CAAI Transactions on Intelligence Technology

Research Article

IPDDF: an improved precision dense descriptor based flow estimation

ISSN 2468-2322
Received on 27th July 2019
Revised on 27th July 2019
Accepted on 27th December 2019
doi: 10.1049/trit.2019.0052
www.ietdl.org

Weiyong Eng ✉, *Voonchet Koo, Tiensze Lim*

Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia
✉ *E-mail: engweiyong@gmail.com*

Abstract: Large displacement optical flow algorithms are generally categorised into descriptor-based matching and pixel-based matching. Descriptor-based approaches are robust to geometric variation, however they have inherent localisation precision limitation due to histogram nature. This work presents a novel method called improved precision dense descriptor flow (IPDDF). The authors introduce an additional pixel-based matching cost within an existing dense Daisy descriptor framework to improve the flow estimation precision. Pixel-based features such as pixel colour and gradient are computed on top of the original descriptor in the authors' matching cost formulation. The pixel-based cost only requires a light-weight pre-computation and can be adapted seamlessly into the matching cost formulation. The framework is built based on the Daisy Filter Flow work. In the framework, Daisy descriptor and a filter-based efficient flow inference technique, as well as a randomised fast patch match search algorithm, are adopted. Given the novel matching cost formulation, the framework enables efficiently solving dense correspondence field estimation in a high-dimensional search space, which includes scale and orientation. Experiments on various challenging image pairs demonstrate the proposed algorithm enhances flow estimation accuracy as well as generate a spatially coherent yet edge-aware flow field result efficiently.

1 Introduction

There are many kinds of literature on the optical flow, in both pixel-based matching and descriptor-based matching. The former is known to deal with the small displacement flow and the latter can deal with large displacement flow of greater geometric variances. Previously, most methods have focused on the smaller displacement optical flow and rigid motion. Most of the small displacement flow estimations are based on the variational method [1], in which a local gradient-based matching of pixel grey values data term is formulated together with a global smoothness assumption. Later, subsequent research extends the work to overcome the limitation of occlusions by adopting the non-quadratic penalisers in both the data term and smoothness term accordingly [2, 3]. Also, the violation of constant brightness assumption in the data term leads to gradient constraint [4], which is a photometric invariant constraint. Furthermore, a local window is adopted in the data term for a point-to-point pixel grey value patch matching. Then, an adaptive local patch is employed to deal with the motion discontinuity. Later, a number of fast edge-preserving filters [5, 6] are designed in order to perform the patch matching in constant time irrespective of window size. Besides, PatchMatch [7, 8] has explored complexity reduction in the search range dimension. Afterwards, PatchMatch filter [9] has extended the works for complexity reduction in both window size and search range dimension by adapting the superpixel [10] as a bridge to link both methods.

Descriptor-based matching is often employed in the image registration task since the advent of the feature detectors [11–13] and descriptors [14–16]. Feature detectors are first adopted to identify some distinctive keypoint locations such as corners and edges in an image. Later, descriptors are handcrafted to describe a local region surrounding that keypoint. Various works [14–16] are explored in the descriptor design in order to deal with geometric distortion such as scale and orientation variances. Histogram of the oriented gradient is first employed in SIFT descriptor [17] and due to its robustness, it has gradually become the basic building block of subsequently advanced descriptors. Later, Daisy descriptor [18] is proposed to compute descriptor densely for every pixel in an image. Unlike previous methods of sparsely detect and describe a few distinctive keypoints, Daisy descriptor can construct descriptor densely thanks to its reuse of the histogram across the pixel. Daisy Filter Flow [19] further adopted the Daisy descriptor in the matching cost computation and adapted PatchMatch Filter [9] in the flow inference.

The main advantage of large displacement optical flow is that it requires sparser image sampling in the time domain than small displacement flow. However, large displacement optical flow is a challenging task as it may contain multi-layered motion and it exhibits significant geometric variances. Large displacement optical flow may involve multiple independent motions due to the motion of individual objects in the scene and the motion of the camera. Specifically, the Moseg dataset [20] that is experimented here contains multi-layered motion and thus, the optical flow cannot be modelled as a single global motion. The multi-layered motion requires flow estimation to be computed densely or at least for some pixels in all different motion layers for further interpolations. Also, large displacement optical flow often contains significant scale and orientation variation as objects in the scene may go through the larger motion.

Previous state-of-the-art methods which work well on small-displacement rigid motion optical flow have difficulties in working on the new challenges of significant geometric transformation. Variational approach implements a coarse-to-fine framework to deal with large-displacement optical flow [21]. Typical coarse-to-fine warping method can solve large displacement optical flow up to a certain extent that the motion is no larger than the scale of the structure as pointed out in [22]. Another technique [22] investigates into this large motion problem by integrating rich descriptors into the variational optical flow setting.

Even though the descriptor-based approach can deal with significant geometric difference, local region descriptor is rarely used in the optical flow estimation. It is due to the top-performance descriptors are mostly spatial histogram-based, which has localisation limitation, and thus affects the estimated flow precision particularly on the motion boundary. Pixel-based methods, on the other hand, does not suffer the localisation limitation as it computes the matching cost on the fine-grain pixel-wise features such as colour and gradient.

One would like to benefit from the typical pixel-based method for their higher matching precision as well as the descriptor-based matching for their ability to deal with significant geometry difference. The contribution of this work is a novel combination of descriptor-based algorithm and pixel-based matching algorithm that can tackle large displacement optical flow with higher accuracy than the previous methods. Fig. 1 highlights the basic idea of concatenating the pixel-based finer grain feature with the descriptor-based higher level histogram representation to improve the precision of a descriptor-based only matching. Precision improvement is clearly shown in Fig. 1 by comparing the straight line on the road and pavement on the bottom right of both images.

This work investigates the improvement on the *precision* of the large displacement optical flow estimation. Pixel-based matching algorithms offer a high precision flow estimation. Descriptor-based matching methods provide high recall flow estimation as they are more robust to the geometric variances.

- A descriptor-based method adopts a patch-based histogram of oriented gradient features in the matching cost formulation. This kind of carefully hand-engineered descriptor such as SIFT [17] is known for its robustness to geometric variances such as scale and orientation distortion. However, this patch-based descriptor is built upon histogram of oriented gradient and thus has inherent localisation limitation due to the histogram nature.
- On the other hand, a pixel-based method adopts pixel-wise finer-grain features such as colour and gradient in the matching score computation. The pixel-based method does not suffer from the localisation limitation as it does not employ a histogram computation. Consequently, a pixel-based approach provides high precision flow estimation. Thus, following the terminology from [23], a descriptor-based matching has high recall (robustness) flow estimation whereas a pixel-based matching provides high precision (spatial discrimination).

In order to retain the descriptor robustness and improve its precision, this work integrates the matching score from both the *pixel-based matching* and *descriptor-based matching*. A novel formulation is proposed to minimise the cost incurred by both pixel-based and descriptor-based matching.

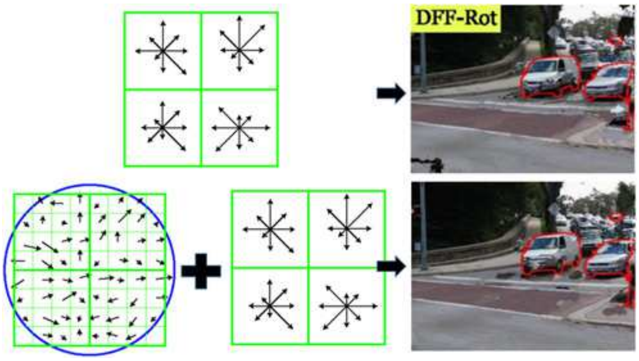


Fig. 1 *Top row shows the idea of histogram-based descriptor matching is used in DFF [19] and it is robust to geometric variation. The bottom row illustrates the idea of concatenating the pixel-based and descriptor-based matching and its result by our method, IPDDF. The finer grain details preserved in the pixel-wise feature is integrated with histogram-based descriptor representation to improve the precision. The precision improvement is clearly shown by comparing the straight line on the road and the curve pavement on the bottom right by zooming into the images*

2 Related works

2.1 *Pixel-based matching: high precision*

Most optical flow estimation algorithms are pixel-based matching, either local method such as Lucas and Kanade [24] or global approach incorporating the regularisation constraint such as smoothness term as in work by Horn and Schunck [1]. Pixel-based matching usually enforces *brightness* and *gradient constancy* between two corresponding image patches. Pixel-based matching is accurate in translational motion and some can even achieve subpixel accuracy. However, pixel-based matching has difficulty to deal with non-translational motion, especially the challenging geometric distortion such as scale and orientation difference.

2.2 *Descriptor-based matching: high recall*

The introduction of the keypoint detector and descriptor [15, 17, 18] has led to sparse descriptor matching, followed by adaptive sparse-to-dense interpolation for optical flow estimation. Patch-based descriptors such as SIFT [17] and SURF [15] are robust to photometric and geometric deformations. However, computational complexity becomes the main design criterion when it comes to dense descriptor-based matching. Recently, researchers have shown that it is possible to compute descriptor on every pixel of the images (e.g. Daisy [18], dense Scale-Invariant Descriptors [25]), thanks to the constructional scheme of reusing shared histogram across pixel [18]. This tackles the limitation of the previous descriptor-based matching which can only be computed sparsely [17].

Among the dense descriptor, Daisy has shown it outperforms SIFT and it runs about 60 times faster in wide-baseline stereo matching. However, Daisy [18] is designed to tackle only with rigid camera motions and subject to the assumption that the two given images are calibrated. On the other hand, Daisy Filter Flow (DFF) [19] generalises the Daisy descriptors and they can deal with scale and rotation changes beyond just translations. DFF [19] computes the descriptor-based correspondence field estimation efficiently in a high-dimensional space which is augmented by scale and orientation. They [19] built this efficient algorithm upon a few well-known methods which are Daisy descriptor, PatchMatch fast search [7, 8] as well as filter-based efficient flow field inference [26, 27].

2.2.1 Histogram-based descriptor localisation limitation: A building block of a typical histogram-based keypoint descriptor is illustrated on the bottom row left side of Fig. 1. Bottom left of Fig. 1 shows a local patch of pixel gradients of black arrows, where its length denotes the gradient magnitude and its direction indicate the gradient orientation. The bottom middle of Fig. 1 shows a histogram of oriented gradients which quantised into eight directions and the histogram is built by accumulating the pixel gradient within that local region. A pixel gradient on the left can shift up to four sample positions while still contributing to the same histogram in the middle, therefore achieving the objective of permitting greater local positional shift. This type of matching gradients while allowing for shifts in their positions permits matching of 3D objects from a range of viewpoint. However, allowing the shift in the position of the gradients reduce the matching location precision especially when the movement is due to translational motion, rather than rotational motion.

3 Improved precision dense descriptor flow (IPDDF)

In this paper, we propose a novel optical flow estimation algorithm: IPDDF, that integrates the benefits of both pixel-based and descriptor-based approaches. Our proposed method can handle complicated non-rigid motion such as scale, orientation and view-point difference thanks to the robustness of daisy descriptor [18]. Also, our proposed method can tackle small object with large motion [21] as well as motion details preserving [28], as our framework performs the flow estimation on a full resolution grid of the image. A typical variational approach with coarse-to-fine framework struggle with those difficulties, as the motion details and the small scale structures, are lost or being smoothed away during the initial process of downsampling to coarse-scale image [21].

Finally, our approach is suitable for large displacement optical flow, which is a limitation for the differential approach with linearisation assumption and variational methods with a coarse-to-fine framework [24]. Our method does not rely on linearisation assumption which applies only to small-displacement. We do not use a coarse-to-fine framework which can only estimate the flow of the object if its structure is no larger than its displacement, as the structure details are smoothed away just at the level when the displacement is small enough to be estimated in a variational setting [21]. Also, as an alternative to regularisation constraint as in a global method to disambiguate weakly descriptive pixel-based data term, we adopt a richly descriptive feature, specifically Daisy feature [18].

3.1 Proposed formulation: matching cost formulation

In this work, the dense daisy-descriptor framework from DFF [19] is chosen as the baseline implementation. In their work, they focused on matching images across different scenes. Here, we focused especially on the large displacement optical flow problem. Given a pair of images $\boldsymbol{I}$ and $\boldsymbol{I}'$, it is to estimate the labelling of the flow field $\boldsymbol{F} = \{f(p) = (u(p), v(p))\}$ for every pixel $p = \{x_p, y_p\} \in \boldsymbol{I}$. In a dense descriptor flow framework, it is to search for the label field $\boldsymbol{L} = \{l(p) = (u(p), v(p), s(p), \theta(p))\}$ within the label range, that minimises the Markov random field formulation. The additional labels, scale $s(p)$ and angle $\theta(p)$ for each pixel p, are accounted for by the robustness of Daisy descriptor.

Fig. 2 gives an overview of the adopted framework. It is comprised of two stages: (i) pre-computation of the image gradient and Daisy descriptor, (ii) online matching of the computed cost. In the descriptor pre-computation phase, a standard upright Daisy is pre-compute for $\boldsymbol{I}'$ while a set of *convolved orientation maps* $\boldsymbol{G}'^{R}_{s,\theta}$ [18] is pre-compute for $\boldsymbol{I}$, where $s \in E$ and $\theta \in \Theta$. In the online matching stage, the convolved orientation maps can then be adopted to generate a rotated and scaled Daisy descriptor $\boldsymbol{D}'^{R}_{l}$ on demand based on a hypothetical candidate label l. The hypothetical label is generated based on the PatchMatchFilter [9, 19] concept, which is by iterative neighbour propagation and random re-sampling. As in [19], the framework is performed on a superpixel level in order to integrate the PatchMatch search algorithm [8] with the efficient filter-based inference [5]. The raw matching cost is computed for each hypothetical label using image colour, gradient as well as the descriptor distances. The raw matching cost is then adaptively aggregated for an edge-preserving flow estimation by using the linear time filter-based inference [5]. The algorithm is repeated for a fixed iteration. An optimal label that gives the minimum matching cost is chosen among the hypothetical label.

3.1.1 Raw matching cost from the dense descriptor:

The matching cost is formulated as in (1). It is to compute the corresponding points similarity based on their Daisy descriptor distance. The smaller the distance, the smaller the matching cost, and thus the more likely a hypothetical label is the match. The first term in (1) is the Daisy descriptor distance between (i) the standard upright unscaled descriptor $\boldsymbol{D}^{R}_{b}(p)$ in the pixel p of image $\boldsymbol{I}$ and, (ii) descriptor $\boldsymbol{D}'^{R}_{s,\theta}(p')$ that is scaled by s and oriented at an angle θ in the pixel p' that is translated by $(u(p), v(p))$ image $\boldsymbol{I}'$, which is exactly the same as in [19]

$$\begin{aligned}\boldsymbol{C}_l(p) = {} & \min(\|\boldsymbol{D}^{R}_{b}(p) - \boldsymbol{D}'^{R}_{s,\theta}(p')\|, t) \\ & + \alpha \cdot \min(\|\boldsymbol{I}(p) - \boldsymbol{I}(p')\|, t_1) \\ & + \beta \cdot \min(\|\nabla \boldsymbol{I}(p) - \nabla \boldsymbol{I}'(p')\|, t_2)\end{aligned} \tag{1}$$

3.1.2 Raw matching cost from non-histogram pixel-based feature:

The second and last terms in (1) are the newly introduced matching cost into the original descriptor distance by taking into account the brightness and gradient constancy, respectively, for a better localisation precision to the original histogram-based descriptor. The smaller the brightness and gradient difference, the lower the matching cost, and thus the more likely the hypothetical label is the match. Both α and β denote the weighting of colour and gradient constancy costs, respectively. The thresholds t, t_1 and t_2 are adopted to account for the outlier, which is either due to occlusion, or the object that appears only in one view. Truncated L_1 distances are used for all the three feature terms: descriptor, colour and gradient terms for its robustness as in (1).

3.1.3 Filtered matching cost:

As in DFF [19], a filtering-based approach is employed to solve this typical multi-labelling task in computer vision. This results in edge-preserving filtering and adaptively smoothing the raw matching cost as in (1) yet provides an efficient label inference. In contrast to DFF [19], we employed the edge-preserving filter not only to the descriptor-based raw label cost, but also to the pixel-wise colour and/or gradient raw matching label cost. The raw matching cost $\boldsymbol{C}_l(p)$ evaluated for the pixel p and label l as in (1) is then adaptively aggregated across all the pixels q within a neighbourhood window $W^{r}(p)$ of the radius r as in (2). The weighting cost $\lambda_{q,p}(\boldsymbol{I})$ denotes the appearance similarity between the pixel p and its neighbour q in the image $\boldsymbol{I}$, in both intensity and spatial dimension closeness. This is a

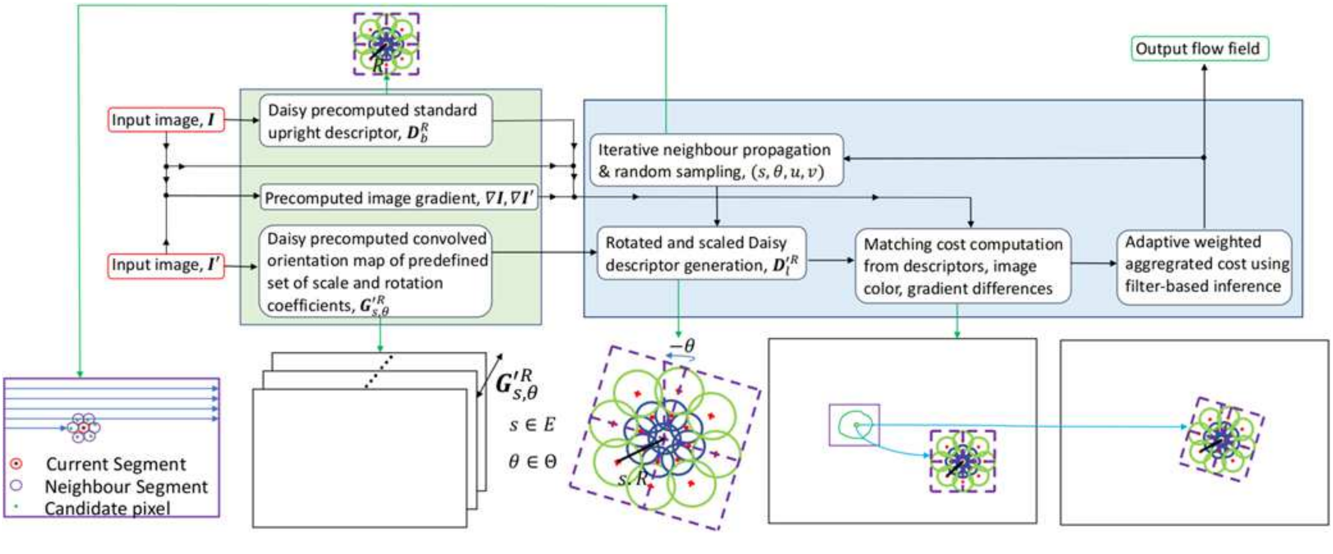
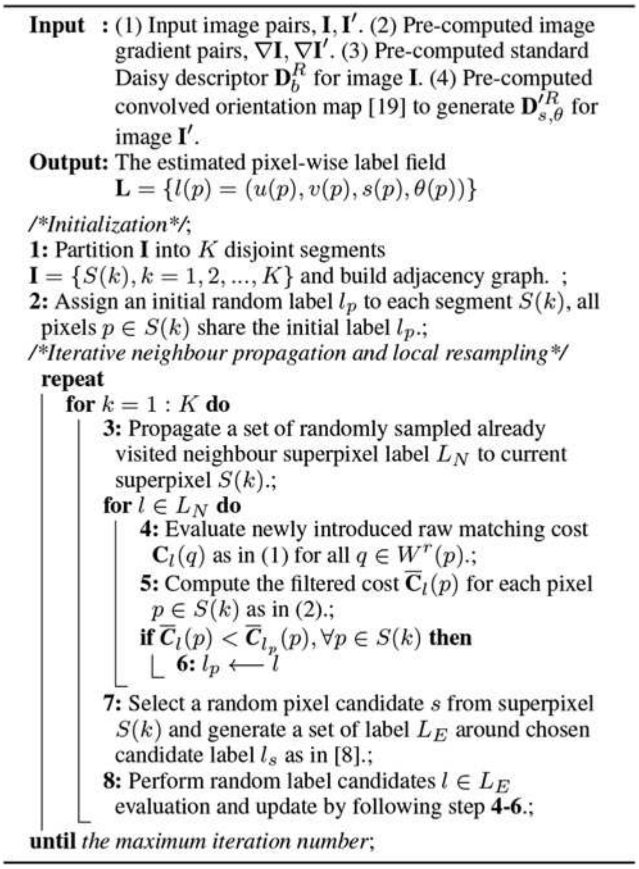

Input image, $\boldsymbol{I}$
Daisy precomputed standard upright descriptor, $\boldsymbol{D}^{R}_{b}$
Precomputed image gradient, $\nabla\boldsymbol{I}, \nabla\boldsymbol{I}'$
Daisy precomputed convolved orientation map of predefined set of scale and rotation coefficients, $\boldsymbol{G}'^{R}_{s,\theta}$
Input image, $\boldsymbol{I}'$
Iterative neighbour propagation & random sampling, (s, θ, u, v)
Rotated and scaled Daisy descriptor generation, $\boldsymbol{D}'^{R}_{l}$
Matching cost computation from descriptors, image color, gradient differences
Adaptive weighted aggregrated cost using filter-based inference
Output flow field
R
Current Segment
Neighbour Segment
Candidate pixel
$\boldsymbol{G}'^{R}_{s,\theta}$
$s \in E$
$\theta \in \Theta$
$-\theta$
s.R

Fig. 2 *Figure shows the framework of our proposed method IPDDF. It consists of two stages: pixel gradient and Daisy pre-computation as well as online matching*

non-linear local varying weight and a brute force computation scale linearly with the local window size. A number of fast edge-aware filters [5, 6, 29] reformulate this bilateral weight so it can be computed in constant time. The authors of [5, 6, 29] can be used to compute the support weight $\lambda_{q,p}(\boldsymbol{I})$ of a neighbourhood pixel q towards p adaptively.



```
Input  : (1) Input image pairs, $\mathbf{I}, \mathbf{I}'$. (2) Pre-computed image
         gradient pairs, $\nabla\mathbf{I}, \nabla\mathbf{I}'$. (3) Pre-computed standard
         Daisy descriptor $\mathbf{D}_b^R$ for image $\mathbf{I}$. (4) Pre-computed
         convolved orientation map [19] to generate $\mathbf{D}'^R_{s,\theta}$ for
         image $\mathbf{I}'$.
Output: The estimated pixel-wise label field
         $\mathbf{L} = \{l(p) = (u(p), v(p), s(p), \theta(p))\}$
/*Initialization*/;
1: Partition $\mathbf{I}$ into $K$ disjoint segments
$\mathbf{I} = \{S(k), k = 1, 2, ..., K\}$ and build adjacency graph. ;
2: Assign an initial random label $l_p$ to each segment $S(k)$, all
  pixels $p \in S(k)$ share the initial label $l_p$.;
/*Iterative neighbour propagation and local resampling*/
repeat
    for $k = 1 : K$ do
        3: Propagate a set of randomly sampled already
          visited neighbour superpixel label $L_N$ to current
          superpixel $S(k)$.;
        for $l \in L_N$ do
            4: Evaluate newly introduced raw matching cost
              $\mathbf{C}_l(q)$ as in (1) for all $q \in W^r(p)$.;
            5: Compute the filtered cost $\overline{\mathbf{C}}_l(p)$ for each pixel
              $p \in S(k)$ as in (2).;
            if $\overline{\mathbf{C}}_l(p) < \overline{\mathbf{C}}_{l_p}(p), \forall p \in S(k)$ then
                6: $l_p \longleftarrow l$
        7: Select a random pixel candidate $s$ from superpixel
          $S(k)$ and generate a set of label $L_E$ around chosen
          candidate label $l_s$ as in [8].;
        8: Perform random label candidates $l \in L_E$
          evaluation and update by following step 4-6.;
until the maximum iteration number;
```

Fig. 3 *Algorithm 1: Improved precision dense descriptor flow*

They usually use the input image $\boldsymbol{I}$ to guide the image filtering process and provide a spatially smooth yet edge-preserving filtered output. Here, we adopt the linear-time guided filter method [5] for its complexity trade-off and filtering quality

$$\bar{\boldsymbol{C}}_l(p) = \sum_{q\in W^r(p)} \lambda_{q,p}(\boldsymbol{I})\boldsymbol{C}_l(q) \qquad (2)$$

The search for the optimal label is simply Winner-Takes-All (WTA), which minimise the aggregated matching cost as follows:

$$l_p = argmin_{l\in L}\bar{\boldsymbol{C}}_l(p) \qquad (3)$$

Exhaustively evaluating the raw and aggregated cost $\boldsymbol{C}_l(p)$ and $\bar{\boldsymbol{C}}_l(p)$ as in (1) and (2) for every single label $l \in L$ is still very time-consuming, even though the filter-based inference gives favourable efficiency. It is due to the complexity is proportional to the high-dimensional label space size, $|L| = L^u \cdot L^v \cdot |E| \cdot H$, where $(L^u, L^v, |E|, H)$ indicate the discrete search spaces for every domain of (u, v, s, θ), respectively.

The generalised PatchMatch algorithm [8] is designed to perform nearest neighbour search over translation, rotation and scale domain with significantly reduced complexity in the search range of $O(\log|L|)$. We follow the main idea of DFF [19] to find the optimal label l_p in (3) by combining filter-based inference [6] with the fast randomised PatchMatch search [8] in the high-dimensional label space. Here, only plausible candidate labels l through *propagation* and *resampling* are evaluated, alleviating from going through every label $l \in L$. Segments or superpixels [10] are adopted as a bridge to link the filter-based inference [6] with the fast randomised PatchMatch [8] algorithms as in DFF [19]. Our algorithm is described in Algorithm 1 (see Fig. 3).

4 Experiment results

The IPDDF algorithm is implemented based on the publicly available DFF code. The parameters set for the dense Daisy

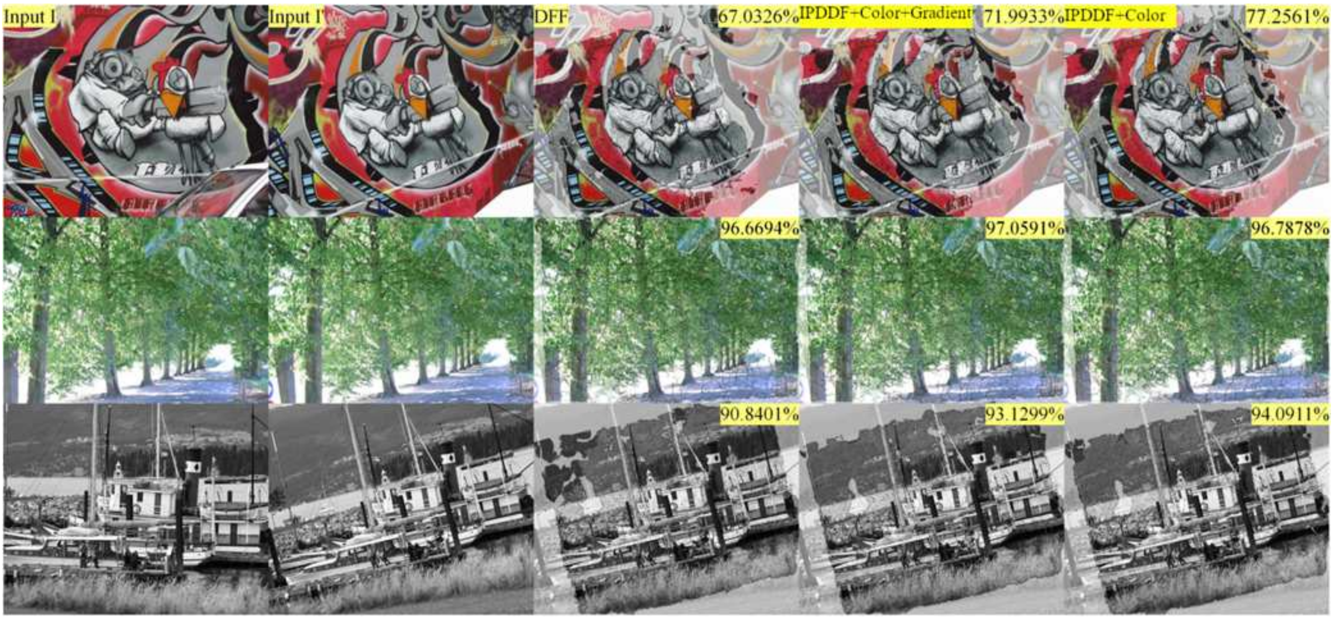


Fig. 4 *Comparison of DFF, IPDDF with colour and gradient feature, and IPDDF with colour only feature, from left to right of the last three columns. The input image pairs that show a large difference in sharpness, planar scale and rotation, and viewpoint are shown in the first two columns. The first image is warped to the second image using the estimated flow field. The warped image is then overlaid directly onto the second image as shown in the last three columns. Only the good matches are highlighted in a darker colour and the accuracy percentages are recorded on the top right corner for all warped images. Our method IPDDF show accuracy percentage improvement over DFF for all three image pairs*

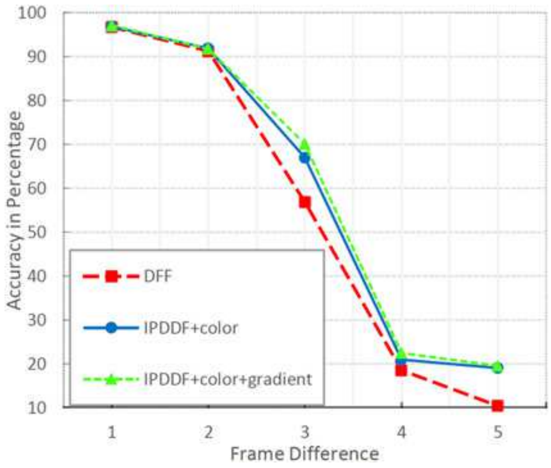


Fig. 5 *Quantitative result with correct matches on the trees sequence of the dataset of Mikolajczyk et al. The accuracy percentages are reported for different image pairs of n or more frames away from the source image **I**. The results are plotted for DFF against the IPDDF with colour only feature, as well as IPDDF with colour and gradient features. Our proposed method IPDDF has higher accuracy than DFF [19] in frame difference of 3 and 5*

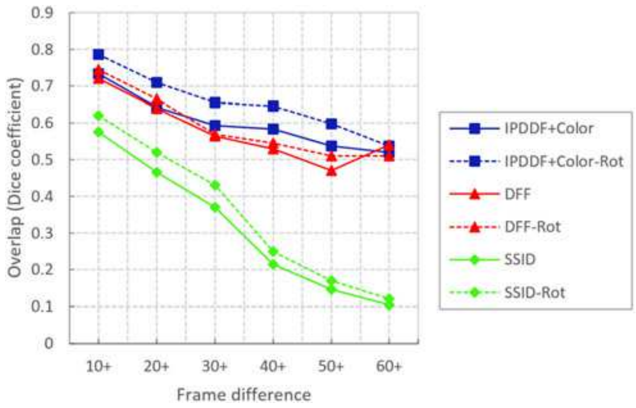


Fig. 6 *Quantitative overlap results on the Moseg dataset. Average results are shown for various '$n+$' cases, namely, using all test pairs of* n *or more frames away from the source image **I**. IPDDF, DFF and SSID, and their rotation-sensitive version with the postfix '-Rot', are reported for comparison. Our method IPDDF + Colour shows clear improvement over the rest of the methods*

feature are set as $R = 16$, $s \in E = \{0.5, 1.0, 1.5, 2.0, 2.5\}$ and $H = 7$, where $\theta \in \Theta = \{15{*}o, o \in [-3, 3]\}$, similar as in DFF [19]. The search range is set to the image size both horizontally and vertically to capture big location changes for large-scaled and oriented motion. The SLIC segment K [10] grows sublinearly with the image size, namely $K = 800$ for 800×480 image. The truncation t in (1) is set to 8.84 empirically. Also, the other two truncation for colour and gradient features, t_1 and t_2 are set empirically to 15 and 4, respectively. The weighting scheme of the colour and gradient cost, α and β in (1) are set empirically to 0.100 and 0.035, respectively. Gradient cost often has a higher weight than the colour cost in the matching cost formulation as it is more robust to photometric distortion. Also, the descriptor cost in (1) contributes the most among the three cost terms, specifically its weight to gradient weight is 10:1. It is due to the new pixel-based cost serves to improve the localisation precision and the descriptor cost is still dominating the matching cost. Thus, both gradient and colour cost weight are just a fraction of the descriptor cost. The number of iterations is set to 20 empirically. The average runtime is 40 s for a 320×480 image. It is close to the DFF (20–38 s for 12–25 iterations) as the proposed introduction of pixel colour and gradient in computing the matching cost only amount to a few additional operations for each matching cost calculation. Finally, the filtering window radius r is set to 9. Note that the $\lambda_{q,p}$ in (2) is the adaptive weight, which is computed automatically by the linear-time edge-aware filter, specifically guided filter [5] here.

Results on the dataset of Mikolajczyk et al. [30]*:* In this dataset, we focus on the image sequences of *graf*, *trees* and *boat* for their challenging geometric changes of scale and orientation variations. We compared the original DFF [19] and our IPDDF algorithm with the colour feature, as well as IPDDF with both colour and gradient features. We apply the similar evaluation technique in [19, 31, 32], namely the estimated match that is located within 15 pixels from the ground truth location in the target image I' as right matches.

Fig. 4 illustrates the typical performance on three types of test cases: viewpoint changes, sharpness difference, as well as planar and rotation variations. Our algorithm shows its robust and favourable performance on this dataset of various challenging settings. It outperforms DFF [19] on the three image sequences of *graf*, *trees* and *boat*. IPDDF with only colour feature performs best in *graf* and *boat* for the challenging geometric changes setting. On the other hand, IPDDF with both the colour and gradient features outperforms in the *trees* sequences on the sharpness changes cases.

Fig. 5 shows the quantitative result with correct matches, especially on the *trees* sequences of Mikolajczyk *et al.* [30]. The comparison is performed on DFF [19], IPDDF with the colour term, as well as IPDDF with colour and gradient features. In the *trees* sequence which tests on sharpness differences, both our variation of enhanced features, namely IPDDF with colour only features, and IPDDF with both colour and gradient features, outperform DFF [19]. It is due to the higher spatial resolution of the colour and gradient features are adopted as compared to the histogram-based descriptor only feature.

Results on the Moseg dataset [20]: The proposed method is tested on this dataset that contains 31 challenging outdoor image pairs with large-displacement and multi-layered motion, following the recent segmentation-aware SID (SSID) work [33]. We follow the evaluation protocol in [33], in which we use the estimated flow to warp the segmentation mask from the target image $\boldsymbol{I}'$ to the source image $\boldsymbol{I}$, and compute the overlap with the ground truth using the Dice coefficient. As in [33], the overlap is measured as $2 \times |A \cap B|/(|A| + |B|)$ for two maps A and B.

Fig. 6 illustrates the overlap results acquired by using an SSID descriptor [33] in the SIFT flow [32] framework and DFF algorithm [19]. As in [33], the postfix 'Rot' which specifies the rotation-invariant feature is switched *off* in the methods, since the foreground object does not have many rotations. In this dataset, we use only the colour feature as denoted by postfix '+Colour'. Both our IPDDF + Colour-Rot and IPDDF + Colour methods outperform the state-of-the-art techniques, and they also work better than the closest competitor DFF-Rot, especially for the challenging pairs of large frame displacements. In addition, Fig. 7 illustrates the visual result of the typical leading techniques. Fig. 7 illustrates the visual quality of our method IPDDF is better in finer grain precision than other methods by zooming into the image. It is clearly shown by carefully examining in Fig. 7 that the straight line on the road and curve pavement on the first row. Also, in the left part of the image in the second row, the small details such as post and tree branches are better reconstructed in the warped image by IPDDF. The visual quality of the warped image by IPDDF has verified that the proposed method improves fine-grain precision.

5 Conclusions

We presented IPDDF – a dense Daisy descriptor-based flow estimation with enhanced colour and gradient data term for various image matching framework. The proposed IPDDF is robust in estimating dense correspondences between difficult image pairs in the presence of the drastic changes of photometric and geometric variations especially on the scale and orientation difference,

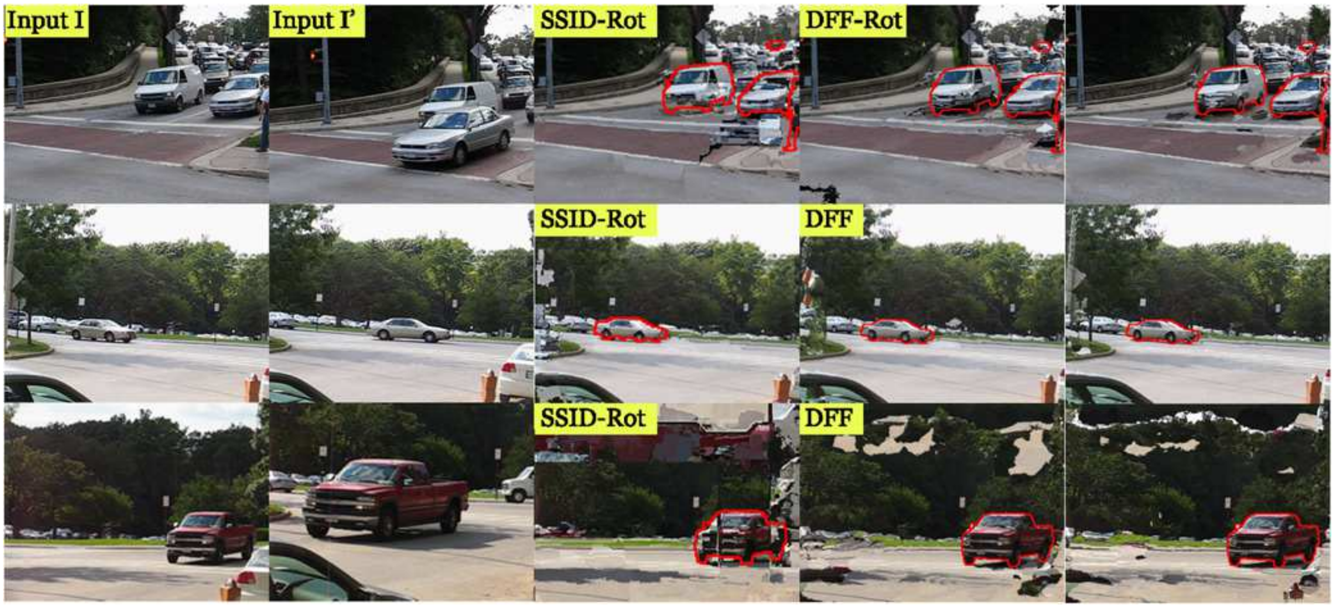


Fig. 7 *Visual comparison of warping the image $\boldsymbol{I}'$ to $\boldsymbol{I}$ on the Moseg pairs of large motion and scale variations. The ground truth segmentation masks of the image $\boldsymbol{I}$ are overlay onto the warped images in red, facilitating the object-mask alignment inspection. Third and fourth columns show the result obtained by SSID-Rot and DFF-Rot. The last column shows our IPDDF + Color-Rot visual result. IPDDF shows clear improvement over the previous method by carefully examining the quality of the warped image. In the first row, the straight line on the road and curve pavement is visually better in IPDDF (last column) by zooming into the image. In the second row, the post and tree branches at the left part of the image are visually better in IPDDF (last column). It verifies that integrating the pixel-wise feature improves in finer grain precision*

brightness changes, and image quality. The addition of enhanced colour and gradient features to the dense-descriptor flow further improve the fine-grain precision of correspondence matching tasks.

6 References

[1] Horn, B.K., Schunck, B.G.: 'Determining optical flow', *Artif. Intell.*, 1981, **17**, (1–3), pp. 185–203

[2] Black, M.J., Anandan, P.: 'The robust estimation of multiple motions: parametric and piecewise-smooth flow fields', *Comput. Vis. Image Underst.*, 1996, **63**, (1), pp. 75–104

[3] Cohen, I.: 'Nonlinear variational method for optical flow computation'. Proc. of the Scandinavian Conf. on Image Analysis, Citeseer, Tromssa, Norway, 1993, vol. 1, pp. 523–523

[4] Brox, T., Bruhn, A., Papenberg, N., *et al.*: 'High accuracy optical flow estimation based on a theory for warping'. European Conf. on Computer Vision, Prague, Czech Republic, 2004, pp. 25–36

[5] He, K., Sun, J., Tang, X.: 'Guided image filtering'. European Conf. on Computer Vision, Crete, Greece, 2010, pp. 1–14

[6] Lu, J., Shi, K., Min, D., *et al.*: 'Cross-based local multipoint filtering'. 2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012, pp. 430–437

[7] Barnes, C., Shechtman, E., Finkelstein, A., *et al.*: 'Patchmatch: a randomized correspondence algorithm for structural image editing', *ACM Trans. Graph.-TOG*, 2009, **28**, (3), p. 24

[8] Barnes, C., Shechtman, E., Goldman, D.B., *et al.*: 'The generalized patchmatch correspondence algorithm'. European Conf. on Computer Vision, Crete, Greece, 2010, pp. 29–43

[9] Lu, J., Yang, H., Min, D., *et al.*: 'Patch match filter: efficient edge-aware filtering meets randomized search for fast correspondence field estimation'. 2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 2013, pp. 1854–1861

[10] Achanta, R., Shaji, A., Smith, K., *et al.*: 'Slic superpixels compared to state-of-the-art superpixel methods', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (11), pp. 2274–2282

[11] Harris, C.G., Stephens, M.: 'A combined corner and edge detector'. Alvey Vision Conf., Citeseer, Manchester, United Kingdom, 1988, vol. 15, pp. 10–5244

[12] Mikolajczyk, K., Schmid, C.: 'An affine invariant interest point detector'. European Conf. on Computer Vision, Copenhagen, Denmark, 2002, pp. 128–142

[13] Shi, J., Tomasi, C.: 'Good features to track' In 1994 Proceedings of IEEE conference on computer vision and pattern recognition, Cornell University, Seattle, WA, USA, 1994, pp. 593–600.

[14] Lowe, D.G.: 'Object recognition from local scale-invariant features'. Int. Conf. on Computer Vision (ICCV), Kerkyra, Greece, 1999, vol. 99, pp. 1150–1157

[15] Bay, H., Tuytelaars, T., Van Gool, L.: 'Surf: speeded up robust features'. European Conf. on Computer Vision, Graz, Austria, 2006, pp. 404–417

[16] Rublee, E., Rabaud, V., Konolige, K., *et al.*: 'Orb: an efficient alternative to sift or surf'. ICCV, Citeseer, vol. 11, Barcelona, Spain, 2011, p. 2

[17] Lowe, D.G.: 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vis.*, 2004, **60**, (2), pp. 91–110

[18] Tola, E., Lepetit, V., Fua, P.: 'Daisy: an efficient dense descriptor applied to wide-baseline stereo', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (5), pp. 815–830

[19] Yang, H., Lin, W.Y., Lu, J.: 'Daisy filter flow: A generalized discrete approach to dense correspondences'. 2014 IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 3406–3413

[20] Brox, T., Malik, J.: 'Object segmentation by long term analysis of point trajectories'. European Conf. on Computer Vision, Crete, Greece, 2010, pp. 282–295

[21] Brox, T., Bregler, C., Malik, J: 'Large displacement optical flow'. IEEE Conf. on Computer Vision and Pattern Recognition, 2009. CVPR 2009, Miami, FL, USA, 2009, pp. 41–48

[22] Brox, T., Malik, J.: 'Large displacement optical flow: descriptor matching in variational motion estimation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (3), pp. 500–513

[23] Fathy, M.E., Tran, Q.H., Zeeshan Zia, M., *et al.*: 'Hierarchical metric learning and matching for 2d and 3d geometric correspondences'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 803–819

[24] Lucas, B.D., Kanade, T.: 'An iterative image registration technique with an application to stereo vision'. IJCAI, 1981, vol. 81, British Columbia, Canada, pp. 674–679

[25] Kokkinos, I., Bronstein, M., Yuille, A.: 'Dense scale invariant descriptors for images and surfaces'. [Research Report] RR-7914, INRIA. 2012 hal-00682775

[26] Rhemann, C., Hosni, A., Bleyer, M., *et al.*: 'Fast cost-volume filtering for visual correspondence and beyond'. 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2011, pp. 3017–3024

[27] Vineet, V., Warrell, J., Torr, P.H.: 'Filter-based mean-field inference for random fields with higher-order terms and product label-spaces', *Int. J. Comput. Vis.*, 2014, **110**, (3), pp. 290–307

[28] Xu, L., Jia, J., Matsushita, Y.: 'Motion detail preserving optical flow estimation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (9), pp. 1744–1757

[29] Paris, S., Kornprobst, P., Tumblin, J., *et al.*: 'Bilateral filtering: theory and applications', *Foundations and Trends in Computer Graphics and Vision*, USA, 2009, **4**, (1), pp. 1–73.

[30] Mikolajczyk, K., Tuytelaars, T., Schmid, C., *et al.*: 'A comparison of affine region detectors', *Int. J. Comput. Vis.*, 2005, **65**, (1–2), pp. 43–72

[31] HaCohen, Y., Shechtman, E., Goldman, D.B., *et al.*: 'Non-rigid dense correspondence with applications for image enhancement', *ACM Trans. Graph. (TOG)*, 2011, **30**, (4), p. 70

[32] Liu, C., Yuen, J., Torralba, A.: 'Sift flow: dense correspondence across scenes and its applications', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (5), pp. 978–994

[33] Trulls, E., Kokkinos, I., Sanfeliu, A., *et al.*: 'Dense segmentation-aware descriptors'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2013, Portland, OR, USA, pp. 2890–2897