Non-Rigid 2D-3D Pose Estimation and 2D Image Segmentation

Romeil Sandhu Samuel Dambreville Anthony Yezzi Allen Tannenbaum
Georgia Institute of Technology
School of Electrical and Computer Engineering
Atlanta, GA, USA 30322
{rsandhu,samuel.dambreville}@gatech.edu {ayezzi,tannenba}@ece.gatech.edu

Abstract

In this work, we present a non-rigid approach to jointly solve the tasks of 2D-3D pose estimation and 2D image segmentation. In general, most frameworks which couple both pose estimation and segmentation assume that one has the exact knowledge of the 3D object. However, in non-ideal conditions, this assumption may be violated if only a general class to which a given shape belongs to is given (e.g., cars, boats, or planes). Thus, the key contribution in this work is to solve the 2D-3D pose estimation and 2D image segmentation for a general class of objects or deformations for which one may not be able to associate a skeleton model. Moreover, the resulting scheme can be viewed as an extension of the framework presented in [7], in which we include the knowledge of multiple 3D models rather than assuming the exact knowledge of a single 3D shape prior. We provide experimental results that highlight the algorithm's robustness to noise, clutter, occlusion, and shape recovery on several challenging pose estimation and segmentation scenarios.

1. Introduction

Two well-studied problems in computer vision are the fundamental tasks of 2D image segmentation and 3D pose estimation from a 2D scene. By leveraging the advantages of certain techniques from each problem, we couple both tasks in a variational and *non-rigid* manner through a single energy functional. This can be viewed as an extension to the framework presented in [7], in which we include the knowledge of multiple 3D shapes rather than assuming the exact knowledge of a single 3D shape prior. However, to appreciate the contributions presented in this note, we briefly revisit some of the key results that have been made pertaining to both fields of interest.

2D-3D pose tracking or pose estimation is concerned with relating the spatial coordinates of an object in the 3D world (with respect to a calibrated camera) to that of a 2D scene [11, 15]. Although a complete literature review is beyond the scope of this note, most methodologies can be described as follows. First, one chooses a local geometric descriptor (e.g., points [19], lines [9, 16], or curves [10, 23]) that can best quantify features on the image to its corresponding 3D counterpart. Then, explicit point correspondences are established in order to solve for the pose transformation. As with most correspondence-based algorithms, which rely on local features, it can be readily seen that these techniques may suffer from the existence of homologies (due to noise, clutter, or occlusions). Moreover, and more importantly, the above methods typically constrain their approaches to the knowledge of a pre-specified 3D model. To overcome this constraint, non-rigid algorithms have appeared in the area of human pose estimation [8, 22, 1]. While we should note that the focus of our note is not specific to this area of computer vision, the proposed framework is closely related if one were to learn a large class of deformations as opposed to rigid objects. However, in contrast to methods such as [22, 8], our approach relies on the surface differential geometry of a 3D model. This

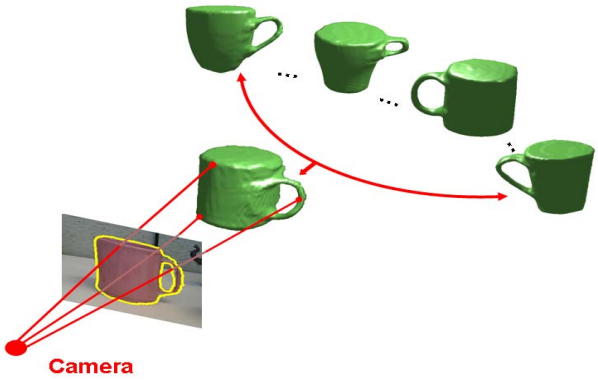


Figure 1. Our non-rigid approach to 2D image segmentation and 3D pose estimation through the use of multiple 3D shapes.

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

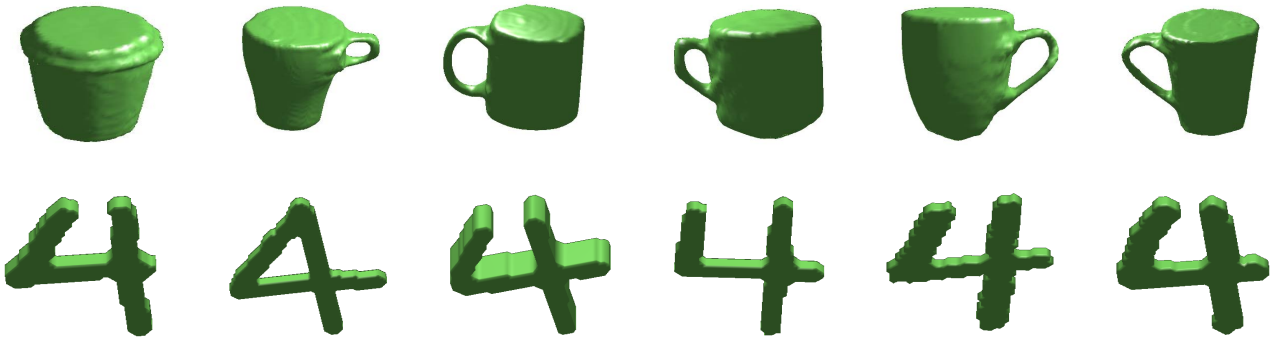


Figure 2. *Top Row:* Different 3D models of commonly used tea cups (6 of the 12 models used for the training are shown). *Bottom Row:* Different 3D models of the number 4 (6 of the 16 models used for the training are shown).

allows us to eliminate the need for point correspondences altogether while still being able to deal with an arbitrary or complex shape. In addition, because we approach the non-rigid problem through a variational manner, we require no costly stochastic optimization process [8, 1], making it ideal for tracking applications. Image segmentation consists of partitioning a scene into an "object" and a "background" [2]. In particular, we will restrict our approach to segmentation to that of the geometric active contour (GAC) framework, whereby a curve is evolved continuously until it satisfies a stopping criterion that coincides with the object's boundaries. Early approaches relied on characterizing the object by local features such as edges to drive the curve evolution [3, 12]. However, these edge-based techniques were shown to be susceptible to noise and missing information. Consequently, an alternative characterization, based on so-called "region-based methods," is to assume the "object" and "background" possess differing image statistics (see [4, 18, 17]). Although this improves segmentation results, the assumption may not hold due to clutter or occlusions. This has resulted in the proposed use of a shape prior to restrict the evolution of the active contour [14, 5, 6, 25]. We should note that even though the framework presented in this paper shares similarities with shape-based methods, one fundamental difference within our methodology is that we derive a novel 3D shape prior from a catalog of 3D shapes to do 2D image segmentation rather than to derive a 2D shape prior from a collection of 2D images. Thus, one benefit is that we are able to reduce computational complexity in statistical shape learning approaches through a compact shape representation. Figure 1 illustrates this notion.

1.1. Relation to Previous Work

It is interesting to note that while 2D-3D pose estimation and 2D image segmentation are closely related, there exist few methodologies that try to couple both tasks in a unified framework. An early attempt to solve the problem of viewpoint dependency for differing aspects of a 3D object is given in [20]. In their work, the authors propose a region-based active contour that employs a unique shape prior, which is represented through a generalized cone based on single reference view of an object. Although the method performs well under different changes in aspect, it is not able to cope with a view of an object that is substantially different from the reference view.

In addition, even though we have restricted our discussion to the GAC framework, we should also note that recent work has been done in simultaneous pose estimation and segmentation via dynamic graph cuts [13]. In their work, the authors propose an articulated shape prior through a stick-man or skeleton model. Then, in order to capture deformations occurring to the object, one must optimize over a set of pre-defined parameters corresponding to specific motions in a model's movement. In relation to this work, our focus is to accurately quantify the deformation through a set of 3D models like that of [1] through PCA. More importantly, we are able to incorporate a general class of shapes for which one may not be able to associate a skeleton model. Also, the above methodologies differ in the segmentation approach used (i.e. graph cuts versus active contours), in which we note that each method has its advantages and disadvantages.

In relation to the framework presented in this note, the authors in [21, 24] also proposed a solution to solve the joint task of pose estimation and segmentation for the case of *rigid* objects. In [21], the authors account for a variation in the projection of the 3D shape by evolving an active contour in conjunction with the 3D pose parameters to minimize a joint energy functional. While this is less restrictive, the algorithm optimizes over an infinite dimensional active contour as well as the set of finite pose parameters. Moreover, in order for one to determine the shape prior and the corresponding 3D pose, costly back projections must be made through ICP-like correspondences. An extension is considered in [24], whereby the authors successfully eliminate the need to evolve the active contour by performing a minimization of 3D pose parameters instead. However, the

costly back-projections and correspondences remain.

Thus, while one may try to expand upon other frameworks such as [21, 24] for multiple shapes, the reasoning for why we have chosen to extend the methodology presented in [7] is as follows. In this work, the authors derive a variational framework to jointly segment a *rigid* object in a 2D image and estimate the corresponding 3D pose through the use of a 3D shape prior. Specifically, the algorithm uses a region-based segmentation method to continuously drive the pose estimation process. This results in a global approach that avoids using local features or ICP-like correspondences by relying on surface differential geometry to link geometric properties of the model surface and its corresponding projection in the 2D image domain. The methodology is motivated by similar approaches that were originally constructed for stereo reconstruction from multiple cameras [27, 28] and further extended for camera calibration [26]. Consequently, the knowledge of a single 3D object is exploited to its full extent within their framework. The question that is to be addressed in this work *is how can one fully exploit the knowledge of a general class of 3D shapes as opposed to a single 3D model?*

Thus, the key contribution in this note is to extend the method in [7] to include the knowledge of multiple 3D shapes. This is done by adopting the shape based approach of [25] to the problem at hand. In other words, we evolve shape parameters or modes of variations that are obtained from performing Principal Component Analysis (PCA) on a collection of 3D shapes. As a result, we approach the non-rigid task through an optimization of a *finite* set of parameters. Moreover, this variational approach can be naturally used to learn the possible deformations of a specified object, which is essential to 3D tracking tasks such as human pose estimation [1].

The remainder of this paper is organized as follows: In the next section, we begin with a generalization of the gradient flow of [7] for an arbitrary set of finite parameters. We then provide details for evolving both the shape parameters, which are obtained from performing PCA on a collection of 3D shapes, and the corresponding pose of an object. Numerical implementation details are given in Section 3. In Section 4, we present experimental results that highlight the robustness of the technique to noise, clutter, and occlusions as well as the ability to segment a novel shape that is not apart of the specified training set. Finally, we discuss future work in Section 5.

2. Proposed Framework

We assume as with many machine learning techniques that we have a catalog of 3D shapes describing a particular object. Specifically, one can use stereo reconstruction methods [27] or range scanners to obtain accurate models as shown in Figure 2. From this, we derive a variational approach to perform the task of non-rigid 3D pose estimation and 2D image segmentation.

2.1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used methodology for shape analysis and dimensional reduction. While there exist many other methods for dimension reduction and shape analysis [5, 6], we restrict ourselves to PCA in this present work for simplicity.

Following the work of [14, 25], we let φ_i represent the signed distance function corresponding to a 3D surface X_i.[1] The average shape $\overline{\varphi}$ can then be computed from the n surfaces as $\overline{\varphi} = (\frac{1}{n}) \sum \psi_i$. From this, we can exploit the variability in the training data through PCA by first creating a mean-offset map $\tau = \{\widetilde{\varphi}_1, \widetilde{\varphi}_2, ..., \widetilde{\varphi}_n\}$ where $\widetilde{\varphi}_i = \varphi_i - \overline{\varphi}$.[2] Each map $\widetilde{\varphi}_i$ is then reorganized into a Nx1 column vector with N being the number of elements within $\widetilde{\varphi}_i$. The resulting column stacking transformation of τ yields an Nxn matrix M. Using Singular Value Decomposition (SVD), the covariance matrix $\frac{1}{n} M M^T$ is decomposed as:

$$U \Sigma U^T = \frac{1}{n} M M^T \tag{1}$$

where $U = \{\psi_1, \psi_2, ..., \psi_n\}$ is a matrix whose column vectors represent the set of orthogonal modes of shape variation and Σ is a diagonal matrix of the corresponding singular values. Rearranging the column vectors back into the structure of φ_i, we can then estimate a novel 3D shape $\hat{\varphi} = \overline{\varphi} + \sum_{i=0}^{i=k} w_i \psi_i$, where w_i is the shape weight and k is the number of principal modes used (see [25] for details). It is important to note here that we are concerned only with the zero level surface of the derived shape. This detail will be essential in solving for the shape parameters. However, before doing so, we describe the notation used throughout the rest of this paper.

2.2. Notation

Let S be the smooth surface in $\mathbb{R}^3$ defining the shape of the object of interest. With a slight abuse of notation, we denote by $\mathbf{X} = [X, Y, Z]^T$, the spatial coordinates that are measured with respect to the referential of the imaging camera. The (outward) unit normal to S at each point $\mathbf{X} \in S$ will then be denoted as $\mathbf{N} = [N_1, N_2, N_3]^T$ Moreover, we assume a pinhole camera realization $\pi : \mathbb{R}^3 \mapsto \Omega; \mathbf{X} \mapsto \mathbf{x}$, where $\mathbf{x} = [x, y]^T = [X/Z,\ Y/Z]^T$, and $\Omega \in \mathbb{R}^2$ denotes the domain of the image I with the corresponding area element $d\Omega$. From this, we define $R = \pi(S)$ to be the region on which the surface S is projected. Similarly, we can

[1] Although we have chosen the signed distance function as a representation, one can equally have chosen to use binary maps.

[2] The catalog of 3D shapes has been aligned so that any variation in shape is not due to a misalignment in the models.

form the complementary region and boundary or "silhouette" curve as $R^c = \Omega \setminus R$ and $\hat{c} = \partial R$, respectively. In other words, if we define the "occluding" curve C to be the intersection of the visible and non-visible region of S, then the image curve is $\hat{c} = \pi(C)$.

Now let $\mathbf{X}_0$ and $S_0 \in \mathbb{R}^3$ be the coordinates and surface that correspond to the 3D world, respectively. S_0 is given as the zero-level surface of the following PCA functional: $\hat{\varphi}(\mathbf{X}_0, w) = \overline{\varphi}(\mathbf{X}_0) + \sum_0^k w_i \psi_i(\mathbf{X}_0)$. That is, $S_0 = \{\mathbf{X}_0 \in \mathbb{R}^3 : \hat{\varphi}(\mathbf{X}_0, w) = 0\}$. Then one can locate the S in the camera referential via the transformation $g \in SE(3)$, such that $S = g(S_0)$. Writing this point-wise yields $\mathbf{X} = g(\mathbf{X}_0) = \mathbf{R}\mathbf{X}_0 + \mathbf{T}$, where $\mathbf{R} \in SO(3)$ and $\mathbf{T} \in \mathbb{R}^3$.

2.3. Generalized Gradient Flow

Let us begin with the assumption that if the correct 3D pose and shape were given, then the projection of the "occluding curve', i.e. $\hat{c} = \pi(C)$, would delineate the boundary that optimally separates or segments a 2D object from its background. Further assuming that the image statistics between the 2D object and its background are distinct, we define an energy functional of the following form:

$$E = \int_R r_o(I(\mathbf{x}), \hat{c}) d\Omega + \int_{R^c} r_b(I(\mathbf{x}), \hat{c}) d\Omega \qquad (2)$$

where $r_o : \chi, \Omega \mapsto \mathbb{R}$ and $r_b : \chi, \Omega \mapsto \mathbb{R}$ are functionals measuring the similarity of the image pixels with a statistical model over the regions R and R^c, respectively. Also, χ corresponds to the photometric variable of interest. In the present work, r_o and r_b are chosen to be region based functionals of [4, 18].

Now we want to optimize Equation (2) with respect to a finite parameter set denoted as $\xi = \{\xi_1, \xi_2, ..., \xi_m\}$. This is given as follows:

$$\frac{\partial E}{\partial \xi_i} = \int_{\hat{c}} \Big(r_o(I(\mathbf{x})) - r_b(I(\mathbf{x})) \Big) \Big\langle \frac{\partial \hat{c}}{\partial \xi}, \hat{\mathbf{n}} \Big\rangle d\hat{s} \qquad (3)$$

where the "silhouette" curve is parameterized by the arc length $\hat{s}$ with the corresponding outward normal $\hat{\mathbf{n}}$. If we further assume that parameter ξ_i acts on the 3D coordinates, the above line integral will be difficult to compute since $\hat{c}$ and $\hat{\mathbf{n}}$ live in the 2D image plane. Thus, it would be much more convenient if we can express the above line integral around the "occluding curve" C that lives in the 3D space and is parameterized by s. We briefly describe this lifting procedure, and refer the reader to [7] for more details. The image plane and surface are then related by

$$\Big\langle \frac{\partial \hat{c}}{\partial \xi}, \hat{\mathbf{n}} \Big\rangle d\hat{s} = \Big\langle \frac{\partial \pi(C)}{\partial \xi}, J \frac{\partial \pi(C)}{\partial s} \Big\rangle ds \qquad (4)$$

where $J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, which yields the following expression

$$\begin{aligned} \Big\langle \frac{\partial \hat{c}}{\partial \xi}, \hat{\mathbf{n}} \Big\rangle d\hat{s} &= \frac{1}{Z^3} \Big\langle \frac{\partial \mathbf{X}}{\partial \xi_i}, \begin{bmatrix} 0 & Z & -Y \\ -Z & 0 & X \\ Y & -X & 0 \end{bmatrix} \frac{\partial \mathbf{X}}{\partial s} \Big\rangle ds \\ &= \frac{1}{Z^3} \Big\langle \frac{\partial \mathbf{X}}{\partial \xi_i}, \frac{\partial \mathbf{X}}{\partial s} \times \mathbf{X} \Big\rangle ds \\ &= \frac{\|\mathbf{X}\|}{Z^3} \sqrt{\frac{\kappa_X \kappa_t}{K}} \Big\langle \frac{\partial \mathbf{X}}{\partial \xi_i}, \mathbf{N} \Big\rangle ds. \end{aligned} \qquad (5)$$

Here K denotes the Gaussian curvature, and κ_X and κ_t denote the normal curvatures in the directions $\mathbf{X}$ and $\mathbf{t}$, respectively, where $\mathbf{t}$ is the vector tangent to the curve C at the point $\mathbf{X}$, i.e. $\mathbf{t} = \frac{\partial \mathbf{X}}{\partial s}$. If we now plug the result of Equation (5) into Equation (3), we arrive at the following flow

$$\frac{\partial E}{\partial \xi_i} = \int_C \Big(r_o\big(I(\pi(\mathbf{X}))\big) - r_b\big(I(\pi(\mathbf{X}))\big) \Big) \cdot \frac{\|\mathbf{X}\|}{Z^3} \sqrt{\frac{\kappa_X \kappa_t}{K}} \Big\langle \frac{\partial \mathbf{X}}{\partial \xi_i}, \mathbf{N} \Big\rangle ds. \qquad (6)$$

Note, that in the above derivation we made no assumptions about the finite set. That is, we show that the overall framework is essentially "blind" to whether we optimize over the shape weights or pose parameters. What is important is how the functional in Equation 3 is lifted from the "silhouette" curve to the "occluding curve" so that the gradient can be readily computed. In particular, the term $\Big\langle \frac{\partial \mathbf{X}}{\partial \xi_i}, \mathbf{N} \Big\rangle$ is what we will focus on in Sections 2.4 and 2.5.

2.4. Evolving the Shape Parameters

In this section, we compute the term $\Big\langle \frac{\partial \mathbf{X}}{\partial \xi_i}, \mathbf{N} \Big\rangle$, when ξ_i corresponds to the shape weights obtained from performing PCA on a collection of 3D models. Let $\xi = \omega = \{\omega_1, \omega_2, ..., \omega_k\}$ with k being the number of principal modes used. In addition, the 3D coordinates X_0, which is derived from the surface S_0, are related by the constraint

$$\begin{aligned} &\hat{\varphi}(\mathbf{X}_0, w) = \overline{\varphi}(\mathbf{X}_0) + \textstyle\sum_0^k w_i \psi_i(\mathbf{X}_0) \\ &\text{s.t.} \qquad \hat{\varphi}(X_0(w), w) = 0. \end{aligned} \qquad (7)$$

The term $\Big\langle \frac{\partial \mathbf{X}}{\partial \omega_i}, \mathbf{N} \Big\rangle$ can then be computed as follows:

$$\begin{aligned} \Big\langle \frac{\partial \mathbf{X}}{\partial \omega_i}, \mathbf{N} \Big\rangle &= \Big\langle \frac{\partial \mathbf{R}\mathbf{X}_0 + \mathbf{T}}{\partial \omega_i}, \mathbf{N} \Big\rangle = \Big\langle \mathbf{R} \frac{\partial \mathbf{X}_0}{\partial \omega_i}, \mathbf{N} \Big\rangle \\ &= \Big\langle \frac{\partial \mathbf{X}_0}{\partial \omega_i}, \mathbf{R}^T \mathbf{N} \Big\rangle = \Big\langle \frac{\partial \mathbf{X}_0}{\partial \omega_i}, \mathbf{R}^T \mathbf{R} \mathbf{N}_0 \Big\rangle \\ &= \Big\langle \frac{\partial \mathbf{X}_0}{\partial \omega_i}, \mathbf{N}_0 \Big\rangle. \end{aligned} \qquad (8)$$

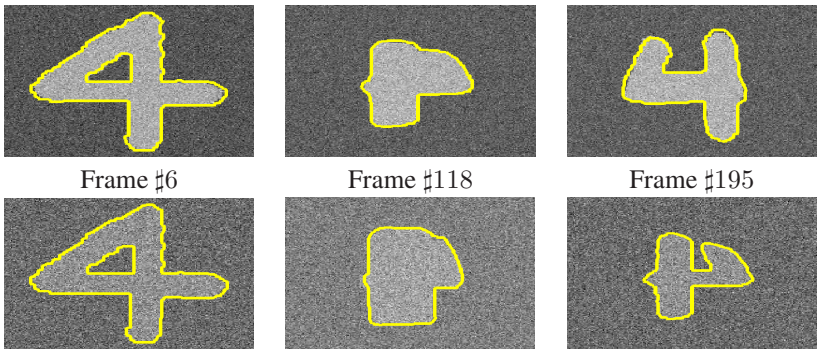


Figure 3. Robustness to deformation (and noise). Visual tracking results for the sequence involving the number 4. *First row:* Tracked sequence with Gaussian noise of standard deviation $\sigma = 25\%$. *Second row:* Tracked sequence for $\sigma = 75\%$.

Using the above constraint on the zero-level surface, and noting that $\frac{\nabla_{\mathbf{X}_0}\hat{\varphi}}{\|\nabla_{\mathbf{X}_0}\hat{\varphi}\|} = \mathbf{N}_0$, we then have that

$$\begin{aligned} 0 &= \frac{\partial}{\partial \omega_i}\hat{\varphi}(X_0(w), w) \\ &= \left\langle \nabla_{\mathbf{X}_0}\hat{\varphi}, \frac{\partial \mathbf{X}_0}{\partial \omega_i} \right\rangle + \frac{\partial \hat{\varphi}}{\partial \omega_i} \\ &= \left\langle \|\nabla_{\mathbf{X}_0}\hat{\varphi}\|\mathbf{N}_0, \frac{\partial \mathbf{X}_0}{\partial \omega_i} \right\rangle + \psi_i(\mathbf{X}_0) \end{aligned}$$

which yields the following compact expression

$$\left\langle \frac{\partial \mathbf{X}_0}{\partial \omega_i}, \mathbf{N}_0 \right\rangle = -\frac{\psi_i(\mathbf{X}_0)}{\|\nabla_{\mathbf{X}_0}\hat{\varphi}\|}. \tag{9}$$

The result presented in equation (9) provides the variation of the energy with respect to the shape parameters, and is one of the major contributions of this work. Next, we discuss how one can evolve the pose parameters.

2.5. Evolving the Pose Parameters

In this section, we revisit the evolution of the pose parameters derived in [7]. Specifically, if we let $\xi = \lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\}^T$, we can then compute the term $\left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \mathbf{N} \right\rangle$ where λ_i is a translation or rotation parameter:

- For $i = 1, 2, 3$ (i.e., λ_i is a translation parameter), and $\mathbf{T} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}$, one has

$$\begin{aligned} \left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \mathbf{N} \right\rangle &= \left\langle \frac{\partial \mathbf{R}\mathbf{X}_0 + \mathbf{T}}{\partial \lambda_i}, \mathbf{N} \right\rangle = \left\langle \frac{\partial \mathbf{T}}{\partial \lambda_i}, \mathbf{N} \right\rangle \\ &= \left\langle \begin{bmatrix} \frac{\partial \lambda_1}{\partial \lambda_i} \\ \frac{\partial \lambda_2}{\partial \lambda_i} \\ \frac{\partial \lambda_3}{\partial \lambda_i} \end{bmatrix}, \mathbf{N} \right\rangle = \left\langle \begin{bmatrix} \delta_{1,i} \\ \delta_{2,i} \\ \delta_{3,i} \end{bmatrix}, \mathbf{N} \right\rangle \\ &= N_i. \end{aligned} \tag{10}$$

where the Kronecker symbol $\delta_{i,j}$ was used ($\delta_{i,j} = 1$ if $i = j$ and 0 otherwise).

- For $i = 4, 5, 6$ (i.e., λ_i is a rotation parameter), and using the expression of the rotation matrix written in exponential coordinates, $\mathbf{R} = \exp\left(\begin{bmatrix} 0 & -\lambda_6 & \lambda_5 \\ \lambda_6 & 0 & -\lambda_4 \\ -\lambda_5 & \lambda_4 & 0 \end{bmatrix}\right)$, one has

$$\begin{aligned} \left\langle \frac{\partial \mathbf{X}}{\partial \lambda_i}, \mathbf{N} \right\rangle &= \left\langle \frac{\partial \mathbf{R}\mathbf{X}_0}{\partial \lambda_i}, \mathbf{N} \right\rangle \\ &= \left\langle \mathbf{R} \begin{bmatrix} 0 & -\delta_{3,i} & \delta_{2,j} \\ \delta_{3,i} & 0 & -\delta_{1,i} \\ -\delta_{2,i} & \delta_{1,i} & 0 \end{bmatrix} \mathbf{X}_0, \mathbf{N} \right\rangle. \end{aligned} \tag{11}$$

We note that one can also expand the rigid body transformation to a more general affine transformation, and hence provide increased flexibility in the proposed approach.

3. Implementation

In Equation (6), the computation of the gradients involve the explicit determination of the occluding curve C. As previously mentioned, one can compute

$$C = \{\mathbf{X} \in \mathcal{V}^+ \cap \mathcal{V}^-, \ \text{such that} \ \pi(\mathbf{X}) \in \hat{c}\} \tag{12}$$

where

$$\begin{aligned} \mathcal{V}^+ &= \{\mathbf{X} \in S, \ \text{s.t.} \ \langle \mathbf{X}, \mathbf{N} \rangle \geq 0\} \ \text{and} \\ \mathcal{V}^- &= \{\mathbf{X} \in S, \ \text{s.t.} \ \langle \mathbf{X}, \mathbf{N} \rangle \leq 0\}. \end{aligned}$$

where the set V (respectively, $V^c = S \setminus V$) of points $\mathbf{X} \in S$ that are visible (respectively, not visible) from the camera center is such that $V \subset \mathcal{V}^+$ (respectively, $V^c \supset \mathcal{V}^-$). Moreover, to save computational time, we approximated the term $\sqrt{\frac{\kappa_X \kappa_i}{K}} \simeq 1$ in Equation (6), which still decreased the energy E.

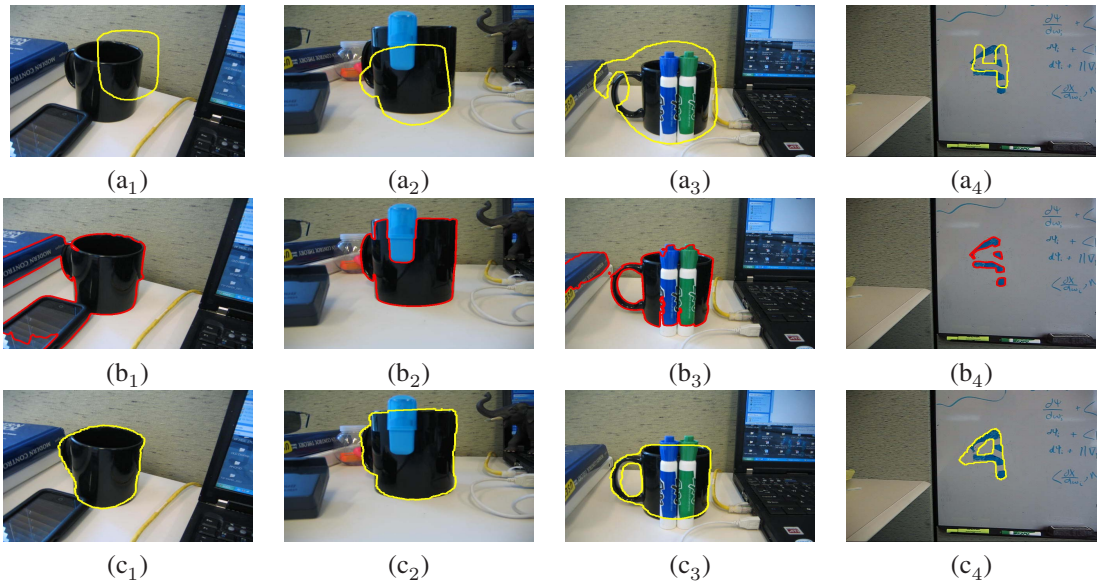


Figure 4. Robustness to Occlusion and Clutter. *Top row:* Initialization *Middle Row:* Unsatisfactory results obtained from using an active contour. *Bottom row:* Results obtained from proposed approach

Lastly, in performing PCA, the authors in [14, 25] note that one can compute the eigenvalues and eigenvalues of $\frac{1}{n}MM^T$ from a much smaller matrix $W = \frac{1}{n}M^TM$, where if $\mathbf{b}$ is an eigenvector of W with the corresponding eigenvalue ς, then $\varsigma\mathbf{b}$ and ς are the eigenvector and eigenvalue of $\frac{1}{n}MM^T$.

4. Experiments

We provide experimental results to demonstrate the algorithm's robustness to **noise**, **deformation**, **occlusion** or **clutter**, and its ability in effective **shape recovery**. Specifically, we generate two 3D training sets corresponding to the number "4" and commonly used tea cups as shown in Figure 2.

4.1. Robustness to Deformation (and Noise)

In the first set of experiments, we demonstrate the algorithm's robustness to noise in the presence of continuous deformation. First, a tracking sequence was generated consisting of 300 frames that were obtained from projecting the number "4" onto the 2D image plane using a simulated camera. Specifically, the variation in the rotation angle was a complete 360° cycle, and the model was varied linearly along its z-axis from 300 to 500 spatial units resulting in a scale in the viewing aspect of the image projection. Moreover, we vary the first three principal modes so that a deformation can be seen. From this basic sequence, two cases of additive Gaussian noise with a standard deviation of $\sigma_n = 25\%$ to $\sigma_n = 75\%$ are formed.

Figure 3 show three frames which exhibit typical tracking results from each of the two noise cases described. Here, we have used the region-based energy of [4], and obtain tracking results by varying the first 6 principal modes, i.e. $k = 6$. Because several 2D-3D pose estimate techniques [19, 9, 16] rely on a correspondence based scheme to estimate the pose, their methodologies are sensitive to noise and outliers as presented in Figure 3. Because of the robustness of region-based active contours (as opposed to local geometric descriptors), the proposed approach yields satisfying visual results such as those of [7]. However, we have not constrained ourselves to know the actual shape of a prespecified number "4". It is straightforward to see (without comparison), that if we were to assume the knowledge of a model, then it would not be possible for us to handle the wide range of deformation seen.

Thus, if it is desirable to track a rigid object that is representative of a certain class (e.g., cars, boats, or planes), then one can learn the different 3D models of a class, and thereby relax the constraint of the prior knowledge needed.

4.2. Robustness to Occlusion and Clutter

In this section, we provide experimental validation that compares our segmentation method, which is an optimization over a finite set, with that of the infinite dimensional geometric active contour (GAC) technique. The reasoning for such a comparison is as follows: In either methodology, we seek to minimize a cost functional of the general form $E(t) = \int_\gamma \Psi(\mathbf{x}, t)d\mathbf{x}$ over a family of curves. For the GAC framework, this family of curves lives only in the image plane when performing 2D image segmentation. Similarly, in the current framework, we are optimizing over a family of 3D "occluding" curves that correspond to 2D "silhou-

ette" curves. That is, we cast the typical infinite dimensional problem of segmentation as a finite optimization problem.

Thus, we benefit from incorporating shape information which results in being able to deal with not only occlusion, but also in clutter environments where the original assumption of separable statistics does not hold. This is shown on four different examples as seen in Figure 4. The top row illustrates the initialization, while the middle row shows the unsatisfactory result of using an active contour. The final row highlights results given by the proposed approach with $k = 6$. Although it is not readily apparent, one can alternatively view the examples in Figure 4 as 3D reconstruction from a single 2D image view exhibiting partial information, which is a fundamental task in computer vision.

4.3. Robustness to Shape Recovery

In this section, we shift the focus of our experiments to segmenting shapes that are not in the original training sets. We want to specifically highlight an important advantage of learning 3D shapes as opposed to a large catalog of 2D shapes to perform the task of 2D segmentation. This is done by segmenting different shapes arising from different 3D models as well as from altering the pose of a 3D object.

In Figure 5, we show the segmentation of different views and shapes that are obtained from a person coloring in the number "4" on a white-board. The top row highlights the initialization while the bottom rows shows the final result. The same experiment is performed for the differing views of a single tea cup that was not in the original training set (see Figure 6). While one may argue that the results are similar to the 2D shape learning approaches, we should note the key differences. First, we not only segment the 2D image, but are able to return the estimated 3D shape and pose from which the 2D object was derived. Moreover, to account for segmentation of objects presented in Figure 5 and Figure 6 with a 2D shape prior, one would have to learn every possible projection of the 3D object onto to the 2D image plane (if no prior knowledge is given about the aspect of the projection). Thus, aside from handling a larger class of scenarios, we greatly reduce the dimensionality of the shape space by using a 3D compact representation. Note, we set $k = 6$.

5. Conclusion and Future Work

In this paper, we derive a geometric and variational approach to perform the task of 2D-3D *non-rigid* pose estimation and 2D image segmentation. This can be seen as a extension of the framework presented in [7], for which we have relaxed the assumption of only a single 3D shape prior. Instead, we infer a 3D shape prior from a catalog of 3D shapes, which may represent a general class of an object or deformations that may occur to the model. As a result, we fully exploit the task of pose estimation and segmentation in a unified framework.

Moreover, because we have shown the above framework is essentially "blind" to the type of finite set, a direction for future work is to incorporate various statistical shape learning techniques such as kernel PCA [5, 6]. This is of particular importance since it has been shown that performing PCA on a collection of models, which exhibit a large shape variation, may yield in an unsatisfactory result. Instead, if we adopt a kernel methodology that allows for us to "cut" the space of shapes, we can further increase the algorithm's robustness to segmentation by allowing only shapes that are close enough to the training data. In doing so, we should then be able to handle a more general class of 3D objects.

Acknowledgements

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. This work was also funded in part by grants from NSF, AFOSR, ARO as well as by a grant from NIH (NAC P41 RR-13218) through Brigham and Women's Hospital.

References

[1] A. Balan, L.Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007.

[2] A. Blake and M. Isard, editors. *Active Contours*. Springer, 1998.

[3] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. In *IJCV*, volume 22, pages 61–79, 1997.

[4] T. Chan and L. Vese. Active contours without edges. *IEEE TIP*, 10(2):266–277, 2001.

[5] D. Cremers, T. Kohlberger, and C. Schnoerr. Shape statistics in kernel space for variational image segmentation. In *Pattern Recognition*, pages 1292–1943, 2003.

[6] S. Dambreville, Y. Rathi, and A. Tannenbaum. A framework for image segmentation using shape models and kernel space shape priors. *TPAMI*, 30(8):1385–1399, 2008.

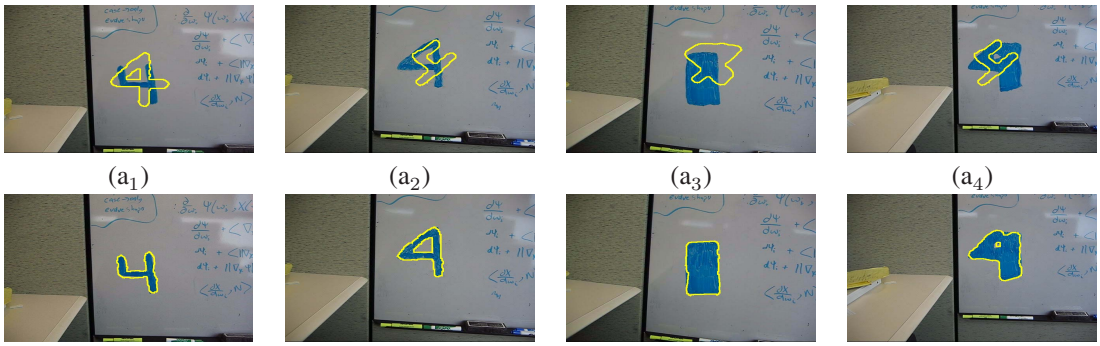
[7] S. Dambreville, R. Sandhu, A. Yezzi, and A. Tannenbaum. Robust 3d pose estimation and efficient 2d region-based segmentation from a 3d shape prior. In *ECCV*, page 2008.

[8] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. In *IJCV*, 2004.

[9] M. Dhome, M. Richetin, and J.-T. Lapreste. Determination of the attitude of 3d objects from a single perspective view. *TPAMI*, 11(12):1265–1278, 1989.

[10] T. Drummond and R. Cipolla. Real-time tracking of multiple articulated structures in multiple views. In *ECCV*, 2000.

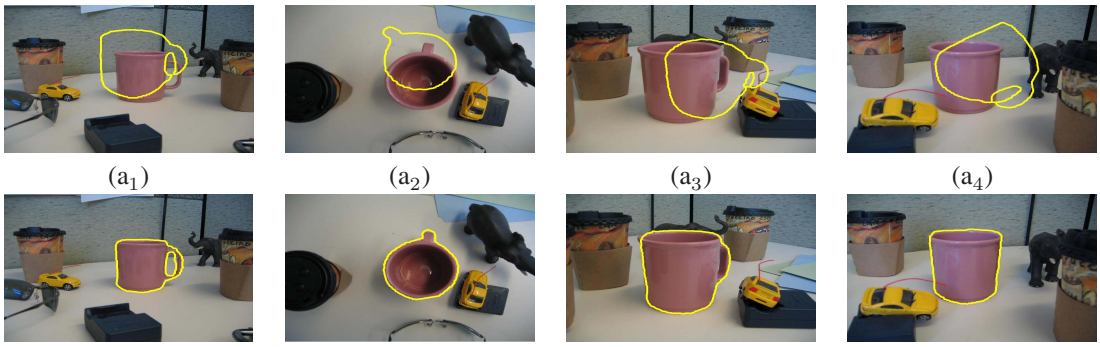
[11] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.



(a_1) (a_2) (a_3) (a_4)

(b_1) (b_2) (b_3) (b_4)

Figure 5. Robustness to Shape Recovery: Segmentation of different views and shapes of the number "4," which are not present in the training set. *Top Row*: Initialization. *Bottom Row*: Final Results obtained for running proposed method to convergence



(a_1) (a_2) (a_3) (a_4)

(b_1) (b_2) (b_3) (b_4)

Figure 6. Robustness to Shape Recovery: Segmentation of different views and shapes of a teacup, which is not present in the training set *Top Row*: Initialization. *Bottom Row*: Final Results obtained for running proposed method to convergence

[12] S. Kichenassamy, S. Kumar, P. Olver, A. Tannenbaum, and A. Yezzi. Conformal curvature flow: From phase transitions to active vision. In *Archives for Rational Mechanics and Analysis*, volume 134, pages 275–301, 1996.

[13] P. Kohli, J. Rihan, M. Bray, and P. Torr. Simultaneous segmentation and 3d pose estimation of humans using dynamic graph cuts. *IJCV*, 79(3):285,298, 2008.

[14] M. Leventon, E. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours. In *Proc. IEEE CVPR*, pages 1316–1324, 2000.

[15] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An invitation to 3D vision*. Springer.

[16] E. Marchand, P. Bouthemy, and F. Chaumette. A 2d-3d model-based approach to real-time visual tracking. *IVC*, 19(13):941–955, 2001.

[17] O. Michailovich, Y. Rathi, and A. Tannenbaum. Image segmentation using active contours driven by the bhattacharyya gradient flow. *IEEE TIP*, 16:2787–2801, 2007.

[18] N. Paragios and R. Deriche. Geodesic active regions: A new paradigm to deal with frame partition problems in computer vision. *VCIR*, 13:249–268, 2002.

[19] L. Quan and Z.-D. Lan. Linear n-point camera pose determination. *IEEE TPAMI*, 21(8):774–780, 1999.

[20] T. Riklin-Raviv, N. Kiryati, and N. Sochen. Prior-based segmentation by projective registration and level sets. In *ICCV*, pages 204–211, 2005.

[21] B. Rosenhahn, T. Brox, and J. Weickert. Three-dimensional shape knowledge for joint image segmentation and pose tracking. *IJCV*, 73(3):243–262., 2007.

[22] B. Rosenhahn, U. Kersting, K. Powel, and H.-P. Seidel. Cloth x-ray: Mocop of people wearing textiles. In *DAGM*, pages 495–504, 2006.

[23] B. Rosenhahn, C. Perwass, and G. Sommer. Pose estimation of free-form contours. *IJCV*, 62(3):267–289., 2005.

[24] C. Schmaltz, B. Rosenhahn, T. Brox, D. Cremers, J. Weickert, L. Wietzke, and G. Sommer. Region-based pose tracking. In *Pattern Recognition and Image Analysis*, pages 56–63, 2007.

[25] A. Tsai, T. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, E. Grimson, and A. Willsky. A shape-based approach to the segmentation of medical imagery using level sets. *IEEE TIP*, 22(2):137–153, 2003.

[26] G. Unal, A. Yezzi, S. Soatto, and G. Slabaugh. A variational approach to problems in calibration of multiple cameras. *TPAMI*, 29:1322–1338., 2007.

[27] A. Yezzi and S. Soatto. Stereoscopic segmentation. *IJCV*, 53(3):31–43., 2003.

[28] A. Yezzi and S. Soatto. Structure from motion for scenes without features. In *Proc. IEEE CVPR*, volume 1, pages 171–178, 2003.