Self-Occlusion and Disocclusion in Causal Video Object Segmentation

Yanchao Yang[1], Ganesh Sundaramoorthi[2], and Stefano Soatto[1]
[1]University of California, Los Angeles, USA [2]King Abdullah University of Science & Technology (KAUST), Saudi Arabia
yyc8912@g.ucla.edu, ganesh.sundaramoorthi@kaust.edu.sa, soatto@ucla.edu

Abstract

We propose a method to detect disocclusion in video sequences of three-dimensional scenes and to partition the disoccluded regions into objects, defined by coherent deformation corresponding to surfaces in the scene. Our method infers deformation fields that are piecewise smooth by construction without the need for an explicit regularizer and the associated choice of weight. It then partitions the disoccluded region and groups its components with objects by leveraging on the complementarity of motion and appearance cues: Where appearance changes within an object, motion can usually be reliably inferred and used for grouping. Where appearance is close to constant, it can be used for grouping directly. We integrate both cues in an energy minimization framework, incorporate prior assumptions explicitly into the energy, and propose a numerical scheme.

1. Introduction

Persistent tracking of three-dimensional (3D) objects in video presents long-standing challenges unless they are flat [33], or the video is short [25]. As surfaces move in 3D relative to the viewer, previously unseen portions of the scene become visible and will have to be attributed to different objects to maintain tracking. Such *disocclusion* phenomena are the focus of our investigation.

Consider a camera rotating around a box in Fig 1: Both the occluded and disoccluded regions involve portions of different objects, in this case just the box and the "background." Occlusions have been addressed by [29, 1]. We focus on disocclusions, by determining the disoccluded area (Sect. 2), partitioning it and grouping each portion with an object (Sect. 3).

Grouping unseen portions of the scene into different objects requires prior assumptions on their properties. One could assume that the "appearance" or "texture" of objects is homogeneous (*i.e*., their reflectance exhibits spatially stationary statistics) and leverage on the similarity of image color histograms to partition and group disoccluded regions. However, this assumption often fails, as in Fig. 1. Alternatively, one could assume that the "apparent motion" of objects is homogeneous (*i.e*., the deformation undergone by the image domain is smooth within objects, and discontinuous across). However, when objects exhibit "textureless" surfaces (*i.e*., constant reflectance), such a deformation is undetermined, and cannot be used for grouping.

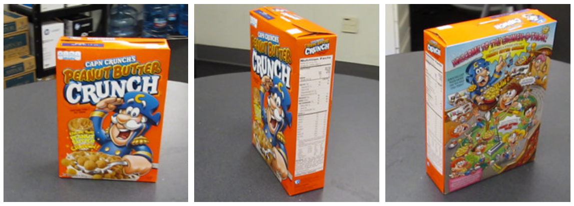


Figure 1. *Relative motion between a three-dimensional scene and the camera (here rotating around the box) causes* disocclusion, i.e., *regions of the image domain where previously unseen portions of the scene project to. Unless objects in the scene are flat, the disocclusion include portions of different objects. Persistent tracking requires detecting the disocclusion and attributing their components to different objects.*

Fortunately, motion and appearance cues are complementary: When one fails to be informative, the other may be. Leveraging such complementarity is central to this paper. When the disoccluded region exhibits complex appearance, motion can be reliably inferred and exploited for grouping. Otherwise, when the disoccluded region is textureless, photometric statistics are spatially homogeneous and can be reliably used for grouping. Of course, both cues can fail if an object has piecewise constant appearance, and the transition happens right at the disocclusion (Fig. 2). However, these are accidental phenomena that do not persist in long temporal sequences.

For us, objects are layouts of piecewise smooth and smoothly deforming surfaces in 3D supporting Lambertian reflection seen under constant illumination throughout a video sequence. There can be multiple objects moving independently, in addition to viewer (or equivalently background) motion. Under these assumptions, the domain of a video image of a scene can be partitioned into two types of regions: Those that are *co-visible*, that under the stated assumptions are a smooth deformation of regions in the previous frame, and those that are *disoccluded*, *i.e.*, whose preimage under perspective projection is a portion of a surface that was not visible in the previous frame(s). In addition, *occluded* regions are subsets of a region that, in the previous frame, was occupied by an object different than the current one. These have been addressed by others [1].

Disoccluded regions in a video are the occluded regions in the video played backwards. Because we eventually aim at real-time closed-loop operation, we wish to process the data *causally*. Furthermore, parts of objects can appear in a frame and disappear in the next, a case which forward-backward sweeps would not address (Sect. 4). With an abuse of nomenclature, we refer to "objects" as both the connected surfaces in 3D, and the subsets of the (2D) image domain where they project.

Contributions: To *detect disocclusions,* we extend the Sobolev framework of [38] to multiple objects (Sect. 2). This framework naturally encompasses coarse-to-fine deformation inference without an explicit regularizer and the associated weighting constant. To *partition and group* disoccluded regions to various objects, we *leverage on the complementarity* of motion and appearance cues by introducing a novel data term that encompasses both (Sect. 3). We derive an efficient numerical scheme and test it against competing methods on benchmark datasets (Sect. 4).

1.1. Related work

Persistent object tracking in video touches upon a large body of work in video segmentation (e.g., [14, 19, 36, 16]), tracking (e.g., [35, 3, 12, 20]), optical flow (e.g., [15, 6, 7, 39, 26]), and motion segmentation (e.g., [33, 27]). In dealing with visibility phenomena, our work relates to occlusion detection. There is a literature on detecting occluding boundaries from static images or short-baseline video (see [29] and references therein). Since we tackle persistent tracking, we do not discuss this further. Our work is related to [1] that partitions the image domain into (flat) layers like [33], but in a convex optimization setting after relaxing the ℓ_0 norm to ℓ_1. We detect occlusions without the need for such a relaxation and without the need for regularization of the deformation field, which can cause over-smoothing in some regions, and under-smoothing in others. Instead, following [38] we employ a Sobolev approach [28] (see also [9, 4]) to infer deformation fields that are by construction smooth in a naturally coarse-to-fine manner. On a short time-scale, such deformation fields are related to optical flow, which we do not review here, except for when the flow is partitioned into regions, as in *motion segmentation*. There, the flow field is often assumed to be piecewise parametric. Here we allow each component to be a diffeomorphism to handle articulated and deforming objects without over-segmenting them. Other motion segmentation approaches perform clustering of optical flow, often non-causally [23, 14].

Although our goal is segmentation, our method produces diffeomorphic warps, and relates to diffeomorphic registration, e.g., [4, 32, 10]. We produce a piecewise diffeomorphism of the image rather than a global diffeomorphism as in [4, 32, 10], an assumption that breaks under (dis)occlusions. Also, our warp computation is parameter-free in contrast to [4, 32, 10].

Taylor et al. [30] perform layer segmentation in longer video sequences leveraging occlusion cues, but do not explicitly address the interplay of motion and intensity cues in disocclusion. Similarly, [27] performs layered segmentation by grouping. Only intensity cues are used for the disocclusion in [8, 38].

This work also relates to dense 3D reconstruction of geometry and photometry [18, 22, 37, 13, 17], since an explicit 3D reconstruction of the scene produces as a side effect a partition of the video into regions. However, it requires a static scene, and does not address deforming objects moving independently, which our work addresses.

2. Sobolev Warps and Occlusions

We seek to partition the domain D of a time-varying color image $I_t : D \subset \mathbb{R}^2 \to \mathbb{R}^3$ for $t = 1, 2, \dots,$ into a collection $\{R_i^t\}_{i=1}^N$ of *regions* R_i^t. We omit the time index hereafter for simplicity. These regions are also called "objects," that *move coherently*, as defined next.

The (apparent) motion of each region R_i, also referred to as a *warp* or a *deformation*, is defined in the domain of the image I_t as the map $w_i : R_i \to D$ that transforms I_{t+1} back to I_t. Assuming the scene is Lambertian, illumination is constant, and the image is corrupted by additive zero-mean Gaussian noise, the maximum-likelihood estimate of w_i is obtained by minimizing $E_{\text{warp}}(w_i, O_i)$, given by

$$E_{\text{warp}} = \int_{R_i \backslash O_i} |I_{t+1}(w_i(x)) - I_t(x)|^2 \, \mathrm{d}x + \beta \int_{O_i} \mathrm{d}x, \quad (1)$$

where $O_i \subset R_i$ is the (unknown) *occluded region* that is visible at time t but not at time $t+1$. Note that, although w_i is defined on all of R_i, the data I_{t+1}, I_t only provides evidence of it in the *co-visible* region $R_i \backslash O_i$. To avoid the trivial solution $O_i = R_i$ and thus w_i undetermined, we put a penalty on the occluded area as in [1].

Eq. (1) is reminiscent of many *optical flow* estimation algorithms [15, 6, 7, 39], but there are important differences: First, each warp is restricted to a subset $R_i \subset D$ with no compatibility condition or relation among the different warps. Second, *there is no regularizer* for the warps. Most motion segmentation or optical flow schemes either assume that each warp belongs to a (small-dimensional) parametric family such as the group of affine transformations, or impose a penalty on the (piecewise) smoothness of w_i. Instead, we leverage on the Sobolev framework [28] to impose regularity in a naturally coarse-to-fine framework, while allowing the warps to be arbitrary diffeomorphisms (smooth maps with a smooth inverse). So, rather than adding a regularizer for the warps in (1), we compute each warp as the integral of a smooth time-varying vector field that, at each instant, belongs to a Sobolev space. This allows us to efficiently optimize (1) without imposing global regularization, which may be too much for fine-scale objects, and too little for large ones.

Given the warp w_i, the optimal occlusion O_i is

$$O_i = \{x \in R_i \, : \, |I_{t+1}(w_i(x)) - I_t(x)|^2 > \beta\}. \tag{2}$$

Substituting the expression above into the energy, we obtain

$$E_{\text{warp}}(w_i) = \int_{R_i} \rho(I_{t+1}(w_i(x)) - I_t(x)) \, \mathrm{d}x, \tag{3}$$

which now depends only on the warp w_i, and where

$$\rho(y) = |y|^2 \text{ for } |y|^2 < \beta \quad \text{and} \quad \rho(y) = \beta \text{ for } |y|^2 \geq \beta \tag{4}$$

With this, we can finally clarify the notion of "coherent motion" used to define the regions R_i: *A region R_i moves coherently if there is a warp w_i that is smooth according to the Sobolev metric, that (locally) minimizes (3).*

The gradient of E_{warp}, $G_i : w_i(R_i) \to \mathbb{R}^2$, with respect to the Sobolev metric has been computed by [38] and is

$$G_i(x) \doteq \nabla_{Sob} E(w_i)(x) = \text{avg}(F_i) + \frac{1}{\alpha} \tilde{G}_i(x), \tag{5}$$

where $\alpha > 0$ is a parameter that will be eliminated below, $F_i : w_i(R_i) \to \mathbb{R}^2$ is

$$F_i = \nabla I_{t+1} \nabla \rho(I_{t+1} - I_t \circ w_i^{-1}) \det \nabla w_i^{-1}, \tag{6}$$

$\text{avg}(F_i)$ is the average over $w_i(R)$, ∇ is the vector of partials, and $\tilde{G}_i$ satisfies the partial differential equation (PDE):

$$\begin{cases} -\Delta \tilde{G}_i(x) = F_i(x) - \text{avg}(F_i) & x \in w_i(R_i) \\ \nabla \tilde{G}_i(x) \cdot N = 0 & x \in \partial w_i(R_i) \\ \text{avg}(\tilde{G}_i) = 0 & \end{cases}, \tag{7}$$

where Δ is the Laplacian, N is normal to $\partial w_i(R_i)$, $\tilde{G}_i$ is the *deformation*, and $\text{avg}(F_i)$ is the *translation*.

To extend the framework to multiple regions, we extend each warp w_i to the entire domain D by imposing $\Delta \tilde{G}_i(x) = 0$ for $x \in D \backslash R_i$ and a Dirichlet condition on ∂R_i. The extension is continuous, but not differentiable across R_i.[1]

Starting with the identity map $w_i(x) = x$, we deform it by the gradient descent (5) as follows. Define $\phi_i^{0,\tau} : D \to D$ and $\phi_i^{\tau,0} : D \to D$ as the evolving warp and its inverse where τ is an artificial time variable parameterizing the evolution. The inverse is needed to compute F_i. The evolution of the warps according to the gradient descent of E_{warp} is

$$G_i^\tau = \nabla_{Sob} E_{\text{warp}}(\phi_i^{0,\tau}), \tag{8}$$

$$\partial_\tau \phi_i^{\tau,0}(x) = \nabla \phi_i^{\tau,0}(x) \cdot G_i^\tau(x), \tag{9}$$

$$\partial_\tau \phi_i^{0,\tau}(x) = -G_i^\tau(\phi_i^{0,\tau}(x)) \tag{10}$$

for all $x \in D$. This gives a coarse-to-fine evolution. One can eliminate the parameter α by noting the independence of the deformation and translation components on α in (5). This gives Algorithm 1, which decreases the energy.

Algorithm 1 *Sobolev Warp Computation*

```
Set $\phi_i^{\tau,0}(x) = \phi_i^{0,\tau}(x) = x$ for $\tau = 0$
repeat
    repeat
        Let $\alpha \to \infty$ so $G_i^\tau = \text{avg}(F_i^\tau)$ is a translation
        Translate: Perform one iteration of (9)-(10)
    until $\text{avg}(F_i^\tau) = 0$.
    Deform: Do one iteration of (9)-(10) with $G_i^\tau = \tilde{G}_i^\tau$
until $\tilde{G}_i^\tau = 0$
Set $w_i = \phi_i^{0,\tau_\infty}$ where $\tau_\infty$ is the convergence time
```

In Section 3.3, we will need to compute the occlusion so that it can be removed in the next frame. It can be computed at the end of the evolution as

$$O_i^{\tau_\infty} = \{x \in R_i \, : \, |I_{t+1}(\phi_i^{0,\tau_\infty}(x)) - I_t(x)|^2 > \beta\}. \tag{11}$$

3. Causal Object Segmentation

If the motion of each region R_i was reliably inferred, one could attempt to propagate forward the R_i to segment the next frame. Unfortunately, regions that become *disoccluded* between t and $t+1$ are not included in any of the R_i. While this is not a major problem if we are interested in only two adjacent frames, t and $t+1$, as the area of the occluded/disoccluded regions is small, as time goes by the disocclusion typically grows. Thus, this phenomenon is hard to ignore when one considers long temporal sequences. The challenge becomes to assign the various components of the disocclusion to existing regions, or to spawn new ones. This is illustrated in Fig. 1: So long as the scene is populated by *non-flat* surfaces, multiple objects contribute to the disoccluded region.

We assume a partition into objects at time $t-1$ and propagate it forward to time t. The disocclusion, *i.e.*, the part of the domain D not covered by the propagated segmentation, is initially assigned to regions based on estimated warps, and this is refined by minimizing the energy in Section 3.1.

[1] While one can define the Sobolev metric over the entire domain D [4], thus naturally having a regular gradient defined over the entire domain D, this is avoided to enable capturing fine-scale structures in a manner that is not influenced by neighboring large-scale structures, for instance an arm swinging near the torso of a person.

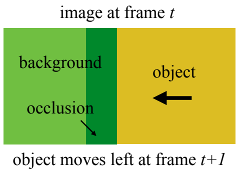


Figure 2. *Illustration of an error that arises in segmentation by grouping pixels only based on motion residuals. The object (dark yellow) moves to the left to occlude a portion of the background (dark green). Pixels in the occluded region are likely to be classified incorrectly in frame* t *if only motion residuals are used since both residuals are large. When the background is constant in the occluded region and around it, classifying by residuals almost certainly leads to misclassifications.*

3.1. Complementarity of Motion and Appearance

Of course both appearance and motion cues are obtained from image irradiance. What we mean by "cues" is bottom-up computation that leverages on the assumption of smooth spatial variation of image irradiance (appearance cues) versus smooth temporal variation of the same (motion cues).

To attribute disoccluded regions to any of the existing objects, we can leverage the photometric regularity and assign each segment to the object that has similar "texture" or motion. We favor the latter, as objects can have spatially-varying appearance, as in the cereal box in Fig. 1. This fails when the object and the background are textureless, as in Figure 2, or when they exhibit similar fine-scale texture. However, in this case grouping by appearance is straightforward. We leverage on this complementarity by exploiting preferentially motion regularity, consistent with our definition of objects, resorting to appearance regularity when the photometry is not suitable to reliably estimate motion.

Textureless regions: To leverage on this complementarity, we use the local standard deviation $\sigma_i(x)$ of I_t in a neighborhood $B_{x,r'} \cap R_i$ where $B_{x,r'} = \{y \in D : |x-y| \leq r'\}$ is the ball of radius r' centered at point x. We can then define a measure of local constancy of any region local to a point x as the minimum standard deviation over all regions that intersect the ball:

$$\underline{\sigma}(x) = \min_{i,\, B_{x,r'} \cap R_i \neq \emptyset} \sigma_i(x). \tag{12}$$

Low values of $\underline{\sigma}(x)$ indicate that the underlying color channels are not *sufficiently exciting* and therefore motion estimates can be expected to be unreliable.

Motion ambiguity function: Grouping by residuals also should not be done when current warp residuals are large. Define the forward, backward and minimum residuals as

$$\text{Res}_i^f(x) = |I_{t+1}(w_i^f(x)) - I_t(x)|^2 \tag{13}$$

$$\text{Res}_i^b(x) = |I_t(w_i^b(x)) - I_{t-1}(x)|^2 \tag{14}$$

$$\text{Res}_i(x) = \min\{\text{Res}_i^f(x), \text{Res}_i^b(x)\} \tag{15}$$

where w_i^f and w_i^b are the current forward and backward warps of region R_i. The backward residual is used to remove some ambiguity in Fig. 2 as sometimes occluded pixels at time $t+1$ are visible at time $t-1$, and hence the backward motion may be reliable. The minimum of Res_i over all regions that intersect with a ball around x,

$$\underline{\text{Res}}(x) = \min_{i,\, B_{x,r'} \cap R_i \neq \emptyset} \text{Res}_i(x), \tag{16}$$

is small when motion cues are reliable. We define the *motion ambiguity function*, maf : $D \to \{0,1\}$, which indicates whether motion cues are unreliable, as

$$\text{maf}(x) = \begin{cases} 1 & \text{if } \underline{\sigma}(x) < k/r' \text{ or } \underline{\text{Res}}(x) > \beta \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

where $k > 0$ is a parameter, the sensitivity to which is studied empirically in Sect. 4. maf is 1 if the pixel is in or borders a constant region or if all motion residuals are large.

Complementary data term: The cost for $x \in R_i$ is

$$f_i(x) = (1 - \text{maf}(x))\text{Res}_i(x) - \text{maf}(x) \log p_{i,x}(I_t(x)), \tag{18}$$

where $p_{i,x}$ are local normalized color histograms of the image I_t within the region R_i. Therefore, if the motion is reliable, as defined by the maf, the cost is the residual of the pixel in the region and if the motion is unreliable, the cost is the fidelity of the pixel to the local intensity distribution of the region R_i. The data energy for region R_i is then:

$$E^i_{\text{data}} = \int_{R_i} f_i(x)\, \mathrm{d}x. \tag{19}$$

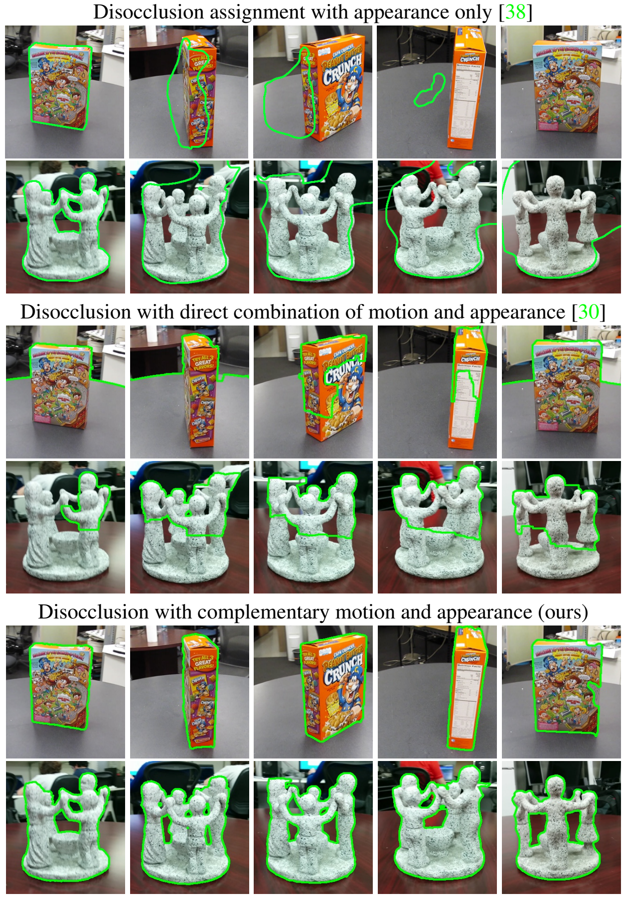
This complementary data term is a key feature in resolving disocclusions (Fig. 3).

3.2. Temporal and Spatial Regularity

To leverage temporal and spatial regularity of the regions, we first note that the warps are regular by construction within the Sobolev framework. We also note that, in between frames, disoccluded regions are small, adjacent to the object they belong to, and typically result in an updated region of similar shape. Thus, if R_i' is the forward warping of the i^{th} region from frame t to $t+1$, we bias the final regions R_i to be close to R_i' in shape and location.

To this end, we construct a local shape similarity prior. Measuring the similarity of R_i and R_i' generally requires knowledge of point correspondences. Similar to ICP [5], we assume that $x \in R_i$ corresponds to its closest point in R_i', $\text{cl}_i(x)$, which can be computed efficiently with Fast Marching [24]. Define the local shape similarity, $S_i : R_i \to \mathbb{R}^+$, of R_i within the ball $B_{r,x}$ to R_i' within $B_{r,\text{cl}_i(x)}$ as follows:

$$S_i(x) = \frac{1}{|B_{x,r}|} \int_{B_{x,r}} |\mathbf{1}_{R_i}(y) - \mathbf{1}_{R_i'}(\text{cl}_i(x) - x + y)| \mathrm{d}y, \tag{20}$$


Disocclusion assignment with appearance only [38]

Disocclusion with direct combination of motion and appearance [30]

Disocclusion with complementary motion and appearance (ours)

Figure 3. *Rotating around an object. Disoccluded parts of an object that have different appearance than the visible parts in the previous frame (cereal box) pose difficulties to existing algorithms. Labeled above are various strategies for addressing disocclusions. Our method also performs well under self-similar appearance (statue), and handles various visibility artifacts from non-convex objects.*

where $\mathbf{1}_R$ is the indicator function of R, and $|B_{x,r}|$ is the area of $B_{r,x}$ (see Fig. 4). The score measures the difference between the shapes $R_i \cap B_{x,r}$ and $R'_i \cap B_{\text{cl}_i(x),r}$ using translation invariant set symmetric difference. The shape similarity energy is:

$$E^i_{\text{shape}} = \int_{R_i} S_i(x)\,\mathrm{d}x. \tag{21}$$

In addition, to bias regions R_i towards being close to R'_i, let $d_{R'_i}$ denote the distance function to $\partial R'_i$, and define

$$E^i_{\text{dist}} = \int_{R_i} d_{R'_i}(x)\,\mathrm{d}x. \tag{22}$$

Finally, we induce spatial regularity of R_i, *i.e.*, nearby points x and y are penalized if they do not belong to the same region. Let

$$W_{R_i} = G_s * (1 - \mathbf{1}_{R_i}) \tag{23}$$

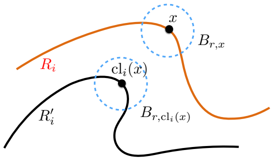


Figure 4. *Illustration of the quantities in the local shape similarity term, S_i. R'_i is the forward warped region and R_i is a candidate in frame $t+1$. The region R_i in a ball around x is compared to R'_i in a ball around $\text{cl}_i(x)$, the closest point on R'_i to x to from $S_i(x)$.*

be a Gaussian smoothing of standard deviation s of the complement of the indicator function of R_i [11]. A large value of $W_{R_i}(x)$ implies that $x \in R_i$ is near many points of $D \backslash R_i$. We induce spatial regularity of R_i by

$$E^i_{\text{smooth}} = \int_{R_i} W_{R_i}(x)\,\mathrm{d}x. \tag{24}$$

3.3. Overall Model and Optimization Method

The assumptions underlying our model are captured by the following energy, which is minimized with respect to the regions R_i:

$$E_{\text{seg}} = \sum_{i=1}^{N} E^i_{\text{data}} + \gamma_{ls} E^i_{\text{shape}} + \gamma_d E^i_{\text{dist}} + \gamma_s E^i_{\text{smooth}}, \tag{25}$$

where $\gamma_{ls}, \gamma_d, \gamma_s > 0$ are weights. We optimize the energy above by a first order approximation to the gradient descent, ignoring terms that involve integrals over R_i. They could be easily included, at a high computational cost and modest performance gain. By defining

$$H_i(x) = f_i(x) + \gamma_{ls} S_i(x) + \gamma_d d_{R'_i}(x) + \gamma_s W_{R_i}(x), \tag{26}$$

we arrive at our optimization scheme in Algorithm 2.

Algorithm 2 *Assigning Disocclusion to Regions*

```
1:  // initialize $R_i$ for gradient descent
2:  Compute propagation of segmentation, $R'_i$ using (27)
3:  Compute disocclusion $\mathbf{D} = D \backslash \cup_i R'_i$
4:  Compute warps of $R'_i$ using Algorithm 1
5:  Compute $H_i$ by substituting $R_i$ with $R'_i \cup \mathbf{D}$
6:  Set $R_i = R'_i \cup \{x \in \mathbf{D} \,:\, H_i(x) \le H_j(x), \forall j\}$
7:  // end initialize
8:  repeat // first order approximation of gradient descent
9:      Update warps of $R_i$ using Algorithm 1
10:     Compute $H_i$
11:     $R_i^{\text{new}} = \{x \in D \,:\, d_{R_i}(x) < \varepsilon,\ H_i(x) \le H_j(x), \forall j\}$
12:     Update regions by $R_i = R_i^{\text{new}}$
13: until $R_i$'s do not change between iterations
```

Algorithm 2 first computes an initialization of regions R_i to the gradient descent (lines 2-6). This is accomplished by

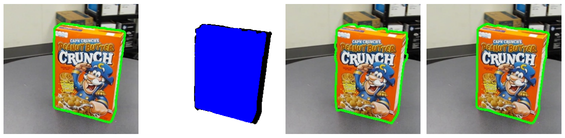


Figure 5. *[Left]: Segmentation from the frame t. [Middle, left]: the propagation of the segmentation from frame t to $t+1$ (black regions indicate disoccluded regions). [Middle, right]: initialization of the regions. [Right]: final segmentation.*

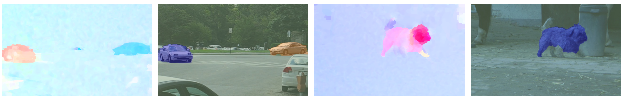


Figure 6. *Illustration of initialization method in the first frame. [Left]: Aggregation of optical flow fields, [Right]: initial segmentation in the first frame.*

propagating forward the segmentation at time $t-1$ to t:

$$R_i' = \{x \in D : \mathbf{1}_{R_i^t \backslash O_i^t}(w_i^{-1}(x)) \geq \mathbf{1}_{R_j^t \backslash O_j^t}(w_j^{-1}(x)), \forall j\} \tag{27}$$

where $O_i^t \subset R_i^t$ is the part of the i^{th} region that is occluded at frame t (11), which is removed, and w_i is the warp from $t-1$ to t. R_i' does not partition all of D because of disocclusion. Therefore, the disoccluded region $\mathbf{D} = D \backslash \cup_i R_i'$ is initially assigned based on motion cues computed from R_i' and other terms in H_i.

With this initialization, the first order approximation to the gradient descent is computed (lines 9-12). Note that the condition, $d_{R_j}(x) < \varepsilon$, is to allow pixel changes only within a band of the boundaries of the current regions so as to approximate the gradient descent. Each step of the warp computation (from t to $t+1$ and from t to $t-1$) in line 9 requires only a few iterations in Algorithm 1 since the warps in the previous iteration of line 9 are close to the final. See Fig. 5 for an example of various stages of this method.

3.4. Initialization for the First Frame

So far we have assumed that, at time t, we have a partition at time $t-1$. This is the case during regime operation when processing a video sequence, but not when $t=0$. For certain applications, such as interactive video segmentation [3, 2], one can assume that the user provides an initial partition. More in general, a number of methods could be employed to obtain an initial partition, using a variety of cues, including semantic labeling from trained detectors. While this process may be costly, it only needs to be performed once as our method affords us the ability to correct initial errors based on motion and appearance regularity.

In the next section, we present results for an initialization performed by clustering optical flow (with regularity (24) using Classic-NL [26]) during a longer initial temporal segment, until enough motion is observed (see Fig. 6).

4. Experiments

Our algorithm aims to segment *objects*, thus we test it on benchmarks with ground truth object annotation: the Freiburg-Berkeley Motion Seg. (FBMS-59) [23], and SegTrack (v1 & v2) [31, 20]. FBMS-59's two sets - training (29 sequences) and test (30 sequences), range between 19-800 frames with multiple objects. SegTrack v2 consists of 14 sequences ranging from 29-279 frames with multiple objects. SegTrack v1 is an earlier version with single objects, which we use to expand the comparison to more methods.

Evaluation: FBMS-59 scores a subset of frames (3-41). Results are reported in terms of precision, recall, F-measure, and the number of objects with $F \geq 0.75$. SegTrack (v1 & v2) evaluates, on all frames, the number of pixels incorrectly classified (v1). Results on v2 are reported as average intersection over union overlap.

Comparisons: On FBMS-59, we compare against a baseline approach [14], one based on clustering motion tracks [23], one segmenting based on occlusion, motion and appearance cues [1], and finally a most recent one integrating motion, appearance, occlusion, and temporal regularity [30]. On SegTrack, we compare to [8] that attempts to solve disocclusions using only appearance and to other state-of-the-art methods [20, 19, 21, 16, 34].

Initialization: On FBMS-59, we report results of our method automatically initialized as described in Sect. 3.4. On SegTrack our method is initialized by the user in frame 1 and compared with similarly initialized methods and also automated methods. Typically, sequences in SegTrack do not have enough object motion in the first few frames to ensure proper initialization.

Parameters: For FBMS-59, we tune the parameters on a few sequences in the training dataset, and then fix them on training and test datasets. On SegTrack, parameters are fixed. Parameters consistent across datasets are $\gamma_{ls} = 0.1, \gamma_d = 0.001, \gamma_s = 5$. Sensitivity of key parameters is addressed later.

Results on FBMS-59 are in Table 1. Figure 7 shows some representative outcomes. Overall our method is more accurate, even compared to non-causal (NC) methods that process the video in batch. This suggests that good disocclusion is key to accurate object segmentation.

Failure Cases on FBMS-59: The main source of error is the automatic initialization in frame 1. This could be mitigated by running our method on multiple candidate initializations, although initialization is not our focus here. To show that better initialization would resolve failures, we show that the results of the 10 most inaccurate cases (typically when an object failed to be detected) improves with user annotation in the first frame (Table 2, Figure 8). Fig. 9 shows that our method recovers from errors in the first frame (short of failed detection).

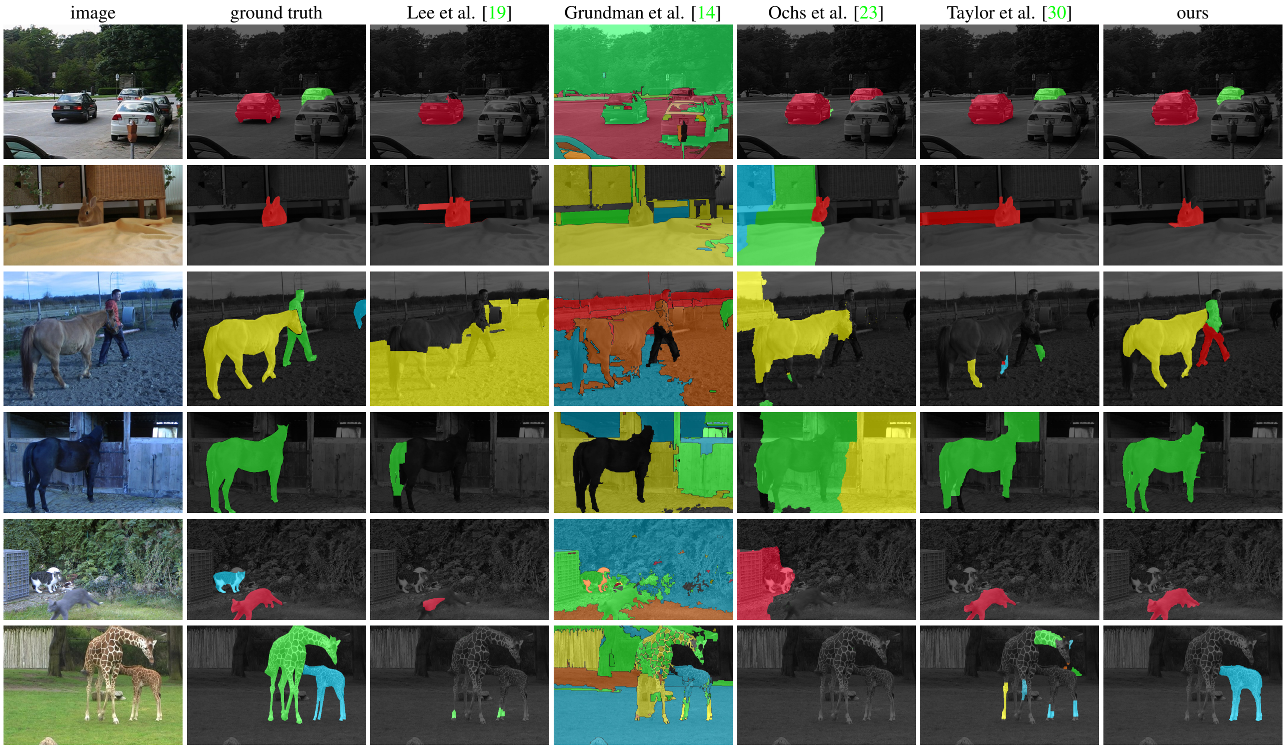


Figure 7. *Sample Visual Results on FBMS-59. Comparison of various state-of-the-art methods. Only a single frame on various sequences are shown. Failure cases (bottom two) in our method typically arise when not enough motion is present in the first few frames.*

	Training set (29 sequences)				Test set (30 sequences)			
	P	R	F	$N/65$	P	R	F	$N/69$
[14]	79.17	47.55	59.42	4	77.11	42.99	55.20	5
[23]	81.50	63.23	71.21	16	74.91	60.14	66.72	20
[1]	87.20	59.60	70.81	17	79.64	50.73	61.98	7
[30]	85.00	67.99	75.55	21	82.37	58.37	68.32	17
[30]-NC	83.00	70.10	76.01	23	77.94	59.14	67.25	15
ours	**89.53**	**70.74**	**79.03**	**26**	**91.47**	**64.75**	**75.82**	**27**

Table 1. *FBMS-59 results. Average precision (P), recall (R), F-measure (F), and number of objects detected (N) over all sequences in the training and test datasets of FMS-59. Higher values indicate superior performance. All methods are fully automatic. [1], [30] and our method are causal; other methods are not.*

	marple9	cats4	farm1	goats1	giraffes1	all
ours (auto)	0.7950	0.7723	0.6730	0.6166	0.7515	0.7217
ours (manual)	0.9782	0.9025	0.7519	0.7505	0.9255	0.8617

Table 2. *Failure cases on FBMS-59 in Fig. 7 can be enhanced with user annotation in the first frame. Thus, the main source of error in our method is the initialization. Results are in terms of F-measure.*

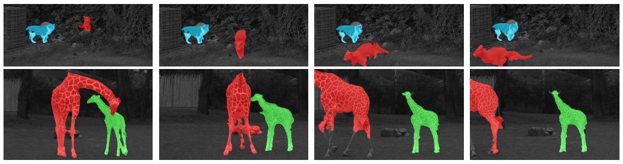


Figure 8. *Sample failure cases (various frames) on FBMS-59 in Fig. 7 are enhanced with user annonation in the first frame.*

Forward-Backward Sweeps on FBMS-59: Although disocclusions are backward-occlusions, addressed extensively in the literature [29, 1], computing disocclusions via forward-backward sweeps followed by a grouping procedure does not perform as well as our method. We compare to the non-causal version of [30], consisting of one forward and one backward pass. Then, advanced grouping is performed based on motion, appearance, temporal continuity, and constraints imposed by occlusions/disocclusions. The result, labeled [30]-NC in Table 1, is worse than ours on all measures. This reaffirms that forward-backward sweeps is not an adequate approach to resolve disocclusions.

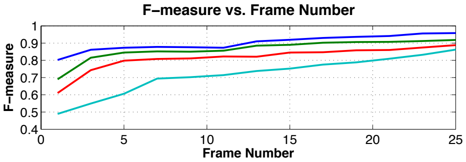


Figure 9. *Results (on FBMS-59) with four different levels of errors in initialization. Errors are mitigated in subsequent frames.*

Results on SegTrack: Table 3. We let the user annotate the first frame, as in [16, 8, 34]. Our method outperforms all others on all but one sequence. That our method outperforms [8] reaffirms that our exploiting complementary motion and appearance cues is beneficial. Results on v2 (Table 4, Fig. 10) show that our method out-performs fully automated ones but also those using user annotation.

	human	ours	[34]	[16]	[8]	[21]	[19]
Mean	347	**409**	535	874	455	677*	740*
Birdfall	130	**144**	163	189	265	189	288
Cheetah	308	623	806	1170	**570**	806	905
Girl	762	**835**	1904	2883	841	1698	1785
Monkeydog	306	**252**	342	333	289	472	521
Parachute	299	**169**	275	228	310	221	201
Penguin	279	**429**	571	443	456	-	136285

Table 3. *SegTrack v1 results. Evaluation is performed in terms of the number of pixels classified incorrectly; smaller values indicate superior results. Note that our method, [34], [16], and [8] use user annotation in frame 1, and [21], [19] do not.*

	ours	[34]	[20]	[19]	[14]
Mean per object	**76.4**	71.8	65.9	45.3	51.8
Mean per sequence	**77.0**	72.2	71.2	57.3	50.8
Girl	**91.6**	84.6	89.2	87.7	31.9
Birdfall	77.3	**78.7**	62.5	49.0	57.4
Parachute	96.1	94.4	93.4	**96.3**	69.1
CheetahDeer	62.4	**66.1**	37.3	44.5	18.8
CheetahCheetah	**52.2**	35.3	40.9	11.7	24.4
Monkeydog-Monkey	**84.1**	82.2	71.3	74.3	68.3
Monkeydog-Dog	**43.7**	21.1	18.9	4.9	18.8
Penguin1	94.0	**94.2**	51.5	12.6	72.0
Penguin2	82.1	**91.8**	76.5	11.3	80.7
Penguin3	78.4	**91.9**	75.2	11.3	75.2
Penguin4	86.3	**90.3**	57.8	7.7	80.6
Penguin5	**77.1**	76.3	66.7	4.2	62.7
Penguin6	**89.0**	88.7	50.2	8.5	75.5
Drifting Car1	**82.3**	67.3	74.8	63.7	55.2
Drifting Car2	**77.6**	63.7	60.6	30.1	27.2
Hummingbird1	39.0	**58.3**	54.4	46.3	13.7
Hummingbird2	69.0	50.7	72.3	**74.0**	25.2
Frog	**76.7**	56.3	72.3	0	67.1
Worm	83.4	79.3	82.8	**84.4**	34.7
Soldier	**84.0**	81.1	83.8	66.6	66.5
Monkey	85.1	**86.0**	84.8	79.0	61.9
Bird of Paradise	**96.1**	93.0	94.0	92.2	86.8
BMXPerson	**92.8**	88.9	85.4	87.4	39.2
BMXBike	32.5	5.70	24.9	**38.6**	32.5

Table 4. *SegTrack v2. The evaluation is performed in terms of the overlap of the best segments; larger values indicate superior results. Our method and [34] uses user annotation in frame 1.*

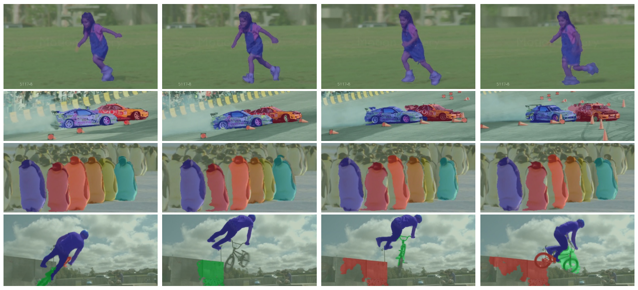


Figure 10. *Sample SegTrack v2 results of our method.*

Sensitivity to Key Parameters: These include the ball size r' and the threshold parameter k in our textureless region detector (12) and (17). To this end, we plot PR curves (measured in terms of correct/incorrectly classified pixels) by fixing one parameter and varying the other and vice-versa. Results (Fig. 11) on the cereal box and statue sequences show that within the operating range, precision does not drop much as recall is increased.

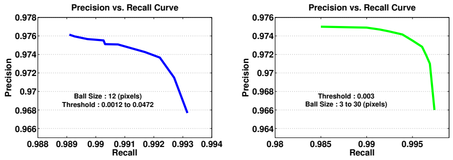


Figure 11. *Analysis of sensitivity of key parameters (the threshold and ball size of the textureless detector). [Left]: ROC curve fixing the ball size and varying the threshold. [Right]: ROC curve fixing the threshold and varying the ball size.*

Computational cost and implementation: Our unoptimized C++ implementation is available[2]. The costliest component is solving for the warps. This requires solving a linear PDE, for which there are many available fast-solvers that could be leveraged. We used conjugate gradient, which can be sped up. The overall cost of our algorithm varies with the amount of deformation between frames. Using a 3.1GHz 12-core processor (with parallelization for the gradient (7)), processing one frame on FBMS-59 takes on average 30 secs.

5. Discussion

We propose a method for handling disocclusion in object tracking that does not require explicit motion regularization, operates naturally in a coarse-to-fine framework, and leverages complementary motion and appearance cues. Our method exhibits reduced dependency on tuning parameters than competing ones, and mitigates typical failures modes.

Our approach assumes that a current estimate of the partition into objects is given at time t to infer the same at $t+1$. If the given partition is nonsensical, most likely so will be the output of our inference scheme. This issue is particularly cogent at time $t = 0$. It can be addressed by spawning multiple trackers corresponding to different initialization hypotheses, later aggregating them through a voting scheme. In many tracking applications, however, the user decides what s/he wants to be tracked, so at least a rough initial partition is available. This is the case for interactive video post-processing [3]. In this case, it would be best to process the entire sequence non-causally, although in some cases processing a sliding batch is still desirable to avoid excessive delay in the interaction with the user.

Real-time operation remains a challenge, but our method has potential since we process data causally, and we use optimization methods that are rapidly evolving, so we can benefit from their improvements.

Acknowledgements

This work was partially supported by ARO W911NF-15-1-0564/66731-CS, AFOSR FA9550-15-1-0229, and KAUST OCRF-2014-CRG3-62140401.

[2] Code: https://github.com/ycyang12/SODVS

References

[1] A. Ayvaci and S. Soatto. Detachable object detection: Segmentation and depth ordering from short-baseline video. *PAMI*, 34(10):1942–1951, 2012. 1, 2, 6, 7

[2] 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. 6

[3] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut: robust video object cutout using localized classifiers. *ACM Transactions on Graphics (TOG)*, 28(3):70, 2009. 2, 6, 8

[4] M. Beg, M. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *IJCV*, 61(2):139–157, 2005. 2, 3

[5] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Robotics-DL*, pages 586–606, 1992. 4

[6] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63(1):75–104, 1996. 2

[7] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pages 25–36. Springer, 2004. 2

[8] J. Chang and J. W. Fisher. Topology-constrained layered tracking with latent flow. In *ICCV*, pages 161–168. IEEE, 2013. 2, 6, 7, 8

[9] G. Charpiat, P. Maurel, J.-P. Pons, R. Keriven, and O. Faugeras. Generalized gradients: Priors on minimization flows. *IJCV*, 73(3):325–344, 2007. 2

[10] A. Cifor, L. Risser, D. Chung, E. M. Anderson, J. Schnabel, et al. Hybrid feature-based diffeomorphic registration for tumor tracking in 2-d liver ultrasound images. *Medical Imaging, IEEE Transactions on*, 32(9):1647–1656, 2013. 2

[11] S. Esedog, Y.-H. R. Tsai, et al. Threshold dynamics for the piecewise constant mumford–shah functional. *Journal of Computational Physics*, 211(1):367–384, 2006. 5

[12] C. Gentile, O. Camps, and M. Sznaier. Segmentation for robust tracking in the presence of severe occlusion. *Image Processing, IEEE Transactions on*, 13(2):166–178, 2004. 2

[13] G. Graber, T. Pock, and H. Bischof. Online 3d reconstruction using convex optimization. In *1st Workshop on Live Dense Reconstruction From Moving Cameras, ICCV 2011*, 2011. 2

[14] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, pages 2141–2148. IEEE, 2010. 2, 6, 7, 8

[15] B. Horn and B. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[16] S. D. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In *Computer Vision–ECCV 2014*, pages 656–671. Springer, 2014. 2, 6, 7, 8

[17] H. Jin, S. Soatto, and A. Yezzi. Multi-view stereo reconstruction of dense shape and complex appearance. *Intl. J. of Comp. Vis.*, 63(3):175–189, 2005. 2

[18] K. Kolev, T. Brox, and D. Cremers. Fast joint estimation of silhouettes and dense 3d geometry from multiple images. *PAMI*, 34(3):493–505, 2012. 2

[19] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, pages 1995–2002. IEEE, 2011. 2, 6, 7, 8

[20] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, pages 2192–2199. IEEE, 2013. 2, 6, 8

[21] T. Ma and L. J. Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, pages 670–677. IEEE, 2012. 6, 8

[22] R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In *CVPR*, pages 1498–1505. IEEE, 2010. 2

[23] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *PAMI*, 36(6):1187–1200, 2014. 2, 6, 7

[24] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. *PNAS*, 93(4):1591–5, 1996. 4

[25] A. Stein and M. Hebert. Incoporating background invariance into feature-based object recognition. In *WACV*, 2005. 1

[26] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, pages 2432–2439. IEEE, 2010. 2, 6

[27] D. Sun, J. Wulff, E. B. Sudderth, H. Pfister, and M. J. Black. A fully-connected layered model of foreground and background flow. In *CVPR*, pages 2451–2458. IEEE, 2013. 2

[28] G. Sundaramoorthi, A. Yezzi, and A. C. Mennucci. Sobolev active contours. *IJCV*, 73(3):345–366, 2007. 2

[29] 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. IEEE, 2011. 1, 2, 7

[30] B. Taylor, V. Karasev, and S. Soatto. Causal video object segmentation from persistence of occlusions. In *CVPR*. IEEE, 2015. 2, 5, 6, 7

[31] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg. Motion coherent tracking using multi-label mrf optimization. *IJCV*, 100(2):190–202, 2012. 6

[32] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Symmetric log-domain diffeomorphic registration: A demons-based approach. In *MICCAI 2008*, pages 754–761. Springer, 2008. 2

[33] J. Y. Wang and E. H. Adelson. Representing moving images with layers. *IEEE TIP*, 3(5):625–638, 1994. 1, 2

[34] L. Wen, D. Du, Z. Lei, S. Z. Li, and M.-H. Yang. JOTS: Joint online tracking and segmentation. In *CVPR*, 2015. 6, 7, 8

[35] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+ 3d active appearance models. In *CVPR (2)*, pages 535–542, 2004. 2

[36] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *Computer Vision–ECCV 2012*, pages 626–639. Springer, 2012. 2

[37] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *ECCV*, 2006. 2

[38] Y. Yang and G. Sundaramoorthi. Shape tracking with occlusions via coarse-to-fine region-based sobolev descent. *PAMI*, 37(5):1053–1066, 2015. 2, 3, 5

[39] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. *Pattern Recognition*, pages 214–223, 2007. 2