Nonrigid Optical Flow Ground Truth for Real-World Scenes with Time-Varying Shading Effects

Wenbin Li, Darren Cosker, Zhihan Lv and Matthew Brown

***Abstract*—In this paper we present a dense ground truth dataset of nonrigidly deforming real-world scenes. Our dataset contains both long and short video sequences, and enables the quantitatively evaluation for RGB based tracking and registration methods. To construct ground truth for the RGB sequences, we simultaneously capture Near-Infrared (NIR) image sequences where dense markers – visible only in NIR – represent ground truth positions. This allows for comparison with automatically tracked RGB positions and the formation of error metrics. Most previous datasets containing nonrigidly deforming sequences are based on synthetic data. Our capture protocol enables us to acquire real-world deforming objects with realistic photometric effects – such as blur and illumination change – as well as occlusion and complex deformations. A public evaluation website is constructed to allow for ranking of RGB image based optical flow and other dense tracking algorithms, with various statistical measures. Furthermore, we present an RGB-NIR multispectral optical flow model allowing for energy optimization by adoptively combining featured information from both the RGB and the complementary NIR channels. In our experiments we evaluate eight existing RGB based optical flow methods on our new dataset. We also evaluate our hybrid optical flow algorithm by comparing to two existing multispectral approaches, as well as varying our input channels across RGB, NIR and RGB-NIR.**

***Index Terms*—Dense Ground Truth, Optical Flow, Near-Infrared Dyes, GRB-NIR Imaging, Multispectral Optical Flow.**

I. Introduction

Tracking is a difficult task involved in many fields e.g. postproduction [1], long term tracking [2], [3], [4], reconstruction [5] and interaction [6]. The quantitative evaluation of optical flow algorithms is a difficult challenge – particularly given long nonrigid scenes with natural noise. The *Middlebury* benchmark [7] and the variations [8], [9], [10] are currently the most widely used *Ground Truth* (GT) in the community. Tracking algorithms which use RGB/Color images may be submitted to the benchmark website for ranking and evaluation. However, this dataset is limited by the lack of object blur, complex nonrigid motion and long image sequences. Most of these limitations are due to the stop-motion method of capture: a scene is first captured under normal lighting; and then a second image of the same scene is captured using ultraviolet lighting. To address these limitations, Butler *et al.* [11] proposed a dataset based on a 3D animated film *Sintel*, which contains inter-frame GT through long sequences and geometric blur under different renderings. However their inherent limitation is the use of synthetic sequences, which lacks real-world photometric effects and textural properties. Similar to *Sintel*, Garg *et al.* [12] rendered synthetic video sequences accompanying with GT by projecting the scene motion (*Motion Capture*) of a realistic waving flag onto an image plane. Although KITTI [13] benchmark enables the evaluation in real-world street scenes, there is still a lack of nonrigid GT for long sequences.

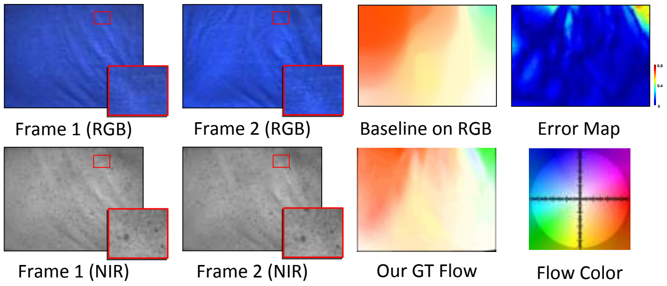


Fig. 1. A baseline algorithm [14] is performed on the RGB channel of one of our ground truth sequences *featureless*. The figure highlights a dense NIR GT patch – used to evaluate RGB based tracking – in an otherwise near-textureless RGB region.

In this paper, we propose such a GT dataset – allowing for the first time the benchmark of dense tracking algorithms on real-world nonrigidly deforming scenes captured at video rate. Sequences may be tracked using the RGB channel, and their performance measured against the GT. The key insight to capture such a dataset is the use of multispectral imaging – in particular, RGB&Near-Infrared (RGB-NIR) imaging which has recently been shown useful in computer vision, e.g. multispectral SIFT [15], image dehazing [16] and registration [17], [18]. A property of such imaging is the ability to apply markers visible in one spectrum (e.g. NIR), but invisible in another (e.g. RGB). Therefore, an algorithm can be applied to the RGB sequence alone, and its performance is then compared to the invisible markers in the NIR channel. To accompany with the data, we provide an evaluation platform which allows users to download the RGB data, upload their tracking results and then view the accuracy versus other methods on our GT.

The second focus of our paper is to investigate how multispectral (RGB-NIR) imaging might improve the quality of tracking, by proposing a multispectral optical flow formulation. The variational optical flow model began with the pioneering

Manuscript received February 28, 2016; Revised June 15, 2016; Accepted June 29, 2016. This paper was recommended for publication by Editor-in-Chief A. Bicchi and Editor J. Kosecka upon evaluation of the reviewers comments. This work was supported by the UK CAMERA EP/M023281/1 and EPSRC projects EP/K023578/1 and EP/K02339X/1. Our public evaluation website: https://vision.cs.bath.ac.uk/flow

Wenbin Li (the corresponding author) and Zhihan Lv with Department of Computer Science, University College London, UK. {w.li, z.lu}@cs.ucl.ac.uk

Darren Cosker with Centre for the Analysis of Motion, Entertainment Research and Applications (CAMERA), University of Bath, UK. d.p.cosker@bath.ac.uk

Matthew Brown with Google, Mountain View, CA 94043, US. m.brown@bath.ac.uk

Digital Object Identifier (DOI): see top of this page.

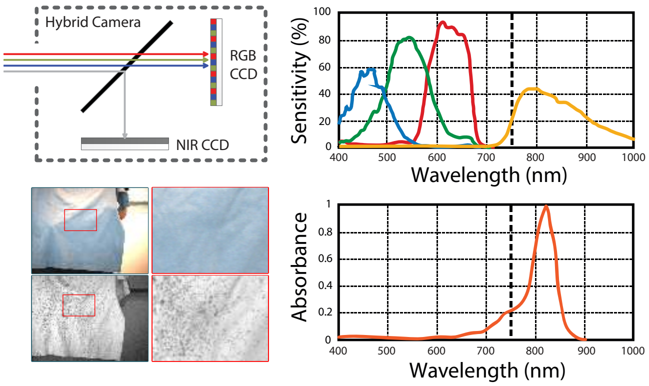


Fig. 2. RGB-NIR Camera and the NIR visible dyes. **Top Left**: The inside structure of the camera. **Bottom Left**: Sample images captured by the RGB CCD sensor and NIR CCD sensor respectively. **Top Right**: The relative transmittance of our RGB CCD sensor and NIR CCD sensor (yellow). **Bottom Right**: The absorbance of the NIR visible dyes respect to various wavelength.

work of Horn and Schunck [19] and Lucas and Kanade [20]. Some complementary concepts have since been developed to deal with the shortcomings of their original models such as spatial discontinuities [21], large displacements [14], motion detail loss through coarse-to-fine minimization [22] and local smoothness [23]. Of these methods, Xu *et al.*'s (MDP) [22] approach is currently amongst top 3 (by average) in the *Middlebury* evaluation while the Li *et al.*'s (LME) [23] approach has state-of-the-art performance given nonrigid surface motion [12]. However, all of these methods are applied on image pairs within the visible spectrum (RGB/Color) and are sensitive to motions in large featureless regions in which the basic *Intensity Consistency* assumption is weakened.

To take an advantage of extra spectrums, Markandey and Flinchbaugh [24] consider the IR image within an optical flow method, which solves a system of two data terms (RGB/Grayscale and IR). They assume an equal contribution (a known fixed weight) from both channels. This may reduce the precision in some cases (Fig. 5). Barron and Klette [25] propose an approach using all three individual color channels, and shows improvement over the grayscale.

Contributions

To summarize there are two major contributions in our paper: ***(1)*** we present a nonrigid GT dataset (Fig. 1) for RGB image based dense tracking (e.g. optical flow) methods, and an evaluation website allowing users to rank the performance of their method versus others. The dataset contains dense inter-frame correspondences from eight short and five long sequences with varying photometric properties; and ***(2)*** we present a multispectral (RGB-NIR) optical flow model (*vnflow*). Within this model, we propose a novel weighting scheme which adoptively selects the best available image features in either the RGB or the complementary NIR channel to enhance motion analysis.

In our experiments, we evaluate ten existing RGB based optical flow methods on our dataset - ranking them based on various statistics (the same presented on our evaluation website). We then turn the attention to our *vnflow* method which illustrates the potential benefit of using combined spectra (e.g. RGB-NIR) for optical flow estimation.

II. Non-Rigid Ground Truth Dataset

Ground Truth (GT) for RGB/Color optical flow is difficult to capture – how does one simultaneously acquire an invisible set of GT positions upon which to evaluate algorithm performance on the visible RGB channel? One important advance in this area was proposed by Baker *et al.* with the introduction of the *Middlebury* benchmark [7]. Due to their contribution, the optical flow community has rapidly developed in recent years. However, Baker *et al.* also point out limitations of their work – the lack of object blur and occluded motion – which are discussed in more recent state-of-the-art datasets [11]. The main limitations of current benchmarks , which we address in this work, are as follows:

Long Image Sequences: As discussed in [11], most of the *Middlebury* sequences are short in length, which leads to a lack of evaluation on long term correspondence. While *Sintel* provides long synthetic sequences (more than 50 frames) and GT for each pair of frames, our dataset provides long sequences from real-world objects – thus exhibiting realistic photometric effects and textural properties.

Realistic Noise: The lack of realistic blur is a common issue in both *Middlebury* and *KITTI*. Our dataset includes realistic camera blur and other noise, e.g. strong shadows, reflectance and illumination changes.

Complex Nonrigid Motions: Unlike *Middlebury* and *Sintel*, our dataset is specifically focused on nonrigid motion, containing examples of stretching, large bends and creases.

A. RGB-NIR Imaging System

In order to acquire our GT, we construct a controllable scene (i.e. lighting and motion properties) using an *RGB-NIR Imaging System* and *NIR Visible Dyes*.

RGB-NIR Camera: In this paper, a hybrid camera (JAI AD-080GE) is used to capture both RGB and NIR images from the same scene simultaneously. Fig. 2 shows internal construction of the camera, where natural light is split onto the RGB and NIR CCD sensors respectively. As opposed to experimental bench-based RGB-NIR beam-splitter setups [26], the overall system is both compact and portable. Such a system simultaneously captures a series of continuous images in both the RGB and NIR channels at 20 FPS.

NIR Visible Dyes: In order to generate dense features on object surfaces for our GT dataset, we utilize *NIR Visible Dyes* (NIR819D, QCR Solutions Corp.) which absorb the spectrum in a range of approximately 700 to 870 nm with a peak at around 819 nm. Our *NIR Visible Dyes* are spread onto object surfaces in order to generate fine patterns of which the diameter is within 1 mm, with a maximum 2 mm distance between any pair of neighboring patterns. Fig. 2 shows dense patches painted by our dyes that are visible in the NIR channel while remaining invisible in the RGB channel. To illustrate the statistical dependencies of the patches between different bands, 20,000 RGB-NIR patches (3×3 pix.) with the dyes applied are randomly selected and plotted as pairwise

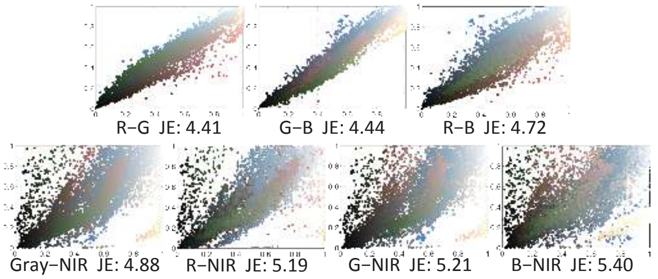


Fig. 3. Pairwise distributions for the RGB and NIR channels of 20,000 sampled patches from our ground truth dataset.

distributions using joint entropy in Fig. 3. Note that we compute the joint entropy as $H(X,Y) = -\sum_{X,Y} P(X,Y)log_2[P(X,Y)]$. It is observed that the joint entropy of $\{$R,G,B,Gray$\}$-NIR is larger than between the visible bands (R,G,B). Therefore, the *NIR Visible Dyes* provide richer visible information apart from RGB channel – making it suitable for a GT basis especially in largely textureless RGB channel regions. The Middleburry benchmark adopts UV-flourescent dye, only visible under the UV lighting. This leads to an issue that they have to stop the object movement and switch to UV lighting when they need to capture the dye-featured image. In this case, their sequences may lose the blurry photometric effects. However, our *NIR Visible Dyes* together with *RGB-NIR Camera* allows a continuous capture (up to 20 FPS) which is able to preserve the blur and other real-world photometric effects possible. In practice, our dyes give the best invisibility on the cotton surface, but are hard to remain completely invisible on other materials. Our dyes cannot be applied to human for long because it would harm the skin in some cases.

Motion Control Component: To precisely control the displacement of objects in our GT scenes, a motion control mechanism is constructed using LEGO NXT Mindstorms robotics kits which produce controllable and uniform inter-frame movements for our GT surfaces.

In the following section, we describe dataset construction together with our proposed evaluation methods.

B. Ground Truth Estimation

Once we have obtained the corresponding pairs of RGB and NIR images, we use the feature-rich NIR channel to construct a dense GT flow field. In this subsection we describe this process and other important properties in detail.

Image Properties: Our *RGB-NIR camera* captures images at 1296×966 pixels. The *Motion Control Component* of our system allows us to precisely range motions from subpixel to 40 pixels. Similar to *Middlebury*, all the captured RGB sequences are downsampled by a factor of 3, resulting in an image size of 432×322 after the *Subpixel Motion Estimation* step (presented later in this subsection).

Data Acquisition: To capture the data properly, we set up a capture system using our *Motion Control Component* and *RGB-NIR Camera*. The motor (NXTMotor) of the component is able to precisely rotate by 1 degree step. Together with the Lego bricks and bars, *Motion Control Component* precisely controls the motion of the object surfaces in the scale of $[1,46]$ cm. Most of the motion represented in the dataset is parallel to the camera plane. Furthermore, Our camera is distortion free and follows a pinhole model. In this case, the calibration aims to find out the relation f between the object movement M_{MCC} (in cm) and the pixel displacement M_p within the image space. Here we have $M_p = f(M_{MCC})$. In practice, we fix the distance (1.5m) between the objects and camera while the camera forward direction is perpendicular to the object surface. We then capture the surface motion with a certain M_{MCC}; and manually measure the M_p from the image. We repeat this process before capturing each of sequences and obtain the size of search window $2M_p \times 2M_p$ pix. for the following *Pixel Correspondence* estimation.

Pixel Correspondence: We use a parfum spray to generate fine patterns on the objects. In most cases, the diameters of such patterns are smaller than 1 mm, corresponding to approximately 1 pixel of the image (Fig. 5 (Left)). And those patterns are still highly variable in terms of intensity and shape. Therefore pixel correspondences are achieved by matching the dye patterns between neighboring NIR images. Unlike the Color-*SSD* tracker used in *Middlebury*, we consider both intensity and shape. A SIFT descriptor with 128 dimensions is computed for each pixel in an image. We nominate a GT match between pixels where the *Euclidean Distance* of their SIFT vectors is smallest within a given search window. This window size ($2M_p \times 2M_p$ pix.) is predefined using the maximum motion in the *Motion Control Component*. To improve robustness we examine the matched results across adjacent frames. A correspondence is labeled with a value "*NAN*" (Not-A-Number) if the intensity difference between the forward matched result and the backward matched result is greater than a threshold. The region mask containing "*NAN*" values is recorded as an occlusion map. Note that we do not apply an existing optical flow method onto the NIR images to estimate the correspondences. Although the optical flow is able to give us the per-pixel dense correspondence, the encoded smoothness term may overly smooth the motion at the object boundaries and small motion details. This would further reduce the precision of the GT.

Subpixel Motion Estimation: After obtaining GT pixel correspondence, we follow the *Middlebury* subpixel motion estimation process. We apply the *Lucas-Kanade* kernel [20] to each search window for subpixel motion using 1/20 pixel precision. We then calculate the average of up to 9 motion vectors in each 3×3 window in order to down-sample the motion field to dimension 432×322.

Realistic Noise: The controllable nature of our *RGB-NIR Imaging System* allows us to incorporate varieties of noise and artefacts into our GT dataset. We increase the exposure time of the RGB CCD sensor to bring object blur into the visible channel, while using a suitably fast exposure time on the NIR CCD sensor to capture a corresponding blur-free image. Alternatively, defocus blur could also be obtained by modifying the aperture settings. Shadow and illumination changes are generated using infrared-free light (LED lighting), leading to realistic shadows in the RGB channel without affecting illumination in NIR channel (Fig. 4).

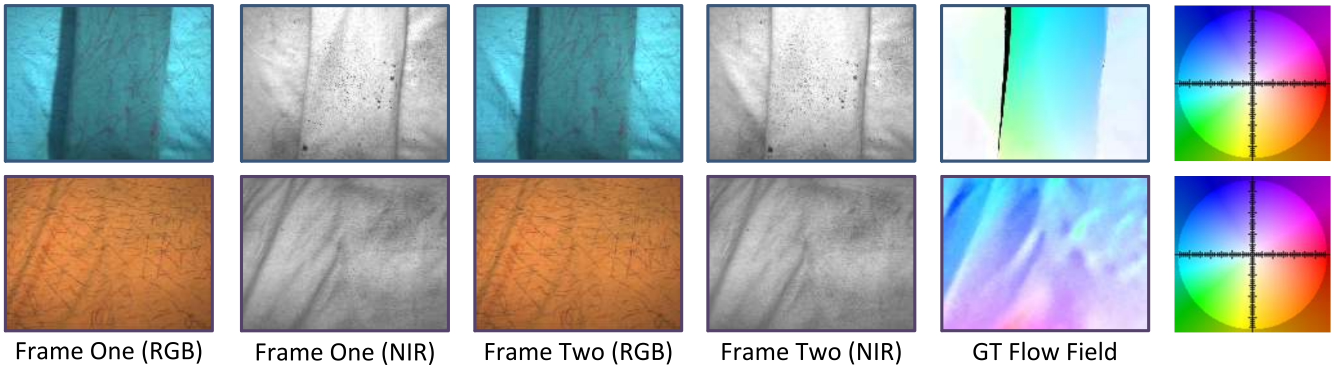


Fig. 4. Examples from our ground truth dataset. **Top Row:** *str.shadow* contains strong shadows and subtle nonrigid motion. **Bottom Row:** *crush* (frames 38 and 39) is a video sequence containing complex nonrigid deformations and self occlusions.

***Sequence Descriptions**:* Here we provide labels for each of the sequences in our dataset, as well as brief descriptions of their characteristics. Our dataset contains two types of GT sequence – *Short Sequences* and *Long Sequences*. We capture eight short sequences in total, each of which contains ten frames with dense GT between middle pair of images. Each sequence is captured so as to include specific common image properties. In terms of sequence naming, *single* contains nonrigid motion of single object. *illumination* contains strong reflectance and illumination change while both *mObjs* and *triObjs* contain multiple objects with nonrigid movement. *featureless* contains small motions for a featureless object surface while *crease* contains a large crease on multiple objects. *blur* and *str.shadow* contain both camera blur and strong shadows respectively. In addition, five longer sequences are captured with dense inter-frame GT for every neighboring image pair. Each sequence contains 50 frames and includes multiple realistic photometric effects and nonrigid motion. *mBlur* contains focus blur, motion blur and large displacements, while *circle* contains complex creases. *crush* presents a large crushing movement with self occlusions and *stretch* shows elastic deformation. Finally, *wave* presents a real-world waving cloth. We also provide training set which contains 3 short and 3 other long GT sequences.

Fig. 4 shows two sample sequences (*str.shadow* and *crush*) from our dataset where tracking algorithms are executed on the RGB data, and the NIR channel - with the aid of NIR visible dyes - provides our GT flow fields upon which to compare to the RGB flow fields. In the following section, we introduce our evaluation methods along with the public website to openly evaluate algorithms.

C. Evaluation Methods and Statistics

Similar to *Middlebury*, we provide tests of *Endpoint Error* (EE) and *Angle Error* (AE). Users are expected to download the RGB data from our evaluation platform, and compute flow fields between all frames in the *Long Sequences* and for one image pair for each sequence in the *Short Sequences*. Users then upload their result and our evaluation system compares it to the GT flow fields calculated on our NIR channel (which includes the NIR visible dyes). For robustness statistics, we perform *Average* (Avg.), *Accumulated* (Acc.), *Standard Deviations* (SD), *RX* and *AX* [7] where *RX* presents the percentage of pixels that have an error reading above X; And *AX* is for the accuracy of the error reading at the Xth percentile, after sorting the errors from low to high. Avg., SD and {A50, A75, A99, A100} are given for both EE and AE; {R0.5, R0.75, R1, R2} are performed for EE; Acc. is calculated for EE in long sequences only; {R2, R5, R7.5, R10} are computed for AE.

As shown in Fig. 6, we generate a comparison table for cross-evaluation of user uploaded flow field results against any other methods previously uploaded to our evaluation system. For long sequences, we can plot results selected by the user with respect to a specific frame index.

III. RGB-NIR Variational Optical Flow Model

In the previous sections, we described a GT dataset and evaluation website for algorithms operating on RGB data. In this section, we now slightly change focus and introduce a novel algorithm which combines both RGB and NIR channels in such a way as to maximize the distinguishing information from each channel. Certain visual information can be poorly represented in an RGB channel. It is therefore prudent in many cases to also consider the NIR channel (and vice-versa). In our evaluation section we examine these properties in more detail. Note that for fairness, our public RGB-only evaluation website does not include results of our *vnflow* or any future multispectral methods.

Our algorithm considers a pair of consecutive frames in an image sequence. The current frame is denoted by $I_1(\mathbf{x})$ and its successor is $I_2(\mathbf{x})$ where $I = (V,N)^T$, $\{V : \Omega \subset \mathbb{R}^3 \to \mathbb{R}\}$ represents a rectangular image in the RGB channel and $\{N : \Omega \subset \mathbb{R}\}$ denotes a rectangular image in the NIR channel. Both V and N are aligned and share the same Cartesian coordinate where $\mathbf{x} = (x,y)^T$ is a pixel location. The optical flow displacement between $I_1(\mathbf{x})$ and $I_2(\mathbf{x})$ is defined as $\mathbf{w} = (u,v)^T$. Our proposed optical flow approach leads to the following energy function:

$$E(\mathbf{w}) = (1-\lambda(\mathbf{x}))E_V(\mathbf{w}) + \lambda(\mathbf{x})E_N(\mathbf{w}) + \gamma E_S(\mathbf{w}) \qquad (1)$$

where the *Visible RGB Energy* $E_V(\mathbf{w})$ contains both *Intensity Constancy* and *Gradient Constancy* assumptions between the visible components $V_1(\mathbf{x})$ and $V_2(\mathbf{x})$ of the images while our main contribution i.e. *Invisible NIR Energy* is represented as the term $E_N(\mathbf{w})$. A high-order regularization $E_S(\mathbf{w})$ is adopted.

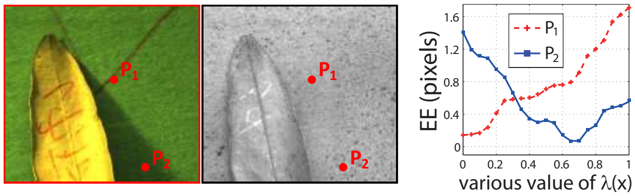


Fig. 5. *Endpoint Error* (EE) affected by varying weight $\lambda(\mathbf{x})$. **Left**: A patch of *LeafShadow* is shown where two points P_1 and P_2 are plotted in RGB and NIR channels respectively. **Right**: EE for both points P_1 and P_2 are plotted by varying weight $\lambda(\mathbf{x})$.

Visible RGB Energy: Following the *Intensity Constancy* assumption, we assume that the intensity of a pixel is not varied by its displacement throughout an image sequence. In addition, we also make a *Gradient Constancy* assumption [14] to provide additional stability where pixel intensity is violated by illumination changes. The *Visible RGB Energy* term encoding these assumptions is thus formulated as:

$$E_V(\mathbf{w}) = \int_\Omega \phi(\|V_2(\mathbf{x}+\mathbf{w}) - V_1(\mathbf{x})\|^2)d\mathbf{x} + \theta \int_\Omega \phi(\|\nabla V_2(\mathbf{x}+\mathbf{w}) - \nabla V_1(\mathbf{x})\|^2)d\mathbf{x} \qquad (2)$$

For robustness against occlusions and boundary blur, we apply the increasing concave function $\phi(s^2) = \sqrt{s^2+\varepsilon^2}$ with $\varepsilon = 0.001$ to solve this formation. The remaining term $\nabla = (\nabla_x, \nabla_y)^T$ is a spatial gradient and $\theta \in [0,1]$ denotes a linear weight. The smoothness term is a dense pixel based regularizer that penalizes global variation. The objective is to produce a globally smooth constraint:

$$E_S(\mathbf{w}) = \int_\Omega \phi(\|\nabla u\|^2 + \|\nabla v\|^2)d\mathbf{x} \qquad (3)$$

Invisible NIR Energy: A *visible RGB Energy* term is widely used in optical flow [14] but error-prone in featureless regions or unclear boundaries. We therefore propose to inspect additional spectral channels given these situations. We include an *Invisible NIR Energy* term as a complementary assumption to the classic framework, namely to introduce additional texture information to optical flow estimation. Similar to the RGB *Intensity Constancy*, we assume that the pixel intensity in the NIR channel is not changed by displacement, which yields an energy term as follows:

$$E_N(\mathbf{w}) = \int_\Omega \phi(\|N_2(\mathbf{x}+\mathbf{w}) - N_1(\mathbf{x})\|^2)d\mathbf{x} \qquad (4)$$

Where the term $E_N(\mathbf{w})$ presents the continuous energy in the NIR channel. Note that both terms $E_V(\mathbf{w})$ and $E_N(\mathbf{w})$ share the same spatial smoothness regularizer.

Detail-Aware Weight $\lambda(\boldsymbol{x})$*:* In Fig. 5 (Left) we show an image patch in which two points P_1 and P_2 are plotted. The small region centered on P_2 contains soft shadow in the RGB channel but has more distinguishing features in the NIR channel. For the point P_1, the situation is opposite. The *Endpoint Error* (EE) with respect to the different $\lambda(\mathbf{x})$ values are plotted in Fig. 5 (Right). We observe that plain texture leads to larger errors in the optical flow estimation. Dynamically taking more contribution from the channel containing more detailed texture is therefore adopted in our method.

A. Minimization Framework

Prior to energy minimization, $\lambda(\boldsymbol{x})$ *Initialization* is performed to improve overall optical flow energy in featureless regions. A numerical scheme is then applied to minimize the continuous RGB-NIR energy within a pyramidal framework. Both steps are described in following sections.

$\lambda(\boldsymbol{x})$ *Initialization.:* Inspired by the kernel-based edge detector where an *Intensity Gradient* is used to represent geometric information in the texture space, we define a weight $\{\lambda(\mathbf{x}) : \mathbb{R} \mapsto [0,1]\}$ using an *Intensity Gradient* as follows:

$$\lambda(\mathbf{x}) = \left(1 + \exp\left\{-a\left(\frac{|\nabla N_1(\mathbf{x})|}{|\nabla V_1(\mathbf{x})| + |\nabla N_1(\mathbf{x})|} - b\right)\right\}\right)^{-1}$$

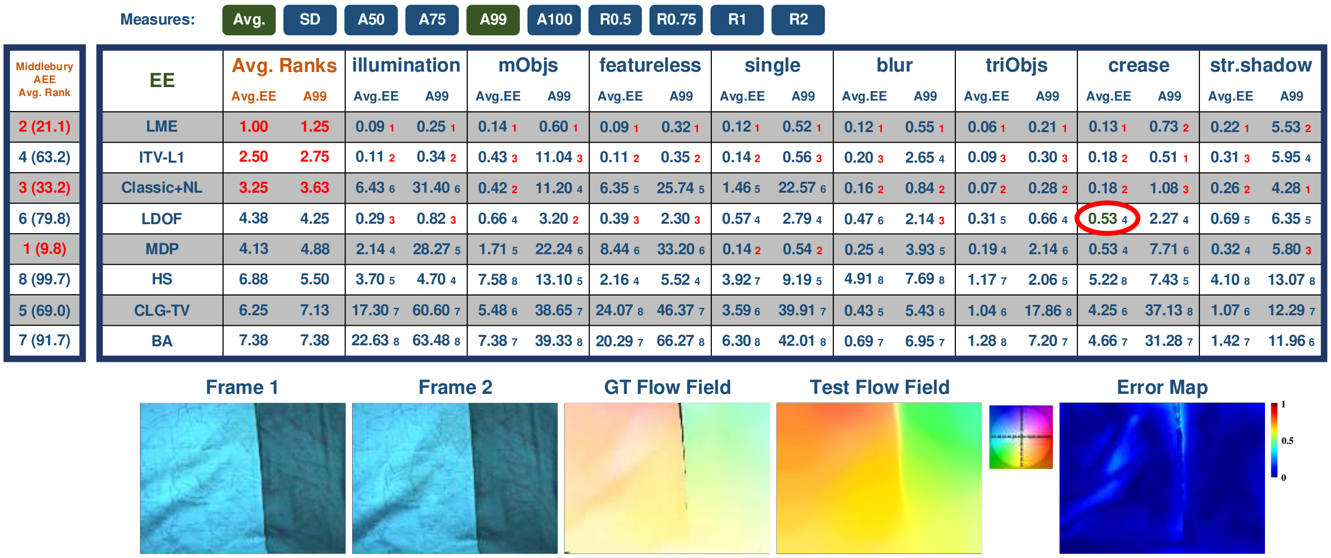
where $\mathbf{x}$ denotes a pixel location while $\nabla = (\nabla_x, \nabla_y)^T$ presents the intensity gradient calculated using a 3×3 *Sobel Kernel*; a and b are parameters chosen to be 10 and 0.5 respectively. The weight $\lambda(\mathbf{x})$ is intensity-dependent and can be precalculated before energy minimization. Given an n-level image pyramid, the input images I_1, I_2 and the weight map $\lambda(\mathbf{x})$ are resized to the same scale on each level. These are denoted by $I_1^i = (V_1^i, N_1^i)^T$, $I_2^i = (V_2^i, N_2^i)^T$ and λ^i, and are used in the following energy minimization phase.

RGB-NIR energy optimization.: In this process, we aim to find the global minimum of the energy in Eq. (1) which is continuous but highly nonlinear. We need to remove the nonlinearity and obtain the final linear system. Thus, we apply nested fixed point iterations on $\mathbf{w}$ by mainly following the numerical scheme in [27]. In the implementation, the image pyramid is constructed using a downsampling of 0.75. The final linear system is solved with successive over-relaxation. For more details of our optimization scheme, please refer to the supplementary document.

IV. Experiments

In this section, we evaluate ***(1)*** eight publicly available optical flow algorithms from *Middlebury* using our nonrigid GT dataset (executed on the RGB channel, and compared against the NIR GT flow fields), and ***(2)*** our proposed multispectral optical flow method (*vnflow*) comparing to two multispectral approaches, highlighting the potential benefits of *Detail-Aware Weighting*.

We consider ten baseline methods in our experiments. Eight of those is executed on the RGB channel of our dataset. The remaining two (MCOF and COF) are evaluated using the NIR channel (and invisible NIR dye GT). Algorithms from Xu *et al.* (MDP) [22] (AEE rank 4/119) and Li *et al.* (LME) [23] (rank 11) are state-of-the-art optical flow methods. The former has leading performance in the *Middlebury* evaluation while the latter achieves the state-of-the-art results on Garg *et al.* [12]. *Combined local-global Optical Flow* (CLG-TV) [28] (AIE rank 10/119) highlights the utility of bilateral filtering and anisotropic regularization, which gives high performance in image interpolation. *Large Displacement Optical Flow*



Measures: Avg. SD A50 A75 A99 A100 R0.5 R0.75 R1 R2

Middlebury AEE Avg. Rank	EE	Avg. Ranks		illumination		mObjs		featureless		single		blur		triObjs		crease		str.shadow	
		Avg.EE	A99	Avg.EE	A99	Avg.EE	A99	Avg.EE	A99	Avg.EE	A99	Avg.EE	A99	Avg.EE	A99	Avg.EE	A99	Avg.EE	A99
2 (21.1)	LME	1.00	1.25	0.09_1	0.25_1	0.14_1	0.60_1	0.09_1	0.32_1	0.12_1	0.52_1	0.12_1	0.55_1	0.06_1	0.21_1	0.13_1	0.73_2	0.22_1	5.53_2
4 (63.2)	ITV-L1	2.50	2.75	0.11_2	0.34_2	0.43_3	11.04_3	0.11_2	0.35_2	0.14_2	0.56_3	0.20_3	2.65_4	0.09_3	0.30_3	0.18_2	0.51_1	0.31_3	5.95_4
3 (33.2)	Classic+NL	3.25	3.63	6.43_6	31.40_6	0.42_2	11.20_4	6.35_5	25.74_5	1.46_5	22.57_6	0.16_2	0.84_2	0.07_2	0.28_2	0.18_2	1.08_3	0.26_2	4.28_1
6 (79.8)	LDOF	4.38	4.25	0.29_3	0.82_3	0.66_4	3.20_2	0.39_3	2.30_3	0.57_4	2.79_4	0.47_6	2.14_3	0.31_5	0.66_4	0.53_4	2.27_4	0.69_5	6.35_5
1 (9.8)	MDP	4.13	4.88	2.14_4	28.27_5	1.71_5	22.24_6	8.44_6	33.20_6	0.14_2	0.54_2	0.25_4	3.93_5	0.19_4	2.14_6	0.53_4	7.71_6	0.32_4	5.80_3
8 (99.7)	HS	6.88	5.50	3.70_5	4.70_4	7.58_8	13.10_5	2.16_4	5.52_4	3.92_7	9.19_5	4.91_8	7.69_8	1.17_7	2.06_5	5.22_8	7.43_5	4.10_8	13.07_8
5 (69.0)	CLG-TV	6.25	7.13	17.30_7	60.60_7	5.48_6	38.65_7	24.07_8	46.37_7	3.59_6	39.91_7	0.43_5	5.43_6	1.04_6	17.86_8	4.25_6	37.13_8	1.07_6	12.29_7
7 (91.7)	BA	7.38	7.38	22.63_8	63.48_8	7.38_7	39.33_8	20.29_7	66.27_8	6.30_8	42.01_8	0.69_7	6.95_7	1.28_8	7.20_7	4.66_7	31.28_7	1.42_7	11.96_6

Fig. 6. Screen shot of our public evaluation website for RGB/Color based dense tracking/optical flow algorithms. Sample scores for the *short* sequences are shown, and we demonstrate the *Endpoint Error* (EE) evaluation. Multiple statistics/measures (Sec. II-C) can be manually selected on the top of the table and illustrated as sub-columns within a sequence where the subscripts show the rank in that sub-column. The user can mouse-click any of the results to show sequence details, the proposed flow field and the error map against the ground truth (as shown on the bottom row). All methods are listed in order of their average rank (Avg. Ranks).

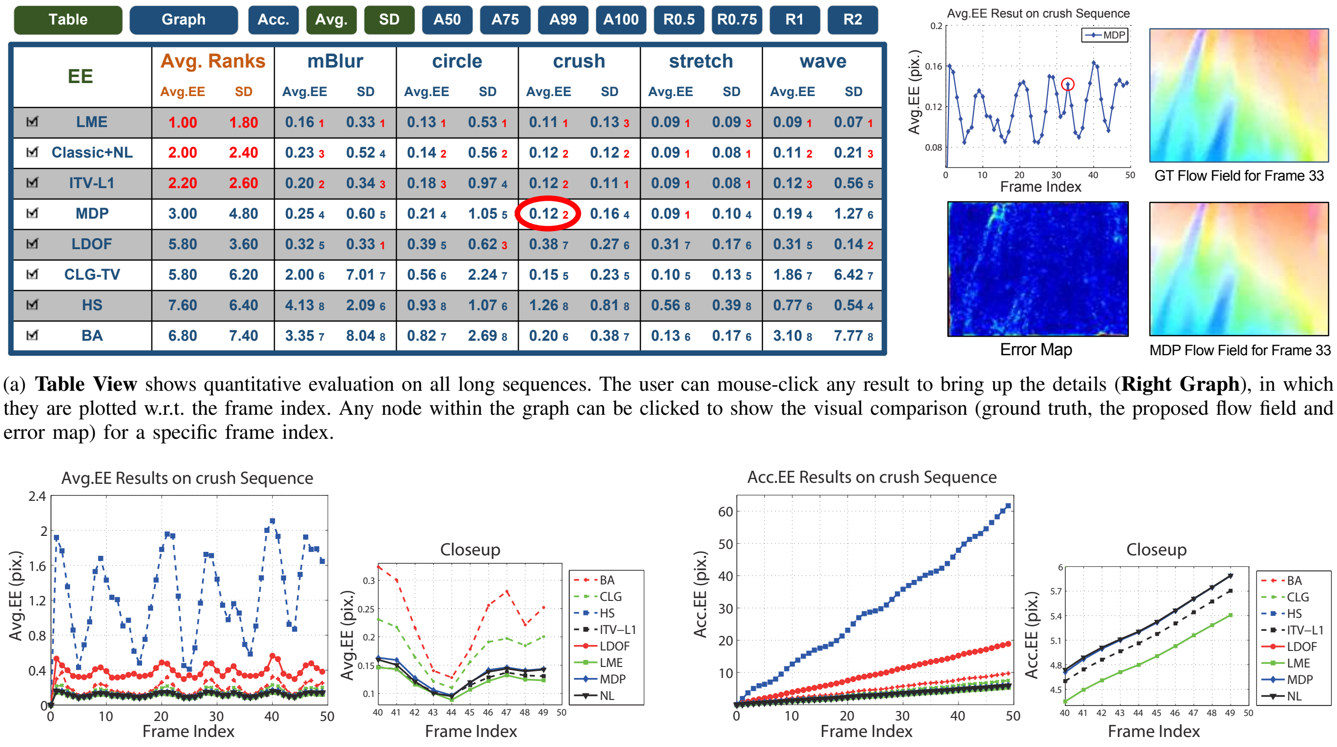


Table Graph Acc. Avg. SD A50 A75 A99 A100 R0.5 R0.75 R1 R2

EE	Avg. Ranks		mBlur		circle		crush		stretch		wave	
	Avg.EE	SD	Avg.EE	SD	Avg.EE	SD	Avg.EE	SD	Avg.EE	SD	Avg.EE	SD
☑ LME	1.00	1.80	0.16_1	0.33_1	0.13_1	0.53_1	0.11_1	0.13_3	0.09_1	0.09_3	0.09_1	0.07_1
☑ Classic+NL	2.00	2.40	0.23_3	0.52_4	0.14_2	0.56_2	0.12_2	0.12_2	0.09_1	0.08_1	0.11_2	0.21_3
☑ ITV-L1	2.20	2.60	0.20_2	0.34_3	0.18_3	0.97_4	0.12_2	0.11_1	0.09_1	0.08_1	0.12_3	0.56_5
☑ MDP	3.00	4.80	0.25_4	0.60_5	0.21_4	1.05_5	0.12_2	0.16_4	0.09_1	0.10_4	0.19_4	1.27_6
☑ LDOF	5.80	3.60	0.32_5	0.33_1	0.39_5	0.62_3	0.38_7	0.27_6	0.31_7	0.17_6	0.31_5	0.14_2
☑ CLG-TV	5.80	6.20	2.00_6	7.01_7	0.56_6	2.24_7	0.15_5	0.23_5	0.10_5	0.13_5	1.86_7	6.42_7
☑ HS	7.60	6.40	4.13_8	2.09_6	0.93_8	1.07_6	1.26_8	0.81_8	0.56_8	0.39_8	0.77_6	0.54_4
☑ BA	6.80	7.40	3.35_7	8.04_8	0.82_7	2.69_8	0.20_6	0.38_7	0.13_6	0.17_6	3.10_8	7.77_8

Avg.EE Resut on crush Sequence
MDP
Avg.EE (pix.)
0.2
0.16
0.12
0.08
0
10
20
30
40
50
Frame Index
GT Flow Field for Frame 33
Error Map
MDP Flow Field for Frame 33

(a) **Table View** shows quantitative evaluation on all long sequences. The user can mouse-click any result to bring up the details (**Right Graph**), in which they are plotted w.r.t. the frame index. Any node within the graph can be clicked to show the visual comparison (ground truth, the proposed flow field and error map) for a specific frame index.

(b) **Graph View** plots details for each sequence. The user can select multiple baseline methods by clicking their checkboxes then clicking the *Graph* option on top of the table. The measure details, e.g. Avg.EE and Acc.EE, are plotted onto the downloadable graphs for each sequence.

Fig. 7. Screen shot of our public evaluation website for long sequences, illustrating the *Endpoint Error* (EE) evaluation.

(LDOF) [14] (AEE rank 89) is a variational model integrating rich feature descriptors and is designed to overcome large displacement issues. Classic+NL [29] (rank 28) improves the TV-L1 framework by combining a Lorentzian penalty and a median filtering heuristic. *Horn and Schunck* (HS) [19] (rank 108), *Black and Anandan* (BA) [21] (rank 101) and *Improved TV-L1* (ITV-L1) [30] (rank 56) are classic modelswidely used in real-world image registration. MCOF [24] is considered as the classic approach using both RGB and NIR channels while COF is a robust implementation of [25] using additional smoothness constraint [19] and coarse-to-fine optimization [27]. Those selected baselines may not cover all the state-of-the-art methods of the community but are able to represent strength/performance in all typical measures.

EE	Avg. Ranks		illumination		mObjs		featureless		single		blur		triObjs		crease		str.shadow	
	Avg.EE	A100	Avg.EE	A100	Avg.EE	A100	Avg.EE	A100	Avg.EE	A100	Avg.EE	A100	Avg.EE	A100	Avg.EE	A100	Avg.EE	A100
vnflow.DA	1.00	1.50	0.01 1	0.10 1	0.02 1	0.53 1	0.02 1	0.27 2	0.02 1	0.19 1	0.02 1	0.45 1	0.01 1	0.23 2	0.04 1	8.72 3	0.03 1	6.48 1
vnflow <1>	4.38	4.38	0.54 5	7.24 5	1.59 5	16.43 5	0.12 4	2.05 4	0.10 3	0.93 3	0.16 5	6.10 5	0.15 5	3.64 5	0.39 5	8.75 5	0.22 3	7.38 3
vnflow <0.5>	6.13	6.13	1.07 6	14.43 6	4.14 6	28.04 7	22.19 7	41.02 7	0.18 5	1.83 6	0.28 7	6.18 7	0.29 6	6.70 7	0.65 6	9.21 7	0.35 6	7.29 2
vnflow <0>	7.75	7.50	2.15 8	27.15 8	5.03 7	58.14 8	44.37 8	82.01 8	0.34 8	3.62 8	0.51 8	6.96 8	0.58 8	13.01 8	1.18 8	14.52 8	0.62 7	7.57 4
MCOF	7.00	6.50	2.11 7	16.70 7	8.56 8	17.11 6	19.88 6	15.52 6	0.30 7	1.66 5	0.26 6	6.09 6	0.52 7	6.06 6	1.06 7	8.73 4	2.12 8	9.97 8
COF	4.50	5.13	0.49 4	4.33 4	0.78 4	3.16 4	12.24 5	8.01 5	0.21 6	2.62 7	0.15 4	4.33 4	0.14 4	2.01 4	0.26 4	9.11 6	0.23 5	8.43 7
LME.NIR	2.75	2.13	0.04 2	0.11 2	0.07 2	0.68 2	0.05 2	0.19 1	0.05 2	0.20 2	0.10 2	3.74 3	0.04 2	0.19 1	0.09 2	4.47 1	0.13 2	8.49 5
LME.RGB	3.13	3.25	0.09 3	0.47 3	0.14 3	2.63 3	0.09 3	0.82 3	0.12 4	1.11 4	0.12 3	1.10 2	0.06 3	0.60 3	0.13 3	5.31 2	0.22 3	9.21 6

Fig. 8. Avg.EE and A100 results of CMOF [24], COF [25] and *vnflow* variations: *Detail-Aware Weight* (**DA**) and the fixed weights (**0**, **0.5** and **1**).

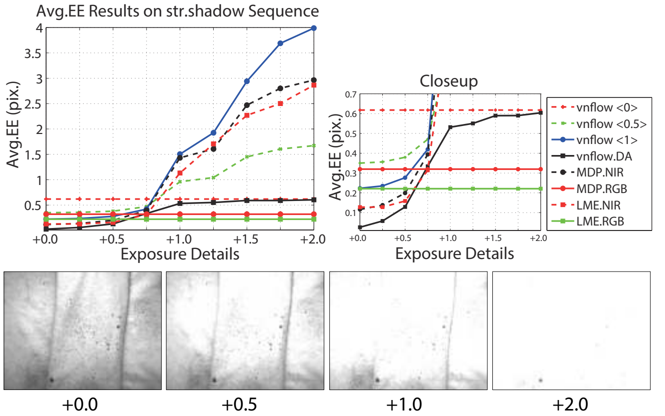


Fig. 9. Avg.EE measures for *vnflow* on *str.shadow* sequence by varying the exposure (feature distribution) in the NIR channel.

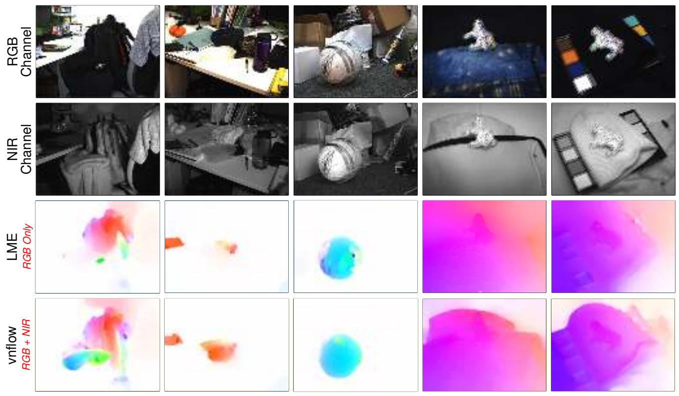


Fig. 10. Visual results of *vnflow* on real-world sequences (**From Left To Right**) *hat*, *office*, *football*, *arts* and *dark*. Note that we show the overlaps of two input images in the top two rows. Hence the blur is not caused by the image quality.

We first perform an evaluation on the short sequences of our GT dataset (i.e. each of the above algorithms are executed on the RGB channel only). Fig. 6 shows a screen shot of our public evaluation website where eight optical flow methods are quantitatively compared to each other using their default parameter settings. Note that the relative *Middlebury* AEE rank (Average rank, captured on Feb. 23, 2016) of the baseline methods is also listed for comparison. We observe that LME leads all trials in Avg.EE. ITV-L1 and Classic-NL respectively rank 2.50 and 3.25 in general Avg.EE. The former outperforms most other algorithms in *featureless* while the latter shows more robust toward flow discontinuities (*mObjs*, *triObjs* and *crease*) and blur motion (*blur*). Note that most methods have a large error (>0.5 Avg.EE.) for *illumination* because the strong illumination change violates the *Intensity Consistency*. In this case, LME (Avg.EE 0.09), ITV-L1 (Avg.EE 0.11) and LDOF (Avg.EE 0.29) give higher performance over the other methods.

Interestingly, compared to *Middlebury* the short sequences of our dataset result in a significantly different ranking. We believe this is due to the range of new photometric effects in our GT which are absent in *Middlebury*. MDP achieves top performance in *Middlebury* but ranks (in relative terms) 6 in *featureless* and 4.13 in Avg.EE by average. This is because large textureless regions in *featureless* provide less SIFT features, in turn weakening its inner motion detail preservation. Additionally, LME ranks higher (in relative terms) than in *Middllebury*. The reason may be due to the local smoothness and deformation penalties [23], which is robust to complex motion (Avg.EE 0.12 in *blur*) and texureless regions (Avg.EE 0.09 in *featureless*).

An evaluation on the RGB channel of the long sequences is also performed as shown in Fig. 7. Similar to the short sequence case, LME provides the best Avg.EE in all trials while Classic+NL, ITV-L1 and MDP yield equally top performance in *stretch*. All the methods display comparatively larger Avg.EE in *mBlur* due to the camera blur and fast motion in the scene. In the robustness test (SD), ITV-L1 reaches the top performance on both *crush* and *stretch* while LME yields the best results on the other sequences. Our graph view in Fig. 7(b) shows that both LME and ITV-L1 give lower accumulated error (Acc.EE) against the other baselines along the entire *crush* sequence.

To evaluate our hybrid RGB-NIR optical flow algorithm – and the potential benefit of using our weighting scheme and multiple spectrums for dense tracking – we compare our method which includes the proposed *Detail-Aware Weight* (*vnflow.DA*) against MCOF, COF as well as three other implementations using fixed weights (0, 0.5 and 1) in Fig. 8. Note that our implementation of COF is applied using all R, G, B and NIR channels. It is observed that *vnflow.DA* outperforms all other baseline methods in Avg.EE in all cases. Our algorithm **without** NIR energy ($\lambda = 0$) shows low overall performance (Avg.EE rank 7.75) while **with only** NIR energy ($\lambda = 1$) it ranks 4.38 in Avg.EE. In addition, LME with NIR imagery achieves comparably lower overall Avg.EE but shows large A100 error in *str.shadow* due to the large shadow that affects the inner detail preservation process. MCOF takes the advantage from additional NIR channel and gives precision generally closed to our methods using fixed weights. In some difficult cases e.g. *mObjs* etc, the precision of MCOF is affected by the primitive optimization scheme. Furthermore, COF yields competitive performance in overall (Avg.EE rank 4.5) and shows the improvement over the case used RGB channel only (Avg.EE rank 7.75).

We then perform an Avg.EE comparison of LME, MDP and four *vnflow* implementations on *str.shadow* by varying the feature distribution in the NIR channel. As shown in Fig. 9, we are ramping up the exposure to reduce the overall number of NIR features in the image. As expected, less NIR information (higher exposure) generally increases the Avg.EE. However, even with a very low quantity of NIR information (+2.0), *vnflow.DA* still shows improvement over other implementations using fixed weights (0, 0.5 and 1).

Fig. 10, a compelling illustration, visualises how switching between RGB and NIR information can contribute to the strong performance of *vnflow.DA*. Our *vnflow.DA* uses texture details invisible in the RGB channel where required (and vice-versa). This provides an explanation to why the algorithm gives higher accuracy against other methods which are using either the RGB or NIR channels alone. However, it should be noted that our evaluation here is a relative one – providing the first insight into how optical flow (and other tracking) can potentially benefit from multiple spectrums. An absolute RGB-NIR evaluation would require a *third* hidden spectrum – in the same way that to evaluate RGB algorithms in our new dataset and evaluation framework we have required NIR for GT (i.e. a second spectrum). Such an evaluation of dedicated RGB/NIR (or other multispectral) methods may not be practical until multispectral tracking, hardware and other suitable dyes become more widespread.

V. Conclusion

In this paper, we present a new publicly available ground truth dataset for evaluating RGB/Color based optical flow algorithms. By leveraging RGB-NIR imaging and NIR visible dyes, our dataset provides dense ground truth for real-world objects in short and long sequences, as well as with nonrigid motion, illumination changes and motion blur. Algorithms are executed on the RGB sequences, and their result is compared to the ground truth obtained by analysing the dense patters only visible in the NIR channel. We also propose a multispectral optical flow framework which utilizes an adoptive weighting scheme to balance the contributions of different channels in order to optimize overall performance. This provides a compelling insight into the potential benefits for tracking in multiple spectra. One further challenge is finding a dye solution which remains invisible in the RGB channel for any object surface. This way, ground truth deformations could be obtained from a wider range of material.

References

[1] W. Li, F. Viola, J. Starck, G. J. Brostow, and N. D. Campbell, “Roto++: Accelerating professional rotoscoping using shape manifolds,” *ACM SIGGRAPH'16*, vol. 35, no. 4, 2016.

[2] R. Tang, D. Cosker, and W. Li, “Global alignment for dynamic 3d morphable model construction,” in *Workshop on Vision and Language (V&LW'12)*, 2012.

[3] W. Li, D. Cosker, and M. Brown, “An anchor patch based optimisation framework for reducing optical flow drift in long image sequences,” in *Proc. of ACCV*, November 2012, pp. 112–125.

[4] W. Li, D. Cosker, and M. Brown, “Drift robust non-rigid optical flow enhancement for long sequences,” *Journal of Intelligent and Fuzzy Systems*, vol. 0, no. 0, p. 12, 2016.

[5] C. Godard, P. Hedman, W. Li, and G. J. Brostow, “Multi-view reconstruction of highly specular surfaces in uncontrolled environments,” in *3DV*, 2015, pp. 19–27.

[6] G. Ren, W. Li, and E. O'Neill, “Towards the design of effective freehand gestural interaction for interactive tv,” *Journal of Intelligent and Fuzzy Systems*, vol. 0, no. 0, p. 12, 2016.

[7] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski, “A database and evaluation methodology for optical flow,” *IJCV*, vol. 92, pp. 1–31, 2011.

[8] W. Li, Y. Chen, J. Lee, G. Ren, and D. Cosker, “Robust optical flow estimation for continuous blurred scenes using rgb-motion imaging and directional filtering,” in *Proc. of WACV*, 2014, pp. 792–799.

[9] W. Li, “Nonrigid surface tracking, analysis and evaluation,” Ph.D. dissertation, University of Bath, 2013.

[10] W. Li and D. Cosker, “Video interpolation using optical flow and laplacian smoothness,” *Neurocomputing*, vol. 0, p. 0, 2016.

[11] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, “A naturalistic open source movie for optical flow evaluation,” in *Proc. of ECCV*, 2012, pp. 611–625.

[12] R. Garg, A. Roussos, and L. Agapito, “Robust trajectory-space tv-l1 optical flow for non-rigid sequences,” *In Proc. of EMMCVPR*, pp. 300–314, 2011.

[13] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Proc. of CVPR*, 2012, pp. 3354–3361.

[14] T. Brox and J. Malik, “Large displacement optical flow: Descriptor matching in variational motion estimation,” *IEEE Trans. on PAMI*, vol. 33, pp. 500–513, 2011.

[15] M. Brown and S. Susstrunk, “Multi-spectral sift for scene category recognition,” *In Proc. of CVPR*, vol. 0, pp. 177–184, 2011.

[16] L. Schaul, C. Fredembach, and S. Susstrunk, “Color image dehazing using the near-infrared,” in *Proc. of ICIP*, 2009, pp. 1629–1632.

[17] D. Firmenichy, M. Brown, and S. Susstrunk, “Multispectral interest points for rgb-nir image registration,” in *Proc. of ICIP*. IEEE, 2011, pp. 181–184.

[18] W. Li, Y. Chen, J. Lee, G. Ren, and D. Cosker, “Blur robust optical flow using motion channel,” *Neurocomputing*, vol. 0, no. 0, p. 12, 2016.

[19] B. Horn and B. Schunck, “Determining optical flow,” *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[20] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proc. of IJCAI*, 1981, pp. 674–679.

[21] M. Black and P. Anandan, “The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields,” *CVIU*, vol. 63, no. 1, pp. 75–104, 1996.

[22] L. Xu, J. Jia, and Y. Matsushita, “Motion detail preserving optical flow estimation,” *IEEE Trans. on PAMI*, vol. 34, no. 9, pp. 1744–1757, 2012.

[23] W. Li, D. Cosker, M. Brown, and R. Tang, “Optical flow estimation using laplacian mesh energy,” in *Proc. of CVPR*, 2013, pp. 2435–2442.

[24] V. Markandey and B. E. Flinchbaugh, “Multispectral constraints for optical flow computation,” in *International Conference on Computer Vision (ICCV)*, Dec 1990, pp. 38–41.

[25] J. Barron and R. Klette, “Quantitative color optical flow,” in *Proc. of ICPR*, vol. 4, 2002, pp. 251–255.

[26] X. Cao, X. Tong, Q. Dai, and S. Lin, “High resolution multispectral video capture with a hybrid camera system,” in *Proc. of CVPR*, 2011, pp. 297–304.

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

[28] M. Drulea and S. Nedevschi, “Total variation regularization of local-global optical flow,” in *Proc. of ITSC*. IEEE, 2011, pp. 318–323.

[29] D. Sun, S. Roth, and M. Black, “Secrets of optical flow estimation and their principles,” in *Proc. of CVPR*, 2010, pp. 2432 –2439.

[30] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, “An improved algorithm for tv-l1 optical flow,” *Proc. of DVMA Workshop*, pp. 23–45, 2009.