Occlusion-Aware Video Registration for Highly Non-Rigid Objects

Bertram Taetz*[1,2], Gabriele Bleser†[1,2], Vladislav Golyanik‡[1], and Didier Stricker§[1,2]

[1]German Research Center for Artificial Intelligence, Kaiserslautern, Germany
[2]Department of Computer Science, Technical University of Kaiserslautern, Kaiserslautern, Germany

Abstract

This paper addresses the problem of video registration for dense non-rigid structure from motion under suboptimal conditions, such as noise, self-occlusions, considerable external occlusions or specularities, i.e. *the computation of optical flow between the reference image and each of the subsequent images in a video sequence when the camera observes a highly deformable object. We tackle this challenging task by improving previously proposed variational optimization techniques for multi-frame optical flow (MFOF) through detection, tracking and handling of uncertain flow field estimates. This is based on a novel Bayesian inference approach incorporated into the MFOF. At the same time, computational costs are significantly reduced through iterative pre-computation of the flow fields. As shown through experiments, the resulting method performs superior to other state-of-the-art (MF)OF methods on video sequences showing a highly non-rigidly deforming object with considerable occlusions.*

1. Introduction

Dense optical flow computation of highly deformable objects is a challenging task that is useful for multiple computer vision applications. Occlusions at the same time are an inherent problem in realistic scenarios that can usually not be avoided. Methods that can handle both aspects have applications ranging from medical imaging and motion compensation over video augmentation, occlusion replacement and video segmentation up to dense non-rigid structure from motion (NRSfM) [9], which is our target application. Given a template or reference image of an object and several input images containing highly non-rigid deformations of the object, the task can be described as finding

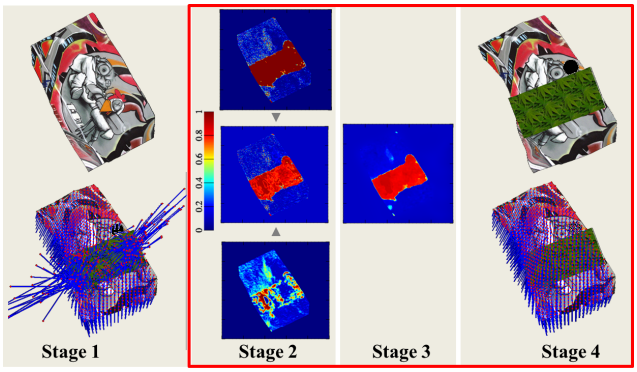


Figure 1: The different stages of our proposed MFOF method with explicit occlusion handling (*cf.* Section 4.1): (1) reference image and result of MFOF without explicit occlusion handling; (2) divergence indicator, counter indicator → initial occlusion probability map; (3) denoised occlusion probability map; (4) occluded input frame and occlusion-aware MFOF result. The blue lines show the pixel correspondences from the reference to the current image.

the optical flow fields (trajectories/warps) [10] that relate all the input images back to the reference image. This is subsequently denoted *multi-frame optical flow (MFOF)*. This complex scenario exhibits a multitude of challenges that are only partially solved in general. Promising approaches for the estimation of optical flow fields in long video sequences consider temporal coherence based on subspace constraints [8, 16, 10]. The underlying idea consists in first deriving a motion basis, either from reliable sparse tracks or via a predefined basis that describes the general motion and deformation of the considered object. Subsequently, the optical flow is described via a linear combination of these basis trajectories. Different approaches directly estimate the coefficients of the linear combination which yields hard constraints to the subspace [8, 16]. An improvement with respect to accuracy and robustness in the presence of noise and highly non-rigid motion was obtained by allowing some deviation from the basis trajectories via the in-

*Bertram.Taetz@dfki.de
†Gabriele.Bleser@dfki.de
‡Vladislav.Golyanik@dfki.de
§Didier.Stricker@dfki.de

clusion as soft constraints [10]. The latter method yields for this scenario currently the most accurate results (to the best of our knowledge), but leaks explicit occlusion handling that is particularly important for large occlusions as shown in Figure 1. Occlusion handling is crucial in long video sequences of real scenarios, since deformable objects typically occlude themselves or are occluded by external objects at some point. Occlusions can influence the accuracy of the overall method considerably, since the optical flow might exhibit random behaviour in occluded and neighbouring regions. This wrong optical flow fields introduces noise for correspondence computations, see Figure 1. Current MFOF approaches only consider occlusion estimation under rather rigid motion with only slightly non-rigid deformations [16, 17] that is not reliable under high non-rigidity [17]. Therefore, we introduce an explicit occlusion detection, tracking and handling that is efficient and gives consistent results in the presence of highly non-rigid deformations. While we focus on handling different types of occlusions, we show that other disturbing effects, such as specularities and noise, are also handled appropriately.

NRSfM methods rely on a decomposition of correspondences to compute (dense) 3D shape and motion, see for instance [9, 13]. As indicated in [13], these methods tend to be sensitive to erroneous correspondences (noise). Stable methods can only compensate for a small amount of noise, typically 1 to 4% of the considered correspondences. However, missing data can be better compensated for, for instance in [13] a 3D reconstruction error of only 5.4% was reported with 32% of missing correspondences. This motivates the usage of our algorithm not only to improve the flow field estimates, but also to provide a method that detects and tracks uncertain correspondences that should be excluded from the 3D shape and motion estimates via NRSfM methods.

2. Related work and contributions

2.1. Optical flow

The variational formulation for the dense optical flow problem dates back to the work of Horn and Schunck [12] in 1980. The problem formulation is composed of a data term that usually accounts for the brightness constancy assumption and a regularization term that allows to fill-in consistent optical flow estimates in low textured areas. With variational methods this problem is formulated as an optimization problem of an energy function in a continuous domain. There have been numerous advances with respect to the method, the data term and the regularization term, especially for *two-frame optical flow (TFOF)* estimation; in [24, 20] an overview of some major approaches can be found. The methods that are most related to our work build on the total variation (TVL1) formulation and use efficient primal-dual methods to solve the optical flow problem with high accuracy [28, 25]. Due to the high parallelizability of these methods, real-time capability was achieved [28]. To consider temporal coherence over many frames with rigid and non-rigid motion, leading to *non-rigid MFOF*, the concept of subspace constraints has proven to be useful. It was introduced in [8], where the flow-fields were reparametrized via a set of basis trajectories. The accuracy and robustness in the presence of highly non-rigid motion, noise and (rather small) occlusions was considerably increased by the recently published method of [10] that includes the subspace constraints as soft constraints. This approach can be interpreted as a generalization of the improved TVL1 method [25] (I-TVL1) for non-rigid objects. The method gives favorable accuracy with respect to other state-of-the-art optical flow methods in the presence of non-rigid motion. However, large occlusions can not be handled as indicated in Figure 1 (left).

2.2. Occlusion handling

There are several *rigid TFOF* methods that address occlusion estimation and handling. Well known approaches use robust norms [4] or forward-backward flow estimation [1]. An efficient and simple occlusion detection via the exploitation of the mapping uniqueness criterion was proposed in [27]. This method is one of the best performing methods on the famous Middlebury dataset [2]. EpicFlow [15] is the currently best performing method on the challenging SINTEL benchmark dataset [6]. It couples occlusion estimation with the optical flow estimation and extrapolates optical flow into occluded regions, which allows to handle even considerable occlusions. Both methods are included in our evaluation. Different methods were proposed for joint motion estimation, segmentation and occlusion modeling [23] as well as temporal consistence and pseudo-depth ordering [21, 19] or local layering, with promissing results even for rather large amount of occlusions in consecutive frames [19]. However, while being a very promising approaches from the modeling point-of-view, these multi-layer approaches are computationally expensive, with a complexity depending on the amount of (local) layers and motion candidates. Both have to be large to accurately compute optical flow in the case of general highly non-rigid deformations as considered in this work. Futhermore, other disturbing effects, like specularities are not considered. *Non-rigid TFOF* methods are *e.g.* [22, 14]. The latter is also robust against large self-occlusions by learning a model for the image distortion and is included in our evaluation. In the case of *MFOF*, contributions to explicit occlusion detection and handling are [16, 17]. Here, [16] presents a variational approach that estimates visibility maps for each pixel, using two reference frames and hard subspace constraints of the pixel displacements. The

same authors improved upon this method in [17] by introducing a different strategy for visibility labeling and even more reference images, but the computational complexity was very high. Furthermore, the authors of [17] explicitly mention that their method can not deal with the highly deformable motion of the waving flag sequence of [10, 8] that we use in our evaluation. In general, the referenced literature contains two general concepts for occlusion modelling and correction. Either occlusions are estimated from a first uncorrected flow estimation and are then used for correcting this afterwards (*e.g.* [1]) or occlusion estimation and correction are done jointly with the optical flow estimation (*e.g.* [23, 26, 3, 19]).

2.3. Our approach and contributions

While great advances have been achieved in the area of accurate and robust dense MFOF for highly non-rigid object motion, as well as in the area of occlusion estimation and handling, mostly for two-frame settings, there are currently no promising approaches for achieving both together under large occlusions. This is, however, crucial when aiming at dense real NRSfM applications. Our main contributions are devoted to overcoming this limitation. First, we develop a novel occlusion estimation framework by fusing different occlusion indicators based on a first flow-field estimation using a Bayesian filter and smoother. This leads to reliable detection and tracking of areas containing uncertain flow information (*e.g.* due to external occlusions and/or self-occlusions) in the presence of highly non-rigid motion. Second, we seamlessly integrate this probabilistic information into the variational framework during a second flow-field estimation. Third, we present an iterative precomputation scheme, which is shown to increase the convergence speed and accuracy of the method. Through experiments we confirm that the proposed approach outperforms state-of-the-art optical flow methods on long video sequences of a single highly deformable object.

3. Proposed method

Our proposed approach consists of several stages. In **Stage 1**, an initial set of optical flow fields is obtained using a MFOF method similar to [10] without occlusion handling that we will call *base scheme*. **Stage 2** consists of computing per frame probabilistic occlusion maps efficiently using Bayesian inference based on the initial flow fields. **Stage 3** consists of a global edge-aware, spatial and temporal variational smoothing. Here, we use an edge-aware version of the global smoothing approach proposed in [16] and based on efficient primal dual algorithms [7]. The exact algorithm can be found in the supplementary material. This stage is not further detailed in this paper. In the final **Stage 4**, the optimized maps are used in our proposed occlusion-aware MFOF method. A convergence speed-up of the base scheme for a pyramid based approach, as in [10], is achieved through iterative pre-computation of the flow fields, as described in Algorithm 1. The underlying idea is to use the rather small motions between consecutive frames to obtain a suitable initialization for the MFOF over long video sequences. The complete process is visualized in Figure 1 and stages 1, 2 and 4 are further detailed below.

3.1. Initial flow field estimation (Stage 1)

It is assumed that the input image sequence has $F-1$ images and a reference image $n_0 = 1$ has been chosen. The image

$$\mathbf{I}(\mathbf{x}, n) : \Omega \times \{1, \dots, F\} \to \mathbb{R}^{N_c} \tag{1}$$

is a vector-valued image with N_c channels, and denotes the n-th image in the sequence, with the image domain $\Omega \subset \mathbb{R}^2$. The point trajectories are represented with the function

$$\mathbf{u}(\mathbf{x}; n) = \begin{bmatrix} u_1(\mathbf{x}, n) \\ u_2(\mathbf{x}, n) \end{bmatrix} : \Omega \times \{2, \dots, F\} \to \mathbb{R}^2. \tag{2}$$

Thus, for every visible point $\mathbf{x} \in \Omega$ in the reference image n_0, the function $\mathbf{u}(\mathbf{x}, \cdot) : \{2, \dots, F\} \to \mathbb{R}^2$ is its discrete-time 2D trajectory over all images of the sequence. At the reference image the trajectories are defined as $\mathbf{u}(\mathbf{x}, n_0) = 0$. Linear subspace constraints that allow for some deviation can be written as

$$\mathbf{u}(\mathbf{x}, n) = \sum_{r=1}^{R} \mathbf{q}_r(n) L_r(\mathbf{x}) + \epsilon(\mathbf{x}, n). \tag{3}$$

This states that each trajectory $\mathbf{u}(\mathbf{x}, \cdot)$ for $\mathbf{x} \in \Omega$ can be represented (up to some error $\epsilon(\mathbf{x}, n) \in \mathbb{R}^2$) by a linear combination of R basis trajectories $\mathbf{q}_1(n), \dots, \mathbf{q}_R(n) : \{2, \dots, F\} \to \mathbb{R}^2$ which are independent from the point location. The trajectories can be predefined or computed from pre-computed tracks, see [10] for more details. In order to estimate the trajectories $\mathbf{u}(\mathbf{x}, n)$, we propose to minimize the following energy

$$E_{\xi(\mathbf{x},n)}[\mathbf{u}(\mathbf{x}, n), \mathbf{L}(\mathbf{x})] = (1 - \xi(\mathbf{x}, n))[\alpha E_{\text{data}}] + \beta E_{\text{link}} + E_{\text{reg}}, \tag{4}$$

$$E_{\text{data}} = \int_\Omega \sum_{n=2}^{F} \Phi(\mathbf{I}(\mathbf{x} + \mathbf{u}(\mathbf{x}, n), n) - \mathbf{I}(\mathbf{x}, n_0))\, d\mathbf{x}, \tag{5}$$

$$E_{\text{link}} = \int_\Omega \sum_{n=2}^{F} |\mathbf{u}(\mathbf{x}, n) - \sum_{i=1}^{R} \mathbf{q}_i(n) L_i(\mathbf{x})|^2\, d\mathbf{x}, \tag{6}$$

$$E_{\text{reg}} = \int_\Omega \sum_{r=1}^{R} g(\mathbf{x}) \Phi(\nabla L_r(\mathbf{x}))\, d\mathbf{x}, \tag{7}$$

with

$$\mathbf{u}(\mathbf{x},n) = \begin{cases} \sum_{i=1}^{R} \mathbf{q}_i(n) L_i(\mathbf{x}) & \text{if} \quad \xi(\mathbf{x},n) = 1 \\ \mathbf{u}(\mathbf{x},n) & \text{if} \quad \xi(\mathbf{x},n) = 0 \end{cases}. \quad (8)$$

Here, $\xi(\mathbf{x},n)$ denotes the given occlusion maps for all frames $n = 2,\ldots,F$ and will be explained in more detail in the following sections. Usually, Φ is chosen to be the $L1$-norm or the Huber-norm. The space varying weight $g(\mathbf{x})$ in the regularization term (7) encourages discontinuities in the flow to coincide with edges of the reference image by reducing the regularization strength near those edges [10, 3]. The energy (4) can be efficiently minimized by alternatingly optimizing with respect to $\mathbf{u}(\mathbf{x},n)$ and $\mathbf{L}(\mathbf{x})$, by applying primal-dual algorithms of [7] as described in the supplementary material. This formulation allows to compute the *initial flow-fields* by setting $\xi(\mathbf{x},n) = 0$ for all frames. In this case the method is reduced to the one proposed in [10]. Furthermore, it allows the computation of the final trajectories, given the computed occlusion maps $\xi(\mathbf{x},n)$ (see next sections). We found that binarized occlusion maps give better results than a soft weighting ($\xi(\mathbf{x},n) \in [0,1]$) in this formulation, since the pre-estimated trajectories $\mathbf{u}(\mathbf{x},n)$ might be very wrong in uncertain regions (*e.g.* see Figure 1, left). This could badly influence the final trajectories by using a soft weighting.

Algorithm 1 Iterative pre-computation for MFOF.

Input: Downsampled image sequence (on highest pyramid level)
Output: Pre-estimate for MFOF (on highest pyramid level)
Set: $\mathbf{u}(\mathbf{x},1) = 0$;
for $n = 2$ to F **do**
  Compute: $\mathbf{u}(\mathbf{x},n)$ by minimizing the energy in Equ. (4) with initial guess $\mathbf{u}(\mathbf{x},n-1)$, by only considering image $\mathbf{I}(1)$ and $\mathbf{I}(n)$
end for

3.2. Occlusion filter (Stage 2)

The major goal of this stage is to obtain a good initial guess for the occlusion maps denoted by $\tilde{\xi}(\mathbf{x},n) \in [0,1]$ for each image of the sequence ($n = 1,\ldots,F$). These initial guesses will be globally smoothed in Stage 3 and binarized to obtain the final occlusion maps $\xi(\mathbf{x},n)$. The latter will then be used in the optimization in Equ. (4), in order to correct the flow-fields. Here, the occlusion map for image n contains the probability of each pixel $\mathbf{x}$ in the reference image being occluded in image n. For efficiently estimating these probabilities, while (1) fusing information from indicators for different types of occlusions and (2) taking temporal and spatial coherence into account, we use discrete Bayes filters and optimal smoothers [18]. More precisely, we use per pixel estimators for tracking the binary state of each pixel in the reference image being occluded or visible in the subsequent images. The local estimates have shown to provide a sufficiently accurate initial guess, while at the same time being efficient and parallelizable (in the spatial domain).

3.2.1 Bayes filter, smoother and state-space model

For each pixel in the reference image we estimate its probability of being occluded in a subsequent image I_n via a Bayes filter, followed by a Bayes smoother, with finite state space $Y_{n,k} \in \{0,1\}$. Here, for a pixel in image n, $Y_{n,1}$ denotes occluded and $Y_{n,0}$ denotes visible. Starting from the prior assumption $P(Y_{1,0}) = 1$ and $P(Y_{1,1}) = 0$ (i.e. the pixel is visible in the reference image), the posterior estimates for $n = 2,\ldots,F$ are recursively calculated from sequences of measurements and control inputs up to time n. These are denoted by $Z_{1:n}$ and $U_{1:n}$ (as further defined in Sections 3.2.2, 3.2.3) respectively. The posterior is thus [18]

$$P_{\mathrm{f}}(Y_{n,k}|Z_{1:n},U_{1:n}) = \eta P(Z_n|Y_{n,k}) \sum_{i\in\{0,1\}} P(Y_{n,k}|Y_{n-1,i},U_n) P(Y_{n-1,i}) \quad (9)$$

with $\eta = 1/\left(\sum_{k\in\{0,1\}} P(Y_{n,k})\right)$. Equ. (9) allows to incorporate different assumptions and occlusion indicators through the state transition probabilities $P(Y_{n,k}|Y_{n-1,i},U_n)$ and measurement likelihoods $P(Z_n|Y_{n,k})$. The above filtering density incorporates measurements and control inputs only up to image n, while we have all information, $Z_{1:F}$ and $U_{1:F}$, available at any time. Hence, for exploiting this fact to obtain even more reliable estimates, we compute the posterior probability via optimal Bayesian smoothing [18] as

$$P(Y_{n,k}|Z_{1:F},U_{1:F}) = P_{\mathrm{f}}(Y_{n,k}|Z_{1:n}) \sum_{i\in\{0,1\}} \frac{P(Y_{n+1,i}|Y_{n,k},U_{n+1}) P(Y_{n+1}|Z_{1:F})}{\sum_{j\in\{0,1\}} P(Y_{n+1,i}|Y_{n,j},U_{n+1}) P_{\mathrm{f}}(Y_{n,j}|Z_{1:n})}. \quad (10)$$

Note, while the per pixel filtering probabilities are computed based on a forward recursion ($n \rightarrow n+1$), the smoothing probabilities are computed based on a backward recursion ($n+1 \rightarrow n$). The occlusion probabilities $P(Y_{n,1}|Z_{1:F},U_{1:F})$ throughout the image sequence for each pixel then comprise the above mentioned occlusion maps $\tilde{\xi}(\mathbf{x},n)$. In the following, the proposed filter models are described.

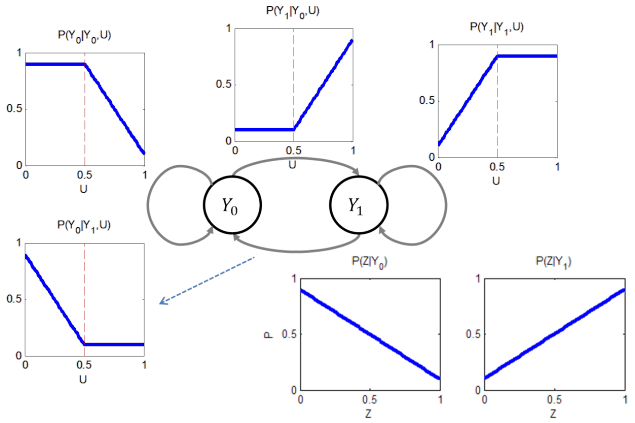


Figure 2: Shown are the conditional probability functions and a diagram of the state transition model (upper row and lower left). On the lower right are the two conditional probability functions for the measurement model. The functions are not normalized.

3.2.2 State transition model

The state transition model is based on the assumption of temporal and spatial coherence as well as flow field information modelled as uncertain control input U_n. The latter utilizes the mapping uniqueness criterion, i.e. the observation that multiple pixels in the reference image mapping to the same point in a target image can indicate an occlusion. For the considered pixel x and a local neighborhood A (typically a 7×7 pixels patch) this criterion can be formalized as [26]

$$U_n(A, \mathbf{x}, \mathbf{u}(\mathbf{x}, n))) := \frac{1}{|A|} \sum_{x \in A} T(m(\mathbf{x}+\mathbf{u}(\mathbf{x}, n))-1, 0, 1). \tag{11}$$

Here $m(\cdot)$ is the counter of reference pixels that map to (a small region around) the pixel $(\mathbf{x} + \mathbf{u}(\mathbf{x}, n))$ in the current image using forward warping. The function $T(a, l, h)$ truncates the value of a to $[l, h]$. The averaging over patch A was included for also taking spatial coherence into account. Equ. (11) in particular provides a good indicator for self-occlusions which is incorporated into the state transition model as uncertain control input U (*cf.* Figure 3). Thus, the four conditional probability functions comprising the full model ($P(Y_{n,k}|Y_{n-1,i}, U_n) = P(Y_k|Y_i, U)$, $i, k \in \{0, 1\}$) have been defined as shown in Figure 2. By design, through the truncation of the linear mappings the model prefers state preservation over state change in order to account for temporal coherence.

3.2.3 Measurement model

The measurement model is based on the divergence of the flow fields that are calculated in the first stage. The divergence is known to measure sinks (negative divergence) and sources (positive divergence) in the flow fields. The negative divergence essentially measures the same as the counter indicator described above was already used to indicate occlusions produced by foreground objects [3]. However, we are not only interested in self-occlusions, but also in external occlusions due to newly incoming objects, which result in positive divergence. Moreover, we observed that local divergences of the flow estimated in Stage 1 show a high variance in external occlusion regions. This can be interpreted as flow uncertainty resulting in spurious sinks and sources. As occlusion measurement Z for pixel $\mathbf{x}$, we therefore propose an indicator that relies on the weighted divergence variance calculated over the local neighborhood A (typically 7×7 pixels) around $\mathbf{x}$

$$Z_n(\mathbf{u}(\mathbf{x}, n)) := w_{div} \operatorname{var}_A(\operatorname{div}(\mathbf{u}(\mathbf{x}, n)). \tag{12}$$

In order to fit the measurements Z_n into the margin $Z \in [0, 1]$ we truncate Z_n at a maximum value $Z_{\max} = 1$ that can be influenced via the weighting w_{div}. This is a parameter that can be adapted in order to allow for different sensitivities with respect to the divergence of the estimated flow-fields. We typically choose $w_{div} \in [0, 1]$, usually $w_{div} = 1$, where larger values amount for higher sensitivity with respect to sinks and sources in the flow-fields. The measurement likelihood model $P(Z_n|Y_{n,k})$ is then defined as shown in Figure 2. The indicator is shown in Figure 3. Note that both the state transition and the measurement probability functions were designed in a way to avoid probabilities close to zero and one in order to account for the inherent uncertainty. In both cases we choose $P = 0.1$ as a lower bound and $P = 0.9$ as an upper bound. Moreover, the continuous functions are discretized through a histogram approximation, in order to be usable in Equ. (9) and (10), while at the same time ensuring that the discrete probabilities sum up to one.

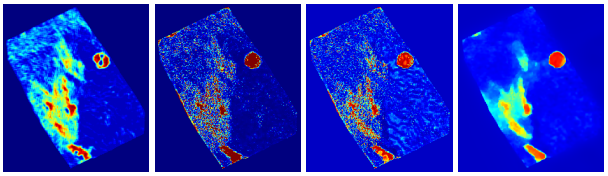


Figure 3: Left to right: Occlusion probabilities of counter indicator, divergence indicator, pre-estimate (Stage 2) and final occlusion map (Stage 3) after global smoothing. While the counter indicator detects self-occlusions, but fails to provide a complete detection of external occlusions, the divergence indicator detects both, but also introduces spurious detections at the borders. Fusing both types of information in combination with global smoothing results in reliable occlusion maps. See Figure 1 for the color scale.

3.3. Occlusion-aware MFOF (Stage 4)

After the global smoothing (Stage 3, see [16] and the supplementary material) has been applied to $\tilde{\xi}(\mathbf{x}, n)$, we binarize the probability map with a threshold of 0.5, resulting

Method	Original	Noise	Occl.	L. occl.
LDOF [5]	1.71	4.35	2.01	—
I-TVL1 [25]	1.43	2.61	1.89	—
S-Occl-R [14]	1.24	1.94	1.27	—
Best MFOF [10]	0.69	1.43	0.76	—
MDP [27]	0.79	2.98	0.90	1.71
EpicFlow [15]	0.55	2.56	0.59	1.08
Base scheme	**0.2939**	**1.19**	**0.49**	**1.44**
MFOF Prop.-1	**0.34**	**1.11**	**0.35**	**0.58**
MFOF Prop.-2	**0.32**	**1.04**	**0.33**	**0.46**
MFOF (sparse)	**0.31**	**0.92**	**0.34**	**0.42**
MFOF Prop.-3	**0.2906**	**0.89**	**0.32**	**0.34**

Table 1: Quantitative evaluation of different versions of our proposed method against other MFOF and TFOF methods in terms of pixel EPE on the waving flag benchmark dataset.

in $\xi(\mathbf{x}, n)$, which is used in Equ. (4) for correcting the optical flow. Thus, the coefficients for the subspace constraints are only computed from reliable data of visible regions. This information is extrapolated into the occluded regions (via the minimization) and, thus, still considers the general non-rigid motion. Therefore, this procedure enables us to robustly compute pixel trajectories of a non-rigidly deforming object, even if the object is largely occluded in several frames. Note that the restriction to the subspace constraints is only strictly imposed in regions that are marked occluded. The accuracy outside of those regions is up to the used optical method and does not influence other potential features of the method such as handling of large displacements.

4. Evaluation

In this section we quantitatively and qualitatively evaluate the proposed method. We start with the quantitative evaluation on a challenging sequence with respect to non-rigid motion in a video sequence of a waving flag with given ground truth data [1]. This dataset has been created to cover the complexity of realistic non-rigid motion with rather small occlusions synthetically added in [10] and larger occlusions added for the evaluation of our work. To the best of our knowledge, there is currently no other publicly available dataset that provides groundtruth optical flow with respect to one reference frame for a long video sequence showing a highly non-rigid object.The popular optical flow datasets Middlebury [2], SINTEL [6] and KITTI [11] only provide ground truth flow for consecutive frames, while we are interested in registering each image in a video sequence with respect to one chosen reference frame as required for dense NRSfM (*e.g.* [9]). Note, when considering only two consecutive frames, our algorithm reduces to the underlying I-TVL1 algorithm [25] with the colour extension of [9] and the presented occlusion handling which is also related to [3]. Both latter algorithms have already been tested on the Middlebury dataset and we do not claim to largely outperform these methods in the typical two-frame scenarios. In order to show the improvement of our approach in the highly non-rigid multi-frame scenario, we adapted different top performing optical flow algorithms to our setting by registering each image separately to the reference image. We compare against EpicFlow [15] and Motion Detail Preserving (MDP) optical flow [27] which both include explicit occlusion handling (*cf.* Section 2.2) as well as against other standard and state-of-the-art algorithms in the benchmark dataset with rather small occlusions. Finally, we present qualitative results including moving specularities and large occlusions on a real dataset of a beating heart during a Bypass surgery[2]. Further results can be found in the supplementary material.

4.1. Test on benchmark dataset (waving flag)

The authors of [10] provide different versions of the waving flag sequence consisting of 60 images with a resolution of 500×500 pixels, showing the flag in colour, as gray-scale images, with different forms of noise and (rather small) occlusions. The dataset has been used to compare different optical flow methods with respect to accuracy in the presence of non-rigid motion. Different variants of the MFOF method published in [10] outperformed other state-of-the-art methods regarding this task, such as, the I-TVL1 [25], the Large Displacement Optical Flow method (LDOF) [5] and an optical flow method based on self-occlusion reasoning [14]. The best performing MFOF variant throughout the tests was the one working on colour images with a PCA basis of different ranks (R $=$ 50, or full rank R $=$ 120) for the subspace constraints. The subspace basis was obtained via PCA from ground truth data for the base scheme, MFOF Prop.-1,-2,-3 or from sparse KLT tracks as described in [16] for MFOF (sparse). We compared our method with the above mentioned ones as well as EpicFlow [15] and the MDP method [27], using the same rank of subspace constraints. We choose the same setting as mentioned in [10], use the same amount of iterations and the same constants for the data and linking terms for the occlusion sequence, namely 20 alternation iterations, $\beta = 0.4$ and $\alpha = 18/\sqrt{3}$, to obtain a fair comparability. For the divergence indicator we use $w_{div} = 1$ for all tests. Table 1 shows the root mean squared end-point error (EPE) in pixels of the different optical flow methods averaged over the sequence, together with our results on different versions of the sequence. As [10] we compute the EPE over all foreground pixels of the whole sequence. (Original) is the original sequence. (Noise) is a version of the sequence with additional Gaus-

[1] http://www0.cs.ucl.ac.uk/staff/lagapito/subspace_flow/

[2] Video is available on http://hamlyn.doc.ic.ac.uk/vision

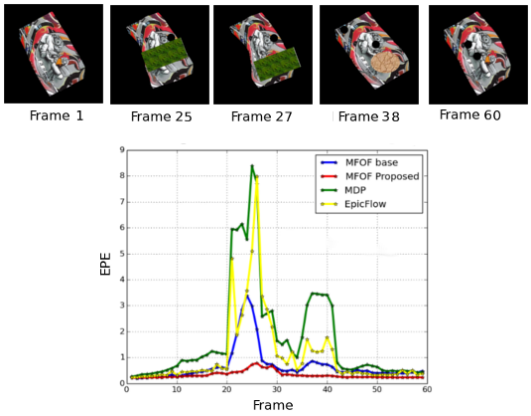


Figure 4: Comparison of the EPE of each frame on the flag sequence with large occlusions. It can be clearly seen that the EPE of our proposed approach is particularly lower in the case of large occlusions and non-rigid motion around frame 25 and 38.

sian noise. (Occl.) denotes a version of the sequence with additional occlusions. All these sequences are presented on the above mentioned website. Moreover, we included a version of the sequence with large occlusions (L. occl.), where some frames with the occlusions can be observed in Figure 4. Our **base scheme** consists in computing the initial flow with the above settings for all frames (*cf.* Section 3.1) with the proposed pre-iterations in Alg. 1. It can be observed that the base scheme gives quite accurate results for the original sequence, without occlusions. However, the accuracy degrades as more occlusions are included. The method **MFOF Prop.-1** consists of a fast version of the whole approach, without pre-iterations and only 5 alternation iterations, in order to compare against the same fast approach **MFOF Prop.-2**, with the pre- iterations. It can be observed that the pre-iterations increase the accuracy of the method **MFOF Prop.-2** in particular in the presence of large occlusions. This originates from the increase in convergence speed due to the initialization via the pre-computed fields. The proposed occlusion estimation and handling increases the accuracy of both methods in particular in the case of occlusions. The method **MFOF Prop.-3 or MFOF Proposed** is the most accurate method and is based on the full approach, with the same amount of alternation iterations as the base scheme. The proposed approach with occlusion handling does not degrade the accuracy of the base scheme in the original sequence without occlusions and yields consistently more accurate results in the presence of noise, and in the presence of small as well as large occlusions. The method **MFOF (sparse)** uses the same parameters as MFOF Prop.-3 and is based on pre-computed sparse tracks. This indicates that the performance of the method does not depend on highly accurate point tracks for the trajectories to obtain the subspace constraints. However, the rank (R) should be appropriate (rather low) for large occlusions,

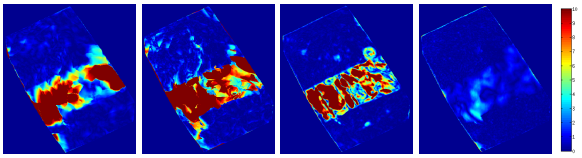


Figure 5: Comparison of the EPE ($\in [0..10]$ pixels, left to right: EpicFlow, MDP, MFOF Base, MFOF Proposed) at frame 25 on non-rigid motion of the flag with small and large occlusions. All methods can reasonably well handle the small occlusion (black disc in Figure 1), but only our proposed scheme is able to estimate an accurate flow in the presence of the large occlusion (green leaves in Figure 1).

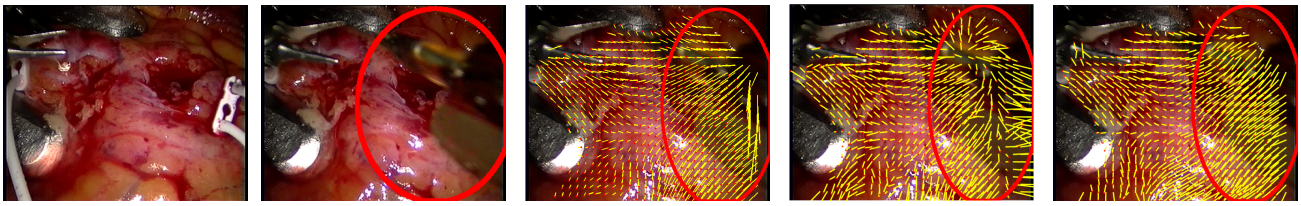
Method	Runtime
MDP [27]	25172 sec. (419 min. 30 sec.)
EpicFlow [15]	750 sec. (12 min. 30 sec.)
Base scheme	308 sec. (5 min. 13 sec.)
MFOF Prop.-3	5101 sec. (85 min. 01 sec.)
MFOF (sparse)	5108 sec. (85 min. 08 sec.)
MFOF Prop.-2	699 sec. (11 min. 39 sec.)

Table 2: Single core runtimes on sequence (L. Occl.).

since this restricts the motion in the occlusion to the most expressive motion. It can be observed that different versions of our method outperform all other methods considerably on the different versions of the dataset, which is due to the superior handling of the challenging task of accurately capturing the non-rigidity, together with an appropriate occlusion handling. Figure 4 shows a diagram of the EPE on each frame of different methods over the whole sequence with large occlusions. It can be seen that the proposed method gives consistently low errors. Figure 5 shows a comparison of the methods on frame 25, where each method has the largest error in the sequence. The occlusion indicators and occlusion maps in the different stages for this case can be observed in Figure 1. Table 2 compares the different runtimes of EpicFlow, MDP and two versions of our method. All methods have been executed, using only one CPU, over the whole sequence (L. Occl.). The time needed to compute initial tracks, if needed (EpicFlow and our methods), was included. The computations were performed on an Intel-Xeon CPU E3-1245 V2 with 3.4 GHz and 16GB RAM. It can be observed that MFOF Prop.-2 yields comparable running time to the highly efficient EpicFlow. MFOF Prop.-3, MFOF (sparse) need a reasonable runtime compared to the MDP method, while providing higher accuracy in all cases. Note that our Bayesian inference approach works on a per pixel bases and is thus highly parallelizable like the base scheme [10] and therefore our proposed schemes.

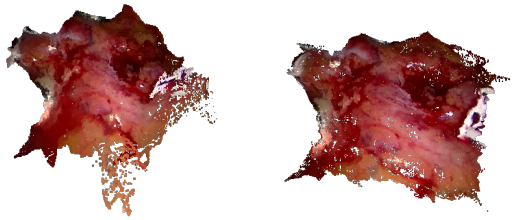
4.2. Test on real data

Please note that the base scheme that we build on has already been applied to several challenging real world exam-



(a) Reference frame (b) Frame 23 (c) EpicFlow (d) Base scheme (e) MFOF Proposed

Figure 6: Comparison of the different methods on a beating heart sequence with a large occlusion due to the robot arm (marked by circle). Note the different performance of the approaches regarding the correspondences (yellow lines). The correspondences of the proposed method are least affected by the occlusion.



(a) Base scheme (b) MFOF Proposed

Figure 7: Dense NRSfM reconstruction of frame 23 with the texture of the reference frame using [13], based on the correspondences from MFOF Base scheme (a) and the proposed method, MFOF (sparse) (b).

ples [10] including dense NRSfM [9]. We tested our method on a challenging dataset of a beating heart during a bypass surgery showing non-rigid deformations (self-occlusions), moving specularities and large occlusions due to a robot arm entering the scene. Since no ground truth data is available, we show qualitative results based on visual effects of the estimated flow fields. We used the same parameters and iterations as mentioned above with the method MFOF Proposed (MFOF Prop.-3), by working on the RGB gradients of the images (a 6 dimensional vector) for the base scheme and the proposed MFOF (sparse) method, due to the better performance on some textureless parts. The basis for the subspace constraints were obtained from sparse KLT tracks ($R = 10$), see [16]. Figure 6 shows every 10-th point correspondence of the dense optical flow computed via EpicFlow, the base scheme and our proposed scheme in a selected region of interest. It can be observed that the point tracks show inconsistent patterns in the occlusion part for the base scheme. The point tracks based on EpicFlow look much better due to an inherent occlusion consideration, but, show distorted tracks at the edges (right part of Figure 6(c)). The proposed method shows a more consistent pattern of estimated correspondences in the region that is occluded by the robot arm (Figure 6(e)). In Figure 7 dense NRSfM reconstructions, based on uncorrected correspondences (from the base scheme) and the proposed MFOF method, are shown. The increase in reconstruction quality, obtained from the correspondences of the proposed method, with no distortions on the right part, is clearly visable. Further results of dense NRSfM reconstructions based on the base scheme and the proposed scheme of the waving flag sequence with large occlusions as well as applications to a "turning head scenario" and a video of a "deforming human back" can be found in the supplementary material.

5. Conclusions

We propose a novel method for accurate video registration of highly deformable objects under sub-optimal conditions including occlusions of different size and type (*e.g.* self-occlusions, external occlusions), noise and specularities. To this end, we developed a new flow uncertainty estimation method based on Bayesian inference that fuses coherence assumptions with (an extendable set of) different occlusion indicators together into an occlusion pre-estimate that is thereafter globally smoothed to obtain a spatially and temporally consistent estimate of an occlusion probability. Furthermore, we include this information into an occlusion-aware MFOF method to compute corrected optical flow fields. This has been shown to increase the accuracy, in particular in regions of large occlusions, in the challenging case of the video registration of highly non-rigid objects. We also proposed an iterative pre-computation scheme that further increases the overall convergence speed and accuracy of the method and thus leads to an efficient occlusion aware video registration method. The latter eases the way to accurate dense NRSfM methods in challenging real world scenarios. Future work will focus on an automatic estimation of the underlying parameters, in particular w_{div} from the data. A suitable integration with dense NRSfM methods on a GPU is also planned.

Acknowledgements

This work was funded by the BMBF project DYNAMICS (01IW15003).

References

[1] L. Alvarez, R. Deriche, T. Papadopoulo, and J. Sánchez. Symmetrical dense optical flow estimation with occlusions detection. *Int. J. Comput. Vision*, 75(3):371–385, Dec. 2007.

[2] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

[3] C. Ballester, L. Garrido, V. Lazcano, and V. Caselles. A tv-l1 optical flow method with occlusion detection. In A. Pinz, T. Pock, H. Bischof, and F. Leberl, editors, *DAGM/OAGM Symposium*, volume 7476 of *Lecture Notes in Computer Science*, pages 31–40. Springer, 2012.

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision (ECCV)*, volume 3024 of *Lecture Notes in Computer Science*, pages 25–36. Springer, May 2004.

[5] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):500–513, 2011.

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

[7] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, May 2011.

[8] R. Garg, L. Pizarro, D. Rueckert, and L. Agapito. Dense multi-frame optic flow for non-rigid objects using subspace constraints. In R. Kimmel, R. Klette, and A. Sugimoto, editors, *Computer Vision – ACCV 2010*, volume 6495 of *Lecture Notes in Computer Science*, pages 460–473. Springer Berlin Heidelberg, 2011.

[9] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[10] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *International Journal of Computer Vision*, pages 1–29, 2013.

[11] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[12] B. K. Horn and B. G. Schunck. Determining optical flow. Technical report, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, 1980.

[13] M. Paladini, A. D. Bue, J. M. F. Xavier, L. de Agapito, M. Stosic, and M. Dodig. Optimal metric projections for deformable and articulated structure-from-motion. *International Journal of Computer Vision*, 96(2):252–276, 2012.

[14] D. Pizarro and A. Bartoli. Feature-based deformable surface detection with self-occlusion reasoning. *Int. J. Comput. Vision*, 97(1):54–70, Mar. 2012.

[15] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. *CoRR*, abs/1501.02565, 2015.

[16] S. Ricco and C. Tomasi. Dense lagrangian motion estimation with occlusions. In *CVPR*, pages 1800–1807. IEEE, 2012.

[17] S. Ricco and C. Tomasi. Video motion for every visible point. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[18] S. Särkkä. *Bayesian filtering and smoothing*. Cambridge University Press, 2013.

[19] D. Sun, C. Liu, and H. Pfister. Local layering for joint motion estimation and occlusion detection. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1098–1105, 2014.

[20] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision (IJCV)*, 106(2):115–137, 2014.

[21] D. Sun, E. B. Sudderth, and M. J. Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2226–2234. Curran Associates, Inc., 2010.

[22] Y. Tian and S. G. Narasimhan. A globally optimal data-driven approach for image distortion estimation. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 1277–1284, 2010.

[23] M. Unger, M. Werlberger, T. Pock, and H. Bischof. Joint motion estimation and segmentation of complex scenes with label costs and occlusion modeling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1878–1885, June 2012.

[24] A. Wedel and D. Cremers. *Stereo Scene Flow for 3D Motion Analysis*. Springer, 2011.

[25] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers. An improved algorithm for tv-l 1 optical flow. In D. Cremers, B. Rosenhahn, A. Yuille, and F. Schmidt, editors, *Statistical and Geometrical Approaches to Visual Motion Analysis*, volume 5604 of *Lecture Notes in Computer Science*, pages 23–45. Springer Berlin Heidelberg, 2009.

[26] L. Xu, J. Jia, and Y. Matsushita. Motion detail preserving optical flow estimation. In *CVPR*, pages 1293–1300. IEEE, 2010.

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

[28] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *Proceedings of the 29th DAGM Conference on Pattern Recognition*, pages 214–223, Berlin, Heidelberg, 2007. Springer-Verlag.