

Modeling Self-Occlusions in Dynamic Shape and Appearance Tracking

Yanchao Yang and Ganesh Sundaramoorthi
King Abdullah University of Science and Technology (KAUST)
{yanchao.yang, ganesh.sundaramoorthi}@kaust.edu.sa

Abstract

We present a method to track the precise shape of a dynamic object in video. Joint dynamic shape and appearance models, in which a template of the object is propagated to match the object shape and radiance in the next frame, are advantageous over methods employing global image statistics in cases of complex object radiance and cluttered background. In cases of complex 3D object motion and relative viewpoint change, self-occlusions and dis-occlusions of the object are prominent, and current methods employing joint shape and appearance models are unable to accurately 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. Experiments on video exhibiting occlusion/dis-occlusion, complex radiance and background show that occlusion/dis-occlusion modeling leads to superior shape accuracy compared to recent methods employing joint shape/appearance models or employing global statistics.

1. Introduction

In many video processing applications, such as post-production of motion pictures, it is important to obtain the *shape* (silhouette) of the object of interest at each frame in a video. Although many methods have been proposed, much work remains. Many existing tracking methods (e.g., [21, 11, 7, 12]) are built on top of partitioning the image into foreground and background based on global image statistics (e.g., color distributions, edges, texture, motion), which is advantageous in obtaining shape of the object. However, in tracking objects with complex radiance and cluttered background, partitioning the image based on global statistics may not yield the object as a partition. An alternative approach is to deform a template (the radiance function defined on the region of the projected object) to match the object in shape and radiance in the next frame (the deformed shape yields the object of interest). We will refer to this alternative approach as *joint shape/appearance matching*.

A difficulty in tracking by joint shape/appearance matching is that 3D object and camera motion imply that parts of the object come into view (*dis-occlusions*) and go out of view (*occlusions*); therefore, an initially accurate template, even when warped through a non-rigid deformation, becomes an inaccurate model of the object in later frames. Thus, it is necessary to update the template by removing occluded regions and including dis-occluded regions.

In this work, we model self-occlusions and dis-occlusions in *tracking by joint shape/appearance matching*. Small frame rate implies moderately large non-rigid deformation of the projected object between frames. Thus, we represent the large non-rigid deformation as an integration of a time-varying vector field (see e.g., [5]) defined on *evolving region* (or *domain of interest*). Since an occlusion is the part of the template that does not correspond to the next frame, occlusions and the deformation are coupled, and thus, a joint optimization problem in the large deformation and occlusion is setup, and a simple, efficient algorithm is derived. We note that dis-occlusions can be detected only with priors on the object. We show how to use a prior that the object radiance is self-similar, so that dis-occluded regions between frames can be detected by measuring image similarity to the current template. To ensure robust estimates of the object's radiance across frames, recursive filtering is used.

Contributions: Our 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 either in a framework with simpler models of radiance (e.g., [7]), i.e., color histograms, or are layered models with complex radiance (e.g., [16]) that can cope with occlusions of one layer on another, but not *self-occlusions* or *dis-occlusions*. Second, we solve dis-occlusions with the similarity prior mentioned above.

1.1. Related Work

Most shape tracking techniques (e.g., [15, 21, 11, 7]) extend image segmentation techniques such as active contours (e.g., [17, 9, 20, 10]). These techniques build on discrim-

inating the foreground and background using global image statistics (e.g., color distributions, texture, edges, motion). However, when the object has complex radiance and is within cluttered background, discriminating global image statistics leads to errors in the segmentation. Some methods try to resolve this issue by using local statistics (e.g., [18]). Other methods use temporal consistency to predict the object location / shape in the next frame (e.g., [15, 21, 25]) to provide better initialization to frame partitioning. In [11], dynamics of shape are modeled from training data, constraining the solution of frame partitioning; however, training data is only available in restricted scenarios. While providing improvements, images with complex object radiance and cluttered background still pose a significant challenge.

We approach shape tracking by joint shape/appearance matching. We use a radiance model that is a dense function defined on the projected object. Dense radiance functions have been used (e.g., [8, 13]) for tracking via matching to the next frame. However, they are box trackers, and do not provide shape. In [16, 3], a joint model of radiance and shape of the object *and* background is used, however, *self-occlusions* and *dis-occlusions* are not modeled.

Occlusions have been considered in optical flow. In [1, 6], forward and backward optical flows are computed, and the occluded region is the set where the composition of these flows is not the identity. In [24, 28], an occlusion is the set where the optical flow residual is large. In [26], occlusion boundaries are detected by discontinuities of optical flow. In [2], joint estimation of the optical flow and occlusions is performed. In [22], dense trajectory estimation across multiple frames with occlusions is solved. We use ideas of occlusions in [2], and apply them to shape tracking where additional considerations must be made for evolving the shape, dis-occlusions, and larger deformations.

2. Dynamic Model of the Projected Object

In this section, we give our dynamic model of the shape and radiance of the 3D object projected in the imaging plane. From this, the notion of occlusions and dis-occlusions is clear. The dynamic model is necessary for the recursive estimation algorithm in Section 5.

Let $\Omega \subset \mathbb{R}^2$, and $I : \{1, 2, \dots, N\} \times \Omega \rightarrow \mathbb{R}^k$ 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 $a_t : R_t \rightarrow \mathbb{R}^k$. Our dynamic model of the region and radiance (see Fig. 1 for a diagram) is

$$R_{t+1} = w_t(R_t \setminus O_t) \cup D_{t+1} \quad (1)$$

$$a_{t+1}(x) = \begin{cases} a_t(w_t^{-1}(x)) + \eta_t(x) & x \in w_t(R_t \setminus O_t) \\ a_{t+1}^d(x) + \eta_t(x) & x \in D_{t+1} \end{cases} \quad (2)$$

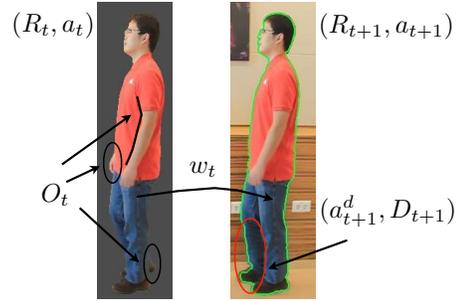


Figure 1: **Diagram illustrating our dynamic model.** Left: template (R_t, a_t) (non-gray), right: I_{t+1} . Self-occlusions O_t , dis-occlusions D_{t+1} and its radiance a_{t+1}^d , the region at frame $t + 1$ is R_{t+1} (inside the green contour), and the warp is w_t , which is defined in $R_t \setminus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

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 disoccluded (comes into view) at frame $t + 1$, $a_{t+1}^d : D_{t+1} \rightarrow \mathbb{R}^k$ is the radiance of the disoccluded 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 a diffeomorphism on the *un-occluded* region $R_t \setminus O_t$ (it will be extended to all of R_t : see Section 3 for details), which is a transformation arising from viewpoint change and 3D deformation.

The region $R_t \setminus O_t$, is warped by w_t and the dis-occlusion of the projected object, D_{t+1} , is appended to the warped region to form R_{t+1} . The relevant portion of the radiance, $a_t|_{(R_t \setminus O_t)}$ is transferred via the warp w_t to R_{t+1} (as usual brightness constancy), noise added, and then 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, etc...).

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 3, we derive the method for determining w_t , the occlusion O_t , and $w_t(R_t \setminus O_t)$ (the warping of the unoccluded region). In Section 4, we derive a method, given $w_t(R_t \setminus O_t)$ and I_{t+1} , to estimate the dis-occlusion of the object, D_{t+1} . In Section 5, we derive a recursive estimation procedure and integrate all steps. See Fig. 2 for a system overview.

3. Occlusions and Deformation Computation

In this section, we model the warp w_t as an integration of a time-varying vector field (see e.g., [5]) to obtain large deformations and (with sufficient regularity) a diffeomorphic registration. While this representation of a warp is standard, there are important differences in this work: 1) the vector field is defined on an *evolving region* and the target region in the next frame is unknown, and 2) part of the region is

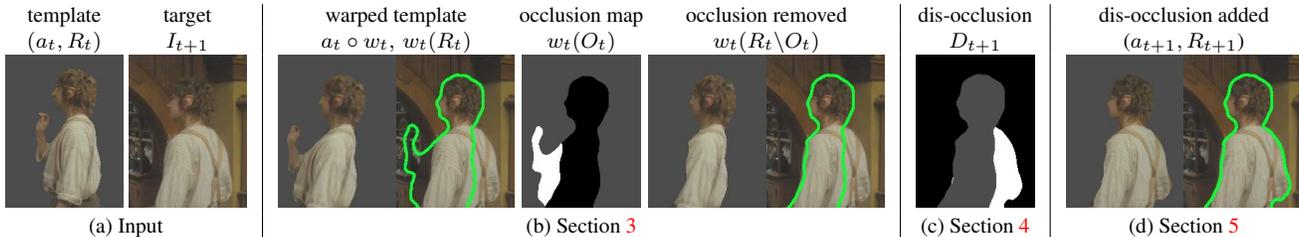


Figure 2: **Illustration of frame processing in our algorithm.** (a): Estimate at frame t of the shape and radiance (a_t, R_t) , and the next image I_{t+1} . (b): Simultaneous non-rigid warping and occlusion estimation is performed (first image: warped template $a_t \circ w_t$, second: boundary of warped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

occluded, and the occlusion must be determined.

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. As suggested in [2] for optical flow, occlusion detection and registration should be computed jointly.

3.1. Energy Formulation

We avoid subscripts t for ease of notation in the rest of this section. We formulate the problem of given a region $R \subset \Omega$, the radiance $a : R \rightarrow \mathbb{R}^k$, and $I : \Omega \rightarrow \mathbb{R}^k$ to compute the occluded part O of R , the warp w defined on $R \setminus O$, and $w(R \setminus O)$ such that $I(x) = a(w^{-1}(x)) + \eta(x)$ for $x \in w(R \setminus O)$ (where η is noise modeled in (2)).

The warp w is a diffeomorphism in the unoccluded region $R \setminus O$. For ease in the optimization (see [14]), we consider w to be a diffeomorphism on all of R ; the warp of interest will be the restriction to $R \setminus O$. The map w is the integration of a smooth time varying velocity field:

$$w(x) = \phi_T(x), \quad \phi_\tau(x) = x + \int_0^\tau v_s(\phi_s(x)) ds, \quad (3)$$

where $x \in R$, $T > 0$, $v_\tau : R_\tau \rightarrow \mathbb{R}^2$ is a velocity field (defined on $R_\tau = \{\phi_\tau(x) : x \in R\}$), and ϕ_τ is defined on R for every $\tau \in [0, T]$. The map ϕ_τ is such that $\phi_\tau(x)$ indicates the mapping of x after it flows along the velocity field for time τ , which is an artificial time parameter.

We formulate the energy (to be optimized in O, w):

$$E_o(O, w; I, a, R) = \int_{R \setminus O} |I(w(x)) - a(x)|^2 dx + \alpha \int_0^1 \int_{R_\tau} |\nabla v_\tau(x)|^2 dx d\tau + \beta_o \text{Area}(O). \quad (4)$$

Regularization of w is needed due to the aperture ambiguity, and velocity v_τ regularization ensures smoothness of w . 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. Note that although w is defined on all of R , a needs only to warp to I in the unoccluded region as the data term excludes O .

3.2. Approximate Optimization of E_o

While the lofty goal is to minimize the energy E_o (4) subject to (3) via a gradient descent, in the interest of computational speed and simplicity, we use a greedy algorithm to obtain a sub-optimal solution rather than computing the full Euler-Lagrange equations. The idea is: starting at $\tau = 0$, solve for the incremental velocity v_τ jointly with an estimate of the occlusion (then linearization of the energy is valid), deform the region R_τ by the velocity v_τ , deform the radiance a by the accumulation of velocity ϕ_τ^{-1} , and repeat the procedure (until convergence of the region R_τ) with the deformed radiance $a_\tau = a \circ \phi_\tau^{-1}$ and deformed region R_τ .

The optimization scheme is the following equations:

$$a_0 = a, \quad \Psi_0(x) = d_R(x), \quad \phi_0^{-1}(x) = x \quad (5)$$

$$R_\tau = \{\Psi_\tau \leq 0\}, \quad (6)$$

$$v_\tau, O_\tau = \arg \min_{v, O} \tilde{E}_o(v, O; I, a_\tau, R_\tau) \quad (7)$$

$$\partial_\tau \Psi_\tau(x) = -\nabla \Psi_\tau(x) \cdot v_\tau(x), \quad x \in B_2(R_\tau), \quad (8)$$

$$\partial_\tau \phi_\tau^{-1}(x) = -\nabla \phi_\tau^{-1}(x) \cdot v_\tau(x), \quad x \in R_\tau, \quad (9)$$

$$a_\tau(x) = a \circ \phi_\tau^{-1}(x), \quad x \in R_\tau, \quad (10)$$

where ∂_τ denotes partial with respect to τ , and $B_2(R_\tau) = \{x \in \Omega : |d_{R_\tau}(x)| \leq 2\}$ where d_{R_τ} is the signed distance function of R_τ . The function $\Psi_\tau : \Omega \rightarrow \mathbb{R}$ is a level set function [19] for the region R_τ , and the evolution of Ψ_τ is given by the transport equation (8), i.e., the region R_τ is updated in direction of the velocity $v_\tau : R_\tau \rightarrow \mathbb{R}^2$. Note v_τ

is extended to $B_2(R_\tau)$ as in narrowband level set methods. The backward warp $\phi_\tau^{-1} : R_\tau \rightarrow R$ is computed by flowing the identity map along the velocity field v_τ up to time τ , and this can be accomplished by the transport equation (9). The radiance in the warped region, $a_\tau : R_\tau \rightarrow \mathbb{R}^k$, is computed at a point by using the value of the original radiance at the back-warping of the point (10).

The energy in (7) is a linearized version of E_o :

$$\begin{aligned} \tilde{E}_o(v, O; I, a_\tau, R_\tau) &= \alpha \int_{R_\tau} |\nabla v(x)|^2 dx + \beta_o \text{Area}(O) \\ &+ \int_{R_\tau \setminus O} |I(x) - a_\tau(x) + \nabla a_\tau(x) \cdot v(x)|^2 dx. \end{aligned} \quad (11)$$

The energy must be optimized jointly in v and O . The global optimum in v can be obtained given O , and vice-versa. Thus, we use an alternating optimization scheme. Given O , the global optimal for v is determined from

$$-\alpha \Delta v(x) = \begin{cases} F(x) \nabla a_\tau(x) & x \in R_\tau \setminus O \\ 0 & x \in O \end{cases} \quad (12)$$

$$F(x) = I(x) - a_\tau(x) + \nabla a_\tau(x) \cdot v(x) \quad (13)$$

with Neumann boundary conditions on ∂R_τ . The above equation is solved efficiently with a conjugate gradient solver. Given v , the solution for O is

$$O = \{x \in R_\tau : (G_\sigma * F^2)(x) > \beta_o\}, \quad (14)$$

where G_σ is a Gaussian smoothing filter. The global optimum for O is when $\sigma = 0$, but smoothing is applied to ensure a spatially regular O . To optimize \tilde{E}_o , O is initially chosen to be the empty set, then (12) is solved, then the occlusion is updated using (14), and the process is iterated until convergence (i.e., the set O does not change).

Due to large displacement of the object between frames, we perform the optimization in a coarse-to-fine manner: at initial time τ the regularity parameter α is chosen large until convergence of the energy \tilde{E}_o , then α is lowered (to capture finer details) at larger time τ , and the process is iterated until convergence of \tilde{E}_o . Fig. 3 shows the evolution procedure (5)-(10) illustrated on an example.

Let $\tau = T$ be the time of convergence, $R_{\tau=T}$ - a warping of R includes a warping of the occluded region $O_{\tau=T}$, and thus the warping of the un-occluded region is $w(R \setminus O) = R'_T = R_{\tau=T} \setminus O_{\tau=T}$, and does not include the disoccluded region, which is computed in the next section from R'_T .

For more details and discretization, see [27].

4. Dis-Occlusion Computation

In this section, we describe the computation of the dis-occlusion $D_{t+1} \subset \Omega$ of the object at frame $t + 1$ given the warped unoccluded part of the region $w_t(R_t \setminus O_t)$ determined from the previous section, and the image I_{t+1} . To



Figure 3: **Occlusion estimation and warping.** [Top to bottom]: Beginning ($\tau = 0$), intermediate, and final stages of evolution. [1st column]: radiance a_τ , [2nd]: target image I and boundary of R_τ , [3rd]: velocity v_τ , [4th]: occlusion estimation F^2 at time τ , [5th]: optical flow color code. The final occluded region is shown in Fig. 2(b).

determine the disoccluded region of the object (the region of the projected object that comes into view in the next frame that is not seen in the current template), 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 disoccluded region of the object is similar to parts of the image I_{t+1} in $w_t(R_t \setminus O_t)$, and for computational 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.

4.1. Energy Formulation of Dis-Occlusions

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 $R' = w_t(R_t \setminus O_t)$. The energy is

$$E_d(D) = - \int_D p(x) dx + \beta_d \text{Area}(D) \quad (15)$$

where $D \subset \Omega \setminus R'$, $p(x) \geq 0$ denotes the likelihood that $x \in \Omega \setminus R'$ belongs to the dis-occluded region, and $\beta_d > 0$

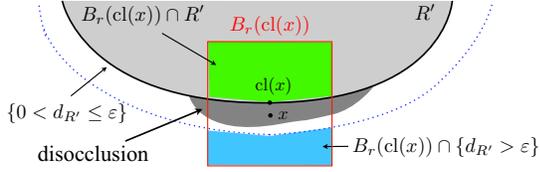


Figure 4: **Diagram of quantities used in the likelihood $p(x)$ of a disoccluded pixel.** The dark gray region is the dis-occlusion to be determined. Light gray region is R' , region before the dis-occlusion is determined. A pixel x within the band $\{0 < d_{R'} \leq \varepsilon\}$ is depicted, and its closest point to R' , $\text{cl}(x)$. The green (blue) region is where the foreground (background) distribution $p_{\text{cl},f}(x)$ ($p_{\text{cl},b}(x)$) is determined.

is a weight. The dis-occluded region, assuming a moderate camera frame rate, is small in area compared to the projected object, hence the penalty on area.

Let $\text{cl}(x)$ 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. 4): one that measures the fit of $I(x)$ to the local distribution of I within $B_r(\text{cl}(x)) \cap R'$ versus the background $B_r(\text{cl}(x)) \cap \{d_{R'} > \varepsilon\}$ in I , and the second that measures nearness of x to R' . One choice of p is

$$p(x) \propto \exp \left[-\frac{d_{R'}(x)^2}{2\sigma_d^2} + p_{\text{cl}(x),f}(I(x)) - p_{\text{cl}(x),b}(I(x)) \right] \quad (16)$$

where $d_{R'}(x)$ indicates the Euclidean distance from x to R' , $p_{\text{cl}(x),f}$, $p_{\text{cl}(x),b}$ are Parzen estimates of the intensity distribution of I in $B_r(\text{cl}(x)) \cap R'$ (resp. $B_r(\text{cl}(x)) \cap \{d_{R'} > \varepsilon\}$) where ε is chosen large enough so that the region includes some background beyond the dis-occlusion.

4.2. Optimization of E_d

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

$$D = \{x : d_{R'}(x) \in (0, \varepsilon], (G_\sigma * p)(x) > \beta_d\} \quad (17)$$

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. 5 shows an example of p , the dis-occlusion detected, and the final estimate of the region.

Computation of $d_{R'}$ in $\{0 < d_{R'} < \varepsilon\}$ is done efficiently with the Fast Marching Method [23], and $\text{cl}(x)$ at each point is simultaneously propagated as the front in the Fast Marching Method evolves. Then p is readily computed.

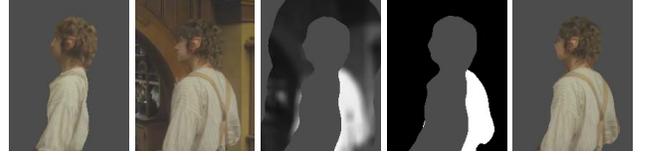


Figure 5: **Illustration of disocclusion detection.** [1st]: warped unoccluded radiance defined on R' (after occlusion and deformation computation), [2nd]: target image I , [3rd]: likelihood of dis-occlusion map p (defined in $B_{R'}(\varepsilon)$), [4th]: computed dis-occlusion D (white), and [5th]: final radiance. Boundary of final region super-imposed on I is in Fig. 2 (d).

5. 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 $R_0 \subset \Omega$, $a_0 : R_0 \rightarrow \mathbb{R}^k$, 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 3.2 with input R_{t-1} , a_{t-1} and I_t .
2. Given $R'_t = w_{t-1}(R_{t-1}) \setminus w_{t-1}(O_{t-1})$, the warping of the unoccluded part of R_{t-1} , and the image I_t , compute the dis-occlusion D_t using (17). The estimate of R_t is then $R'_t \cup D_t$.
3. The radiance is then updated as

$$a_t(x) = \begin{cases} (1 - K_a)a'_t(x) + K_a I_t(x) & x \in R'_t \\ I_t(x) & x \in D_t \end{cases} \quad (18)$$

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

The averaging of the warped radiance and the current image (18) 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.

6. Experiments and Comparisons

We demonstrate our method on a variety of videos that contain self-occlusions/disocclusions. All examples shown have over 100 frames (all videos are in Supplementary Material). To demonstrate that occlusion/dis-occlusion modeling aids joint shape/appearance tracking, we compare to Adobe After Effects CS6 2012 (AAE) (based on [4]), 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 using global statistics, we compare to [12] (publicly available code), which employs global statistics in addition to other advanced techniques.

Parameters are chosen as: $\sigma = 5$ in (17) and (14), $\sigma_d = 100$ in the likelihood, p in (16), 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 $\beta_d = 0.5$ when p is normalized to be a probability. The gain in the radiance update (18) is $K_a = 0.8$. Most parameters can be fixed for the whole video, and work on a wide range. Most significant parameters are the β 's, and sensitivity analysis is shown in [27].

The first experiment (Fig. 6) 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/disoccluded. Ignoring occlusions (setting $O = \emptyset$ in Section 3.2) 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), disoccluded parts of the body are detected. However, irrelevant regions of the background (that can be removed in the occlusion stage) are captured. Best results (last row) are achieved when both the occlusion and dis-occlusions are modeled. The fourth row shows the result of [12], which has trouble discriminating between face and the background, which share similar radiance. The fifth row shows the result of Adobe After Effects 2012 (AAE), which captures irrelevant background.

Fig. 7 shows tracking of a fish and a skater. When foreground/ background global histograms are easily separable, [12] performs well, and when occlusions are minor AAE, performs well as does the proposed method.

In Fig. 8, 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, the methods [12] and Adobe After Effects 2012 (AAE) have trouble discriminating be-

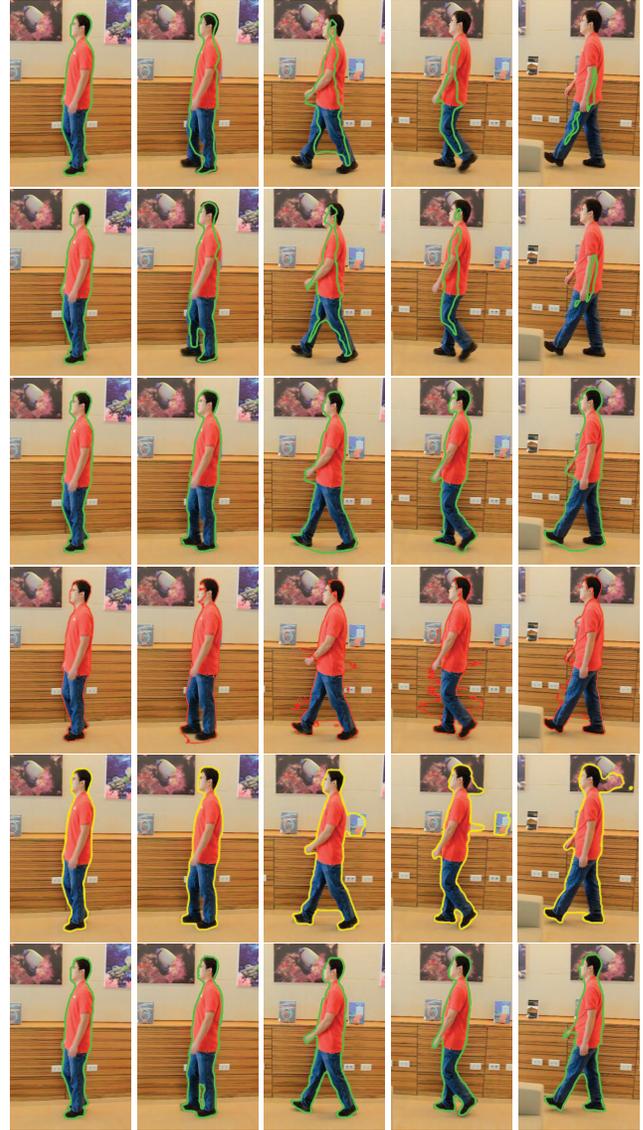


Figure 6: **Modeling Occlusions/Dis-Occlusions is Necessary.** [1st row]: occlusion/dis-occlusion detection are turned off in our method. [2nd]: occlusion modeling done, but not dis-occlusions in our method. [3rd]: dis-occlusions detected but not occlusions. [4th]: result of [12]. [5th]: result of AAE [4]. [6th]: accurate tracking when both occlusion and dis-occlusion modeling is performed (our final result).

tween 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, [12] 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.

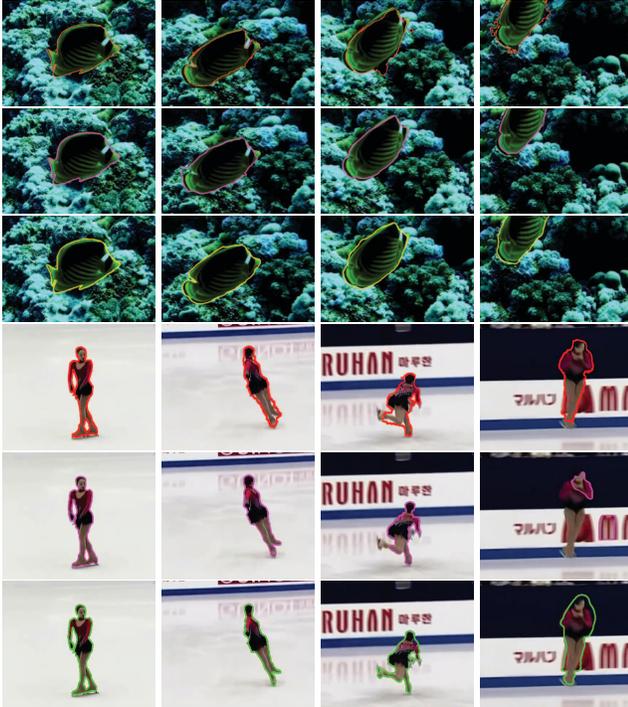


Figure 7: **Distinctive foreground/background global statistics.** [Top]: [12], [Middle]: AAE, [Bottom]: proposed method. When fore/background global statistics are separable, [12], and AAE, for minor occlusions, performs well.

Sequence	Scribbles [12]	Adobe Effects 2012 [4]	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

Figure 9: **Quantitative performance analysis.** Average F-measure (over all frames) computed from ground truth are shown. Larger F-measure means better performance.

Our method is able to capture the shape of the objects quite well (quantitative assessment is in Fig. 9). 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 do not share similar radiance as the current template (sole of shoe) are not detected.

Lastly, we state the running time of our algorithm on a standard Intel 2.8GHz dual core processor. Note that the speed will depend on a variety of factors such as the size of the object and amount of deformation between frames. On HD 720 video, it is on average 8 seconds per frame for sequences in Fig. 8 (in C++), while AAE takes 1 second. Speed-ups are possible, e.g., the joint velocity and occlusion

computation can be sped up using a multi-scale procedure.

7. Conclusion

The proposed technique for shape tracking is based on jointly matching shape and complex radiance (defined as a function on the region) of the object across frames. Self-occlusions and dis-occlusions pose a challenge for joint shape/appearance tracking, which were modeled and computed in a principled framework in this work. Experiments demonstrated the criticality of modeling occlusions and dis-occlusions. Comparison to recent methods built on global image statistics foreground/background separation and joint shape/appearance modeling without occlusion modeling demonstrated the effectiveness of the proposed algorithm in situations of complex object/background radiance, and self-occlusions/dis-occlusions.

Future work includes full occlusions of the object by other objects, and improving dis-occlusion detection.

References

- [1] L. Alvarez, R. Deriche, T. Papadopoulos, and J. Sánchez. Symmetrical dense optical flow estimation with occlusions detection. *ECCV 2002*, pages 721–735, 2002. 2
- [2] A. Ayvaci, M. Raptis, and S. Soatto. Sparse occlusion detection with optical flow. *International Journal of Computer Vision*, pages 1–17, 2011. 2, 3
- [3] X. Bai, J. Wang, and G. Sapiro. Dynamic color flow: a motion-adaptive color model for object segmentation in video. *ECCV 2010*, pages 617–630, 2010. 2
- [4] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapshot: robust video object cutout using localized classifiers. *ACM Transactions on Graphics (TOG)*, 28(3):70, 2009. 5, 6, 7
- [5] M. Beg, M. Miller, A. Trounev, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, 2005. 1, 2
- [6] R. Ben-Ari and N. Sochen. Variational stereo vision with sharp discontinuities and occlusion handling. In *ICCV*, pages 1–7. IEEE, 2007. 2
- [7] C. Bibby and I. Reid. Real-time tracking of multiple occluding objects using level sets. In *CVPR*, pages 1307–1314. IEEE, 2010. 1
- [8] M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998. 2
- [9] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *IJCV*, 22(1):61–79, 1997. 1
- [10] T. Chan and L. Vese. Active contours without edges. *Image Processing, IEEE Transactions on*, 10(2):266–277, 2001. 1
- [11] D. Cremers. Dynamical statistical shape priors for level set-based tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(8):1262–1273, 2006. 1, 2
- [12] J. Fan, X. Shen, and Y. Wu. Scribble tracker: a matting-based approach for robust tracking. *IEEE Transactions on Pattern*

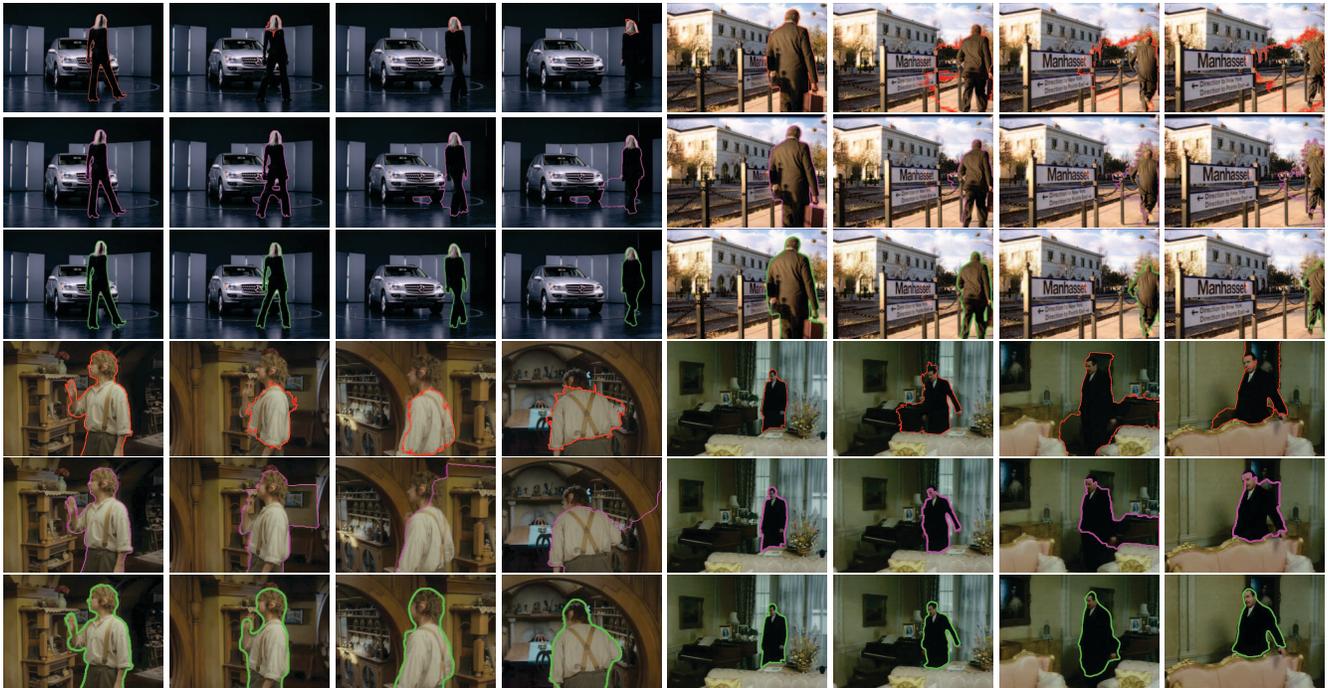


Figure 8: **Occlusions/dis-occlusions, violations of brightness constancy, and foreground/background not easily separable.** [Top]: [12], [Middle]: Adobe After Effects 2012, [Bottom]: proposed method. Methods based on foreground/background image statistic discrimination leak into the background. Note 4 (out of about 100) frames are selected for display in each sequence (see Supplementary Material for video).

Analysis and Machine Intelligence, 34(8):1633–1644, August 2012. 1, 6, 7, 8

[13] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(10):1025–1039, 1998. 2

[14] B.-W. Hong, Z. Lu, and G. Sundaramoorthi. A new model and simple algorithms for multi-label mumford-shah problems. In *CVPR*, pages 1219–1226. IEEE, 2013. 3

[15] M. Isard and A. Blake. Condensation: conditional density propagation for visual tracking. *International journal of computer vision*, 29(1):5–28, 1998. 1, 2

[16] J. Jackson, A. Yezzi, and S. Soatto. Dynamic shape and appearance modeling via moving and deforming layers. *Int. Journal of Computer Vision*, 79(1):71–84, 2008. 1, 2

[17] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *IJCV*, 1(4):321–331, 1988. 1

[18] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989. 2

[19] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *J. Comp. Physics*, 79(1):12–49, 1988. 3

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

[21] Y. Rathi, N. Vaswani, A. Tannenbaum, and A. Yezzi. Tracking deforming objects using particle filtering for geometric active contours. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(8):1470–1475, 2007. 1, 2

[22] S. Ricco and C. Tomasi. Dense lagrangian motion estimation with occlusions. In *CVPR*, pages 1800–1807. IEEE, 2012. 2

[23] J. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591, 1996. 5

[24] C. Strecha, R. Fransens, and L. Van Gool. A probabilistic approach to large displacement optical flow and occlusion detection. *Statistical methods in video processing*, pages 25–45, 2004. 2

[25] 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 Journal on Imaging Sciences*, 2011. 2

[26] P. Sundberg, T. Brox, M. Maire, P. Arbeláez, and J. Malik. Occlusion boundary detection and figure/ground assignment from optical flow. In *CVPR*, pages 2233–2240, 2011. 2

[27] <http://arxiv.org/abs/1208.4391>. 4, 6

[28] J. Xiao, H. Cheng, H. Sawhney, C. Rao, and M. Isnardi. Bilateral filtering-based optical flow estimation with occlusion detection. *ECCV*, pages 211–224, 2006. 2