistics. The present approach clearly states all model assumptions in a single energy functional. This keeps the model manageable and allows further extensions for the future. While common alternative segmentation approaches that integrate 2D shape knowledge face the problem that an object can look very different from various viewpoints, a 3D free form model ensures that for each view the model can perfectly fit the data in the image. Moreover, one solves the higher level problem of determining the object pose including its distance to the camera. Experiments demonstrate the performance of the method.

1 Introduction

Pose estimation and image segmentation are principal problems in computer and robot vision. The task of 2D-3D pose estimation is to estimate a rigid motion which fits a 3D object model to 2D image data [8]. In this context it is crucial which features are used for the object model as they must be fit to corresponding features in the image to determine the pose. One such feature is the object surface with the object silhouette as its 2D counterpart in the image. The task of pose estimation is to find a rigid motion that minimizes the error between the projected object surface and the region encircled by the contour in the image. As the common role of image segmentation is exactly to extract the contour of objects in the image, this shows the possible connection between 2D-3D pose estimation and image segmentation.

Image segmentation can become very difficult, as the image gray value or color alone are rarely good indicators for object boundaries due to noise, texture, shading, occlusion, or simply because the color of two objects is nearly the same. Recent segmentation approaches therefore integrate 2D shape information in order to employ additional

⋆ We gratefully acknowledge funding by the DFG projects We2602/1-1, We2602/1-2, Ro2497/1-1, and Ro2497/1-2.

constraints that force the contour to more desirable solutions. An early example can be found in [9] where shape information influences the evolution of an active contour model. This basic concept has been extended and modified in [16, 6, 13, 5] and provides a good framework for the sound integration of 2D shape prior in segmentation processes. However, the real world has three spatial dimensions. This fact is responsible for an inherent shortcoming of 2D shape models: they cannot exactly describe the image of an object from arbitrary views. This problem is solved when the 2D shape model is replaced by a 3D surface model, as is suggested in the present paper.

For the integration of 3D shape information, the object model has to be projected onto the image plane, and for this its pose in the scene has to be known. We realize again the connection between image segmentation and pose estimation, yet now the connection points into the other direction: a pose estimate is needed in order to integrate the surface model. Note that a pose estimation problem appears in the case of 2D shape knowledge as well. Also there, it is necessary to estimate the translation, rotation, and scaling of the shape knowledge, before it can constrain the contour in the image. This is either achieved by explicit estimation of the pose parameters [16], or by an appropriate normalization of the shapes [5]. Extensions to perspective transformations of 2D shapes have recently been proposed in [13]. However, all these approaches only aim on the use of shape knowledge in order to yield improved segmentations. The 2D pose estimates do not allow a location of the object in the real 3D world but only in the 2D projection of this world. In contrast, the 2D-3D pose estimation employed in our model allows the exact location of the object in the scene.

We now have a classical chicken-and-egg problem: a contour is needed for pose estimation, and the pose estimates are necessary to integrate the shape prior into the segmentation that determines the contour. Such situations are in general best handled by solving both problems simultaneously. We achieve this by formulating an energy minimization problem that contains both the image contour and the pose parameters as unknowns. The minimization is done by alternating both image segmentation and pose estimation in an iterative manner, see Fig. 1. For the experiments we concentrate on the segmentation and pose estimation of a rigid object. It is demonstrated that the model yields promising results both for the contour and the 3D pose parameters even in complex scenarios.

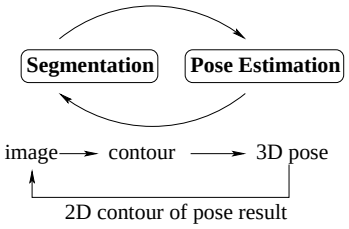

Fig. 1. Basic idea: Iterating segmentation and pose estimation. The projected pose result is used as a-priori knowledge for segmentation.

Paper organization. The next section contains a brief review of the level set based image segmentation model used in our approach. Section 3, on the other side, explains the concept of 2D-3D pose estimation. In Section 4 we then introduce our idea to combine image segmentation and 3D pose estimation in a joint energy functional. Experiments in Section 5 show the performance of the proposed technique. The paper is concluded by a brief summary in Section 6.

2 Image Segmentation

2.1 Level Set Formulation

Our approach is based on image segmentation with level sets [7, 11, 3, 12, 4], in particular on the method described in [1]. 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.
The segmentation should maximize 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. Further on, the boundary between both regions should be as small as possible. This can be expressed by the following energy functional:

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

where $\nu > 0$ is a weighting parameter 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. Minimization with respect to the region boundary can be performed according to the gradient descent equation

$$\partial_t \Phi = H'(\Phi)\left(\log\frac{p_1}{p_2} + \nu\,\mathrm{div}\left(\frac{\nabla\Phi}{|\nabla\Phi|}\right)\right) \tag{2}$$

where $H'(s)$ is the derivative of $H(s)$ with respect to its argument. The contour converges to a minimum for the numerical evolution parameter $t \to \infty$.

2.2 Region Statistics

For the curve evolution, still the probability densities p_1 and p_2 have to be determined. Our segmentation is driven by the texture feature space proposed in [2] which yields $M = 5$ feature channels I_j for gray scale images, and $M = 7$ channels if color is available. We assume that the probability densities of the feature channels are independent, thus $p_i = \prod_{j=1}^{M} p_{ij}(I_j)$.
The probability densities p_{ij} are estimated according to the *expectation-maximization principle*. Having the level set function initialized with some partitioning, the probability densities can be approximated by a Gaussian density estimate:

$$p_{ij}(s,x) \propto \frac{1}{\sqrt{2\pi}\sigma_{ij}(x)}\exp\left(\frac{(s-\mu_{ij}(x))^2}{2\sigma_{ij}(x)^2}\right). \tag{3}$$

Note that these are *local* estimates of the probability densities. This can be useful particularly in complicated scenes where differences between regions are only locally visible. Consequently, the parameters $\mu_{ij}(x)$ and $\sigma_{ij}(x)$ are computed in a local neighborhood K_ρ of x by:

$$\mu_{ij}(x) = \frac{\int_{\Omega_i} K_\rho(\zeta - x) I_j(\zeta)\,d\zeta}{\int_{\Omega_i} K_\rho(\zeta - x)\,d\zeta} \qquad \sigma_{ij}(x) = \frac{\int_{\Omega_i} K_\rho(\zeta - x)(I_j(\zeta) - \mu_{ij}(x))^2\,d\zeta}{\int_{\Omega_i} K_\rho(\zeta - x)\,d\zeta}. \tag{4}$$

The densities are used for the level set evolution according to (2), leading to a further update of the probability densities, and so on. This iterative process converges to a local

minimum, showing that the initialization matters. In order to attenuate this dependency on the initialization, it is recommendable to apply a coarse-to-fine strategy. Starting with a down-sampled image, there are less local minima, so the segmentation is more robust. The resulting segmentation can then be used as initialization for a finer scale, until the original segmentation problem is solved.

3 2D-3D Pose Estimation

2D-3D pose estimation [8] means to estimate a rigid body motion which maps a 3D surface model to an image of a calibrated camera. The scenario is visualized in Fig. 2.

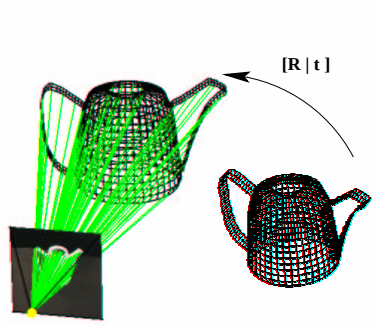

Fig. 2. The pose scenario: the aim is to estimate the pose $\boldsymbol{R}$, $\boldsymbol{t}$.

The core algorithm is based on a point-based constraint equation, which has been derived in the language of Clifford Algebras. We assume a set of point correspondences (X_i, x_i), with 4D (homogeneous) model points X_i and 3D (homogeneous) image points x_i. Each image point is reconstructed to a Plücker line $L_i = (n_i, m_i)$, with a (unit) direction n_i, and moment m_i [10]. The 3D rigid motion is represented as exponential form

$$\boldsymbol{M} = \exp(\theta\hat{\xi}) = \exp\begin{pmatrix} \hat{\omega} & \boldsymbol{v} \\ 0_{3\times 1} & 0 \end{pmatrix} \tag{5}$$

where $\theta\hat{\xi}$ is the matrix representation of a twist $\xi = (\omega_1, \omega_2, \omega_3, v_1, v_2, v_3) \in se(3) = \{(\boldsymbol{v}, \omega) | \boldsymbol{v} \in \mathbb{R}^3, \hat{\omega} \in so(3)\}$, with $so(3) = \{\boldsymbol{A} \in \mathbb{R}^{3\times 3} | \boldsymbol{A} = -\boldsymbol{A}^T\}$. In fact, $\boldsymbol{M}$ is an element of the one-parametric Lie group $SE(3)$, known as the group of direct affine isometries. 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 [10] for more details. The corresponding Lie algebra to $SE(3)$ is denoted as $se(3)$. A twist contains six parameters and can be scaled to $\theta\xi$ with a unit vector ω. 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. To reconstruct a group action $\boldsymbol{M} \in SE(3)$ from a given twist, the exponential function $\exp(\theta\hat{\xi}) = \boldsymbol{M} \in SE(3)$ must be computed. It can be calculated efficiently by using the Rodriguez formula [10],

$$\exp(\hat{\xi}\theta) = \begin{pmatrix} \exp(\theta\hat{\omega}) & (I - \exp(\hat{\omega}\theta))(\omega \times \boldsymbol{v}) + \omega\omega^T \boldsymbol{v}\theta \\ 0_{1\times 3} & 1 \end{pmatrix} \quad \text{for } \omega \neq 0 \tag{6}$$

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

$$\exp(\theta\hat{\omega}) = I + \hat{\omega}\sin(\theta) + \hat{\omega}^2(1 - \cos(\theta)). \tag{7}$$

Note that only sine and cosine functions of real numbers need to be computed.

For pose estimation we combine the reconstructed Plücker lines with the screw representation for rigid motions and apply a gradient descent method: Incidence of the transformed 3D point X_i with the 3D ray $L_i = (n_i, m_i)$ can be expressed as

$$(\exp(\theta\hat{\xi})X_i)_{3\times 1} \times n_i - m_i = 0. \tag{8}$$

Indeed, X_i is a homogeneous 4D vector, and after multiplication with the 4×4 matrix $\exp(\theta\hat{\xi})$ we neglect the homogeneous component (which is 1) to evaluate the cross product with n_i. Note, that this constraint equations expresses the perpendicular error vector between the Plücker line and the 3D point. The aim is to minimize this spatial error. Therefore we linearize the equation 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. This results in

$$((\boldsymbol{I} + \theta\hat{\xi})X_i)_{3\times 1} \times n_i - m_i = 0 \tag{9}$$

which can be reordered into an equation of the form $\boldsymbol{A}\xi = \boldsymbol{b}$. Collecting a set of such equations (each is of rank two) leads to an over-determined linear system of equations in ξ. The Rodriguez formula can be applied to reconstruct the group action $\boldsymbol{M}$ from the twist ξ. Then, the 3D points can be transformed and the process is iterated until the gradient descent approach converges. In recent years, this technique has been extended to higher order curves, free-form contours and free-form surfaces, see [14, 15]. The surface based pose estimation procedure is basically an ICP-algorithm, which has the problem to get trapped in local minima. For this reason we use a sampling method with different (neighboring) start poses and use the resulting pose with minimum error. This can be seen as a simple particle filter during pose estimation. Note that the constraint equations express a spatial distance measure in 3D. In [14] we have shown that each equation can be rescaled individually to an equivalent 2D distance measure. For combining segmentation and pose estimation we make use of this property to get a single energy functional.

4 Coupling Image Segmentation and 2D-3D Pose Estimation

In order to couple pose estimation and image segmentation in a joint optimization problem, the energy functional for image segmentation in (1) is extended by an additional term that integrates the object model:

$$\begin{aligned} E(\Phi, \theta\xi) = &- \int_\Omega \big(H(\Phi)\log p_1 + (1 - H(\Phi))\log p_2\big)\, dx + \nu \int_\Omega |\nabla H(\Phi)|\, dx \\ &+ \underbrace{\lambda \int_\Omega (\Phi - \Phi_0(\theta\xi))^2\, dx}_{\text{Shape}}\,. \end{aligned} \tag{10}$$

The quadratic error measure in the shape term has been proposed in the context of 2D shape priors, e.g. in [16]. The prior $\Phi_0 \in \Omega \to \mathbb{R}$ is assumed to be represented by the signed distance function. This means in our case, $\Phi_0(x)$ yields the distance of x to the silhouette of the projected object surface.

In detail, Φ_0 is constructed as follows: let X_S denote the set of points $\boldsymbol{X}$ on the object surface. Projection of the transformed points $\exp(\theta\xi)X_S$ into the image plane yields the set x_S of all (homogeneously scaled) 2D points x on the image plane that correspond to a 3D point on the surface model

$$x = P\exp(\theta\xi)\boldsymbol{X}, \qquad \forall \boldsymbol{X} \in X_S \tag{11}$$

Fig. 3. From left to right: (a) Initialization. (b) Segmentation result with object knowledge. (c) Pose result. (d) Segmentation result without object knowledge.

where P denotes a projection with known camera parameters. The level set function Φ_0 can then be constructed from x_S by setting $\Phi_0(x) = 1$ if $x \in x_S$, $\Phi_0(x) = -1$ otherwise, and applying the distance transform.

Note that the distance $(\Phi(x) - \Phi_0(x))^2$ is exactly the distance used in the pose estimation method. Given the contour Φ, the pose estimation method thus minimizes the shape term in (10). Minimizing (10) with respect to the contour Φ, on the other hand, leads to the gradient descent equation

$$\partial_t \Phi = H'(\Phi) \left(\log \frac{p_1}{p_2} + \nu \operatorname{div} \left(\frac{\nabla \Phi}{|\nabla \Phi|} \right) \right) + 2\lambda \left(\Phi_0(\theta\xi) - \Phi \right). \tag{12}$$

In order to minimize the total energy, an iterative approach is suggested: keeping the contour Φ fixed, the optimum pose parameters $\theta\xi$ are determined as described in Section 3 and yield the silhouette of the object model Φ_0. Retaining in the opposite way the pose parameters, (12) determines an update on the contour. Both iteration steps thereby minimize the distance between Φ and Φ_0. While the pose estimation method draws Φ_0 towards Φ, thereby respecting the constraint of a rigid motion, (12) in return draws the curve Φ towards Φ_0, thereby respecting the data in the image.

5 Experiments

Fig. 3 - 5 show tracking results with a tea box as object model and cluttered backgrounds. Fig. 3 demonstrates the advantage of integrating object knowledge into the segmentation process. Without object knowledge, parts of the tea box are neglected 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.

In Fig. 4, the motion of the object causes severe reflections on the metallic surface of the tea box. Nevertheless, the results remain stable. Further note some smaller occlusions due to the fingers that do not disturb the pose estimation.

In Fig. 5 the amount of occlusion is far more eminent. This experiment also demonstrates the straightforward extension of the method to multiple cameras. The non-occluded parts of both views provide enough information for pose recognition of the object.

However, we also do not want to conceal a decisive drawback of the method, this is the dependency of the result on the initialization. As the pose estimation of the object prior is based on the segmentation, the object model cannot help to initially find the

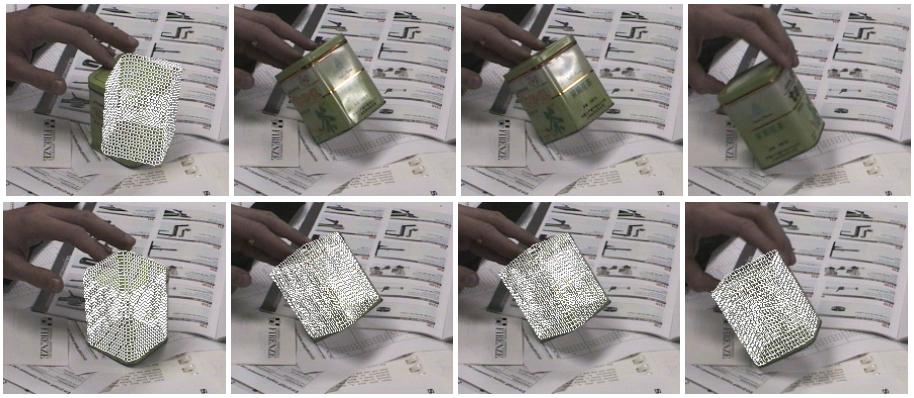

Fig. 4. Top row: Initialization at the first frame. Frames 49, 50, and 117 of the sequence. **Bottom row:** Tracking results at frames 0, 49, 50, and 117. The tea box is moved in 3D, causing partially severe reflections on the box.

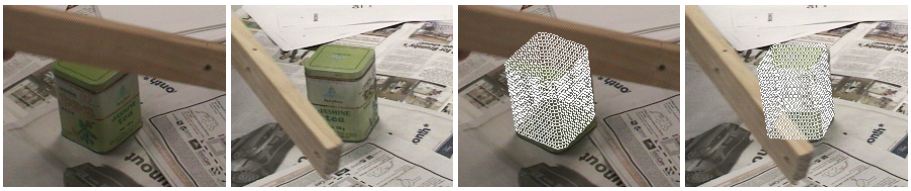

Fig. 5. Tracking result of a stereo sequence. In both views the object is partially occluded but the pose can be reconstructed from the remaining information (frame 98 from 210 frames).

object in the image. It can only improve the *tracking* of the object, once a good pose initialization has been found. How to find such an initialization automatically, i.e. how to *detect* objects in cluttered scenes, is a topic on its own.

6 Conclusion

We presented a technique that integrates 3D shape knowledge into a variational model for level set based image segmentation. While the utilization of 2D shape knowledge has been investigated intensively in recent time, the presented approach accommodates the three-dimensional nature of the world. The technique is based on a powerful image-driven segmentation model on one side, and an elaborated method for 2D-3D pose estimation on the other side. The integration of both techniques improves the robustness of contour extraction and, consequently, also the robustness of pose estimation that relies on the contour. It allows for the tracking of three-dimensional objects in cluttered scenes with inconvenient illumination effects. The strategy to model the segmentation in the image plane, whereas the shape model is given in three-dimensional space, has the advantage that the image-driven part can operate on its natural domain as provided by the camera, while the 3D object model offers the full bandwidth of perspective views. Moreover, in contrast to 2D techniques, it gives the extracted object a position in space.

References

1. T. Brox, M. Rousson, R. Deriche, and J. Weickert. Unsupervised segmentation incorporating colour, texture, and motion. In N. Petkov and M. A. Westenberg, editors, *Computer Analysis of Images and Patterns*, volume 2756 of *Lecture Notes in Computer Science*, pages 353–360. Springer, Berlin, 2003.
2. T. Brox and J. Weickert. A TV flow based local scale measure for texture discrimination. In T. Pajdla and J. Matas, editors, *Proc. 8th European Conference on Computer Vision*, volume 3022 of *Lecture Notes in Computer Science*, pages 578–590. Springer, Berlin, May 2004.
3. V. Caselles, F. Catté, T. Coll, and F. Dibos. A geometric model for active contours in image processing. *Numerische Mathematik*, 66:1–31, 1993.
4. T. Chan and L. Vese. An active contour model without edges. In M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, editors, *Scale-Space Theories in Computer Vision*, volume 1682 of *Lecture Notes in Computer Science*, pages 141–151. Springer, 1999.
5. D. Cremers, S. Osher, and S. Soatto. A multi-modal translation-invariant shape prior for level set segmentation. In C.-E. Rasmussen, H. Bülthoff, M. Giese, and B. Schölkopf, editors, *Pattern Recognition*, volume 3175 of *Lecture Notes in Computer Science*, pages 36–44. Springer, Berlin, Aug. 2004.
6. D. Cremers, F. Tischhäuser, J. Weickert, and C. Schnörr. Diffusion snakes: introducing statistical shape knowledge into the Mumford-Shah functional. *International Journal of Computer Vision*, 50(3):295–313, Dec. 2002.
7. A. Dervieux and F. Thomasset. A finite element method for the simulation of Rayleigh–Taylor instability. In R. Rautman, editor, *Approximation Methods for Navier–Stokes Problems*, volume 771 of *Lecture Notes in Mathematics*, pages 145–158. Springer, Berlin, 1979.
8. W. E. L. Grimson. *Object Recognition by Computer.* The MIT Press, Cambridge, MA, 1990.
9. M. E. Leventon, W. E. L. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours. In *Proc. 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 316–323, Hilton Head, SC, June 2000. IEEE Computer Society Press.
10. R. Murray, Z. Li, and S. Sastry. *Mathematical Introduction to Robotic Manipulation.* CRC Press, Boca Raton, FL, 1994.
11. S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *Journal of Computational Physics*, 79:12–49, 1988.
12. N. Paragios and R. Deriche. Unifying boundary and region-based information for geodesic active tracking. In *Proc. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 300–305, Forth Collins, Colorado, June 1999.
13. T. Riklin-Raviv, N. Kiryati, and N. Sochen. Unlevel-sets: geometry and prior-based segmentation. In T. Pajdla and J. Matas, editors, *Proc. 8th European Conference on Computer Vision*, volume 3024 of *Lecture Notes in Computer Science*, pages 50–61. Springer, Berlin, May 2004.
14. B. Rosenhahn. Pose estimation revisited. Technical Report TR-0308, Institute of Computer Science, University of Kiel, Germany, Oct. 2003.
15. B. Rosenhahn and G. Sommer. Pose estimation of free-form objects. In T. Pajdla and J. Matas, editors, *Proc. 8th European Conference on Computer Vision*, volume 3021 of *Lecture Notes in Computer Science*, pages 414–427, Prague, May 2004. Springer.
16. M. Rousson and N. Paragios. Shape priors for level set representations. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Computer Vision – ECCV 2002*, volume 2351 of *Lecture Notes in Computer Science*, pages 78–92. Springer, Berlin, 2002.