Shape Tracking with Occlusions via Coarse-to-Fine Region-Based Sobolev Descent

Yanchao Yang and Ganesh Sundaramoorthi

Abstract—We present a method to track the shape of an object from video. The method uses a joint shape and appearance model of the object, which is propagated to match shape and radiance in subsequent frames, determining object shape. Self-occlusions and dis-occlusions of the object from camera and object motion pose difficulties to joint shape and appearance models in tracking. They are unable to adapt to new shape and appearance information, leading to inaccurate shape detection. In this work, we model self-occlusions and dis-occlusions in a joint shape and appearance tracking framework. Self-occlusions and the warp to propagate the model are coupled, thus we formulate a joint optimization problem. We derive a coarse-to-fine optimization method, advantageous in tracking, that initially perturbs the model by coarse perturbations before transitioning to finer-scale perturbations seamlessly. This coarse-to-fine behavior is automatically induced by gradient descent on a novel infinite-dimensional Riemannian manifold that we introduce. The manifold consists of planar parameterized *regions*, and the metric that we introduce is a novel Sobolev metric. Experiments on video exhibiting occlusions/dis-occlusions, complex radiance and background show that occlusion/dis-occlusion modeling leads to superior shape accuracy.

Index Terms—Object segmentation from video, object tracking, deformable templates, occlusions, shape metrics, optical flow

1 Introduction

In many applications (e.g., post-production of motion pictures, 3D video, robotics, augmented reality), it is important to determine the precise shape of the object of interest at each frame of a video. Many existing tracking methods that are designed to obtain object shape (e.g., [1], [2], [3], [4]) use a step that aims to partition the image into object and background by discriminating elementary image statistics (e.g., color, edges, texture, motion) into two groups. These approaches have the advantage of pixel-wise accuracy when the object and the background have simple and distinguishable radiance. Additional constraints from motion models (e.g., [5], [6]) and prior object shape information (e.g., [2]) have led to improvements over a basic partitioning approach in more complex scenarios. However, in tracking objects with complex radiance in a cluttered background, the underlying assumption that the image (or even a neighborhood around the object) consists of elementary statistics that fit in two groups is not always valid. Often times that assumption, even if augmented with additional constraints, leads to errors in shape detection.

One way to cope with complex object radiance is to use a dynamic model of the object shape and radiance. The model is a template, which is the dense radiance function of the object defined on the region of the projected object. The template at frame t is deformed to match the object shape and radiance in frame $t+1$, thereby obtaining the object segmentation in frame $t+1$. We refer to this approach as *joint shape/appearance matching*. One difficulty in deforming a template to match the object in the next frame is that object and camera motion may induce parts of the object to come into view (*dis-occlusions*) and go out of view (*occlusions*). If the template is not updated to account for occlusions and dis-occlusions, the deformed template may not capture the object shape accurately.

This work addresses the problem of self-occlusions and dis-occlusions within a joint shape/appearance matching framework. Our approach computes the deformation of the template to match the next frame while detecting occlusions of the template and dis-occluded parts of the object. The template is updated to remove the occlusion and include the dis-occluded region. Since the frame-rate of typical video induces non-infinitesimal deformation of the projected object between frames, we model the deformation as the integration of a time-varying vector field following a standard representation from fluid mechanics [7]. In contrast to standard representations, since we are interested in only the deforming the object of interest, the time varying vector field is defined on an evolving region (not the entire image domain). Following observations in [8] for computing optical flow, we note that an occlusion is the part of the template that does not deform to the next frame, and therefore, occlusions and the deformation are coupled. We thus setup a joint optimization problem for the deformation and the

- *Y. Yang is with the Department of Electrical Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. E-mail: yanchao.yang@kaust.edu.sa.*
- *G. Sundaramoorthi is with the Department of Electrical Engineering and the Department of Applied Mathematics and Computational Science, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. E-mail: ganesh.sundaramoorthi@kaust.edu.sa.*

Manuscript received 23 Dec. 2013; revised 18 Aug. 2014; accepted 17 Sept. 2014. Date of publication 24 Sept. 2014; date of current version 4 Apr. 2015.
Recommended for acceptance by R. Vidal.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2360380

occlusion. We note that dis-occlusions can only be detected with prior assumptions on the object, and show that a self-similarity prior on the object radiance can be used to determine dis-occlusions.

The optimization problem for the deformation and occlusion is non-convex and cannot easily be written in a convex form since the deformation is non-infinitesimal, precluding linearization used in optical flow methods. Therefore, we introduce a novel coarse-to-fine optimization method that avoids many undesirable local minima. The method is a gradient descent on a novel Riemannian manifold. The manifold consists of parameterized *regions*, represented as warps from an initial region to arbitrary regions defined in the plane. The choice of regions is suitable since the object in the imaging plane is described by both its shape and its radiance. The latter is defined on a *region* in the imaging plane. We introduce a Sobolev-type Riemannian metric defined on vector fields on regions. The gradient descent with respect to this metric induces a beneficial property for tracking: an initial region is deformed according to coarse deformations before transitioning automatically and seamlessly to finer deformations.

1.1 Key Contributions

Our main contributions are two-fold: modeling and theory. The first main contribution is to formulate self-occlusions and dis-occlusions in tracking by joint shape/appearance matching. Occlusions have been modeled in shape tracking, but existing works do so with simpler models of radiance, i.e., color histograms (e.g., [3]), or are layered models (e.g., [9]) that can cope with occlusions of one layer on another, but not *self*-occlusions or dis-occlusions. We also solve dis-occlusions with the similarity prior mentioned above. The second main contribution is a novel optimization scheme for energies defined on deformations that has an automatic coarse-to-fine behavior. This scheme is based on new theoretical advances, including our novel Riemannian manifold of regions, and a novel Sobolev metric on infinitesimal perturbations of regions.

This work extends our conference paper [10]. One extension in this paper (Section 5) is defining a novel Sobolev metric on a new Riemannian manifold of regions, leading to the automatic coarse-to-fine optimization scheme. In contrast, the scheme in [10] was only an approximation of the coarse-to-fine property, and not based on a unified energy. Also, the new optimization avoids a joint problem for the infinitesimal deformation (not the large deformation), which speeds up the technique by a factor of 2.

2 Related Work

2.1 Tracking and Occlusions

A video consists of a sequence of images, and thus, many approaches for shape tracking (e.g., [1], [2], [3], [5], [11]) have built on image segmentation techniques such as active contours (e.g., [12], [13], [14], [15], [16], [17]) and more recently, convex relaxations of active contour energies (e.g., [18], [19]). These approaches aim to determine the object of interest and the background by separating elementary image statistics (e.g., color, texture, edges, motion) into two groups. However, when the object has complex radiance and is within cluttered background, grouping elementary image statistics leads to errors in the segmentation. Some methods try to resolve this issue in tracking by using space-varying local statistics to perform the grouping (e.g., [20]). Other methods use motion models to predict the object location/shape in the next frame (e.g., [1], [5], [6], [21]) to provide more accurate initialization to frame partitioning. Dynamic models of the shape are constructed from training data in [2], extending active shape and appearance models [22], and used to constrain the solution of frame partitioning. Training data is not always available. While these extensions provide improvements to basic frame partitioning, complex object radiance and cluttered background still pose a challenge.

Our approach uses a model of the object that is a dense radiance function defined on the projected object. Other tracking methods (e.g., [23], [24]) also use dense radiance functions. However, they only obtain bounding boxes around the object, and do not provide shape. Joint dynamic models of radiance and shape for tracking have been considered in [9], [25]. However, [25] does not consider occlusions, and while [9] considers occlusions of an object by another object, it does not consider *self*-occlusions and *dis-occlusions*.

Since occlusions arise from object/camera motion, occlusions have been computed from optical flow. In [26], [27], occluded regions are defined to be the set there the composition of the forward and backward optical flow is not the identity map. In [28], [29], occlusions are detected by detecting regions where the optical flow residual is large. Occlusion boundaries are detected by discontinuities of optical flow in [30]. Noting that reliable optical flow depends on knowledge of occluded regions, and that occlusions are regions where optical flow does not exist, joint estimation of the optical flow and occlusions is performed in [8]. In [31], dense trajectory estimation across multiple frames with occlusions is solved. We use ideas of occlusions in [8], and apply them to shape tracking where considerations must be made for evolving the shape, large deformations, and dis-occlusions.

2.2 Shape Metrics

The optimization technique for deformation and occlusion estimation that we introduce is a gradient descent on a Riemannian shape manifold. Thus, our work relates to the literature on *shape metrics* defined on a Riemannian manifold of shapes. There have been two primary uses for shape metrics. One is *shape optimization*, that is, minimization of energies defined on shapes, e.g., to segment shapes from images. The other is *shape matching and analysis*, i.e., computing morphs between already segmented shapes or decomposing shapes into constituent components (e.g., via PCA).

Active contours (e.g., [12], [14], [15], [17], [32]), where shape is defined as a planar contour, are an instance of shape optimization. Active contours are usually based on a gradient descent of an energy, and the gradient depends on a choice of a metric on perturbations of planar contours. The metric typically chosen is a geometric $\mathbb{L}^2$ metric. Other metrics for active contours, in particular Sobolev-type metrics on contours, were considered by [33], [34]. These metrics favor spatially regular flows for gradient descent and avoid undesirable local minima due to fine structures in an image. In [35], it was shown that Sobolev-type metrics are suited for tracking applications since they have an automatic coarse-to-fine property in comparison to the $\mathbb{L}^2$ metric. The new Riemannian metric introduced in this paper is motivated by the coarse-to-fine property noticed in [35]. The energies considered in this paper are *not defined on contours*, but on parameterized *regions* since the radiance of an object is defined on a *region*. Thus, the framework of [35] does not apply. We define a new Riemannian manifold and a Sobolev-type metric on parameterized regions, i.e., warps of a region to arbitrary regions.

In shape matching and analysis, several Riemannian metrics have been proposed. In [36], an $\mathbb{L}^2$ Riemannian metric is proposed on the tangent vector field of planar curves. In [37], [38], Sobolev-type metrics are proposed on planar curves, which induces meaningful shape morphings as geodesic paths (shortest paths on the manifold of shapes), unlike the $\mathbb{L}^2$ metric, which does not yield geodesics [39], [40]. Deformable templates [41], [42], [43] defines a Riemannian manifold on the space of warps (diffeomorphisms) from the entire domain of the image to itself. Shape matching can be performed by diffeomorphisms that map an indicator function (defined on the entire domain of the image) of one shape onto another. Sobolev metrics on vectors fields of the fixed domain are defined, and geodesic paths are computed.

Our work relates to deformable templates, since we also define a Riemannian metric on a space of warps. However, there are two differences, besides the obvious fact that we are interested in object tracking rather than image registration or shape matching of already segmented shapes as in [41], [42], [44], [45]. First, our set of warps are defined on a region of an object to all regions in the imaging domain. This choice is natural since we model only the object of interest. Modeling the entire image presents difficulties, as the image consists of objects and the background that have differing motion. The smoothness assumption on entire domain made in [41], [42], [44] is more suited to medical images than video from natural scenes where there are discontinuities in deformation between boundaries of objects. Moreover, occlusions are not considered in [41], [42], [44]. The second difference from [41], [42] is that we are not interested in computing geodesic paths on the Riemannian manifold of warps, rather we compute a gradient descent on warps. The latter may be computationally more efficient since computing geodesics requires searching for a minimal path over all paths, whereas a gradient descent simply chooses a path based on the energy and the metric. The gradient descent with respect to the metric we introduce also induces a coarse-to-fine evolution of the region.

Lastly, the work of [46] introduces a Sobolev-type Riemannian metric on regions for shape matching rather than *shape optimization*, which is the focus of this work. The particular form of the Sobolev-metric that we construct is different than [46] as it has a natural decomposition of perturbations of a region into translations and orthogonal deformations, which is well suited for object tracking.

3 Dynamic Model of the Projected Object

We now define a dynamic model of the object shape and radiance in the imaging plane. From the model, the notion of occlusions and dis-occlusions is clear. The model is also needed for the recursive estimation algorithm in Section 7.

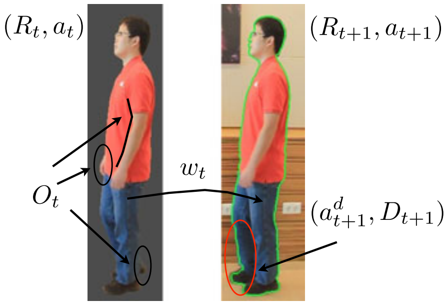


Fig. 1. Diagram illustrating our dynamic model. Left: template (R_t, a_t) (non-gray), right: I_{t+1}. Self-occlusions O_t, dis-occlusions D_{t+1} and its radiance a_d^{t+1}, the region at frame $t+1$ is R_{t+1} (inside the green contour), and the warp is w_t, which is defined in $R_t \backslash O_t$. The curved black line is a self-occlusion since the arm moves towards the left.

Let $\Omega \subset \mathbb{R}^2$, and $I : \{1, 2, \dots, N\} \times \Omega \to \mathbb{R}^k$ denote the image sequence (N frames) that has k channels. We denote frame t by I_t. The camera projection of visible points on the 3D object at time t is denoted by R_t, which we refer to as "shape" or region. The projected object's radiance is denoted a_t, and $a_t : R_t \to \mathbb{R}^k$. Our dynamic model of the region and radiance (see Fig. 1 for a diagram) is

$$R_{t+1} = w_t(R_t \backslash O_t) \cup D_{t+1}, \tag{1}$$

$$a_{t+1}(x) = \begin{cases} a_t(w_t^{-1}(x)) + \eta_t(x), & x \in w_t(R_t \backslash O_t), \\ a^d_{t+1}(x), & x \in D_{t+1}, \end{cases} \tag{2}$$

where O_t denotes the subset of R_t that is occluded from view in frame $t+1$, D_{t+1} denotes the subset of the projected object that is dis-occluded (comes into view) at frame $t+1$, $a^d_{t+1} : D_{t+1} \to \mathbb{R}^k$ is the radiance of the dis-occluded region, and w_t maps points that are not occluded in R_t to R_{t+1} in the next frame. The warp w_t is an invertible map on the *unoccluded* region $R_t \backslash O_t$, which is a transformation arising from viewpoint change and deformation. The warp will be extended to all of R_t (see Section 4.1 for details).

The region $R_t \backslash O_t$, is warped by w_t and the dis-occlusion of the object, D_{t+1}, is appended to the warped region to form R_{t+1}. The relevant portion of the radiance, $a_t|(R_t \backslash O_t)$ is transferred via the warp w_t to R_{t+1}, as brightness constancy, and noise added. A newly visible radiance is obtained in D_{t+1}. The noise models deviation from brightness constancy, e.g., non-Lambertian reflectance, small illumination change, noise.

Organization of the rest of the paper. A template (a_0, R_0) of the object is given. Our goal is, given an estimate of R_t, a_t, and I_{t+1} to estimate R_{t+1} in I_{t+1}. In Section 4.1, we formulate an optimization problem to determine w_t and the occlusion O_t given a_t, R_t, and I_{t+1}. In Section 4.2, we formulate an optimization problem to determine the dis-occlusion D_{t+1} given $w_t(R_t \backslash O_t)$ and I_{t+1}. The joint energy for w_t and O_t presented in Section 4.1 involves an alternating optimization. In Section 5, we present a new general optimization scheme for energies defined on warps, which requires introducing a new Riemannian manifold and a novel *Sobolev-type region based metric*. The induced gradient descent is shown to have a coarse-to-fine property. This optimization scheme is a

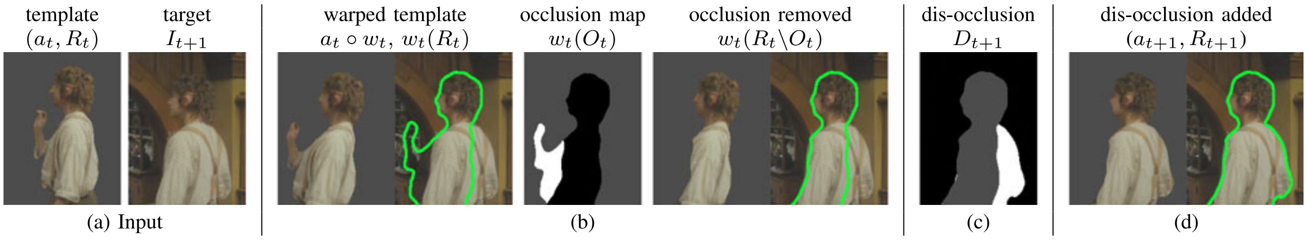


Fig. 2. Illustration of frame processing in our algorithm. (a): Estimate at frame t of the shape and radiance (a_t, R_t), and the next image I_{t+1}. (b): Simultaneous non-rigid warping and occlusion estimation is performed (first image: warped template $a_t \circ w_t$, second: boundary of warped template in I_{t+1}, third: warped occlusion $w_t(O_t)$ determined, fourth: warped template with warped occlusion removed $w_t(R_t\backslash O_t)$, fifth: boundary of $w_t(R_t\backslash O_t)$). (c): Dis-Occlusion D_{t+1} in I_{t+1} determined from input $w_t(R_t\backslash O_t)$. (d): Final shape and radiance (a_{t+1}, R_{t+1}) in frame $t+1$ (adding dis-occlusion D_{t+1} to $w_t(R_t\backslash O_t)$). Shaded gray regions indicates not defined.

relevant sub-problem for the energy of interest in Section 4.1. The full optimization scheme for the joint energy in the warp and occlusion is presented in Section 6.1. The optimization for the dis-occlusion energy is presented in Section 6.2. Finally, in Section 7, we derive a recursive estimation procedure and integrate all steps. See Fig. 2 for a system overview.

4 Energy Formulation

The first section in this section formulates a joint energy for the warp of a template to an unknown subset in an image, and the occluded subset of the template. The next section formulates an energy for the dis-occlusion.

4.1 Joint Energy for the Warp and Occlusion

We model the warp w_t as a diffeomorphism from $R_t\backslash O_t$, the co-visible region, to an unknown target set in the domain of I_{t+1}, which must be determined. A diffeomorphism is a smooth invertible non-rigid transformation whose inverse is also smooth. An occlusion of region R_t is the subset of R_t that goes out of view in frame $t+1$. We compute occlusions as the subset of R_t that *does not register* to I_{t+1} under a viable warp. Thus, the occlusion depends on the warp, but to determine an accurate warp, data from the occluded region must be excluded, hence a circular problem. Therefore, occlusion detection and registration should be computed jointly.

We avoid subscripts t for ease of notation in the rest of this section, and all sections until Section 7. Given a region $R \subset \Omega$, the radiance $a : R \to \mathbb{R}^k$, and $I : \Omega \to \mathbb{R}^k$, we formulate the problem of computing the occluded part O of R, the warp w, and $w(R\backslash O)$. Note that these quantities must satisfy $I(x) = a(w^{-1}(x)) + \eta(x)$ for $x \in w(R\backslash O)$, where η is the noise modeled in (2).

The warp w is a diffeomorphism in the un-occluded region $R\backslash O$. For ease in the optimization, we consider w to be extended to a diffeomorphism on all of R. The warp of interest will be the restriction to $R\backslash O$. We setup an optimization problem to determine w so that $w(R\backslash O)$ is the object region in I, i.e., $a|R\backslash O$ should correspond to $I|w(R\backslash O)$ via the warp w. We formulate the energy, to be minimized in O, w, as

$$E_o(O, w; I, a, R) = \int_R f(w(x), x)\,\mathrm{d}x + \beta_o \mathrm{Area}(O), \tag{3}$$

$$f(y, z) = \rho((I(y) - a(z))^2)\overline{\chi}_O(z), \tag{4}$$

where $\beta_o > 0$ is a weight, $\overline{\chi}_O(x) = 1 - \chi_O(x)$, χ_O is the indicator or characteristic function of O, and $\rho : \mathbb{R}^+ \to \mathbb{R}$ is some monotonic function. For example, $\rho(x) = x$ for a quadratic penalty or $\rho(x) = \sqrt{x + \delta}$, where $\delta > 0$ for a robust penalty [47]. The choice of ρ will depend on the noise model η in (2). The first term penalizes deviation of the object radiance, a, to the pull-back of the image intensity $I\,|\,w(R)$ under w onto the region R. The factor $\overline{\chi}_O(x)$ implies that w is only required to warp the radiance to match the image intensity I in the *un-occluded region* $R\backslash O$. The occlusion area penalty is needed to avoid the trivial solution $O = R$. Given a moderate frame rate of the camera, it is realistic to assume that the occlusion is small in area compared to the object.

Due to the aperture problem, multiple warps w can optimize the energy E_o. Typically a regularization term is added directly into the energy (e.g., for small warps as in optical flow [48], or for large warps [41]), changing the energy. In contrast, we regularize the *flow optimizing* E_o in a way that optimizes E_o without changing it, leading to a favorable solution. This is described in Section 5.

4.2 Energy Formulation of Dis-Occlusion

We now describe the energy formulation of the dis-occlusion $D_{t+1} \subset \Omega$ of the object at frame $t+1$ given the warped co-visible region $w_t(R_t\backslash O_t)$ determined from the optimization of the energy in the previous section, and the image I_{t+1}. To determine the dis-occluded region of the object, the region of the object that comes into view in the next frame, it is necessary to make a prior assumption on the 3D object.

A realistic assumption is self-similarity of the 3D object's radiance, that is, the radiance of the 3D object in a patch is similar to other patches. To translate this prior into determining the dis-occlusion of the object D_{t+1}, we assume that the image in the dis-occluded region of the object is similar to parts of the image I_{t+1} in $w_t(R_t\backslash O_t)$. For computationally efficiency, we assume similarity to close-by parts of the template. This is true in many cases, and is effective as shown in the experiments.

Although dis-occlusions in image I_{t+1} are parts of the image that do not correspond to I_t, i.e., an occlusion backward in time, these parts may be a dis-occlusion of the object or the *background*. It is not possible to determine without additional priors which dis-occlusions are of the object of interest. Our method works directly from the prior without having to compute a backward warp.

We now setup an optimization problem for the dis-occlusion. To simplify notation, we avoid subscripts in D_{t+1} and I_{t+1}, and denote $R' = w_t(R_t\backslash O_t)$. The energy is

$$E_d(D) = -\int_D \log p(x)\,\mathrm{d}x + \beta_d \mathrm{Area}(D), \tag{5}$$

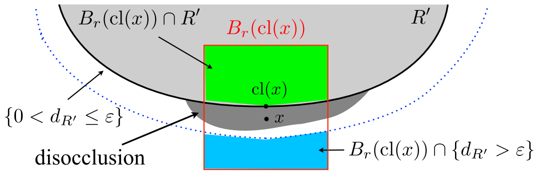

$B_r(\mathrm{cl}(x)) \cap R'$
$B_r(\mathrm{cl}(x))$
R'
$\mathrm{cl}(x)$
x
$\{0 < d_{R'} \leq \varepsilon\}$
disocclusion
$B_r(\mathrm{cl}(x)) \cap \{d_{R'} > \varepsilon\}$

Fig. 3. Diagram of quantities used in the likelihood $p(x)$ of a dis-occluded pixel. The dark gray region is the dis-occlusion to be determined. Light gray region is R', region before the dis-occlusion is determined. A pixel x within the band $\{0 < d_{R'} \leq \varepsilon\}$ is depicted, and its closest pixel to R', $\mathrm{cl}(x)$. The green (blue) region is where the foreground (background) distribution $p_{\mathrm{cl,f}}(x)$ ($p_{\mathrm{cl,b}}(x)$) is determined.

where $D \subset \Omega \backslash R'$, $p(x) \geq 0$ denotes the likelihood that $x \in \Omega \backslash R'$ belongs to the dis-occluded region, and $\beta_d > 0$. The dis-occluded region, assuming a moderate frame rate, has small area compared to the object, hence the penalty on area.

Let $\mathrm{cl}(x)$ denote the closest point of R' to x, and let $B_r(x)$ denote the ball of radius r about the point x. We choose $p(x)$ to have two components (see diagram in Fig. 3.). One measures the fit of $I(x)$ to the local distribution of I within $B_r(\mathrm{cl}(x)) \cap R'$ versus the background $B_r(\mathrm{cl}(x)) \cap \{d_{R'} > \varepsilon\}$ in I, and the second that measures nearness of x to R':

$$p(x) \propto \exp\left[-\frac{d_{R'}(x)^2}{2\sigma_d^2} + \log\frac{p_{\mathrm{cl}(x),f}(I(x))}{p_{\mathrm{cl}(x),b}(I(x))}\right], \tag{6}$$

where $d_{R'}(x)$ indicates the euclidean distance from x to R', $\sigma_d > 0$ is a weighting factor, and $p_{\mathrm{cl}(x),f}$, $p_{\mathrm{cl}(x),b}$ are Parzen estimates of the intensity distribution of I in the foreground $B_r(\mathrm{cl}(x)) \cap R'$ (respectively in the background $B_r(\mathrm{cl}(x)) \cap \{d_{R'} > \varepsilon\}$). A Parsen estimator [49] robustly estimates a distribution from samples by summing kernels (e.g., Gaussians) centered at the samples. ε is chosen large enough so that the region includes some background beyond the dis-occlusion.

5 Coarse-to-Fine Optimization of Energies Defined on Warps

In order to optimize E_o, we will apply an alternating scheme, alternating between optimization of O and w. This will be presented in Section 6. This section will focus on optimizing an energy defined on warps of the form

$$E(w) = \int_R f(w(x), x)\,\mathrm{d}x, \tag{7}$$

where $f : \Omega \times \Omega \to \mathbb{R}$. Note that this sub-problem is relevant in optimizing E_o. The optimization with respect to w is done using a steepest descent scheme. Steepest descent depends on a Riemannian metric on the space of warps, w. The Riemannian metric is defined on infinitesimal perturbations of the warp w, and the metric controls the type of motions/deformations that are favored in optimizing the energy. We will design a novel Sobolev-type metric, and use it in the steepest descent of E.

The motivation for the design of this metric comes from the active contours literature [33], [35]. It was shown that Sobolev-type metrics defined on *curves* (boundaries of regions) result in flows that optimize the energy in a coarse-to-fine manner, initially optimizing the energy with respect to coarse perturbations, and then moving to finer perturbations when coarse deformations no-longer optimize the energy. This coarse-to-fine behavior is automatically induced by the gradient descent with respect to the Sobolev metric. Motivated by this coarse-to-fine property, we design a new-Sobolev metric that is suited for energies defined on warps, that is, a *region-based metric*. The metric used in [33], [35] does not apply to the energy E in this paper as E is defined on the space of warps, and the point-wise correspondence of the interior is essential.

5.1 Sobolev Region-Based Metric and Gradient

We start by presenting some theoretical background so that the metric can be defined and the gradient of the energy with respect to the metric can be computed. The space where our energy is defined is

$$M = \{w : R \to \Omega \,|\, w : R \to w(R) \text{ is a diffeomorphism}\}, \tag{8}$$

where $R \subset \Omega \subset \mathbb{R}^2$ is a compact set with smooth boundary, and thus also the range of w's are compact and have smooth boundary. The range of $w \in M$ need not be all of Ω, but rather an arbitrary subset of Ω. We refer to M as the *space of parameterized regions* since elements $w \in M$ parameterize regions $w(R)$ via the fixed region R. Note that the parameterization of a region is important as the energy of interest E depends on the parameterization.

Infinitesimal perturbations of w are smooth vector fields $h : R \to \mathbb{R}^2$, which form the tangent space to w and is denoted T_wM. An infinitesimal perturbation of w is w_ε, given by

$$w_\varepsilon(x) = w(x) + \varepsilon h(x). \tag{9}$$

Note that if $\varepsilon > 0$ is small enough, then $w_\varepsilon \in M$, i.e., w_ε is a diffeomorphism, which implies that M is a manifold. Thus, we may define a Riemannian metric on T_wM, which in turn allows us to define gradients of the energy. Perturbations h are defined on R, and by right translation, i.e., $h \circ w^{-1} : w(R) \to \mathbb{R}^2$, they are also defined on $w(R)$. We now specify an inner product on T_wM, which makes M a Riemannian manifold:

Definition 1 (Sobolev-type Inner Product on M). *The inner product on the set of perturbations of w (i.e., the metric) that we consider is defined as follows:*

$$\langle h_1, h_2 \rangle_{Sob,w} = \overline{\hat{h}_1} \cdot \overline{\hat{h}_2} + \alpha \int_{w(R)} \mathrm{tr}\{\nabla \hat{h}_1(x)^T \nabla \hat{h}_2(x)\}\,\mathrm{d}x, \tag{10}$$

where $\alpha > 0$, $\hat{h} := h \circ w^{-1}$ when $h : R \to \mathbb{R}^2$, $\nabla \hat{h}_1(x)$ denotes the spatial Jacobian of $\hat{h}_1(x)$, tr denotes the trace of a matrix, $\mathrm{d}x$ is the area measure on $w(R)$, and

$$\overline{\hat{h}} = \frac{1}{|w(R)|}\int_{w(R)} \hat{h}(x)\,\mathrm{d}x. \tag{11}$$

The first term in (10) uses the mean value of the perturbations rather than the $\mathbb{L}^2$ inner product of the perturbations as in standard Sobolev inner products [50]. This change is for convenience in the algorithm that we present to optimize E, and an easy decomposition of the gradient into orthogonal components as we shall see. The second term of (10) is the $\mathbb{L}^2$ inner product of the Jacobian of the perturbations.

The goal now is to define a gradient (or steepest) descent approach to minimize E. It should be noted that the gradient of an energy depends on the choice of inner product on the space of perturbations of the warp. The typical choice (either implicitly or explicitly) is the $\mathbb{L}^2$ inner product, but this does not have desirable properties for tracking. We therefore, compute the gradient with respect to the Sobolev inner product defined above in (10). First, we state the definition of the gradient, which shows the dependence on the inner product.

Definition 2 (Gradient of Energy). *Let $E : M \to \mathbb{R}$, $w \in M$, $h \in T_wM$, and $\langle , \rangle_w$ denote the inner product on T_wM. The* directional derivative *of E at w in the direction h denoted, $\mathrm{d}E(w) \cdot h$, is*

$$\mathrm{d}E(w) \cdot h = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} E(w + \varepsilon h)|_{\varepsilon=0}. \tag{12}$$

The gradient of E, denoted $\nabla E(w) \in T_wM$, is the perturbation that satisfies the relation

$$\mathrm{d}E(w) \cdot h = \langle \nabla E(w), h \rangle_w, \tag{13}$$

for all $h \in T_wM$.

To show how the choice of inner product affects the gradient, we give another interpretation of the gradient, i.e., it is a perturbation that maximizes the following ratio:

$$\frac{\mathrm{d}E(w) \cdot h}{\|h\|_w}, \tag{14}$$

where $\|h\|_w = \sqrt{\langle h, h \rangle_w}$ is the norm induced by the inner product. That is, the gradient is a perturbation h that maximizes the change in energy by perturbing in direction h divided by the norm of the perturbation. Therefore, while it is often stated that the gradient is the direction that maximizes the energy the fastest, it is actually the direction that maximizes energy while *minimizing its cost* (measured by the norm).

Since non-smooth perturbations cost a lot according to the Sobolev norm, they are not typically Sobolev gradients. Coarse perturbations are favored for Sobolev gradients when they can increase the energy. Note that moving in the negative Sobolev gradient direction reduces the energy for any α.

The Sobolev gradient of E, denoted $G = \nabla_{Sob} E(w)$, is a linear combination of two orthogonal (w.r.t. (10)) components, the translation and the deformation:

$$G(x) = \overline{G} + \frac{1}{\alpha} \tilde{G}(x),\ x \in w(R), \tag{15}$$

where $\tilde{G}$, which is independent of α, satisfies the following Poisson partial differential equation (PDE):

$$\begin{cases} -\Delta \tilde{G}(x) = f_1(x, w^{-1}(x)) \det(\nabla w^{-1}(x)) & \\ \qquad\qquad - \overline{f_1(\cdot, w^{-1}(\cdot)) \det(\nabla w^{-1}(\cdot))} & x \in w(R) \\ \nabla \tilde{G}(x) \cdot N = 0 & x \in \partial w(R) \\ \overline{\tilde{G}} = 0 & \end{cases}, \tag{16}$$

where Δ is the Laplacian, N is the unit normal to $\partial w(R)$, $\overline{\tilde{G}}$ is the average of $\tilde{G}$ over $w(R)$, $\overline{G}$ is given by

$$\overline{G} = \int_{w(R)} f_1(x, w^{-1}(x)) \det(\nabla w^{-1}(x))\, \mathrm{d}x, \tag{17}$$

and f_1 denotes the partial derivative of f with respect to the first argument of f. Details of the derivations for these expressions can be found in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2014.2360380. The numerical scheme to solve (16) is given in Appendix B, available in the online supplemental material. Note that larger α (implying more spatial regularity) implies the gradient approaches a translation (the smoothest transformation). Smaller α implies a non-rigid deformation, which is spatially smooth and the amount of smoothness depends on the data.

5.2 Optimizing the Energy via Gradient Descent

The gradient flow to optimize E is then given by the following partial differential equation:

$$\begin{cases} \partial_\tau \phi_\tau(x) = -\nabla_{Sob} E(\phi_\tau)(\phi_\tau(x)), & x \in R, \\ \phi_0(x) = x, & x \in R, \end{cases} \tag{18}$$

where τ indicates an artificial time parameter parameterizing the evolution of the warp $\phi_\tau : R \to \Omega$ at a given frame in the image sequence (not to be confused with the frame number t). The final converged ϕ_τ is a local optimizer of the energy E. It should be noted that the above equation maintains that $\phi_\tau \in M$, i.e., that the final converged result is a diffeomorphism. This can be seen since $\nabla_w E$ is smooth (it is the solution of a Poisson equation and thus, H^2 [50]), and integrating a smooth vector field results in diffeomorphism using classical results [7], and in particular, [51] for first order Sobolev regularity. Precise details for this fact are out of the scope of this paper.

In implementing the gradient flow (18), we are interested in the final converged region, and thus we keep track of $R_\tau = \phi_\tau(R)$. For numerical ease and accuracy, we keep track of R_τ using a level set method [52], although it is not required. We also keep track of the backward map ϕ_τ^{-1}, which is needed to evaluate the gradient $\nabla_{Sob} E(\phi_\tau)(\phi_\tau(x))$.

The level set function will be denoted $\Psi_\tau : \Omega \to \mathbb{R}$. Its evolution is described by a transport PDE. The backward map ϕ_τ^{-1} also satisfies a transport equation. Therefore, the optimization of E is given by the coupled PDE:

$$\Psi_0(x) = d_R(x), x \in B_2(R) \tag{19}$$

$$\phi_0^{-1}(x) = x, x \in R_0 = R \tag{20}$$

$$G_\tau = \nabla_{Sob} E(\phi_\tau) \tag{21}$$

$$\partial_\tau \phi_\tau^{-1} = \nabla \phi_\tau^{-1}(x) \cdot G_\tau(x), x \in R_\tau \tag{22}$$

$$\partial_\tau \Psi_\tau = \nabla \Psi_\tau(x) \cdot G_\tau(x), x \in B_2(R_\tau) \tag{23}$$

$$R_\tau = \{\Psi_\tau < 0\}, \tag{24}$$

where ∂_τ denotes partial with respect to τ, and $B_2(R_\tau) = \{x \in \Omega : |d_{R_\tau}(x)| \le 2\}$ where d_{R_τ} is the signed distance function of R_τ. The region R_τ is updated in minus the gradient of E, $-G_\tau : R_\tau \to \mathbb{R}^2$, direction via the level set evolution. Note G_τ is extended to $B_2(R_\tau)$ as in narrowband level set methods. The backward warp $\phi_\tau^{-1} : R_\tau \to R$ is computed by flowing the identity map along the velocity field $-G_\tau$ up to time τ, and this is accomplished by the transport equation (22). The convergence time, i.e., the time at which E does not decrease, is denoted by τ_∞. At τ_∞, $w = \phi_{\tau_\infty} : R \to R_{\tau_\infty}$ is a local minimum of E, and $R_{\tau_\infty} = w(R)$ is the region matched in image I. Note that w can be computed as $w = (\phi_{\tau_\infty}^{-1})^{-1}$.

The evolution above is automatically coarse-to-fine for any choice of α, that is, the gradient descent favors coarse motions/deformations initially before transitioning to finer scale deformations. See Fig. 5 in Section 5.4 for an experimental verification of this property.

5.3 Parameter Independent Optimization

One of the advantages of the form of the Sobolev-type metric chosen in (10) besides the coarse-to-fine property is that one can eliminate the need for choosing the parameter α, while optimizing E. One can take $\alpha \to \infty$, in which case $G \to \overline{G}$, a translation motion. One can optimize by translating in the direction $-G \to -\overline{G}$ when $\alpha \to \infty$, until convergence. At convergence, $\overline{G} = 0$, then one can evolve the warp infinitesimally in the negative gradient $-G = -\tilde{G}/\alpha$ direction for any finite α. Since the gradient depends only on α by a scale factor, the choice of α is just a time re-parameterization of the evolution, not changing the geometry of the evolution. It does not impact the final converged warp nor the converged region. The algorithm to optimize E that is not dependent on the choice of α is summarized in the following steps:

1) Perform the initializations (19)-(20).
2) Repeat the evolution (21)-(24) with $\alpha \to \infty$, in which case $G_\tau = \overline{G_\tau}$, until convergence (when $\overline{G_\tau} = 0$).
3) Perform one time step (21)-(24) with the deformation $G_\tau \propto \tilde{G}_\tau$. One may choose $\alpha = 1$, but any choice would give the same result.
4) Repeat Steps 2-3 until convergence (when E does not decrease).

The procedure above optimizes with respect to translations until convergence, then optimizes with respect to deformations that are not translations (favoring coarse deformations that optimize the energy), and the process is iterated. This results in a scheme independent of a regularity parameter α. The scheme favors a coarse-to-fine evolution, like the gradient descent with any fixed α, of the region R_τ and coarse-to-fine motion/deformation estimation.

5.4 Discussion

We relate our approach to Lucas and Kanade [53] and energy regularization methods of optical flow (e.g., [48], [47]).

Since there are multiple optimizers of E, which contains only data fidelity, regularization is needed to determine a viable solution. Lucas and Kanade [53] restrict the possible warps to a smaller set rather than the space of diffeomorphisms, i.e., translations, affine motions, or other parametric groups. While providing a unique optimizer of E, this restricts the possible warps w and thus also the shape of the region. One may consider optimizing E with respect to translations first, thus obtaining a coarse estimate of the desired region in image I, then resort to optimizing in finer transformations, e.g., euclidean transformations, then affine transformations. However, one may go up to the projective group, and then it becomes unclear what group to choose to optimize further. The algorithm that we have presented optimizes the energy by using coarse perturbations initially, it then transitions *continuously* and *automatically* to finer-scale perturbations, in fact, it transitions through all possible scales of motions/deformations, eliminating the need to choose groups of motions to optimize. This property of Sobolev-type metrics for contours was shown analytically using a Fourier analysis in [35]. Since this paper deals with regions, the property is harder to show analytically since a shape-dependent Fourier basis would need to be derived on the region. We therefore demonstrate the property in an experiment.

Energy regularization methods (e.g., [48], [47]), deal with multiple optimizers of E by *changing* the original energy by adding regularization of the warp directly into the energy. The energy in [47] for infinitesimal warps is

$$E_{BA}(v; a, I, R) = \int_R \rho(|I(x) - a(x) + \nabla a(x) \cdot v(x)|)\,\mathrm{d}x + \gamma \int_R \rho(|\nabla v(x)|)\,\mathrm{d}x. \tag{25}$$

An advantage of this approach over [53] is that motions/deformations are not restricted to finitely parameterized groups. The parameter γ controls the scale of the motion: large γ implies coarse motion, and small γ implies finer motion.

One can use Black and Anandan [47] optical flow (or other energy regularization approaches) determined from data within a template to match a template to the next frame. This is accomplished by deforming the region R by v infinitesimally to obtain R_τ, then recalculating v based on the warped radiance $a \circ \phi_\tau^{-1}$, and iterating the process. This is the same as (19) to (24), but replacing G_τ in (21) with the minimizer of $E_{BA}(v; a \circ \phi_\tau^{-1}, I, R_\tau)$, the Black and Anandan velocity determined from data within the template. We will call this approach template B&A. Template B&A allows the cumulative warp w to be more flexible than [53], obtaining arbitrarily shaped regions, but γ must be chosen. Large γ yields only coarse approximations of the region shape, and small γ yields finer details of shape, but is likely to be trapped in fine details of the image before reaching the desired region. No one scale of motions/deformations, i.e., no one γ is sufficient. Further, template B&A does not optimize a common energy for the warp w.

One ad-hoc solution to choosing γ is to attempt a coarse-to-fine scheme by starting with γ large until the region converges, reduce γ and then deform the region until convergence, reduce γ, etc., which is the scheme considered in our conference paper [10]. While the procedure solves the issue of choosing γ and is coarse-to-fine, Sobolev descent has three advantages. First, Sobolev does not rely on an ad-hoc scheme to reduce the parameter γ. Second, Sobolev *automatically* and *continuously* traverses through *all* scales of

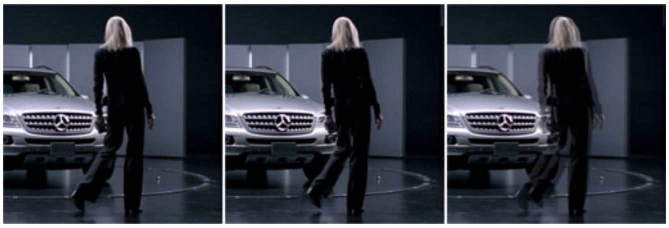


Fig. 4. Images I_1 (left) and I_2 (middle) used in the experiment in Fig. 5, and an overlay of I_1 on I_2 to show the motion/deformation between frames, which is non-rigid and contains both coarse and fine motion/deformations.

motions/deformations roughly favoring coarse-to-fine transition, whereas the ad-hoc scheme only traverses through a discrete number of scales (chosen by the scheme to reduce γ) and the transition is not automatic. Reducing γ monotonically in the ad-hoc scheme may not always be beneficial (e.g., when new coarse structure is "discovered" from the data during evolution and larger γ is then needed). Sobolev chooses the appropriate scale of deformation implicit in the computation of the gradient. Computing the Sobolev gradient is fast; it has similar computational cost as computing velocity in Horn and Schunck [48]. Sobolev is thus more convenient for practical applications. Lastly, our scheme is minimizes E, while the ad-hoc scheme does not necessarily minimize an energy.

We illustrate the coarse-to-fine behavior of region-based Sobolev by matching a template of the object (woman) obtained from image 1 to image 2 shown in Fig. 4. The motion contains both coarse and fine-scale deformations. The evolution (at various snapshots) of region-based Sobolev and template B&A is shown in Fig. 5. Final objects detected with these schemes in a zoomed region of interest is shown in Fig. 6. The displacement, $d_{\tau_i,\tau_{i+1}}(x) = \phi_{\tau_i} \circ \phi^{-1}_{\tau_{i+1}}(x) - x$, between two time instances τ_i and τ_{i+1} is shown in color code [54] (the color indicates direction and darkness indicates magnitude; magnitude should not be compared across images as they are re-scaled in each image) in Fig. 5. Region-based Sobolev moves according to coarse motions (constant color) before resorting to finer deformations whereas template B&A has roughly the same scale of motions/deformation at all stages of the evolution for each γ. Small γ does not capture regions of coarse deformation and is stuck in intermediate structures. Larger γ captures regions of coarse deformation, but regions of finer motion (e.g., the legs) are not captured. Other energy regularization optical flow approaches, e.g., Horn and Schunck [48], similarly cannot recover the desired object, as no one γ is able to recover deformation at multiple degrees of locality. Notice that the Sobolev evolution starts with a translation, moves to coarse deformation of both legs, and ends with fine deformations of the feet.

6 Occlusion/Dis-Occlusion Computation and Alternating Optimization

We now describe the alternating optimization scheme to optimize E_o, combining the coarse-to-fine optimization scheme described in the previous section, and optimization in the occlusion, which we describe next. We then present the optimization scheme to determine the dis-occlusion.

6.1 Joint Occlusion and Warp Optimization

Given an estimate w, one can solve for a global optimizer of the energy E_o. The energy can be written as (with $O \subset R$)

$$E_o(O \,|\, w; I, a, R) = \int_{R\backslash O} \rho((I(w(x)) - a(x))^2)\,\mathrm{d}x + \int_O \beta_o\,\mathrm{d}x. \quad (26)$$

The optimization problem can be thought of as an assignment problem where points $x \in R$ are assigned to the occlusion O or the co-visible region $R\backslash O$. If x is assigned to O, then it adds to the energy an amount β_o, whereas, if it is assigned to $R\backslash O$, it adds to the energy an amount $\rho((I(w(x)) - a(x))^2)$. Therefore to minimize the energy, we assign pixels to the occlusion based on

$$O = \left\{x \in R \,:\, \rho\big((I(w(x)) - a(x))^2\big) > \beta_o\right\} \quad (27)$$

$$= w^{-1}\{x \in w(R) \,:\, \rho((I(x) - a(w^{-1}(x))^2)) > \beta_o\}, \quad (28)$$

which is a global optimizer of E_o conditioned on w.

The alternating scheme to optimize E_o in both O and w is a modification of the scheme presented in Section 5.2 to update the occlusion during the evolution. It is initialized as

$$\Psi_0(x) = d_R(x),\ x \in B_2(R) \quad (29)$$

$$\phi_0^{-1}(x) = x,\ x \in R_0 = R \quad (30)$$

$$\tilde{O}_0 = \emptyset. \quad (31)$$

Then the following is iterated until convergence:

$$G_\tau = \nabla_{Sob} E(\phi_\tau \,|\, O_\tau, R_\tau, I) \quad (32)$$

$$\partial_\tau \phi_\tau^{-1} = \nabla\phi_\tau^{-1}(x) \cdot G_\tau(x), x \in R_\tau \quad (33)$$

$$\partial_\tau \Psi_\tau = \nabla\Psi_\tau(x) \cdot G_\tau(x), x \in B_2(R_\tau) \quad (34)$$

$$R_\tau = \{\Psi_\tau < 0\} \quad (35)$$

$$\tilde{O}_\tau = \{x \in R_\tau \,:\, \rho((I(x) - a(\phi_\tau^{-1}(x))^2)) > \beta_o\}, \quad (36)$$

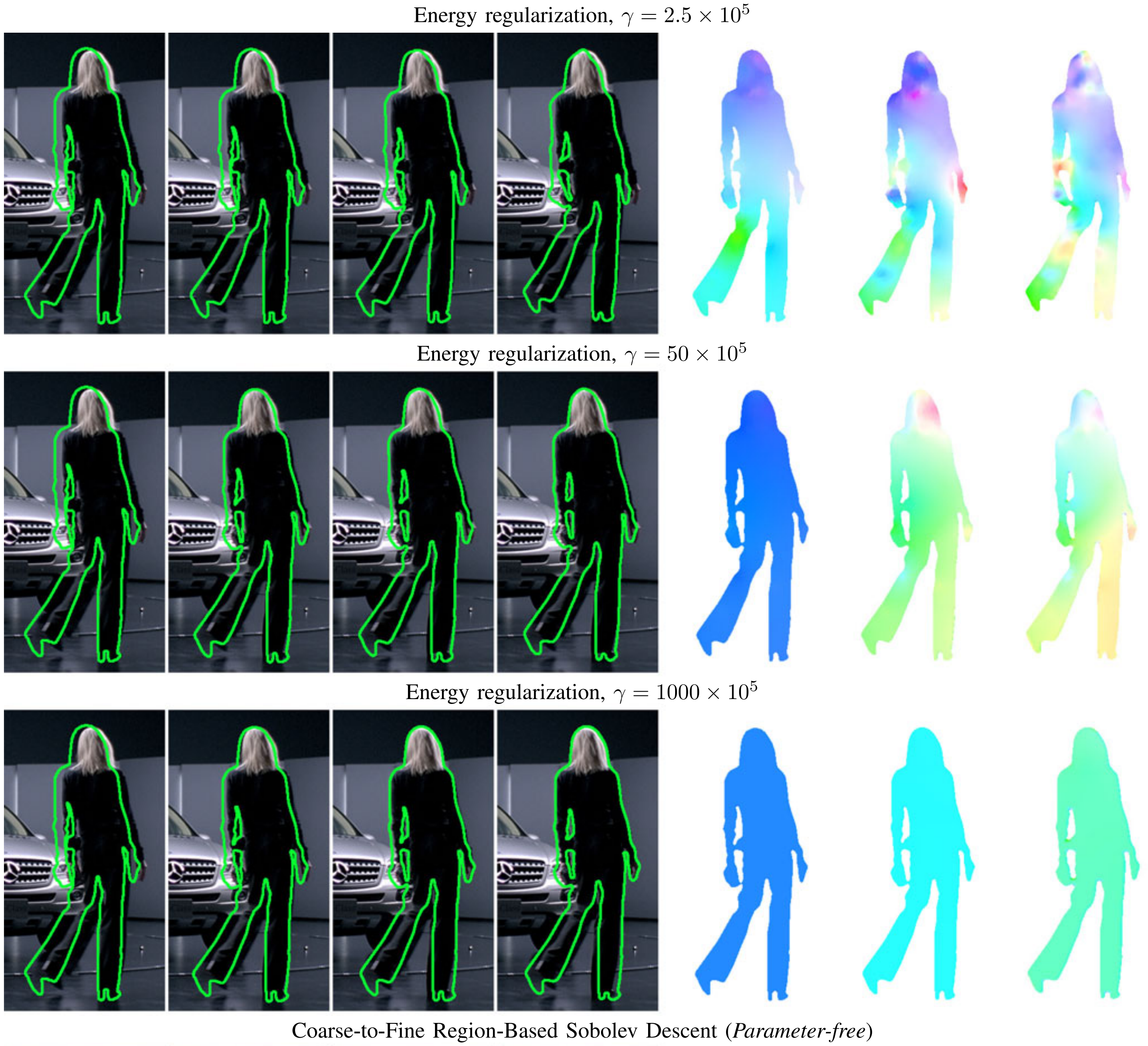
where $\tilde{O}_\tau = \phi_\tau(O_\tau)$ indicates the current estimate of the warped occlusion. Note that only $\tilde{O}_\tau$ is needed to compute the gradient G_τ, and thus we do not explicitly compute O_τ. Note that G_τ is specified by $\overline{G_\tau}$ and $\tilde{G}_\tau$, where $\tilde{G}_\tau$ satisfies the Poisson equation (16) with $w^{-1} := \phi_\tau^{-1}$, and

$$\begin{aligned} f_1\big(x, \phi_\tau^{-1}(x)\big) &= \rho'(|I(x) - a_\tau(x))|^2) \\ &\quad \times (I(x) - a_\tau(x))\nabla I(x)\chi_{\tilde{O}_\tau}(x),\ x \in R_\tau \end{aligned} \quad (37)$$

$$a_\tau(x) = a\big(\phi_\tau^{-1}(x)\big), \quad x \in R_\tau. \quad (38)$$

Discretization of (29)-(36) and numerical implementation is given in Appendix B, available in the online supplemental material.

Let $\tau = \tau_\infty$ be the time of convergence. R_{τ_∞}, a warping of R, includes a warping of the occluded region O_{τ_∞}, and thus the warping of the un-occluded region is $w(R\backslash O_{\tau_\infty}) = R'_{\tau_\infty} =$



Energy regularization, $\gamma = 2.5 \times 10^5$

Energy regularization, $\gamma = 50 \times 10^5$

Energy regularization, $\gamma = 1000 \times 10^5$

Coarse-to-Fine Region-Based Sobolev Descent (*Parameter-free*)

Fig. 5. Coarse-to-fine behavior of region-based sobolev descent. Matching a template (obtained from I_1) to I_2 from Fig. 4 using regularization of the velocity field in the energy, and Sobolev descent. In each row, the evolution (until convergence) is shown. [First four images]: ∂R_τ on I_2 for various snapshots τ. [Last three images]: displacement of object between adjacent snapshots (in optical flow color code). Small γ favors fine deformations and is sensitive to intermediate structures, whereas large γ favors only coarse deformations and cannot capture regions with fine-scale deformations, e.g., legs. Sobolev descent captures all scales of deformation without being sensitive to intermediate structures.

$R_{\tau_\infty} \backslash \tilde{O}_{\tau_\infty}$. This does not include the dis-occluded region, which is computed in the next section from R'_{τ_∞}. To ensure spatial regularity of R'_{τ_∞}, at convergence of (29)-(36), we induce spatial regularity into O_{τ_∞} by using the estimate

$$\tilde{O}_{\tau_\infty} = \{x \in R_{\tau_\infty} : (G_\sigma * \text{Res})(x) > \beta_o\} \tag{39}$$

$$\text{Res}(x) = \rho((I(x) - a(\phi^{-1}_{\tau_\infty}(x)))^2), \tag{40}$$

where G_σ denotes an isotropic Gaussian kernel.

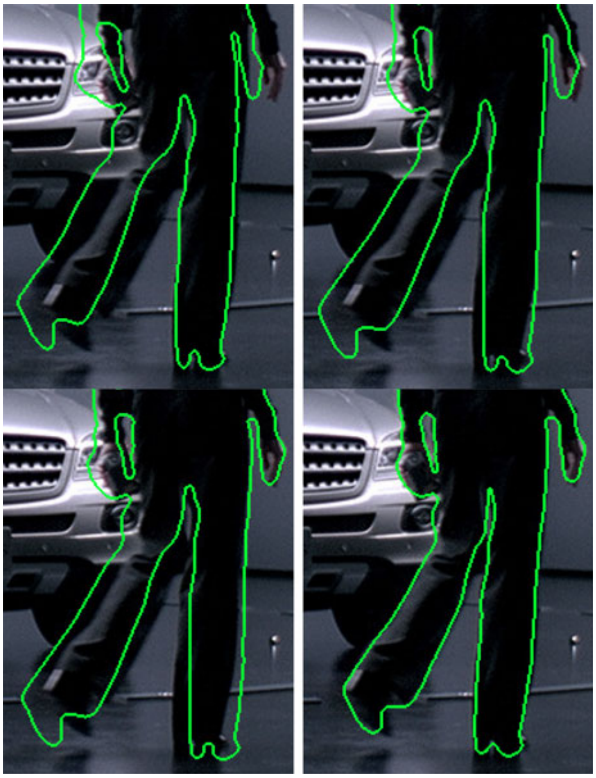


Fig. 6. Zoom of converged results of experiment of Fig. 5. Boundary of converged region on I_2. [Top-left]: energy regularization $\gamma = 2.5 \times 10^5$, [Top-right]: energy regularization $\gamma = 50 \times 10^5$, [Bottom-left]: energy regularization $\gamma = 1{,}000 \times 10^5$, [Bottom-right]: region-based Sobolev. Notice that small γ misses regions of coarse motion, larger γ obtains regions of coarse motion, but misses regions where finer deformation occurs. Sobolev obtains both coarse and fine deformations.

Fig. 7 shows the evolution (29)-(36) on an example, and the final co-visible region R'_{τ_∞}.

6.2 Dis-Occlusion Optimization

We show how to optimize the dis-occlusion energy E_d (5). The global minimum of E_d is computed in a thresholding step from the likelihood p. Since p decreases exponentially with distance to R', we assume that $D \subset \{0 < d_{R'} < \varepsilon\}$. The dis-occlusion is computed as

$$D = \{x \,:\, d_{R'}(x) \in (0, \varepsilon], (G_\sigma * p)(x) > \exp(\beta_d)\}, \tag{41}$$

where $\sigma = 0$ corresponds to the global optimum, but to ensure spatial regularity of D, we choose $\sigma > 0$. The choice of β_d is based on the frame-rate of the camera and the speed of the object (the more the speed and the less the frame-rate, the smaller β_d). Fig. 8 shows an example of p, the dis-occlusion detected, and the final estimate of the region.

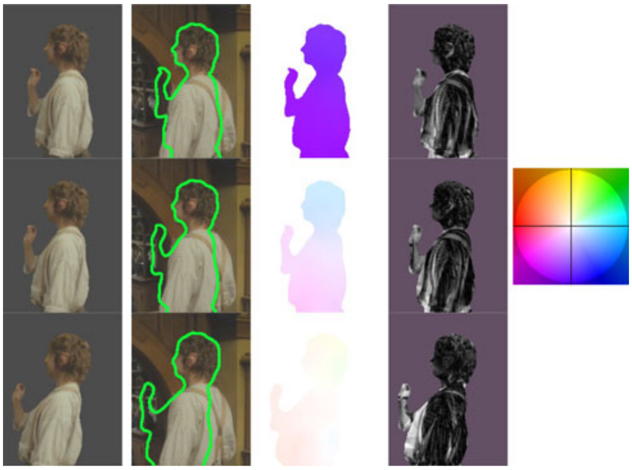


Fig. 7. Occlusion estimation and warping. [Top to bottom]: Beginning ($\tau = 0$), intermediate, and final stages of evolution. [first column]: radiance a_τ, [second]: target image I and boundary of R_τ, [third]: velocity $-G_\tau$, [fourth]: occlusion estimation Res at time τ, [fifth]: optical flow color code. The final occluded region is shown in Fig. 2b.

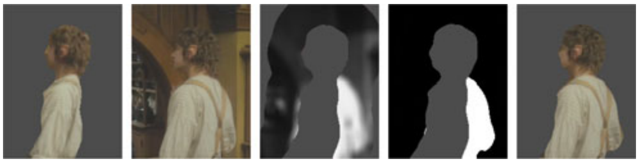


Fig. 8. Illustration of disocclusion detection.[first]: warped un-occluded radiance defined on R' (after occlusion and deformation computation), [second]: target image I, [third]: likelihood of dis-occlusion map p (defined in $B_{R'}(\varepsilon)$), [fourth]: computed dis-occlusion D (white), and [fifth]: final radiance. Boundary of final region super-imposed on I is in Fig. 2d.

Computation of $d_{R'}$ in $\{0 < d_{R'} < \varepsilon\}$ is done efficiently with the Fast Marching Method [55], and $\mathrm{cl}(x)$ at each point is simultaneously propagated as the front in the Fast Marching Method evolves. Then p is readily computed.

7 Filtering Radiance Across Frames

We integrate the results of occlusion/deformation estimation and dis-occlusion estimation into a final estimate of the shape and radiance in each frame. To deal with modeling noise (specified in (2)), we filter the radiance in time.

Given the image sequence I_t, $t = 1 \dots, N$ and an initial template $R_0 \subset \Omega$, $a_0 : R_0 \to \mathbb{R}^k$, the final algorithm is as follows. For $t = 1, \dots, N$, the following steps are repeated:

1) Compute the warping of R_{t-1} and O_{t-1}: $w_{t-1}(R_{t-1})$ and $w_{t-1}(O_{t-1})$, resp., and $a'_t = a_{t-1} \circ w_{t-1}^{-1}$ defined on $w_{t-1}(R_{t-1})$ using the optimization scheme described in Section 6.1 with input R_{t-1}, a_{t-1} and I_t.
2) Given $R'_t = w_{t-1}(R_{t-1}) \backslash w_{t-1}(O_{t-1})$, the warping of the un-occluded part of R_{t-1}, and the image I_t, compute the dis-occlusion D_t using (41). The estimate of R_t is then $R'_t \cup D_t$.
3) The radiance is then updated as

$$a_t(x) = \begin{cases} (1 - K_a)a'_t(x) + K_a I_t(x), & x \in R'_t, \\ I_t(x), & x \in D_t, \end{cases} \tag{42}$$

where $K_a \in [0, 1]$ is the gain, a constant.

The averaging of the warped radiance and the current image (42) combats modeling noise η in (2). In practice, K_a is chosen large if the image is reliable (e.g., no specularities, illumination change, noise, or any other deviations from brightness constancy), and small otherwise.

8 Experiments and Comparisons

We demonstrate our method on a variety of videos that contain self-occlusions/dis-occlusions. All examples

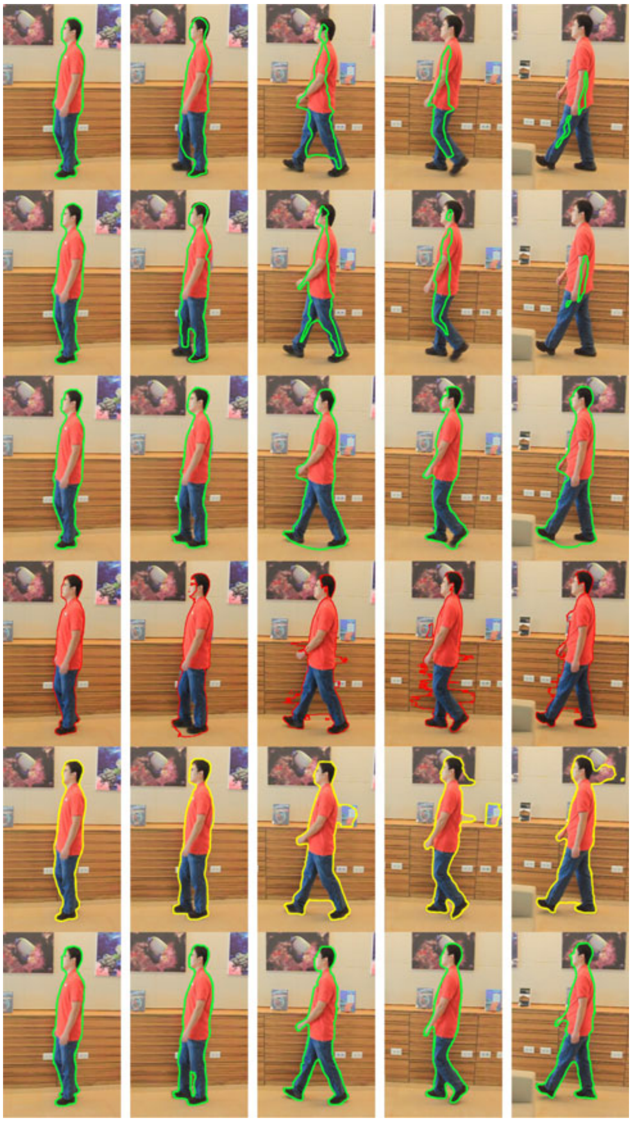


Fig. 9. Modeling occlusions/dis-occlusions is necessary. [first row]: occlusion/dis-occlusion detection are turned off in our method. [second]: occlusion modeling done, but not dis-occlusions in our method. [third]: dis-occlusions detected but not occlusions. [fourth]: result of Scribbles. [fifth]: result of AAE. [sixth]: accurate tracking when both occlusion and dis-occlusion modeling is performed (our final result).

shown have over 100 frames.[1] To demonstrate that occlusion/dis-occlusion modeling aids joint shape/appearance tracking, we compare to Adobe After Effects CS6 2012 (AAE) (based on [20], but significantly extended), which employs localized joint shape and appearance information without explicit occlusion modeling. Note that AAE has an interactive component to correct errors in the automated component; we compare to the automated component to show less interaction would be required with our approach. To show advantages over tracking by partitioning elementary image statistics, we compare to Scribbles [4] (publicly available code and results optimized for parameters), which is a recent technique that employs global statistics in addition to other advanced techniques.

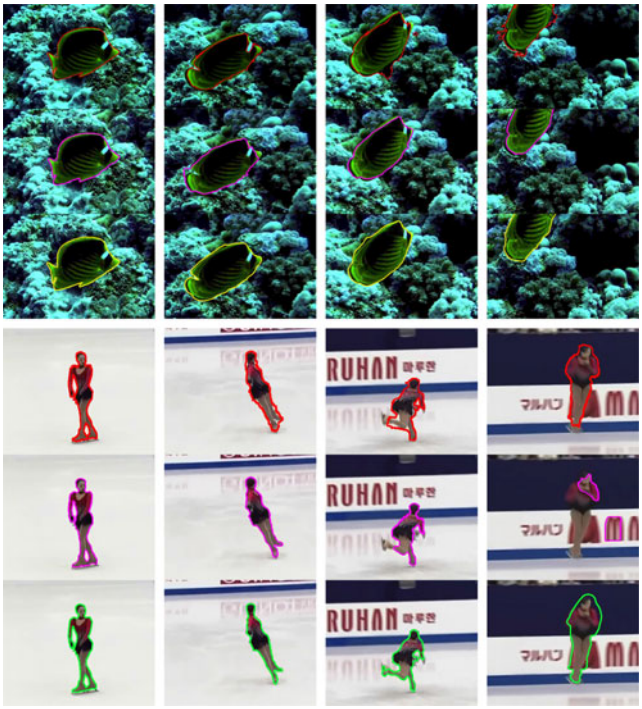


Fig. 10. Distinctive foreground/background global statistics. [Top]: Scribbles, [Middle]: AAE, [Bottom]: proposed method. When fore/background global statistics are separable, Scribbles, and AAE, for minor occlusions, performs well.

Parameters are chosen as: $\sigma = 5$ in (41) and (39), $\sigma_d = 100$ in the likelihood, p in (6), the band thickness for the domain of p is $\varepsilon = 30$, and the radius of B_r in $p_{f,x}$ and $p_{b,x}$ is $r = 3\varepsilon$ (i.e., a $6\varepsilon \times 6\varepsilon$ window). The threshold for the occlusion stage is $\beta_o = \text{Res}_{min} + 0.3 \times (\text{Res}_{max} - \text{Res}_{min})$ where Res_{max} (Res_{min}) denotes the maximum (minimum) value of smoothed residual. The threshold for the dis-occlusion stage is $\exp\beta_d = 0.5$ when p is normalized to be a probability. The gain in the radiance update (42) is $K_a = 0.8$. Parameters are fixed for the whole video. Parameter sensitivity analysis is shown at the end of the Section.

Initialization. Precise initialization in frame 1 is not needed, as initialization inside the object can be corrected by running dis-occlusion detection. Outside initialization will be self-corrected in joint warp/occlusion estimation as the background moves differently than the object, and would be detected as occlusion and removed. This is true when the background area in the initialization is less than the object area.

The first experiment (Fig. 9) shows that occlusion and dis-occlusion modeling is vital. As the man in the sequence walks forward, his legs, arms and back are self-occluded/dis-occluded. Ignoring occlusions (setting $\tilde{O}_\tau = \emptyset$ in Section 6.1) and dis-occlusion detection, the shape is inaccurate (first row). Using occlusion modeling but not dis-occlusions (second row), it is possible to discard the portion of the background between the legs, and the occluded right hand in the first frame is removed. Using the dis-occlusion modeling but not occlusions (third row), dis-occluded parts of the body are detected. However, irrelevant regions of the background (that can be removed in the occlusion stage) are

[1] Videos for all experiments are available on the following website: http://vision.ucla.edu/ ganeshs/articulated_object_tracking_html/pami_supp.html

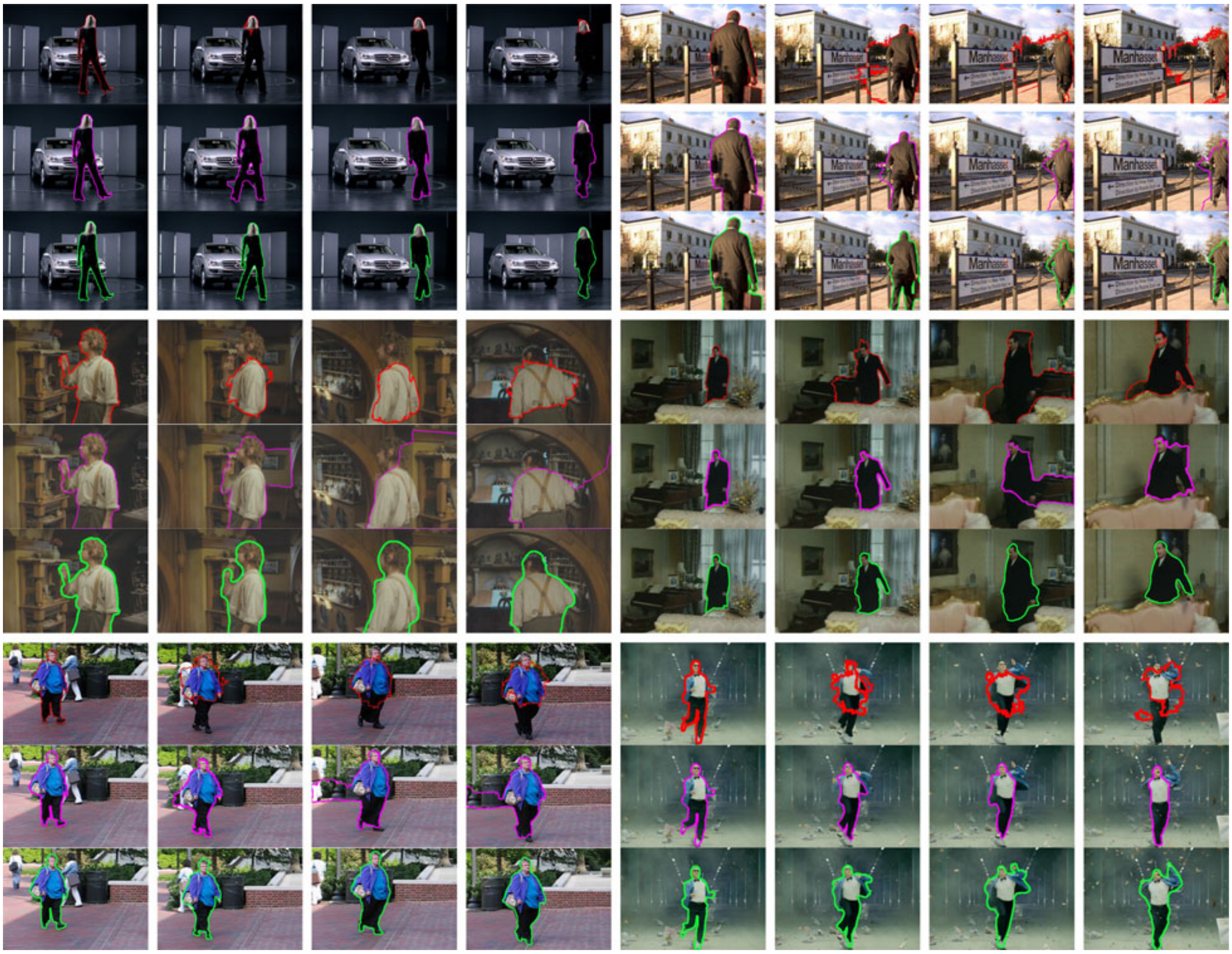


Fig. 11. Occlusions/dis-occlusions, violations of brightness constancy, and foreground/background not easily separable. [Top]: Scribbles, [Middle]: Adobe After Effects 2012, [Bottom]: proposed method. Methods based on foreground/background image statistic discrimination leak into the background. Note 4 (out of about 100-200 for most sequences) frames are selected for display in each sequence.

captured. Best results (last row) are achieved when both the occlusion and dis-occlusions are modeled. The fourth row shows the result of Scribbles. It has trouble discriminating between face and the background, which have similar radiance. The fifth row shows the result of AAE, which captures irrelevant background.

Fig. 10 shows tracking of a fish and a skater. When foreground/ background global histograms are easily separable, Scribbles performs well, and when occlusions are minor AAE, performs well as does the proposed method.

In Fig. 11, we have tested our algorithm on challenging video (more than 100 frames per sequence) exhibiting self-occlusions and dis-occlusion (crossing legs, viewpoint change, rotations in depth), complex object radiance and background in which it becomes difficult to discriminate between foreground and background global statistics (e.g., the woman's pants have same radiance as car tires). Deviations from brightness constancy are clearly visible (small illumination change, specular reflections, and even shadows). The latter are handled with our dynamic radiance update. In these sequences, Scribbles and Adobe After Effects 2012 have trouble discriminating between object and background which share portions of similar intensity, and occlusions (e.g., crossing of legs). In the "Lady Mercedes," sequence (top left), after a few frames, Scribbles can only track the head of the lady. This is because the lady's clothing shares similar intensity as the tires of the car and some of the background. Thus, the tracker confuses the clothing with the background and only tracks the head, which has different statistics from the rest of the images. Our method is able to capture the shape of the objects quite well (quantitative assessment is in Table 1). The man at the station (top right group) at the fourth column shows a limitation of our dis-occlusion detection: dis-occluded parts of the object that

TABLE 1
Quantitative Performance Analysis

Sequence	Scribbles [4]	Adobe Effects 2012 [20]	Ours
Library	0.8926	0.9193	0.9654
Fish	0.9239	0.9513	0.9792
Skater	0.8884	0.6993	0.9086
Lady	0.2986	0.8243	0.9508
Station	0.5367	0.8258	0.9216
Hobbit	0.7312	0.5884	0.9335
Marple	0.6942	0.8013	0.9186
Lady 2	0.7457	0.7909	0.9584
Psy	0.6163	0.8845	0.9329

Average F-measure (over all frames) computed from ground truth are shown. larger F-measure means better performance.

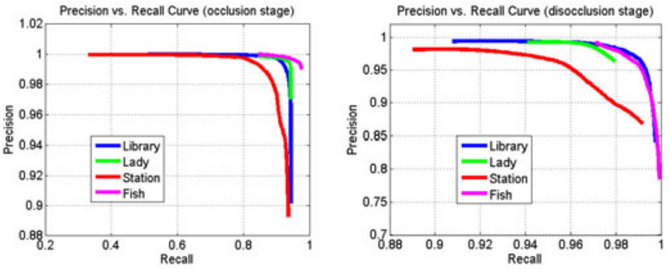


Fig. 12. Sensitivity of key parameters. Quantitative assessment of the sensitivity of the key parameters (i.e., thresholds), β_o and β_d, of the proposed algorithm for the occlusion and disocclusion stages in the sequences above. The Precision/Recall curves indicate robustness to a wide range of thresholds.

do not share similar radiance as the current template (sole of shoe) are not detected. A variety of other videos are processed, and our method performs well.

In Fig. 12, we show sensitivity analysis of the key parameters. We analyze β_o and β_d in the occlusion and disocclusion detection stages using a precision / recall (PR) curve. For four image sequences, we choose a pair of images so that significant occlusion and disocclusion are present between the frames (about five frames apart on a 30 fps video), and significant deformation and motion is present. Given a hand cutout in the first frame, we run our algorithm to obtain the cutout in the next frame. The first image in Fig. 12 shows the PR curve as the parameter β_o is varied between its valid range (the minimum value of the residual, Res, and its maximum value), and the threshold of the disocclusion stage β_d is kept fixed. The second image in Fig. 12 shows the PR curve as the parameter β_d in the disocclusion stage is varied between its valid range (the minimum and maximum value of p), and the threshold in the occlusion stage β_o is kept fixed. High precision and recall is maintained for a wide range of β_o, β_d.

We state the running time of our algorithm on a standard Intel 2.8 GHz dual core processor. The speed depends on a variety of factors such as object size and amount of deformation between frames. On HD 720 video, it is on average 5 seconds per frame for sequences in Fig. 11 (in C++), while AAE takes 1 second. Speed-ups are possible, e.g., the deformation computation can be sped up using a multiscale procedure.

9 Conclusion

The proposed technique for shape tracking is based on jointly matching shape and complex radiance of the object across frames. Self-occlusions and dis-occlusions, which pose a challenge to this approach, were modeled in this work.

To compute self-occlusions and the warp of a template to the next frame, a joint energy was formulated, and a novel optimization scheme was derived. The scheme has an automatic coarse-to-fine property, which is beneficial in tracking. The method was based on constructing a novel infinite dimensional Riemannian manifold of parameterized regions and a novel Sobolev-type metric. The optimization is a gradient descent with respect to a Sobolev metric. The coarse-to-fine property was demonstrated empirically.

Experiments demonstrated the criticality of modeling occlusions and dis-occlusions. Comparison to methods of partitioning image statistics and shape/appearance matching without occlusion modeling demonstrated the effectiveness of the proposed algorithm. Future work includes full occlusions of the object by other objects, and addressing limitations of the self-similarity assumption in dis-occlusion detection.

Acknowledgments

This work was funded by KAUST Baseline and Visual Computing Center funding.

References

[1] Y. Rathi, N. Vaswani, A. Tannenbaum, and A. Yezzi, “Tracking deforming objects using particle filtering for geometric active contours,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1470–1475, Aug. 2007.

[2] D. Cremers, “Dynamical statistical shape priors for level set-based tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1262–1273, Aug. 2006.

[3] C. Bibby and I. Reid, “Real-time tracking of multiple occluding objects using level sets,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 1307–1314.

[4] J. Fan, X. Shen, and Y. Wu, “Scribble tracker: A matting-based approach for robust tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1633–1644, Aug. 2012.

[5] M. Isard and A. Blake, “Condensation: Conditional density propagation for visual tracking,” *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.

[6] G. Sundaramoorthi, A. Mennucci, S. Soatto, and A. Yezzi, “A new geometric metric in the space of curves, and applications to tracking deforming objects by prediction and filtering,” *SIAM J. Imag. Sci.*, vol. 4, pp. 109–145, 2011.

[7] D. G. Ebin and J. Marsden, “Groups of diffeomorphisms and the motion of an incompressible fluid,” *Ann. Math.*, vol. 92, no. 1, pp. 102–163, 1970.

[8] A. Ayvaci, M. Raptis, and S. Soatto, “Sparse occlusion detection with optical flow,” *Int. J. Comput. Vis.*, vol. 97, pp. 1–17, 2011.

[9] J. Jackson, A. Yezzi, and S. Soatto, “Dynamic shape and appearance modeling via moving and deforming layers,” *Int. J. Comput. Vis.*, vol. 79, no. 1, pp. 71–84, 2008.

[10] Y. Yang and G. Sundaramoorthi, “Modeling self-occlusions in dynamic shape and appearance tracking,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 201–208.

[11] M. Klodt and D. Cremers, “A convex framework for image segmentation with moment constraints,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2236–2243.

[12] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.

[13] D. Mumford and J. Shah, “Optimal approximations by piecewise smooth functions and associated variational problems,” *Commun. Pure Appl. Math.*, vol. 42, no. 5, pp. 577–685, 1989.

[14] V. Caselles, R. Kimmel, and G. Sapiro, “Geodesic active contours,” *Int. J. Comput. Vis.*, vol. 22, no. 1, pp. 61–79, 1997.

[15] S. Kichenassamy, A. Kumar, P. Olver, A. Tannenbaum, and A. Yezzi, “Gradient flows and geometric active contour models,” in *Proc. IEEE 5th Int. Conf. Comput. Vis.*, 1995, pp. 810–815.

[16] N. Paragios and R. Deriche, “Geodesic active regions: A new framework to deal with frame partition problems in computer vision,” *J. Vis. Commun. Image Representation*, vol. 13, nos. 1/2, pp. 249–268, 2002.

[17] T. Chan and L. Vese, “Active contours without edges,” *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, Feb. 2001.

[18] T. F. Chan, S. Esedoglu, and M. Nikolova, “Algorithms for finding global minimizers of image segmentation and denoising models,” *SIAM J. Appl. Math.*, vol. 66, no. 5, pp. 1632–1648, 2006.

[19] T. Pock, A. Chambolle, D. Cremers, and H. Bischof, “A convex relaxation approach for computing minimal partitions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 810–817.

[20] X. Bai, J. Wang, D. Simons, and G. Sapiro, “Video SnapCut: Robust video object cutout using localized classifiers,” *ACM Trans. Graph.*, vol. 28, no. 3, p. 70, 2009.

[21] M. Niethammer, P. Vela, and A. Tannenbaum, "Geometric observers for dynamically evolving curves," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1093–1108, Jun. 2008.

[22] T. Cootes, C. Taylor, D. Cooper, J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[23] M. Black and A. Jepson, "EigenTracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Comput. Vis.*, vol. 26, no. 1, pp. 63–84, 1998.

[24] G. Hager and P. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 10, pp. 1025–1039, Oct. 1998.

[25] X. Bai, J. Wang, and G. Sapiro, "Dynamic color flow: A motion-adaptive color model for object segmentation in video," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 617–630.

[26] L. Alvarez, R. Deriche, T. Papadopoulo, and J. Sánchez, "Symmetrical dense optical flow estimation with occlusions detection," in *Proc. 7th Eur. Conf. Comput. Vis.*, 2002, pp. 721–735.

[27] R. Ben-Ari and N. Sochen, "Variational stereo vision with sharp discontinuities and occlusion handling," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–7.

[28] C. Strecha, R. Fransens, and L. Van Gool, "A probabilistic approach to large displacement optical flow and occlusion detection," in *Proc. Wrokshop Statist. Methods Video Process.*, 2004, pp. 71–82.

[29] J. Xiao, H. Cheng, H. Sawhney, C. Rao, and M. Isnardi, "Bilateral filtering-based optical flow estimation with occlusion detection," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 211–224.

[30] P. Sundberg, T. Brox, M. Maire, P. Arbeláez, and J. Malik, "Occlusion boundary detection and figure/ground assignment from optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 2233–2240.

[31] S. Ricco and C. Tomasi, "Dense lagrangian motion estimation with occlusions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 1800–1807.

[32] N. Paragios and R. Deriche, "Geodesic active contours and level sets for the detection and tracking of moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 3, pp. 266–280, Mar. 2000.

[33] G. Sundaramoorthi, A. Yezzi, and A. C. Mennucci, "Sobolev active contours," *Int. J. Comput. Vis.*, vol. 73, no. 3, pp. 345–366, 2007.

[34] G. Charpiat, P. Maurel, J.-P. Pons, R. Keriven, and O. Faugeras, "Generalized gradients: Priors on minimization flows," *Int. J. Comput. Vis.*, vol. 73, no. 3, pp. 325–344, 2007.

[35] G. Sundaramoorthi, A. Yezzi, and A. C. Mennucci, "Coarse-to-fine segmentation and tracking using Sobolev active contours," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 851–864, May 2008.

[36] E. Klassen, A. Srivastava, M. Mio, and S. Joshi, "Analysis of planar shapes using geodesic paths on shape spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 372–383, Mar. 2004.

[37] P. Michor, D. Mumford, J. Shah, and L. Younes, "A metric on shape space with explicit geodesics," *Rendiconti Lincei, Matematica e Applicazioni*, vol. 9, pp. 25–27, 2008.

[38] P. W. Michor and D. Mumford, "An overview of the Riemannian metrics on spaces of curves using the hamiltonian approach," *Appl. Comput. Harmonic Anal.*, vol. 23, no. 1, pp. 74–113, 2007.

[39] P. W. Michor and D. Mumford, "Riemannian geometries on spaces of plane curves," *J. Eur. Math. Soc.*, vol. 8, no. 1, pp. 1–48, 2003.

[40] A. Yezzi and A. Mennucci, "Conformal metrics and true," in *Proc. IEEE 10th Int. Conf. Comput. Vis.*, 2005, vol. 1, pp. 913–919.

[41] M. Beg, M. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 139–157, 2005.

[42] M. I. Miller, A. Trouvé, and L. Younes, "Geodesic shooting for computational anatomy," *J. Math. Imag. Vis.*, vol. 24, no. 2, pp. 209–228, 2006.

[43] X. Huang, N. Paragios, and D. Metaxas, "Shape registration in implicit spaces using information theory and free form deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1303–1318, Aug. 2006.

[44] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, "Deformable templates using large deformation kinematics," *IEEE Trans. Image Process.*, vol. 5, no. 10, pp. 1435–1447, Oct. 1996.

[45] A. Trouvé, "Diffeomorphisms groups and pattern matching in image analysis," *Int. J. Comput. Vis.*, vol. 28, no. 3, pp. 213–221, 1998.

[46] B. Wirth, L. Bar, M. Rumpf, and G. Sapiro, "A continuum mechanical approach to geodesics in shape space," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 293–318, 2011.

[47] M. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Comput. Vis. Image Understanding*, vol. 63, no. 1, pp. 75–104, 1996.

[48] B. Horn and B. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, 1981.

[49] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.

[50] L. C. Evans, "Partial differential equations. graduate studies in mathematics," *Amer. Math. Soc.*, vol. 2, 1998.

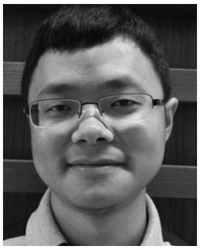
[51] A. Mennucci, A. Yezzi, and G. Sundaramoorthi, "Properties of Sobolev-type metrics in the space of curves," *Interfaces Free Boundaries*, vol. 10, no. 4, pp. 423–445, 2008.

[52] S. Osher and J. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations," *J. Comput. Phys.*, vol. 79, no. 1, pp. 12–49, 1988.

[53] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell.*, 1981, vol. 81, pp. 674–679.

[54] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, 2011.

[55] J. Sethian, "A fast marching level set method for monotonically advancing fronts," *Proc. Nat. Acad. Sci.*, vol. 93, no. 4, pp. 1591–1595, 1996.



Yanchao Yang received the BSc degree from the University of Science and Technology of China (USTC) in 2011 and the MSc degree from the King Abdullah University of Science and Technology (KAUST) in 2013, both in electrical engineering. He is currently working as a research assistant in the Visual Computing Center at KAUST. His research interests include object tracking from video and image matching.

Ganesh Sundaramoorthi received the PhD degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, USA. He was a postdoctoral researcher in computer science at the University of California, Los Angeles, between 2008 and 2010. In 2011, he was appointed an assistant professor of electrical engineering and an assistant professor of applied mathematics and computational science at the King Abdullah University of Science and Technology (KAUST). His research interests include computer vision with recent interest in shape and motion analysis, texture analysis, and invariant representations for visual tasks.

Appendix A
Computing Region-Based Sobolev Gradients

We now show how to compute the gradient of an energy with respect to the Sobolev inner product defined in (10). For generality, we compute the gradient of

$$E(w) = \int_R f(w(x), x)\,\mathrm{d}x \tag{43}$$

where $f : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$. The directional derivative in the direction $h : R \to \mathbb{R}^2$ is

$$\mathrm{d}E(w) \cdot h = \int_R f_1(w(x), x) \cdot h(x)\,\mathrm{d}x = \\ \int_{w(R)} f_1(x, w^{-1}(x)) \cdot h \circ w^{-1}(x) \det\left(\nabla w^{-1}(x)\right) \mathrm{d}x \tag{44}$$

and since by definition $\mathrm{d}E(w) \cdot h = \langle G, h \rangle_w$ for all $h \in T_w M$, where $G = \nabla_w E$ is the gradient with respect to the Sobolev inner product, we have that

$$\int_{w(R)} f_1(x, w^{-1}(x)) \cdot \hat{h}(x) \det\left(\nabla w^{-1}(x)\right) \mathrm{d}x = \\ \overline{G} \cdot \overline{\hat{h}} + \alpha \int_{w(R)} \mathrm{tr}\left\{ \nabla G(x)^T \nabla \hat{h}(x) \right\} \mathrm{d}x \tag{45}$$

By integrating by parts, one finds that

$$\int_{w(R)} f_1(x, w^{-1}(x)) \cdot \hat{h}(x) \det\left(\nabla w^{-1}(x)\right) \mathrm{d}x = \\ \alpha \int_{\partial w(R)} (\nabla G(x) \cdot N) \cdot \hat{h}(x)\,\mathrm{d}x - \\ \int_{w(R)} \left(\frac{1}{|w(R)|} \overline{G} - \alpha \Delta G(x) \right) \cdot \hat{h}(x)\,\mathrm{d}x. \tag{46}$$

Therefore, G can be obtained by solving

$$\begin{cases} \frac{\overline{G}}{|w(R)|} - \alpha \Delta G(x) = \tilde{F}(x) & x \in w(R) \\ \tilde{F}(x) := f_1(x, w^{-1}(x)) \det \nabla w^{-1}(x) & x \in w(R) \\ \nabla G(x) \cdot N = 0 & x \in \partial w(R) \end{cases}. \tag{47}$$

Integrating both sides of the first equation above over R, we find that

$$\overline{G} = \int_R f_1(x, w^{-1}(x)) \det\left(\nabla w^{-1}(x)\right) \mathrm{d}x. \tag{48}$$

Therefore, the solution for G is expressed as

$$G = \overline{G} + \frac{1}{\alpha} \tilde{G} \tag{49}$$

where $\tilde{G}$ (independent of α) satisfies

$$\begin{cases} -\Delta \tilde{G}(x) = F(x) & x \in R \\ \nabla \tilde{G}(x) \cdot N = 0 & x \in \partial w(R) \\ \overline{\tilde{G}} = 0 \end{cases}, \tag{50}$$

and

$$F(x) = f_1(x, w^{-1}(x)) \det\left(\nabla w^{-1}(x)\right) - \\ \overline{f_1(\cdot, w^{-1}(\cdot)) \det\left(\nabla w^{-1}(x)\right)}. \tag{51}$$

We consider f of the form

$$f(y, z) = \frac{1}{2} \rho(|I(y) - a(z)|^2) \bar{\chi}_O(z) \tag{52}$$

where $\rho : \mathbb{R} \to \mathbb{R}^+$. This gives

$$f_1(y, z) = \rho'(|I(y) - a(z)|^2)(I(y) - a(z)) \nabla I(y) \bar{\chi}_O(z). \tag{53}$$

Appendix B
Numerical Discretization

A. *Sobolev Gradient Discretization*

We show how to discretize (50), the Poisson equation. The discretization of the Laplacian is

$$-\Delta \tilde{G}(x) = -\sum_{y \sim x} \tilde{G}(y) - \tilde{G}(x) = F(x), \tag{54}$$

where F is defined in (51), and $y \sim x$ indicates that y is a 4-neighbor of x. Discretizing the boundary condition $\nabla \tilde{G}(x) \cdot N = \tilde{G}(y) - \tilde{G}(x) = 0$, when $y \sim x$, $y \notin R$, and substituting it above, we have that

$$-\sum_{y \sim x, y \in R} \tilde{G}(y) - \tilde{G}(x) = F(x). \tag{55}$$

This can be solved using the conjugate gradient method. Indeed, the operator on the left is positive definite on the set of mean zero vector fields. One starts with an initialization such that $\overline{\tilde{G}} = 0$.

B. *Discretization of Transport Equations*

We describe the discretizations of the transport equations used in the gradient descent of the warp ϕ_τ^{-1} and the warped region R_τ, which for the most part, are standard.

Let $\Psi_\tau : \Omega \to \mathbb{R}$ denote the level set function at time τ such that $\{x \in \Omega : \Psi_\tau(x) < 0\} = R_\tau$. The level set evolution equation (34) (shown here again for convenience):

$$\partial_\tau \Psi_\tau(x) = \nabla G_\tau(x) \cdot \nabla \Psi_\tau(x) \tag{56}$$

is discretized using an up-winding difference scheme:

$$\Psi_{\tau_{i+1}}(x) = \Psi_{\tau_i}(x) + \\ \Delta t \left(G^1_{\tau_i}(x) D_{x_1}[\Psi_{\tau_i}, G^1_{\tau_i}, x] + G^2_{\tau_i}(x) D_{x_2}[\Psi_{\tau_i}, G^2_{\tau_i}, x] \right) \tag{57}$$

where $\Delta t > 0$ is the time step,

$$D_{x_j}[\Psi_{\tau_i}, G^j_{\tau_i}, x] = \begin{cases} D^+_{x_j} \Psi_{\tau_i}(x) & \text{if } G^j_{\tau_i}(x) < 0 \\ D^-_{x_j} \Psi_{\tau_i}(x) & \text{if } G^j_{\tau_i}(x) \geq 0 \end{cases} \tag{58}$$

where $D^+_{x_j}$ ($D^-_{x_j}$) denotes the forward (backward, resp.) difference with respect to the j^{th} coordinate, and $G_\tau(x) = (G^1_\tau(x), G^2_\tau(x))$. Note that $G_\tau | \partial R_\tau$ is extended to the narrowband of the level set function by choosing G_τ at a point x in the narrowband to be the same as that of the closest point on ∂R_τ from x.

The discretization of the transport equation (33) for the backward map:

$$\partial_\tau \phi_\tau^{-1}(x) = \nabla G_\tau(x) \cdot \nabla \phi_\tau^{-1}(x) \tag{59}$$

is

$$\phi^{-1}_{\tau_{i+1}}(x) =$$
$$\begin{cases} \phi^{-1}_{\tau_i}(x) + \Delta t \left(G^1_{\tau_i}(x) D_{x_1}[\phi^{-1}_{\tau_i}, G^1_{\tau_i}, x] \right. \\ \quad \left. + G^2_{\tau_i}(x) D_{x_2}[\phi^{-1}_{\tau_i}, G^2_{\tau_i}, x] \right), & x \in R_{\tau_{i+1}} \cap R_{\tau_i} \\ \frac{\sum_{y \in N_x \cap R_{\tau_i}} d_{\Psi_{\tau_i}}(x,y) \phi^{-1}_{\tau_i}(y)}{\sum_{y \in N_x \cap R_{\tau_i}} d_{\Psi_{\tau_i}}(x,y)}, & x \in R_{\tau_{i+1}} \backslash R_{\tau_i} \end{cases} \tag{60}$$

where N_x denotes the eight neighbors of x, and $d_{\Psi_{\tau_i}}(x, y)$ denotes the distance between x and the zero crossing of the level set Ψ_{τ_i} between x and y (zero if there is no zero crossing). In the computation of the forward/backward difference, if the relevant neighbor of x is not in R_{τ_i}, then the difference is set to zero. It should be noted that the step size is chosen to satisfy the stability criteria, which means that the level set may not move more than one pixel and thus x will always have a neighbor that is in R_{τ_i}, and so the second case in (60) is well-defined. The step size Δt is chosen to satisfy $\Delta t < 0.5 / \max_{x \in R_{\tau_i}, j=1,2} |G^j_{\tau_i}(x)|$.