On the Spatial Statistics of Optical Flow

Stefan Roth Michael J. Black
Department of Computer Science, Brown University, Providence, RI, USA
{roth,black}@cs.brown.edu

Preprint - September 25, 2006

Abstract

We present an analysis of the spatial and temporal statistics of "natural" optical flow fields and a novel flow algorithm that exploits their spatial statistics. Training flow fields are constructed using range images of natural scenes and 3D camera motions recovered from hand-held and car-mounted video sequences. A detailed analysis of optical flow statistics in natural scenes is presented and machine learning methods are developed to learn a Markov random field model of optical flow. The prior probability of a flow field is formulated as a Field-of-Experts model that captures the spatial statistics in overlapping patches and is trained using contrastive divergence. This new optical flow prior is compared with previous robust priors and is incorporated into a recent, accurate algorithm for dense optical flow computation. Experiments with natural and synthetic sequences illustrate how the learned optical flow prior quantitatively improves flow accuracy and how it captures the rich spatial structure found in natural scene motion.

Keywords: optical flow, database of flow fields, spatial statistics, natural scenes, Markov random fields, machine learning

1 Introduction

In this paper we study the statistics of optical flow in natural imagery and exploit recent advances in machine learning to obtain a rich probabilistic model for optical flow fields. This extends work on the analysis of image statistics in natural scenes and range images to the domain of image motion. In doing so we make connections to previous robust statistical formulations of optical flow smoothness priors and learn a new Markov random field prior over large neighborhoods using a *Field-of-Experts* model (Roth and Black, 2005a). We extend a recent (and very accurate) optical flow method (Bruhn et al., 2005) with this new prior and provide an algorithm for estimating optical flow from pairs of images. We quantitatively compare the learned prior with more traditional robust priors and find that in our experiments the accuracy is improved by about 10% while removing the need for tuning the scale parameter of the traditional priors.

Figure 1: Flow fields generated for an outdoor (left) and an indoor (right) scene. The horizontal motion **u** is shown on the left, the vertical motion **v** on the right of each figure; dark/light means negative/positive motion (independently scaled to $0 \ldots 255$ for display).

Natural image statistics have received intensive study (Huang, 2000; Lee et al., 2001; Ruderman, 1994), but the spatial and temporal statistics of optical flow are relatively unexplored because databases of natural scene motions are currently unavailable. One of the contributions of this paper is the development of such a database.

The spatial statistics of the motion field (i. e., the ideal optical flow) are determined by the interaction of 1) camera motion; 2) scene depth; and 3) the independent motion of objects. Here we focus on rigid scenes and leave independent motion for future work (though we believe the statistics from rigid scenes are useful for scenes with independent motion and will show experimental results with independent motion). To generate a realistic database of optical flow fields we exploit the Brown range image database (Lee and Huang, 2000), which contains depth images of complex scenes including forests, indoor environments, and generic street scenes. Given 3D camera motions and range images we generate flow fields that have the rich spatial statistics of "natural" flow fields[1]. A set of natural 3D motions was obtained from both hand-held and car-mounted cameras performing a variety of motions including translation, rotation, and active fixation. The 3D motion was recovered from these video sequences using commercial software (2d3 Ltd., 2002). Figure 1 shows two example flow fields generated using the 3D motions and the range images.

Our study shows that the first-derivative statistics of optical flow fields are very heavy-tailed as are the statistics of natural images. We observe that the first derivative statistics are well modeled by heavy tailed distributions such as the Student-t distribution. This provides a connection to previous robust statistical methods for recovering optical flow that modeled spatial smoothness using robust functions (Black and Anandan, 1996) and suggests that the success of robust methods is due to the fact that they capture the first-order statistics of optical flow.

Our goal here is to go beyond such local (first derivative) models and formulate a Markov random field (MRF) prior that captures richer spatial statistics present in larger neighborhoods. To that end, we exploit

[1] By "natural" here we primarily mean spatially structured and relevant to humans. Consequently the data represents scenes humans inhabit and motions they perform. We do not include non-rigid or textural motions resulting from independent movement in the environment (though these are also "natural"). Motions of a person's body and motions they cause in other objects are also excluded at this time.

a "Field of Experts" (FoE) model (Roth and Black, 2005a), that represents MRF clique potentials in terms of various linear filter responses on each clique. We model these potentials as a product of t-distributions and we learn both the parameters of the distribution and the filters themselves using contrastive divergence (Hinton, 2002; Roth and Black, 2005a).

We compute optical flow using the learned prior as a smoothness term. The log-prior is combined with a data term and we minimize the resulting energy (log-posterior). While the exact choice of data term is not relevant for the analysis, here we use the recent approach of Bruhn et al. (2005), which replaces the standard optical flow constraint equation with a tensor that integrates brightness constancy constraints over a spatial neighborhood. We present an algorithm for estimating dense optical flow and compare the performance of standard robust spatial terms with the learned FoE model on both synthetic and natural imagery.

1.1 Previous work

There has been a great deal of work on modeling natural image statistics (Huang, 2000; Lee et al., 2001; Olshausen and Field, 1996; Ruderman, 1994; Srivastava et al., 2003) facilitated by the existence of large image databases. One might expect optical flow statistics to differ from image statistics in that there is no equivalent of "surface markings" in motion and all structure in rigid scenes results from the shape of surfaces and the discontinuities between them. In this way it seems plausible that flow statistics share more with depth statistics. Unlike optical flow, direct range sensors exist and a time-of-flight laser was used in (Huang et al., 2000) to capture the depth in a variety of scenes including residential street scenes, forests, and indoor environments. Scene depth statistics alone, however, are not sufficient to model optical flow, because image motion results from the combination of the camera motion and depth. While models of self-motion in humans and animals (Betsch et al., 2004; Lewen et al., 2001) have been studied, we are unaware of attempts to learn or exploit a database of camera motions captured by a moving camera in natural scenes.

The most similar work to ours also uses the Brown range image database to generate realistic synthetic flow fields (Calow et al., 2004). The authors use a gaze tracker to record how people view the range images and then simulate their motion into the scene with varying fixation points. Their focus is on human perception of flow and consequently they analyze a retinal projection of the flow field. They also limit their analysis to first-order statistics and do not propose an algorithm for exploiting these statistics in the computation of optical flow.

Previous work on learning statistical models of video focuses on the statistics of the changing brightness patterns rather than the flow it gives rise to. For example, adopting a classic sparse-coding hypothesis, video sequences can be represented using a set of learned spatio-temporal filters (van Harteren and Ruderman, 1998). Other work has focused on the statistics of the classic brightness constancy assumption (and how it is violated) rather than the spatial statistics of the flow field (Fermüller et al., 2001; Simoncelli et al., 1991).

The lack of training data has limited research on learning spatial models of optical flow. One exception is the work by Fleet et al. (2000) in which the authors learn local models of optical flow from examples using principal component analysis (PCA). In particular, they use synthetic models of moving occlusion boundaries and bars to learn linear models of the flow for these motion features. Local, non-overlapping models such as these may be combined in a spatio-temporal Bayesian network to estimate coherent global flow fields (Fleet et al., 2002). While promising, these models cover only a limited range of the variation in natural flow fields.

There is related interest in the statistics of optical flow in the video retrieval community; for example, Fablet and Bouthemy (2001) learn statistical models using a variety of motion cues to classify videos based on their spatio-temporal statistics. These methods, however, do not focus on the estimation of optical flow.

The formulation of smoothness constraints for optical flow estimation has a long history (Horn and Schunck, 1981), as has its Bayesian formulation in terms of Markov random fields (Black and Anandan, 1991; Heitz and Bouthemy, 1993; Konrad and Dubois, 1988; Marroquin et al., 1987; Murray and Buxton, 1987). Such formulations of optical flow estimate the flow using Bayesian inference based on a suitable posterior distribution. Given an image sequence $\mathbf{f}$, the horizontal flow $\mathbf{u}$ and the vertical flow $\mathbf{v}$ are typically found by maximizing the posterior distribution $p(\mathbf{u}, \mathbf{v} \mid \mathbf{f})$ with respect to $\mathbf{u}$ and $\mathbf{v}$. Other inference methods estimate the flow using sampling (Barbu and Yuille, 2004) or by computing expectations. Using Bayes rule the posterior can be broken down into a likelihood term $p(\mathbf{f} \mid \mathbf{u}, \mathbf{v})$, also called the data term, and an optical flow prior $p(\mathbf{u}, \mathbf{v})$, also called the spatial term:

$$\arg\max_{\mathbf{u},\mathbf{v}} \; p(\mathbf{u}, \mathbf{v} \mid \mathbf{f}) = \arg\max_{\mathbf{u},\mathbf{v}} \; p(\mathbf{f} \mid \mathbf{u}, \mathbf{v}) \cdot p(\mathbf{u}, \mathbf{v}). \tag{1}$$

The data term enforces the brightness constancy assumption, which underlies most flow estimation techniques (e. g., Horn and Schunck, 1981); the spatial term enforces spatial smoothness for example using a Markov random field model. Previous work, however, has focused on very local prior models that are typically formulated in terms of the first differences in the optical flow (i. e., the nearest neighbor differences). This can model piecewise constant or smooth flow but not more complex spatial structures. A large class of techniques make use of spatial regularization terms derived from the point of view of variational methods, PDEs, and nonlinear diffusion (Alvarez et al., 2000; Ben-Ari and Sochen, 2006; Bruhn et al., 2005; Cremers and Soatto, 2005; Papenberg et al., 2006; Proesmans et al., 1994; Scharr and Spies, 2005; Weickert and Schnörr, 2001). These methods also exploit only local measures of the flow gradient and are not directly motivated by statistical properties of natural flow fields. Other work has imposed geometric rather than spatial smoothness constraints on multi-frame optical flow (Irani, 1999).

Weiss and Adelson (1998) propose a Bayesian model of motion estimation to explain human perception of visual stimuli. In particular, they argue that the appropriate prior prefers "slow and smooth" motion (see also (Lu and Yuille, 2006)). Their stimuli, however, are too simplistic to probe the nature of flow priors in complex scenes. We find these statistics are more like those of natural images in that the motions are piecewise smooth; large discontinuities give rise to heavy tails in the first derivative statistics. Our analysis suggests that a more appropriate flow prior is "mostly slow and smooth, but sometimes fast and discontinuous".

2 Spatial Statistics of Optical Flow

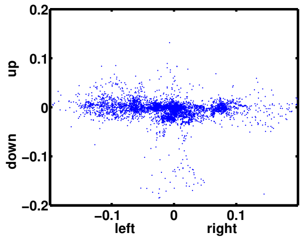
2.1 Obtaining training data

One of the key challenges in analyzing and learning the spatial statistics of optical flow is to obtain suitable optical flow data. The issue here is that optical flow cannot be directly measured, which makes the statistics of optical flow a largely unexplored field. Synthesizing realistic flow fields is thus the only viable option for studying, as well as learning the statistics of optical flow. Optical flow has previously been generated using computer graphics techniques in order to obtain benchmarks with ground truth, for example in case of the Yosemite sequence (Barron et al., 1994). In most cases, the underlying scenes are very simple however. Our goal here is to create a database of realistic optical flow fields as they arise in natural as well as man-made scenes. It is unlikely that the rich statistics will be captured by any manual construction of the training data as in (Fleet et al., 2000), which uses simple polygonal object models. We also cannot rely on optical flow as computed from image sequences, because then we would learn the statistics of the algorithm used to compute the flow. It would be possible to use complex scene models from computer graphics, but such scenes would have to be very complex to capture the structure of real scenes. Instead we rely on range images from the Brown range image database (Lee and Huang, 2000), which provides accurate scene depth information for a set of 197 indoor and outdoor scenes (see Figure 3). Optical flow has been generated from range images by Calow et al. (2004), but they focus on a synthetic model of human gaze and ego-motion. Unlike Calow et al. (2004), we focus on optical flow fields as they arise in machine vision applications as opposed to human vision[2].

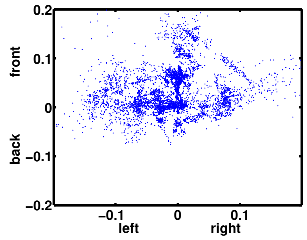
The range image database we use captures information about surfaces and surface boundaries in natural scenes, but it is completely static. Hence we focus only on the case of ego-motion in static scenes. A rigorous study of the optical flow statistics of independently moving objects will remain the subject of future work. Despite this limitation, the range of motions represented is broad and varied.

Camera motion. Apart from choosing appropriate 3D scenes, finding suitable camera motion data is another challenge. In order to cover a broad range of possible frame-to-frame camera motions, we used a database of 67 video clips of approximately 100 frames, each of which was shot using a hand-held or car-mounted video camera. The database is comprised of various kinds of motion, including forward walking and moving the camera around an object of interest. The extrinsic and intrinsic camera parameters were recovered using the *boujou* software system (2d3 Ltd., 2002) from a number of tracked feature points; the underlying camera model is a simple perspective pinhole camera. Figure 2 shows empirical distributions of the camera translations and rotations in the database. The plots reveal that left-right movements are more common than up-down movements and that moving into the scene occurs more frequently than moving out

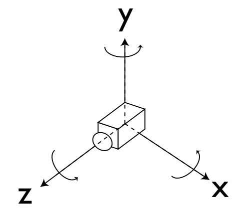
[2] It is worth noting that our goal is to learn the prior probability of the motion field (i. e., the projected scene flow) rather than the apparent motion (i. e., the optical flow) since most applications of optical flow estimation seek this "true" motion field.



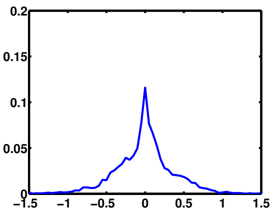
(a) Left-right (x) and up-down (y) translational motion.



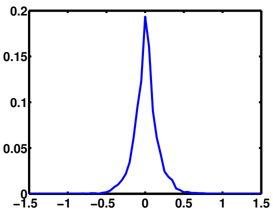
(b) Left-right (x) and forward-backward (z) translational motion.



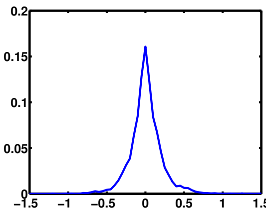
(c) Camera coordinate system.



(d) Yaw (rotation around y).



(e) Pitch (rotation around x).



(f) Roll (rotation around z).

Figure 2: Statistics of camera motion in our database: *(a-b)* Scatter plots of the translational camera motion between subsequent frames (scale in meters). *(d-f)* Histograms of camera rotation between subsequent frames (angle in degrees).

of the scene. Similarly the empirical distributions of camera rotation show that yaw occurs more frequently than pitch and roll motions of the camera.

Synthesizing optical flow. To generate optical flow from range and camera motion data we use the following procedure: Given a range scene, we build a triangular mesh from the depth measurements, which allows us to view the scene from any view point. However, only viewpoints near the location of the range scanner will lead to the intended appearance. Given two camera transformations of subsequent frames, rays are then cast through every pixel of the first frame to find the corresponding scene point from the range image. Each of these scene points is then projected onto the image plane of the second frame. The optical flow is simply given by the difference in image coordinates under which a scene point is viewed in each of the two cameras. We used this procedure to generate a database of 800 optical flow fields, each 250×200 pixels large[3]. Figure 1 shows example flow fields from this database. Note that we do not explicitly represent the regions of occlusion or disocclusion in the database. While occlusions for a particular camera motion can be detected based on the polygonal scene model, they are not explicitly modeled in our spatial model of optical flow. While our learned MRF model will capture motion boundaries implicitly, a rigorous treatment of occlusions (cf. Fleet et al., 2002; Ross and Kaelbling, 2005) will remain future work.

Additionally, we have to select appropriate combinations of range scenes and 3D camera motions. In previous work (Roth and Black, 2005b), we independently sampled range scenes as well as camera motions.

[3] The database is available at `http://www.cs.brown.edu/~roth/research/flow/downloads.html`.

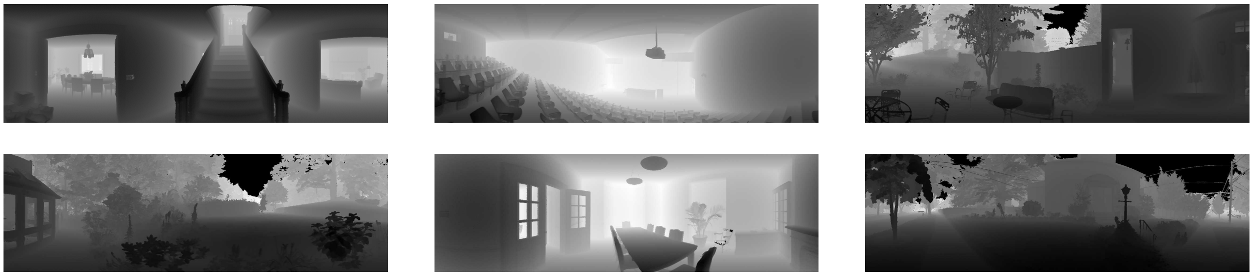


Figure 3: Example range scenes from the Brown range image database (Lee and Huang, 2000). Intensity here codes for depth with distant objects appearing brighter. Regions for which no range estimate could be obtained are shown in black.

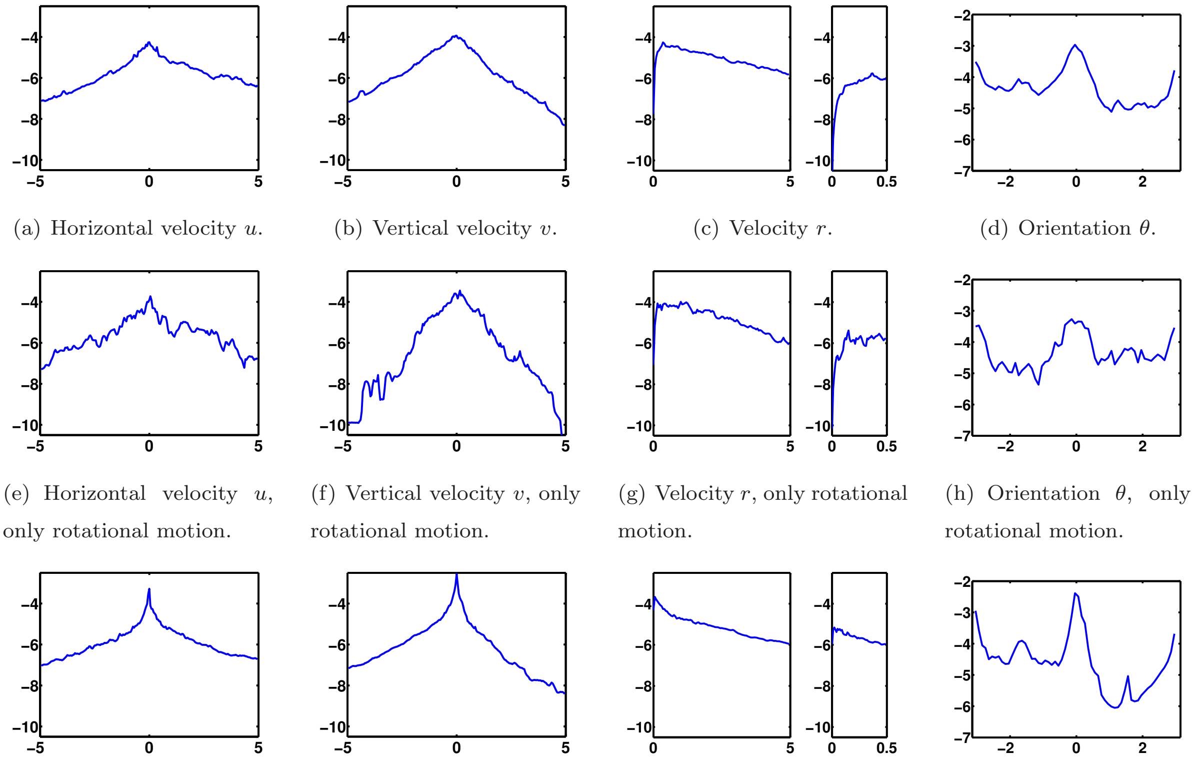
This is a simplifying assumption, however, because the kind of camera motion executed may depend on the kind of scene that is viewed. In particular, the amplitude of the camera motion may be dependent on the scale of the viewed scene. It seems likely, for example, that camera motion tends to be slower when objects are nearby, than when very distant objects are viewed. To account for this potential dependency, we go beyond our previous work (Roth and Black, 2005b) and introduce a coupling between camera motion and scene structure. We first sample a random camera motion and compute a depth histogram of the feature points being tracked in the original sequence to recover the camera motion. These feature points are provided by the *boujou* system and give a coarse description of the scene being filmed. We then compute depth histograms for 4925 potential views of various range scenes (25 different views of each scene), and find the Bhattacharyya distance (Kailath, 1967)

$$d(\mathbf{p}, \mathbf{q}^{(j)}) = \sqrt{1 - \sum_i \sqrt{p_i \cdot q_i^{(j)}}} \tag{2}$$

between the depth histogram of the scene corresponding to the camera motion ($\mathbf{p}$) and each of the candidate views of the range scenes ($\mathbf{q}^{(j)}$). We define a weight $w_j := (1 - d(\mathbf{p}, \mathbf{q}^{(j)}))^2$ for each candidate range scene, which assigns greater weight to range scenes with similar depth statistics to the camera motion scene. A range scene view is finally sampled according to these weights, which achieves a coupling of the depth and camera motion statistics. In practice we found that many of the range scenes have very distant objects compared to the scenes for which we have camera motion. We found that if the range scenes are scaled down using a global scaling factor of 0.25, then the range statistics much better matched those for the scenes used to capture the camera motion. We hence used this scaling factor when sampling scene/motion pairs.

2.2 Velocity statistics

Using this database, we are able to study several statistical properties of optical flow. Figure 4 shows log-histograms of the image velocities in various forms. In addition to the optical flow database described in the previous section, we also study two further flow databases: one resulting from purely translational camera motion and one from purely rotational camera motion. These allow us to attribute various properties to each



(a) Horizontal velocity u.
(b) Vertical velocity v.
(c) Velocity r.
(d) Orientation θ.
(e) Horizontal velocity u, only rotational motion.
(f) Vertical velocity v, only rotational motion.
(g) Velocity r, only rotational motion.
(h) Orientation θ, only rotational motion.
(i) Horizontal velocity u, only translational motion.
(j) Vertical velocity v, only translational motion.
(k) Velocity r, only translational motion.
(l) Orientation θ, only translational motion.

Figure 4: Velocity and orientation histograms of the optical flow in our database (log scale). The right plot in parts *(c)*, *(g)*, and *(k)* shows details of the left plot.

type of motion. We observe that the vertical velocity in Figure 4(b) is roughly distributed like a Laplacian distribution; the horizontal velocity in 4(a) shows a slightly broader histogram that falls off less quickly. This is consistent with our observations that horizontal camera motions are more common. If we factor the motion into rotational and translational components, we find that the heavy tails of the velocity histograms are due to the translational camera motion (Fig. 4(i,j)), while the image velocity from rotational camera motion covers a smaller range (Fig. 4(e,f)). Maybe somewhat surprisingly the vertical image motion due to translational camera motion in 4(j) appears to be biased toward negative values, which correspond to motion toward the bottom of the image. This nevertheless has a quite simple explanation: From the camera motion statistics we know that moving in the viewing direction occurs quite frequently; for hand-held and car-mounted cameras this motion will typically be along the ground plane. Finally, the part of the image below the horizon typically occupies more than half of the whole image, which causes the focus of expansion to be in the top half of the image. Because of that, downward image motion dominates upward image motion for this common type of camera motion.

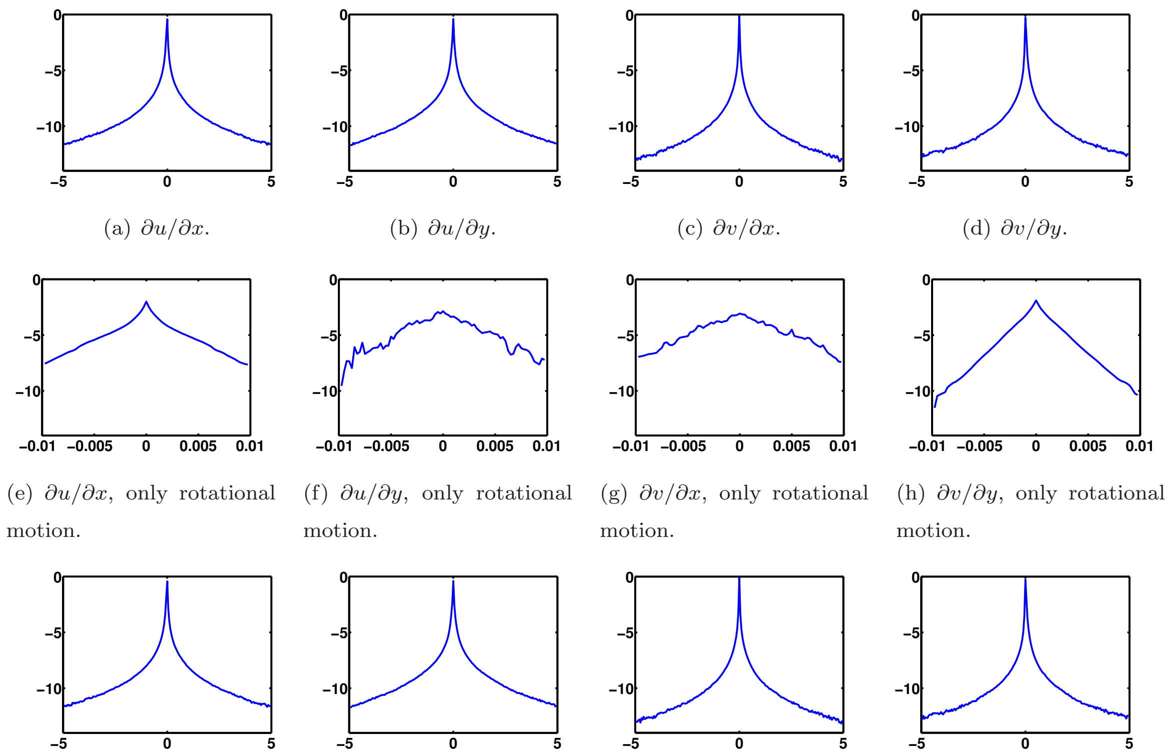
Figure 4(c) shows the magnitude of the velocity, which falls off in a manner similar to a Laplacian distribution (Simoncelli et al., 1991). The statistics of natural image motion hence suggest that image motions are typically slow. The results of Weiss and Adelson (1998) suggest that humans may exploit this prior information in their estimation of image motion. The heavy-tailed nature of the velocity histograms indicates that while slow motions are typical, this assumption is still violated fairly frequently for "natural" optical flow. From the histograms we can also see that very small motions (near zero) seem to occur rather infrequently, which is mainly attributable to rotational camera motion (Fig. 4(g)). This suggests that the camera is rarely totally still and is often rotated at least a small amount. The orientation histogram in 4(d) again shows the preference for horizontal motion (the bumps at 0 and $\pm\pi$). However, there are also smaller spikes indicating somewhat frequent up-down motion (at $\pm\pi/2$); we observe similar properties for rotational (Fig. 4(h)) and translational camera motion (Fig. 4(l)).

2.3 Derivative statistics

Figure 5 shows the first derivative histograms of the spatial derivatives for both horizontal and vertical image motion. For flow from translational and combined translational/rotational camera motion, the distributions are all heavy-tailed and strongly resemble Student t-distributions (Fig. 5(a-d) and 5(i-l)). Such distributions have also been encountered in the study of natural images (e. g., Grenander and Srivastava, 2001; Huang, 2000; Lee et al., 2001). In natural images, the image intensity is often locally smooth, but occasionally shows large jumps at object boundaries or in fine textures, which give rise to substantial probability mass in the tails of the distribution. Furthermore, the study of range images (Huang et al., 2000) has shown similar derivative statistics. For scene depth the heavy-tailed distributions arise from depth discontinuities mostly at object boundaries. Because the image motion from camera translation is directly dependent on the scene depth, it is not surprising to see similar distributions for optical flow. Optical flow from purely rotational camera motion on the other hand is independent of scene depth, which is reflected in the much less heavy-tailed derivative histograms (Fig. 5(e-h)).

The large peaks at zero of the derivative histograms show that optical flow is typically very smooth. Aside from prior information on the flow magnitude, the work by Weiss and Adelson (1998) suggested that humans also use prior information about the smoothness of optical flow. The heavy-tailed nature of the statistics of natural optical flow suggests that flow discontinuities still occur with considerable frequency. Hence an appropriate prior would be "mostly slow" and "mostly smooth". This suggests a direction for further study in human vision to see if such priors are used.

Furthermore, the observed derivative statistics likely explain the success of robust statistical formulations for optical flow computation based on M-estimators (e. g., Black and Anandan, 1991). In (Black and Anandan, 1991) the spatial smoothness term was formulated as a robust (Lorentzian) function of the horizontal and vertical partial derivatives of the flow field. This robust function fits the marginal statistics of these partial derivatives very well.



(a) $\partial u/\partial x$. (b) $\partial u/\partial y$. (c) $\partial v/\partial x$. (d) $\partial v/\partial y$.

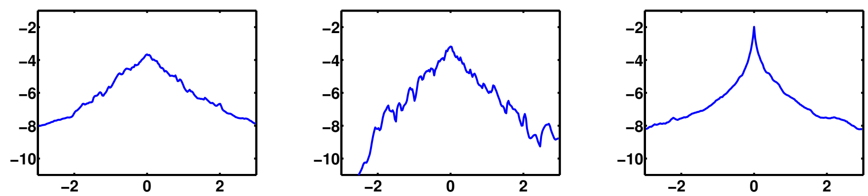
(e) $\partial u/\partial x$, only rotational motion. (f) $\partial u/\partial y$, only rotational motion. (g) $\partial v/\partial x$, only rotational motion. (h) $\partial v/\partial y$, only rotational motion.

(i) $\partial u/\partial x$, only translational motion. (j) $\partial u/\partial y$, only translational motion. (k) $\partial v/\partial x$, only translational motion. (l) $\partial v/\partial y$, only translational motion.

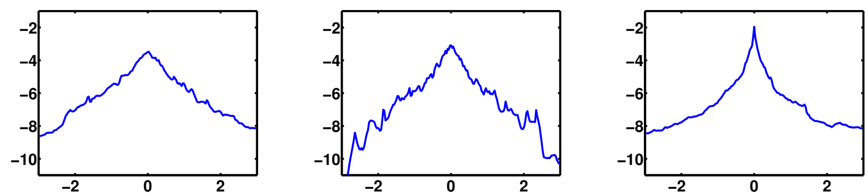
Figure 5: Spatial derivative histograms of optical flow in our database (log scale): $\partial u/\partial x$, for example, denotes the horizontal derivative of the horizontal flow.

Temporal statistics. Even though this paper is mainly concerned with analyzing and modeling the spatial statistics of optical flow, we also performed an analysis of the temporal properties of the flow. In particular, we generated pairs of flow fields from three subsequent camera viewpoints, and computed the difference of the optical flow between subsequent pairs in the image coordinate system. Figure 6 shows the statistics of this temporal derivative, again for all three types of camera motion. We find that similar to the spatial derivatives, the temporal derivatives of flow from translational and combined translational/rotational camera motion are more heavy-tailed (Fig. 6(a,c,d,f)) than for flow from purely rotational motion (Fig. 6(b,e)). In the remainder of this work, we do not pursue the temporal statistics any further, but the extension of the model proposed in Section 3 to the temporal domain will be an interesting avenue for future work.

Joint statistics. We also obtained joint empirical histograms of derivatives of the horizontal and vertical flow. Figure 7 shows log-histograms of various derivatives, where a derivative of the horizontal flow is plotted against a derivative of the vertical flow. We can see from the shape of the empirical histograms that the derivatives of horizontal and vertical flow are largely independent. This observation is reinforced quantitatively when considering the mutual information (MI) between the various derivatives. We estimated

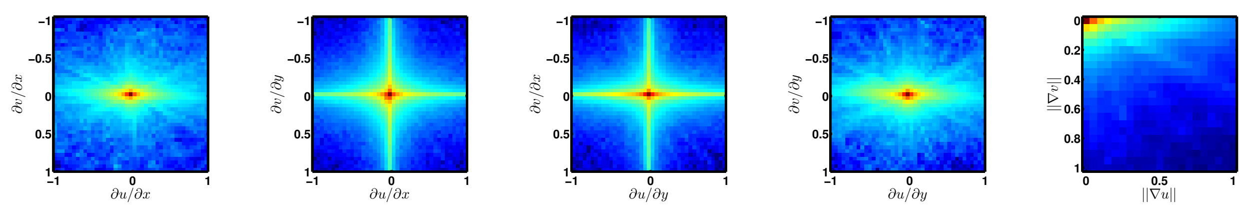


(a) $\partial u/\partial t$. (b) $\partial u/\partial t$, only rotational motion. (c) $\partial u/\partial t$, only translational motion.



(d) $\partial v/\partial t$. (e) $\partial v/\partial t$, only rotational motion. (f) $\partial v/\partial t$, only translational motion.

Figure 6: Temporal derivative histograms of optical flow in our database (log scale).



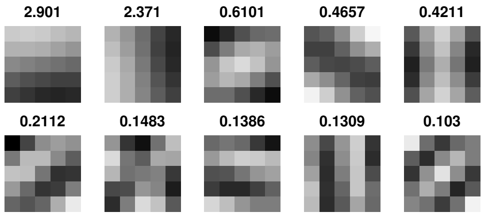
(a) $\partial u/\partial x$ and $\partial v/\partial x$ (MI 0.11 bits) (b) $\partial u/\partial x$ and $\partial v/\partial y$ (MI 0.06 bits) (c) $\partial u/\partial y$ and $\partial v/\partial x$ (MI 0.03 bits) (d) $\partial u/\partial y$ and $\partial v/\partial y$ (MI 0.13 bits) (e) $||\nabla u||$ and $||\nabla v||$ (MI 0.12 bits)

Figure 7: Joint log-histograms of derivatives of horizontal and vertical flow and their mutual information (MI). The mutual information expresses how much information (in bits) one derivative value contains about the other; the small values here indicate approximate statistical independence.

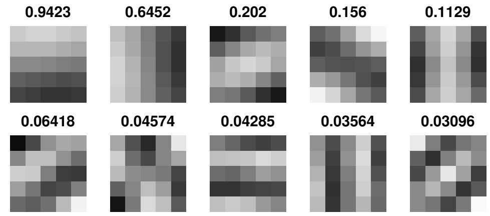
the mutual information directly from the empirical histograms. Figure 7 gives the MI values for all considered joint histograms.

2.4 Principal component analysis

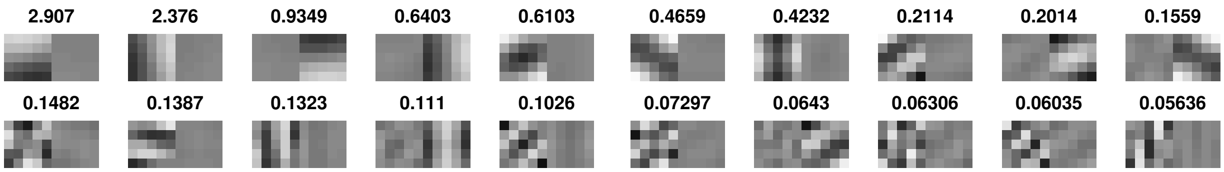
We also performed principal component analysis on small patches of flow of various sizes. Figure 8 shows the results for horizontal flow $\mathbf{u}$ and vertical flow $\mathbf{v}$ in 5×5 patches. The principal components of horizontal and vertical flow look very much alike, but the variance of the vertical flow components is smaller due to the observed preference for horizontal motion. We can see that a large portion of the flow variance is focused on the first few principal components which, as with images, resemble derivative filters of various orders (cf. Fleet et al., 2000). A joint analysis of horizontal and vertical flow as shown in 8(c) reveals that the principal components treat horizontal and vertical flow largely independently, i. e., one of the components is constant



(a) Horizontal velocity $\mathbf{u}$.



(b) Vertical velocity $\mathbf{v}$.



(c) Joint horizontal velocity $\mathbf{u}$ & vertical velocity $\mathbf{v}$.

Figure 8: First 10 principal components (20 for $\mathbf{u}$ & $\mathbf{v}$) of the image velocities in 5×5 patches. The numbers denote the variance accounted for by the principal component.

while the other is not.

3 Modeling Optical Flow

We capture the spatial statistics of optical flow using the *Fields-of-Experts* (FoE) approach (Roth and Black, 2005a), which models the prior probability of images (and here optical flow fields) using a Markov random field. In contrast to many previous MRF models, it uses larger cliques of, for example, 3×3 or 5×5 pixels, and allows learning the appropriate clique potentials from training data. We argue here that spatial regularization of optical flow will benefit from prior models that capture interactions beyond adjacent pixels. Markov random fields of high-order have been shown to be quite powerful for certain low-level vision applications. The FRAME model by Zhu et al. (1998), for example, has been very successful in the area of texture modeling. To our knowledge, neither this nor other learned high-order models have been applied to the problem of optical flow estimation. In Section 2.3 we have seen that the derivatives of horizontal and vertical motion are largely independent, hence for simplicity, we will typically treat horizontal and vertical image motions separately, and learn two independent models.

In the Markov random field framework, the pixels of an image or flow field are assumed to be represented by nodes V in a graph $G = (V, E)$, where E are the edges connecting nodes. The edges are typically defined through a neighborhood system, such as all spatially adjacent pairs of pixels. We will instead consider neighborhood systems that connect all nodes in a square $m \times m$ region. Every such neighborhood centered on a node (pixel) $k = 1, \ldots, K$ defines a maximal clique $\mathbf{x}_{(k)}$ in the graph; $\mathbf{x}_{(k)}$ denotes the vector of pixels

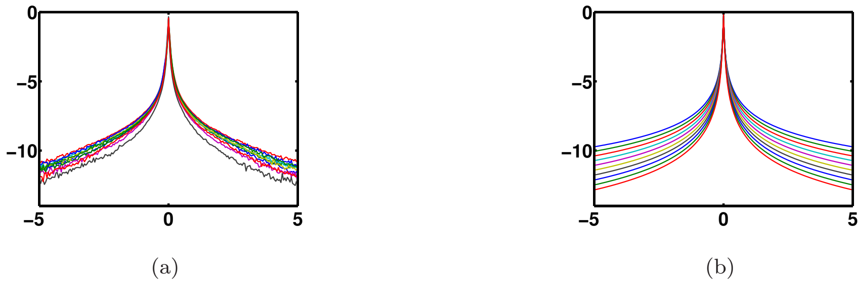


Figure 9: *(a)* Log-histograms of filter responses of 10 zero-mean, unit-variance random filters applied to the horizontal flow in our optical flow database. *(b)* Log-densities of 10 different members of the family of Student t-distributions illustrating the qualitative similarity to the shape of the filter statistics.

in the clique. We write the probability of a flow field component $\mathbf{x} \in \{\mathbf{u}, \mathbf{v}\}$ under the MRF as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{k=1}^{K} \psi(\mathbf{x}_{(k)}), \tag{3}$$

where $\psi(\mathbf{x}_{(k)})$ is the so-called potential function for clique $\mathbf{x}_{(k)}$ (we assume homogeneous MRFs here) and Z is a normalization term. Because of our assumptions, the joint probability of the flow field $p(\mathbf{u}, \mathbf{v})$ is simply the product of the probabilities of the two components $p(\mathbf{u}) \cdot p(\mathbf{v})$ under the MRF model. As we will discuss in more detail below, most typical prior models for optical flow can be expressed in this MRF framework.

When considering MRFs with cliques that do not just consist of pairs of nodes, finding suitable potential functions $\psi(\mathbf{x}_{(k)})$ and training the model on a database become much more challenging. In our experiments we have observed that responses of linear filters applied to flow components in the optical flow database show histograms that are typically well fit by t-distributions. Figure 9 illustrates this with the response statistics of 10 zero-mean random filters on our optical flow database, as well as 10 different members of the family of Student t-distributions. The FoE model that we propose uses the Products-of-Experts framework (Hinton, 1999; Teh et al., 2003), and motivated by the above observation, models the clique potentials with products of Student t-distributions. Each expert distribution works on the response to a linear filter $\mathbf{J}_i$. Even though the filter applies to a square region of pixels, we assume for mathematical convenience that $\mathbf{J}_i$ is expressed as a vector. The cliques' potential under this model is written as:

$$\psi(\mathbf{x}_{(k)}) = \prod_{i=1}^{N} \phi(\mathbf{J}_i^{\mathrm{T}} \mathbf{x}_{(k)}; \alpha_i), \tag{4}$$

where each expert is a t-distribution with parameter α_i:

$$\phi_i(\mathbf{J}_i^{\mathrm{T}} \mathbf{x}_{(k)}; \alpha_i) = \left(1 + \frac{1}{2}(\mathbf{J}_i^{\mathrm{T}} \mathbf{x}_{(k)})^2\right)^{-\alpha_i}. \tag{5}$$

The FoE optical flow prior is hence written as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{k=1}^{K} \prod_{i=1}^{N} \phi(\mathbf{J}_i^{\mathrm{T}} \mathbf{x}_{(k)}; \alpha_i). \tag{6}$$

We should note here that this model can easily be extended to a joint model of horizontal and vertical flow by considering cliques of size $m \times m \times 2$.

The FRAME model (Zhu et al., 1998) takes a somewhat similar form as Eq. (6); the key difference is that the FRAME model uses hand-defined filters and discrete "expert" functions, whereas the FoE model relies on continuous experts and learns the filters from training data. The filters $\mathbf{J}_i$ as well as the expert parameters α_i are jointly learned from training data using maximum likelihood estimation. Because there is no closed form expression for the partition function Z in case of the FoE model, maximum likelihood estimation relies on Markov chain Monte Carlo sampling, which makes the training process computationally expensive. To speed up the learning process, we use the contrastive divergence algorithm (Hinton, 2002), which approximates maximum likelihood learning, but only relies on a fixed, small number of Markov chain iterations in the sampling phase. Convergence of the algorithm is difficult to establish automatically due to its stochastic nature, hence we manually monitored convergence. For our experiments, we trained various FoE models: 3×3 cliques with 8 filters, 5×5 cliques with 24 filters, as well as a joint $3 \times 3 \times 2$ model with 16 filters. We restrict the filters so that they do not capture the mean velocity in a patch and are thus only sensitive to relative motion. The training was done on 2000 flow fields of size 15×15. These small flow fields were selected and cropped uniformly at random from the synthesized flow fields described in Section 2. The filters were initialized randomly by drawing from the joint Gaussian distribution of flow patches with the same size as the filters; the expert parameters were initialized uniformly with $\alpha_i = 1$. However, our experience with contrastive divergence learning indicates that the model performance is not very sensitive to initialization. For more details about how FoE models can be trained, we refer the reader to (Roth and Black, 2005a).

In the context of optical flow estimation, the prior knowledge about the flow field is typically expressed in terms of an energy function $E(\mathbf{x}) = -\log p(\mathbf{x})$. Accordingly, we can express the energy for the FoE prior model as

$$E_{\mathrm{FoE}}(\mathbf{x}) = -\sum_{k=1}^{K} \sum_{i=1}^{N} \log \phi(\mathbf{J}_i^{\mathrm{T}} \mathbf{x}_{(k)}; \alpha_i) + \log Z. \tag{7}$$

Note that for fixed parameters α and $\mathbf{J}$, the partition function Z is constant, and can thus be ignored when estimating flow.

The optical flow estimation algorithm we propose in the next section relies on the gradient of the energy function with respect to the flow field. Since Fields of Experts are log-linear models, expressing and computing the gradient of the energy is relatively easy. Following Zhu and Mumford (1997), the gradient of the energy can be written as

$$\nabla_{\mathbf{x}} E_{\mathrm{FoE}}(\mathbf{x}) = -\sum_{i=1}^{N} \mathbf{J}_i^{-} * \xi_i(\mathbf{J}_i * \mathbf{x}), \tag{8}$$

where $\mathbf{J}_i * \mathbf{x}$ denotes the convolution of image velocity component $\mathbf{x} \in \{\mathbf{u}, \mathbf{v}\}$ with filter $\mathbf{J}_i$. We also define $\xi_i(y) = \partial/\partial y \ \log \phi(y; \alpha_i)$ and let $\mathbf{J}_i^{-}$ denote the filter obtained by mirroring $\mathbf{J}_i$ around its center pixel (Zhu and Mumford, 1997).

3.1 Comparison to traditional regularization approaches

Many traditional regularization approaches for optical flow can be formalized in a very similar way, and some of them are in fact restricted versions of the more general FoE model introduced above. One widely used regularization technique is based on the gradient magnitude of the flow field components (Bruhn et al., 2005). Because the gradient magnitude is typically computed using finite differences in horizontal and vertical directions, the maximal cliques of the corresponding MRF model consist of 4 pixels arranged in a diamond-like shape. Assuming that the flow field component $\mathbf{x} \in \{\mathbf{u}, \mathbf{v}\}$ is indexed as $x_{i,j}$, the model can be formalized as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{i,j} \psi(x_{i+1,j}, x_{i-1,j}, x_{i,j+1}, x_{i,j-1}) = \frac{1}{Z} \prod_{i,j} \hat{\psi}\left(\sqrt{(x_{i+1,j} - x_{i-1,j})^2 + (x_{i,j+1} - x_{i,j-1})^2}\right). \tag{9}$$

If $\hat{\psi}$ is chosen to be a Gaussian distribution, one obtains a standard linear regularizer; heavy-tailed potentials on the other hand will lead to robust, non-linear regularization (Bruhn et al., 2005). Because computing the gradient magnitude involves a non-linear transformation of the pixel values, this model cannot be directly mapped into the FoE framework introduced above.

A second regularization approach that has been used frequently throughout the literature is based on component derivatives directly as opposed to the gradient magnitude (e. g., Black and Anandan, 1996; Horn and Schunck, 1981). Assuming the same notation as above, this prior model can be formalized as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{i,j} \psi(x_{i+1,j}, x_{i,j}) \cdot \prod_{i,j} \psi(x_{i,j+1}, x_{i,j}) = \frac{1}{Z} \prod_{i,j} \hat{\psi}(x_{i+1,j} - x_{i,j}) \cdot \prod_{i,j} \hat{\psi}(x_{i,j+1} - x_{i,j}). \tag{10}$$

As above, $\hat{\psi}$ can be based on Gaussian or robust potentials and parametrized accordingly. In contrast to the gradient magnitude prior, the potentials of this prior are based on a linear transformation of the pixel values in a clique. Because of that, it is a special case of the FoE model in Eq. (6); here, the filters are simple derivative filters based on finite differences. While the type of the filter, i. e., the direction of the corresponding filter vector $\mathbf{J}_i$, is fixed in this simple pairwise case, the norm of the filters $\mathbf{J}_i$ is not fixed and neither are the expert parameters α_i. As for the general FoE model, these have to be chosen appropriately, which we can do by learning them from data. As before, we do this based on approximate maximum likelihood estimation using contrastive divergence learning.

4 Optical Flow Estimation

In order to demonstrate the benefits of learning the spatial statistics of optical flow, we integrate our model with a recent, competitive optical flow method and quantitatively compare the results. As a baseline algorithm we chose the combined local-global method (CLG) as proposed by Bruhn et al. (2005). Global methods typically estimate the horizontal and vertical image velocities $\mathbf{u}$ and $\mathbf{v}$ by minimizing an energy of the form (Horn and Schunck, 1981)

$$E(\mathbf{u}, \mathbf{v}) = \int_I \rho_{\mathrm{D}}(I_x \mathbf{u} + I_y \mathbf{v} + I_t) + \lambda \cdot \rho_{\mathrm{S}}\left(\sqrt{|\nabla \mathbf{u}|^2 + |\nabla \mathbf{v}|^2}\right) \,\mathrm{d}x \,\mathrm{d}y. \tag{11}$$

ρ_{D} and ρ_{S} are robust penalty functions, such as the Lorentzian (Black and Anandan, 1996); I_x, I_y, I_t denote the spatial and temporal derivatives of the image sequence. The first term in Eq. (11) is the so-called data term that enforces the brightness constancy assumption (here written as the first-order optical flow constraint). The second term is the so-called spatial term, which enforces (piecewise) spatial smoothness. Since in this model, the data term relies on a local linearization of the brightness constancy assumption, such methods are usually used in a coarse-to-fine fashion (e. g., Black and Anandan, 1996), which allows the estimation of large displacements. For the remainder, we will assume that large image velocities are handled using such a coarse-to-fine scheme with appropriate image warping (see also Appendix B).

The combined local-global method extends this framework through local averaging of the brightness constancy constraint by means of a structure tensor. This connects the method to local optical flow approaches that spatially integrate image derivatives to estimate the image velocity (Lucas and Kanade, 1981). Using $\nabla I = (I_x, I_y, I_t)^{\mathrm{T}}$ we can define a spatio-temporal structure tensor as $\mathbf{K}_\sigma(I) = G_\sigma * \nabla I \nabla I^{\mathrm{T}}$, where G_σ denotes a Gaussian convolution kernel with width σ. To compute the image derivatives, we rely on optimized $4 \times 4 \times 2$ filters as developed by Scharr (2004), unless otherwise remarked. The CLG approach estimates the optical flow by minimizing the energy

$$E_{\mathrm{CLG}}(\mathbf{w}) = \int_I \rho_{\mathrm{D}} \left(\sqrt{\mathbf{w}^{\mathrm{T}} \mathbf{K}_\sigma(I) \mathbf{w}} \right) + \lambda \cdot \rho_{\mathrm{S}} \left(|\nabla \mathbf{w}| \right) \,\mathrm{d}x\,\mathrm{d}y, \tag{12}$$

where $\mathbf{w} = (\mathbf{u}, \mathbf{v}, 1)^{\mathrm{T}}$. In the following we will only work with spatially discrete flow representations. Consequently we assume that the data term of $E_{\mathrm{CLG}}(\mathbf{w})$ is given in discrete form $E_{\mathrm{D}}(\mathbf{w})$, where $\mathbf{w}$ is now a vector of all horizontal and vertical velocities in the image (see Appendix A for details). Experimentally, the CLG approach has been shown to be one of the best currently available optical flow estimation techniques. The focus of this paper is the spatial statistics of optical flow; hence, we will only make use of the 2D-CLG approach, i. e., only two adjacent frames will be used for flow estimation.

We refine the CLG approach by using a spatial regularizer that is based on the learned spatial statistics of optical flow. Many global optical flow techniques enforce spatial regularity or "smoothness" by penalizing large spatial gradients. In Section 3 we have shown how to learn high-order Markov random field models of optical flow, which we use here as a spatial regularizer for flow estimation. Our objective is to minimize the energy

$$E(\mathbf{w}) = E_{\mathrm{D}}(\mathbf{w}) + \lambda \cdot E_{\mathrm{FoE}}(\mathbf{w}). \tag{13}$$

Since $E_{\mathrm{FoE}}(\mathbf{w})$ is non-convex, minimizing Eq. (13) is generally difficult. Depending on the choice of the robust penalty function, the data term may in fact be non-convex, too. We will not attempt to find the global optimum of the energy function, but instead perform a simple local optimization. At any local extremum of the energy it holds that

$$0 = \nabla_{\mathbf{w}} E(\mathbf{w}) = \nabla_{\mathbf{w}} E_{\mathrm{D}}(\mathbf{w}) + \lambda \cdot \nabla_{\mathbf{w}} E_{\mathrm{FoE}}(\mathbf{w}). \tag{14}$$

The gradient of the spatial term is given by Eq. (8); the gradient for the data term is equivalent to the data term discretization from (Bruhn et al., 2005). As explained in more detail in Appendix A, we can rewrite

Method with spatial and data terms	mean AAE	AAE std. dev.
(1) Quadratic + quadratic	3.75°	3.34°
(1b) Quadratic + quadratic (modified)	2.97°	2.91°
(2) Charbonnier + Charbonnier	2.69°	2.91°
(2b) Charbonnier + Charbonnier (modified)	2.21°	2.38°
(3) Charbonnier + Lorentzian	2.76°	2.72°
(3b) Charbonnier + Lorentzian (modified)	2.33°	2.32°
(4) 3×3 FoE (separate $\mathbf{u}$ & $\mathbf{v}$) + Lorentzian	2.04°	2.31°
(5) 5×5 FoE (separate $\mathbf{u}$ & $\mathbf{v}$) + Lorentzian	2.08°	2.41°
(6) $3 \times 3 \times 2$ FoE (joint $\mathbf{u}$ & $\mathbf{v}$) + Lorentzian	1.98°	2.12°

Table 1: Results on synthetic test data set of 36 image sequences generated using our optical flow database: Average angular error (AAE) for best parameters.

the gradient of the data term using $\nabla_{\mathbf{w}} E_{\mathrm{D}}(\mathbf{w}) = \mathbf{A}_{\mathrm{D}}(\mathbf{w})\mathbf{w} + \mathbf{b}_{\mathrm{D}}(\mathbf{w})$, where $\mathbf{A}_{\mathrm{D}}(\mathbf{w})$ is a large, sparse matrix and $\mathbf{b}_{\mathrm{D}}(\mathbf{w})$ is a vector, both of which depend on $\mathbf{w}$. Similarly, we can rewrite the gradient of the spatial term as $\nabla_{\mathbf{w}} E_{\mathrm{FoE}}(\mathbf{w}) = \mathbf{A}_{\mathrm{FoE}}(\mathbf{w})\mathbf{w}$, where the matrix $\mathbf{A}_{\mathrm{FoE}}(\mathbf{w})$ is sparse and depends on $\mathbf{w}$. This is also shown in more detail in Appendix A. The gradient constraint in Eq. (14) can thus be rewritten as

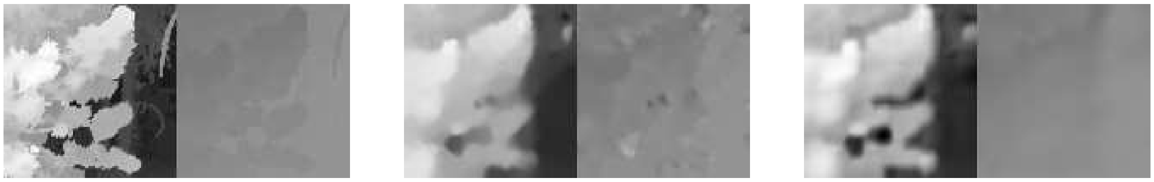
$$[\mathbf{A}_{\mathrm{D}}(\mathbf{w}) + \lambda \cdot \mathbf{A}_{\mathrm{FoE}}(\mathbf{w})]\, \mathbf{w} = -\mathbf{b}_{\mathrm{D}}(\mathbf{w}). \tag{15}$$

In order to solve for $\mathbf{w}$, we make Eq. (15) linear by keeping $\mathbf{A}_{\mathrm{D}}(\mathbf{w}) + \lambda \cdot \mathbf{A}_{\mathrm{FoE}}(\mathbf{w})$ and $\mathbf{b}_{\mathrm{D}}(\mathbf{w})$ fixed, and solve the resulting linear equation system using a standard technique (Davis, 2004). After finding a new estimate for $\mathbf{w}$, we linearize Eq. (15) around the new estimate and repeat this linearization procedure until a fixed point of the non-linear equation system is reached.

For all the experiments conducted here, we select the smoothness weight λ by hand as described in more detail alongside each experiment. Recently, Krajsek and Mester (2006) proposed a technique for automatically selecting this parameter for each input sequence, however did so for simpler models of spatial smoothness. It seems worthwhile to study whether their technique can be extended to high-order flow priors as introduced here.

4.1 Experimental evaluation

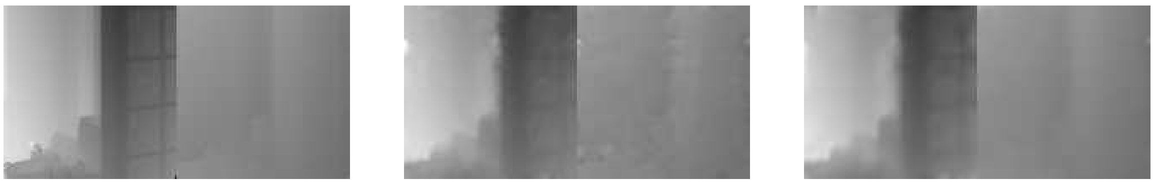
To evaluate the proposed method, we performed a series of experiments with both synthetic and real data. The quantitative evaluation of optical flow techniques suffers from the problem that only a few image sequences with ground truth optical flow data are available. The first part of our evaluation thus relies on synthetic test data. To provide realistic image texture we randomly sampled 36 (intensity) images from a database of natural images (Martin et al., 2001) and cropped them to 140×140 pixels. The images were warped with randomly sampled, synthesized flow from a separate test set, and the final image pair was



(a) Ground truth flow.

(b) Standard 2D-CLG ($AAE = 10.86°$).

(c) 2D-CLG with FoE spatial term ($AAE = 8.92°$).



(d) Ground truth flow.

(e) Standard 2D-CLG ($AAE = 1.57°$).

(f) 2D-CLG with FoE spatial term ($AAE = 1.46°$).

(g) Ground truth flow.

(h) Standard 2D-CLG ($AAE = 1.59°$).

(i) 2D-CLG with FoE spatial term ($AAE = 1.83°$).

Figure 10: Optical flow estimation for 3 representative test scenes from 36 used for evaluation. The horizontal motion is shown on the left side of each flow field; the vertical motion is displayed on the right. The middle column shows the flow estimation results from algorithm *(2b)*, the right column shows the flow estimation results from algorithm *(4)*. The bottom row shows an example where algorithm *(4)* does not perform as well as algorithm *(2b)*.

cropped to 100×100 pixels. The flow fields for testing were generated from range data and camera motion using the same algorithm as used to generate the training data; the details of this procedure are described in Section 2.1. While not identical, the testing data is thus built on the same set of assumptions such as rigid scenes and no independent motion.

We ran 6 different algorithms on all the test image pairs: *(1)* The 2D-CLG approach with quadratic data and spatial terms; *(2)* The 2D-CLG approach with Charbonnier data and spatial terms as used in (Bruhn et al., 2005); *(3)* The 2D-CLG approach with Lorentzian data term and Charbonnier spatial term; algorithms *(4-6)* are all based on the 2D-CLG approach with Lorentzian data term and a FoE spatial term. *(4)* uses a 3×3 FoE model with separate models for $\mathbf{u}$ and $\mathbf{v}$; *(5)* uses a 5×5 FoE model again with separate models for $\mathbf{u}$ and $\mathbf{v}$; *(6)* uses a $3 \times 3 \times 2$ FoE model that jointly models $\mathbf{u}$ and $\mathbf{v}$. Two different variants of algorithms *(1-3)* were used; one using the discretization described in (Bruhn et al., 2005), and the other using the (correct) discretization actually used by the authors in their experiments (Bruhn, 2006). Details are provided in Appendix B. The corrected versions are marked as *(1b)*, *(2b)*, and *(3b)* respectively.

The Charbonnier robust error function has the form $\rho(x) = 2\beta^2\sqrt{1 + x^2/\beta^2}$, where β is a scale parameter. The Lorentzian robust error function is related to the t-distribution and has the form $\rho(x) = \log(1+\frac{1}{2}(x/\beta)^2)$, where β is its scale parameter. For all experiments in this paper, we chose a fixed integration scale for the structure tensor ($\sigma = 1$). For methods *(2)* and *(3)* we tried $\beta \in \{0.05, 0.01, 0.005\}$[4] for both the spatial and the data term and report the best results. For methods *(3-6)* we fixed the scale of the Lorentzian for the data term to $\beta = 0.5$. For each method we chose a set of 10 candidate λ values (in a suitable range), which control the relative weight of the spatial term. Each algorithm was run using each of the candidate λ values. Using a simple MATLAB implementation, running method *(1)* on one frame pair takes on average $1.8s$, method *(2)* takes $15.7s$, and method *(4)* averages at $218.9s$. The increase in computational effort from using the FoE model stems from the fact that the linear equation systems corresponding to Eq. (15) are less sparse.

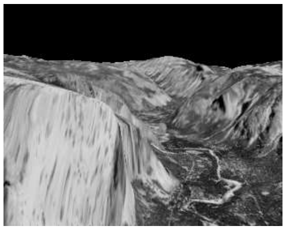
We measure the performance of the various methods using the average angular error (Barron et al., 1994)

$$AAE(\mathbf{w}_e, \mathbf{w}_t) = \arccos\left(\frac{\mathbf{w}_e^{\mathrm{T}}\mathbf{w}_t}{||\mathbf{w}_e|| \cdot ||\mathbf{w}_t||}\right), \tag{16}$$

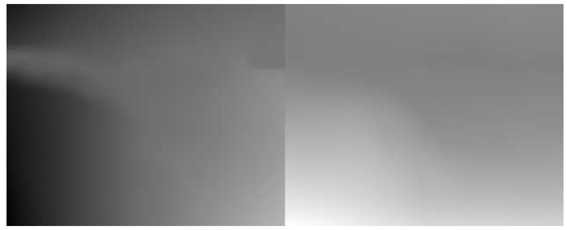
where $\mathbf{w}_e = (\mathbf{u}_e, \mathbf{v}_e, 1)$ is the estimated flow and $\mathbf{w}_t = (\mathbf{u}_t, \mathbf{v}_t, 1)$ is the true flow. We exclude 5 pixels around the boundaries when computing the AAE. Table 1 shows the mean and the standard deviation of the average angular error for the whole test data set (36 flow fields). The error is reported for the λ value that gave the lowest average error on the whole data set, i. e., the parameters are not tuned to each individual test case to emphasize generality of the method. Figure 10 shows 3 representative results from this benchmark. We can see that the FoE model recovers smooth gradients in the flow rather well, and that the resulting flow fields look less "noisy" than those recovered by the baseline CLG algorithm. On the other hand, motion boundaries sometimes appear blurred in the results from the FoE model. We presume that this is in part due to local minima in the non-convex energy from Eq. (13). Future work should address the problem of inference with such complex non-convex energies as discussed in more detail in Section 5.

The quantitative results in Table 1 show that the FoE flow prior improves the flow estimation error on this synthetic test database. A Wilcoxon signed rank test for zero median shows that the difference between the results for *(2b)* and *(4)* is statistically significant at a 95% confidence level ($p = 0.0015$). We furthermore find that a FoE prior with 5×5 cliques does not provide superior performance over a 3×3 model; the performance in fact deteriorates slightly, which may be attributable to local minima in the inference process, but the difference is not statistically significant (using the same test as above). Given that the derivatives of horizontal and vertical flow are largely independent (see Section 2.3), it is not surprising that the joint model for horizontal and vertical flow only gives a small performance advantage, which is furthermore not statistically significant. In contrast to methods *(2)* and *(3)* all FoE priors do not require any manual tuning of the parameters of the prior; instead the parameters are learned from training data. Only the λ value requires tuning for all 6 techniques.

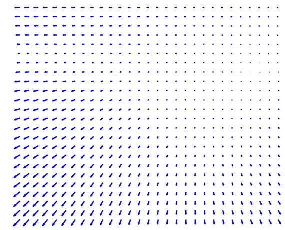
[4] This parameter interval is suggested in (Bruhn et al., 2005).



(a) Frame 8 from image sequence.



(b) Estimated optical flow with separate **u** and **v** components. Average angular error 1.43°.

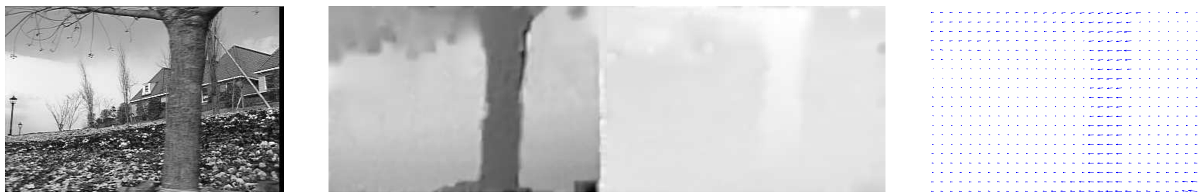


(c) Estimated optical flow as vector field.

Figure 11: Optical flow estimation: Yosemite fly-through.

Learned pairwise model. We also evaluated the performance of a learned pairwise model of optical flow based on Eq. (10) with t-distribution potentials, roughly equivalent to the model of (Black and Anandan, 1996). We trained the pairwise model using contrastive divergence as described in Section 3, and estimated flow as above; aside from the learned prior, the setup was the same as for algorithm *(3b)*. Using the same benchmark set we found that the learned pairwise model lead to worse results than the hand-tuned CLG model (the average AAE rose to 4.28°). Compared to the convex prior employed in *(3b)* the learned prior model is non-convex, which may cause the algorithm to suffer from local optima in the objective. To investigate this further, we initialized both algorithms with the ground truth flow to compare their behavior; both algorithms were run with the best parameters as found based on the regular (non-ground truth) initialization. We found that both models perform almost equally well in this case (0.88° AAE for the learned pairwise model compared to 0.91° for method *(3b)*). This illustrates the need for developing better inference techniques for optical flow estimation with non-convex models.

Other sequences. In a second experiment, we learned a FoE flow prior for the Yosemite sequence (Barron et al., 1994), a computer generated image sequence (version without the cloudy sky). First we trained the FoE prior on the ground truth data for the Yosemite sequence, omitting frames 8 and 9 which were used for evaluation. To facilitate comparisons with other methods, we used the image derivative filters given in (Bruhn et al., 2005) for this particular experiment. Estimating the flow with the learned model and the same data term as above gives an average angular error of 1.43° (standard deviation 1.51°). This is 0.19° better than the result reported for the standard two frame CLG method (see (Bruhn et al., 2005)), and 0.15° better than the result of Mémin and Pérez (2002), which is as far as we are aware currently the best result for a two frame method. While training on the remainder of the Yosemite sequence may initially seem unfair, most other reported results for this sequence rely on tuning the method's parameters so that one obtains the best results on a particular frame pair. Figure 11 shows the computed flow field. We can see that it seems rather smooth, but given that the specific training data contains only very few discontinuities, this is not very surprising. In fact, changing the λ parameter of the algorithm so that edges start to appear leads to numerically inferior results.



(a) Frame 1 from image sequence.

(b) Estimated optical flow with separate **u** and **v** components.

(c) Estimated flow.

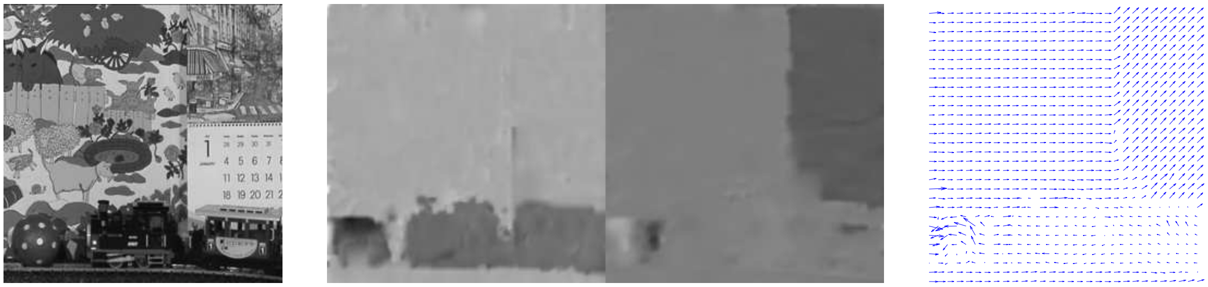
Figure 12: Optical flow estimation: Flower garden sequence.

An important question for a learned prior model is how well it generalizes. To evaluate this, we used the model trained on our synthetic flow data to estimate the Yosemite flow. Using the same parameters described above, we found that the accuracy of the flow estimation results decreased to $1.60°$ average angular error (standard deviation $1.63°$). While this result is not as good as that from the Yosemite-specific model, it still slightly outperforms the regular 2D-CLG approach, while not requiring any manual tuning of the prior parameters. This raises two interesting questions: Is the generic training database not fully representative of the kinds of geometries or motions that occur in the Yosemite sequence? Or is it the case that the Yosemite sequence is not representative of the kinds of image motions that occur naturally? This further suggests that future research should be devoted to determining what constitutes generic optical flow and to developing more comprehensive benchmark data sets for flow estimation. In particular, a better data set would include realistic models of the appearance of the scene in addition to the motion. Moreover, this suggests that particular care must be taken when designing a representative optical flow database.

In a final experiment, we evaluated the FoE flow prior on two real image sequences. Figure 12 shows the first frame from a "flower garden" sequence as well as the estimated flow. The sequence has two dominant motion layers, a tree in the foreground and a background, with different image velocities. Figure 13 shows one frame and flow estimation results for a "mobile & calendar" sequence (downsampled to 256×256), which exhibits independent motion of the calendar, the train, and the ball in front of the train. In both cases we applied the FoE model as trained on the synthetic flow database and used the parameters as described above for model *(4)*. Both figures show that the obtained flow fields qualitatively capture the motion and object boundaries well, even in case of independent object motion, which was not present in the training set.

5 Conclusions and Future Work

We have presented a novel database of optical flow as it arises when realistic scenes are captured with a hand-held or car-mounted video camera. This database allowed us to study the spatial and temporal statistics of optical flow. We found that "natural" optical flow is mostly slow, but sometimes fast, as well as mostly smooth, but sometimes discontinuous. Furthermore, the derivative statistics were found to be very heavy-tailed, which likely explains the success of previous flow estimation techniques based on robust



(a) Frame 1 from image sequence.

(b) Estimated optical flow with separate **u** and **v** components.

(c) Estimated flow.

Figure 13: Optical flow estimation: Mobile & calendar sequence.

potential functions. Moreover, the flow data enabled us to learn prior models of optical flow using the *Fields-of-Experts* framework. We have integrated the learned FoE flow prior into a recent, accurate optical flow algorithm and obtained statistically significant accuracy improvements on a synthetic test set. While our experiments suggest that the training database may not yet be representative of the image motion in certain sequences, we believe that this is an important step towards studying and learning the spatial statistics of optical flow.

There are many opportunities for future work that build on the proposed prior and the database of flow fields. For example, Calow et al. (2004) point out that natural flow fields are inhomogeneous; for example, in the case of human motion, the constant presence of a ground plane produces quite different flow statistics in the lower portion of the image than in the upper portion. The work of Torralba (2003) on estimating global scene classes could be used to apply scene and region appropriate flow priors. In this work we proposed a homogeneous flow prior but it is also possible, with sufficient training data, to learn an inhomogeneous FoE model. Moreover, it may be fruitful to study the scale invariance properties of optical flow and relate them to the scale invariance of natural scene depth. Scale invariance could, for example, be modeled using hierarchical multi-scale models.

It may also be desirable to learn application-specific flow priors (e. g., for automotive applications). This suggests the possibility of learning multiple categories of flow priors (cf. Torralba and Oliva, 2003) and using these to classify scene motion for applications in video databases.

So far, inference with this optical flow model has relied on iteratively solving linear equation systems, which suffers from the fact that the data and spatial term are non-convex. In case of pairwise MRFs for regularization, other work has relied on annealing techniques to partly overcome this problem (Black and Anandan, 1996). However, from recent research on stereo problems, it has become apparent that better inference techniques such as graph cuts or belief propagation are particularly good at minimizing the corresponding energies (Szeliski et al., 2006). We think that this along with the findings in this paper suggest the need for non-convex regularization models in optical flow computation. It is clear that this complicates the computational techniques needed to recover optical flow, but the success in stereo problems and some initial successes in optical flow (Barbu and Yuille, 2004) suggest that research into better inference techniques for optical flow should be a fruitful area for future work.

A natural extension of our work is the direct recovery of structure from motion. We can exploit our training set of camera motions to learn a prior over 3D camera motions and combine this with a spatial prior learned from the range imagery. The prior over 3D motions may help regularize the difficult problem of recovering camera motion given the narrow field of view and small motions present in common video sequences.

Future work must also consider the statistics of independent, textural, and non-rigid motion. Here obtaining ground truth is more problematic. Possible solutions involve obtaining realistic synthesized sequences from the film industry or hand-marking regions of independent motion in real image sequences.

Finally, a more detailed analysis of motion boundaries is warranted. In particular, our current flow prior does not explicitly encode information about the occluded/unoccluded surfaces or the regions of the image undergoing deletion/accretion. Future work may also explore the problem of jointly learning motion and occlusion boundaries using energy-based models such as Fields of Experts (cf. Ross and Kaelbling, 2005).

Acknowledgements We would like to thank David Capel (2d3) and Andrew Zisserman for providing camera motion data, Andrés Bruhn for providing additional details about the algorithm in (Bruhn et al., 2005), John Barron for clarifying details about the Yosemite sequence ground truth, Étienne Mémin for providing the "Mobile & Calendar" sequence, Hanno Scharr for helpful discussions on optical flow and optimal filters, and Frank Wood for helpful discussions on energy-based models. This research was supported by Intel Research, NSF IIS-0535075, and NIH-NINDS R01 NS 50967-01 as part of the NSF/NIH Collaborative Research in Computational Neuroscience Program.

A Derivation Details

This section describes in more detail how the non-linear equation system for flow estimation in Eq. (15) can be derived.

A.1 Data term

Starting from the continuous data term of the CLG approach (Bruhn et al., 2005)

$$\int_I \rho_{\mathrm{D}} \left(\sqrt{\mathbf{w}^{\mathrm{T}} \mathbf{K}_\sigma(I) \mathbf{w}} \right) \, \mathrm{d}x \, \mathrm{d}y,$$

we first obtain a simple discretization of the data term energy

$$E_{\mathrm{D}}(\mathbf{w}) = \sum_{i=1}^{K} \rho_{\mathrm{D}} \left(\sqrt{(u_i, v_i, 1) \cdot \mathbf{K}_i \cdot (u_i, v_i, 1)^{\mathrm{T}}} \right). \tag{17}$$

Here, $\mathbf{w} = (u_1, \ldots, u_K, v_1, \ldots, v_K)^{\mathrm{T}}$ is the stacked vector of all horizontal velocities $\mathbf{u}$ and all vertical velocities $\mathbf{v}$; $\mathbf{K}_i$ is the structure tensor $\mathbf{K}_\sigma(I)$ evaluated at pixel i. The structure tensor $\mathbf{K}_i$ is a 3×3 tensor with entries K_{kli}. We can now take partial derivatives with respect to u_i and v_i:

$$\begin{aligned} \frac{\partial}{\partial u_i} E_{\mathrm{D}}(\mathbf{w}) &= \tilde{\rho}_{\mathrm{D}} \left(\sqrt{(u_i, v_i, 1) \cdot \mathbf{K}_i \cdot (u_i, v_i, 1)^{\mathrm{T}}} \right) (K_{11i} u_i + K_{12i} v_i + K_{13i}) \qquad (18) \\ \frac{\partial}{\partial v_i} E_{\mathrm{D}}(\mathbf{w}) &= \tilde{\rho}_{\mathrm{D}} \left(\sqrt{(u_i, v_i, 1) \cdot \mathbf{K}_i \cdot (u_i, v_i, 1)^{\mathrm{T}}} \right) (K_{21i} u_i + K_{22i} v_i + K_{23i}), \qquad (19) \end{aligned}$$

where we define $\tilde{\rho}_{\mathrm{D}}(y) = \rho'_{\mathrm{D}}(y)/y$. Note that this formulation is equivalent to the data term parts of Eqs. (38) and (39) in (Bruhn et al., 2005).

Finally, we set the partial derivatives to 0 and regroup the terms in matrix-vector form as

$$0 = \nabla_{\mathbf{w}} E_{\mathrm{D}}(\mathbf{w}) = \mathbf{A}_{\mathrm{D}}(\mathbf{w})\mathbf{w} + \mathbf{b}_{\mathrm{D}}(\mathbf{w}). \qquad (20)$$

In this notation $\mathbf{A}_{\mathrm{D}}(\mathbf{w})$ is a large, sparse matrix that depends on $\mathbf{w}$, and the matrix-vector product $\mathbf{A}_{\mathrm{D}}(\mathbf{w})\mathbf{w}$ is used to express all terms of Eqs. (18) and (19) that contain K_{11i}, K_{12i}, K_{21i}, and K_{22i}. $\mathbf{b}_{\mathrm{D}}(\mathbf{w})$ is a vector that also depends on $\mathbf{w}$ and contains all terms with K_{13i} and K_{23i}.

A.2 Spatial term

From Eq. (8), we know that we can write the gradient of the energy of the FoE spatial term with respect to the flow field component as follows:

$$\nabla_{\mathbf{x}} E_{\mathrm{FoE}}(\mathbf{x}) = -\sum_{i=1}^{N} \mathbf{J}_i^- * \xi_i(\mathbf{J}_i * \mathbf{x}).$$

Because convolution is a linear operation, we can express the convolution $\mathbf{J}_i * \mathbf{x}$ as a matrix-vector product of a filter matrix $\mathbf{F}_i$ with the vectorized flow field component $\mathbf{x}$. Similarly, convolution with the mirrored filter $\mathbf{J}_i^-$ can be expressed as a matrix-vector product; we call the corresponding filter matrix $\mathbf{G}_i$. If we assume that $\xi_i(\mathbf{y})$ is the element-wise application of the non-linearity ξ_i to the elements of $\mathbf{y}$, we can rewrite the preceding equation as

$$\nabla_{\mathbf{x}} E_{\mathrm{FoE}}(\mathbf{x}) = -\sum_{i=1}^{N} \mathbf{G}_i \cdot \xi_i(\mathbf{F}_i \cdot \mathbf{x}). \qquad (21)$$

Furthermore, we can exploit the form of the non-linearity and express it as $\xi_i(y) = \zeta_i(y) \cdot y$. If we assume that ξ_i and ζ_i can be applied to vectors in an element-wise fashion, we can express the non-linearity for vectors as follows:

$$\xi_i(\mathbf{y}) = \mathrm{diag}\{\zeta_i(\mathbf{y})\} \cdot \mathbf{y}, \qquad (22)$$

where $\mathrm{diag}\{\mathbf{z}\}$ denotes a diagonal matrix with the entries of vector $\mathbf{z}$ on its diagonal. When combining this with the previous step, we obtain that the gradient of the FoE spatial term can be written as

$$\nabla_{\mathbf{x}} E_{\mathrm{FoE}}(\mathbf{x}) = -\sum_{i=1}^{N} \mathbf{G}_i \cdot \mathrm{diag}\{\zeta_i(\mathbf{F}_i \cdot \mathbf{x})\} \cdot \mathbf{F}_i \cdot \mathbf{x} = \left[-\sum_{i=1}^{N} \mathbf{G}_i \cdot \mathrm{diag}\{\zeta_i(\mathbf{F}_i \cdot \mathbf{x})\} \cdot \mathbf{F}_i \right] \mathbf{x}. \qquad (23)$$

Note that the term in brackets is a large matrix that depends on $\mathbf{x}$, which we denote as $\mathbf{A}_{\mathrm{FoE}}(\mathbf{x})$. It thus follows that $\nabla_{\mathbf{x}} E_{\mathrm{FoE}}(\mathbf{x}) = \mathbf{A}_{\mathrm{FoE}}(\mathbf{x})\mathbf{x}$.

B Incremental Flow Estimation using Pyramids

Global techniques for optical flow such as the ones presented here often use multi-resolution methods based on image pyramids to overcome the limitations of the local linearization of the optical flow constraint, and to estimate flows with large displacements. One issue that arises when employing incremental multi-resolution schemes for optical flow estimation is how to properly take into account the flow estimate from coarser scales. Usually the input sequence is pre-warped with the flow as estimated at a coarser scale. This result is then incrementally refined at a finer scale (Black and Anandan, 1996). The data term only considers the incremental flow and is thus easy to deal with; the spatial term on the other hand has to consider the combination of the incremental flow and the estimation from the coarser scales, since the spatial term is a prior model of the combined flow and not of the incremental flow.

To that end, we combine the flow estimate $\mathbf{w}$ from the next coarser scale with the incremental flow $\Delta\mathbf{w}$ and obtain the energy

$$E(\Delta\mathbf{w}) = E_{\mathrm{D}}(\Delta\mathbf{w}) + \lambda \cdot E_{\mathrm{FoE}}(\mathbf{w} + \Delta\mathbf{w}) \tag{24}$$

that is to be minimized with respect to $\Delta\mathbf{w}$. As in Section 4, we set the gradient to zero and rewrite the gradient terms as $\nabla_{\mathbf{w}} E_{\mathrm{D}}(\Delta\mathbf{w}) = \mathbf{A}_{\mathrm{D}}(\Delta\mathbf{w})\Delta\mathbf{w} + \mathbf{b}_{\mathrm{D}}(\Delta\mathbf{w})$ and $\nabla_{\mathbf{w}} E_{\mathrm{FoE}}(\mathbf{w}+\Delta\mathbf{w}) = \mathbf{A}_{\mathrm{FoE}}(\mathbf{w}+\Delta\mathbf{w})\cdot(\mathbf{w}+\Delta\mathbf{w})$. The resulting equation system can be written as

$$\left[\mathbf{A}_{\mathrm{D}}(\Delta\mathbf{w}) + \lambda \cdot \mathbf{A}_{\mathrm{FoE}}(\mathbf{w} + \Delta\mathbf{w})\right] \Delta\mathbf{w} = -\mathbf{b}_{\mathrm{D}}(\Delta\mathbf{w}) - \lambda \cdot \mathbf{A}_{\mathrm{FoE}}(\mathbf{w} + \Delta\mathbf{w})\mathbf{w}, \tag{25}$$

which we solve iteratively as before.

The 2D-CLG discretization given in (Bruhn et al., 2005) (Eqs. (42) and (43)) is missing a term corresponding to $-\lambda \cdot \mathbf{A}_{\mathrm{FoE}}(\mathbf{w} + \Delta\mathbf{w})\mathbf{w}$ (on the RHS of Eq. (25)). This is problematic, because without this term the regularization is not properly applied to the combined flow. The original implementor of (Bruhn et al., 2005) has confirmed that this is simply an oversight in the manuscript (Bruhn, 2006), and that the original implementation is in fact based on the correct discretization. Our implementation of the algorithm confirms this in case of the Yosemite sequence. In the notation of (Bruhn et al., 2005), Eq. (42) can be corrected as follows:

$$0 = \sum_{j \in \mathcal{N}(i)} \frac{\psi'^{m}_{2i} + \psi'^{m}_{2j}}{2} \frac{u^m_j + \delta u^m_j - u^m_i - \delta u^m_i}{h^2} - \frac{\psi'^{m}_{1i}}{\alpha}(J^m_{11i}\delta u^m_i + J^m_{12i}\delta v^m_i + J^m_{13i}); \tag{26}$$

Eq. (43) can be adapted accordingly. For completeness, Section 4.1 gives results for both discretizations.

References

2d3 Ltd. boujou. http://www.2d3.com, 2002.

L. Alvarez, J. Weickert, and J. Sánchez. Reliable estimation of dense optical flow fields with large displacements. *Int. J. Comput. Vision*, 39(1):41–56, Aug. 2000.

A. Barbu and A. Yuille. Motion estimation by Swendsen-Wang cuts. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, vol. 1, pp. 754–761, June 2004.

J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *Int. J. Comput. Vision*, 12(1):43–77, Feb. 1994.

R. Ben-Ari and N. Sochen. A general framework and new alignment criterion for dense optical flow. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, vol. 1, pp. 529–536, June 2006.

B. Y. Betsch, W. Einhäuser, K. P. Körding, and P. König. The world from a cat's perspective - Statistics of natural videos. *Biological Cybernetics*, 90(1):41–50, Jan. 2004.

M. J. Black and P. Anandan. Robust dynamic motion estimation over time. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, pp. 296–302, June 1991.

M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Comput. Vis. Image Und.*, 63(1):75–104, Jan. 1996.

A. Bruhn. Personal communication, 2006.

A. Bruhn, J. Weickert, and C. Schnörr. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *Int. J. Comput. Vision*, 61(3):211–231, Feb. 2005.

D. Calow, N. Krüger, F. Wörgötter, and M. Lappe. Statistics of optic flow for self-motion through natural scenes. In U. Ilg, H. Bülthoff, and H. Mallot, eds., *Dynamic Perception*, pp. 133–138, 2004.

D. Cremers and S. Soatto. Motion competition: A variational approach to piecewise parametric motion segmentation. *Int. J. Comput. Vision*, 62(3):249–265, May 2005.

T. A. Davis. A column pre-ordering strategy for the unsymmetric-pattern multifrontal method. *ACM Transactions on Mathematical Software*, 30(2):165–195, June 2004.

R. Fablet and P. Bouthemy. Non parametric motion recognition using temporal multiscale Gibbs models. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, vol. 1, pp. 501–508, Dec. 2001.

C. Fermüller, D. Shulman, and Y. Aloimonos. The statistics of optical flow. *Comput. Vis. Image Und.*, 82 (1):1–32, Apr. 2001.

D. J. Fleet, M. J. Black, Y. Yacoob, and A. D. Jepson. Design and use of linear models for image motion analysis. *Int. J. Comput. Vision*, 36(3):171–193, Feb. 2000.

D. J. Fleet, M. J. Black, and O. Nestares. Bayesian inference of visual motion boundaries. In G. Lakemeyer and B. Nebel, eds., *Exploring Artificial Intelligence in the New Millennium*, pp. 139–174. Morgan Kaufmann Pub., 2002.

U. Grenander and A. Srivastava. Probability models for clutter in natural images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(4):424–429, Apr. 2001.

F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(12):1217–1232, Dec. 1993.

G. E. Hinton. Products of experts. In *Int. Conf. on Art. Neur. Netw. (ICANN)*, vol. 1, pp. 1–6, Sept. 1999.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8): 1771–1800, Aug. 2002.

B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1–3):185–203, Aug. 1981.

J. Huang. *Statistics of Natural Images and Models*. PhD thesis, Brown University, 2000.

J. Huang, A. B. Lee, and D. Mumford. Statistics of range images. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, vol. 1, p. 1324ff, June 2000.

M. Irani. Multi-frame optical flow estimation using subspace constraints. In *IEEE Int. Conf. on Comp. Vis. (ICCV)*, vol. 1, pp. 626–633, Sept. 1999.

T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, COM-15(1):52–60, Feb. 1967.

J. Konrad and E. Dubois. Multigrid Bayesian estimation of image motion fields using stochastic relaxation. In *IEEE Int. Conf. on Comp. Vis. (ICCV)*, pp. 354–362, Dec. 1988.

K. Krajsek and R. Mester. On the equivalence of variational and statistical differential motion estimation. In *Southwest Symposium on Image Analysis and Interpretation*, pp. 11–15, Denver, Colorado, Mar. 2006.

A. B. Lee and J. Huang. Brown range image database. http://www.dam.brown.edu/ptg/brid/index.html, 2000.

A. B. Lee, D. Mumford, and J. Huang. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *Int. J. Comput. Vision*, 41(1–2):35–59, Jan. 2001.

G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck. Neural coding of naturalistic motion stimuli. *Network: Comp. Neural*, 12(3):317–329, Mar. 2001.

H. Lu and A. L. Yuille. Ideal observers for detecting motion: Correspondence noise. In *Adv. in Neur. Inf. Proc. Sys. (NIPS)*, vol. 18, pp. 827–834, 2006.

B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Int. J. Conf. on Art. Intel. (IJCAI)*, pp. 674–679, Apr. 1981.

J. Marroquin, S. Mitter, and T. Poggio. Probabilistic solutions of ill-posed problems in computational vision. *J. Am. Stat. Assoc.*, 82(397):76–89, Mar. 1987.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE Int. Conf. on Comp. Vis. (ICCV)*, vol. 2, pp. 416–423, July 2001.

É. Mémin and P. Pérez. Hierarchical estimation and segmentation of dense motion fields. *Int. J. Comput. Vision*, 46(2):129–155, Feb. 2002.

D. W. Murray and B. F. Buxton. Scene segmentation from visual motion using global optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9(2):220–228, Mar. 1987.

B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network: Comp. Neural*, 7 (2):333–339, May 1996.

N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert. Highly accurate optic flow computation with theoretically justified warping. *Int. J. Comput. Vision*, 67(2):141–158, Apr. 2006.

M. Proesmans, L. J. Van Gool, E. J. Pauwels, and A. Oosterlinck. Determination of optical flow and its discontinuities using non-linear diffusion. In J.-O. Eklundh, ed., *Eur. Conf. on Comp. Vis. (ECCV)*, vol. 801 of *Lect. Notes in Comp. Sci.*, pp. 295–304, 1994.

M. G. Ross and L. P. Kaelbling. Learning static object segmentation from motion segmentation. In *Nat. Conf. on Art. Int (AAAI)*, pp. 956–961, Menlo Park, California, 2005. AAAI Press.

S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, vol. 2, pp. 860–867, June 2005a.

S. Roth and M. J. Black. On the spatial statistics of optical flow. In *IEEE Int. Conf. on Comp. Vis. (ICCV)*, vol. 1, pp. 42–49, Oct. 2005b.

D. L. Ruderman. The statistics of natural images. *Network: Comp. Neural*, 5(4):517–548, Nov. 1994.

H. Scharr. Optimal filters for extended optical flow. In *First International Workshop on Complex Motion*, vol. 3417 of *Lect. Notes in Comp. Sci.* Springer, 2004.

H. Scharr and H. Spies. Accurate optical flow in noisy image sequences using flow adapted anisotropic diffusion. *Signal Processing: Image Communication*, 20(6):537–553, July 2005.

E. P. Simoncelli, E. H. Adelson, and D. J. Heeger. Probability distributions of optical flow. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, pp. 310–315, June 1991.

A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *J. Math. Imaging Vision*, 18(1):17–33, Jan. 2003.

R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields. In A. Leonardis, H. Bischof, and A. Prinz, eds., *Eur. Conf. on Comp. Vis. (ECCV)*, vol. 3952 of *Lect. Notes in Comp. Sci.*, pp. 16–29. Springer, 2006.

Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton. Energy-based models for sparse overcomplete representations. *J. Mach. Learn. Res.*, 4(Dec.):1235–1260, 2003.

A. Torralba. Contextual priming for object detection. *Int. J. Comput. Vision*, 53(2):169–191, July 2003.

A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Comp. Neural*, 14(2):391–412, Aug. 2003.

J. H. van Harteren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *J. Roy. Stat. Soc. B*, 265(1412): 2315–2320, Dec. 1998.

J. Weickert and C. Schnörr. Variational optic flow computation with a spatio-temporal smoothness constraint. *J. Math. Imaging Vision*, 14(3):245–255, May 2001.

Y. Weiss and E. H. Adelson. Slow and smooth: A Bayesian theory for the combination of local motion signals in human vision. Technical Report AI Memo 1624, MIT AI Lab, Cambridge, Massachusetts, Feb. 1998.

S. C. Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(11):1236–1250, Nov. 1997.

S. C. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *Int. J. Comput. Vision*, 27(2):107–126, Mar. 1998.