Int J Comput Vis (2008) 79: 71–84
DOI 10.1007/s11263-007-0097-1

Dynamic Shape and Appearance Modeling via Moving and Deforming Layers

Jeremy D. Jackson · Anthony J. Yezzi · Stefano Soatto

Received: 15 May 2006 / Accepted: 20 September 2007 / Published online: 29 December 2007

Abstract We propose a model of the shape, motion and appearance of a scene, seen through a sequence of images, that captures occlusions, scene deformations, unconstrained viewpoint variations and changes in its radiance. This model is based on a collection of overlapping layers that can move and deform, each supporting an intensity function that can change over time. We discuss the generality and limitations of this model in relation to existing ones such as traditional optical flow or motion segmentation, layers, deformable templates and deformotion. We then illustrate how this model can be used for inference of shape, motion, deformation and appearance of the scene from a collection of images. The layering structure allows for automatic inpainting of partially occluded regions. We illustrate the model on synthetic and real sequences where existing schemes fail, and show how suitable choices of constants in the model yield existing schemes, from optical flow to motion segmentation and inpainting.

Keywords Shape modeling · Appearance modeling · Non-rigid registration · Rigid registration · Active contours

1 Introduction

We are interested in modeling video sequences where changes occur over time due to viewer motion, motion or deformation of objects in the scene—including occlusions—and appearance variations due to the motion of objects relative to the light sources. A suitable model will trade off generality, by allowing variations of shape, motion and appearance, with tractability, by being amenable to inference and analysis. The goal of modeling is to support inference, and depending on the application one may be more interested in recovering shape (e.g. in shape analysis, classification, recognition, registration), or recovering motion (e.g. tracking, optical flow), or appearance variations (e.g. segmentation) including restoration (inpainting). Traditionally, the modeling task has been approached by making strict assumptions on some of the unknowns in order to recover the others, for instance the brightness-constancy assumption in optical flow, or the affine warping in shape analysis and registration. This is partly justified because in any image-formation model there is ambiguity between the three factors—shape, motion and appearance—and therefore the most general inference problem is ill-posed. In some applications, for instance video compression, the ambiguity is moot since all that matters is for the model to capture the sequence as faithfully and parsimoniously as possible. Nevertheless, since all three factors affect the generation of the image, a more germane approach would call for modeling all three jointly, then letting complexity dictate the responsibility of each factor, and the application dictate the choice of suitable regularizers to make the inference algorithms well posed. We therefore concentrate our attention on modeling, not on any particular application.

We propose a model of image formation that is general enough to capture shape, motion and appearance variations (Sect. 2), and simple enough to allow inference (Sect. 4). We want to be able to capture *occlusion phenomena*, hence our model will entail a notion of *hierarchy* or *layering*; we want to capture image variability due to arbitrary *changes*

J.D. Jackson (✉) · A.J. Yezzi
School of Electrical Engineering, Georgia Institute of Technology, 777 Atlantic Drive NW, Atlanta, GA 30332-0250, USA
e-mail: jeremydjackson@gmail.com

S. Soatto
Department of Computer Science, University of California at Los Angeles, 405 Hilgard Ave., Los Angeles, CA 90095-1596, USA

in viewpoint for non-planar objects, hence our model will entail *infinite-dimensional deformations* of the image domain. Such deformations can be due to changes in viewpoint for a rigid scene, or changes of shape of the scene seen from a static viewpoint, or any combination thereof. Our model will not attempt to resolve this ambiguity, since that requires higher-level knowledge. Furthermore, we want to capture large-scale motion of objects in the scene, as opposed to deformations, hence we will allow for a choice of a finite-dimensional group, e.g. Euclidean or affine, separate from infinite-dimensional deformations. An added benefit of this approach is that higher-level knowledge of viewpoint changes may be incorporated through an added prior on this finite-dimensional group to resolve the ambiguity addressed above. Finally, we want to capture changes in appearance, hence scene radiance will be one of the unknowns in our model. Changes in radiance can come from changes in reflectance or changes in illumination, including changes in the mutual position between the light sources and the scene; again we do not attempt to resolve this ambiguity, since that requires higher-level knowledge. The image-formation model we propose is not the most general that one can conceive; far from it. Indeed, it is far less general than the simplest models considered acceptable in Computer Graphics, and we illustrate the lack of generality in Sect. 3. Nevertheless, it is more general than any other model used so far for motion analysis in Computer Vision, as we discuss also in Sect. 3, and is complex enough to be barely tractable with the analytical and computational tools at our disposal today. We pose the inference problem within a variational framework, involving partial differential equations, integrated numerically in the level set framework (Osher and Sethian 1988), although any other computational scheme of choice would do, including stochastic gradients or Markov-chain Monte Carlo. The point of this paper is to propose a model for shape and appearance of layers and therefore a scene and show that it can be inferred with at least one particular computational scheme, not to advocate a particular optimization technique.

1.1 Relation to Existing Work

This work relates to a wide body of literature in scene modeling, motion estimation, shape analysis, segmentation, and registration which cannot be properly reviewed in the limited space available. In Sect. 3 we illustrate the specific relationship between the model we propose and existing models. These include Layers (Wang and Adelson 1994; Hsu et al. 1992), which only model affine deformations of the domain and can therefore only capture planar scenes under small viewer motion or small aperture, and where there is no explicit spatial consistency within each layer and the appearance of each layer is fixed. As we will illustrate, our model allows deformations that can model arbitrary viewpoint variation, model layer deformation and enforce spatial coherence within each layer. One could think of our work as a generalization of existing work on Layers to arbitrary viewpoint changes, or arbitrary scene shape, and to changes in radiance (texture), all cast within a variational framework.

Our work relates to a plethora of variational algorithms for optical flow computation, for instance (Schnörr 1992; Alvarez et al. 1999; Deriche et al. 1995) and references therein, except that we partition the domain and allow arbitrary smooth deformations as well as changes in appearance (that would violate the brightness constancy constraints that most work on optical flow is based on, with a few exceptions, e.g. Haussecker and Fleet 2001). It also relates to various approaches to motion segmentation, where the domain is also partitioned and allowed to move with a simple motion, e.g. Euclidean or affine, see for instance (Cremers 2003) and references therein. Such approaches do not allow deformations of the region boundaries, or changes in the intensity within each region. Furthermore, they realize a partition, rather than a hierarchy, of domain deformations, so our model can be thought of as motion segmentation with moving and deforming layers with changes in intensity and inpainting (Bertalmio et al. 2000). In this, our work relates to (Soatto and Yezzi 2002), except that we allow layers to overlap. So, our work can be thought of as a layered version of Deformotion with changes in region intensities. Also relevant to our work is (Paragios et al. 2003) where one distance function is registered to another using rigid and non-rigid transformations. Our work relates to deformable templates (Grenander 1993; Miller and Younes 1999), in the sense that each of our layers will be a deformable template. However, we do not know the shape and intensity profile of the template, so we estimate that along with the layering structure. A one-layer version of our work is similar to (Trouve and Younes 2005) where the author describes energies on the manifold $G \times M$ where $g \in G$ is a group action (possibly a C^∞ diffeomorphism or an affine transformation) and M is a manifold consisting of a collection of landmark points or images). For the example of G being the set of C^∞ diffeomorphisms and M being the set of images, the geodesic between two points $(g_1, m_1), (g_2, m_2) \in G \times M$ describes metamorphoses from one "group-image" pair to another. Our work is also related to active appearance models (Cootes et al. 1998; Baker et al. 2003), in that we seek the same goal, although rather than imposing regularization of shape and appearance by projection onto suitably inferred linear subspaces we employ generic regularizers. One can therefore think of our work as a generalization of active appearance models to smooth shape and intensity deformations, cast in a variational framework. Of course this work relates more generically to active contours, e.g. (Blake and Isard 1998;

Kichenassamy et al. 1995; Caselles et al. 1997; Paragios and Deriche 2000) and references therein. Finally, our joint estimation of both geometric and photometric unknowns follows a similar spirit found in (Marks et al. 2005) where pose and texture are jointly estimated through the use of conditionally Gaussian filters. In the next section we introduce our model, and in Sect. 4 we illustrate our approach to infer its (infinite-dimensional) constitutive elements.

2 Modeling

We represent a scene as a collection of L overlapping *layers*. Each layer, labeled by an index $k = 1, \ldots, L$, is a function that has associated with it a domain, or *shape* $\Omega^k \subset \mathbb{R}^2$, and a range, or *radiance* $\rho^k : \Omega^k \to \mathbb{R}^+$. Layer boundaries model the occlusion process, and each layer k undergoes a *motion*, described by a (finite-dimensional) group action g^k, for instance $g^k \in \mathbb{SE}(2)$ (the group of rigid motion on the plane) or the affine group $\mathbb{A}(2)$, and a *deformation*, or *warping*, described by a diffeomorphism $w^k : \Omega^k \to \mathbb{R}^2$, in order to generate an image I at a given time t. The warping models changes of viewpoint for non-planar scenes, or actual changes in the shape of objects in the scene. Since each image is obtained from the given scene after a different motion and deformation, we index each of **the image's corresponding variables** by t: g_t^k, w_t^k, and I_t. Finally, since layers occlude each other, there is a natural ordering in k which, without loss of generality, we will assume to coincide with the integers: Layer $k = 1$ is occluded by layer $k = 2$ and so on. But since this occlusion model could change, say layer $k = 2$ goes behind layer $k = 3$ and then later layer $k = 2$ is in front of layer $k = 3$, there is a function $l = \max\{k \mid x \in \Omega^k\}$ that indicates the layer that will contribute to the intensity at a pixel in a given image which is the frontmost **layer** that intersects the warped domain. For simplicity we assume that $\Omega^0 = \mathbb{R}^2$ (the backmost layer, or "the background"). With this notation, the model of how the value of the generic image $I_t : \Omega^0 \to \mathbb{R}^+$ at the location $x \in \Omega^0 \subset \mathbb{R}^2$ is generated can be summarized as $I_t(g_t^l \circ w_t^l(x)) = \rho^l(x)$, with $x \in \Omega^l$, $l = \max\{k \mid x \in \Omega^k\}$. To simplify the notation, we call $x_t^l \doteq g_t^l \circ w_t^l(x)$, which sometimes we indicate, for simplicity, as x_t, so that

$$\begin{cases} I_t(x_t^l) = \rho^l(x), & x \in \Omega^l, \\ x_t^l = g_t^l \circ w_t^l(x), & l = \max\{k \mid x \in \Omega^k\}. \end{cases} \tag{1}$$

Our goal in this work is to infer, to the extent possible, the radiance family $\{\rho^k\}_{k=1,\ldots,L}$, the shape family $\{\Omega^k\}_{k=1,\ldots,L}$, the motions $\{g_t^k\}_{k=1,\ldots,L;t=1,\ldots,N}$ and the deformations $\{w_t^k\}_{k=1,\ldots,L;t=1,\ldots,N}$ that minimize the discrepancy of the measured images from the ideal model (1), subject to generic regularity constraints. Such a discrepancy is measured by a cost functional $\phi(\Omega^k, \rho^k, w_t^k, g_t^k)$ to be minimized

$$\begin{aligned} \phi \doteq & \sum_{t=1}^{N} \int_{\Omega^0} (I_t(x_t) - \rho^l(w_t^{l-1} \circ g_t^{l-1}(x_t)))^2 dx_t \\ & + \zeta \sum_{k=1}^{L} \int_{\partial\Omega^k} ds + \lambda \sum_{k=1}^{L} \int_{\Omega^k} \|\nabla \rho^k(x)\|^2 dx \\ & + \mu \sum_{k,t=1}^{L,N} \int_{\Omega^l} r(w_t^k)(x) dx \end{aligned}$$

$$\text{subject to } l = \max\{k \mid x \in \Omega^k\}. \tag{2}$$

Here r is a regularizing functional, for instance $r(w) \doteq |\dot{w}| + \frac{1}{|\dot{w}|}$ where $|\dot{w}|$ is the determinant of the Jacobian Matrix (with respect to x) of w. Since it is desirable to keep w to be a one-to-one function this regularizer r keeps $|\dot{w}|$ close to one. If $|\dot{w}|$ deviates from 1 then either one of the terms $|\dot{w}|$ and $\frac{1}{|\dot{w}|}$ gets bigger. λ, μ, and ζ are positive constants. Note that l is a function, specifically $l : \Omega^0 \to \mathbb{Z}^+$. We have chosen the two-norm for the data-dependent term and the regularizer for simplicity, but other choices would of course do as well.

3 Generality of the Model

It can be easily shown that (1) models images of 3-D scenes with piecewise smooth geometry exhibiting Lambertian reflection with piecewise smooth albedo[1] viewed under diffuse illumination from an arbitrarily changing viewpoint. It does not capture global or indirect illumination effects, such as cast shadows or inter-reflections, complex reflectance, such as specularities, anisotropies or sub-surface scattering. These are treated as modeling errors and are responsible for the discrepancy between the model and the images, which is measured by ϕ in (2). We lump these discrepancies together with sensor errors and improperly call them "noise." Although far from general, (1) is nevertheless a more ambitious model than has ever been used in the context of motion estimation and tracking. In fact, many existing models are special cases of (1).

We start by showing how the model includes traditional **optical flow** as a special case. In particular, if we assume a single layer to represent the whole image domain (i.e. $L = 0$), a trivial group action (i.e. $g = Id$) and no regularity in the modeled radiance $\rho = \rho^0$ (i.e. $\lambda = 0$) then the resulting minimization problem includes only the radiance ρ and the warps $w_1 = w_1^0$ and $w_2 = w_2^0$ as unknowns (we consider

[1] The model can be further generalized by allowing ρ^l to be vector-valued to capture a set of radiance statistics such as the coefficients of a filter bank or other texture descriptors, but this is beyond the scope of this paper.

the case of just two images I_1 and I_2 for now). We are therefore left with the much simpler energy

$$\phi(\rho, w_1, w_2) = \sum_{t=1}^{2} \int_{\Omega^0} (I_t(x_t) - \rho({w_t}^{-1}(x_t)))^2 dx_t + \mu \sum_{t=1}^{2} \int_{\Omega^0} r(w_t)(x) dx. \quad (3)$$

If our goal is just to find the warp $w = w_2 \circ w_1^{-1}$ that registers I_1 to I_2 (through the common radiance model ρ), then we may further simplify things by setting $w_1 = Id$ and $w_2 = w$, thereby eliminating yet another unknown and yielding (up to a change of measure corresponding to the Jacobian Matrix of w, which is $\dot{w}$)

$$\phi(\rho, w) = \int_{\Omega^0} (I_1(x) - \rho(x))^2 + (I_2(w(x)) - \rho(x))^2 dx + \mu \int_{\Omega^0} r(w)(x) dx. \quad (4)$$

Since we have omitted the smoothness penalty on ρ, it is straightforward to show for a given choice of w that (4) is minimized by setting $\rho(x)$ to the mean of $I_1(x)$ and $I_2(w(x))$. Thus, in this special case (no smoothness on ρ) we may replace the joint optimization in (4) with a direct optimization of w through this substitution of ρ. The resulting energy

$$\phi(w) = \frac{1}{2} \int_{\Omega^0} (I_1(x) - I_2(w(x)))^2 dx + \mu \int_{\Omega^0} r(w)(x)\, dx, \quad (5)$$

depending upon the choice of the regularizer r (note that r typically depends on the derivatives of w rather than its direct values), corresponds to either the classical optical flow in (Horn and Schunk 1981) or to one of its many variants.

Our model has the advantage of not enforcing global regularization (regularization is imposed within layers, but not across layers), of not comparing images to each other, but to an underlying model (this carries significant advantages when it comes to robustness to noise, as we illustrate with experiments), and of having an explicit model of the appearance of the scene, which allows "inpainting" individual layers while preserving their motion boundaries.

Choosing $L = 1$, $w = Id$, $\lambda = 0$, $\mu = 0$ yields **motion segmentation,** that has also been addressed by many, see for instance (Cremers 2003) and references therein for the case of affine motion $g \in \mathbb{A}(2)$. In motion segmentation one partitions the domain into a number of individually moving segments, each of which is assumed to move with a constant (finite-dimensional) motion. Like in optical flow, there is no model of appearance, and the data-dependent term consists of the brightness constancy constraint which forces direct image-to-image comparison:

$$\phi(g^k, \Omega^k) = \int_{\Omega^k} (I_0(x) - I_1(g^k(x)))^2 dx + \int_{\Omega_{0\setminus k}} (I_0(x) - I_1(x))^2 dx + \zeta \int_{\partial\Omega^k} ds \quad (6)$$

where

$$\Omega_{0\setminus k} = \Omega_0 \setminus \{\Omega^k \cup g^k(\Omega^k)\}. \quad (7)$$

Note that, in this case, we have allowed Ω^k to be one of the unknowns since w^k is no longer part of the inference, although one could easily define $\Omega^k \doteq w^k(\Omega_0)$, as we have discussed in the previous section.

Choosing $L = 1$, $\rho = \text{const}$, and $r(w) = \langle \nabla u, \nabla u \rangle + \langle \nabla v, \nabla v \rangle$ yields a model called **Deformotion** in (Soatto and Yezzi 2002), and has also been extended to grayscale images $L = 1$, $r(w) = \langle \nabla u, \nabla u \rangle + \langle \nabla v, \nabla v \rangle$. Our work is the natural extension of Deformotion to layers.

Choosing $L > 1$, $w = Id$, Ω^k unconstrained and $g \in \mathbb{A}(2)$ would yield a variational version of the **Layers** model (Wang and Adelson 1994), that to the best of our knowledge has never been attempted. Note that this is different than simpler variational multi-phase motion segmentation, since in that case the motion of a phase affects the shape of neighboring phases, whereas in the model (1) layers can overlap without distorting underlying domains. One can think of the Layer model as a multi-phase motion segmentation with inpainting (Bertalmio et al. 2000) of occluded layers and shape constraints.

The model also relates to **deformable templates**, where $\rho = \text{const}$ in the traditional model (Grenander 1993) and $\rho = \text{smooth}$ in the more general version (Miller and Younes 1999). Another relevant approach is **Active Appearance Models** where the regions, warping and radiances are modeled as points in a linear space.

$$w_t^k(x) = w_0^k(x) + W^k(x) s_t \quad (8)$$

where $w_0 : \Omega^k \to \mathbb{R}^2$ and $W^k : \Omega^k \to \mathbb{R}^n$ denotes a set of basis functions or principal components, and $s_t \in \mathbb{R}^n$, $t = 1, \ldots, N$ is a vector of shape coefficients. Similarly,

$$\rho^k(x) = \bar{\rho}^k(x) + P^k(x)\alpha^k \quad (9)$$

where $\bar{\rho}^k : \Omega^k \to \mathbb{R}$ and $P^k(x) : \Omega^k \to \mathbb{R}^n$ is a vector of principal components, and $\alpha^k \in \mathbb{R}^n$ a vector of appearance parameters. Note that the functions P^k and W^k have to satisfy orthogonality constraints, and these have to be enforced during the inference of the bases. The model (1) does not impose such restrictions, and render the problem well-posed by generic regularization instead.

Finally, by virtue of the regularization imposed on ρ, our scheme relates to **image inpainting**, except that we perform inpainting *both* by layer transfer from multiple images and by regularization. The advantage of our method is that it can exploit whatever information is there: If multiple views are available, their contribution is weighted relative to the harmonic interpolation term. If only one image is available, then intensity regularization dictates the filling process.

4 Inference

Minimizing the cost functional in (2) is a tall order. It depends upon each domain Ω^k and its boundary (a closed planar contour), its deformation (a flow of planar diffeomorphisms) w_t^k, the radiance (a piecewise smooth function) ρ^k, all of which are infinite-dimensional unknowns. In addition, it depends on a group action per layer per instant, g_t^k, and on the occlusion model, which is represented by the discrete-valued function $l(x) = \max\{k \mid x \in \Omega^k\}$, and all of this for each layer $k = 1, \ldots, L$.

We proceed by minimizing the functional (2) using simultaneous gradient flows with respect to the groups (motion), the radiances (appearance) and the diffeomorphisms (deformation). The detailed evolution equations are a bit complicated depending upon the number of layers and the occlusion structure between layers. To help avoid excessive subscripting and superscripting and multiple-case definitions according to occlusion relationships, we will outline some of the key properties of the various gradient terms for the case of a background layer Ω^0, a single image I, and a single foreground layer Ω^1. We will also, to help keep the illustration simple, assume that the group action g^0 and the warp w^0 for the background layer are simply the identity transforms. This is the simplest possible scenario that will allow us to still show the key properties of the gradient flows.

We first use another simplification, letting $g = g^1$ and $w = w^1$. Let $\hat{x} = g(w(x))$ and $\hat{\Omega}^1 = g(w(\Omega^1))$. With this notation, we may write our energy functional as follows.

$$E = \int_{\hat{\Omega}^1} (I(\hat{x}) - \rho^1(w^{-1} \circ g^{-1}(\hat{x})))^2 d\hat{x} + \int_{\Omega^0 \setminus \hat{\Omega}^1} (I(\hat{x}) - \rho^0(\hat{x}))^2 d\hat{x} + \int_{\Omega^1} r(w)(x) + \int_{\partial\Omega^1} ds. \quad (10)$$

If η denotes any single parameter (e.g. horizontal translation) of the group g, then differentiating yields

$$\frac{\partial E}{\partial \eta} = \int_{\partial\hat{\Omega}^1} \left\langle \frac{\partial \hat{x}}{\partial \eta}, \hat{N} \right\rangle ((I(\hat{x}) - \rho^1(w^{-1} \circ g^{-1}(\hat{x})))^2 - (I(\hat{x}) - \rho^0(\hat{x}))^2) d\hat{s} + 2 \int_{\hat{\Omega}^1} (I(\hat{x}) - \rho^1(w^{-1} \circ g^{-1}(\hat{x}))) \times \left\langle \nabla\rho(w^{-1} \circ g^{-1}(\hat{x})), [(w^{-1})'] \frac{\partial}{\partial \eta} g^{-1}(\hat{x}) \right\rangle d\hat{x} \quad (11)$$

where $\hat{N}$ and $d\hat{s}$ denote the outward unit normal and the arclength element of $\partial\hat{\Omega}^1$ respectively. For multi-dimensional group, the procedure can be repeated for each parameter in the local coordinate representation of the group.

We are able to note two things. First, the update equations for the group involve measurements both along the boundary of its corresponding layer (first integral) as well as measurements within the layer's interior (second integral). Notice that this latter integral vanishes if a constant radiance ρ is utilized for the layer. We also see that it is not necessary to differentiate the image data I. Derivatives land on the estimated smooth radiance ρ instead, which is a significant computational perk of our model that results in considerable robustness to image noise (Yezzi and Soatto 2001).

A similar gradient structure arises for the case of the infinite dimensional warp w (boundary-based terms and region-based terms for each layer are similar to previous integrals). However, additional terms arise in the gradient flow equations for w depending upon the choice of regularization terms in the energy functional (smoothness penalties, magnitude penalties, etc.).

Here we solve for the transformations of layer 1 so the superscript is dropped on w and g. Let $w(x) = [x + u(x), y + v(x)]^T$ and $r(w)(x) = \langle \nabla u(x), \nabla u(x) \rangle + \langle \nabla v(x), \nabla v(x) \rangle$.

To solve for u and v at time n we use the following iterative explicit method:

$$\begin{bmatrix} u^n(x) \\ v^n(x) \end{bmatrix} = \begin{bmatrix} u^{n-1}(x) - dt * (\delta(\partial\Omega^1) * u_c^{n-1}(x) + u_r^{n-1}(x) - \Delta u^{n-1}(x)) \\ v^{n-1}(x) - dt * (\delta(\partial\Omega^1) * v_c^{n-1}(x) + v_r^{n-1}(x) - \Delta v^{n-1}(x)) \end{bmatrix} \quad (12)$$

where

$$\begin{bmatrix} u_c(x) \\ v_c(x) \end{bmatrix} = \begin{bmatrix} \hat{N}^x(x) \\ \hat{N}^y(x) \end{bmatrix} \times [(I \circ g \circ w(x) - \rho^1(x))^2 - ((I - \rho^0) \circ g \circ w(x))^2] \quad (13)$$

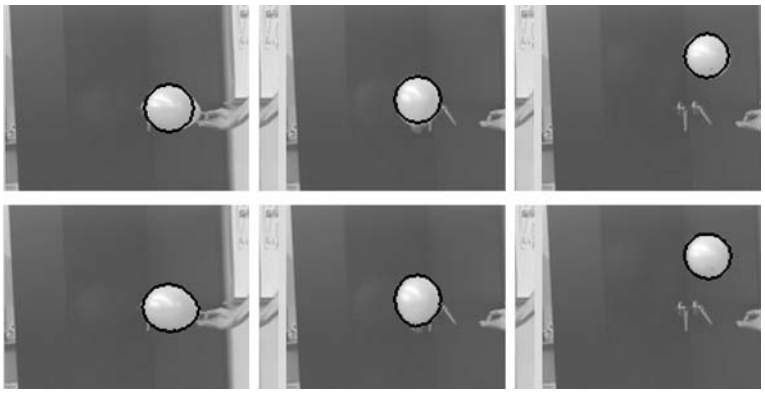

Fig. 1 Tracking a balloon: Three sample views are shown from a sequence of a deflating balloon moving with an erratic motion while changing its shape from a drop-like shape to a circle. In the *top row* we show the boundary of the first layer as estimated by a rigid layer model with a single scaling term that does not allow for layer deformation, akin to a variational implementation of traditional layer models. As it can be seen, the model tracks the motion of the layer, but it fails to capture its deformation. On the *bottom row* we show the same three images with the first layer superimposed, where the layer is allowed to both move (rigidly) and deform (diffeomorphically), yielding 82% lower RMS residual error, and capturing the subtler shape variations

and

$$\begin{bmatrix} u_r(x) \\ v_r(x) \end{bmatrix} = -2|g'||w'(x)| \times \begin{bmatrix} (I \circ g \circ w(x) - \rho^1(x))(-\rho^1_x(x)) \\ (I \circ g \circ w(x) - \rho^1(x))(-\rho^1_y(x)) \end{bmatrix}. \quad (14)$$

Here

$$R = \begin{bmatrix} c\theta & s\theta \\ -s\theta & c\theta \end{bmatrix}$$

is the rotation matrix,

$$S = \begin{bmatrix} xs & 0 \\ 0 & ys \end{bmatrix}$$

is the scaling matrix and

$$J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

is the 90° rotation matrix and

$$\begin{bmatrix} \hat{N}^x(x) \\ \hat{N}^y(x) \end{bmatrix} = S^T R^T J g' w'(x) J^T \begin{bmatrix} N^x(x) \\ N^y(x) \end{bmatrix}. \quad (15)$$

The curve evolution is also similar to the boundary-based term for the evolution of g:

$$\frac{\partial C}{\partial t} = -((I(\hat{x}) - \rho^1(w^{-1} \circ g^{-1}(\hat{x})))^2 - (I(\hat{x}) - \rho^0(\hat{x}))^2)\hat{N}. \quad (16)$$

Finally, the optimality conditions for the smooth radiance functions ρ^0 and ρ^1 are given by the following Poisson-type

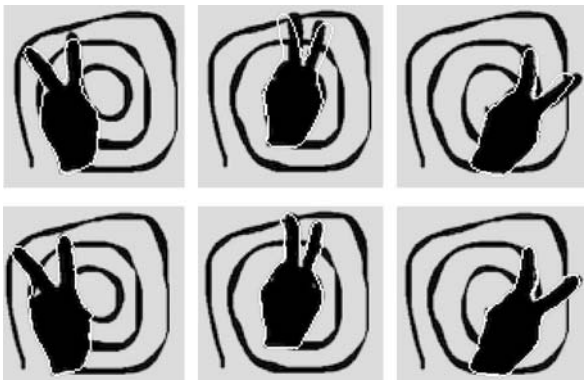

Fig. 2 Victory sign, with deforming hand, moving in front of a partially occluded background portraying a spiral. The goal here is to recover the radiance of each layer (the spiral in the background and the constant black intensity of the hand), as well as the motion and deformation of the foreground layer. Note that current layer models based only on affine motion would fail to capture the phenomenology of this scene by over-segmenting the region into three regions, each moving with independent affine motion. Our model captures the overall motion of the layer with an affine group, and then the relative motion between the fingers as a deformation, as we illustrate in the next Fig. 3

equations

$$\Delta\rho^1(x) = \lambda(\rho^1(x) - I(g \circ w(x))), \quad x \in \Omega^1, \quad (17)$$

$$\Delta\rho^0(x) = \begin{cases} 0, & x \in \hat{\Omega}^1 \\ \lambda(\rho^0(x) - I(x)), & x \in \Omega^0 \setminus \hat{\Omega}^1. \end{cases} \quad (18)$$

Notice that the background radiance ρ^0 is "inpainted" in regions occluded by the foreground layer Ω^1 by harmonic interpolation from the boundary of $\hat{\Omega}^1$, since ρ satisfies Laplace's equation $\Delta\rho^0 = 0$. Once all the terms are put together we can generate a gradient flow that simultaneously evolves all layer assignments, boundaries and intensities. In

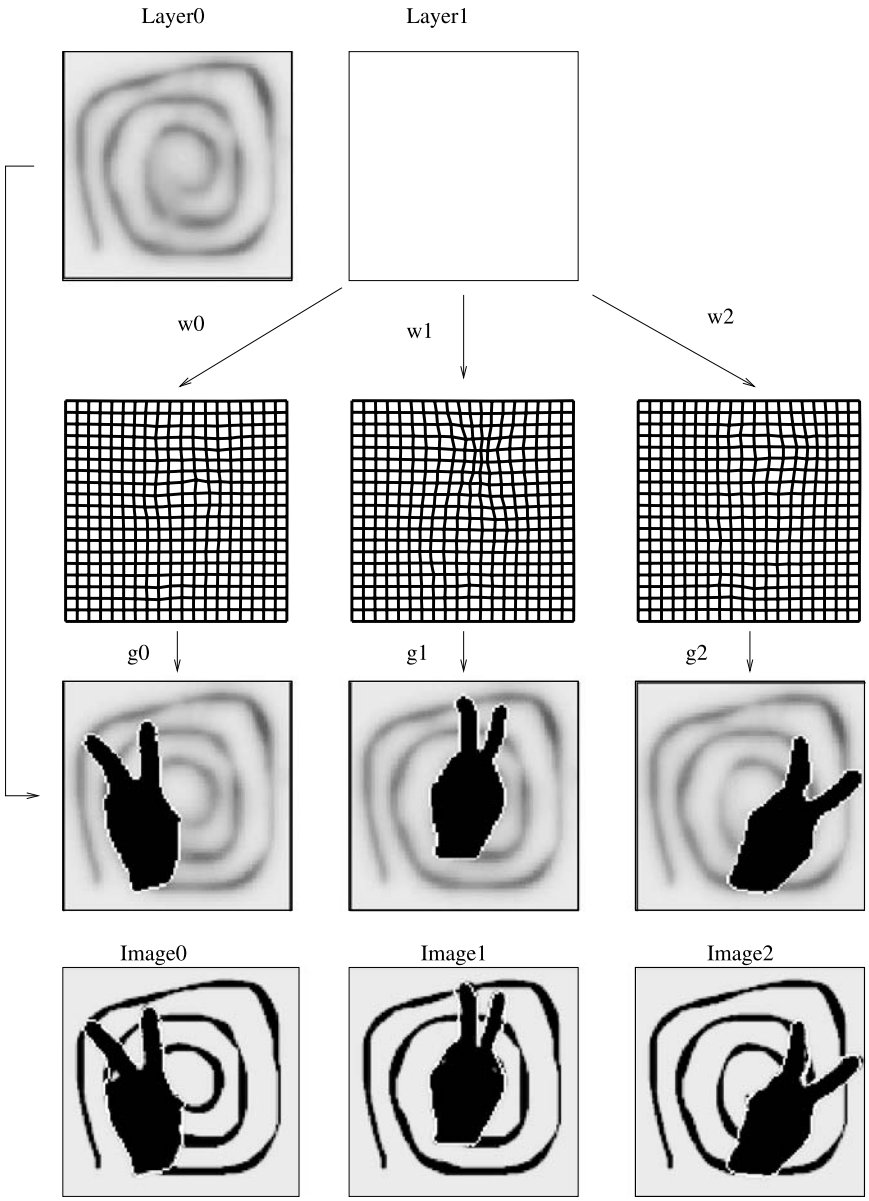

Fig. 3 Multiple layers mapping onto multiple images: The inference process returns an estimate of the albedo in each layer (*top*). Since we are assuming smooth albedo, the spiral is smoothed. The deformation of each layer is estimated (*second row*) together with its affine motion, to yield an approximation of the image (*third row*). This is used for comparison with the measured images (*bottom row*) that drives the optimization scheme

the next section we illustrate some of the features of the model and the resulting optimization, as it compares with existing schemes.

Here for completeness, we expand (11). For a single parameter η from the mapping g_t^k from layer k to an image t:

$$\frac{\partial E}{\partial \eta} = \int_{\partial \Omega_t^k} \delta_l(k, x_t) \left\langle \frac{\partial x_t}{\partial \eta}, N_t \right\rangle ((f^k(x_t))^2 - (h^m(x_t))^2) ds_t + 2 \int_{\Omega_t^k} \delta_l(k, x_t)(f^k(x_t)) \times \left\langle \nabla \rho^k(w_t^{k^{-1}} \circ g_t^{k^{-1}}(x_t)), (w_t^{k^{-1}})' \frac{\partial}{\partial \eta} g_t^{k^{-1}}(x_t) \right\rangle dx_t \tag{19}$$

where

$$f^k(x_t) = I(x_t) - \rho^k(w_t^{k^{-1}} \circ g_t^{k^{-1}}(x_t)) \tag{20}$$

and

$$h^m(x_t) = I(x_t) - \rho^m(w_t^{m^{-1}} \circ g_t^{m^{-1}}(x_t)) \tag{21}$$

and $\delta_l(k, x_t)$ is 1 when $l(x_t) = k$ and 0 otherwise and $m = l - 1$ when $k = l$.

5 Experiments

In the first experiment we illustrate the capability of our model to track deforming layers. In Fig. 1 we show three

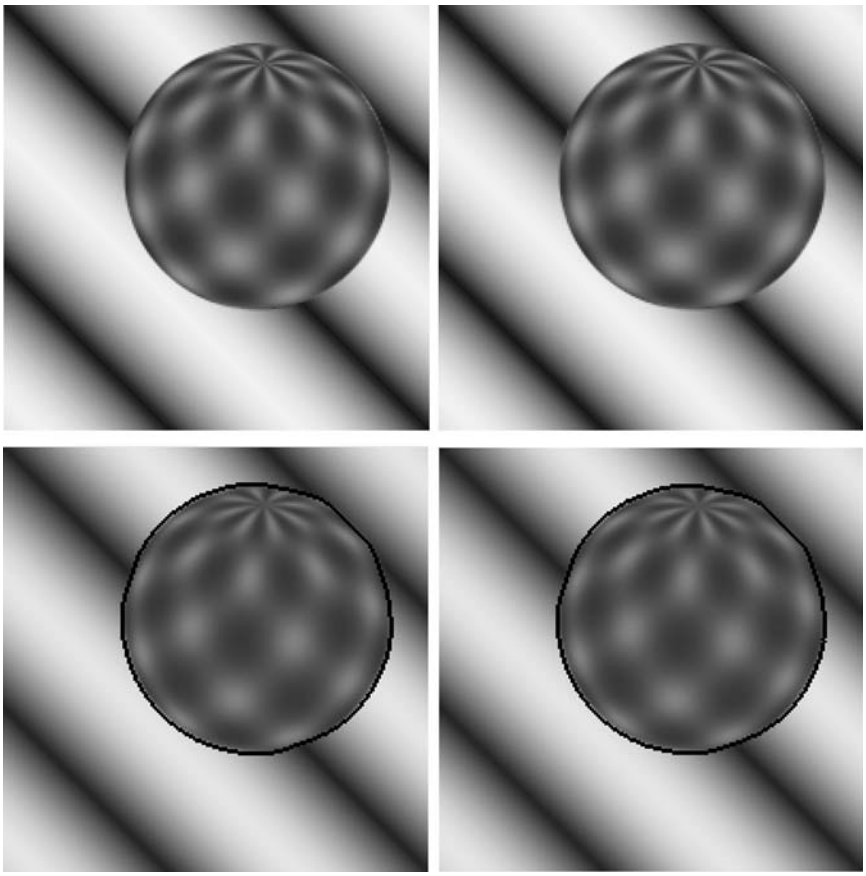

Fig. 4 Rotating sphere and segmentation obtained using deformotion

images of a sequence where a deflating balloon is undergoing a rather erratic motion while deforming from an initial waterdrop shape to a circular one, finally to a drop-like shape. On the top row of Fig. 1 we show the layer boundaries for a model that only allows for rigid deformations of the initial contour (a circle) using a single scaling term. This is essentially a variational implementation of the model of (Wang and Adelson 1994). As it can be seen, it captures the gross motion of the balloon, but it cannot capture the subtler shape variations. The second row shows the same three sample images with the boundary of the first layer superimposed, where the layer is allowed to deform according to the model we have introduced. The data fidelity term used is a Mumford-Shah term so the radiances representing each layer are smooth functions. As it can be seen, the layer changes shape to adapt to the deforming balloon, all while capturing its rather erratic motion. The average RMS (root mean squared) error per image for the affine layer model is 30.87, whereas the residual for the case of the deforming layers is 5.51. More importantly, the phenomenology of the scene, visible in the figure, has been correctly captured.

In the next experiment we illustrate all the features of our model by showing how it recovers the background behind partially occluded layers while recovering their motion and deformation. In Fig. 2 we show a few samples from a dataset where the silhouette of a moving hand forms a victory sign while moving the relative position between the fingers. The background, which is partially occluded, is a spiral. Here we use an average hand shape (constructed from segmentations of other hand images) as the initial shape of the foreground layer to find its affine motion, and then the diffeomorphic warp w_i. Again we assume smooth radiance within each layer, so when we recover the background layer we show a slightly smoothed version of the spiral (of course we could further segment the black spiral from the background and thus obtain sharp boundaries, but this is standard and would not help us illustrate the feature of the model, therefore we do not illustrate it here.)

In Fig. 3 we illustrate the results of this experiments, arranged to summarize the modeling process. On the top row we show the recovered layers. Since we are assuming a smooth radiance within each layer, we can only recover a smoothed version of the spiral. These layers are deformed according to a diffeomorphism, one per layer, defined on the domain of the layer (second row) and then moved according to an affine motion. The third row shows the image generated by the model, which can therefore be though of as a deterministic generative model since it performs comparisons

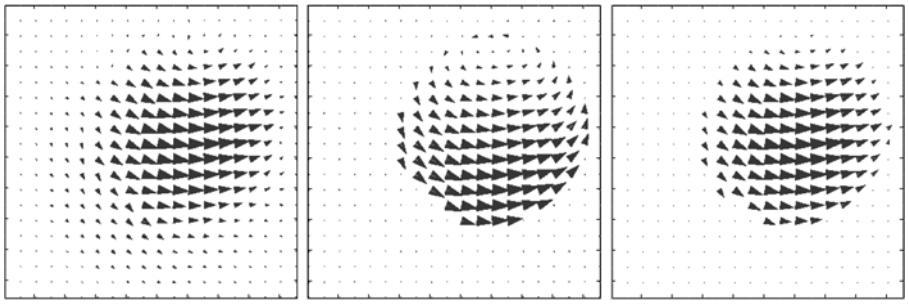

Fig. 5 Optical flow; ground truth; deformotion: Standard optical flow (*left*) imposes global regularization, which results in errors at the boundary (the vector field is more spread out than the model proposed, *on the right*). The ground truth is in the middle. The average angular errors for optical flow and deformotion are 11.49° and 6.31° respectively. The standard deviation for the angular errors are 1.37° and 1.44°. The parameters and regularization constants used were $dt = 0.2$, iterations $= 10000$, $\alpha = 10$ (data fidelity), $\mu = 0.5$ (smoothness of w), $\lambda = 5$ (smoothness of ρ)

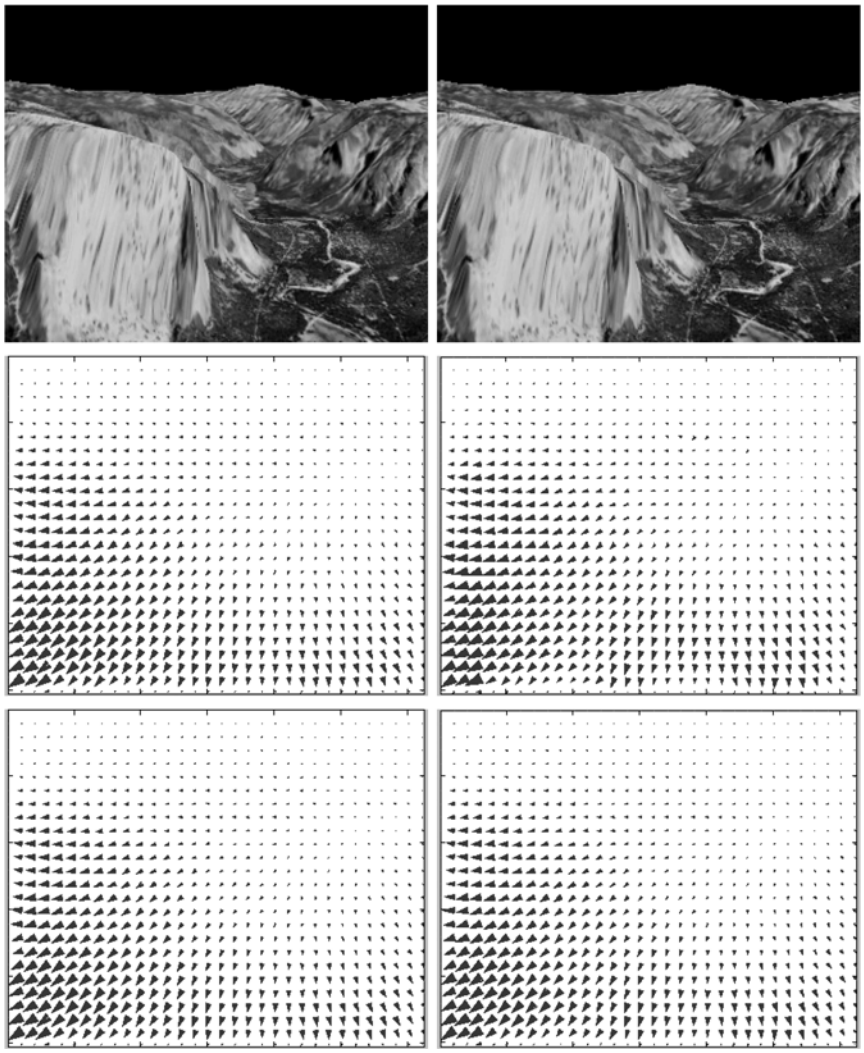

Fig. 6 Our model to optical flow: Optical flow (*left*) can be obtained from the general model (*right*) by allowing $\lambda \to 0$. Compare the results with parameters $dt = 0.028$, iterations $= 71000$, $\alpha = 20$, $\mu = 0.55$, $\lambda = 20$ on the *bottom row* with 200 on the *middle row*. Note that the two models (*left* and *right*) are closer on the bottom row. In comparison to the ground truth vector field, the vector field given by optical flow has an average angular error of 8.12. Our model with a smoothness weight of 200 gives an average angular error of 9.99. Reducing the smoothness weight to 20 gives an average angular error of 8.11 which is closer to the result of optical flow

at the image level, not via some intermediate representation. The corresponding images are displayed in the last row, with the layers superimposed for comparison.

The next set of experiments, using standard sequences used for optical flow analysis, is designed to illustrate the difference between our model and standard optical flow. A representative set of results of the motion field estimated by optical flow (left) and our model (right) is reported in Fig. 5. Our model does not rely on global regularization, but only regularization within each layer segmented in Fig. 4.

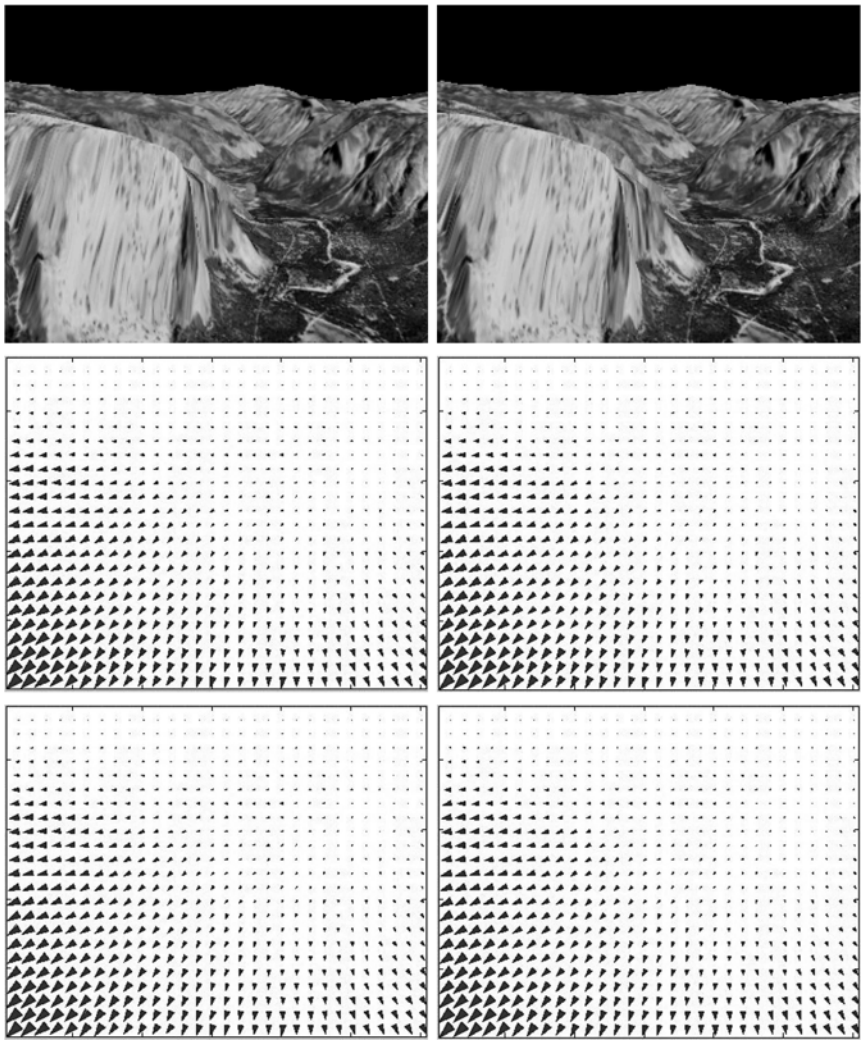

Fig. 7 L1 and L2 versions of our model and optical flow: L2 is on the *middle row* with our model first and optical flow second. The average angular errors are 4.83° and 4.74° respectively. L1 is on the *bottom row* with average angular errors of 4.80° and 4.67°. The parameters used are $dt = 0.1$, iterations $= 20000$, $\alpha = 1$, $\mu = 500$, $\lambda = 1$

Therefore, the boundaries of the motion field are better resolved.

Naturally, our model is a superset of those commonly used for optical flow computation. We illustrate this point by reducing the weight of the smoothness term for ρ in Fig. 6, which yields results closer to standard optical flow. In comparison to the ground truth vector field, the vector field given by optical flow has an average angular error of 8.12°. Our model, with a smoothness weight λ of 200, gives an average angular error of 9.99°. Reducing the smoothness weight λ to 20 gives an average angular error of 8.11° which is closer to the result of optical flow.

There can be some benefit in changing $r(w)$ from an L2 type norm to an L1 norm which has been done in (Papenberg et al. 2006) to improve optical flow. Instead of an L2 norm

$$r(w)(x) = \int \langle \nabla u(x), \nabla u(x) \rangle + \langle \nabla v(x), \nabla v(x) \rangle dx,$$

we use an L1 norm

$$r(w)(x) = \int \sqrt{\langle \nabla u(x), \nabla u(x) \rangle + \langle \nabla v(x), \nabla v(x) \rangle + \epsilon^2} dx$$

where ϵ is small, $\epsilon = .001$. We compare using the two norms combined with our model and then optical flow in Fig. 7.

A beneficial side-effect of having an explicit model of the scene, simple as it is (a regular irradiance pattern, with smoothness controlled by λ), is the possibility of comparing individual images to a (noiseless) model, rather than comparing noisy images to each other. The effects are visible in Fig. 8, where the flow field obtained with our model with L1 on artificially corrupted sequences is closer to the cleaner version of the sequence than using L1 standard optical flow.

The comparison with optical flow illustrates the necessity for partitioning the domain into independently moving objects. This is a motion segmentation task. Therefore, here we compare our model with more standard ones that partition the flow into affine segments, while still relying on the brightness constancy constraint and without an explicit

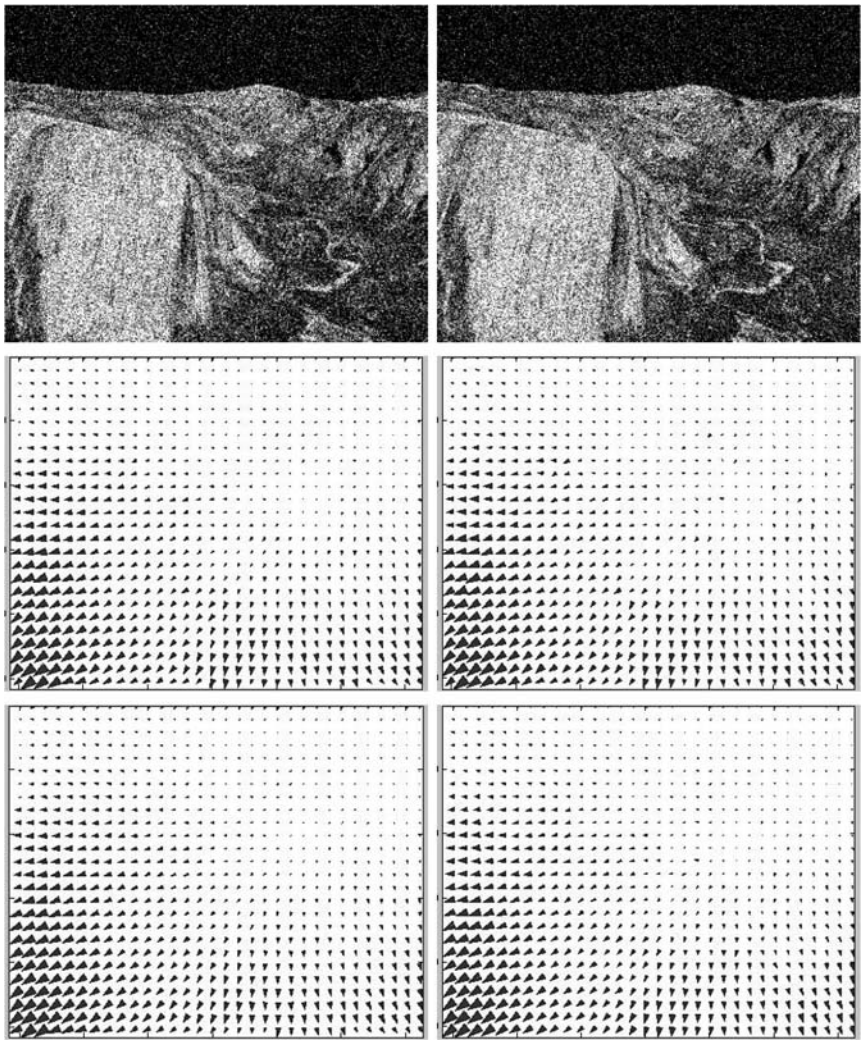

Fig. 8 Noisy Case of L1 vs. L2 versions of our model and optical flow: The images have been corrupted with Gaussian noise of zero mean and a variance of 0.05. L2 is on the *middle row* with our model first and optical flow second. The average angular errors are 11.93° and 13.23° respectively. L1 is on the *bottom row* with average angular errors of 11.26° and 11.88°. The L1 version of our model attains the best result with the angular error of 11.26°. The parameters used are $dt = 0.1$, iterations $= 20000$, $\alpha = 1$, $\mu = 5000$, $\lambda = 30$

model of the appearance of the scene. Such models can be obtained simply by increasing the regularization of the layer deformation (i.e. the entire layer moves with the same finite-dimensional motion: translational, Euclidean or affine). Figure 9 illustrates this effect.

Note that our model, by virtue of having an explicit representation of the appearance of each layer, can automatically fill in the appearance of underlying layers, as we illustrate in Fig. 10.

In Fig. 11 we illustrate inpainting using our model. In this example there is some camera motion, which makes it so the whiteboards in the two images are not quite lined up. Also there has been some corruption of the images which is modeled as the foreground layer that is moved around via an affine group. The whiteboard (background layer) is recovered with its own affine registration and the inpainted whiteboard is shown.

The conclusion we would like to draw from these experiments is that our model, being a superset of existing schemes (optical flow, motion segmentation, deformotion, inpainting), allows the user to apply existing algorithms simply by proper choice of constants. Naturally the price to pay for such flexibility and for the added power stemming from a richer model is computational complexity. However, all the experiments we have shown have been run on a Pentium M 2 GHz PC and takes five minutes per 1000 iterations.

6 Discussion

We have presented a generative model of the appearance (piecewise smooth albedo), motion (affine transformation) and deformation (diffeomorphism) of a sequence of images that exhibit occlusions. We have used this model as a basis for a variational optimization algorithm that simultaneously tracks the motion of a number of overlapping layers, estimates their deformation, and estimates the albedo of each layer, including portions that were partially occluded. Where no information is available, the layers are implicitly inpainted by their regularizers.

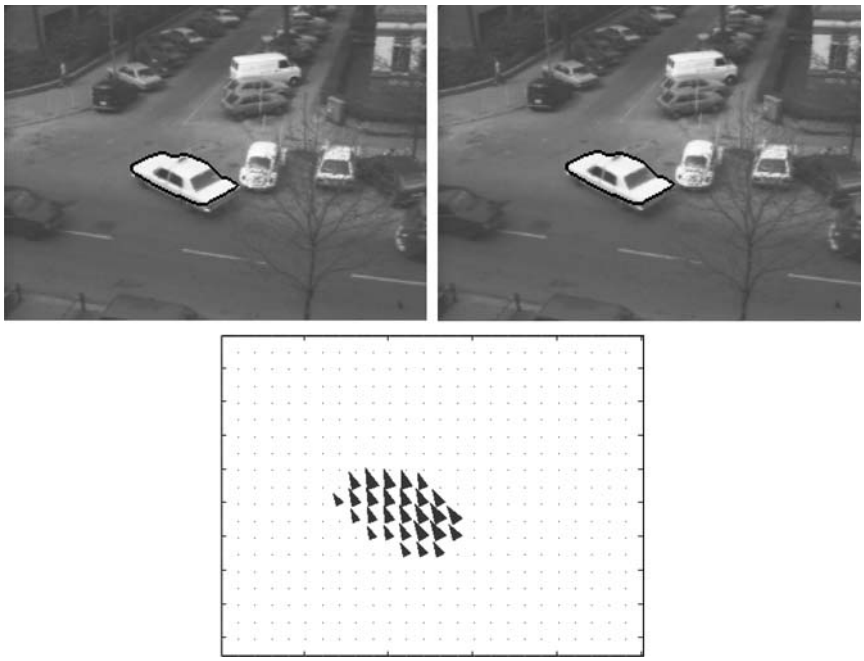

Fig. 9 The model proposed can be used to perform motion segmentation by increasing the regularization μ of the domain deformation for each layer (parameters used: $dt = \frac{0.2}{4000}$, iterations $= 20000$, $\alpha = 20$, $\mu = 4000$, $\lambda = 2000$)

Fig. 10 Our model yields "inpainted" layers. The *top row* shows the boundaries of layers, the *middle row* the reconstructed appearance of the layers (ρ) and the *bottom row* the warpings (w). Parameters used: $dt = 0.2$, iterations $= 2000$, $\alpha = 20$, $\mu = 1.0$, $\lambda = 0.8$, $\zeta = 3.0$ (arclength weight)

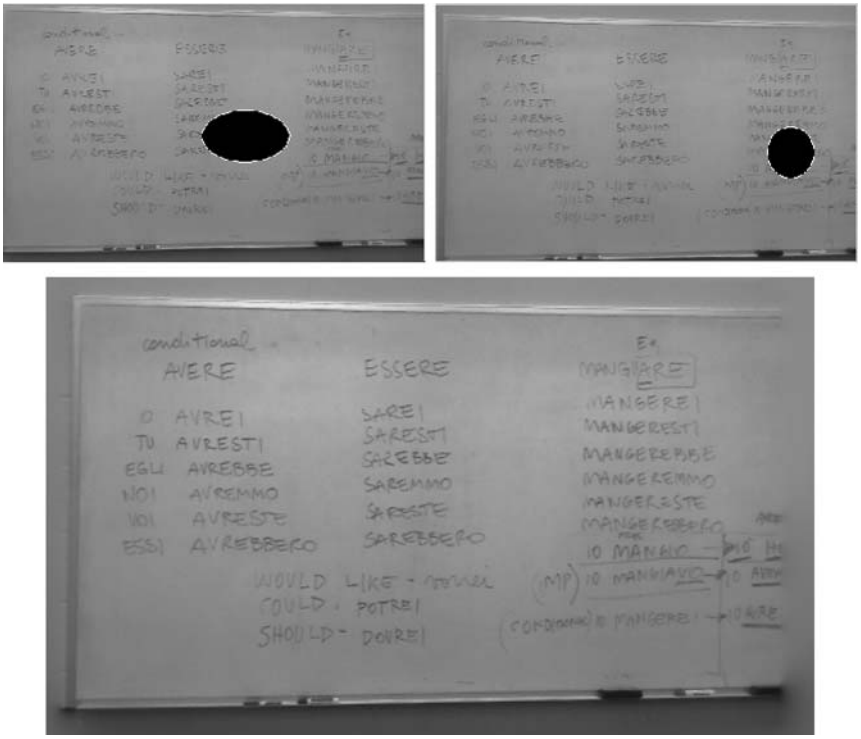

Fig. 11 Image Inpainting with our model. *First two images*: corrupted images of a teacher's whiteboard with some camera jitter, *Last image*: Image inpainting result

This model generalizes existing layer models to the case of deforming layers. Alternatively, one can think of our algorithm as a layered version of deformable tracking algorithms, or as a generalized version of optical flow or motion segmentation where multiple layers are allowed to occlude each other without disturbing the estimate of adjacent and occluded ones.

Our numerical implementation of the flow-based algorithm uses level set methods, and is realized without taking derivatives of the image, a feature that yields significant robustness when compared with boundary-based approaches to estimating optical flow. We have illustrated our approach on simple but representative sequences where existing methods fail to capture the phenomenology of the scene by either over-segmenting it, or by failing to capture its deformation while only matching its affine motion.

Acknowledgements This work was supported by NIH grants RO1-HL-68904/R01-NS-037747, NSF grant CCR-0133736, and an AFOSR MURI grant.

References

Alvarez, L., Weickert, J., & Sanchez, J. (1999). A scale-space approach to nonlocal optical flow calculations. In *ScaleSpace '99* (pp. 235–246).

Baker, S., Matthews, I., & Schneider, J. (2003). *Image coding with active appearance models* (Technical report). Carnegie Mellon University, The Robotics Institute.

Bertalmio, M., Sapiro, G., Caselles, V., & Ballester, C. (2000). Image inpainting. In K. Akeley (Ed.), *Siggraph 2000, computer graphics proceedings* (pp. 417–424). Longman: Addison-Wesley.

Blake, A., & Isard, M. (1998). *Active contours*. Berlin: Springer.

Caselles, V., Kimmel, R., & Sapiro, G. (1997). Geodesic active contours. *International Journal of Computer Vision*, *22*(1), 61–79.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (1998). Active appearance models. In *Proceedings of the European conference on computer vision* (pp. 484–496).

Cremers, D. (2003). Multiphase levelset framework for variational motion segmentation. In *International conference on scale-space theories in computer vision* (pp. 599–614), June 2003.

Deriche, R., Kornprobst, P., & Aubert, G. (1995). Optical flow estimation while preserving its discontinuities: a variational approach. In *Proceedings of ACCV* (Vol. 2, pp. 290–295).

Grenander, U. (1993). *General pattern theory*. Oxford: Oxford University Press.

Haussecker, H. W., & Fleet, D. J. (2001). Computing optical flow with physical models of brightness variation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 661–673.

Horn, B. K. P., & Schunk, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*, 185–203.

Hsu, S., Anandan, P., & Peleg, S. (1992). Accurate computation of optical flow by using layered motion representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1621–1626).

Kichenassamy, S., Kumar, A., Olver, P., Tannenbaum, A., & Yezzi, A. (1995). Gradient flows and geometric active contour models. In *ICCV '95: proceedings of the fifth international conference on computer vision* (pp. 810–815), Washington, DC, USA. New York: IEEE Computer Society.

Marks, T., Hershey, J., Roddey, J., & Movellan, J. (2005). Joint tracking of pose, expression, and texture using conditionally Gaussian filters. In *Advances in neural information processing systems* (Vol. 17). Cambridge: MIT Press.

Miller, M. I., & Younes, L. (1999). Group action, diffeomorphism and matching: a general framework. In *Proceedings of SCTV*.

Osher, S., & Sethian, J. (1988). Fronts propagating with curvature-dependent speed: algorithms based on Hamilton–Jacobi equations. *Journal of Computational Physics*, *79*, 12–49.

Papenberg, N., Bruhn, A., Brox, T., Didas, S., & Weickert, J. (2006). Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, *67*(2), 141–158.

Paragios, N., & Deriche, R. (2000). Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(3), 266–280.

Paragios, N., Rousson, M., & Ramesh, V. (2003). Non-rigid registration using distance functions. *Computer Vision and Image Understanding*, *89*(2–3), 142–165.

Schnörr, C. (1992). Computation of discontinuous optical flow by domain decomposition and shape optimization. *International Journal of Computer Vision*, *8*(2), 153–165.

Soatto, S., & Yezzi, A. (2002). Deformotion: deforming motion, shape average and the joint segmentation and registration of images. In *Proceedings of the European conference on computer vision (ECCV)* (Vol. 3, pp. 32–47).

Trouve, A., & Younes, L. (2005). Metamorphoses through lie group action. *Foundations of Computational Mathematics*, *5*(2), 173–198.

Wang, J., & Adelson, E. (1994). Representing moving images with layers. *IEEE Transactions on Image Processing*, *3*(5), 625–638.

Yezzi, A., & Soatto, S. (2001). Stereoscopic segmentation. In *Proceedings of the international conference on computer vision* (pp. 59–66).