Optical Flow Estimation Using Learned Sparse Model

Kui Jia*
Department of Information Engineering
The Chinese University of Hong Kong
kjia@ie.cuhk.edu.hk

Xiaogang Wang
Department of Electronic Engineering
The Chinese University of Hong Kong
xgwang@ee.cuhk.edu.hk

Xiaoou Tang
Department of Information Engineering
The Chinese University of Hong Kong
xtang@ie.cuhk.edu.hk

Abstract

Optical flow estimation is a fundamental and ill-posed problem in computer vision. To recover a dense flow field, appropriate spatial constraints have to be enforced. Recent advances exploit higher order spatial regularization, and achieve the top performance on the Middlebury benchmark. In this work, we revisit learning-based approach, and propose a learned sparse model to patch-wisely regularize the flow field. In particular, our method is based on multi-scale spatial regularization, which benefits from first-order spatial regularity and our learned, higher order sparse model. To obtain accurate flow estimation, we propose a sequential optimization scheme to solve the corresponding energy minimization problem. Moreover, as the errors in intermediate flow estimates are usually dense with large variations, we further propose flow-driven and image-driven approaches to address the problem of outliers. Experiments on the Middlebury benchmark show that our method is competitive with the state-of-the-art.

1. Introduction

Optical flow estimation is one of the fundamental problems in computer vision. It concerns with computing the motion of pixels between consecutive image frames. Such a dense correspondence problem arises not just in motion estimation, but also in image registration, 3D reconstruction, and visual tracking. Similar to many computer vision techniques, optical flow is inherently ill-posed due to the aperture problem [3], i.e., using only data constraint leads to an under-determined system of equations. To recover a dense flow field, it is necessary to consider some sorts of spatial regularization to constrain the flow varying patterns in a plausible way.

In the past two decades, although the accuracy of optical flow estimation has been steadily improved, it remains challenging especially when dealing with tough situations in various natural image sequences. To this date, the challenges that dominate optical flow research includes: (1) propagating the flow into untextured regions, (2) accurate estimation at flow boundaries, and (3) preserving small-scale motion structures in the estimated flow field.

Numerous optical flow techniques have been developed to address these challenges. A large portion of them followed the seminal work of Horn and Schunck (HS) [1], which defined optical flow estimation as minimizing an energy functional. The energy functional consists of a data term that assumes image intensities (or other advanced image properties) do not change over time, and a spatial term typically inducing a (piece-wise) smooth flow field. At the time of HS, due to computational reasons, quadratic functions were used to penalize deviations in both data and spatial terms. The limitations are obvious as they cannot robustly handle data outliers and preserve discontinuities in the flow field. Instead, Black and Anandan [2] proposed to use robust, non-convex functions and greatly improved the results. Later, different robust functions [4, 5, 6, 9] have been explored that compromise between robustness, convexity and differentiability. Among them, the TV-L1 framework [11, 10] is a popular one, which used total variation (TV) like regularization and a robust L^1 norm in the data term. Based on the observation that motion discontinuities often coincide with object boundaries in images, some researchers proposed to adapt the isotropic spatial regularization to local image structures [13, 6]. For data similarity measures, more advanced ones such as image gradient [4] and normalized cross correlation [16, 17], have also been proposed to improve over image intensities.

Learning-based approaches have been attempted in optical flow literature. In particular, Roth and Black [18] learned the spatial statistics of optical flow, which was shown to be heavy-tailed. They used the learned prior model to regularize flow estimation. In their work, they considered spatial interactions up to 3×3 pixels. In [6], Sun et al. further learned statistical models of both data constancy error, and image structure-adaptive flow derivatives, resulting in a complete probabilistic model of optical

*This work is partly supported by the National Natural Science Foundation of China (Grant No. 60903115).

flow.

Recently, several works exploited higher order or non-local spatial terms [19, 7, 17], and achieved the top performance on the Middlebury optical flow benchmark [20]. Common to these approaches is a weighted non-local term, which robustly (using L^1 norm) penalizes the pairwise differences of flow vectors in a local neighborhood. The weight for each pair is determined based on bilateral filtering [21] by combining information of color similarity, spatial proximity, and/or occlusion condition. Although the state-of-the-art results were obtained, however, they are limited in: (1) still considering pairwise flow relations in a local neighborhood, (2) using purely geometric spatial priors, and (3) their regularization cannot be across flow boundaries.

In this work, we revisit learning-based approach and propose a *learned sparse model* (LSM) to regularize the flow field. Different from early attempts [18, 6], which typically learn the statistics of first-order flow derivatives, our model is higher order, i.e., we patch-wisely constrain how the flow is expected to vary across the whole field. In particular, our model is motivated by recent success in image restoration [22, 23, 24], which used sparse representation over learned, possibly over-complete image dictionaries (or basis functions), and achieved the state-of-the-art in image denoising and demosaicking [24]. In this work, we consider learning an optical flow dictionary that adapts to the training ground truth flow fields. For spatial regularization, our assumption is that each flow patch can be encoded via a sparse representation over the learned over-complete flow dictionary. Note that by doing so, we actually solve the aperture problem in a way distinct from [1, 25]. Compared with [1, 25], our model does not need to regularize smooth motions and motion discontinuities separately.

Different from situations in image denoising, the noises in intermediate flow estimates are in general dense with large variations. We further propose a multi-scale spatial regularizer, which benefits from first-order spatial regularity and the learned, higher order sparse model. Multi-scale spatial regularization stabilizes the estimation process, and enable our model to be easily embedded in a coarse-to-fine/warping framework [26, 27], to cope with large motions. Together with a robust data term, flow field recovery is formulated as an energy minimization problem. We propose to decompose the optimization into a sequence of simpler ones, with each alternating in satisfying data constraints, and spatial regularization via sparse coding. Moreover, except for dense noises, some intermediate flow estimates can be completely corrupted and become outliers, which degrade the performance of learned sparse model. In this work, we also propose flow-driven and image-driven approaches to address the problem of outliers. Experiments on the Middlebury benchmark show that our method is competitive with the state-of-the-art.

Note that we are not the first to introduce sparsity priors into optical flow estimation. In [28], Shen and Wu assumed that flow field can be estimated by finding its sparsest representation in other domains. They showed plausible results in subsampled image frames with small motions. Our method is different from [28] in the following aspects.

1. We propose a learned sparse model, and get improved performance over generic ones such as wavelet or DCT, which were used in [28].

2. To robustify higher order spatial regularization, we propose flow-driven and image-driven approaches to address the problem of outliers. Experiments show the effectiveness.

3. We propose multi-scale spatial regularization and a sequential optimization scheme. We adapt the learned sparse model in a coarse-to-fine/warping framework, and obtain accurate results on the original frame size with large motions. Our results are competitive with the state-of-the-art.

The rest of this paper is organized as follows. In Section 2, we present in details our learned sparse model and its multi-scale extension. Section 3 introduces robust higher-order spatial regularization. Our sequential optimization scheme will be explained in Section 4. Section 5 presents experiments, followed by conclusion and future works in Section 6.

2. Flow field regularization using the learned sparse model

Optical flow estimation is commonly formulated as an energy minimization problem. The objective function is

$$E(\mathbf{u}) = E_D(\mathbf{u}) + \lambda E_S(\mathbf{u}), \tag{1}$$

where $\mathbf{u} = [u, v]^\top \in \Re^{2N}$ is the vectorized flow field to be estimated, N is the number of image pixels, and λ is a regularization parameter. [1] For a given $\mathbf{u}$, the data term $E_D(\mathbf{u}) = \sum_{\mathbf{x}} \psi_D(I_1(\mathbf{x}) - I_2(\mathbf{x}+\mathbf{u}_\mathbf{x}))$ measures the similarity between two consecutive image frames I_1 and I_2, ψ_D is a properly chosen penalty function, and $\mathbf{x} = [x, y]^\top$ indexes the image coordinates. When the unknown motion $\mathbf{u}$ is in a small proximity of a given point $\mathbf{u}^0$, we can linearize the image residual $\rho(\mathbf{x}) = I_1(\mathbf{x}) - I_2(\mathbf{x}+\mathbf{u}_\mathbf{x})$, which leads to the classical optical flow equation $\rho(\mathbf{x}) = \nabla I_2^\top(\mathbf{u}_\mathbf{x} - \mathbf{u}_\mathbf{x}^0) + I_t$, where ∇I_2 denotes the horizontal and vertical partial derivatives at $\mathbf{x} + \mathbf{u}_\mathbf{x}^0$, and $I_t = I_2(\mathbf{x} + \mathbf{u}_\mathbf{x}^0) - I_1(\mathbf{x})$ is the temporal derivative. Since optical flow is highly underdetermined if only based on the assumption of intensity

[1] Throughout this paper, we will use spatially discrete and vectorized representation to denote the optical flow field.

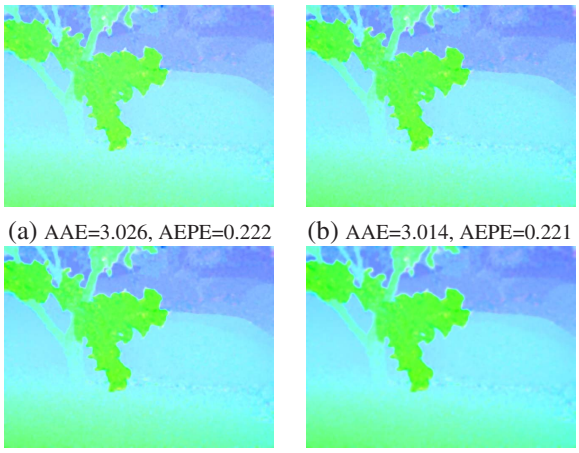
constancy, i.e., it suffers from the aperture problem. Additional constraints are needed in order to obtain a dense and accurate flow field. This brings the spatial term $E_S(\mathbf{u})$ in, which essentially constrains how the flow is expected to vary across the image. Originating from the HS model [1], most of the spatial terms proposed in literature take the form like $E_S(\mathbf{u}) = \sum_{\mathbf{x}} \psi_S(\nabla \mathbf{u}_{\mathbf{x}})$, which favors a smooth flow field, and is edge-preserving by using some robust penalty function ψ_S [2]. Alternatively, Lucas and Kanade [25] addressed the aperture problem by assuming that the flow vectors are constant in a local neighborhood. However, this assumption fails in regions with multiple motions.

As introduced in Section 1, Shen and Wu [28] recently proposed to use a sparsity prior to regularize the flow field. They assumed a flow patch can be described via a sparse representation over some basis functions. From the perspective of compressive sensing, this amounts to recover a dense flow field from much fewer measurements, thus solving the aperture problem. As pointed out in [28], although the flow patterns may be complex and varying across the whole field, they are much simpler compared with those of natural images. By assuming the sparsity of local flow patches, ideally we can unify the different treatments of smooth or discontinuous motions, and various motion models such as affine transformation and rotation.

In [28], generic basis functions (dictionaries) such as Wavelet and DCT are used for sparse coding. Motivated by the success of learned dictionaries over off-the-shelf ones in image restoration [22, 23, 24], in this work, we consider learning an adapted, possibly over-complete, optical flow dictionary using training ground truth flow fields. We expect through learning, the dictionary can encode more flow statistics and as a consequence, leads to a sparser and more accurate representation. Specifically, we propose to regularize the flow field using a *learned sparse model*. Adapting the sparsity assumption with the learned dictionary in an energy model, we get

$$E(\mathbf{u}) = \sum_{\mathbf{x}} \psi_D\big(\rho(\mathbf{x})\big) + \lambda \|T^h_{\mathbf{x}}\mathbf{u} - \mathbf{D}^h \mathbf{a}^h_{\mathbf{x}}\|_2^2 + \beta \psi_S(\mathbf{a}^h_{\mathbf{x}}), \tag{2}$$

where $T^h_{\mathbf{x}} \in \Re^{2n \times 2N}$ is a binary operator that extracts the flow patch centering at position $\mathbf{x}$ from $\mathbf{u}$, n is the size of the patch. $\mathbf{D}^h = [\mathbf{D}^h_u \, \mathbf{0}; \mathbf{0} \, \mathbf{D}^h_v] \in \Re^{2n \times 2p}$ represents the learned flow dictionary with the dictionary size p, and $\mathbf{a}^h_{\mathbf{x}} \in \Re^{2p}$ is the sparse coefficient vector when decomposing $T^h_{\mathbf{x}}\mathbf{u}$ on $\mathbf{D}^h$, β is a sparsity inducing parameter. Here we want to emphasize that, different from most of existing first-order spatial terms that typically penalize the difference between neighboring flow vectors, and some recently proposed higher order spatial terms that adaptively and robustly penalize the difference among non-local flow vectors in an expanded neighborhood [19, 17, 7], the spatial term in (2) assumes some prior on the spatially varying pattern of local flow patches, and such a pattern can be sparsely encoded and reconstructed by the learned flow dictionary. In this work, we follow [7] and use a generalized Charbonnier data penalty function $\psi_D(x) = (x^2 + \epsilon^2)^\gamma$, and set $\gamma = 0.45$ to make it slightly non-convex. ϵ is fixed as 0.001. The spatial penalty can be chosen as $\psi_S(\cdot) = \|\cdot\|_1$.



(a) AAE=3.026, AEPE=0.222 (b) AAE=3.014, AEPE=0.221

(c) AAE=2.828, AEPE=0.206 (d) AAE=2.775, AEPE=0.198

Figure 1. Effectiveness of the learned sparse model on the "Grove2" sequence of Middlebury training set. (a) Initialization. (b) Result using HS method [1]. (c) Result using higher order sparse model with a DCT dictionary. (d) Result using the learned sparse model. Average angular error (AAE) and average end-point error (AEPE) are shown below each color coded image result.

To learn the flow dictionary $\mathbf{D}^h = [\mathbf{D}^h_u \, \mathbf{0}; \mathbf{0} \, \mathbf{D}^h_v]$, we simplify the problem by treating the horizontal and vertical motions separately. We will use $\mathbf{D}^h_u$ as an example to present how the flow dictionary can be learned, and $\mathbf{D}^h_v$ is learned similarly. Given a large training set of ground truth flow data $\{\mathbf{z}^i_u\}$, with each $\mathbf{z}^i_u \in \Re^n$ represents an extracted patch of horizontal flow fields, the learning of $\mathbf{D}^h_u \in \Re^{n \times p}$ amounts to solve the following optimization problem

$$\min_{\{\mathbf{D}^h_u, \{\mathbf{a}^i_u\}\}} \sum_i \frac{1}{2}\|\mathbf{z}^i_u - \mathbf{D}^h_u \mathbf{a}^i_u\|_2^2 + \beta \|\mathbf{a}^i_u\|_1$$
$$s.t. \quad \|\mathbf{d}^h_{u,j}\|_2^2 \le 1 \quad \forall \quad j = 1, \ldots, p, \tag{3}$$

where $\mathbf{a}^i_u \in \Re^p$ is the sparse coefficient vector of $\mathbf{z}^i_u$ to be optimized, and $\mathbf{d}^h_{u,j} \in \Re^n$ represents a dictionary atom which is a column of $\mathbf{D}^h_u$ and constrained to be unit norm. Note the objective function (3) is not convex w.r.t. $\mathbf{D}^h_u$, but it is convex w.r.t. $\mathbf{D}^h_u$ or $\{\mathbf{a}^i_u\}$ when the other one is fixed. To optimize, we follow the sparse coding literature [31], and use an iterative approach that alternates between the sparse coding stage (solving $\{\mathbf{a}^i_u\}$) and the dictionary update stage (updating $\mathbf{D}^h_u$). In this work, we choose the LARS algorithm [32] for sparse coding, and Lee et al.'s Lagrange dual method [31] for dictionary learning.

Note that when the data penalty function ψ_D is chosen as

a L^2 norm, the first two terms in (2) can be merged, yielding a standard sparse coding problem, which is equivalent to the optical flow formulation as proposed in [28]. For any flow patch centering at $\mathbf{x}$, let $\mathbf{B}_\mathbf{x} \in \Re^{n\times 2n}$ be the diagonalized matrix representation of the horizontal and vertical derivatives ensemble $\{\nabla I_2(\mathbf{x}+\mathbf{u}^0_\mathbf{x})\}$ of the pixels in this patch, and $\mathbf{y}_\mathbf{x} \in \Re^n$ be the vectorized ensemble $\{\nabla I_2^\top \mathbf{u}^0_\mathbf{x} - I_t(\mathbf{x})\}$, sparse coding amounts to minimize

$$\|\mathbf{y}_\mathbf{x} - \mathbf{B}_\mathbf{x}\mathbf{D}^h\mathbf{a}^h_\mathbf{x}\|_2^2 + \beta\|\mathbf{a}^h_\mathbf{x}\|_1. \tag{4}$$

When optimal sparse coefficient vectors $\{\mathbf{a}^h_\mathbf{x}\}$ for all flow patches are obtained, which normally overlap each other, a common way to reconstruct the flow field is by computing

$$\mathbf{u} = \frac{1}{n}\sum_\mathbf{x} R^h_\mathbf{x}\mathbf{D}^h\mathbf{a}^h_\mathbf{x}, \tag{5}$$

where $R^h_\mathbf{x} \in \Re^{2N\times 2n}$ is a binary operator which places each flow patch at its proper position in the flow field. This process essentially averages flow patches at overlapping pixels. In Figure 1, we demonstrate the effectiveness of learned sparse model starting from an initialization $\mathbf{u}^0$. And $\mathbf{y}_\mathbf{x}$ and $\mathbf{B}_\mathbf{x}$ at each position $\mathbf{x}$ are computed based on $\mathbf{u}^0$. We solve equation (4) to get $\{\mathbf{a}^h_x\}$, and use equation (5) to reconstruct the estimated flow field. The size of flow patch is 5×5. Figure 1 shows that the learned sparse model is generally better than those using generic dictionaries such as DCT.

2.1. Multi-scale spatial regularization

The learned sparse model in (2) exploits higher order spatial regularization. It works when either an initial flow field estimate $\mathbf{u}^0$ is given, or the displacements between frames I_1 and I_2 are small. However, in optical flow computation, the errors in intermediate flow estimates are normally dense with large variations. In fact, as the data term in (2) relies on the assumption of intensity constancy, which can be easily violated due to sensor noises, illumination changes, reflections, and shadows. Any advanced alternatives [4, 16] may only alleviate, but not eliminate the problem. When the flow noises become dense and large, higher order spatial terms generally suffer from instability and being trapped in local minima, neither learned dictionaries nor generic ones can provide a good constraint. This is a fundamental difference from image denoising if we look optical flow estimation as a flow field denoising process.

In order to stabilize the flow estimation process, and also to enable our model to cope with large displacements, we extend the model (2) and propose a multi-scale spatial term to regularize the flow field. The new spatial term is composed of a purely geometric first-order regularizer and our higher order learned sparse model. To derive the new model, we start from the commonly taken spatial regularity form $E_S(\mathbf{u}) = \sum_\mathbf{x} \psi_S(\nabla\mathbf{u}_\mathbf{x})$. If we choose $\psi_S(\cdot) = \|\cdot\|_1$ as used in the TV-L1 framework [4, 10], this is equivalent to let the flow gradient field being sparse. In fact, if we use simple horizontal and vertical kernels $[1\ -1]$ and $[1\ -1]^\top$, we can approximate the flow gradient computation as a linear combination of the flow field. We thus can get a variant of the TV like energy model as

$$E(\mathbf{u}) = \sum_\mathbf{x} \psi_D\big(\rho(\mathbf{x})\big) + \lambda\|T^l_\mathbf{x}\mathbf{u} - \mathbf{D}^l\mathbf{a}^l_\mathbf{x}\|_2^2 + \beta\|\mathbf{a}^l_\mathbf{x}\|_1, \tag{6}$$

where $T^l_\mathbf{x}$ is defined similarly as in (2), $\mathbf{D}^l$ denotes the pseudo-inverse of the linearized first-order derivative operator, it applies to a flow patch $T^l_\mathbf{x}\mathbf{u}$ centering at position $\mathbf{x}$. Combining with our proposed learned sparse model, we arrive at the following energy function to minimize

$$E(\mathbf{u}) = \sum_\mathbf{x} \Big\{\psi_D\big(\rho(\mathbf{x})\big) + \sum_{s\in\{l,h\}} \lambda_s\|T^s_\mathbf{x}\mathbf{u} - \mathbf{D}^s\mathbf{a}^s_\mathbf{x}\|_2^2 + \beta_s\|\mathbf{a}^s_\mathbf{x}\|_1\Big\}. \tag{7}$$

Note that the new model exploits statistics of different spatial scales, which may complement each other. Indeed, while the structure of a flow patch can be sparsely represented by the learned flow dictionary, flow vectors inside the patch is not necessary to be (piece-wise) smooth, which can be ensured by the added first-order sparsity constraint. Moreover, first-order spatial constraint stabilizes optical flow estimation process, and makes it easier to adapt into a coarse-to-fine/warping framework, which has proven itself to be very effective in optical flow estimation. Based on a sequential optimization scheme and robust higher order regularization (will be introduced in the following sections), our method can produce high quality results competitive with the current state-of-the-art.

3. Robust higher order spatial regularization

In Section 2.1, we have discussed the types of noises generally encountered in optical flow estimation, which are dense and large, the estimates at some pixels may be completely corrupted. We have thus introduced the first-order spatial regularizer to stabilize the estimation process. Together with a robust penalty function, it can reduce the errors at most of the pixels. However, due to data constraint violations caused by illumination changes, it inevitably leaves gross errors or outliers at some pixels, which can degrade the performance of the learned sparse model.

On the other hand, sparse signal recovery with dense and large errors is still an open problem in sparse coding literature. Among those relevant methods, Wright et al. [29] first showed that when the corrupted measurements are sparse, accurate recovery can be achieved via an extended L^1 minimization. They further proved that the same approach is

possible to cope with dense corruption [30]. However, their proving conditions on a *highly correlated dictionary*, which is in general true in face recognition [29], but not applicable in both image restoration and optical flow estimation using learned dictionaries. In this work, we take a more direct approach to address the problem of outliers. That is, we consider identifying those more reliable pixels and in each flow patch, we use them to do sparse coding regularization. Since flow patches, no matter smooth or discontinuous, always have simple structures and are indeed sparse signals, accurate recovery using partial measurements is the inbuilt property of sparse coding.

Our approach is based on the observation that optical flow is in general piece-wise smooth. Both flow estimates deviating from their surrounding ones in smooth regions, and flow boundary estimates are less reliable and can be treated as outliers. Formally, for each estimated flow vector $\mathbf{u}_\mathbf{x}$, we compute an associated weight $w_\mathbf{x}$ based on normalized flow similarities and spatial distances w.r.t. its surrounding pixels

$$w_\mathbf{x} = \frac{1}{m} \sum_{\tilde{\mathbf{x}} \in \mathcal{N}(\mathbf{x})} \exp\Big\{ - \frac{\|\mathbf{u}_\mathbf{x} - \mathbf{u}_{\tilde{\mathbf{x}}}\|^2}{2\sigma_1^2} - \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|^2}{2\sigma_2^2} \Big\}, \quad (8)$$

where $\mathcal{N}(\mathbf{x})$ denotes a neighborhood of $\mathbf{x}$, m is the size of $\mathcal{N}(\mathbf{x})$, σ_1 and σ_2 are tuning parameters. When doing higher order spatial regularization, for each flow patch $T^h_\mathbf{x}\mathbf{u}$ with n pixels, we use those $\alpha n (0 < \alpha < 1)$ pixels having the top weights to perform robust partial sparse coding, and get an optimal $\mathbf{a}^h_\mathbf{x}$. Then all pixels of this patch are updated as $\mathbf{D}^h\mathbf{a}^h_\mathbf{x}$. The expression (8) is motivated from bilateral filtering [21], but it is flow-driven, and is embedded in a learned and robust sparse model. Moreover, it can treat both smooth regions and regions having multiple motions. In Figure 2, we demonstrate the effectiveness of robust regularization on the "RubberWhale" sequence in the Middlebury training set.

We have introduced the common way to update the flow field as in (5), which averages flow patches at overlapping pixels. However, motivated by recent optical flow works using non-local spatial regularization [17, 7], we find it is better to consider local image structures when reconstructing the flow field. More specifically, for each patch in higher order spatial regularization, we compute a weight mask $M^h_\mathbf{x} \in \Re^{2n}$ based on color similarity

$$M^h_\mathbf{x}(\mathbf{x}') = \exp\{-\|I_1(\mathbf{x}) - I_1(\mathbf{x}')\|^2 / 2\sigma_3^2\}, \quad (9)$$

where $\mathbf{x}'$ is a pixel of the patch centering at $\mathbf{x}$, and σ_3 is a tuning parameter. The color value $I_1(\cdot)$ is measured in the Lab space. The following weighted flow reconstruction scheme generally improves performance

$$\mathbf{u} = \mathbf{diag}\Big(\sum_\mathbf{x} R^h_\mathbf{x} M^h_\mathbf{x}\Big)^{-1} \sum_\mathbf{x} R^h_\mathbf{x} \mathbf{diag}(M^h_\mathbf{x}) \mathbf{D}^h \mathbf{a}^h_\mathbf{x}. \quad (10)$$

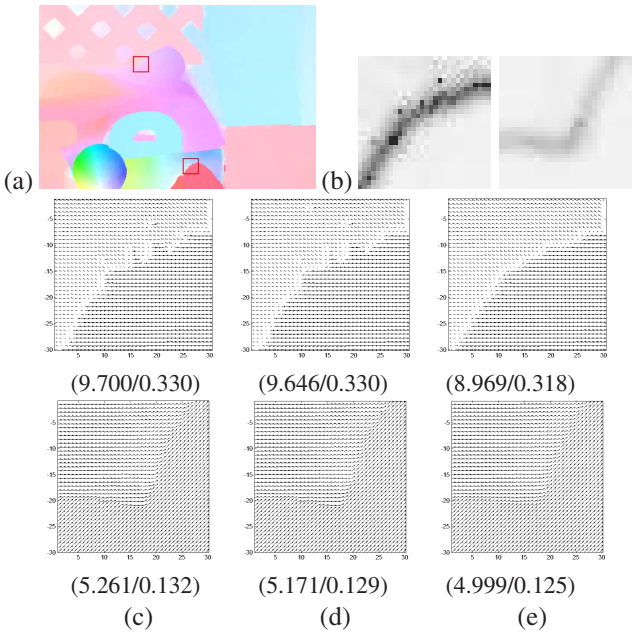


Figure 2. Effectiveness of the proposed robust approach for higher order spatial regularization. (a) is a color coded intermediate flow estimate of the "RubberWhale" sequence in [20]. Two local regions of (a) are plotted in (c). Their corresponding weight maps (computed by (8)) are shown in (b), where darker points are less reliable. Results in (d) are based on standard sparse coding. Resuls in (e) are based on the proposed robust approach. Average angular error (AAE) and average end-point error (AEPE) are shown in bracket below each plot (AAE/AEPE).

4. Sequential optimization

Due to a robust penalty function used in the data term and sparsity priors for multi-scale spatial regularization, the energy function (7) is neither convex nor continuously differentiable. To optimize, we propose to decompose the problem into a sequence of simpler ones, while each subproblem involves alternating updates and iterating until convergence, similar to the quadratic splitting scheme commonly used in recent optical flow works [11, 13, 14]. Specifically, our algorithm proceeds with the initial $\mathbf{u} = \mathbf{u}^0$ and the following iterations:

- For $\mathbf{u}$ being fixed, solve a sparse coding problem for each flow patch centering at $\mathbf{x}$

$$\lambda_l \|T^l_\mathbf{x}\mathbf{u} - \mathbf{D}^l\mathbf{a}^l_\mathbf{x}\|_2^2 + \beta_l \|\mathbf{a}^l_\mathbf{x}\|_1. \quad (11)$$

  Optimal $\{\mathbf{a}^l_\mathbf{x}\}$ can be computed using LARS [32] or Lee et al.'s method [31]. To update the whole field $\mathbf{u}$, we simply average the reconstructed flow patches $\{\mathbf{D}^l\mathbf{a}^l_\mathbf{x}\}$ at overlapping pixels, similar to the equation (5) as for the higher order case.

- For $\{\mathbf{a}_{\mathbf{x}}^l\}$ being fixed, minimize

$$\sum_{\mathbf{x}} \psi_D(\nabla I_2^\top(\mathbf{u}_{\mathbf{x}} - \mathbf{u}_{\mathbf{x}}^0) + I_t) + \lambda_l \|T_{\mathbf{x}}^l \mathbf{u} - \mathbf{D}^l \mathbf{a}_{\mathbf{x}}^l\|_2^2. \tag{12}$$

  Since function (12) is differentiable, we follow [6] and pursue a local minimum by setting its derivative zero w.r.t. $\mathbf{u}$, and solve the corresponding linear system of equations.

When the optimization concerning first-order spatial regularity is stable, our algorithm continues with the following iterations:

- For $\mathbf{u}$ being fixed, solve a robust partial sparse coding problem as proposed in Section 3, using the learned dictionary $\mathbf{D}^h$ [2]

$$\lambda_h \|T_{\mathbf{x}}^h \mathbf{u} - \mathbf{D}^h \mathbf{a}_{\mathbf{x}}^h\|_2^2 + \beta_h \|\mathbf{a}_{\mathbf{x}}^h\|_1. \tag{13}$$

  Again, Lee et al.'s method or LARS can be used to compute $\{\mathbf{a}_{\mathbf{x}}^h\}$. The updating of whole field $\mathbf{u}$ is based on the proposed weighted flow reconstruction scheme (10).

- For $\{\mathbf{a}_{\mathbf{x}}^h\}$ being fixed, minimize

$$\sum_{\mathbf{x}} \psi_D(\nabla I_2^\top(\mathbf{u}_{\mathbf{x}} - \mathbf{u}_{\mathbf{x}}^0) + I_t) + \lambda_h \|T_{\mathbf{x}}^h \mathbf{u} - \mathbf{D}^h \mathbf{a}_{\mathbf{x}}^h\|_2^2, \tag{14}$$

  which can be solved similarly as (12).

Our algorithm proceeds with a sequence of iterative steps, and alternates in minimizing functions (11), (12) and (13), (14) until convergence. Similar to [11], the parameters λ_l in (12) and λ_h in (14) are initially set small to allow warm starting, and then logarithmically increased in their iterations.

Note that by writing the energy model as the form (7) and optimizing using (13), we implicitly assume that the overlapping flow patches are independent from each other, this is obviously questionable. However, this approximation makes the optimization easier and in practice, leads to improved performance. It is also interesting to compare with the popularly used TV-L1 framework [11, 9]. While their spatial regularization steps can be interpreted as total variation based noise removal, our model and optimization step in (13) borrow ideas from learning adapted, sparse and redundant image models, which is currently most competitive in image restoration.

[2]Equation (13) does not explicitly account for partial sparse coding to keep consistent with the main energy function (7).

4.1. Implementation

To allow for illumination changes between image frames, we pre-process the images using the structure-texture decomposition proposed in [12]. Our method is embedded in a coarse-to-fine/warping framework to cope with large displacements. We use a downsampling factor of 0.8 when constructing image pyramids. On each pyramid level, we perform 10 warping steps. In each warping step, the parameters λ_l in (12) and λ_h in (14) are logarithmically increased from 10^{-4} to 10^2. For sparse coding regularization, β_l/λ_l in (11) is set as 0.1. Instead of fixing β_h/λ_h in (13), we set the number of nonzero elements for each $\mathbf{a}_{\mathbf{x}}^h$ in (13) as 10, i.e., $\|\mathbf{a}_{\mathbf{x}}^h\|_0 = 10$.

First-order spatial regularization is applied on 8×8 blocks of the flow field, then results are averaged at overlapping pixels. Following [11, 7], we perform a 5×5 median filtering after each step of first-order regularization. For higher order regularization, we use 5×5 ($n = 25$) flow patches. The horizontal and vertical flow dictionaries are separately trained, with the size of 4 times over-completeness, thus $p = 100$ and $\mathbf{D}^h \in \Re^{50\times 200}$. Currently we only apply higher order regularization on the pyramid level of original frame size. For the proposed robust approach, we consider a 9×9 neighborhood, thus $m = 81$ in (8). The tuning parameters σ_1 and σ_2 are set as 0.5 and 4 respectively, and $\alpha = 0.8$ for partial sparse coding. Finally, we fix the weighted flow reconstruction parameter as $\sigma_3 = 10$.

5. Experiments

In this section, we quantitatively evaluate our proposed contributions for optical flow estimation. We used the Middlebury benchmark [20], which provides a training set with given ground truth flow fields, and an evaluation set for comparison between different methods. Since our method is based on learning, when comparing with other methods on the evaluation set, we used all 8 ground truth flow fields in the training set to learn the flow dictionary. When testing on the training set, we used "leave-one-out" methodology. That is, we used 7 ground truth flow fields to learn the dictionary, and used the left one for evaluation. In the following, we will first give separate evaluation of key contribution factors proposed in this work. We then show overall performance on the evaluation set of the Middlebury benchmark. Throughout these evaluations, parameters were set as in Section 4.1 for all testing sequences.

5.1. Contribution evaluation

In Table 1, we use the Middlebury training set to show the contribution of higher order spatial regularization for accurate flow estimation. Accuracies in terms of average angular error (AAE) are presented. While results using multi-

Measure	DCT Dict.	Learned Dict.	Dimetrodon	Grove2	Grove3	Hydrangea	RubberWhale	Urban2	Urban3	Venus
AAE	×	×	2.505	2.132	6.169	1.795	2.682	2.572	**4.629**	4.150
AAE	v	×	2.511	2.063	6.043	1.774	2.672	2.498	4.633	4.123
AAE	×	v	**2.481**	**2.012**	**6.011**	**1.758**	**2.629**	**2.481**	4.630	**4.095**

Table 1. Evaluation results on the Middlebury training set. Comparisons are made between methods using first-order spatial regularity only (first row), first-order plus higher order using DCT dictionary, and first-order plus higher order using learned dictionary. Measure is in terms of the average angular error (AAE).

Measure	RobustLSM	Weighted Recon.	Dimetrodon	Grove2	Grove3	Hydrangea	RubberWhale	Urban2	Urban3	Venus
AAE	v	×	2.551	1.595	5.112	1.811	2.300	2.036	2.685	3.357
AAE	v	v	**2.541**	**1.511**	**5.005**	**1.803**	**2.285**	**2.004**	**2.599**	**3.297**

Table 2. Evaluation results on the Middlebury training set. Results in both rows are based on robust higher order regularization using learned dictionary. Using a weighted flow reconstruction scheme, the results in the second row are further improved. Measure is in terms of the average angular error (AAE).

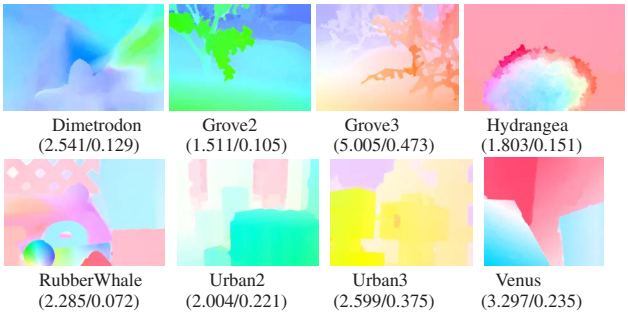


Figure 3. Color coded flow results of the 8 sequences in the Middlebury training set. Average angular error (AAE) and average end-point error (AEPE) are given in brackets below each image (AAE/AEPE).

scale spatial regularization are generally better than those using first-order spatial regularity only, our results based on learned flow dictionaries further improve over those using DCT. Note that in these experiments, we have not used the proposed robust higher order regularization yet, the effectiveness of which is demonstrated in Table 2. From Table 2 we can see that robust partial sparse coding indeed reduces the influence of outliers and improves performance. Finally, the image-driven, weighted flow field reconstruction scheme pushes the accuracies a step further. Figure 3 gives the color coded flow results of the 8 Middlebury training sequences.

5.2. Overall performance

Figure 4 compares our method with other methods using screenshots from the Middlebury evaluation homepage, where our method is denoted as LSM. Only top-performing methods are shown for comparison. At the time of publication, our results rank third for AAE and fourth for average EPE, among the methods listed there. Figure 4 shows that under all three criteria, i.e., the whole flow field (all), flow boundaries (disc), and smooth regions (untext), our method is highly competitive with the state-of-the-art.

The first ranking method, MDP-Flow2 [15], exploited extended flow initialization on each image scale to preserve small-scale motion structures, which are often lost in traditional coarse-to-fine/warping framework. The second method, Layers++ [8], proposed a probabilistic layered model that can address occlusions between different motion layers. We have not addressed these problems in this paper. Nevertheless, we mainly aim to show the effectiveness of learning-based sparse representation for optical flow estimation. Our method gives better results than both previous learning-based approaches [6, 18], and those recently proposed methods using higher order spatial regularization [19, 17, 7]. The techniques in [15, 8] may be combined with ours to further improve performance, we leave these issues for future research.

6. Conclusion

In this work, we showed the effectiveness of learned sparse representation for accurate optical flow estimation. Our method is based on multi-scale spatial regularization, which benefits from first-order spatial regularity and our proposed, learned sparse model. We used a sequential optimization scheme to solve the energy minimization problem. To address the problem of outliers in intermediate flow estimates, we further proposed flow-driven and image-driven approaches for robust spatial regularization. Experiments show that accuracies are significantly improved. Currently we have not addressed the recovery of small-scale motion structures. In future research, we plan to combine our method with extended flow initialization on each image scale, to further improve the accuracy.

References

[1] B.K.P. Horn and B.G. Schunck, Determining optical flow, *Artificial Intelligence*, 17:185-203, 1981. 1, 2, 3

Average angle error	avg. rank	Army (Hidden texture) GT all	Army im0 all	Army im0 disc	Army im0 untext	Army im1 all	Army im1 disc	Army im1 untext	Mequon (Hidden texture) GT all	Mequon im0 all	Mequon im0 disc	Mequon im0 untext	Mequon im1 all	Mequon im1 disc	Mequon im1 untext	Schefflera (Hidden texture) GT all	Schefflera im0 all	Schefflera im0 disc	Schefflera im0 untext	Schefflera im1 all	Schefflera im1 disc	Schefflera im1 untext	Wooden (Hidden texture) GT all	Wooden im0 all	Wooden im0 disc	Wooden im0 untext	Wooden im1 all	Wooden im1 disc	Wooden im1 untext	Grove (Synthetic) GT all	Grove im0 all	Grove im0 disc	Grove im0 untext	Grove im1 all	Grove im1 disc	Grove im1 untext	Urban (Synthetic) GT all	Urban im0 all	Urban im0 disc	Urban im0 untext	Urban im1 all	Urban im1 disc	Urban im1 untext	Yosemite (Synthetic) GT all	Yosemite im0 all	Yosemite im0 disc	Yosemite im0 untext	Yosemite im1 all	Yosemite im1 disc	Yosemite im1 untext	Teddy (Stereo) GT all	Teddy im0 all	Teddy im0 disc	Teddy im0 untext	Teddy im1 all	Teddy im1 disc	Teddy im1 untext
MDP-Flow2 [40]	6.5	3.32 7	8.76 5	2.85 8	**2.18** 1	7.47 3	1.85 4	2.77 3	6.95 3	2.06 9	3.25 14	17.3 18	1.59 12	2.87 6	3.73 5	2.32 5	3.15 3	11.1 8	2.65 3	2.04 7	3.64 11	1.60 6	1.88 6	4.49 7	1.49 2
Layers++ [38]	6.6	3.11 2	8.22 2	2.79 6	2.43 5	**7.02** 1	2.24 11	**2.43** 1	**5.77** 1	2.18 12	**2.13** 1	**9.71** 1	**1.15** 1	**2.35** 1	**3.02** 1	**1.96** 1	3.81 12	11.4 9	3.22 15	2.74 21	4.01 25	2.35 20	1.45 2	**3.05** 1	1.79 6
LSM [41]	8.0	3.12 3	8.62 3	2.75 4	3.00 16	10.5 13	2.44 16	3.43 7	8.85 8	2.35 14	2.66 3	13.6 3	1.44 2	2.82 4	3.68 3	2.36 7	3.38 6	9.41 3	2.81 9	2.69 20	3.52 9	2.84 27	1.59 3	3.38 3	1.80 7
TC-Flow [48]	8.6	**2.91** 1	**8.00** 1	2.34 2	**2.18** 1	8.77 7	**1.52** 1	3.84 13	10.7 14	**1.49** 1	3.13 10	16.6 14	1.46 4	2.78 3	3.73 5	**1.96** 1	3.08 2	11.4 9	2.66 4	1.94 6	3.43 5	3.20 40	3.06 15	7.04 15	4.08 33
Classic+NL [31]	9.5	3.20 5	8.72 4	2.81 7	3.02 17	10.6 14	2.44 16	3.46 9	8.84 7	2.38 16	2.78 4	14.3 4	1.46 4	2.83 5	3.68 3	2.31 4	3.40 7	**9.09** 1	2.76 8	2.87 26	3.82 15	2.86 31	1.67 5	3.53 4	2.26 11
SimpleFlow [52]	11.2	3.35 8	9.20 9	2.98 12	3.18 19	10.7 16	2.71 24	5.06 20	12.6 19	2.70 20	2.95 6	15.1 7	1.58 10	2.91 7	3.79 7	2.47 11	3.59 11	9.49 4	2.99 11	2.39 13	3.46 6	2.24 18	1.60 4	3.56 5	1.57 3

Average endpoint error	avg. rank	Army (Hidden texture) GT all	Army im0 all	Army im0 disc	Army im0 untext	Army im1 all	Army im1 disc	Army im1 untext	Mequon (Hidden texture) GT all	Mequon im0 all	Mequon im0 disc	Mequon im0 untext	Mequon im1 all	Mequon im1 disc	Mequon im1 untext	Schefflera (Hidden texture) GT all	Schefflera im0 all	Schefflera im0 disc	Schefflera im0 untext	Schefflera im1 all	Schefflera im1 disc	Schefflera im1 untext	Wooden (Hidden texture) GT all	Wooden im0 all	Wooden im0 disc	Wooden im0 untext	Wooden im1 all	Wooden im1 disc	Wooden im1 untext	Grove (Synthetic) GT all	Grove im0 all	Grove im0 disc	Grove im0 untext	Grove im1 all	Grove im1 disc	Grove im1 untext	Urban (Synthetic) GT all	Urban im0 all	Urban im0 disc	Urban im0 untext	Urban im1 all	Urban im1 disc	Urban im1 untext	Yosemite (Synthetic) GT all	Yosemite im0 all	Yosemite im0 disc	Yosemite im0 untext	Yosemite im1 all	Yosemite im1 disc	Yosemite im1 untext	Teddy (Stereo) GT all	Teddy im0 all	Teddy im0 disc	Teddy im0 untext	Teddy im1 all	Teddy im1 disc	Teddy im1 untext
MDP-Flow2 [40]	5.2	0.09 6	0.23 3	0.07 3	0.16 2	**0.52** 1	0.13 3	0.22 2	0.46 3	0.17 4	0.17 11	0.93 19	0.09 4	0.65 4	0.98 5	0.43 4	**0.29** 1	**0.91** 1	0.26 3	0.11 7	0.13 9	0.17 8	0.51 7	1.11 9	0.72 7
Layers++ [38]	6.1	0.08 2	**0.21** 1	0.07 3	0.19 5	0.56 3	0.17 13	**0.20** 1	**0.40** 1	0.18 8	**0.13** 1	**0.58** 1	**0.07** 1	**0.48** 1	**0.70** 1	0.33 2	0.47 12	1.01 3	0.33 11	0.15 25	0.14 18	0.24 24	**0.46** 1	**0.88** 1	0.72 7
TC-Flow [48]	9.0	**0.07** 1	**0.21** 1	**0.06** 1	**0.15** 1	0.59 4	**0.11** 1	0.31 13	0.78 14	**0.14** 1	0.16 6	0.86 12	0.08 2	0.75 11	1.11 12	0.54 10	0.42 6	1.40 16	0.25 2	0.11 7	0.12 4	0.29 38	0.62 13	1.35 14	0.93 25
LSM [41]	9.0	0.08 2	0.23 3	0.07 3	0.22 13	0.73 15	0.18 15	0.28 8	0.64 6	0.19 13	0.14 2	0.70 2	0.09 4	0.66 5	0.97 4	0.48 7	0.50 15	1.06 5	0.33 11	0.15 25	0.12 4	0.29 38	0.50 4	0.99 3	0.73 10
Classic+NL [31]	9.9	0.08 2	0.23 3	0.07 3	0.22 13	0.74 16	0.18 15	0.29 10	0.65 10	0.19 13	0.15 3	0.73 4	0.09 4	0.64 3	0.93 3	0.47 5	0.52 17	1.12 7	0.33 11	0.16 31	0.13 9	0.29 38	0.49 3	0.98 2	0.74 12
IROF-TV [57]	10.8	0.09 6	0.25 10	0.08 10	0.22 13	0.77 17	0.19 19	0.30 12	0.70 12	0.19 13	0.18 14	0.93 19	0.11 16	0.73 8	1.04 7	0.56 12	0.44 8	1.69 28	0.31 7	0.09 3	**0.11** 1	0.12 3	0.50 4	1.08 7	0.73 10

Figure 4. Screenshots from the Middlebury optical flow benchmark (http://vision.middlebury.edu/flow). Our proposed method is denoted as LSM.

[2] M.J. Black and P. Anandan, The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields, *CVIU*, 63(1):75-104, 1996. 1, 3

[3] M. Bertero, T.A. Poggio, and V. Torre, Ill-posed problems in early vision, *Proc. of the IEEE*, 76(8):869-889, 1988. 1

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, High accuracy optical flow estimation based on a theory for warping, *Proc. of ECCV*, pp. 25-36, 2004. 1, 4

[5] V. Lempitsky, S. Roth, and C. Rother, FusionFlow: Discrete-continuous optimization for optical flow estimation, *Proc. of CVPR*, 2008. 1

[6] D. Sun, S. Roth, J.P. Lewis, and M.J. Black, Learning optical flow, *Proc. of ECCV*, Vol III, pp. 83-97, 2008. 1, 2, 6, 7

[7] D. Sun, S. Roth, and M. Black, Secrets of optical flow estimation and their principles, *Proc. of CVPR*, 2010. 2, 3, 5, 6, 7

[8] D. Sun, E. Sudderth, and M.J. Black, Layered Image Motion with Explicit Occlusions, Temporal Consistency, and Depth Ordering, *NIPS*, 2010. 7

[9] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, Anisotropic Huber-L1 Optical Flow, *Proc. of BMVC*, 2009. 1, 6

[10] C. Zach, T. Pock, and H. Bischof, A duality based approach for realtime TV-L1 optical flow, *Proc. of Pattern Recognition, DAGM*, pp. 214-223, 2007. 1, 4

[11] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, An improved algorithm for TV-L1 optical flow computation, *Proc. of DVMA Workshop*, 2008. 1, 5, 6

[12] L. Rudin, S.J. Osher, and E. Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D*, 60:259-268, 1992. 6

[13] A. Wedel, D. Cremers, T. Pock, and H. Bischof, Structure- and motion-adaptive regularization for high accuracy optic flow, *Proc. of ICCV*, 2009. 1, 5

[14] L. Xu, J. Jia, and Y. Matsushita, Motion detail preserving optical flow estimation, *Proc. of CVPR*, 2010. 5

[15] L. Xu, J. Jia, and Y. Matsushita, Motion detail preserving optical flow estimation, *Submitted to PAMI*, 2010. 7

[16] F. Steinbruecker, T. Pock, and D. Cremers, Advanced data terms for variational optic flow estimation, *Vision, Modeling, and Visualization Workshop*, 2009. 1, 4

[17] M. Werlberger, T. Pock, and H. Bischof, Motion estimation with non-local total variation regularization, *Proc. of CVPR*, 2010. 1, 2, 3, 5, 7

[18] S. Roth and M. J. Black, On the spatial statistics of optical flow, *Proc. of ICCV*, 2005. 1, 2, 7

[19] K. Lee, D. Kwon, I. Yun, and S. Lee, Optical flow estimation with adaptive convolution kernel prior on discrete framework, *Proc. of CVPR*, 2010. 2, 3, 7

[20] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M.J. Black, and R. Szeliski, A database and evaluation methodology for optical flow, *Proc. of ICCV*, 2007. 2, 5, 6

[21] C. Tomasi and R. Manduchi, Bilateral Filtering for Gray and Color Images, *Proc. of ICCV*, 1998. 2, 5

[22] M. Elad and M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, *IEEE Trans. on TIP*, 54(12), pp. 3736-3745, 2006. 2, 3

[23] J. Mairal, M. Elad, and G. Sapiro, Sparse representation for color image restoration, *IEEE Trans. on TIP*, 17(1), pp. 53-69, 2008. 2, 3

[24] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, Non-local Sparse Models for Image Restoration, *Proc. of ICCV*, 2009. 2, 3

[25] B. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, *Proc. of IJCAI*, pp. 674-679, 1981. 2, 3

[26] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani, Hierarchical model-based motion estimation, *Proc. of ECCV*, 1992. 2

[27] A. Bruhn, J. Weickert, and C. Schnorr, Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods, *IJCV*, 63(3), 2005. 2

[28] X. Shen and Y. Wu, Sparsity model for robust optical flow estimation at motion discontinuities, *Proc. of CVPR*, 2010. 2, 3, 4

[29] J. Wright, A.Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, Robust face recognition via sparse representation, *IEEE TPAMI*, 2008. 4, 5

[30] J. Wright and Y. Ma, Dense Error Correction via L1-Minimization, *IEEE Trans. Info. Theory*, 2009. 5

[31] H. Lee, A. Battle, R. Raina, and A.Y. Ng, Efficient sparse coding algorithms, *NIPS*, 2007. 3, 5

[32] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression, *Ann. Stat.*, 32(2), 2004. 3, 5