Learning Optical Flow

Deqing Sun[1], Stefan Roth[2], J.P. Lewis[3], and Michael J. Black[1]

[1] Department of Computer Science, Brown University, Providence, RI, USA
{dqsun,black}@cs.brown.edu
[2] Department of Computer Science, TU Darmstadt, Darmstadt, Germany
sroth@cs.tu-darmstadt.de
[3] Weta Digital Ltd., New Zealand
zilla@computer.org

Abstract. Assumptions of brightness constancy and spatial smoothness underlie most optical flow estimation methods. In contrast to standard heuristic formulations, we learn a statistical model of both brightness constancy error and the spatial properties of optical flow using image sequences with associated ground truth flow fields. The result is a complete probabilistic model of optical flow. Specifically, the ground truth enables us to model how the assumption of brightness constancy is violated in naturalistic sequences, resulting in a probabilistic model of "brightness inconstancy". We also generalize previous high-order constancy assumptions, such as gradient constancy, by modeling the constancy of responses to various linear filters in a high-order random field framework. These filters are free variables that can be learned from training data. Additionally we study the spatial structure of the optical flow and how motion boundaries are related to image intensity boundaries. Spatial smoothness is modeled using a Steerable Random Field, where spatial derivatives of the optical flow are steered by the image brightness structure. These models provide a statistical motivation for previous methods and enable the learning of all parameters from training data. All proposed models are quantitatively compared on the Middlebury flow dataset.

1 Introduction

We address the problem of learning models of optical flow from training data. Optical flow estimation has a long history and we argue that most methods have explored some variation of the same theme. Particularly, most techniques exploit two constraints: brightness constancy and spatial smoothness. The brightness constancy constraint (data term) is derived from the observation that surfaces usually persist over time and hence the intensity value of a small region remains the same despite its position change [1]. The spatial smoothness constraint (spatial term) comes from the observation that neighboring pixels generally belong to the same surface and so have nearly the same image motion. Despite the long history, there have been very few attempts to *learn* what these terms should be [2]. Recent advances [3] have made sufficiently realistic image sequences with ground truth optical flow available to finally make this practical. Here we revisit

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part III, LNCS 5304, pp. 83–97, 2008.

several classic and recent optical flow methods and show how training data and machine learning methods can be used to train these models. We then go beyond previous formulations to define new versions of both the data and spatial terms.

We make two primary contributions. First we exploit image intensity boundaries to improve the accuracy of optical flow near motion boundaries. The idea is based on that of Nagel and Enkelmann [4], who introduced oriented smoothness to prevent blurring of flow boundaries across image boundaries; this can be regarded as an anisotropic diffusion approach. Here we go a step further and use training data to analyze and model the statistical relationship between image and flow boundaries. Specifically we use a Steerable Random Field (SRF) [5] to model the conditional statistical relationship between the flow and the image sequence. Typically, the spatial smoothness of optical flow is expressed in terms of the image-axis-aligned partial derivatives of the flow field. Instead, we use the local image edge orientation to define a *steered* coordinate system for the flow derivatives and note that the flow derivatives along and across image boundaries are highly kurtotic. We then model the flow field using a Markov random field (MRF) and formulate the steered potentials using Gaussian scale mixtures (GSM) [6]. All parameters of the model are learned from examples thus providing a rigorous statistical formulation of the idea of Nagel and Enkelmann.

Our second key contribution is to learn a statistical model of the data term. Numerous authors have addressed problems with the common brightness constancy assumption. Brox *et al.* [7], for example, extend brightness constancy to high-order constancy, such as gradient and Hessian constancy in order to minimize the effects of illumination change. Additionally, Bruhn *et al.* [8] show that integrating constraints within a local neighborhood improves the accuracy of dense optical flow. We generalize these two ideas and model the data term as a general high-order random field that allows the principled integration of local information. In particular, we extend the Field-of-Experts formulation [2] to the spatio-temporal domain to model temporal changes in image features. The data term is formulated as the product of a number of experts, where each expert is a non-linear function (GSM) of a linear filter response. One can view previous methods as taking these filters to be fixed: Gaussians, first derivatives, second derivatives, etc. Rather than assuming known filters, our framework allows us to learn them from training data.

In summary, by using naturalistic training sequences with ground truth flow we are able to learn a complete model of optical flow that not only captures the spatial statistics of the flow field but also the statistics of brightness inconstancy and how the flow boundaries relate to the image intensity structure. The model combines and generalizes ideas from several previous methods and the resulting objective function is at once familiar and novel. We present a quantitative evaluation of the different methods using the Middlebury flow database [3] and find that the learned models outperform previous models, particularly at motion boundaries. Our analysis uses a single, simple, optimization method throughout to focus the comparison on the effects of different objective functions. The results

suggest the benefit of learning standard models and open the possibility to learn more sophisticated ones.

2 Previous Work

Horn and Schunck [9] introduced both the brightness constancy and the spatial smoothness constraints for optical flow estimation, however their quadratic formulation assumes Gaussian statistics and is not robust to outliers caused by reflection, occlusion, motion boundaries etc. Black and Anandan [1] introduced a robust estimation framework to deal with such outliers, but did not attempt to *model* the true statistics of brightness constancy errors and flow derivatives. Fermüller *et al.* [10] analyzed the effects of noise on the estimation of flow, but did not attempt to learn flow statistics from examples. Rather than assuming a model of brightness constancy we acknowledge that brightness can change and, instead, attempt to explicitly model the statistics of *brightness inconstancy*.

Many authors have extended the brightness constancy assumption, either by making it more physically plausible [11,12] or by linear or non-linear pre-filtering of the images [13]. The idea of assuming constancy of first or second image derivatives to provide some invariance to lighting changes dates back to the early 1980's with the Laplacian pyramid [14] and has recently gained renewed popularity [7]. Following a related idea, Bruhn *et al.* [8] replaced the pixelwise brightness constancy model with a spatially smoothed one. They found that a Gaussian-weighted spatial integration of brightness constraints results in significant improvements in flow accuracy. If filtering the image is a good idea, then we ask what filters should we choose? To address this question, we formulate the problem as one of learning the filters from training examples.

Most optical flow estimation methods encounter problems at motion boundaries where the assumption of spatial smoothness is violated. Observing that flow boundaries often coincide with image boundaries, Nagel and Enkelmann [4] introduced oriented smoothness to prevent blurring of optical flow across image boundaries. Alvarez *et al.* [15] modified the Nagel-Enkelmann approach so that less smoothing is performed close to image boundaries. The amount of smoothing along and across boundaries has been determined heuristically. Fleet *et al.* [16] learned a statistical model relating image edge orientation and amplitude to flow boundaries in the context of a patch-based motion discontinuity model. Black [17] proposed an MRF model that coupled edges in the flow field with edges in the brightness images. This model, however, was hand designed and tuned. We provide a probabilistic framework within which to learn the parameters of a model like that of Nagel and Enkelmann from examples.

Simoncelli *et al.* [18] formulated an early probabilistic model of optical flow and modeled the statistics of the deviation of the estimated flow from the true flow. Black *et al.* [19] learned parametric models for different classes of flow (e.g. edges and bars). More recently, Roth and Black [2] modeled the spatial structure of optical flow fields using a high-order MRF, called a Field of Experts (FoE), and learned the parameters from training data. They combined their learned prior

model with a standard data term [8] and found that the FoE model improved the accuracy of optical flow estimates. While their work provides a learned prior model of optical flow, it only models the spatial statistics of the optical flow and not the data term or the relationship between flow and image brightness.

Freeman *et al.* [20] also learned an MRF model of image motion but their training was restricted to simplified "blob world" scenes; here we use realistic scenes with more complex image and flow structure. Scharstein and Pal [21] learned a full model of stereo, formulated as a conditional random field (CRF), from training images with ground truth disparity. This model also combines spatial smoothness and brightness constancy in a learned model, but uses simple models of brightness constancy and spatially-modulated Potts models for spatial smoothness; these are likely inappropriate for optical flow.

3 Statistics of Optical Flow

3.1 Spatial Term

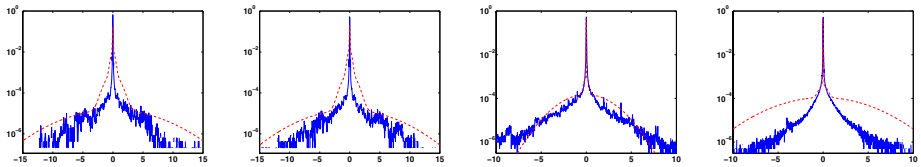
Roth and Black [2] studied the statistics of horizontal and vertical optical flow derivatives and found them to be heavy-tailed, which supports the intuition that optical flow fields are typically smooth, but have occasional motion discontinuities. Figure 1 (a, b (solid)) shows the marginal log-histograms of the horizontal and vertical derivatives of horizontal flow, computed from a set of 45 ground truth optical flow fields. These include four from the Middlebury "other" dataset, one from the "Yosemite" sequence, and ten of our own synthetic sequences. These synthetic sequences were generated in the same way as, and are similar to, the other Middlebury synthetic sequences (Urban and Grove); two examples are shown in Fig. 2. To generate additional training data the sequences were also flipped horizontally and vertically. The histograms are heavy-tailed with high peaks, as characterized by their high kurtosis ($\kappa = E[(x-\mu)^4]/E[(x-\mu)^2]^2$).

We go beyond previous work by also studying the steered derivatives of optical flow where the steering is obtained from the image brightness of the reference (first) frame. To obtain the steered derivatives, we first calculate the local image orientation in the reference frame using the structure tensor as described in [5]. Let $(\cos\theta(\mathbf{I}), \sin\theta(\mathbf{I}))^T$ and $(-\sin\theta(\mathbf{I}), \cos\theta(\mathbf{I}))^T$ be the eigenvectors of the structure tensor in the reference frame $\mathbf{I}$, which are respectively orthogonal to and aligned with the local image orientation. Then the orthogonal and aligned derivative operators $\partial_O^{\mathbf{I}}$ and $\partial_A^{\mathbf{I}}$ of the optical flow are given by

$$\partial_O^{\mathbf{I}} = \cos\theta(\mathbf{I}) \cdot \partial_x + \sin\theta(\mathbf{I}) \cdot \partial_y \quad \text{and} \quad \partial_A^{\mathbf{I}} = -\sin\theta(\mathbf{I}) \cdot \partial_x + \cos\theta(\mathbf{I}) \cdot \partial_y, \quad (1)$$

where ∂_x and ∂_y are the horizontal and vertical derivative operators. We approximate these using the 2×3 and 3×2 filters from [5].

Figure 1 (c, d) shows the marginal log-histograms of the steered derivatives of the horizontal flow (the vertical flow statistics are similar and are omitted here). The log-histogram of the derivative orthogonal to the local structure orientation has much broader tails than the aligned one, which confirms the intuition that large flow changes occur more frequently across the *image* edges.



(a) $\partial_x \mathbf{u}$, $\kappa = 420.5$ (b) $\partial_y \mathbf{u}$, $\kappa = 527.3$ (c) $\partial_{\mathcal{O}}^{\mathbf{I}} \mathbf{u}$, $\kappa = 340.1$ (d) $\partial_{\mathcal{A}}^{\mathbf{I}} \mathbf{u}$, $\kappa = 636.4$

Fig. 1. Marginal filter response statistics (log scale) of standard derivatives (left) and derivatives steered to local image structure (right) for the horizontal flow $\mathbf{u}$. The histograms are shown in solid blue; the learned experts in dashed red. κ denotes kurtosis.

These findings suggest that the steered marginal statistics provide a statistical motivation for the Nagel-Enkelmann method, which performs stronger smoothing along image edges and less orthogonal to image edges. Furthermore, the non-Gaussian nature of the histograms suggest that non-linear smoothing should be applied orthogonal to *and* aligned with the image edges.

3.2 Data Term

To our knowledge, there has been no formal study of the statistics of the brightness constancy error, mainly due to the lack of appropriate training data. Using ground truth optical flow fields we compute the brightness difference between pairs of training images by warping the second image in each pair toward the first using bi-linear interpolation. Figure 2 shows the marginal log-histogram of the brightness constancy error for the training set; this has heavier tails and a tighter peak than a Gaussian of the same mean and variance. The tight peak suggests that the value of a pixel in the first image is usually nearly the same as the corresponding value in the second image, while the heavy tails account for violations caused by reflection, occlusion, transparency, etc. This shows that modeling the brightness constancy error with a Gaussian, as has often been done, is inappropriate, and this also provides a statistical explanation for the robust data term used by Black and Anandan [1]. The Lorentzian used there has a similar shape as the empirical histogram in Fig. 2.

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

Fig. 2. (a) Statistics of the brightness constancy error: The log-histogram (solid blue) is fit with a GSM model (dashed red). (b)-(e) two reference (first) images and their associated flow fields from our synthetic training set.

We should also note that the shape of the error histogram will depend on the type of training images. For example, if the images have significant camera noise, this will lead to brightness changes even in the absence of any other effects. In such a case, the error histogram will have a more rounded peak depending on how much noise is present in the images. Future work should investigate adapting the data term to the statistical properties of individual sequences.

4 Modeling Optical Flow

We formulate optical flow estimation as a problem of probabilistic inference and decompose the posterior probability density of the flow field $(\mathbf{u}, \mathbf{v})$ given two successive input images $\mathbf{I}_1$ and $\mathbf{I}_2$ as

$$p(\mathbf{u}, \mathbf{v}|\mathbf{I}_1, \mathbf{I}_2; \Omega) \propto p(\mathbf{I}_2|\mathbf{u}, \mathbf{v}, \mathbf{I}_1; \Omega_{\mathrm{D}}) \cdot p(\mathbf{u}, \mathbf{v}|\mathbf{I}_1; \Omega_{\mathrm{S}}), \tag{2}$$

where Ω_{D} and Ω_{S} are parameters of the model. Here the first (data) term describes how the second image $\mathbf{I}_2$ is generated from the first image $\mathbf{I}_1$ and the flow field, while the second (spatial) term encodes our prior knowledge of the flow fields given the first (reference) image. Note that this decomposition of the posterior is slightly different from the typical one, e. g., in [18], in which the spatial term takes the form $p(\mathbf{u}, \mathbf{v}; \Omega_{\mathrm{S}})$. Standard approaches assume conditional independence between the flow field and the image structure, which is typically not made explicit. The advantage our formulation is that the conditional nature of the spatial term allows for more flexible methods of flow regularization.

4.1 Spatial Term

For simplicity we assume that horizontal and vertical flow fields are independent; Roth and Black [2] showed experimentally that this is a reasonable assumption. The spatial model thus becomes

$$p(\mathbf{u}, \mathbf{v}|\mathbf{I}_1; \Omega_{\mathrm{S}}) = p(\mathbf{u}|\mathbf{I}_1; \Omega_{\mathrm{Su}}) \cdot p(\mathbf{v}|\mathbf{I}_1; \Omega_{\mathrm{Sv}}). \tag{3}$$

To obtain our first model of spatial smoothness, we assume that the flow fields are independent of the reference image. Then the spatial term reduces to a classical optical flow prior, which can, for example, be modeled using a pairwise MRF:

$$p_{\mathrm{PW}}(\mathbf{u}; \Omega_{\mathrm{PWu}}) = \frac{1}{Z(\Omega_{\mathrm{PWu}})} \prod_{(i,j)} \phi(u_{i,j+1} - u_{ij}; \Omega_{\mathrm{PWu}}) \cdot \phi(u_{i+1,j} - u_{ij}; \Omega_{\mathrm{PWu}}), \tag{4}$$

where the difference between the flow at neighboring pixels approximates the horizontal and vertical image derivatives (see e. g., [1]). $Z(\Omega_{\mathrm{PWu}})$ here is the partition function that ensures normalization. Note that although such an MRF model is based on products of very local potential functions, it provides a global probabilistic model of the flow. Various parametric forms have been used to model the potential function ϕ (or its negative log): Horn and Schunck [9] used

Gaussians, the Lorentzian robust error function was used by Black and Anandan [1], and Bruhn *et al.* [8] assumed the Charbonnier error function. In this paper, we use the more expressive Gaussian scale mixture (GSM) model [6], i. e.,

$$\phi(x;\Omega) = \sum_{l=1}^{L} \omega_l \cdot \mathcal{N}(x;0,\sigma^2/s_l), \tag{5}$$

in which $\Omega = \{\omega_l | l = 1,\ldots,L\}$ are the weights of the GSM model, s_l are the scales of the mixture components, and σ^2 is a global variance parameter. GSMs can model a wide range of distributions ranging from Gaussians to heavy-tailed ones. Here, the scales and σ^2 are chosen so that the empirical marginals of the flow derivatives can be represented well with such a GSM model and are not trained along with the mixture weights ω_l.

The particular decomposition of the posterior used here (2) allows us to model the spatial term for the flow conditioned on the measured image. For example, we can capture the oriented smoothness of the flow fields and generalize the Steerable Random Field model [5] to a steerable model of optical flow, resulting in our second model of spatial smoothness:

$$p_{\mathrm{SRF}}(\mathbf{u}|\mathbf{I}_1;\Omega_{SRFu}) \propto \prod_{(i,j)} \phi\left((\partial_O^{\mathbf{I}_1} u)_{ij};\Omega_{SRFu}\right) \cdot \phi\left((\partial_A^{\mathbf{I}_1} u)_{ij};\Omega_{SRFu}\right). \tag{6}$$

The steered derivatives (orthogonal and aligned) are defined as in (1); the superscript denotes that steering is determined by the reference frame $\mathbf{I}_1$. The potential functions are again modeled using GSMs.

4.2 Data Term

Models of the optical flow data term typically embody the brightness constancy assumption, or more specifically model the deviations from brightness constancy. Assuming independence of the brightness error at the pixel sites, we can define a standard data term as

$$p_{\mathrm{BC}}(\mathbf{I}_2|\mathbf{u},\mathbf{v},\mathbf{I}_1;\Omega_{\mathrm{BC}}) \propto \prod_{(i,j)} \phi(I_1(i,j) - I_2(i+u_{ij}, j+v_{ij});\Omega_{\mathrm{BC}}). \tag{7}$$

As with the spatial term, various functional forms (Gaussian, robust, etc.) have been assumed for the potential ϕ or its negative log. We again employ a GSM representation for the potential, where the scales and global variance are determined empirically before training the model (mixture weights).

Brox *et al.* [7] extend the brightness constancy assumption to include high-order constancy assumptions, such as gradient constancy, which may improve accuracy in the presence of changing scene illumination or shadows. We propose a further generalization of these constancy assumptions and model the constancy of responses to several general linear filters:

$$p_{\mathrm{FC}}(\mathbf{I}_2|\mathbf{u},\mathbf{v},\mathbf{I}_1;\Omega_{\mathrm{FC}}) \propto \prod_{(i,j)} \prod_{k} \phi_k\{(J_{k1} * I_1)(i,j) - (J_{k2} * I_2)(i+u_{ij}, j+v_{ij});\Omega_{\mathrm{FC}}\}, \tag{8}$$

where the J_{k1} and J_{k2} are linear filters. Practically, this equation implies that the second image is first filtered with J_{k2}, after which the filter responses are warped toward the first filtered image using the flow $(\mathbf{u}, \mathbf{v})$ [1]. Note that this data term is a generalization of the Fields-of-Experts model (FoE), which has been used to model prior distributions of images [22] and optical flow [2]. Here, we generalize it to a spatio-temporal model that describes brightness (in)constancy.

If we choose J_{11} to be the identity filter and define $J_{12} = J_{11}$, this implements brightness constancy. Choosing the J_{k1} to be derivative filters and setting $J_{k2} = J_{k1}$ allows us to model gradient constancy. Thus this model generalizes the approach by Brox *et al.* [7] [2]. If we choose J_{k1} to be a Gaussian smoothing filter and define $J_{k2} = J_{k1}$, we essentially perform pre-filtering as, for example, suggested by Bruhn *et al.* [8]. Even if we assume fixed filters using a combination of the above, our probabilistic formulation still allows learning the parameters of the GSM experts from data as outlined below. Consequently, we do not need to tune the trade-off weights between the brightness and gradient constancy terms by hand as in [7]. Beyond this, the appeal of using a model related to the FoE is that we do not have to fix the filters ahead of time, but instead we can learn these filters alongside the potential functions.

4.3 Learning

Our formulation enables us to train the data term and the spatial term separately, which simplifies learning. Note though, that it is also possible to turn the model into a conditional random field (CRF) and employ conditional likelihood maximization (cf. [23]); we leave this for future work. To train the pairwise spatial term $p_{\mathrm{PW}}(\mathbf{u}; \Omega_{\mathrm{PWu}})$, we can estimate the weights of the GSM model by either simply fitting the potentials to the empirical marginals using expectation maximization, or by using a more rigorous learning procedure, such as maximum likelihood (ML). To find the ML parameter estimate we aim to maximize the log-likelihood $\mathcal{L}_{\mathrm{PW}}(\mathcal{U}; \Omega_{\mathrm{PWu}})$ of the horizontal flow components $\mathcal{U} = \{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(t)}\}$ of the training sequences w. r. t. the model parameters Ω_{PWu} (i. e., GSM mixture weights). Analogously, we maximize the log-likelihood of the vertical components $\mathcal{V} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(t)}\}$ w. r. t. Ω_{PWv}. Because ML estimation in loopy graphs is generally intractable, we approximate the learning objective and use the contrastive divergence (CD) algorithm [24] to learn the parameters.

To train the steerable flow model $p_{\mathrm{SRF}}(\mathbf{u}|\mathbf{I}_1; \Omega_{\mathrm{SRF}})$ we aim to maximize the conditional log-likelihoods $\mathcal{L}_{\mathrm{SRF}}(\mathcal{U}|\mathcal{I}_1; \Omega_{\mathrm{SRFu}})$ and $\mathcal{L}_{\mathrm{SRF}}(\mathcal{V}|\mathcal{I}_1; \Omega_{\mathrm{SRFv}})$ of the

[1] It is, in principle, also possible to formulate a similar model that warps the image first and then applies filters to the warped image. We did not pursue this option, as it would require the application of the filters at each iteration of the flow estimation procedure. Filtering before warping ensures that we only have to filter the image once before flow estimation.

[2] Formally, there is a minor difference: [7] penalizes changes in the gradient magnitude, while the proposed model penalizes changes of the flow derivatives. These are, however, equivalent in the case of Gaussian potentials.

training flow fields given the first (reference) images $\mathcal{I}_1 = \{\mathbf{I}_1^{(1)}, \ldots, \mathbf{I}_1^{(t)}\}$ from the training image pairs w. r. t. the model parameters Ω_{SRFu} and Ω_{SRFv}.

To train the simple data term $p_D(\mathbf{I}_2|\mathbf{u}, \mathbf{v}, \mathbf{I}_1; \Omega_D)$ modeling brightness constancy, we can simply fit the marginals of the brightness violations using expectation maximization. This is possible, because the model assumes independence of the brightness error at the pixel sites. For the proposed generalized data term $p_{\text{FC}}(\mathbf{I}_2|\mathbf{u}, \mathbf{v}, \mathbf{I}_1; \Omega_{\text{FC}})$ that models filter response constancy, a more complex training procedure is necessary, since the filter responses are not independent. Ideally, we would maximize the conditional likelihood $\mathcal{L}_{\text{FC}}(\mathcal{I}_2|\mathcal{U}, \mathcal{V}, \mathcal{I}_1; \Omega_{\text{FC}})$ of the training set of the second images $\mathcal{I}_2 = \{\mathbf{I}_2^{(1)}, \ldots, \mathbf{I}_2^{(t)}\}$ given the training flow fields and the first images. Due to the intractability of ML estimation in these models, we use a conditional version of contrastive divergence (see e. g., [5,23]) to learn both the mixture weights of the GSM potentials as well as the filters.

5 Optical Flow Estimation

Given two input images, we estimate the optical flow between them by maximizing the posterior from (2). Equivalently, we minimize its negative log

$$E(\mathbf{u}, \mathbf{v}) = E_{\text{D}}(\mathbf{u}, \mathbf{v}) + \lambda E_{\text{S}}(\mathbf{u}, \mathbf{v}), \tag{9}$$

where E_{D} is the negative log (i. e., energy) of the data term, E_{S} is the negative log of the spatial term (the normalization constant is omitted in either case), and λ is an optional trade-off weight (or regularization parameter).

Optimizing such energies is generally difficult, because of their non-convexity and many local optima. The non-convexity in our approach stems from the fact that the learned potentials are non-convex and from the warping-based data term used here and in other competitive methods [7]. To limit the influence of spurious local optima, we construct a series of energy functions

$$E_{\text{C}}(\mathbf{u}, \mathbf{v}, \alpha) = \alpha E_{\text{Q}}(\mathbf{u}, \mathbf{v}) + (1 - \alpha) E(\mathbf{u}, \mathbf{v}), \tag{10}$$

where E_{Q} is a quadratic, convex, formulation of E that replaces the potential functions of E by a quadratic form and uses a different λ. Note that E_{Q} amounts to a Gaussian MRF formulation. $\alpha \in [0, 1]$ is a control parameter that varies the convexity of the compound objective. As α changes from 1 to 0, the combined energy function in (10) changes from the quadratic formulation to the proposed non-convex one (cf. [25]). During the process, the solution at a previous convexification stage serves as the starting point for the current stage. In practice, we find using three stages produces reasonable results.

At each stage, we perform a simple local minimization of the energy. At a local minimum, it holds that

$$\nabla_{\mathbf{u}} E_{\text{C}}(\mathbf{u}, \mathbf{v}, \alpha) = 0, \text{ and } \nabla_{\mathbf{v}} E_{\text{C}}(\mathbf{u}, \mathbf{v}, \alpha) = 0. \tag{11}$$

Since the energy induced by the proposed MRF formulation is spatially discrete, it is relatively straightforward to derive the gradient expressions. Setting these

to zero and linearizing them, we rearrange the results into a system of linear equations, which can be solved by a standard technique. The main difficulty in deriving the linearized gradient expressions is the linearization of the warping step. For this we follow the approach of Brox *et al.* [7] while using the derivative filters proposed in [8].

To estimate flow fields with large displacements, we adopt an incremental multi-resolution technique (e. g., [1,8]). As is quite standard, the optical flow estimated at a coarser level is used to warp the second image toward the first at the next finer level and the flow increment is calculated between the first image and the warped second image. The final result combines all the flow increments. At the first stage where $\alpha = 1$, we use a 4-level pyramid with a downsampling factor of 0.5. At other stages, we only use a 2-level pyramid with a downsampling factor of 0.8 to make full use of the solution at the previous convexification stage.

6 Experiments and Results

6.1 Learned Models

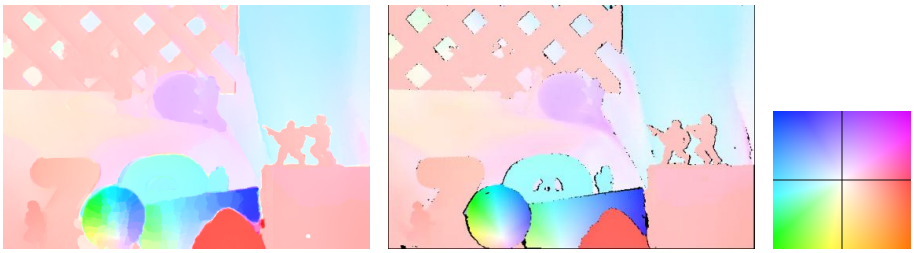
The spatial terms of both the pairwise model (**PW**) and the steerable model (**SRF**) were trained using contrastive divergence on $20,000$ 9×9 flow patches that were randomly cropped from the training flow fields (see above). To train the steerable model, we also supplied the corresponding $20,000$ image patches (of size 15×15 to allow computing the structure tensor) from the reference images. The pairwise model used 5 GSM scales; and the steerable model 4 scales.

The simple brightness constancy data term (**BC**) was trained using expectation-maximization. To train the data term that models the generalized filter response constancy (**FC**), the CD algorithm was run on $20,000$ 15×15 flow patches and corresponding 25×25 image patches, which were randomly cropped from the training data. 6-scale GSM models were used for both data terms. We investigated two different filter constancy models. The first (**FFC**) used 3 fixed 3×3 filters: a small variance Gaussian ($\sigma = 0.4$), and horizontal and vertical derivative filters similar to [7]. The other (**LFC**) used 6 3×3 filter pairs that were learned automatically. Note that the GSM potentials were learned in either case. Figure 3 shows the fixed filters from the FFC model, as well as two of the learned filters from the LFC model. Interestingly, the learned filters do not look like ordinary derivative filters nor do they resemble the filters learned in an FoE model of natural images [22]. It is also noteworthy that even though the J_{k2} are not enforced to be equal to the J_{k1} during learning, they typically exhibit only subtle differences as Fig. 3 shows.

Given the non-convex nature of the learning objective, contrastive divergence is prone to finding local optima, which means that the learned filters are likely not optimal. Repeated initializations produced different-looking filters, which however performed similarly to the ones shown here. The fact that these "non-standard" filters perform better (see below) than standard ones suggests that more research on better filters for formulating optical flow data terms is warranted.

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

Fig. 3. Three fixed filters from the FFC model: (a) Gaussian, (b) horizontal derivative, and (c) vertical derivative. (d,e) Two of the six learned filter pairs of the LFC model and the difference between each pair (left: J_{k1}, middle: J_{k2}, right: $J_{k1} - J_{k2}$).



(a) Estimated flow (b) Ground truth (c) Key

Fig. 4. Results of the SRF-LFC model for the "Army" sequence

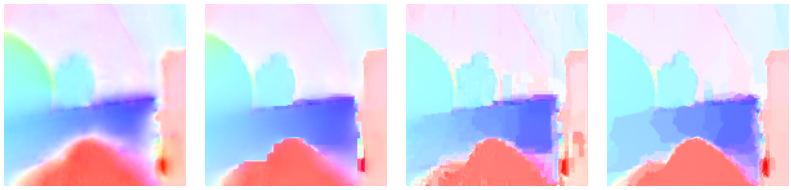
For the models for which we employed contrastive divergence, we used a hybrid Monte Carlo sampler with 30 leaps, $l = 1$ CD step, and a learning rate of 0.01 as proposed by [5]. The CD algorithm was run for 2000 to 10000 iterations, depending on the complexity of the model, after which the model parameters did not change significantly. Figure 1 shows the learned potential functions alongside the empirical marginals. We should note that learned potentials and marginals generally differ. This has, for example, been noted by Zhu *et al.* [26], and is particularly the case for the SRFs, since the derivative responses are not independent within a flow field (cf. [5]).

To estimate the flow, we proceeded as described in Section 5 and performed 3 iterations of the incremental estimation at each level of the pyramid. The regularization parameter λ was optimized for each method using a small set of training sequences. For this stage we added a small amount of noise to the synthetic training sequences, which led to larger λ values and increased robustness to novel test data.

6.2 Flow Estimation Results

We evaluated all 6 proposed models using the test portion of the Middlebury optical flow benchmark [3][3]. Figure 4 shows the results on one of the sequences along with the ground truth flow. Table 1 gives the average angular error (AAE)

[3] Note that the Yosemite frames used for testing as part of the benchmark are not the same as those used for learning.



(a) HS [9] (b) BA [1] (c) PW-BC (d) SRF-BC

(e) PW-FFC (f) SRF-FFC (g) PW-LFC (h) SRF-LFC

Fig. 5. Details of the flow results for the "Army" sequence. HS=Horn & Schunck; BA=Black & Anandan; PW=pairwise; SRF=steered model; BC=brightness constancy; FFC=fixed filter response constancy; LFC=learned filter response constancy.

Table 1. Average angular error (AAE) on the Middlebury optical flow benchmark for various combinations of the proposed models

	Rank	**Average**	Army	Mequon	Schefflera	Wooden	Grove	Urban	Yosemite	Teddy
HS[9]	16.4	8.72	8.01	9.13	14.20	12.40	4.64	8.21	4.01	9.16
BA[1]	9.8	7.36	7.17	8.30	13.10	10.60	4.06	6.37	2.79	6.47
PW-BC	13.6	8.36	8.01	10.70	14.50	8.93	4.35	7.00	3.91	9.51
SRF-BC	10.0	7.49	6.39	10.40	14.00	8.06	4.10	6.19	3.61	7.19
PW-FFC	12.6	6.91	4.60	4.63	9.96	9.93	5.15	7.84	3.51	9.66
SRF-FFC	9.3	6.26	4.36	5.46	9.63	9.13	4.17	7.11	2.75	7.43
PW-LFC	10.9	6.06	4.61	3.92	7.56	7.77	4.76	7.50	3.90	8.43
SRF-LFC	8.6	5.81	4.26	4.81	7.87	8.02	4.24	6.57	2.71	8.02

of the models on the test sequences, as well as the results of two standard methods [1,9]. Note that the standard objectives from [1,9] were optimized using exactly the same optimization strategy as used for the learned models. This ensures fair comparison and focuses the evaluation on the model rather than the optimization method. The table also shows the average rank from the Middlebury flow benchmark, as well as the average AAE across all 8 test sequences. Table 2 shows results of the same experiments, but here the AAE is only measured near motion boundaries. From these results we can see that the steerable flow model (SRF) substantially outperforms a standard pairwise spatial term (PW), particularly also near motion discontinuities. This holds no matter what data term the respective spatial term is combined with. This can also be seen visually in Fig. 5, where the SRF results exhibit the clearest motion boundaries.

Among the different data terms, the filter response constancy models (FFC & LFC) very clearly outperform the classical brightness constancy model (BC), particularly on the sequences with real images ("Army" through "Schefflera"), which are especially difficult for standard techniques, because the classical

Table 2. Average angular error (AAE) in motion boundary regions

	Average	Army	Mequon	Schefflera	Wooden	Grove	Urban	Yosemite	Teddy
PW-BC	16.68	14.70	20.70	24.30	26.90	5.40	20.70	5.26	15.50
SRF-BC	15.71	13.40	20.30	23.30	26.10	5.07	19.00	4.64	13.90
PW-FFC	16.36	12.90	17.30	20.60	27.80	6.43	24.00	5.05	16.80
SRF-FFC	15.45	12.10	17.40	20.20	27.00	5.16	22.30	4.24	15.20
PW-LFC	15.67	12.80	16.00	18.30	27.30	6.09	22.80	5.40	16.70
SRF-LFC	15.09	11.90	16.10	18.50	27.00	5.33	21.50	4.30	16.10

brightness constancy assumption does not appear to be as appropriate as for the synthetic sequences, for example because of stronger shadows. Moreover, the model with learned filters (LFC) slightly outperforms the model with fixed, standard filters (FFC), particularly in regions with strong brightness changes. This means that learning the filters seems to be fruitful, particularly for challenging, realistic sequences. Further results, including comparisons to other recent techniques are available at *http://vision.middlebury.edu/flow/*.

7 Conclusions

Enabled by a database of image sequences with ground truth optical flow fields, we studied the statistics of both optical flow *and* brightness constancy, and formulated a fully learned probabilistic model for optical flow estimation. We extended our initial formulation by modeling the steered derivatives of optical flow, and generalized the data term to model the constancy of linear filter responses. This provided a statistical grounding for, and extension of, various previous models of optical flow, and at the same time enabled us to learn all model parameters automatically from training data. Quantitative experiments showed that both the steered model of flow as well as the generalized data term substantially improved performance.

Currently a small number of training sequences are available with ground truth flow. A general purpose, learned, flow model will require a fully general training set; special purpose models, of course, are also possible. While a small training set may limit the generalization performance of a learned flow model, we believe that training the parameters of the model is preferable to hand tuning (particularly to individual sequences) which has been the dominant approach.

While we have focused on the objective function, the optimization method may also play an important role [27] and some models may may admit better optimization strategies than others. In addition to improved optimization, future work may consider modulating the steered flow model by the strength of the image gradient similar to [4], learning a model that adds spatial integration to the proposed filter-response constancy constraints and thus extends [8], extending the learned filter model beyond two frames, automatically adapting the model to the properties of each sequence, and learning an explicit model of occlusions and disocclusions.

Acknowledgments. This work was supported in part by NSF (IIS-0535075, IIS-0534858) and by a gift from Intel Corp. We thank D. Scharstein, S. Baker, R. Szeliski, and L. Williams for hours of helpful discussion about the evaluation of optical flow.

References

1. Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. CVIU 63, 75–104 (1996)
2. Roth, S., Black, M.J.: On the spatial statistics of optical flow. IJCV 74, 33–50 (2007)
3. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., Szeliski, R.: A database and evaluation methodology for optical flow. In: ICCV (2007)
4. Nagel, H.H., Enkelmann, W.: An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. IEEE TPAMI 8, 565–593 (1986)
5. Roth, S., Black, M.J.: Steerable random fields. In: ICCV (2007)
6. Wainwright, M.J., Simoncelli, E.P.: Scale mixtures of Gaussians and the statistics of natural images. In: NIPS, pp. 855–861 (1999)
7. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)
8. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. IJCV 61, 211–231 (2005)
9. Horn, B., Schunck, B.: Determining optical flow. Artificial Intelligence 16, 185–203 (1981)
10. Fermüller, C., Shulman, D., Aloimonos, Y.: The statistics of optical flow. CVIU 82, 1–32 (2001)
11. Gennert, M.A., Negahdaripour, S.: Relaxing the brightness constancy assumption in computing optical flow. Technical report, Cambridge, MA, USA (1987)
12. Haussecker, H., Fleet, D.: Computing optical flow with physical models of brightness variation. IEEE TPAMI 23, 661–673 (2001)
13. Toth, D., Aach, T., Metzler, V.: Illumination-invariant change detection. In: 4th IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 3–7 (2000)
14. Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J., Ogden, J.M.: Pyramid methods in image processing. RCA Engineer 29, 33–41 (1984)
15. Alvarez, L., Deriche, R., Papadopoulo, T., Sanchez, J.: Symmetrical dense optical flow estimation with occlusions detection. IJCV 75, 371–385 (2007)
16. Fleet, D.J., Black, M.J., Nestares, O.: Bayesian inference of visual motion boundaries. In: Exploring Artificial Intelligence in the New Millennium, pp. 139–174. Morgan Kaufmann Pub., San Francisco (2002)
17. Black, M.J.: Combining intensity and motion for incremental segmentation and tracking over long image sequences. In: Sandini, G. (ed.) ECCV 1992. LNCS, vol. 588, pp. 485–493. Springer, Heidelberg (1992)
18. Simoncelli, E.P., Adelson, E.H., Heeger, D.J.: Probability distributions of optical flow. In: CVPR, pp. 310–315 (1991)
19. Black, M.J., Yacoob, Y., Jepson, A.D., Fleet, D.J.: Learning parameterized models of image motion. In: CVPR, pp. 561–567 (1997)

20. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. IJCV 40, 25–47 (2000)
21. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: CVPR (2007)
22. Roth, S., Black, M.J.: Fields of experts: A framework for learning image priors. In: CVPR, vol. II, pp. 860–867 (2005)
23. Stewart, L., He, X., Zemel, R.: Learning flexible features for conditional random fields. IEEE TPAMI 30, 1145–1426 (2008)
24. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput 14, 1771–1800 (2002)
25. Blake, A., Zisserman, A.: Visual Reconstruction. The MIT Press, Cambridge, Massachusetts (1987)
26. Zhu, S., Wu, Y., Mumford, D.: Filters random fields and maximum entropy (FRAME): To a unified theory for texture modeling. IJCV 27, 107–126 (1998)
27. Lempitsky, V., Roth, S., Rother, C.: FusionFlow: Discrete-continuous optimization for optical flow estimation. In: CVPR (2008)