Skin and Bones: Multi-layer, Locally Affine, Optical Flow and Regularization with Transparency

Shanon X. Ju* **Michael J. Black†** **Allan D. Jepson* ‡**

* Department of Computer Science, University of Toronto, Toronto, Ontario M5S 1A4 Canada
† Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304
‡ Canadian Institute for Advanced Research
{juxuan/jepson}@vis.toronto.edu, black@parc.xerox.com

Abstract

This paper describes a new method for estimating optical flow that strikes a balance between the flexibility of local dense computations and the robustness and accuracy of global parameterized flow models. An affine model of image motion is used within local image patches while a spatial smoothness constraint on the affine flow parameters of neighboring patches enforces continuity of the motion. We refer to this as a "Skin and Bones" model in which the affine patches can be thought of as rigid "bones" connected by a flexible "skin". Since local image patches may contain multiple motions we use a layered representation for the affine bones. To regularize this layered motion representation we develop a new framework for regularization with transparency.

1 Introduction

Recent work on optical flow can be seen as trying to find a balance between local dense optical flow schemes and global parameterized approaches [2, 9, 14]. Dense optical flow methods require only local image measurements and integrate information over larger areas via regularization. While these methods have the advantage of being able to cope with complex and varying flow fields and can be extended to model motion discontinuities in a relatively straightforward fashion [8], they remain somewhat inaccurate. Global parameterized approaches, on the other hand, assume that the optical flow within some image region (possibly the entire image) can be modeled by a low-order polynomial [4]. When the model is a good approximation to the image motion these methods are very accurate since one only has to estimate a small number of parameters given hundreds or thousands of constraints. The problem with these methods is that large image regions are typically not well modeled by a single parametric motion due to the complexity of the motion or the presence of multiple motions. Smaller regions on the other hand may not provide sufficient constraints for estimating the motion. This problem of choosing a region size has been referred to as the *generalized aperture problem* (GAP) [11]. The work described here combines features of both the regularized and parameterized methods to obtain nearly the accuracy of the parametrized motion approaches but with the generality and flexibility of the regularized approaches.

The approach tiles the image with a fixed set of rectangular patches and assumes that the motion within the regions can be represented by a small number of affine motions that can be thought of as "layers" [10, 16]. The approach assigns pixels to layers and estimates the motion of each layer using a robust mixture model formulation [2, 11, 13] that accounts for outliers which cannot be represented by any of the layers. The assignment to layers and the estimation of the motions is achieved using a variant of the EM algorithm [13].

Within image regions of fixed size the affine motion model may be underconstrained, and therefore we add a regularization term that embodies the assumption that the affine parameters of a patch should be similar to its neighbors' parameters. We refer to this formulation as "Skin and Bones" where the parameterized patches can be thought of as rigid pieces of bone that are connected by a flexible skin. Standard regularization techniques, however, cannot cope with this situation since there may be multiple affine motion estimates in each patch.

Consider a single patch with multiple motion estimates and its four nearest neighbors which may also have multiple affine motion estimates. Our approach "connects" every layer in the center patch with every layer in all the neighboring patches. To regularize a particular layer one considers all possible neighboring motions within a robust statistical framework. In such a framework, neighboring layers that have similar motions with have a strong influence on the solution while layers with dissimilar motions will be treated as outliers with little, or no, influence. We call this method *regularization with transparency.*

The following section reviews related work on layered motion estimation. Sections 3 and 4 introduce single-layer Bones and Skin respectively and show how the skin improves the motion estimates. The model is then extended to include multi-layer bones in Section 5 and transparent regularization in Section 6.

2 Related Work

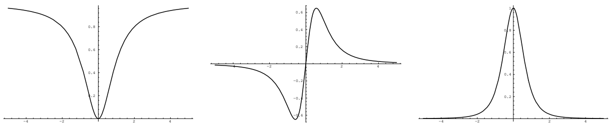
Parameterized optical flow methods assume that the spatial variation of the image motion within a region can be represented by a low-order polynomial (eg. affine motion). With many motion constraints and few parameters to estimate these approaches can recover accurate motion estimates when the motion model is a good approximation to the image motion. The problem with this approach is that parametric motion models applied over arbitrary image regions are rarely valid in real scenes due to surfaces at varying depths or the independent motion of objects.

Approaches have been devised which ameliorate some of the problems of "global" parametric models. One set of approaches estimates a fixed number of parametric motions within a given image region using a variety of regression techniques [5, 6, 11, 18]. Another set of approaches applies parametric models to coarse flow fields by grouping the flow vectors into consistent regions [1, 16]. Both sets of approaches can cope with a small number of motions within a region but not with general flow fields. They do not address how to select appropriate image regions in which to apply the parametric models nor how to select the appropriate number of motions or layers. These limitations can lead to inaccuracies and instabilities in the recovered motions.

A number of methods have addressed the problem of how to choose the appropriate number of parameterized motions that are necessary to represent the motion in the scene. One set of approaches [2, 10] uses a minimum description length encoding principle to strike a balance between accurate encoding of the motion and the number of layers needed to represent it. While these methods provide a segmentation of the image based on the support of pixels for each of the layers, they still operate over fixed image regions (typically the entire image).

There have been a number of recent attempts to find appropriate image regions within which to apply parameterized motion models. For example, Black and Jepson [9] first segment an image into regions using brightness information and then fit the motion within the regions using parameterized flow models. When a good segmentation is available, the motion can be estimated accurately but brightness information alone cannot be guaranteed to provide a good segmentation.

Szeliski and Shum [14] take an approach based on "quadtree splines" that treats the image as a set of patches of varying size which are connected in a spline-based representation that enforces smooth motion. The motion within a patch is determined by a parametrized motion model and the patch size varies based on how well the motion in a region can be approximated by a single flow model. The approach can only model a single motion within a patch which precludes the representation of transparent motion and fragmented occlusion. Additionally, the spline-based representation does not readily admit spatial discontinuities.



$\rho(x,\sigma)$ $\psi(x,\sigma)$ $l(x,\sigma)$

Figure 1: A robust error norm.

In contrast, we take fixed sized regions of the image and model multiple motions within each region using a layered motion estimation scheme [2, 6, 11, 18]. To model spatial smoothness we add a constraint on the affine parameters of neighboring patches. This is similar in spirit to the constraints used in oriented particle systems [15]. In our case we must extend standard regularization schemes to deal with the multi-layer data. Madarasmi *et al.* [12] approached a similar problem of regularization with multiple depth measurements at each point using a stochastic minimization framework. Our solution is deterministic and is a straightforward extension of the robust regularization scheme described by Black and Anandan [8].

3 Locally Affine Motion (Bones)

For a small image region, an affine (linear) transformation can well approximate the image motion of a smooth surface. This model is defined as

$$u(x,y) = a_0 + a_1(x - x_c) + a_2(y - y_c), \quad (1)$$
$$v(x,y) = a_3 + a_4(x - x_c) + a_5(y - y_c), \quad (2)$$

where $\mathbf{u}(\mathbf{x},\mathbf{a}) = [u(x,y), v(x,y)]^T$ are the horizontal and vertical components of the image velocity at the image point $\mathbf{x} = [x,y]^T$, and $\mathbf{a} = [a_0, a_1, a_2, a_3, a_4, a_5]^T$ denotes the vector of parameters to be estimated relative to some region center (x_c, y_c).

The assumption of brightness constancy for a given region and a particular flow model gives rise to the optical flow constraint equation

$$\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}(\mathbf{s})) + I_t = 0, \quad \forall \mathbf{x} \in \mathcal{R}(\mathbf{s}) \quad (3)$$

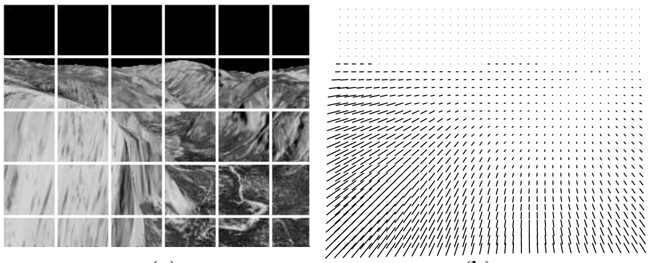
where $\mathbf{a}(\mathbf{s})$ denotes the affine model for region $\mathbf{s}$, $\mathcal{R}(\mathbf{s})$ denotes the points in region $\mathbf{s}$, I is the image brightness function and t represents time. $\nabla I = [I_x, I_y]$, and the subscripts indicates partial derivatives of image brightness with respect to the spatial dimensions and time at the point $\mathbf{x}$.

To estimate the parameters $\mathbf{a}(\mathbf{s})$, we minimize

$$E(\mathbf{s}) = \sum_{\mathbf{x} \in \mathcal{R}(\mathbf{s})} \rho(\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}(\mathbf{s})) + I_t,\ \sigma), \quad (4)$$

with respect to the affine parameters $\mathbf{a}(\mathbf{s})$. The value σ is a scale parameter and ρ is some robust error norm. For the examples in this paper, ρ is taken to be

$$\rho(x,\sigma) = x^2/(\sigma + x^2) \quad (5)$$



(a) (b)

Figure 2: Yosemite Sequence, ground truth. (a) Image 11 in the sequence; (b) flow field.

which is used in [6] and is shown in Figure 1. The shape of ρ is such that it "rejects", or down-weights, large residual errors. The function $\psi(x, \sigma)$, also shown in Figure 1, is the derivative of ρ and characterizes the influence of the residuals. As the magnitudes of residuals $\nabla I \cdot \mathbf{u}(\mathbf{a}) + I_t$ grow beyond a point their influence on the solution begins to decrease and the value of $\rho(\cdot)$ approaches a constant.

The value σ effects the point at which the influence of residuals begins to decrease. This down-weighting of residuals begins where the second derivative of ρ is zero; that is $\pm\sqrt{\sigma/3}$ for the norm used here. Following [6] we consider residual errors, $\nabla I \cdot \mathbf{u}(\mathbf{a}) + I_t$, to be *outliers* if their magnitude is greater than $\sqrt{\sigma}$ times $\sqrt{\sigma/3}$; that is, $\sigma/\sqrt{3}$.

To minimize Equation (4) we use a simple gradient descent scheme with a continuation method that begins with a high value of σ and lowers it gradually during the minimization until it reaches the desired value [8]. To cope with large motions a coarse-to-fine strategy is employed [6].

3.1 Bones Example

To illustrate the behavior of local affine "bones" we apply the method to two images in the synthetic Yosemite sequence[1], the first of which is shown in Figure 2(a). Figure 2(b) shows the known vector-field for the true motion. The image is segmented into fixed rectangular patches (51×48 pixels) and the affine motion of each patch is estimated independently. A four-level Gaussian pyramid was used in the coarse-to-fine processing. The value of σ began at 35 and was lowered by a factor of 0.95 at each iteration to a minimum of 10, and 30 iterations of gradient descent were used at each level. These parameters, except for patch size and levels, remain fixed for the experiments in the remainder of this section and the next.

The affine motions $\mathbf{a}(\mathbf{s})$, for each region $\mathbf{s}$, specify the motion of every pixel $\mathbf{x} \in \mathcal{R}(\mathbf{s})$ and we can use this computed affine motion to produce a dense flow field with a vector at every pixel as shown in Figure 3 (a).

Since the sequence is synthetic, we can compute the error in the flow using the angular error measure of Barron *et al.* [3]. The performance of the algorithm can be quantified as shown in Table 1 (Bones). "Average Error" refers to the mean angular error over the non-sky portion of the image.

By visual inspection, it is clear that the motion field in Figure 3(a) is not as smooth as the actual flow and shows a clear block structure. In some regions, most notably at the boundaries, the estimated motion is incorrect. The following section illustrates how a regularization term (skin) improves on these local affine estimates.

4 Regularization (Skin)

Regardless of the region size chosen for optical flow estimation, there is the possibility that the solution will be ill-conditioned due to the lack of sufficient brightness variation within the region. It is therefore useful to regularize the optical flow estimation problem by adding a spatial coherence constraint that favors solutions which are "smooth". Traditionally, this constraint is formulated to minimize the difference between neighboring optical flow vectors but, when the local flow estimation is performed by affine bones, we instead need to formulate a notion of spatial coherence between the parameters of neighboring affine patches.

We define the Skin & Bones model by adding a spatial coherence term to the to Equation (4)

$$E(\mathbf{s}) = \frac{1}{|\mathcal{R}(\mathbf{s})|}[\sum_{\mathbf{x}\in\mathcal{R}(\mathbf{s})} \rho(\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}(\mathbf{s})) + I_t,\ \sigma_D)] + \frac{\lambda}{|\mathcal{G}(\mathbf{s})|}[\sum_{\mathbf{t}\in\mathcal{G}(\mathbf{s})} \rho(||\mathbf{a}(\mathbf{s}) - \mathbf{a}^*(\mathbf{t})||, \sigma_S)] \quad (6)$$

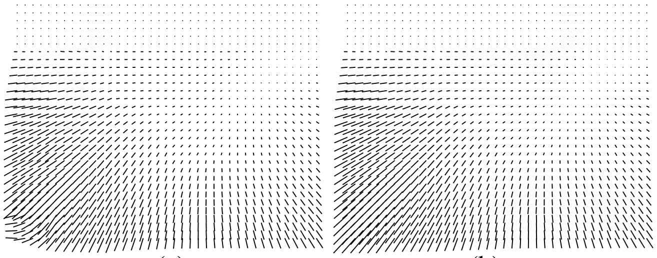
where $\mathbf{s}$ is an image region, λ controls the relative importance of the two terms, $\mathcal{R}(\mathbf{s})$ and $\mathbf{a}(\mathbf{s})$ are the pixels and the affine parameters of region $\mathbf{s}$ respectively, $\mathcal{G}(\mathbf{s})$ are the neighboring patches of $\mathbf{s}$, and some appropriate norm is defined on the neighboring affine parameters. The neighboring affine motion $\mathbf{a}(\mathbf{t})$ is dependent on the region center $(x_c(\mathbf{t}), y_c(\mathbf{t}))$ and to be compared with $\mathbf{a}(\mathbf{s})$ must be transformed as explained below. This transformed affine motion is $\mathbf{a}^*(\mathbf{t})$. The data and spatial terms of E are normalized with respect to the size of $\mathcal{R}(\mathbf{s})$ and $\mathcal{G}(\mathbf{s})$ respectively and each has its own scale parameter. The use of a robust error norm, ρ allows spatial discontinuities between neighboring affine patches.

To compare the affine parameters of neighboring patches, it is necessary to transform these parameters so that they are defined with respect to the center of the central patch, $\mathbf{s}$. If the center of patch $\mathbf{s}$ is $(x_c(\mathbf{s}), y_c(\mathbf{s}))$ and the center of a neighboring patch $\mathbf{t}$ is $(x_c(\mathbf{t}), y_c(\mathbf{t}))$ then a point $\mathbf{x}$ in region $\mathbf{t}$ can be described, with respect ot the center of $\mathbf{s}$, as $((x - x_c(\mathbf{s})) + (x_c(\mathbf{t}) - x_c(\mathbf{s})), (y - y_c(\mathbf{t})) + (y_c(\mathbf{t}) - y_c(\mathbf{s})))$. Substituting this into the affine motion equations (1) and (2) and simplifying gives shifted affine parameters, $\mathbf{a}^*$, of patch

[1] This sequence was generated by Lynn Quam and provided by David Heeger.

	Average Error	Standard Deviation	Percent of flow vectors with error less than:				
			$< 1^\circ$	$< 2^\circ$	$< 3^\circ$	$< 5^\circ$	$< 10^\circ$
Bones:	2.77°	3.4°	23.7%	49.9%	69.2%	89.1%	98.3%
Skin&Bones:	2.16°	2.0°	33.0%	61.3%	76.3%	91.6%	99.6%

Table 1: Error results for the Yosemite fly-through sequence.



(a) (b)

Figure 3: Yosemite flow results. (a) Local affine patches; (b) Affine patches with spatial coherence.

Technique	Average Error	Standard Deviation	Density
Anandan	15.84°	13.46°	100%
Singh	13.16°	12.07°	100%
Nagel	11.71°	10.59°	100%
Horn and Schunck (modified)	11.26°	16.41°	100%
Uras *et al.*	10.44°	15.00°	100%
Fleet and Jepson	4.29°	11.24°	34.1%
Lucas and Kanade	4.10°	9.58°	35.1%
Weber and Malik [17]	3.42°	5.35°	45.2%
Black and Anandan [8]$*$	4.47°	3.90°	100%
Black [7]$*$	3.52°	3.25°	100%
Black and Jepson [9]$*$	2.29°	2.25°	100%
Skin & Bones$*$	2.16°	2.0°	100%

Table 2: Comparison of various optical flow algorithms.

t as

$$\begin{aligned} a_0^* &= a_0 + a_1(x_c(\mathbf{t}) - x_c(\mathbf{s})) + a_2(y_c(\mathbf{t}) - y_c(\mathbf{s})), \\ a_3^* &= a_3 + a_4(x_c(\mathbf{t}) - x_c(\mathbf{s})) + a_5(y_c(\mathbf{t}) - y_c(\mathbf{s})), \end{aligned}$$

and $a_i^* = a_i$, $i \neq 0, 3$.

In practice we have found that minimizing the sum of the differences in the neighboring individual affine parameters works as well as minimizing the norm and is simpler to implement. The spatial term then becomes

$$\sum_{\mathbf{t} \in \mathcal{G}(\mathbf{s})} \sum_{i=0}^{5} \rho(a_i(\mathbf{s}) - a_i^*(\mathbf{t}), \sigma_i) \tag{7}$$

where $a_i(\mathbf{s})$ is the i^{th} affine parameter of patch **s** and the scale parameter may vary depending on the parameter.

We minimize this function using the same gradient descent scheme and continuation method described in the previous section and in [6, 8]. Unlike traditional parametric motion estimation schemes, the addition of the spatial coherence constraint on the affine parameters means that each step in the optimization takes into account both the optical flow constraints within the region and the parameters of the neighboring regions (cf. [14]). This results in more accurate motion estimates and a more stable optimization problem.

4.1 Example: Skin & Bones

To illustrate the effect of regularizing the affine parameters we add skin to the Yosemite sequence example from the previous section. The recovered optical flow using Equation (6) is shown in Figure 3 (b). Comparing the results to those in Figure 3 (a) reveals that the unstable results near the boundaries are gone and that the flow appears slightly smoother. Quantitatively, the addition of "skin" improves the average angular error by 22% as seen in Table 1 (Skin & Bones).

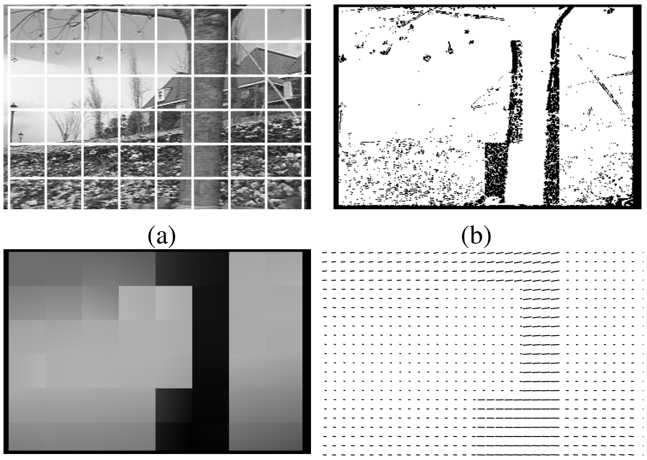
All parameter values were the same as those in the previous section and, for the new parameters, $\sigma_0 = \sigma_3$ started at 4.0 and were lowered to 0.2 by a factor of 0.88 per frame. The remaining σ_i were a factor of 100 smaller than this and λ was taken to be 0.05.

The results of the Skin & Bones approach are compared with other published results for the Yosemite sequence in Table 2 [3]. Methods followed by a "$*$" have errors computed without the sky region. In [3], when the sky is ignored, the accuracy of the other methods improves by approximately 25% which is still below the accuracy of the Skin & Bones model. The Skin & Bones model also provides a flow vector at every point (100% density).

In [9], Black and Jepson perform a similar parametrized fit, but do so in regions obtained by segmenting the brightness images. They allow deformations from the fitted motions using a robust regularization scheme in which the parameterized motion of the patches is treated as a prior. If we allow similar local deformations from the Skin & Bones fit, the average angular error decreases to 1.82° with as standard deviation of 1.58° and 100% density.

4.2 Limitations of Single-Layer Bones

The Skin & Bones model exploits the accuracy of area-based regression techniques locally and does so reliably through the use of a regularizing skin. When the affine flow model is a reasonable approximation for the motion in a region this results in very accurate motion estimates as were seen with the Yosemite sequence. In practice, flow fields



(a) (b) (c) (d)

Figure 4: (a) Image with patches shown; (b) Outliers (in black) where the estimated motion did not conform to the parameterized model; (c) Horizontal component of flow (darkness is proportional leftward velocity); (d) Flow field.

are rarely smoothly varying and typically contain discontinuities.

Consider the "flower garden" sequence shown in Figure 4 (a). The 43×40 pixel regions in the figure span surfaces at a number of depths. In this case the robust estimation scheme of the Skin & Bones model will tend to recover the dominant motion within a region. This can be seen in the horizontal flow estimates in Figure 4 (c) (there is very little vertical motion). Regions that span the boundary of the tree choose one of the two motions in the region and, where this occurs, pixels corresponding to the other motion are treated as outliers (Figure 4 (b)). If the patch size is increased sufficiently, the motion of the foreground may eventually be ignored completely.

5 Mixtures of Robust Bones

We deal with several motions within a single region using a straightforward extension of the mixture model approach described in [11] (cf. [2]). That is, for a given image region we model the flow using several affine layers. In addition, to accommodate data which cannot be accounted for by any of these layers, we include an outlier process. The data at any given pixel $\mathbf{x}$ is assigned to the i^{th} layer with an ownership weight $m_i(\mathbf{x}, \sigma)$. The estimation problem, then, involves recovering the affine parameters for each layer, say $\mathbf{a}_i$ for $1 \leq i \leq \mathcal{L}$, along with the appropriate layer assignment weights, $m_i(\mathbf{x}, \sigma)$ for $1 \leq i \leq \mathcal{L}+1$. Here we denote the outliers as layer $\mathcal{L}+1$.

The estimation process we use is a variant of the EM-algorithm, which is an iterative process involving two separate steps at each iteration. The first step involves the estimation of the ownership weights, while the second uses these ownership weights to solve for the affine parameters of each layer.

Ownership Weights. We use a soft assignment of data to layers based on the discrepancies between the data and each of the layers. In particular, the assignment weights are defined in terms of the robust error norm ρ, from which we derive the likelihood function

$$l(x, \sigma) = \frac{1}{2x}\frac{\partial}{\partial x}\rho(x, \sigma) = \frac{\psi(x, \sigma)}{2x} = \frac{\sigma}{(\sigma + x^2)^2}. \quad (8)$$

This is the same ρ-function as used earlier and its associated likelihood function is shown in Figure 1. For a given pixel, we consider the likelihood that the pixel $\mathbf{x}$ belongs to layer i in region $\mathbf{s}$ to be

$$\begin{aligned} l_i(\mathbf{x}, \sigma) &= l(\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_i(\mathbf{s})) + I_t, \sigma) \\ &= \frac{\sigma}{(\sigma + (\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_i(\mathbf{s})) + I_t)^2)^2}. \end{aligned} \quad (9)$$

As we see from Figure 1, data having a smaller error is considered to have a higher likelihood of belonging to layer i, and this likelihood decays to zero as the error increases.

We will also need the likelihood, $l_{\mathcal{L}+1}(\mathbf{x}, \sigma)$, that the data at a given pixel arises from the outlier process. Following [11] we take any data item to be equally likely to be produced from this outlier process. Moreover, the value of this likelihood is taken to be the weight given by ρ to the smallest possible outlying residual, namely

$$l_{\mathcal{L}+1}(\sigma) = \frac{\sigma}{(\sigma + (\sigma/\sqrt{3})^2)^2} = \frac{9}{\sigma(3+\sigma)^2}. \quad (10)$$

Finally, we set L to be the sum of the likelihoods for each layer, including the outliers; that is $L = \sum_{i=1}^{\mathcal{L}+1} l_i(\mathbf{x}, \sigma)$.

Given these likelihoods $l_i(x, \sigma)$, $1 \leq i \leq \mathcal{L}+1$, the ownership weights $m_i(\mathbf{x}, \sigma)$ are determined by rescaling the likelihoods so that the results sum to one. That is,

$$m_i(\mathbf{x}, \sigma) = l_i(\mathbf{x}, \sigma)/L, \quad (11)$$

for $1 \leq i \leq \mathcal{L}+1$. This rescaling is particularly useful in situations where the layers are close enough so that a data item has a significant likelihood of coming from two or more layers. In such a situation the reweighting can reduce or eliminate a bias towards the mean of nearby layers (see [13]).

Layer Parameters. Given the soft assignment of the data into the different layers by Equation (11), we solve for layer parameters, $\mathbf{a}_i$, using a reweighted least squares formulation

$$E(\mathbf{s}) = \sum_{\mathbf{x} \in \mathcal{R}(\mathbf{s})} \sum_{i=1}^{\mathcal{L}} m_i(\mathbf{x}, \sigma)(\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_i) + I_t)^2. \quad (12)$$

This formulation can be expected to be robust to outliers since the ownership for large errors will be small.

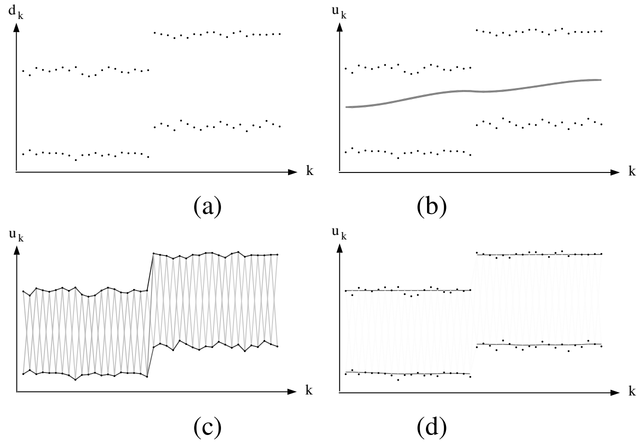


Figure 5: Transparent regularization. (a) Transparent data; (b) Single-layer regularization; (c) Weight of the connection between neighboring points in all layers; (d) Transparent regularization, piecewise smooth result.

New estimates for the affine parameters $\mathbf{a}_i$, $1 \leq i \leq \mathcal{L}$, obtained by minimizing E($\mathbf{s}$) are then used to re-estimate the ownership weights, and so on, as in the iterative EM-algorithm.

Unlike the approach presented in [11] we choose a likelihood function based on a robust error norm rather than the standard Gaussian component densities. Additionally, rather than attempt to estimate the probability, averaged over the image region, that a data item will belong to each of the layers, we simply take it to be equally likely.

6 Regularization with Transparency

The need to regularize noisy data arises in many computer vision and image processing problems. Here we will consider what happens when there are multiple measurements at a given point. To illustrate what this means we will consider a 1D example which extends a simple regularization problem to the transparent case.

Consider the noisy data in Figure 5(a). At each spatial position, k, there are multiple values, $d_{k,1}$ and $d_{k,2}$ which might, for example, be derived from depth measurements of two transparent surfaces. Fitting a single surface to this data using a least-squares formulation does not provide a useful solution as shown in 5(b).

Our goal is to regularize the measurements to derive two piecewise-smooth approximations $u_{k,1}$ and $u_{k,2}$ *without knowing* a priori *which measurements are grouped with which other measurements*. A given point $u_{k,1}$ has *two* neighbors to its left: $u_{k-1,1}$ and $u_{k-1,2}$. It is important to note that we do not know which, if either, of these measurements belongs to the same "surface" as $u_{k,1}$. If we knew the segmentation of the data points into surfaces, these surfaces could be regularized independently.

When the segmentation is not known a priori, we can still regularize by minimizing

$$E(\mathbf{u}, \mathbf{d}) = \sum_{k=0}^{K} \sum_{i=1}^{\mathcal{L}} [\rho(u_{k,i} - d_{k,i}, \sigma_D) + \sum_{j=1}^{\mathcal{L}} \rho(u_{k,i} - u_{k-1,j}, \sigma_S)] \quad (13)$$

with respect to each surface point $u_{k,i}$, where $\mathcal{L}$ is the number of layers. This means that we smooth a point with respect to *all* its neighbors in all surfaces. If any of these points are similar, they will be treated as inliers by the robust norm ρ and will have a strong influence on the solution. If they differ, they will be treated as outliers and will be *automatically* ignored. Minimizing Equation (13) smooths the data without explicitly assigning data to particular layers.

To illustrate this, Figure 5(c) shows the "weight" that the ρ-function gives to each neighbor. The dark lines indicate a strong connection between the surface points while the light lines indicate a weak connection. Note that we could threshold these values to derive a segmentation of the data into surfaces, but that there is no need to do this explicitly. As Equation (13) is minimized, the values of σ_i are gradually lowered, and outlying points receive lower and lower weight. Figure 5(d) shows the result of minimizing Equation (13) in this way. The solution converges to the desired piecewise-smooth, and transparent, surface interpretation.

6.1 Optical Flow

The transparent regularization theory can incorporated into the optical flow problem in a straightforward way to allow the regularization of multi-layer bones. We modify Equation (6) which combined single-layer affine motion estimates with standard robust regularization and define a new objective function, $E_i(\mathbf{s})$, for layer i of patch $\mathbf{s}$ as

$$E_i(\mathbf{s}) = \frac{1}{|\mathcal{R}(\mathbf{s})|}[\sum_{\mathbf{x} \in \mathcal{R}(\mathbf{s})} m_i(\mathbf{x}, \sigma_D)(\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_i) + I_t)^2] + \frac{\lambda}{|\mathcal{G}(\mathbf{s})|}[\sum_{\mathbf{t} \in \mathcal{G}(\mathbf{s})} \sum_{l \in \mathcal{L}(\mathbf{t})} \sum_{j=0}^{5} \rho(a_{i,j}(\mathbf{s}) - a^*_{l,j}(\mathbf{t}), \sigma_j)]. \quad (14)$$

The first term is simply a multi-layer motion model. The smoothness term considers each of the neighboring patches $\mathbf{t}$ and, for each of these patches, considers the layers $\mathcal{L}(\mathbf{t})$ present in that patch (here $a^*_{l,j}(\mathbf{t})$ refers to the j^{th} coefficient of the transformed affine parameters for layer l in the patch $\mathbf{t}$). For each of these neighbors and layers, the robust smoothness term is applied to the affine parameters. Motions that are similar will tend to reinforce each other while dissimilar motions will be ignored as outliers.

Equation (14) can be minimized in exactly the same way as all the previous objective functions considered so far.

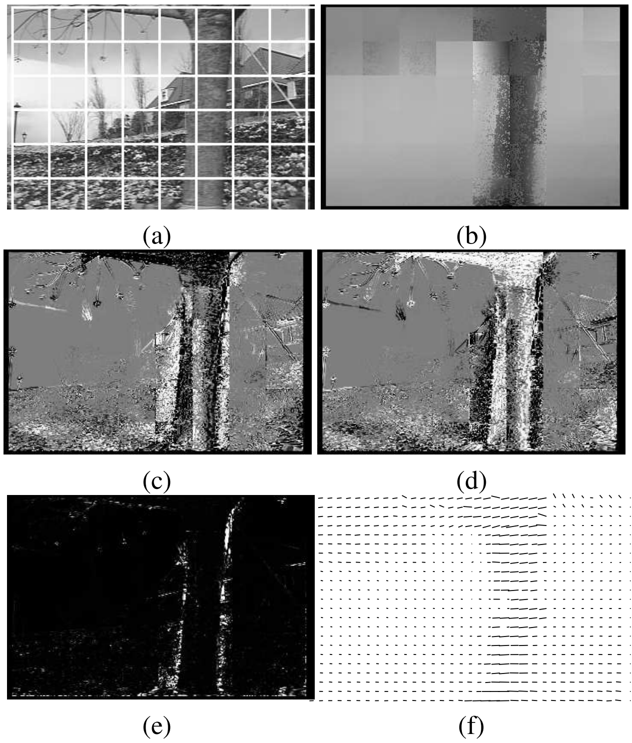


Figure 6: (a) Image with segmented regions shown; (b) Horizontal component of flow; (c) Weights for layer 1; (d) Weights for layer 2; (e) Weights for outlier layer; (f) Flow field.

This process alternates between solving for the $\mathbf{a}_i$ in each layer taking into account the smoothness term and solving for the weights $m_i(\mathbf{x}, \sigma_D)$. In our current implementation we assume that the number of layers is known and is taken to be two (plus outliers). If more motions are present, they will be treated as outliers. If fewer motions are present, which is quite likely, both layers converge to the same motion and the weights assigning pixels to layers become close to 0.5.

The motion parameters are estimated using a coarse-to-fine strategy in which the affine transformations are computed at a coarse level and then, at the next finer level, the estimated transformations are used to register the two images by warping one towards the other (note that this must be done for each of the layers). This process is repeated down to the finest level in the pyramid while the transformations are updated at each stage.

Experimental Results. Since the data term is different from that used in Section 4, some the parameters used for the multi-layer case differ from the single layer case. In particular, σ_D decreases from 85.0 to 15.0 by a factor of 0.9 at each stage in the continuation method and λ is taken to be 1.0. All other parameters remained the same.

Figure 6 revisits the flower garden sequence of Figure 4. In the single-layer case regions containing multiple motions chose only one of the motions. In the multi-layer case, regions are assumed to contain two motions. This can be seen in the horizontal motion at the boundary of the tree in Figure 6 (b). The regions boardering the tree have two clearly distinct motions which are smoothly connected to their neighbors.

Figure 6 (c) and (d) show the weights for the two motion layers within each region. Gray areas correspond to a weight of 0.5 where only one motion was present. Regions that span a motion boundary have two distinct sets of weights. One portion of the region has high weights (white areas in the figure) while the other has low weights within a particular layer. This pattern is reversed in the other layer. Figure 6 (d) shows those points that were not accounted for by either layer and were treated as outliers. These occur predominantly at the boundary between the tree and the background. A flow field (Figure 6 (e)) can be generated by taking the most likely motion at each pixel (given the weights $m_i(\mathbf{x}, \sigma_D)$).

Figure 7 shows multi-layer results for the SRI tree sequence. The weights indicate that the ground plane is treated as a single layer while the branches of the tree and the background are assigned to different layers when they both appear in the same region.

In evaluating these motion estimates it is important to keep in mind that this is not a "dense" method in the standard sense but rather a cross between the parametric and dense approaches. The flow for the SRI tree, for example, does not exhibit smoothness at the pixel level, but rather at the region level.

As mentioned in Section 4, the Skin & Bones method can provide an initial guess for, and a constraint on, a more traditional dense method. For example, the approach in [9] was applied to the multi-layer results to produce a dense flow field, the horizontal component of which is shown in Figure 8. This result is more accurate than that obtained by a dense method alone.

7 Conclusions

Estimating optical flow accurately involves pooling information over a large area. Parametric motion models do this well and can cope with multiple motions in certain cases but are not applicable globally. When applied locally, however, insufficient constraints may result in an unstable solution. We have shown how regularization can be extended to constrain these local affine flow parameters. Moreover, we have provided a general framework for regularization with transparency that extends regularization to cope with multiple local motion estimates. The methods have been tested on synthetic and natural images and provide accurate flow estimates common to parametric approaches, while maintaining the flexibility of regularization schemes.

Acknowledgements. We thank Eric Mjolsness for his sug-

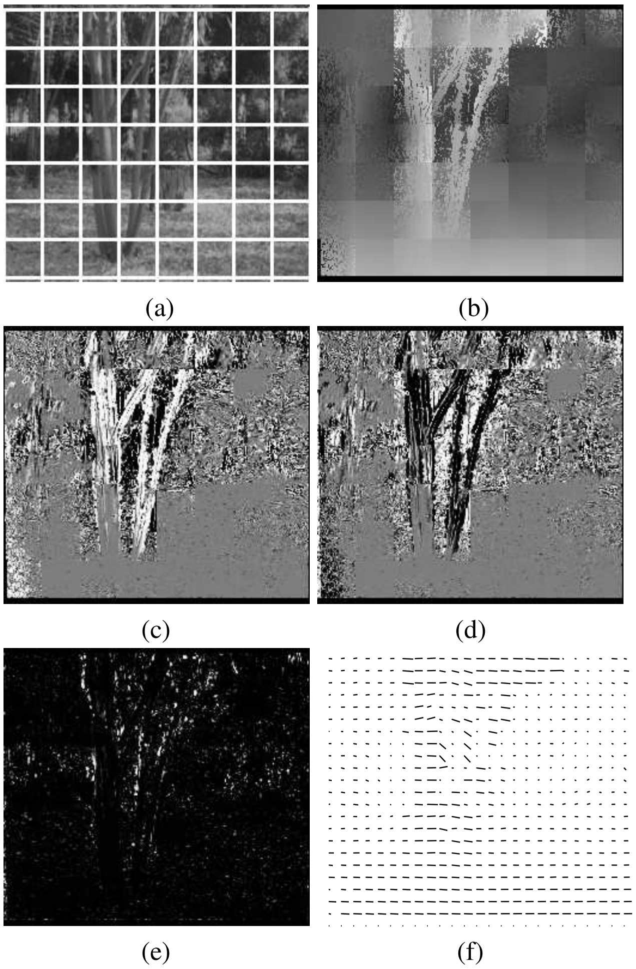


Figure 7: (a) Image with segmented regions shown; (b) Horizontal component of flow; (c) Weights for layer 1; (d) Weights for layer 2; (e) Weights for outlier layer; (f) Flow field.

gestions.

References

[1] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. PAMI-7(4):384–401, July 1985.

[2] S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. *ICCV'95*, pp. 777–784, Boston, MA, 1995.

[3] J.L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *IJCV*, 12(1), 1994.

[4] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. *ECCV'92*, pp. 237–252, Italy, May 1992.

[5] J. R. Bergen, P. J. Burt, R. Hingorani, and S. Peleg. Computing two motions from three frames. *ICCV'90*, pp. 27–30, Osaka, Japan, Dec. 1990.

Figure 8: Horizontal component of flow after combining the Skin & Bones method with a dense method.

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

[7] M. J. Black. Recursive non-linear estimation of discontinuous flow fields. *ECCV'94*, pp. 138–145, 1994.

[8] M. J. Black and P. Anandan. A framework for the robust estimation of optical flow. *ICCV'93*, pp. 231–236, Berlin, May 1993.

[9] M. J. Black and A. Jepson. Estimating multiple independent motions in segmented images using parametric models with local deformations. *Workshop on Motion of Non-rigid and Articulated Objects*, pp. 220–227, Austin, Nov. 1994.

[10] T. Darrell and A. Pentland. Robust estimation of a multi-layer motion representation. *Workshop on Visual Motion*, pp. 173–178, Princeton, Oct. 1991.

[11] A. Jepson and M. J. Black. Mixture models for optical flow computation. In I. Cox, P. Hansen, and B. Julesz, eds, *Partitioning Data Sets: With Applications to Psychology, Vision and Target Tracking*, pp. 271–286, AMS Pub., Providence, RI, Apr. 1993.

[12] S. Madarasmi, D. Kersten, and T. C. Pong. Multilayer surface segmentation using energy minimization. *CVPR'93*, pp. 774–775, New York, June 1993.

[13] G.J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker Inc., N.Y., 1988.

[14] R. Szeliski and H. Shum. Motion estimation with quadtree splines. *ICCV*, pp. 757–763, Boston, 1995.

[15] R. Szeliski and D. Tonnesen. Surface modeling with oriented particle systems. *Computer Graphics*, 26(2):185–194, July 1992.

[16] J. Y. A. Wang and E. H. Adelson. Representing Moving Images with Layers. *IEEE Trans. on Image Processing*, 3(5): 625–638, Sept. 1994.

[17] J. Weber and J. Malik. Robust computation of optical flow in a multi-scale differential framework. *ICCV'93*, pp. 12–20, Berlin, May 1993.

[18] A. Yuille, T. Yang, and D. Geiger. Robust statistics, transparency and correspondence. Tech. Rep. 90–7, Harvard Robotics Lab.