Robust 3D Pose Estimation and Efficient 2D Region-Based Segmentation from a 3D Shape Prior

Samuel Dambreville, Romeil Sandhu, Anthony Yezzi, and Allen Tannenbaum

Georgia Institute of Technology

Abstract. In this work, we present an approach to jointly segment a rigid object in a 2D image and estimate its 3D pose, using the knowledge of a 3D model. We naturally couple the two processes together into a unique energy functional that is minimized through a variational approach. Our methodology differs from the standard monocular 3D pose estimation algorithms since it does not rely on local image features. Instead, we use global image statistics to drive the pose estimation process. This confers a satisfying level of robustness to noise and initialization for our algorithm, and bypasses the need to establish correspondences between image and object features. Moreover, our methodology possesses the typical qualities of region-based active contour techniques with shape priors, such as robustness to occlusions or missing information, without the need to evolve an infinite dimensional curve. Another novelty of the proposed contribution is to use a unique 3D model surface of the object, instead of learning a large collection of 2D shapes to accommodate for the diverse aspects that a 3D object can take when imaged by a camera. Experimental results on both synthetic and real images are provided, which highlight the robust performance of the technique on challenging tracking and segmentation applications.

1 Motivation and Related Work

2D image segmentation and 2D-3D pose estimation are ubiquitous tasks in computer vision applications and have received a great deal of attention in the past few years. These two fundamental techniques are usually studied separately in the literature. In this work, we combine both approaches in a variational framework. To appreciate the contribution of this work, we recall some of the results and specifics of both fields.

2D-3D pose estimation aims at determining the pose of a 3D object relative to a calibrated camera from one unique or a collection of 2D images. By knowing the mapping between the world coordinates and image coordinates from the camera calibration matrix, and after establishing correspondences between 2D features in the image and their 3D counterparts on the model, it is then possible to solve the pose transformation (from a set of equations that express these correspondences). The literature concerned with 3D pose estimation is very large and a complete survey is beyond the scope of this paper. However, most methods can be distinguished by the type of *local* image features used to establish correspondences, such as points [1], lines or segments [2,3], multi-part curve segments [4], or complete contours [5,6].

Segmentation consists of separating an object from the background in an image. The geometric active contour (GAC) framework, in which a curve is evolved continuously to

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 169–182, 2008.

capture the boundaries of an object, has proven to be quite successful at performing this task. Originally, the method focused on extracting local image features such as edges to perform segmentation; see [7,8] and the references therein. However, edge-based techniques can suffer from the typical drawbacks that arise from using local image features: high sensitivity to noise or missing information, and a multitude of local minima that result in poor segmentations. Region-based approaches, which use global image statistics inside and outside the contour, were shown to drastically improve the robustness of segmentation results [9,10,11,12]. Region-based techniques are able to deal with various statistics of the object and background such as distinct mean intensities [10], Gaussian distributions [11,12] or intensity histograms [13,14] as well as a wide variety of photometric descriptors such as grayscale values, color or texture [15]. Further improvement of the GAC approach consists of learning the shape of objects and constrain the contour evolution to adopt familiar shapes, to make up for poor segmentation results obtained in the presence of noise, clutter, occlusion or when the statistics of the object and background are difficult to distinguish (see e.g., [16,17,18,19]).

Motivation/Contribution. Our goal is to combine the strengths of both techniques and to avoid some of their typical weaknesses, to robustly both segment 2D images and estimate the pose of an arbitrary 3D object which shape is known.

In particular, we use a region-based approach to continuously drive the pose estimation process. This global approach avoids using local image features and, hence, addresses two shortcomings that typically arise from doing so in many 2D-3D pose estimation algorithms: Firstly, finding the correspondence between local features in the image and on the model is a non-trivial task, due for instance to their viewpoint dependency - no local correspondences need to be found in our global approach. Secondly, local image features may not even exist or can be difficult to detect in a reliable and robust fashion in the presence of noise clutter or occlusion. Furthermore, simplifying assumptions usually need to be made on the class of shapes that a 2D-3D pose estimation technique can handle. Many approaches are limited to simple shapes that can be described using geometric primitives such as corners, lines, circles or cylinders. Recent work focused on free-form objects, which admit a manageable parametric description as in [5]. However, even this type of algebraic approaches can become unmanageable for objects of arbitrary and complex shape. Our approach can deal with rigid object of *arbitrary* shape, represented by a 3D level set [20] or a 3D cloud of points (Figure 1).

Conversely, a shortcoming of the GAC framework using shape priors is that 2D shapes are usually learned to segment 2D images. Hence, a large collection of 2D shapes needs to be learned to represent the wide variation in aspect that most natural 3D objects take, when projected onto the 2D image plane. Our region-based approach benefits from the knowledge of the object shape that is compactly described by a *unique* 3D model. Acquisition of 3D models can be readily accomplished using range scans [21] or structure from motion approaches [22], notably. In addition, and in contrast to the GAC framework, the proposed method does not involve the evolution of an infinite dimensional contour to perform segmentation, but only solves for the finite dimensional pose parameters (as is common for 2D-3D pose estimation approaches).

This results in a much simplified framework that avoids dealing with problems such as infinite dimensional curve representation, evolution and regularization .

Relation to Previous Work. In this paper, we exploit many ideas from recent variational approaches that address the problem of structure from motion and stereo reconstruction from multiple cameras ([22,23] or [24]). Originally, the authors in [22,23] presented a method to reconstruct the 3D shape of an object from multiple 2D views obtained from calibrated cameras. The present contribution aims at performing a somewhat opposite task: given the 3D model of an object, perform the segmentation of 2D images and recover the 3D pose of the object relative to a *unique* camera. This is the first time that the framework in [22,23] is adapted and employed in the specific context of segmenting 2D images from a unique camera, using the knowledge of a 3D model. The framework in [22] has also recently been extended in [25] to address the problem of multiple camera calibration. In the present work, the camera is assumed to be calibrated. However, this assumption could easily be dropped by also solving for the optimal camera calibration parameters as presented in [25].

We note that, although the use of 3D shape knowledge to perform the 2D segmentation of regions presents obvious advantages, the literature dealing with this type of approaches is strikingly thin. The pieces of work closest to the proposed contribution are probably [26] and [27]. In [26], the authors evolve an (infinite dimensional) active contour as well as 3D pose parameters to minimize a joint energy functional encoding both image information and 3D shape knowledge. Our method differs from the aforementioned approach in many crucial aspects: We optimize a *unique* energy functional, which allows us to circumvent the need to determine ICP-like correspondences and to perform costly back-projections between the segmenting contour and the shape model at each iteration. Also, we perform optimization *only* in the finite dimensional space of the Euclidean pose parameters. In addition to being computationally efficient, this allows our technique to be less likely to be trapped in local minima, resulting in robust performances as demonstrated in the experimental part. In [27], the method of [26] is successfully simplified by performing energy minimization only in the space of 3D pose parameters. Thus, the method of [27] and our contribution present some similarities. However, the energy minimization method and resulting algorithms are radically different: In [27], an algebraic approach is used that involves establishing correspondences and back-projections between the 3D and 2D world, as well as linearizing the resulting system of equations. Consequently, important information about the geometry of the 3D model is lost through the algebraic approach. In contrast, our approach relies on surface differential geometry (see e.g., [28]) to link geometric properties of the model surface and its projection in the image domain. This allows us to derive the partial differential equations necessary to perform energy optimization, as well as exploit the knowledge of the 3D object to its full extent.

Our technique uses a 3D shape prior in a region-based framework, and can thereby be expected to be robust to noise or occlusion. Hence, an obvious application of the proposed approach is the robust tracking of 3D rigid objects in 2D image sequences. Our approach is, thus, also related to a wealth of methods concerned with the problem of model-based monocular tracking (see [29] for a recent survey).

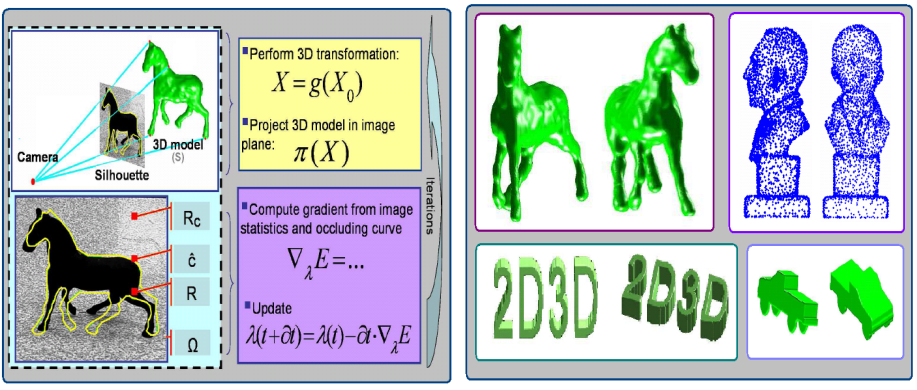


Fig. 1. *Left:* Schema summarizing our segmentation/pose estimation approach from a 3D model, in 4 steps. *Right:* Different views of the 3D models used (rendered surfaces or cloud of points).

2 Proposed Approach

We suppose we have at our disposal the 3D surface model of an object. Our goal is to find the (Euclidean) transformation that needs to be applied to the model so that it coincides with the object of interest in the referential attached to a calibrated camera. We now describe the proposed approach in details, starting with our choice of notation. An overview of the method can be found in Figure 1.

2.1 Notation

Let $\mathbf{X} = [X, Y, Z]^T$ denote the coordinates of a point in $\mathbb{R}^3$, measured with respect to a referential attached to the imaging camera. We denote by I the image, by $\Omega \subset \mathbb{R}^2$ the image domain, and by $d\Omega$ its area element. We assume the camera is modeled as an ideal perspective projection[1]: $\pi : \mathbb{R}^3 \mapsto \Omega; \mathbf{X} \mapsto \mathbf{x}$, where $\mathbf{x} = [x, y]^T = [X/Z,\ Y/Z]^T$ denotes coordinates in Ω.

Let S be the smooth surface in $\mathbb{R}^3$ defining the shape of the object of interest. The (outward) unit normal to S at each point $\mathbf{X} \in S$ will be denoted by $\mathbf{N} = [N_1, N_2, N_3]^T$. To determine the pose of S with respect to the camera, we define the identical reference surface S_0, whose pose is known.[2] Denoting by X_0 the coordinates of points on S_0, one can locate S in the camera referential via the transformation $g \in SE(3)$, such that $S = g(S_0)$, or written point-wise $\mathbf{X} = g(\mathbf{X}_0) = \mathbf{R}\mathbf{X}_0 + \mathbf{T}$, with $\mathbf{R} \in SO(3)$ and $\mathbf{T} \in \mathbb{R}^3$. The parameters of the rigid motion g will be denoted by $\lambda = [\lambda_1, \ldots, \lambda_6]^T = [t_x, t_y, t_z, \omega_1, \omega_2, \omega_3]^T$ (rotations are represented in exponential coordinates [32]).

Let $R = \pi(S) \subset \Omega$, be the region of the image on which the surface S projects (i.e., the region of Ω corresponding to imaging S). Let $R^c = \Omega \setminus R$ and $\hat{c} = \partial R$ denote the

[1] More general models of cameras (see [30,31]) can straightforwardly be handled. We make this assumption here to simplify the presentation.

[2] One can assume that the center of gravity of S_0 coincides with the camera center and the rotation is known.

complement and the boundary of R, respectively (Figure 1). The curve $\hat{c} \subset \Omega$ is the projection of the curve $C \subset S$ that delineates the visible part of S from the camera: $\hat{c} = \pi(C)$. The 2D curve $\hat{c}$ and 3D curve C will be respectively referred to as the "silhouette" and the "occluding curve".

2.2 Energy Functional

In [22], the authors used an image formation approach to define a cost functional measuring the discrepancy between the photometric properties of the surface S (as well as the 3D-background), and the pixel intensities of multiple images. The resulting energy involved back-projections to the surface S to guarantee the coherence between the measurements obtained from multiple cameras. In this present work, we are interested in segmenting a *unique* image and adopt a somewhat different approach directly inspired from region-based active contours techniques [10,11,12,13]. Most region-based approaches assume that the pixels corresponding to the object of interest or the background are distinct with respect to a certain grouping criterion. To perform segmentation, a closed curve is evolved to increase the discrepancy between the statistics of the pixels located in its interior and exterior. Accordingly, we define an energy of the form:

$$E = \int_{\mathrm{R}} r_O\big(I(\mathbf{x}), \hat{c}\big) d\Omega + \int_{\mathrm{R}^c} r_B\big(I(\mathbf{x}), \hat{c}\big) d\Omega, \tag{1}$$

where $r_O : \mathcal{Z}, \Omega \mapsto \mathbb{R}$ and $r_B : \mathcal{Z}, \Omega \mapsto \mathbb{R}$ are two monotonically decreasing functions measuring the matching quality of the image pixels with a statistical model over the regions R and R^c, respectively. The space $\mathcal{Z}$ corresponds to the photometric variable (such as gray-scale intensity, color or texture vector) chosen to perform segmentation. Hence, depending on the choices for r_O, r_B, and $\mathcal{Z}$ a larger class of images than the ones fitting the specific hypotheses made in [22] can be dealt with. The energy E measures the discrepancy between the statistical properties of the pixels located inside and outside the curve $\hat{c}$, and does not involve any back-projections.

It is expected that E is minimal when R and R^c correspond to the object and background in I, respectively. Most region-based approaches evolve an infinite dimensional curve, which amounts to exploring unconstrained shapes of the segmenting contour. Since we assume that the 3D shape of the rigid object is known, we want to minimize E by exploring only the possible regions R and R^c that result from projecting the surface S onto the image plane. For a calibrated camera, these regions are functions of the transformation g only. Solving for the transformation that minimize E can be undertaken via gradient descent over the parameters λ, as described below.

2.3 Gradient Flow

The partial differentials of E with respect to the pose parameters λ_i's can be computed using the chain-rule:

$$\begin{aligned}\frac{dE}{d\lambda_i} = \int_{\hat{c}} \Big(r_O\big(I(\mathbf{x})\big) - r_B\big(I(\mathbf{x})\big) \Big) \left\langle \frac{\partial \hat{c}}{\partial \lambda_i}, \hat{\mathbf{n}} \right\rangle d\hat{s} + \int_R \left\langle \frac{\partial r_O}{\partial \hat{c}}, \frac{\partial \hat{c}}{\partial \lambda_i} \right\rangle d\Omega \\ + \int_{R^c} \left\langle \frac{\partial r_B}{\partial \hat{c}}, \frac{\partial \hat{c}}{\partial \lambda_i} \right\rangle d\Omega,\end{aligned} \tag{2}$$

where $\hat{s}$ is the arc-length parametrization of the silhouette $\hat{c}$ and $\hat{\mathbf{n}}$ the (outward) normal to the curve at $\mathbf{x}$.

Using the arc-length s of C and the direct $\frac{\pi}{2}$-rotation matrix J, one has

$$\left\langle \frac{\partial \hat{c}}{\partial \lambda_i}, \hat{\mathbf{n}} \right\rangle d\hat{s} = \left\langle \frac{\partial \hat{c}}{\partial \lambda_i}, J\frac{\partial \hat{c}}{\partial \hat{s}} \right\rangle d\hat{s} = \left\langle \frac{\partial \pi(C)}{\partial \lambda_i}, J\frac{\partial \pi(C)}{\partial s} \right\rangle \frac{ds}{d\hat{s}} d\hat{s} = \left\langle \frac{\partial \pi(C)}{\partial \lambda_i}, J\frac{\partial \pi(C)}{\partial s} \right\rangle ds. \tag{3}$$

Let $\mathcal{J}$ denote the Jacobian of $\pi(\mathbf{X})$ with respect to the spatial coordinates, one has $\mathcal{J} = \frac{1}{Z^2}\begin{bmatrix} Z & 0 & -X \\ 0 & Z & -Y \end{bmatrix}$. From (3), one gets

$$\begin{aligned} \left\langle \frac{\partial \hat{c}}{\partial \lambda_i}, \hat{\mathbf{n}} \right\rangle d\hat{s} &= \left\langle \mathcal{J}\frac{\partial \mathbf{X}}{\partial \lambda_i}, J\mathcal{J}\frac{\partial \mathbf{X}}{\partial s} \right\rangle ds = \left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \mathcal{J}^T J \mathcal{J}\frac{\partial \mathbf{X}}{\partial s} \right\rangle ds \\ &= \frac{1}{Z^3}\left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \begin{bmatrix} 0 & Z & -Y \\ -Z & 0 & X \\ Y & -X & 0 \end{bmatrix} \frac{\partial \mathbf{X}}{\partial s} \right\rangle ds = \frac{1}{Z^3}\left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \frac{\partial \mathbf{X}}{\partial s} \times \mathbf{X} \right\rangle ds. \end{aligned} \tag{4}$$

In the equation above, the point $\mathbf{X}$ belongs to the occluding curve C. A necessary condition for a point $\mathbf{X}$ to belong to the occluding curve is that $\langle \mathbf{X}, \mathbf{N} \rangle = 0$ (since the associated vector $\mathbf{X}$, with origin at the center of the camera, corresponds to the projection/viewing direction and is tangent to the surface S at $\mathbf{X}$). The vector $\mathbf{t} = \frac{\partial \mathbf{X}}{\partial s}$ is the tangent to the curve C at the point $\mathbf{X}$. Since the vectors $\mathbf{t}$ and $\mathbf{X}$ belong to the tangent plane to S at $\mathbf{X}$, one has $\frac{\partial \mathbf{X}}{\partial s} \times \mathbf{X} = \|\mathbf{X}\|\mathbf{N}\sin(\theta)$, with $\theta = \widehat{(\mathbf{t}, \mathbf{X})}$ the angle between $\mathbf{t}$ and $\mathbf{X}$. For $\mathbf{X} \in C$, one has

$$\frac{\partial}{\partial s}\langle \mathbf{X}, \mathbf{N} \rangle = \underbrace{\left\langle \frac{\partial \mathbf{X}}{\partial s}, \mathbf{N} \right\rangle}_{=0} + \left\langle \frac{\partial \mathbf{N}}{\partial s}, \mathbf{X} \right\rangle = 0 = \langle d\mathbf{N}(\mathbf{t}), \mathbf{X} \rangle = \mathrm{II}(\mathbf{t}, \mathbf{X}).$$

Hence, since the second fundamental form $\mathrm{II}(\mathbf{t}, \mathbf{X}) = 0$, the vectors $\mathbf{t}$ and $\mathbf{X}$ are conjugate (see [28]). Hence, using the Euler formula, one can show that $K\sin^2\theta = \kappa_X \kappa_t$, where K is the Gaussian curvature, and κ_X and κ_t denote the normal curvatures in the directions $\mathbf{X}$ and $\mathbf{t}$ at $\mathbf{X} \in S$, respectively. Plugging into Equation (4), one gets

$$\left\langle \frac{\partial \hat{c}}{\partial \lambda_i}, \hat{\mathbf{n}} \right\rangle d\hat{s} = \frac{\|\mathbf{X}\|}{Z^3}\sqrt{\frac{\kappa_X \kappa_t}{K}} \left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \mathbf{N} \right\rangle ds. \tag{5}$$

The energy in (1) is valid for most functions r_O and r_B used in the literature. However, in this paper, we chose

$$r_O = -\log(\sigma_O) - \frac{(I(\mathbf{x}) - \mu_O)^2}{\Sigma_O} \quad \text{and} \quad r_B = -\log(\sigma_B) - \frac{(I(\mathbf{x}) - \mu_B)^2}{\Sigma_B} \tag{6}$$

as in the region-based active contour technique presented in [11]. Using these functions, the two last terms in Equation (2) collapse for $\mu_{\mathrm{O/B}} = \frac{\int_{R/R^c} I(\hat{\mathbf{x}}) d\Omega}{\int_{R/R^c} d\Omega}$ and $\Sigma_{\mathrm{O/B}} = \frac{\int_{R/R^c} \left(I(\hat{\mathbf{x}}) - \mu_{\mathrm{O/B}}\right)^2 d\Omega}{\int_{R/R^c} d\Omega}$.[3]

[3] For gray-scale images $\mu_{\mathrm{O/B}}$ and $\Sigma_{\mathrm{O/B}}$ are scalars. For color images, $\mu_{\mathrm{O/B}} \in \mathbb{R}^3$ and $\Sigma_{\mathrm{O/B}} \in \mathbb{R}^{3\times 3}$. Texture can also be used, see [15].

Thus, the flow becomes a simple line integral on C

$$\frac{dE}{d\lambda_i} = \int_C \Big(r_O\big(I(\pi(\mathbf{X}))\big) - r_B\big(I(\pi(\mathbf{X}))\big) \Big) \cdot \frac{\|\mathbf{X}\|}{Z^3} \sqrt{\frac{\kappa_X \kappa_t}{K}} \left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \mathbf{N} \right\rangle ds. \quad (7)$$

For $i = 1, 2, 3$ (i.e., λ_i is a translation parameter)

$$\left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \mathbf{N} \right\rangle = \left\langle \frac{\partial \mathbf{R}\mathbf{X_0} + \mathbf{T}}{\partial \lambda_i}, \mathbf{N} \right\rangle = \left\langle \frac{\partial \mathbf{T}}{\partial \lambda_i}, \mathbf{N} \right\rangle = N_i.$$

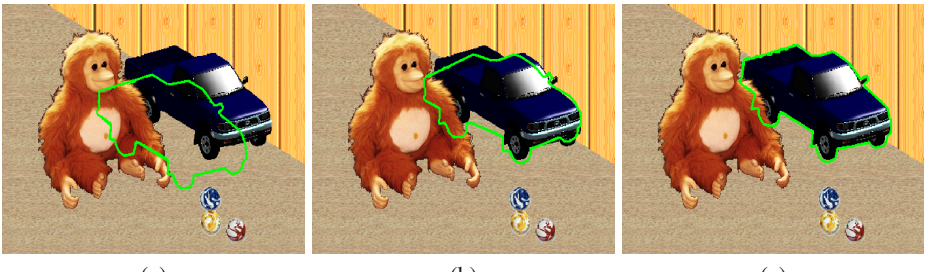
For $i = 4, 5, 6$ (i.e., λ_i is a rotation parameter), one can similarly show that $\left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \mathbf{N} \right\rangle = \langle \mathcal{M}_i, \mathbf{N} \rangle$, with $\mathcal{M}_i$ the i^{th} column of $\mathcal{M} = \mathbf{R} \begin{bmatrix} 0 & Z_0 & -Y_0 \\ -Z_0 & 0 & X_0 \\ Y_0 & -X_0 & 0 \end{bmatrix}$.

3 Experiments

We now report a few experimental results obtained for both synthetic and real datasets. Four different 3D models of rigid objects (see Figure 1) were used to perform segmentation and tracking tasks that highlight the robustness of our technique to **initialization**, **noise** and **missing or imperfect information**. To save computational time, we used the approximation $\sqrt{\frac{\kappa_X \kappa_t}{K}} \simeq 1$ in our implementation of Equation (7), which still decreased the energy E. Also, we note that the shape of the objects, notably the Horse and the Van Gogh Bust, cannot readily be described in terms of geometric primitives (lines, ellipses, etc.) or even algebraically, and thus do not satisfy the working hypotheses of standard pose estimation techniques ([2,3,5,6]).

3.1 Robustness to Initialization

Figure 2 shows segmentation results (and 3D coordinates recoveries) obtained using our approach for a synthetic color image. Figure 3 shows results for diverse natural color images. Despite initializations that are quite far from the truth (e.g., large errors in translation or angular position) accurate segmentations are obtained.



(a) (b) (c)

Fig. 2. Robustness to initialization - Segmentation of a synthetic color image. (a): Initialization; (b): Intermediate step of the evolution; (c): Final result.

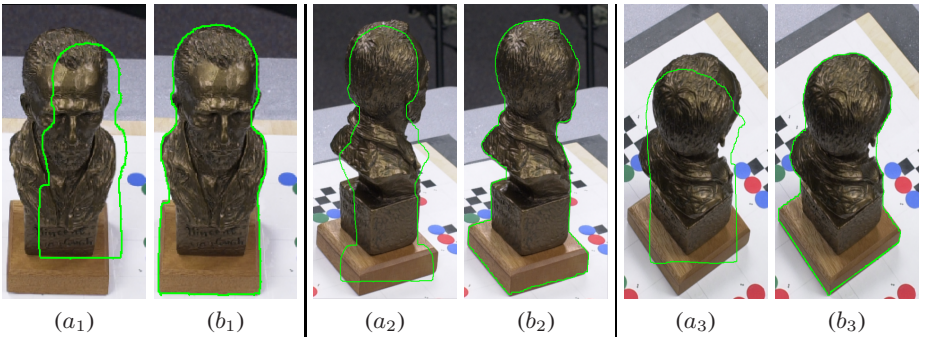


Fig. 3. Robustness to initialization - Segmentation of natural color images. (a_n)'s: Challenging initializations (e.g., large error in translation or angular positions - green curve). $(b_n$'s): Final results with the proposed approach (green curve).

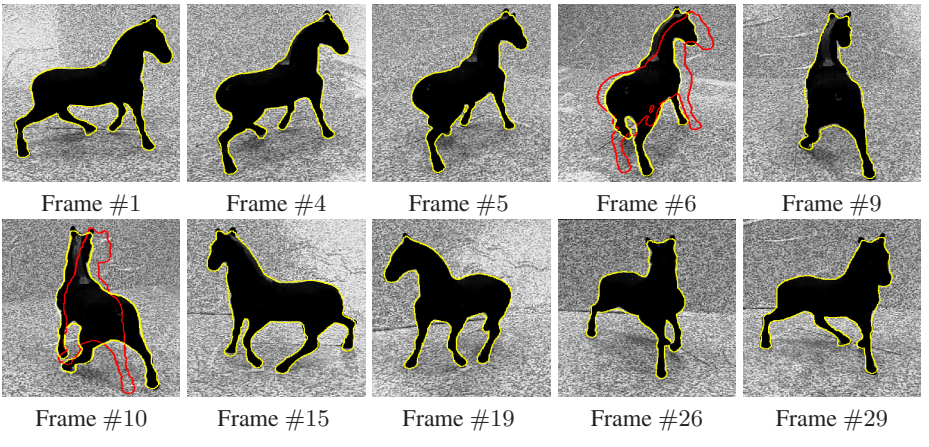


Fig. 4. Robustness to initialization - Tracking a natural sequence *Yellow contours*: final results after convergence; *Red contours*: Initializations from the result of the preceding image (see text for our tracking scheme). The aspect of the object changes drastically throughout the sequence. The position of the object undergoes large changes between successive images.

Figure 4 shows tracking results obtained for a real sequence. The sequence is composed of 32 images of a rigid toy Horse. The images were taken from discrete positions of a calibrated camera that underwent a complete rotation around the object. The camera "jumps" between successive images creating large changes in the pose of the object that need to be recovered (e.g., changes in the angular position of the camera can exceed 15° between frames). Tracking this sequence would be challenging for many 3D pose estimation techniques available in the literature: Many techniques using local features such as points or edges (e.g., [1,3]) are likely to be thrown off by the textured/noisy background (false features) and get trapped in local minima. The sequence was tracked with our technique, using a very simple scheme: For each image, initialization was performed using the pose parameters corresponding to the minimum of the energy

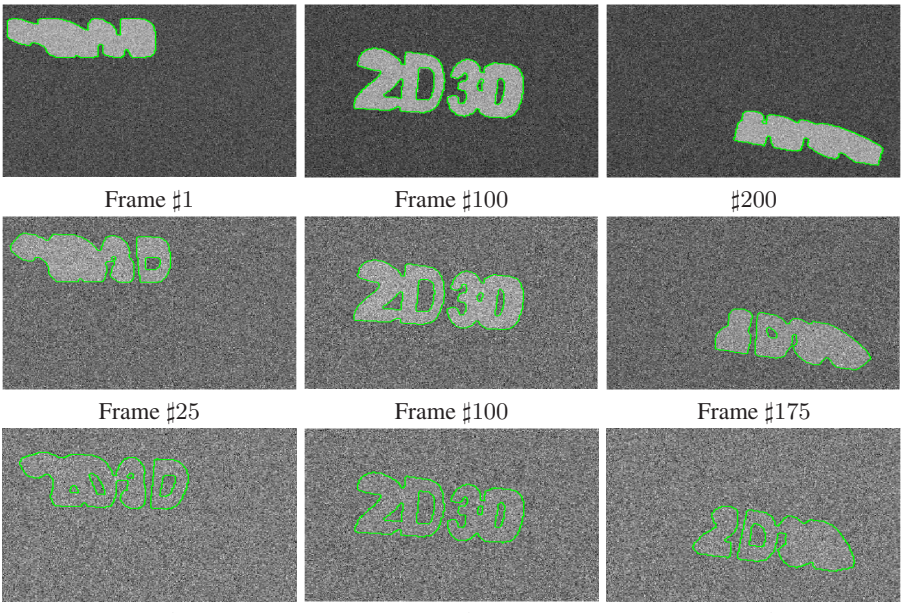


Fig. 5. Robustness to noise. Visual tracking results for the sequences involving the "2D3D" logo (Green curves). *First row:* Tracked sequence with Gaussian noise of standard deviation $\sigma = 10\%$. *Second row:* Tracked sequence for $\sigma = 30\%$. *Third row:* Tracked sequence for $\sigma = 100\%$.

obtained for the preceding image, and our approach was run until convergence. Despite the difficulties described above, very satisfying tracking performances were observed. This highlights the robustness of the technique to initialization since the large cameras jumps are accommodated and the method is not trapped in local minima. We note that region-based active contour techniques such as [10], would lead to satisfying segmentations on this particular sequence. However, these approaches would not also determine the pose of the object, which is valuable information for tracking applications.

3.2 Robustness to Noise

To test the robustness of our technique to noise, a sequence of 200 images was constructed by continuously transforming the 3D model of the "2D3D" logo and projecting it into the image plane using the parameters of a simulated calibrated camera (e.g., focal length f=200). The translation parameters, rotation axis and angle were continuously varied (e.g., the total angle variation over the sequence exceeded $160°$) to ensure a large variation of the aspect and position of the object throughout the sequence. From the basic sequence obtained, diverse level of Gaussian noise were added of standard deviation ranging from $\sigma = 10\%$ to $\sigma = 100\%$ (See Figure 5).

Typical visual results obtained using our approach (and the tracking scheme above) are reproduced in Figure 5. For all noise levels, which can be rather large (e.g., in the case $\sigma = 100\%$ object and background are barely distinguishable), tracking was

Noise Level	**mean error** (in %)	**std. dev. error** (in %)	**max error** (in %)
$\sigma = 10\%$	**T**: 0.85 ; **R**: 0.96	**T**: 0.23 ; **R**: 0.45	**T**: 1.43 ; **R**: 2.60
$\sigma = 30\%$	**T**: 0.97 ; **R**: 1.09	**T**: 0.21 ; **R**: 0.47	**T**: 1.50 ; **R**: 2.94
$\sigma = 60\%$	**T**: 0.95 ; **R**: 1.30	**T**: 0.30 ; **R**: 0.52	**T**: 2.39 ; **R**: 2.60
$\sigma = 100\%$	**T**: 1.02 ; **R**: 2.12	**T**: 0.39 ; **R**: 0.87	**T**: 2.18 ; **R**: 4.36

Fig. 6. Robustness to noise. Quantitative tracking results for the "2D3D" sequences with diverse levels of noise. Table displaying %-**absolute** error statistics over the 200 images of the sequences.

maintained throughout the whole sequence. Figure 6 reproduces the results of the pose estimation procedure. For each image, percent *absolute* errors with respect to the ground-truth were computed for both the translation and rotation as: $\text{Error} = \frac{\|\mathbf{v}_{\text{measured}} - \mathbf{v}_{\text{truth}}\|}{\|\mathbf{v}_{\text{truth}}\|}$, with $\mathbf{v}$ a translation or quaternion (see [32]) vector. From the pose estimation point of view, the method appears to perform quite well: average error and standard deviation computed over the 200 frames of each sequence rarely exceed 2% and 1%, respectively, for *both* translation and rotation. This highlights the accuracy and reliability of the method, and suggests that it is quite resilient to large amounts of noise (very little deterioration of the results is observed with increasing noise levels).

3.3 Robustness to Missing/Imperfect Image Information

To test the robustness to missing information, we created two sequences by adding two different occlusions in the basic sequence featuring the "2D3D" model (see Figure 7). The first occlusion is a grey rectangle that can mask more than $2/3$ of the "2D3D" logo. The second occlusion is the word "SHAPE" written in black letters that can mask the object at several places. Gaussian noise of standard deviation 30% was also added to both resulting sequences. Figure 7 presents the results of tracking the sequences of 200

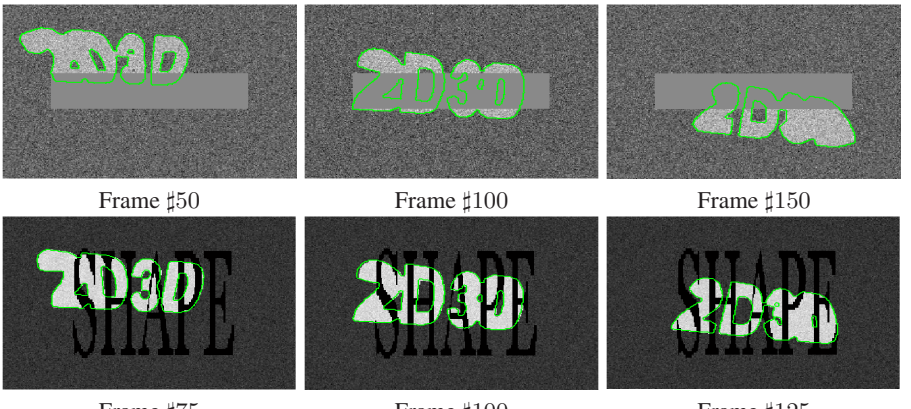


Fig. 7. Robustness to missing information. Tracking results (green curves) for the "2D3D" sequences with occlusions. *First row:* Rectangular occlusion. *Second row:* Word "SHAPE" as occlusion. Gaussian noise with $\sigma = 30\%$ was added.

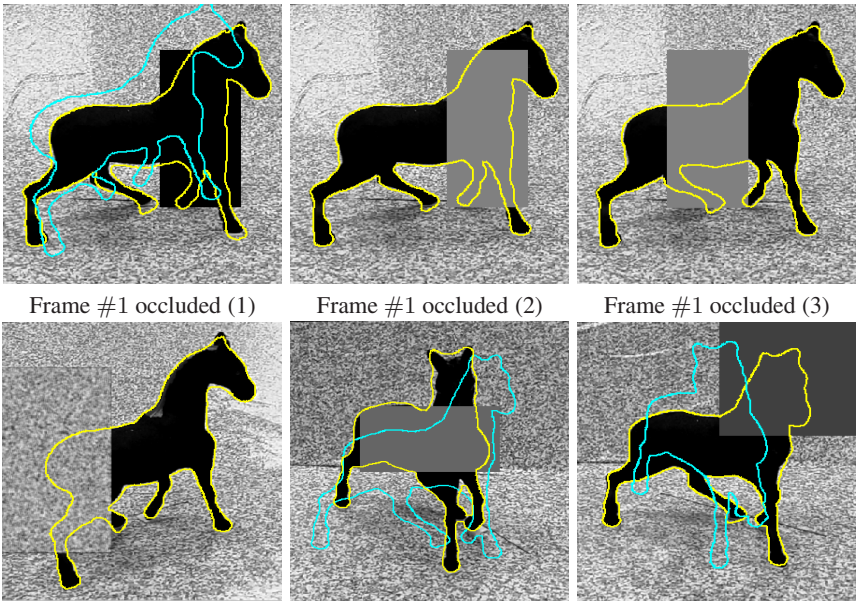


Fig. 8. Robustness to missing information. Segmentation results with occlusions. *Cyan contours*: Some of the initializations tested (note the large errors in angular position); *Yellow contours*: Final results (almost identical to results in Figure 4 with no occlusion).

frames with our approach. One notes that despite the occlusions (and noise), accurate segmentations are obtained: In particular, missing letters or parts are accurately localized and reconstructed. Track was maintained throughout both sequences. For the first sequence mean %-absolute-error (over the 200 frames) in the transformation parameters was 1.08% for translation ($\mathbf{T}$) and 1.57% for rotation ($\mathbf{R}$) - standard deviation 0.45% for $\mathbf{T}$ and 0.75% for $\mathbf{R}$. For the second sequence mean %-absolute-error was 0.87% for $\mathbf{T}$ and 1.19% for $\mathbf{R}$ (std. dev. 0.34% for $\mathbf{T}$ and 0.53% for $\mathbf{R}$).

In Figure 8, we used images extracted from the Horse sequence and occluded different parts of the Horse body (e.g., the legs that have valuable information about its angular position). Diverse pose parameters quite far from the truth were used as initializations (e.g., angular position could be off by more than 30°). Despite the occlusions with various pixel intensities or texture (and poor initializations), very convincing segmentations were obtained. Also, the positions of the object in the camera referential were accurately recovered. As can be noticed by comparing with Figure 4, the results in the presence of occlusion are very comparable to the ones without occlusion.

In Figure 9, we present segmentation results where the background and object are difficult to distinguish based on pixel statistics only (due to specularities on the object and similar colors in object and background). The results obtained with the (infinite dimensional) active contour flow of [11], which is the region-based segmentation technique underlying our approach, are not satisfying since the contour leaks into the background. Robust results are obtained using our approach.

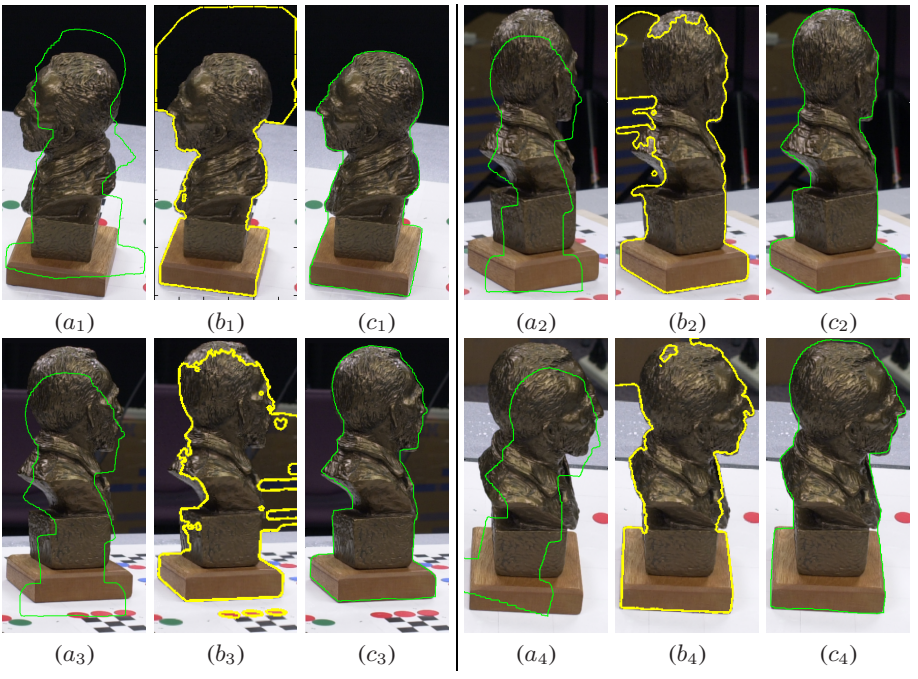


Fig. 9. Robustness to imperfect information. Comparative segmentation results. (a_n's): Initializations. (b_n's): Final results obtained with (infinite dimensional) active contour flow as in [11], which is the region-based segmentation technique underlying our approach. (c_n's): Final results with our approach. In these images, statistical distinction between object and background is difficult due to specularities on the object and similar colors in object and background.

The experiments of Figures 7, 8 and 9 would pose a major challenge to most region-based active contour techniques, even using shape priors [16,18,19]: Statistics only are not sufficient to segment the images, and the aspect of the object changes drastically from one image to the other. Hence, a large catalogue of 2D shapes would need to be learnt to achieve similar performances using the method in [16,18,19], for instance.

4 Conclusion and Future Work

In this work, we presented a region-based approach to the 3D pose estimation problem. This approach differs from other 3D pose estimation algorithms since it does not rely on local image features. Our method allows one to employ global image statistics to drive the pose estimation process. This confers a satisfying level of robustness to noise and initialization to our framework and bypasses the need to establish correspondences between image and object features, contrary to most 3D pose estimation approaches.

Furthermore, the approach possesses the typical qualities of a region-based active contour technique with shape prior, such as robustness to occlusion or missing information, without the need to evolve an infinite dimensional contour. Also, the prior

knowledge of the shape of the object is compactly represented by a unique 3D model, instead of a dense catalogue of 2D shapes.

The main advantage of the proposed technique is that it enables to locate the object not only in 2D images (a typical task handled by GAC approaches) but also in the world (a typical task handled by 2D-3D pose estimation algorithms). This makes the method particularly suitable for tracking applications involving a unique calibrated camera.

A possible direction for future research is to extend the proposed approach to include the knowledge of multiple 3D shapes. The method in [17] (where evolution of parameters in the shape space is performed in addition to pose parameters) could be adapted to the problem at hand. It is expected that the resulting framework will allow to learn the possible deformations of the object and lead to robust performances for non-rigid registration and tracking tasks.

Acknowledgement. This work was supported in part by grants from NSF, AFOSR, ARO, MURI, as well as by a grant from NIH (NAC P41 RR- 13218) through Brigham and Womens Hospital. This work is part of the National Alliance for Medical Image Computing (NAMIC), funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54 EB005149. Information on the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics. Allen Tannenbaum is also supported with a Marie Curie Grant through the Technion, Israel.

References

1. Quan, L., Lan, Z.D.: Linear n-point camera pose determination. IEEE Transactions on Pattern Analysis and Machine Intelligence 21, 774–780 (1999)
2. Dhome, M., Richetin, M., Lapreste, J.T.: Determination of the attitude of 3d objects from a single perspective view. IEEE Trans. Pattern Anal. Mach. Intell. 11, 1265–1278 (1989)
3. Marchand, E., Bouthemy, P., Chaumette, F.: A 2d-3d model-based approach to real-time visual tracking. Image and Vision Computing 19, 941–955 (2001)
4. Zerroug, M., Nevatia, R.: Pose estimation of multi-part curved objects. In: ISCV 1995: Proceedings of the International Symposium on Computer Vision, p. 431 (1995)
5. Rosenhahn, B., Perwass, C., Sommer, G.: Pose estimation of free-form contours. IJCV 62, 267–289 (2005)
6. Drummond, T., Cipolla, R.: Real-time tracking of multiple articulated structures in multiple views. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 20–36. Springer, Heidelberg (2000)
7. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. IJCV, 22, 61–79 (1997)
8. Kichenassamy, S., Kumar, S., Olver, P., Tannenbaum, A., Yezzi, A.: Conformal curvature flow: From phase transitions to active vision. Archives for Rational Mechanics and Analysis 134, 275–301 (1996)
9. Zhu, S.C., Yuille, A.L.: Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE Trans. PAMI 18, 884–900 (1996)
10. Chan, T., Vese, L.: Active contours without edges. IEEE TIP 10, 266–277 (2001)
11. Paragios, N., Deriche, R.: Geodesic active regions: A new paradigm to deal with frame partition problems in computer vision. Journal of Visual Communication and Image Representation 13, 249–268 (2002)

12. Dambreville, S., Yezzi, A., Niethammer, M., Tannenbaum, A.: A variational framework combining level-sets and thresholding. In: BMVC, pp. 266–280 (2007)
13. Michailovich, O., Rathi, Y., Tannenbaum, A.: Image segmentation using active contours driven by the bhattacharyya gradient flow. IEEE TIP, 2787–2801 (2007)
14. Kim, J., Fisher, J., Yezzi, A., Cetin, M., Willsky, A.: Nonparametric methods for image segmentation using information theory and curve evolution. In: Proc. ICIP, vol. 3, pp. 797–800 (2002)
15. Paragios, N., Deriche, R.: Geodesic active regions for supervised texture segmentation. In: ICCV (2), pp. 926–932 (1999)
16. Leventon, M., Grimson, E., Faugeras, O.: Statistical shape influence in geodesic active contours. In: Proc. IEEE CVPR, pp. 1316–1324 (2000)
17. Tsai, A., Yezzi, T., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, E., Willsky, A.: A shape-based approach to the segmentation of medical imagery using level sets. IEEE Trans. on Medical Imaging 22, 137–153 (2003)
18. Cremers, D., Kohlberger, T., Schnoerr, C.: Shape statistics in kernel space for variational image segmentation. Pattern Recognition 36, 1292–1943 (2003)
19. Dambreville, S., Rathi, Y., Tannenbaum, A.: Shape-based approach to robust image segmentation using kernel pca. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 977–984 (2006)
20. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Springer, Heidelberg (2003)
21. Turk, G., Levoy, M.: Zippered polygon meshes from range images. In: Siggraph, pp. 311–318 (1994)
22. Yezzi, A., Soatto, S.: Structure from motion for scenes without features. In: Proc. IEEE CVPR, vol. 1, pp. 171–178 (2003)
23. Yezzi, A., Soatto, S.: Stereoscopic segmentation. International Journal of Computer Vision (IJCV) 53, 31–43 (2003)
24. Faugeras, O.D., Keriven, R.: Variational principles, surface evolution pdes, level set methods and the stereo problem. INRIA Tech. report 3021, 1–37 (1996)
25. Unal, G., Yezzi, A., Soatto, S., Slabaugh, G.: A variational approach to problems in calibration of multiple cameras. Trans. Pattern Analysis and Machine Intelligence 29, 1322–1338 (2007)
26. Rosenhahn, B., Brox, T., Weickert, J.: Three-dimensional shape knowledge for joint image segmentation and pose tracking. IJCV 73, 243–262 (2007)
27. Schmaltz, C., Rosenhahn, B., Brox, T., Cremers, D., Weickert, J., Wietzke, L., Sommer, G.: Region-based pose tracking. In: Pattern Recognition and Image Analysis, pp. 56–63 (2007)
28. DoCarmo, M.P.: Diferential Geometry of Curves and Surfaces. Prentice Hall, Englewood Cliffs (1976)
29. Lepetit, V., Fua, P.: Monocular model-based 3d tracking of rigid objects: A survey. Fundations and Trends in Computer Graphics and Vision 1, 1–89 (2005)
30. Forsyth, D., Ponce, J.: Computer Vision. Prentice Hall, Englewood Cliffs (2003)
31. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press, Cambridge (2000)
32. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.: An invitation to 3D vision. Springer, Heidelberg