A Naturalistic Open Source Movie for Optical Flow Evaluation

Daniel J. Butler[1], Jonas Wulff[2], Garrett B. Stanley[3], and Michael J. Black[2]

[1] University of Washington, Seattle, WA, USA
djbutler@cs.washington.edu
[2] Max-Planck Institute for Intelligent Systems, Tübingen, Germany
{jonas.wulff,black}@tuebingen.mpg.de
[3] Georgia Institute of Technology, Atlanta, GA, USA
garrett.stanley@bme.gatech.edu

Abstract. Ground truth optical flow is difficult to measure in real scenes with natural motion. As a result, optical flow data sets are restricted in terms of size, complexity, and diversity, making optical flow algorithms difficult to train and test on realistic data. We introduce a new optical flow data set derived from the open source 3D animated short film *Sintel*. This data set has important features not present in the popular Middlebury flow evaluation: long sequences, large motions, specular reflections, motion blur, defocus blur, and atmospheric effects. Because the graphics data that generated the movie is open source, we are able to render scenes under conditions of varying complexity to evaluate where existing flow algorithms fail. We evaluate several recent optical flow algorithms and find that current highly-ranked methods on the Middlebury evaluation have difficulty with this more complex data set suggesting further research on optical flow estimation is needed. To validate the use of synthetic data, we compare the image- and flow-statistics of Sintel to those of real films and videos and show that they are similar. The data set, metrics, and evaluation website are publicly available.

1 Introduction

In recent years, large-scale data sets and benchmark evaluations have driven innovation in computer vision, have enabled learning-based methods, and have provided a way to quantitatively evaluate progress in the field. In many areas, including stereo, 3D reconstruction, segmentation, and object recognition, ground truth data can be captured with specialized sensors or obtained by manual labeling. Optical flow is different. There is currently no sensor that provides direct measurements of scene motion, and manual labeling is both impractical and inaccurate. These facts have meant that previous flow evaluation data sets have had significant limitations. Here we describe *MPI-Sintel*, a new data set for optical flow evaluation that addresses many of these limitations. Figure 1 shows example images from our data set, together with the corresponding ground truth flow (color coding as in [1]). This data is derived from the animated short film *Sintel* [2]. It contains richly varied motion, illumination, scene structure, material properties, atmospheric effects, blur, etc. Sintel was created in Blender [3] by

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 611–625, 2012.

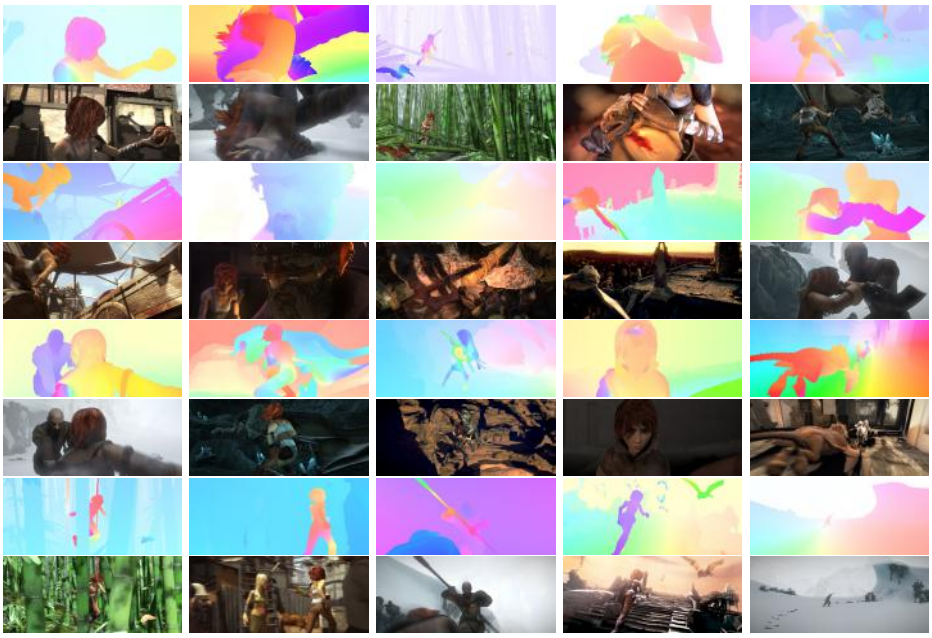

Fig. 1. MPI-Sintel Flow data set. Sample ground truth flow fields and corresponding images. The two bottom rows show frames from 10 of the 12 testing set sequences

the Durian Open Source Movie Project [2], which released all source files used during its creation as *open source*. Consequently, the scientific community has access to all the data used to *create* the film including all 3D graphics elements. We exploit this to create a data set that enables the evaluation and analysis of optical flow methods in ways that were not previously possible.

Sintel was made for entertainment, not science. One of our contributions is to extract the necessary information from Sintel for the scientific evaluation of optical flow algorithms. One of the key novelties of the MPI-Sintel data set is that we render the *same scenes* with *different render settings*, gradually increasing complexity. These different rendering *passes* build on each other and are illustrated in Fig. 2. This allows us to evaluate flow algorithms in different conditions to analyze where and how they fail, suggesting directions for future research.

With synthetic data the question always arises: Is it sufficiently realistic to tell us how optical flow algorithms will behave on real scenes? To address this we analyze the MPI-Sintel data set and find that first-order image and motion statistics are similar to those reported in the literature for natural scenes [5,6,7]. A novelty of our analysis is the use of a companion set of "Lookalike" image sequences from real films and videos. The Lookalikes have semantically similar scene content to clips in the MPI-Sintel data set (Fig. 4) and we compare the first-order image and motion statistics computed on both datasets. These analysises suggest Sintel is a reasonable proxy for a range of challenging real-world videos.

Fig. 2. Render passes. By row from the top: **Albedo Pass**: Flat, unshaded, surfaces exhibit constant albedo over time. **Clean Pass**: Illumination including smooth shading and specular reflections adds realism. **Final Pass**: Full rendering with all effects including blur due to camera depth of field (column 1) and motion (columns 2 and 3), and atmospheric effects (columns 3 and 4). The Final pass resembles the original rendered movie; for a description of changes see [4].

The new data set has several advantages over the widely used Middlebury flow benchmark [1]; a detailed comparison appears in Sec. 3. Briefly, the key advantages are: longer image sequences, ground truth flow for all frames, large non-rigid motions, more complexity (blur, atmosphere, specular surfaces, etc.), and a larger training set for machine learning methods. The dataset contains more than two orders of magnitude more flow data than the Middlebury dataset.

Like Middlebury, we withhold some ground truth flow for evaluation. Because the Sintel 3D data is open source, an unscrupulous researcher could, with effort, reconstruct the ground truth from the Blender files. Another novelty of our data set is the use of perturbed sequences for fraud detection. We render a subset of the evaluation sequences with slightly modified camera and object motion. This yields sequences that appear similar to the original Sintel data but have optical flow that is not public. A flow algorithm that performs much worse on perturbed sequences than non-perturbed ones suggests possible fraud.

We provide a website [8] for uploading results, computing of various error metrics, and ranking methods. While an exhaustive comparison of current optical flow methods is beyond the scope of this paper, we provide an initial analysis with six publicly available algorithms. Because the test data is much richer than previous datasets, we are able to evaluate optical flow methods in new ways that we hope will drive innovation in the field by focusing attention on critical failure cases. In addition to average endpoint error (EPE) for the entire test data set, we measure error as a function of speed, error as a function of distance to occlusion boundaries, and error in unmatched regions (defined below). We find that the existing flow methods perform rather poorly on MPI-Sintel. Algorithms with EPEs of less than 0.5 pixel on Middlebury have EPE of approximately 10 pixels on MPI-Sintel; a 20-fold increase. In unmatched regions we observe EPEs of more than 40 pixels. This opens room for significant improvements. Section 6 provides a comparison of the methods and the new evaluation measures.

In summary, our contributions include: 1) the introduction of MPI-Sintel, a new data set based on an open source animated film; 2) an analysis of the statistical properties of the data suggesting it is sufficiently representative of natural movies to be useful; 3) new evaluation measures; 4) an initial comparison of public-domain flow algorithms; 5) an evaluation website that maintains the current ranking and analysis of methods [8].

2 Previous Data Sets

There is a long history of optical flow evaluation, beginning with the work of Barron et al. [9]. They introduced the use of synthetic sequences with "ground truth" optical flow and proposed an evaluation metric based on average angular error. They independently implemented many existing methods and performed an extensive quantitative evaluation. McCane et al. [10] introduced more realistic graphics sequences with independent motion. The Middlebury evaluation [1] uses some graphics sequences as well, the limitations of which are discussed below. All these methods fall short on several dimensions but, most importantly, they offer a limited range of scenes and motions and limited visual realism.

Real images have also been used for flow evaluation. Otte and Nagel [11] used real scenes with simple geometry. The simple geometric structure made it possible to compute ground truth motion. The scenes however were limited to polyhedral objects, small motions, and simple textures. Middlebury also uses "real" sequences with stop-action motion; these are discussed below. Liu et al. [12] introduced a data set based on natural image sequences. Their key innovation is to hand-segment the images and use a flow algorithm to compute the motions for the segments. The assumption is that flow algorithms perform badly at motion boundaries and that human segmentation solves this problem. However, there are several concerns. First, it is not clear that humans are good at segmenting scenes and may inconsistently label regions such as shadows. Second, the ground truth flow will always be biased towards a particular algorithm used to compute it. Such sequences may be very useful as a "sanity check" to make sure optical flow algorithms generalize to realistic data, but a detailed quantitative comparison with such data seems problematic.

Roth and Black [5] took an approach between synthetic and real. They used laser scans of real scenes and real camera motions to synthesize optical flow fields. The data contained no image sequences – just optical flow in rigid scenes. Recently, two new datasets with natural imagery have been introduced [13,14] that use range sensing or stereo to derive ground truth flow. Unlike the Sintel flow dataset, these focus on automotive driving scenarios.

There are four key lessons to take away from previous studies. First, good data sets facilitate technological progress in the field and are therefore worth developing. Second, the lifespan of any data set is limited. At some point it can no longer be used to differentiate methods because their performance saturates. Third, any data set makes compromises and focuses on a subset of issues in the field. The Barron et al. sequences, for example, did not contain significant

motion discontinuities, which meant that methods dealing with these could not be effectively compared. Similarly, Middlebury does not contain large motions and motion blur. Fourth, a centralized public comparison is important to fairly summarize the state of the art and encourage innovation through competition.

3 Design Decisions and Comparison to Middlebury

Research on optical flow estimation has advanced rapidly in recent years in part due to the Middlebury flow benchmark which provides an impartial quantitative evaluation and ranking of methods [1]. At the time of writing 73 optical flow methods are listed and the accuracy of the top methods is significantly better than when the benchmark was released in 2007. The Middlebury data set and evaluation has been enormously valuable so our critique here has to be taken in this context. Many of the limitations we address here were pointed out by the designers of that study. The main aspects we address with our new data set are:

Difficulty. At present, top methods on Middlebury are tightly grouped with small variations in one sequence greatly affecting the rankings. Thus, a key motivation of the Sintel data set is to introduce sequences that are sufficiently varied and challenging to create some "room at the top" to provide a clearer evaluation of existing and new methods.

Sequence Length. Most of the Middlebury sequences are 8 frames long, with several only being 2 frames long. Ground truth flow is provided only for one pair of frames in each sequence. In contrast, most Sintel sequences are 50 frames long with 49 ground truth flow fields. This should encourage the development of methods that use longer sequences and integrate information over time.

Amount of Data. Middlebury was not designed to provide training data for machine learning methods. The limited amount of non-test data included has meant that very few machine learning approaches have been applied to the flow problem (with a few exceptions [15]). The success of such methods in other areas of vision has grown from an abundance of training data. The Sintel flow data set provides 1628 frames of ground truth flow (100 times Middlebury) in separate test (564 frames, withheld) and training sets (1064 frames).

Image Resolution. Middlebury images range from 548×388 to 640×480 (plus the Yosemite sequence which is only 316×252). While, in principle, we can render Sintel frames at any spatial resolution, we render at 1024×436 to be similar in height to existing data but with a wide-screen format to give more pixel data. Consequently each Sintel frame has $45\% - 100\%$ more pixels than the Middlebury frames (ignoring Yosemite).

Large Motions. While Middlebury has some large motions (up to 12 pixels per frame (ppf) in the real imagery and 35 ppf in the synthetic) most are quite small. Possibly as a consequence, there are few modern methods that explicitly address large motion; [16] is a notable exception which we evaluate here. The new Sintel data set has many large motions, including small objects moving quickly. The

maximum flow velocity is well over 100 ppf. This exposes limitations of current methods that are unseen in the Middlebury evaluation.

Blur. Middlebury uses stop-action for the "real" sequences and renders graphics scenes with no motion or defocus blur. The Sintel sequences are rendered with and without these effects in different passes.

Motion Boundaries and Occluded Regions. The graphics sequences in Middlebury have precisely localized motion boundaries but the real sequences do not. Some of the top methods are now accurate enough that the errors in the Middlebury occlusion regions are a cause for concern. We introduce a novel definition of motion boundaries and a new error measure that computes flow error as a function of distance from boundaries. In addition to motion boundaries, we know which pixels were unmatched in each frame and define an "Unmatched" mask to evaluate flow accuracy in these regions; this is important for sequences with large motion and distinguishes occluded *regions* from motion *boundaries*.

Real-World Challenges. Real world scenes contain non-rigid motion, complex scene structure, lighting variation and shadows, complex materials with specular reflections, and atmospheric effects such as fog. These are all properties of Sintel that are not present in previous data sets. Note that other new datasets also try to capture these effects [13,14].

Transparency. Like all previous optical flow evaluations, we exclude transparency here. This is not because transparency is unimportant but rather because one would need to define multiple ground truth flow values at each pixel, complicating all aspects of the evaluation. Consequently, we generate renderings of Sintel that do not contain transparency.

Ranking. Middlebury ranks methods by the average rank of their performance on every test sequence. This has a significant flaw if one of the sequences is "easy" in the sense that all methods perform roughly equally. In this case, insignificant changes in performance on one sequence can drastically change the overall ranking (arguably this is true for the Yosemite sequence in Middlebury). How one ranks methods affects the problems addressed by researchers. Here we define several "challenge" problems in the field and rank each method on these according to average endpoint error (EPE) across all test sequences. The challenge problems include flow accuracy: overall for different passes, for various distances from occlusion boundaries, in unmatched regions, for small, medium, and fast motions. The website provides summary rankings of each method in each challenge category [8].

4 The Sintel Data Set

Given Sintel's graphics elements, their motion, and the camera parameters, one can compute how every pixel moves from one frame to the next. We modified Blender's internal motion blur pipeline to give us accurate motion vectors at each pixel [4]. These vectors provide the ground truth optical flow maps.

Movies, however, are made for entertainment, not science. What matters is that the final shot "looks right." Blender is a very flexible tool and animators can achieve the effect they want in many ways, not all of which result in physically meaningful flow fields. Consequently, we selected "clips" where the optical flow is well defined and the rendering passes are intuitively correct. In several instances we modified the rendering parameters to produce output that is appropriate for optical flow evaluation. For example, in the original film, the main character's hair is rendered as a particle system, giving it a transparent appearance and making it difficult to define the ground truth flow. In our renders we have replaced such hair with opaque strands. The resulting frames are still visually pleasing and similar to the original, but more appropriate for flow evaluation. For a complete description of all changes see [4].

Clips: Training and Testing. All the data is available on the MPI-Sintel Flow website [8]. From the full movie we selected 35 clips. Apart from six shorter action sequences, each sequence is 50 frames long, giving 49 flow fields per clip. These clips were split into a training set of 23 clips and a test set of 12 clips. Two sequences of the test set were perturbed as described below. The test set was chosen to be sufficiently challenging, yet maintain similar image and flow statistics to the training set. In total, the data set contains 1064 training and 564 test frames. The images are saved as 8-bit PNG files and the frame rate is 24 frames per second. Further details about the image generation are given in [4]. The full dataset can be obtained from [8].

Perturbed Sequences. In the context of an evaluation system, the openness of the Sintel data becomes a disadvantage. While it would require some effort, a malicious user could extract the ground truth optical flow from the available data and submit this ground truth, thus achieving a perfect score. To discourage users from doing so, we randomly perturbed the scene geometry and camera motion of two sequences from the test set. At every 10 frames, random offsets $\in [-0.1, +0.1]$ meter equivalents were added along all dimensions to the location of the camera and stationary (non-animate) objects. Between the keyframes, the locations were linearly interpolated.

Due to the perturbation, the amount of motion in the perturbed scenes is higher. Additionally, for some objects, physics is violated by objects penetrating each other because of the perturbations. Thus, estimating the motion of the altered sequences is somewhat harder than of the original sequences. Using `Classic+NL-Fast` [17] and the "Final" pass (see below), the average endpoint error computed on the unperturbed sequences is 1.48 pixels. On the perturbed sequences it is 1.63 pixels. A cheating attempt (where one were to submit the ground truth of the unperturbed sequences) would result in an average endpoint error of 2.78 pixels. We believe that publishing the results on the perturbed sequences is a strong enough indicator of cheating to discourage it.

Render Passes. Blender generates each Sintel frame in series of steps called "passes" that simulate different aspects of image formation: smooth shading, specular reflections, and inter-reflections are some examples. After the initial

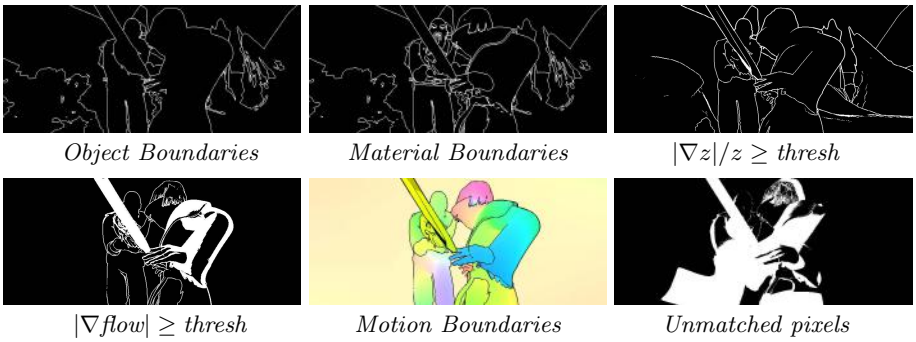

Fig. 3. Occlusions and Unmatched pixels. This shows the elements that are combined to produce an motion boundary mask (see text). The final boundaries are shown superimposed on the flow as black lines. Note that the unmatched pixels result from large inter-frame motion and are not used in computing the motion boundaries

render, the artist can apply additional effects including motion blur, defocus blur, atmospheric scattering, and color correction. The render passes we use in this work are illustrated in Fig. 2 and are:

Albedo. The simplest pass has roughly piecewise constant colors with no illumination effects. The motivation for including this pass in the evaluation is that it adheres to brightness constancy everywhere except at occlusion regions. This enables us to test how well optical flow methods perform when the common brightness constancy assumption holds almost everywhere.

Clean. This pass adds complexity by introducing illumination of various kinds. Surfaces exhibit smooth shading, self shadowing, darkening in cavities, and darkening where an object is close to a surface. Additionally this pass includes more complex illumination and reflectance properties including specular reflections, inter-reflections, and mirroring effects.

Final. This pass is similar to the final artistic rendering included in the released film. Beyond the Clean pass it adds atmospheric effects, depth of field blur, motion blur, color correction, and possibly other artistic embellishments to lighting and overall appearance.

Ground Truth Motion Boundaries. Defining ground truth motion boundaries is more subtle than it may at first appear. Simply thresholding the flow gradient (as in [1]) is not sufficient with large motions since this would classify smooth regions with a strong gradient (e.g. a ground plane) as full of occlusions. Our assumption is that occlusion boundaries occur at physical boundaries in the world which will coincide with object, material, or depth boundaries. Consequently we detect object boundaries and material boundaries from the graphics elements. We compute depth boundaries by thresholding the depth gradient divided by the depth (to make distant regions well behaved). We take the union of these to produce an *overestimate* of where motion boundaries could occur. We then take the intersection of this with a thresholded gradient magnitude of the flow field (threshold=2ppf). See Fig. 3 for an example.

Middlebury uses a fixed region around motion boundaries in which to compute error statistics. In contrast, we compute a distance transform from the boundaries and plot error as a function of distance. For evaluation purposes, we compute the average error in areas closer than 10 pixels, between 10 and 60 pixels, and larger than 60 pixels, and report these errors.

Unmatched Pixels. We additionally compute a mask containing "Unmatched" pixels seen in one image in a pair but not in the other (Fig. 3). With large motions, these are common and we find that nearly 8.5% of the pixels in Sintel are unmatched. Our evaluation makes these regions explicit and we report errors using Matched, Unmatched and All pixels.

Physics Violations. In movies, the laws of physics are often ignored. This also holds true for Sintel, resulting in a number of violations of physics. Specifically, objects sometimes interpenetrate each other and the lighting, especially for the characters, is often physically implausible. For example, Sintel's head is lit separately from the rest of the scene. Note that real films often contain computer-generated elements, composites of multiple images, or other special effects that result in non-physical lighting. Additionally, scenes with live actors often use lighting and reflectors that move with the actors, violating natural lighting conditions, but this usually goes unnoticed by the audience. Sintel is perceptually realistic enough to be enjoyed by human audiences, so we believe that an optical flow algorithm should be able to deal with it as well. While in this sense Sintel is not unlike video data "in the wild", one should be cautious when using this dataset to train and evaluate algorithms that are strongly reliant on real-world laws of physics.

Evaluation Methodology. For the evaluation, we expect people to compute the flow for all 564 frames of the test set for the Clean and Final passes. Given the current state of technology, this would make result uploading difficult. Hence, we provide binary files for Windows, Mac OS X, and Linux, which compute a sparse, consistent subsampling to reduce the amount of data by a factor of 10. The data is analyzed using the metrics and ranking methodology described in Sec. 6 and, with the submitter's permission, is published on the website.

The MPI-Sintel Flow Data Set. In summary, the full publicly available data set includes the image sequences for the Albedo, Clean, and Final passes. Additionally, for the training set, forward flow fields (floating point and color image visualizations), occlusion boundary masks, unmatched pixel masks, and invalid pixel maps are provided. Software is also provided to compute the various error statistics on the training data (users can, for example, use a portion of the training data for cross-validation). All the data as well as the evaluation site is available at `http://sintel.is.tue.mpg.de`.

5 Statistics of Sintel and Natural Movies

Any use of graphics for the analysis of computer vision algorithms raises important questions. Is the graphics data representative of real world data? In the case

Fig. 4. "Lookalike" clips. Representative frames for 5 out of the 49 clips are shown. First row: clips from films and web video featuring scene descriptions similar to Sintel. Second row: similar Sintel clips.

of optical flow this is difficult to verify without complex real-world flow ground truth. We hypothesize that the world of Sintel is sufficiently realistic to advance the study of optical flow, especially compared to the Middlebury benchmark. We offer two pieces of evidence. The first compares the image statistics of Sintel to natural scenes while the second compares the statistics of optical flow computed on Sintel with flow computed on similar real movies. We do not claim that Sintel is representative of all "natural" movies. Rather, we argue that it is sufficiently rich to be a useful challenge for the community and that algorithms that are successful on Sintel are likely to be useful on a relatively rich class of natural movies. We also compare the statistics with those of the Middlebury data set.

Lookalikes. We created a set of "Lookalike" clips that were selected from a mixture of feature films, TV productions, and amateur movies for having similar semantic (not necessarily visual) scene content to scenes in Sintel (Fig. 4). The clips were sorted into six clusters (bamboo, cave, indoor, outdoor, mountain, snowfight). From each cluster random frames were drawn in order to achieve the same relative number of frames as in the corresponding clusters in the Sintel data set. In total, the Lookalike data consists of 1473 frames, uniformly scaled to a height of 436 pixels, the same height as the Sintel sequences.

Image Statistics. To evaluate and compare the luminance statistics for the different data sets we first convert the images to gray-scale, $I(x, y) \in [0, 255]$. Histograms of $I(x, y)$ across all frames and all pixels are shown in Fig. 5. The Kullback-Leibler-Divergence from Sintel to the Lookalikes is 0.058, which is smaller than from Middlebury to the Lookalikes (0.176). Overall, both differences are fairly small, indicating that both Middlebury and Sintel are a good match for the Lookalikes in terms of their luminance distribution.

Spatial power spectra were estimated from the 2D FFT of the 436×436 pixel patch in the center of each frame (after removing the mean), and averaged across all frames. The power spectra for Sintel, the Lookalikes, and the Middlebury sequences all exhibit an approximately linear decrease in power with frequency (in log-log). Natural scene movies usually exhibit such a decrease [6,7], with a power spectrum slope around -2 (equivalent to a $1/f^2$ falloff). The linear least-squares fit to the log-log relationships reveals a power spectrum slope in x-direction of -2.27 for Sintel, -2.36 for the Lookalikes, and -2.17 for Middlebury.

Additionally, spatial and temporal derivatives of the images were computed using first differences. The log-histogram for the horizontal derivative is shown

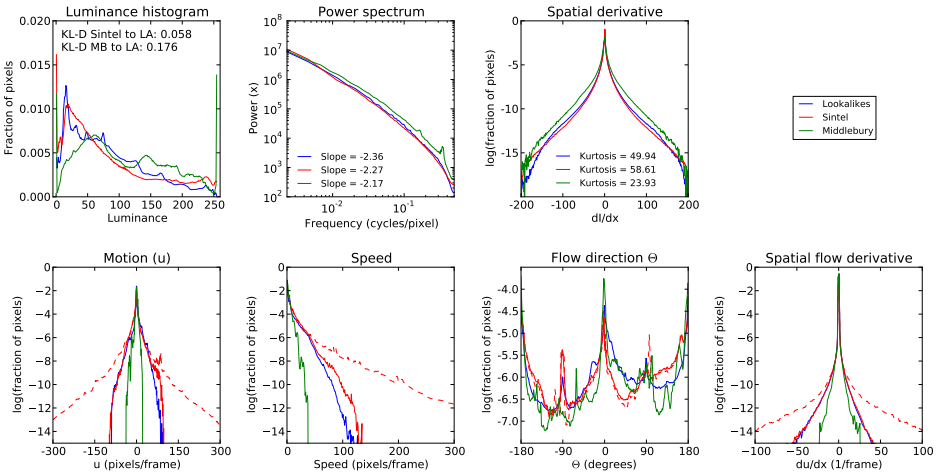

Fig. 5. Comparison of image and flow statistics: Sintel (red), Lookalikes (LA) (blue) and Middlebury (MB) (green); (see text). Top row: image statistics. Bottom row: flow statistics – solid lines represent the estimated optical flow, dashed lines represent the ground truth optical flow for Sintel.

in Fig. 5. The log-histograms exhibit characteristic signatures of natural scenes (peaked at zero, heavy tails) [6]. The Kurtosis of Sintel's distribution (58.61) is closer than Middlebury (23.93) to the Lookalike sequences (49.94).

Flow Statistics. Since there is no ground truth flow for the natural sequences, we compare the *estimated* optical flow for Sintel (Final pass), the Lookalikes, and all available Middlebury frames. The flow fields were computed using the `Classic+NL-Fast` method. We argue that, if the statistics of the computed flow fields are similar, it is likely that the underlying scenes and motions are similar. Let the horizontal and vertical flow be denoted $u(x, y)$ and $v(x, y)$ respectively (pixels/frame). We compute the speed $r(x, y) = \sqrt{u(x, y)^2 + v(x, y)^2}$ and angle $\theta(x, y) = \tan^{-1}(v(x, y)/u(x, y))$ as well as the spatial and temporal derivatives of u and v. Figure 5 shows the log histograms of these. The flow statistics of the Sintel ground truth are visually very similar to those in [5] but with much larger motion derivatives. As expected, the flow magnitudes for Middlebury are smaller than in real sequences or Sintel. The important points are: 1) there exist "natural" movies that have similar flow statistics to Sintel, and 2) the complexity of Sintel is much greater than Middlebury. More extensive statistics can be found in [4].

6 Analysis

To seed the evaluation we computed flow on the test sequences with several publicly available methods: `Classic+NL` [17] (Middlebury rank = 13/65), representative of a good modern method; `Classic++` [17] (rank=32), a standard

Table 1. Algorithm results for different challenges. $d10$: $\leq$ 10 pixels from an occlusion boundary; $d10$-60: 10-60 pixels; $d60$: $\geq$ 60 pixels; $s10$: speed $\leq$ 10 ppf; $s10$-40: 10 to 40 ppf; $s40$: $\geq$ 40 ppf

	EPE	match	unmatch	$d10$	$d10$-60	$d60$	$s10$	$s10$-40	$s40$
LDOF [16]									
Final	9.15	5.11	42.45	11.13	8.64	8.59	1.49	4.84	57.33
Clean	7.59	3.49	41.21	9.77	6.98	7.05	0.94	2.91	51.74
Albedo	7.43	3.18	42.27	9.53	6.64	7.15	1.02	2.57	50.83
Classic+NL [17]									
Final	9.18	4.87	44.60	11.64	8.88	8.09	1.11	4.50	60.31
Clean	7.99	3.83	42.23	10.40	7.76	6.84	0.57	2.70	57.41
Albedo	8.28	4.08	42.89	10.04	8.07	7.49	0.65	2.67	59.54
HS [17]									
Final	9.64	5.49	43.83	12.27	9.58	8.13	1.88	5.34	58.29
Clean	8.77	4.59	43.12	11.86	8.91	6.74	1.14	3.86	58.27
Albedo	9.72	5.33	45.83	12.04	9.73	8.26	1.50	4.51	62.85
Classic++ [17]									
Final	9.99	5.49	47.08	12.76	9.78	8.58	1.40	5.10	64.16
Clean	8.75	4.33	45.18	11.61	8.61	7.21	0.90	3.30	60.69
Albedo	9.22	4.64	46.94	11.61	9.02	8.04	1.08	3.33	63.63
Classic+NL-Fast [17]									
Final	10.12	5.74	46.24	12.37	9.82	9.13	1.09	4.67	67.82
Clean	9.16	4.81	45.12	11.44	9.02	7.97	0.56	2.82	66.96
Albedo	9.30	4.89	45.70	11.06	9.15	8.43	0.55	2.82	68.15
H-L1 [19]									
Final	11.95	7.41	49.52	13.89	11.90	10.85	1.16	7.97	74.80
Clean	12.67	8.07	50.62	14.84	12.91	11.05	0.75	9.98	77.84
Albedo	12.63	7.98	51.03	14.72	12.90	11.04	0.66	9.67	78.79

robust method; `LDOF` [16] (rank=50), specifically designed to deal with large displacement optical flow; Horn and Schunck (`HS`) [18] (rank=67), a classic method reimplemented in [17]; and Anisotropic Huber-L1 Flow (`H-L1`) [19] (rank=33), a GPU-based optical flow method. These methods have quite low EPEs on Middlebury: `Classic+NL`=0.32, `H-L1`=0.4, `Classic++`=0.41, and `LDOF`=0.5. Their EPEs on Sintel are an order of magnitude higher (Table 1).

Table 1 provides a snapshot of the data; see [8] for full details. We observe two effects: First, although brightness constancy is mostly conserved in the Albedo pass, results for this pass are generally worse than on the Clean pass. The images in the Albedo pass contain large homogeneous regions; these are bad for flow estimation. Stable illumination information in the Clean pass may add enough structure to provide useful cues for computing flow. Second, the Final pass is generally much more challenging than the Clean pass. The one exception is `H-L1`, which generally does worse than the other methods. For small motions ($s10$), `H-L1` is very competitive but for motions larger than 10ppf does badly. This suggests that, for large motions, `H-L1` is actually benefiting from the motion blur in the Final pass.

Table 1 reveals other properties of modern methods. Large errors occur in unmatched regions where a pixel appears only in the first of two adjacent frames; errors here are an order of magnitude higher than in matched regions. The methods also all fail fairly catastrophically for speeds above 40 ppf. It may be possible to tune some of these methods to do better on large motions, but here we used them as provided by their authors.

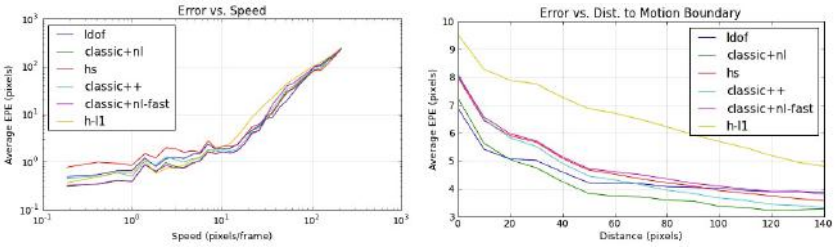

Fig. 6. Error as a function of speed and distance to nearest motion boundary

Fig. 7. Example results: `LDOF`. Top row: ground truth. Bottom row: estimated flow fields. Left to right: EPE: 1.57, 2.30, 6.00, 38.41.

`LDOF` is designed for large motions and it does slightly better on the largest motions but surprisingly not consistently better on intermediate ones (s10-40). Overall `LDOF` performs the best on the challenging Final pass, Fig. 7 shows some examples. `Classic+NL-Fast` performs best on small motions, particularly on the Clean pass. One surprise is that `HS` performed respectably well, even beating the robust methods on the largest motions. In all cases, EPEs consistently increases with speed. This can be seen in Fig. 6 (left).

As expected, all methods have trouble near motions boundaries, with EPE being consistently higher than the average EPE within the same frame. As distance from the boundary increases, error decreases, shown in Fig. 6 (right). We truncate the x-axis at 140 pixels, where the number of pixels per bin becomes too small to provide meaningful results. In computing error near motion boundaries we exclude unmatched pixels. To get a full picture of the difficulty caused by occlusions one should separately consider errors both near motion boundaries and inside unmatched regions.

7 Conclusions

It is common for films today to be largely synthetic. Graphics sequences have the advantage that everything is known – the motion, lighting, 3D shape, etc. Unfortunately, high quality feature films are rarely available for scientific analysis and dissemination. Realistic open-source films like Sintel provide a new opportunity for computer vision research. We have focused on optical flow because

there is no sensor that provides direct measurements of flow in natural scenes. Consequently realistic graphics is possibly the best solution available today. We have presented a carefully designed dataset for optical flow evaluation and found that the statistics of Sintel are similar to those of natural movies, making it a reasonable proxy for real scenes. We further found that recent optical flow algorithms that perform well on the Middlebury benchmark do significantly worse on the MPI-Sintel Flow data set, suggesting room for new methods. The use of graphics enables several innovations such as rendering the scene in multiple passes to explore and expose different problems with flow algorithms.

There are several ways we can expand the current dataset. These include rendering the scenes at higher spatial and temporal resolutions, with different levels of motion blur, with additional degradations such as rolling shutter effects, or with a virtual stereo camera. We plan a companion data set that includes depth maps and camera parameters (intrinsic and extrinsic) as well as scene flow. We may also offer non-motion evaluations of segmentation and the estimation of illumination, material, and shape.

Acknowledgments. We thank T. Roosendaal for his assistance. We also thank O. Broscaru and J. Anning for creating the evaluation website. MJB and GBS were supported in part by NSF CRCNS Grant IIS-0904630.

References

1. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., Szeliski, R.: A database and evaluation methodology for optical flow. IJCV 92, 1–31 (2011)
2. Roosendaal, T. (Producer): Sintel. Blender Foundation, Durian Open Movie Project (2010), `http://www.sintel.org/`
3. `http://www.blender.org/`
4. Butler, D., Wulff, J., Stanley, G., Black, M.: MPI-Sintel optical flow benchmark: Supplemental material. MPI-IS-TR-006, MPI for Intelligent Systems (2012)
5. Roth, S., Black, M.: On the spatial statistics of optical flow. IJCV 74, 33–50 (2007)
6. Field, D.: Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. A 4, 2379–2394 (1987)
7. Simoncelli, E., Olshausen, B.: Natural image statistics and neural representation. Annu. Rev. Neurosci. 24, 1193–1216 (2001)
8. `http://sintel.is.tue.mpg.de`
9. Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. IJCV 12, 43–77 (1994)
10. McCane, B., Novins, K., Crannitch, D., Galvin, B.: On benchmarking optical flow. CVIU 84, 126–143 (2001)
11. Otte, M., Nagel, H.-H.: Optical Flow Estimation: Advances and Comparisons. In: Eklundh, J.-O. (ed.) ECCV 1994. LNCS, vol. 800, pp. 51–60. Springer, Heidelberg (1994)
12. Liu, C., Freeman, W., Adelson, E., Weiss, Y.: Human-assisted motion annotation. In: CVPR, pp. 1–8 (2008)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: CVPR, pp. 3354–3361 (2012)

14. Meister, S., Jaehne, B., Kondermann, D.: An outdoor stereo camera system for the generation of real-world benchmark datasets. Opt. Eng. 51, 021107 (2012)
15. Sun, D., Roth, S., Lewis, J.P., Black, M.J.: Learning Optical Flow. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 83–97. Springer, Heidelberg (2008)
16. Brox, T., C., Malik, J.: Large displacement optical flow: Descriptor matching in variational motion estimation. PAMI 33, 500–513 (2011)
17. Sun, D., Roth, S., Black, M.: Secrets of optical flow estimation and their principles. In: CVPR, pp. 2432–2439 (2010)
18. Horn, B., Schunck, B.: Determining optical flow. AIJ 16, 185–203 (1981)
19. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic Huber-L1 optical flow. In: BMVC, pp. 1–11 (2009)