MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Technical Report No. 6 2 November 2012

# MPI-SINTEL OPTICAL FLOW BENCHMARK: SUPPLEMENTAL MATERIAL

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black

## 1 Introduction

This Technical Report contains the supplemental material to the main report on the MPI-Sintel optical flow dataset and evaluation [1]. In particular, we provide details of the image and optical flow statistics that are mentioned in the main paper. Additionally we provide details of the initial evaluation of optical flow algorithm performance on the dataset. Additional details and the dataset itself can be found on the MPI-Sintel website:

`http://sintel.is.tue.mpg.de`

## 2 Technical background

The technical background of MPI-Sintel is described in a separate paper [2]. It describes the technical choices we made in generating the dataset, as well as the steps and modifications necessary to re-generate the dataset from raw data. In particular it addresses how to transform the raw Sintel graphics data, which was designed for entertainment, to a dataset appropriate for the evaluation of optical flow.

## 3 Image and flow statistics

Unless stated otherwise, all statistics were computed on the "Final" pass of the MPI-Sintel dataset, which contains the most realistic and complex images including motion blur, focal blur, atmospheric effects and complex illumination.

### 3.1 Image statistics

This section shows a comparison of image statistics of the MPI-Sintel dataset, the Middlebury dataset, and a volume of "lookalike" video footage. We compare these sequences on the basis of histograms of luminance, spatial and temporal derivatives, gradient magnitude, and power spectra. All statistics were computed on a grayscale representation of the images, scaled to the range of $[0, 255]$.

#### 3.1.1 Luminance

The luminance distribution (Figure 1) shows that natural video sequences contain more dark areas than present in Middlebury. On the other hand, a relatively large number of Middlebury pixels are overexposed/saturated. MPI-Sintel is a fairly good match to the lookalike videos in terms of overall luminance.

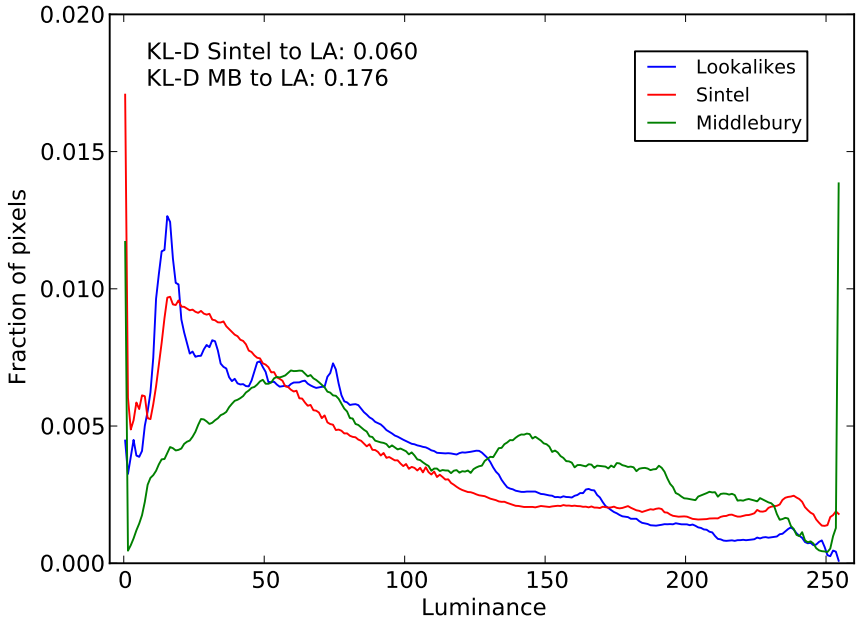


Figure 1: Luminance distribution

The KL divergence (KL-D in the figure) makes this quantitative with MPI-Sintel being much more similar to the lookalike sequences (LA). Since the lookalikes were chosen to semantically match the scenes from MPI-Sintel, this was to be expected.

#### 3.1.2 Gradient and spatial derivatives

Shown are log-histograms of the spatial derivatives in the image horizontal ($x$) and vertical ($y$) directions (Figure 2) and of the image gradient (Figure 3). The Kurtosis values shown in the plots indicate that the MPI-Sintel image derivatives more closely match the lookalikes than do the Middlebury images.

Both visually and in terms of Kurtosis the MPI-Sintel dataset is closer to the natural lookalikes than Middlebury.

#### 3.1.3 Temporal derivative

Figure 4 shows the log-histogram of the temporal derivative, i.e. the change in luminance for a given pixel between two adjacent frames. Pixels here are camera pixels, not world pixels; we do not use optical flow to track the real-world motion of pixels between frames. MPI-Sintel and the lookalikes are more similar for small motions (hence the more similar Kurtosis) but MPI-Sintel differs in the tails of the distribution. A more meaningful measure might compute the flow-warped temporal difference.

#### 3.1.4 Power spectra

Figure 5 and Figure 6 show the power spectra $P(f_x, f_y) = \mathtt{Re}(I(f_x, f_y))^2 + \mathtt{Im}(I(f_x, f_y))^2$ for $f_y = 0$ and $f_x = 0$, respectively, computed in a center patch of size $436 \times 436$ pixels, and averaged over all frames. In order to not violate the periodicity assumption, each patch was weighted with a Hamming window

$$w(x,y) = \begin{cases} 0.54 - 0.46\cos\left(\pi\left(1 - \frac{r(x,y)}{r_{max}}\right)\right) & \text{if } r(x,y) \le r_{max} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

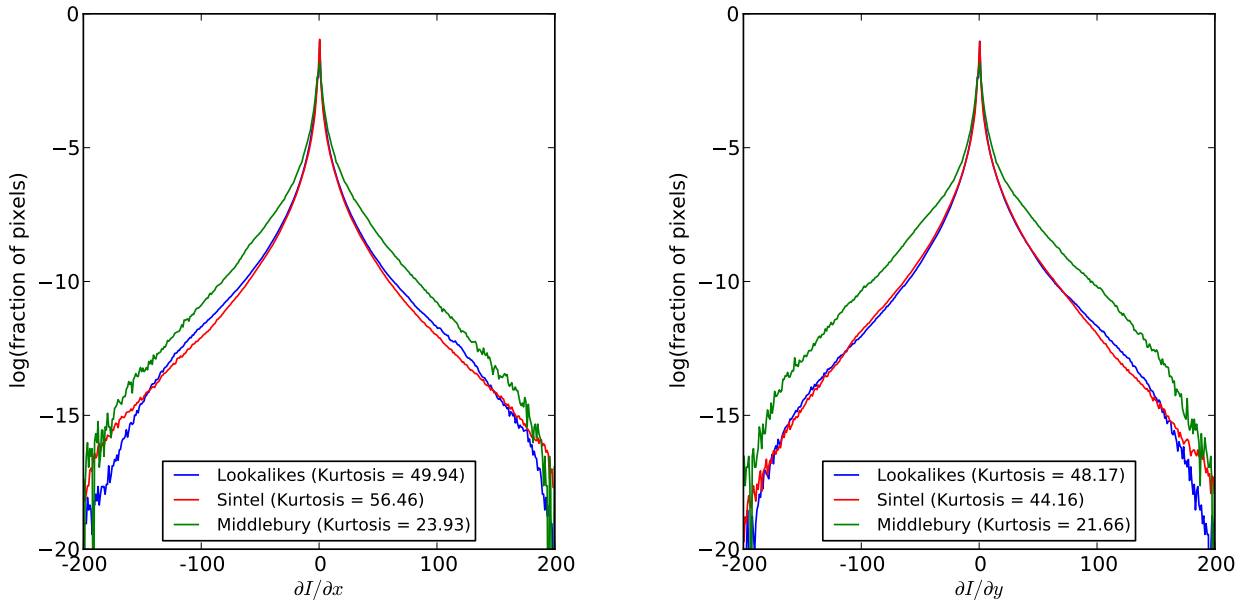


Figure 2: Spatial derivatives

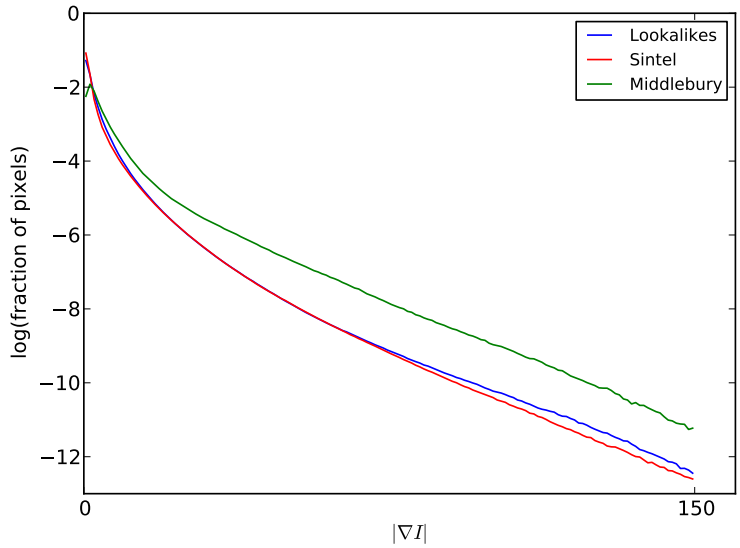


Figure 3: Image gradient

with $r(x,y) = \sqrt{(x-218)^2 + (y-218)^2}$ being the distance to the image center, and $r_{max} = 218$.

This ensures that the edges of the chosen patch have zero intensity and do not cause artifacts on the cardinal axes. To minimize spectral leaking caused by the windowing, we normalize the image, as described in [3]. Given the original image patch $I(x,y)$, the input to the FFT is thus the image

$$\hat{I}(x,y) = \frac{I(x,y) - \mu}{\mu} w(x,y) \tag{2}$$

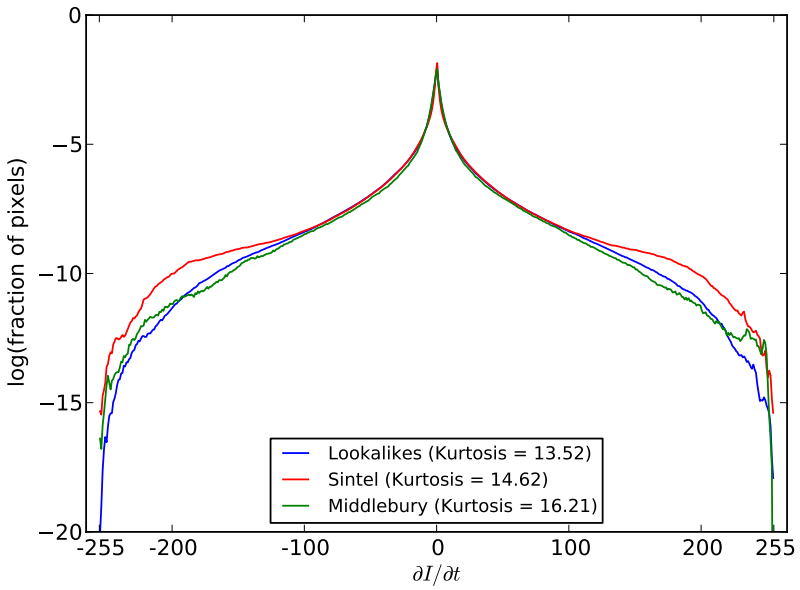


Figure 4: Temporal derivative

with the normalizing factor [3]

$$\mu = \frac{\sum_{x,y} I(x,y)w(x,y)}{\sum_{x,y} w(x,y)}. \tag{3}$$

Figure 5 shows the power in $x$ and $y$ direction, while Figure 6 shows an azimuthal average over all orientations. We computed the slopes using linear least squares fitting in the log-log space in the range $0 < f \leq 0.35$ cycles/pixel, since above these range aliasing and pixelation artifacts dominate [3, 4]. The linear fits are shown as dashed lines in Figure 5 and Figure 6.

We find that the power spectra are fairly similar, with Middlebury showing the lowest slope ($m = 2.52$), the lookalikes showing the highest slope ($m = 2.9$), and Sintel falling in between ($m = 2.66$). In natural images, power spectra are reported to fall off with $1/f^{-\alpha}$, with an often assumed $\alpha \approx 2$. However, reports also show that $\alpha$ can be quite varied, depending on factors such as the gist of the scene, or imaging modalities.

For example, Tolhurst et al. [5] found $\alpha$ to lie in the range $[1.6, 2.9]$, Field et al. [6] found slopes as high as $\alpha = 3.6$, van der Schaaf & van Hateren [3] report $\alpha = 1.88 \pm 0.43$, and Torralba & Oliva [4] distinguish between $\alpha \approx 2.2$ for indoor and $\alpha \approx 2.5$ for outdoor scenes. Both MPI-Sintel and the lookalikes are at the high end of these observations.

To better understand the slope of the MPI-Sintel power spectra we look at how images from the dataset compare with images typically used for the study of natural image statistics. Figure 7 illustrates the character of such "typical" natural images; they exhibit high depth of field and are consequently in crisp focus across the image. This is in contrast to MPI-Sintel images and stills from movies where it is common for the camera to focus on a subject and blur the background; see Figure 8.

Figure 9 shows a comparison between the set of 6 natural images with a high depth-of-field, shown in Figure 7, the set of 6 natural images with a low depth-of-field as shown in Figure 8, the Final pass of MPI-Sintel, and the Clean pass of MPI-Sintel. The Clean pass is included here because it does not include any focus blur, and thus corresponds to an infinitely large depth-of-field. Additionally, motion blur is absent, another possible source of attenuation of high frequencies.

Figure 9 indicates that the presence of focus blur can indeed result in a much steeper slope of the power spectrum, both in the case of the natural images, and in the case of MPI-Sintel. This makes sense since focal or motion blur will have the effect of attenuating the high frequencies, causing a steeper slope to the

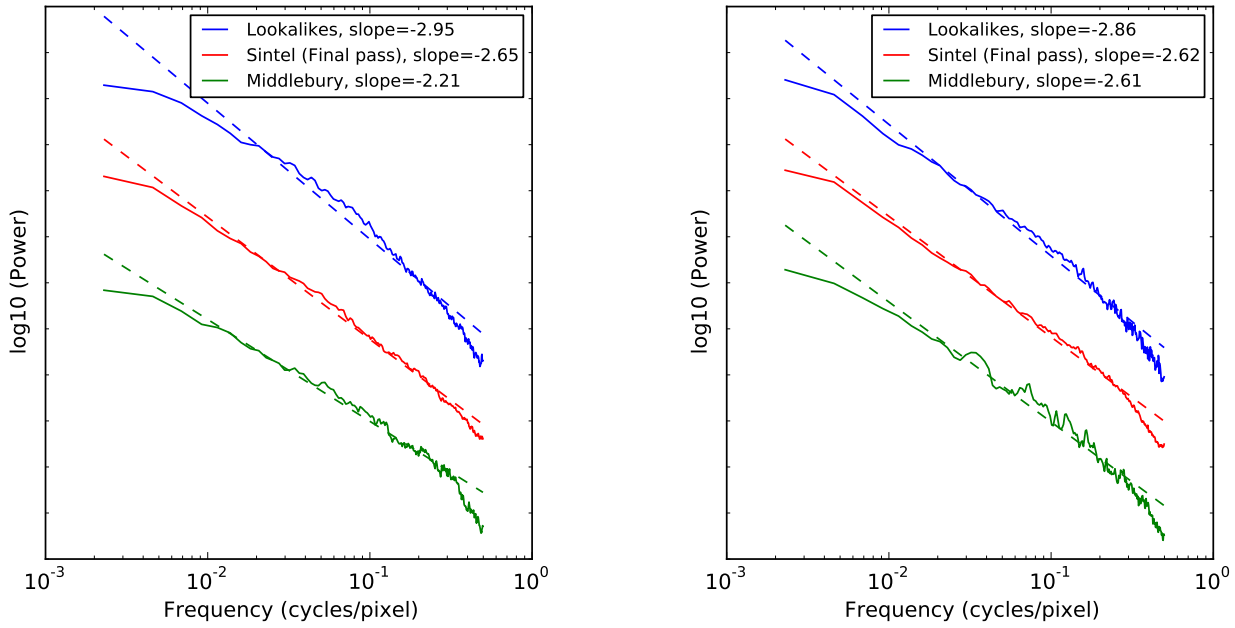


Figure 5: Power spectra in x and y directions. The dashed lines represent the best linear fit. For clarity, the graphs for Sintel and the lookalikes have been shifted by $10^2$ and $10^4$, respectively.

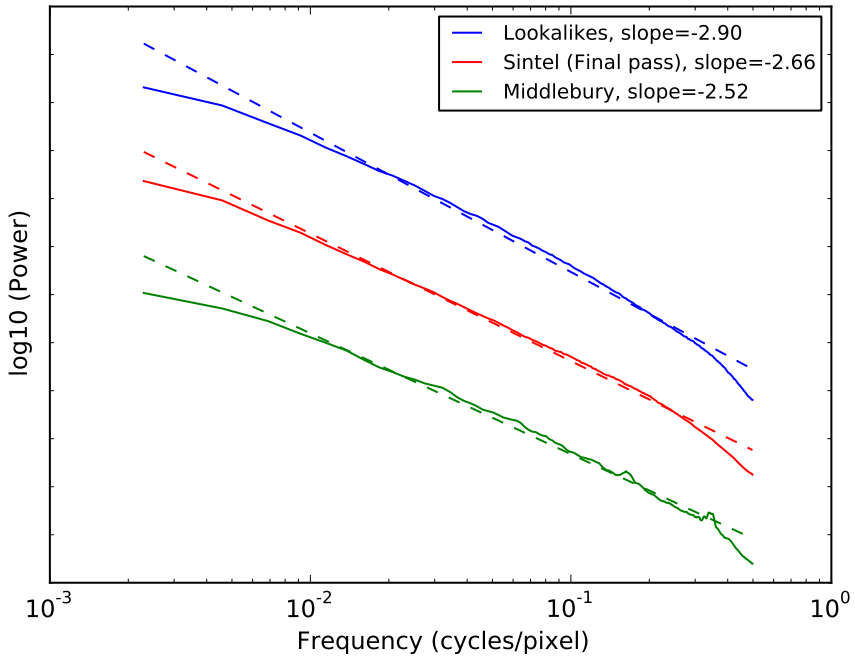


Figure 6: Power spectrum, azimuthal average. The dashed lines represent the best linear fit. For clarity, the graphs for Sintel and the lookalikes have been shifted by $10^2$ and $10^4$, respectively.

power spectrum.

Figure 7: Examples for typically used natural images.

Figure 8: Examples for natural images exhibiting a low depth-of-field and a more movie-like quality.

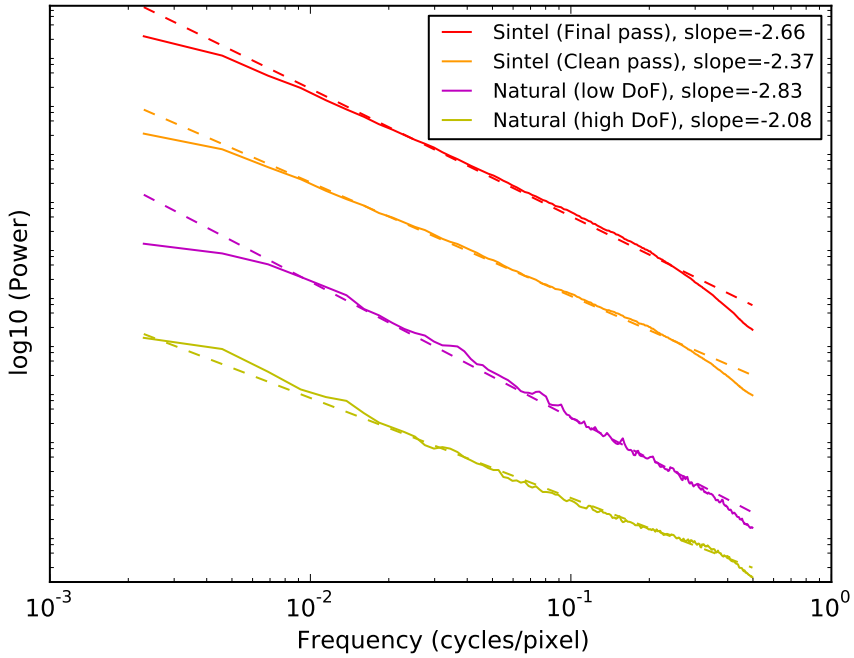


Figure 9: Comparison of power spectra of image sequences with different depths-of-field, azimuthal average. The dashed lines represent the best linear fit. Graphs have been shifted to increase readability.

## 3.2 Optical flow statistics

This section presents a comparison of the optical flow statistics. We are assuming that statistics of optical flow computed by a particular algorithm can be used as a basis for comparing the motions present in different sequences. The optical flow is computed using the `Classic+NL-fast` algorithm [7] using the default parameters. It should be noted that in case of Middlebury, the flow was computed on all publicly available frames, not only on the 8 frames for which ground truth flow is available.

For reference, we also show the statistics of the ground truth optical flow of MPI-Sintel as dashed lines.

### 3.2.1 Flow values

Figure 10 shows log-histograms of the optical flow values in the horizontal ($u$) and vertical ($v$) direction, respectively. As expected, the Middlebury dataset lacks large motions. Figure 10 shows that the MPI-Sintel dataset contains a large variety in motion that is similar to the lookalike motion. Note however that the

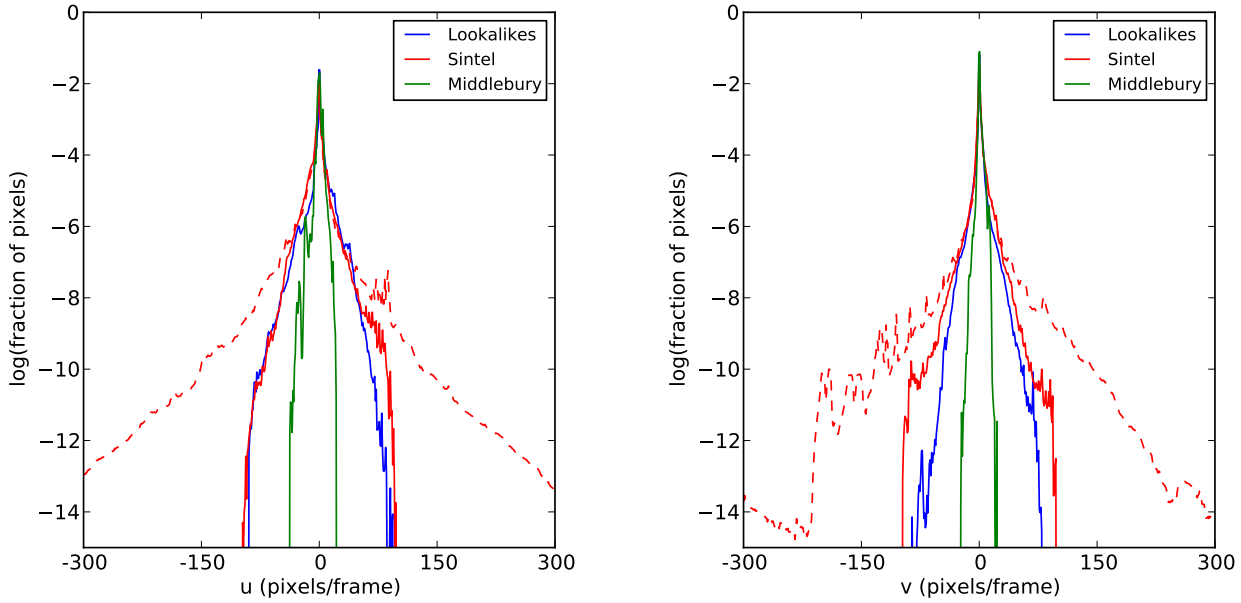


Figure 10: Log histograms of velocities in the horizontal and vertical direction.

`Classic+NL-fast` algorithm is unable to capture the largest motions seen in the heavy tails of the ground truth MPI-Sintel flow. Regardless, based on the computed flow, we argue that MPI-Sintel is more complex and more like the lookalikes than Middlebury.

### 3.2.2 Speed and direction

Figure 11 shows the overall speed and direction of the flow. The speed is defined as $s = \sqrt{u(x,y)^2 + v(x,y)^2}$, the direction as $\theta = \tan^{-1}(v(x,y)/u(x,y))$. Similar to the previous section, the speed plot shows the general lack of large motion in Middlebury.

While noisy, the histogram of directions in MPI-Sintel shows broad peaks around 0 and 180 degrees and smaller at +90 and -90 degrees. This is consistent both with the lookalike sequences, as well as with statistics of optical flow in general natural image sequences [8].

### 3.2.3 Gradients and spatial derivatives

Figure 12 shows the spatial gradients of the flow in the $u$ and $v$ direction respectively. Figure 13 shows the spatial derivatives of the flow fields along the $x$ and $y$ direction for both flow fields.

In all cases, the estimated flow for the MPI-Sintel dataset is more similar to the lookalikes than the estimated flow for Middlebury, pointing to a higher and more realistic variety of motion discontinuities and spatial variations.

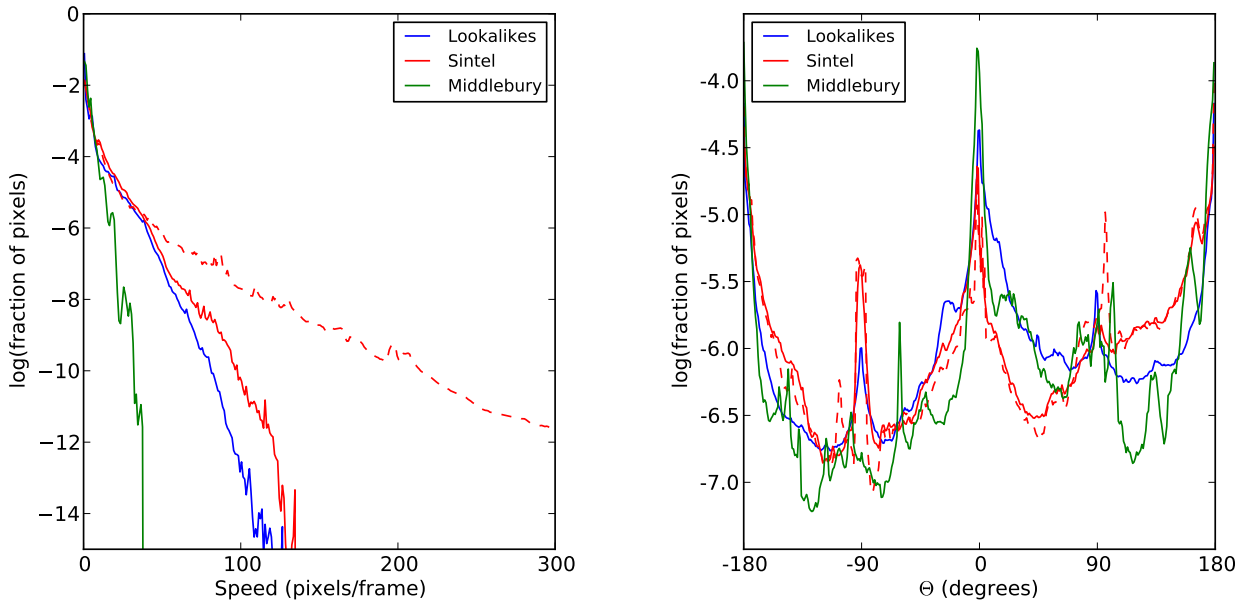


Figure 11: Flow speed and direction.

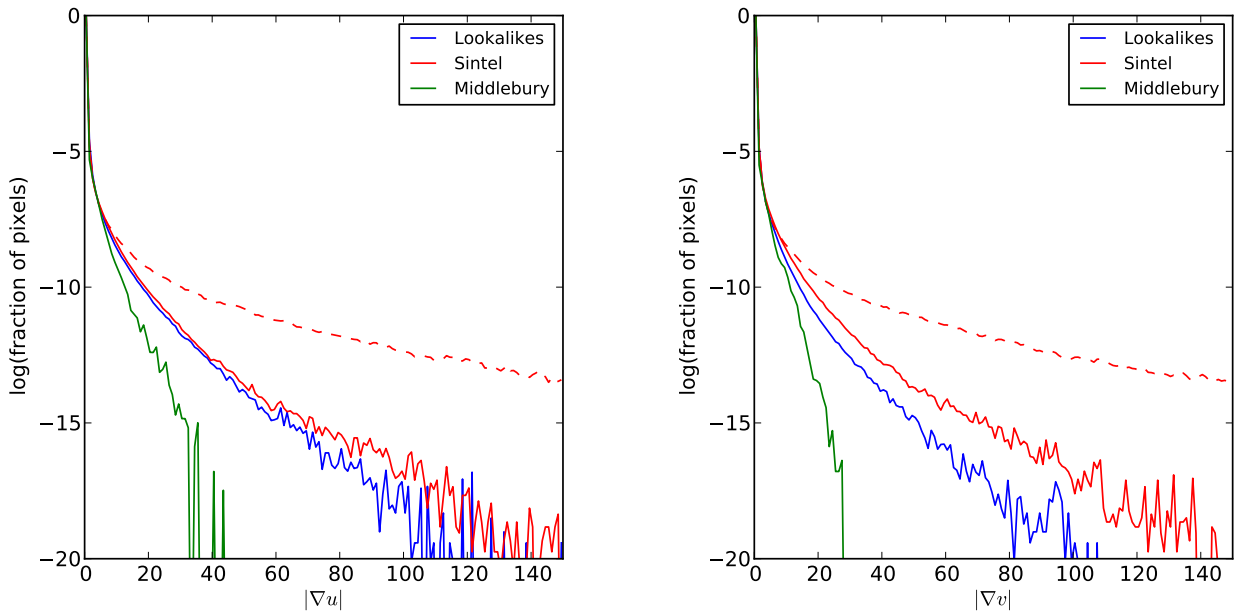


Figure 12: Gradients of flow.

# 4 Numerical results

As described in the main paper, we compute optical flow for the MPI-Sintel test sequences using six different optical flow algorithms:

- `Large-Displacement Optical Flow (LDOF)` [9]

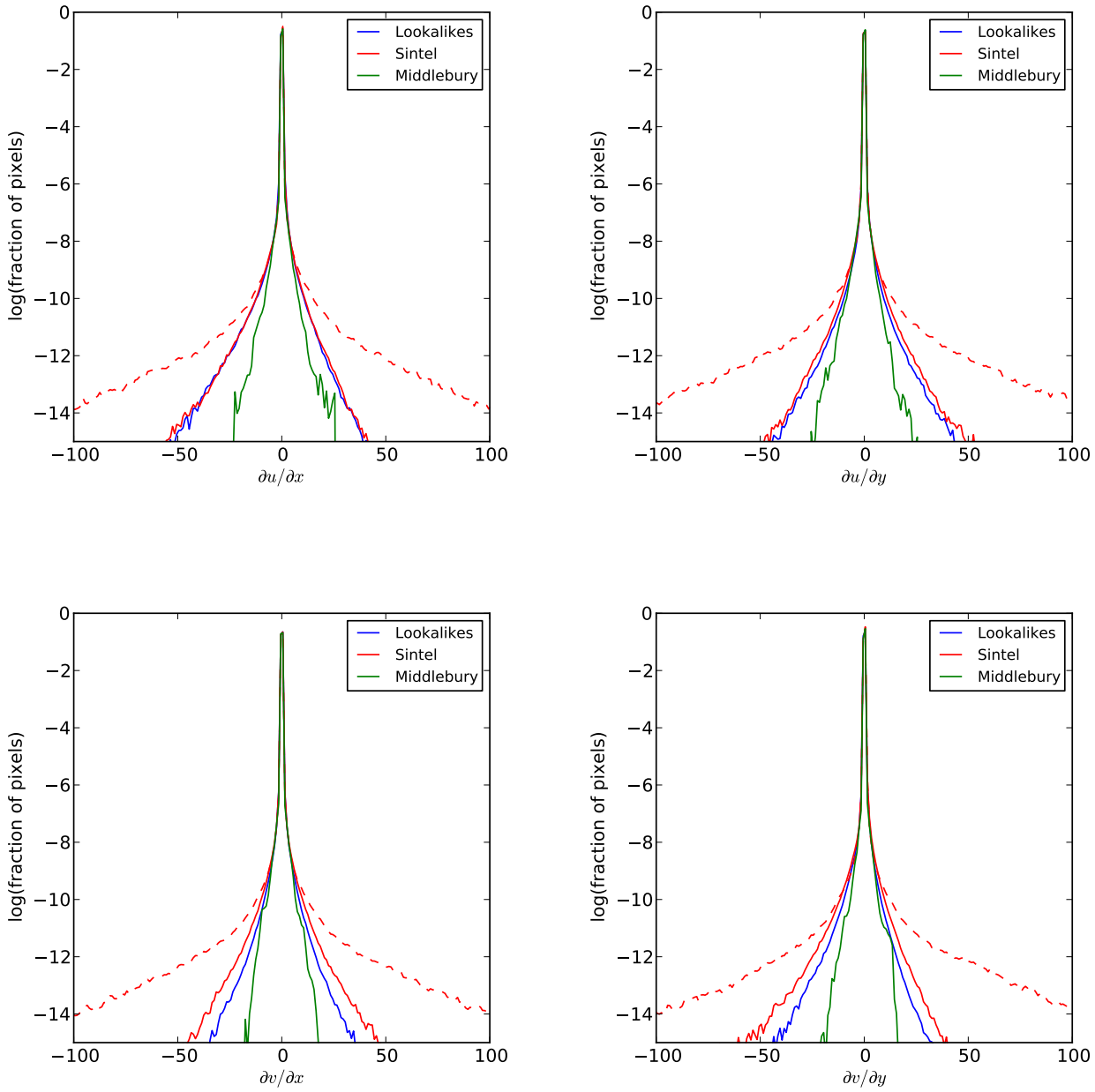


Figure 13: Spatial derivatives of flow.

- `Classic+NL` [7]
- `Classic+NL-fast` [7]
- `Classic++` [7]
- `Horn & Schunck` [10], using the reference implementation from [7]
- `Anisotropic Huber-L1` [11], using the reference implementation from `http://gpu4vision.icg.tugraz.at/`.

Table 1 shows the performance fo the methods across the whole test set for both the Clean and Final pass. All metrics are computed over all frames. The metrics are:

- `EPE`. Average endpoint error.
- `EPE matched`. Average endpoint error in matched regions; i.e. at pixels which are visible in adjacent frames.
- `EPE unmatched`. Average endpoint error in unmatched regions; i.e. at pixels which are only visible in the first frame. The main cause for this are image regions becoming occluded or moving out of frame.
- `d0-10`. Average endpoint error in regions closer than 10 pixels to the nearest motion boundary, taking only matched pixels into account. For the definition of motion boundaries, see [1].
- `d10-60`. Average endpoint error in regions between 10 and 60 pixels to the nearest motion boundary, taking only matched pixels into account.
- `d60-140`. Average endpoint error in regions between 60 and 140 pixels to the nearest motion boundary, taking only matched pixels into account.
- `s0-10`. Average endpoint errors in regions moving slower than 10 pixels per frame.
- `s10-40`. Average endpoint errors in regions moving between 10 and 40 pixels per frame.
- `s40+`. Average endpoint errors in regions moving more than 40 pixels per frame.

**Note:** Different from what is described in [1], the distance metrics (`d0-10`, `d10-60`, `d60-140`) only take matched regions into account. We found that, if unmatched regions are taken into account, the error values are dominated by the high errors present in unmatched regions. Additionally, instead of `d60+`, as reported in the paper, we now use the metric `d60-140`, taking only pixels into consideration that are closer than 140 pixels to the nearest motion boundary. The reason for this is that image regions further away from motion boundaries usually depict large unstructured elements, such as the sky or fields of snow. The errors in these regions are fairly high, again skewing the results. We therefore decided to exclude them.

For a more in-depth description of how the endpoint error varies with increasing speed and increasing distance from the motion boundaries, see [1].

Table 2 shows the average endpoint error per *sequence*, taking all pixels into account. This reveals the varying difficulty of the individual clips in the test set. Note that **PMarket_3** and **PShaman_1** refer to the perturbed sequences, designed to catch cheating attempts (see [1]).

To evaluate how the algorithms extrapolate to unmatched regions, Table 3 gives the average endpoint error per sequence, including only unmatched pixels.

# 5 Visual results

Table 4 and Table 5 show visual results for the methods we evaluated. The left column of both tables contains the ground truth optical flow for each sequence in the testing set; a single "canonical" frame is shown to illustrate the ground truth and the results.

Each column corresponds to a method and, for each sequence, we show the computed optical flow field for the canonical frame and below this an image of the absolute error with respect to the ground truth. The error displayed is the log of the EPE (which is always positive), scaled to the range $[0, 1]$ independently for each frame.

Table 4 shows results for the Final pass, Table 5 for the Clean pass.

# References


[1] Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), ed.: European Conf. on Computer Vision (ECCV). Part IV, LNCS 7577, Springer-Verlag (2012) 611–625

[2] Wulff, J., Butler, D.J., Stanley, G.B., Black, M.J.: Lessons and insights from creating a synthetic optical flow benchmark. In A. Fusiello et al. (Eds.), ed.: ECCV Workshop on Unsolved Problems in Optical Flow and Stereo Estimation. Part II, LNCS 7584, Springer-Verlag (2012) 168–177

[3] van der Schaaf, A., van Hateren, J.: Modelling the power spectra of natural images: Statistics and information. Vision Research **36** (1996) 2759 – 2770

[4] Torralba, A., Oliva, A.: Statistics of natural image categories. Network: computation in neural systems **14** (2003) 391–412

[5] Tolhurst, D.J., Tadmor, Y., Chao, T.: Amplitude spectra of natural images. Ophthalmic and Physiological Optics **12** (1992) 229–232

[6] Field, D.J. In: Scale-invariance and Self-similar 'Wavelet' Transforms: an Analysis of Natural Scenes and Mammalian Visual Systems. Oxford University Press. (1993) 151–193

[7] Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: IEEE Conf. on Computer Vision and Pattern Recognition, CVPR. (2010) 2432–2439

[8] Roth, S., Black, M.: On the spatial statistics of optical flow. International Journal of Computer Vision **74** (2007) 33–50

[9] Brox, T., Malik, J.: Large displacement optical flow: Descriptor matching in variational motion estimation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33** (2011) 500 –513

[10] Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial Intelligence **17** (1981) 185 – 203

[11] Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic huber-l1 optical flow. In: Proceedings of the British machine vision conference. Volume 34. (2009) 1–11

| *Method/Pass* | *Method* | | | | | |
|---|---|---|---|---|---|---|
| | LDOF | Classic+NL | Classic+NL-fast | Classic++ | HS | Aniso-HL1 |
| **EPE** | | | | | | |
| final | 9.116 | 9.153 | 10.088 | 9.959 | 9.610 | 11.927 |
| clean | 7.563 | 7.961 | 9.129 | 8.721 | 8.739 | 12.642 |
| **EPE matched** | | | | | | |
| final | 5.037 | 4.814 | 5.659 | 5.410 | 5.419 | 7.323 |
| clean | 3.432 | 3.770 | 4.725 | 4.259 | 4.525 | 7.983 |
| **EPE unmatched** | | | | | | |
| final | 42.344 | 44.509 | 46.145 | 47.000 | 43.734 | 49.366 |
| clean | 41.170 | 42.079 | 44.956 | 45.047 | 43.032 | 50.472 |
| **d10-** | | | | | | |
| final | 6.849 | 7.215 | 8.010 | 8.072 | 7.950 | 9.464 |
| clean | 5.353 | 6.191 | 7.157 | 6.983 | 7.542 | 10.457 |
| **d10-60** | | | | | | |
| final | 4.928 | 4.822 | 5.738 | 5.554 | 5.658 | 7.692 |
| clean | 3.284 | 3.911 | 4.974 | 4.494 | 5.045 | 8.675 |
| **d60-140** | | | | | | |
| final | 4.003 | 3.427 | 4.160 | 3.750 | 3.976 | 5.929 |
| clean | 2.454 | 2.509 | 3.331 | 2.753 | 2.891 | 6.320 |
| **s10-** | | | | | | |
| final | 1.485 | 1.113 | 1.092 | 1.403 | 1.882 | 1.155 |
| clean | 0.936 | 0.573 | 0.558 | 0.902 | 1.141 | 0.753 |
| **s10-40** | | | | | | |
| final | 4.839 | 4.496 | 4.666 | 5.098 | 5.335 | 7.966 |
| clean | 2.908 | 2.694 | 2.812 | 3.295 | 3.860 | 9.976 |
| **s40+** | | | | | | |
| final | 57.296 | 60.291 | 67.801 | 64.135 | 58.274 | 74.796 |
| clean | 51.696 | 57.374 | 66.935 | 60.645 | 58.243 | 77.835 |

Table 1: Errors per evaluation metric over all sequences.

| *Sequence/Pass* | *Method* | | | | | |
|---|---|---|---|---|---|---|
| | LDOF | Classic+NL | Classic+NL-fast | Classic++ | HS | Aniso-HL1 |
| **PMarket_3** | | | | | | |
| final | 2.832 | 1.550 | 1.696 | 1.786 | 2.118 | 2.908 |
| clean | 1.176 | 0.898 | 1.122 | 1.218 | 1.450 | 3.155 |
| **PShaman_1** | | | | | | |
| final | 2.269 | 1.212 | 1.545 | 1.600 | 2.463 | 4.135 |
| clean | 1.612 | 0.953 | 1.212 | 1.336 | 1.895 | 3.658 |
| **Ambush_1** | | | | | | |
| final | 44.960 | 44.626 | 48.991 | 48.026 | 40.549 | 52.362 |
| clean | 34.703 | 34.801 | 42.061 | 37.862 | 32.974 | 50.875 |
| **Ambush_3** | | | | | | |
| final | 14.134 | 13.722 | 14.188 | 14.764 | 15.382 | 15.247 |
| clean | 8.960 | 9.630 | 9.966 | 10.675 | 10.604 | 14.282 |
| **Bamboo_3** | | | | | | |
| final | 1.107 | 1.093 | 1.098 | 1.221 | 1.419 | 1.302 |
| clean | 1.036 | 0.997 | 1.019 | 1.139 | 1.339 | 2.188 |
| **Cave_3** | | | | | | |
| final | 9.227 | 12.415 | 14.452 | 13.997 | 13.007 | 16.290 |
| clean | 7.550 | 10.695 | 13.284 | 12.484 | 12.409 | 18.505 |
| **Market_1** | | | | | | |
| final | 4.179 | 5.339 | 7.775 | 6.737 | 5.397 | 9.084 |
| clean | 3.233 | 4.278 | 6.377 | 5.252 | 4.650 | 12.592 |
| **Market_4** | | | | | | |
| final | 39.210 | 39.132 | 42.164 | 41.176 | 40.005 | 49.522 |
| clean | 38.431 | 39.937 | 44.332 | 41.660 | 42.941 | 53.121 |
| **Mountain_2** | | | | | | |
| final | 1.618 | 1.453 | 1.430 | 1.490 | 1.544 | 1.702 |
| clean | 1.179 | 0.395 | 0.267 | 0.417 | 0.233 | 1.271 |
| **Temple_1** | | | | | | |
| final | 1.606 | 1.567 | 1.624 | 1.802 | 2.069 | 1.764 |
| clean | 1.460 | 1.278 | 1.364 | 1.571 | 2.056 | 1.690 |
| **Tiger** | | | | | | |
| final | 1.637 | 1.561 | 1.553 | 1.633 | 1.584 | 2.330 |
| clean | 1.254 | 0.843 | 0.846 | 0.908 | 1.064 | 2.754 |
| **Wall** | | | | | | |
| final | 7.294 | 6.554 | 6.731 | 7.134 | 7.889 | 9.367 |
| clean | 5.372 | 5.923 | 6.005 | 6.544 | 6.950 | 8.984 |

Table 2: Average EPE per test sequence.

| *Sequence/Pass* | *Method* | | | | | |
|---|---|---|---|---|---|---|
| | LDOF | Classic+NL | Classic+NL-fast | Classic++ | HS | Aniso-HL1 |
| **PMarket_3** | | | | | | |
| final | 8.263 | 4.968 | 5.479 | 5.926 | 6.050 | 7.719 |
| clean | 4.640 | 4.073 | 4.504 | 5.546 | 5.401 | 8.416 |
| **PShaman_1** | | | | | | |
| final | 11.711 | 8.916 | 10.595 | 11.572 | 15.579 | 15.751 |
| clean | 9.925 | 8.154 | 9.797 | 10.846 | 12.820 | 14.995 |
| **Ambush_1** | | | | | | |
| final | 82.346 | 84.098 | 87.434 | 88.229 | 74.622 | 89.089 |
| clean | 73.419 | 74.310 | 80.540 | 79.506 | 66.931 | 87.289 |
| **Ambush_3** | | | | | | |
| final | 44.736 | 51.408 | 50.247 | 52.507 | 49.240 | 50.306 |
| clean | 39.397 | 43.129 | 43.241 | 46.664 | 43.945 | 49.548 |
| **Bamboo_3** | | | | | | |
| final | 5.064 | 4.952 | 4.858 | 5.621 | 5.727 | 5.590 |
| clean | 4.999 | 4.558 | 4.563 | 5.480 | 5.569 | 5.451 |
| **Cave_3** | | | | | | |
| final | 28.916 | 32.652 | 34.124 | 35.759 | 33.064 | 34.766 |
| clean | 27.457 | 29.592 | 32.558 | 33.663 | 31.737 | 36.475 |
| **Market_1** | | | | | | |
| final | 13.516 | 19.211 | 20.785 | 22.166 | 18.467 | 19.736 |
| clean | 11.369 | 15.557 | 18.706 | 19.687 | 14.873 | 22.472 |
| **Market_4** | | | | | | |
| final | 85.665 | 87.722 | 91.556 | 91.126 | 86.743 | 101.195 |
| clean | 88.672 | 88.173 | 94.245 | 91.103 | 91.379 | 104.244 |
| **Mountain_2** | | | | | | |
| final | 5.756 | 5.113 | 5.307 | 5.341 | 5.599 | 5.666 |
| clean | 5.222 | 2.041 | 1.548 | 2.387 | 1.529 | 5.041 |
| **Temple_1** | | | | | | |
| final | 6.133 | 6.289 | 6.307 | 7.403 | 7.344 | 7.248 |
| clean | 5.548 | 5.624 | 5.768 | 7.081 | 6.929 | 7.049 |
| **Tiger** | | | | | | |
| final | 9.688 | 9.584 | 9.528 | 10.946 | 10.386 | 11.478 |
| clean | 9.072 | 7.517 | 7.614 | 8.978 | 8.901 | 10.883 |
| **Wall** | | | | | | |
| final | 23.551 | 27.657 | 27.764 | 29.874 | 27.618 | 27.321 |
| clean | 22.199 | 25.514 | 25.281 | 28.140 | 25.073 | 28.129 |

Table 3: Average EPE per sequence, unmatched regions.

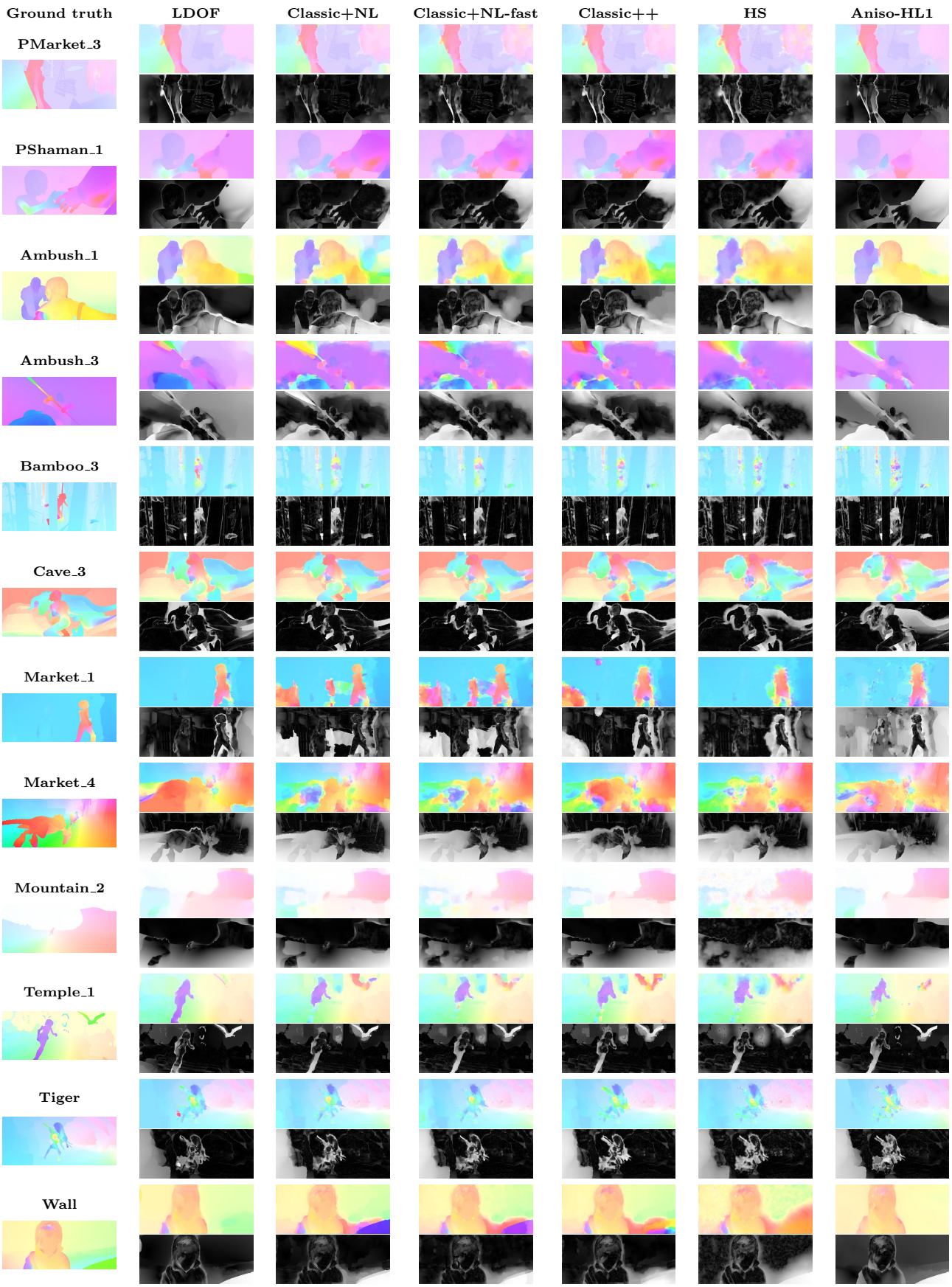


Table 4: Visual overview (final pass).

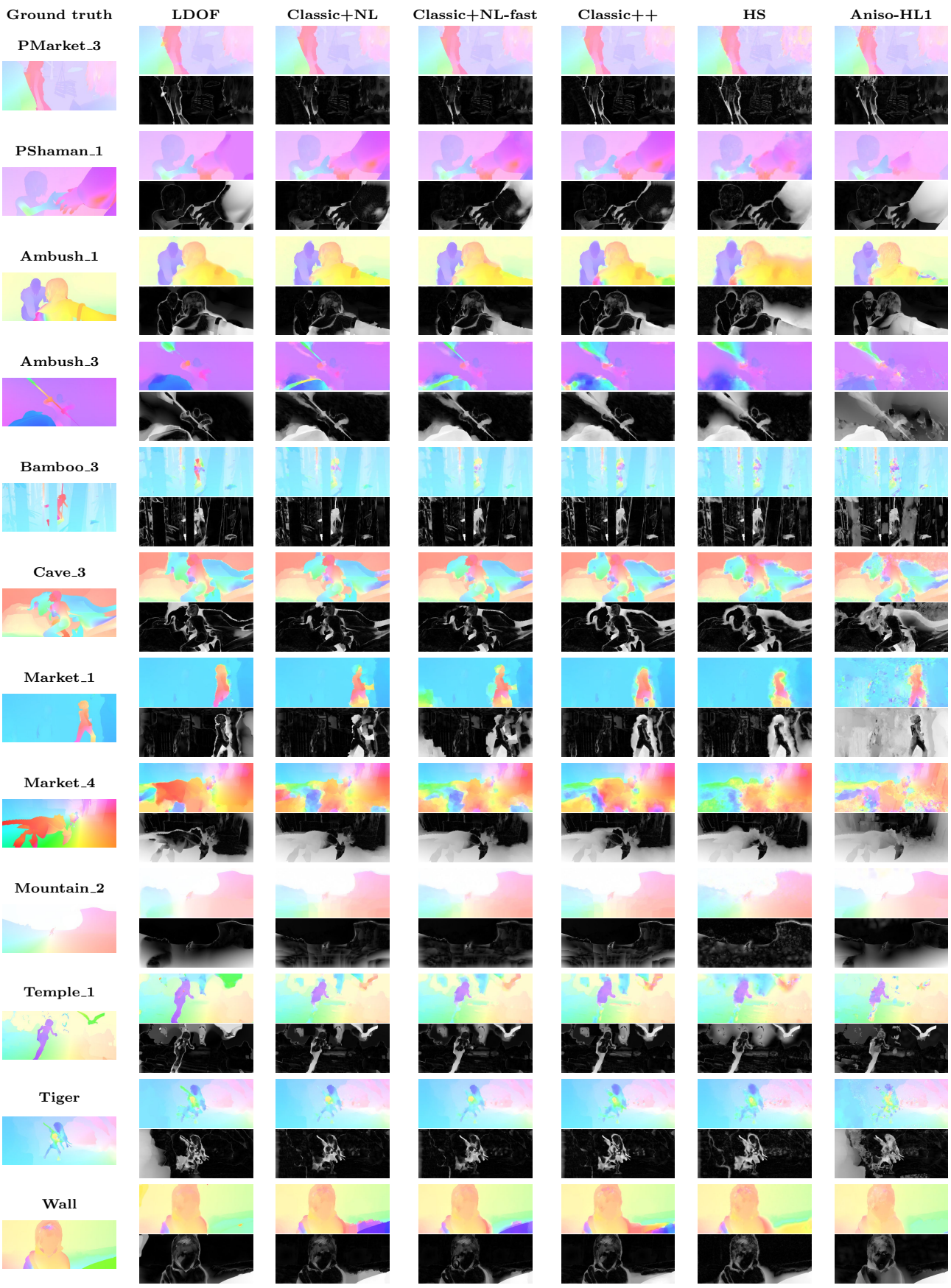


Table 5: Visual overview (clean pass).