ESTIMATING OPTICAL FLOW USING
A GLOBAL MATCHING FORMULATION AND GRADUATED OPTIMIZATION

Ming Ye  Robert M. Haralick  Linda G. Shapiro

Electrical Engineering  Computer Science  Computer Science
Univ. of Washington  CUNY Graduate Center  Univ. of Washington
Seattle, WA 98195  New York, NY 10016  Seattle, WA 98195
mingye@u.washington.edu  haralick@gc.cuny.edu  shapiro@cs.washington.edu

ABSTRACT

In this paper we consider the problem of optimal optical flow estimation assuming brightness conservation and piecewise smoothness. We propose a formulation based on three-frame matching and global optimization allowing local variation. It is superior to popular gradient-based models and more justifiable than existing global methods. We also develop an efficient technique to minimize the resultant global energy. It takes advantage of local gradient, global gradient and global matching methods and overcomes their limitations. Experiments on various synthetic and real data and comparison with state-of-the-art techniques show that the method achieves discontinuity preserving capability and sub-pixel accuracy.

1. INTRODUCTION

Optical flow is a 2D image motion measure which has a wide range of applications in computer vision [9], video coding [14] and computer graphics [17]. Its accurate and efficient estimation is a long-standing difficult problem.

The fundamental assumption enabling optical flow estimation is brightness conservation, i.e.,

$$I(x, y, t) = I(x + u, y + v, t + 1), \quad (1)$$

where $V = (u, v)'$ is the optical flow vector. Methods making direct use of this constraint are called matching-based. Due to computational difficulties, they can yield unsatisfactory accuracy [2, 5]. For this reason, gradient-based methods have become popular. They use the linear approximation of Eq. (1)

$$I_x u + I_y v + I_t = 0, \quad (2)$$
a.k.a. the Optical Flow Constraint Equation (OFCE), where $(I_x, I_y, I_t)'$ is the spatiotemporal image intensity gradient. Despite better overall performance, gradient-based approaches are sub-optimal and break down in the presence of large motions and bad derivative estimates [20].

Additional constraints on the flow are obtained from various flow field models [2, 3]. Among them we are particularly interested in piecewise smooth models [4], because they are applicable to general scenes, and also provide a starting point for object-based modeling [14, 17] when it becomes appropriate and necessary. Piecewise smooth models include two types, local parametric and global optimization. A simple local parametric approach assumes constant flow in an area and amounts to solving a group of OFCEs. Its variants are reported to achieve the best overall accuracy and efficiency [2, 8, 20], but they have difficulties in coping with the generalized aperture problem [4]. Global optimization methods try to strike a balance between overall matching error and smoothness error by minimizing a global energy, which might be developed from regularization [4, 12] or Bayesian (MAP, MRF) [11, 5, 15] perspectives. Some techniques of this type [4, 15] have shown nice results with dense boundary-preserving flows, but their accuracy is somewhat disappointing due to formulation defects and lack of suitable optimization methods.

In this paper we discuss a matching-based global optimization method with a practical solution technique. The formulation is more complete than previous attempts in that it uses three-frame matching to overcome visibility problems at occlusions, and the strength of brightness and smoothness errors are automatically balanced according to local data variation. In order to solve the resultant large-scale nonconvex problem, we develop a graduated optimization strategy. Specifically, we begin with a robust local gradient method for initial flow and variance estimates, then refine the results using a global gradient-based method, and finally minimize the original energy by fastest descent. In this process, merits are inherited and drawbacks are largely avoided from all three formulations. As a result, high accuracy is achieved both on and off motion boundaries. The content presented here extends our previous work [21] in several different aspects. In particular, the global energy is fully redesigned, a global gradient-based step is introduced to the solution, and the performance is significantly improved.
2. GLOBAL ENERGY DESIGN

We assume the optical flow $V$ minimizes the global energy

$$E = \sum_{all\ pixels\ i} E_B(V_i) + E_S(V_i) \quad (3)$$

where $V_i$ is the flow vector at pixel $i$; $E_B$ and $E_S$ represent brightness conservation and flow smoothness respectively.

The traditional assumption that pixels are visible in all frames is a major source of gross errors in areas of occlusion. We resolve the ambiguity by letting the matching error be the minimum of backward and forward warping errors in 3 frames, i.e.,

$$e_W(V_i) = \min(|I_0(V_i) - I_1|, |I_j(V_i) - I_j|)$$

where $I_0(V_i)$, $I_j(V_i)$ are warped intensities in the previous and the next frames respectively. This error design is simple yet very effective and was recently proposed in independent studies [16, 21, 10]. We further use a robust error function $\rho(x, \sigma)$ to resist other sources of outliers, yielding the brightness energy

$$E_B(V_i) = \rho(e_W(V_i), \sigma_B) \quad (4)$$

where $\sigma_B$ is the local brightness variation scale. We define the smoothness error to be

$$E_S(V_i) = \sum_{j \in N^2_i} \rho(|V_i - V_j|, \sigma_S) \quad (5)$$

where $\sigma_S$ is the local flow variation scale. It requires a vector $V_i$ to be consistent with its 8-connected neighbors $V_j, j \in N^2_i$ and the robust error function prevents smoothing across motion boundaries.

We choose to use the Geman-McClure robust error function [4],

$$\rho(x, \sigma) = \frac{x^2}{(\sigma^2 + \sigma^2)}$$

in both $E_B$ and $E_S$ terms such that they are normalized to the range $[0, 1]$ and their relative strengths can be adjusted by local scales $\sigma_B, \sigma_S$; where the observation is not trustworthy ($\sigma_B$ is large), stronger smoothness is enforced, and vice versa. We gradually learn the scales during the optimization process (Section 3). In previous similar formulations [2, 4, 12], there is usually a control parameter $\lambda$ between $E_B$ and $E_S$, and $\sigma_B, \sigma_S, \lambda$ are global tuning parameters. Compared to this approach, our locally adaptive scheme is more reasonable and reduces parameter tuning in experiments.

3. GRADUATED OPTIMIZATION

Minimizing a global energy like Eq. (3) is very hard. Stochastic methods such as simulated annealing remain impractical [5]. Deterministic methods including graduated nonconvexity (GNC) [5] and multigrid [12] were applied to gradient-based formulations with limited success.

Our approach to this problem is graduated optimization. It bears a certain similarity to GNC in that we start from an initial estimate and progressively minimize a series of finer approximations to the original energy. We create a $P$-level image pyramid $p = 0, \ldots, p = P - 1$ and begin estimation from the top (coarsest) level $p = 0$ with a zero initial flow field [3]. At each level the algorithm proceeds in three steps. Step I solves a group of OFCEs under high-breakdown robust criteria for initial flow estimates and calculates scales from inlier variances. Step II is a gradient-based approximation to Eq. (3) for incremental flow $\Delta V$:

$$E(\Delta V) = \sum_{i} \rho(e_G(\Delta V_i), \sigma_B) + \sum_{j \in N^2_i} \rho(|\Delta V_i - \Delta V_j|, \sigma_S)$$

where $e_G$ is the OFCE error (Eq. (2)). We solve it using Successive OverRelaxation (SOR) [5]. At the nth iteration, each $u$ component (and $v$ similarly) is updated as

$$u_i^n = u_i^{n-1} - \omega \frac{1}{\gamma} \frac{\partial E}{\partial u_i},$$

where

$$T(u_i) = T_i^2/\sigma_B^2 + 8/\sigma_S^2.$$ 

The SOR procedure converges fast because of the locally adaptive step sizes and the good initial estimate which contains mainly high-frequency errors. Step III is basically greedy propagation: neighbors’ values and their average are tried on the center pixel and a replacement happens if the clique energy drops. More details of I, III are available in [21]. Fig. 1 gives the system diagram of the new approach.

Fig. 1. System diagram (operations at each pyramid level)

From a practical point of view, the graduated scheme inherits the best of all three popular optical flow approaches. The local gradient step produces high-quality initial flow and scale estimates. The global gradient step efficiently improves the flow consistency. The global matching step works on the original pyramid images and corrects gross errors introduced by derivative computation and the hierarchical process. Such advantages are verified in experiments on various synthetic and real data.
4. EXPERIMENTS

We estimate optical flow in the middle of every three frames. The number of pyramid levels is empirically determined and no other parameters are tuned. We also compare results with those produced by BA—a good representative of previous dense regularization techniques proposed by Black and Anandan [5] whose code is publicly available. BA calculates flow on the second of two frames. It uses the same number of pyramid levels as ours and other parameters are set as suggested in [5].

4.1. Synthetic Data

Five data sets with flow groundtruth are used for quantitative evaluation. We report two average error measures. One is the popular Barron angular error $e_\angle$ [2]. The other is the error vector magnitude $e_{1.1} = |V - V_0|$, where $V_0$ is the correct flow vector.

The TS sequence (Fig. 2) was created to examine the optimality of formulations. It is well textured and contains two occluding squares translating at exactly 1 pixel/frame. Hence an optimal formulation assuming brightness conservation and piecewise smoothness should fully recover the flow. Our method does achieve that. Now let us look at results from two gradient-based methods, BA and our Step I, which is a robust local gradient technique by itself [21, 1, 13]. They both produce gross errors at motion boundaries due to gradient evaluation failure. Step I also shows rounded corners because the background motion becomes dominant there. BA's poor accuracy also attributes to oversmoothing and slow convergence of its SOR procedure.

![Central frame](image1)
![Groundtruth and our result](image2)
![BA](image3)
![Step I](image4)

Fig. 2. TS sequence results (cropped).

The famous Translating Tree (TT), Diverging Tree (DT) and Yosemite (YOS) sequences are obtained from John Barron [2]. TT and DT simulate translational camera motion with respect to a textured planar surface. TT's motion is horizontal and DT's is divergent. YOS's motion is mostly divergent. The cloud part is excluded from evaluation. See [2] for detailed descriptions of these data. We use 2 levels of pyramid for TT and DT, and 3 levels for YOS. All error measures are summarized in Table 1. Most optical flow papers published after [2] report the $e_\angle$ error on YOS (see [1, 15, 12] and references therein). To our knowledge, ours is the smallest error among all reported by dense nonparametric estimation methods.

<table>
<thead>
<tr>
<th>Data</th>
<th>Technique</th>
<th>$e_\angle$ (°)</th>
<th>$e_{1.1}$ (pix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>BA</td>
<td>8.04</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>1.1e-2</td>
<td>3.5e-4</td>
</tr>
<tr>
<td>TT</td>
<td>BA</td>
<td>2.60</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>DT</td>
<td>BA</td>
<td>6.36</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>2.60</td>
<td>0.08</td>
</tr>
<tr>
<td>YOS</td>
<td>BA</td>
<td>2.71</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>1.92</td>
<td>0.12</td>
</tr>
<tr>
<td>DTTT</td>
<td>BA</td>
<td>10.9</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>4.03</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 1. Quantitative measures

We illustrate the discontinuity-preserving capability of our method on the DTTT sequence (Fig. 3). It was generated from TT, DT and "cookie cutters": data inside the cookie cutters come from TT and those outside come from DT. We display the horizontal flow component as intensity images to show more details. Brighter pixels represent larger speeds to the right. Two pyramid levels are used. Our result is smooth with crisp motion boundaries, and is better than BA both visually and quantitatively (Table 1). As a by-product of our method, motion boundaries are easily located in the smoothness error map. It is expectable that such results will greatly ease higher-level scene analysis, e.g. contour-based [12] and layered [17] representation.

However we do notice some gross errors near the boundaries. For example, the right corner of the triangle is smoothed into the background. A closer look reveals that most of such errors happen in textureless regions, where even human viewers are unable to resolve the ambiguity. In such situations, the correctness of "groundtruth" is questionable and confidence measures [2, 19] are needed for quantitative evaluation. Also noticeable is that our motion boundaries are not as smooth as one would like. This exhibits the weakness of the simple optimization method in Step III. The above observations suggest three future work directions: improving the optimization method, deriving confidence measures and developing more convincing evaluation techniques.
Two pyramid levels are used for TAXI and three for FG. Our method achieves the maximum speed about 3.0 pixels/frame. Motion in FG is caused by camera translation and scene depth. The image speed of the front tree is as large as about 7 pixels/frame. Two pyramid levels are used for TAXI and three for FG. Our technique shows consistent performance, yielding clear-cut motion boundaries and smooth flows within each layer.

4.2. Real Data

Finally, results on two well-recognized real image sequences, Hamburg Taxi (TAXI) and Flower Garden (FG), are given in Fig. 4. TAXI [2] mainly contains three moving cars with the maximum speed about 3.0 pixels/frame. Motion in FG is caused by a tree moving with a camera. The image speed of the front tree is as large as about 7 pixels/frame. Two pyramid levels are used for TAXI and three for FG. Our technique shows consistent performance, yielding clear-cut motion boundaries and smooth flows within each layer.

5. CONCLUSION

Optical flow estimation assuming brightness conservation and piecewise smoothness has been studied for over a decade and a good collection of techniques is available. But their accuracy and efficiency remain somewhat unsatisfactory, which has led to some belief that the problem is insolvable unless other sources of information, such as flow models and pre-segmentation, are introduced. In this paper we have discussed a formulation based on three-frame matching and global optimization allowing local variation, and solve it under a graduated minimization strategy. Extensive experiments verify that the new method out-performs its competitors and yields good accuracy on a wide variety of data. This supports our belief that by optimizing problem formulation and solution technique, it is possible to achieve significant improvement on general motion estimation.

In the future work, we will examine more sophisticated general optimization methods such as graph cuts [7] and Bayesian belief propagation [18], and also look into potential applications in automatic motion interpretation [6].

4.2. Real Data

Finally, results on two well-recognised real image sequences, Hamburg Taxi (TAXI), and Flower Garden (FG), are given in Fig. 4. TAXI [2] mainly contains three moving cars with the maximum speed about 3.0 pixels/frame. Motion in FG is caused by camera translation and scene depth. The image speed of the front tree is as large as about 7 pixels/frame. Two pyramid levels are used for TAXI and three for FG. Our technique shows consistent performance, yielding clear-cut motion boundaries and smooth flows within each layer.

5. CONCLUSION

Optical flow estimation assuming brightness conservation and piecewise smoothness has been studied for over a decade and a good collection of techniques is available. But their accuracy and efficiency remain somewhat unsatisfactory, which has led to some belief that the problem is insolvable unless other sources of information, such as flow models and pre-segmentation, are introduced. In this paper we have discussed a formulation based on three-frame matching and global optimization allowing local variation, and solve it under a graduated minimization strategy. Extensive experiments verify that the new method out-performs its competitors and yields good accuracy on a wide variety of data. This supports our belief that by optimizing problem formulation and solution technique, it is possible to achieve significant improvement on general motion estimation.

In the future work, we will examine more sophisticated general optimization methods such as graph cuts [7] and Bayesian belief propagation [18], and also look into potential applications in automatic motion interpretation [6].

II - 292