NRSfM-Flow: Recovering Non-Rigid Scene Flow from Monocular Image Sequences

Vladislav Golyanik[1,2]
http://av.dfki.de/members/golyanik/
Aman Shankar Mathur[1]
asmathur@rhrk.uni-kl.de
Didier Stricker[1,2]
http://av.dfki.de/members/stricker/

[1] Department of Computer Science
University of Kaiserslautern
Kaiserslautern, Germany

[2] Department Augmented Vision
German Research Center for Artificial Intelligence (DFKI GmbH)
Kaiserslautern, Germany

Abstract

Scene flow recovery from monocular image sequences is an emerging field in computer vision. While existing Monocular Scene Flow (MSF) methods extend the classical optical flow formulation to estimate depths/disparities and 3D motion, we propose a framework based on Non-Rigid Structure from Motion (NRSfM) technique — NRSfM-Flow. Therefore, both problems are formulated in the continuous domain and relation between them is established. To cope with real data, we propose two preprocessing steps for image sequences — redundancy removal and translation resolution — which increase quality of reconstructions and speedup computations. In contrast to the existing MSF methods which can cope with non-rigid deformations, our solution makes no strong assumptions about a scene such as known camera motion or camera velocity constancy and can handle occlusions. NRSfM-Flow is qualitatively evaluated on challenging real-world data. Experiments provide evidence that the proposed approach achieves high accuracy and outperforms state of the art in terms of the ability to reconstruct MSF with less prior knowledge about a scene.

1 Introduction

Scene flow recovery from monocular image sequences is an emerging field in computer vision. As of today, this topic was sparsely discussed in literature and only few works exist. Scene flow refers to a dense 3D velocity vector field of a moving and possibly non-rigidly deforming scene, see Fig. 1 for an example. The concept is similar to optical flow, i.e. a dense 2D velocity vector field in an image plane. Scene flow finds applications in autonomous driving, motion segmentation, motion capture, egomotion, 4D reconstruction, scientific visualization and other domains. In real applications, scene flow computation is mostly based on stereo/multi view camera settings or sensors directly outputting depth.

The earliest work on monocular scene flow (MSF) recovery was carried out by Birkbeck *et al*. The variational method proposed in [5] supports short image sequences and can handle rigid, articulated and non-rigid motion. Several aspects may restrict applicability of the approach in practice, i.e. camera motion is assumed to be known in advance and constant in

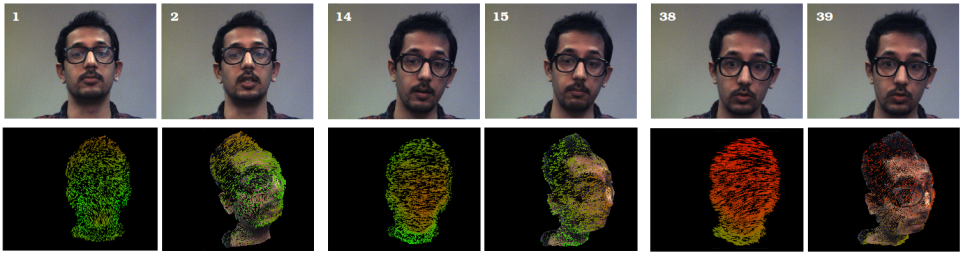

Figure 1: Results of the proposed NRSfM-Flow framework on the *human face* sequence, for several pairs of frames. For every pair of frames: input frames (top row), recovered scene flow (down left) and geometry with overlayed 3D motion fields from a new viewpoint. Better viewed in colour. See supplementary material for video. The proposed approach is able to recover scene flow from monocular image sequences depicting non-rigid scenes and does not make any assumptions on the scene or type of camera motion. It is robust to occlusions, among other things.

a short temporal frame. Another limitation is a high sensitivity to occlusions. In [6], the same authors proposed a solution without the requirement of a known camera motion, but with a known rigidly moving base-mesh geometry approximation of the scene. Mitiche *et al*. recently proposed a variational method for concurrent recovery of structure and scene flow [17]. As shown experimentally, the algorithm can handle noiseless scenarios with rigid motion including scenes with few moving objects. Tikhonov regularization used in the algorithm tends to oversmooth depth and motion discontinuities in the recovered 3D flow fields. Further studies are required to make the technique applicable in more complex and realistic scenarios. Motivated by remote patient monitoring and driver assistance systems, Xiao *et al*. proposed an MSF estimation algorithm based on energy functional minimization [31]. Along with brightness and gradient constancy assumptions, velocity constancy over a short period of time is reflected in the energy functional. The algorithm can use at least three frames and does not require optical flow as an input. In experiments, an application of the proposed technique in a challenging real-world driving scenario was demonstrated which is commonly tackled by stereo based methods. However, support of scenes exhibiting non-rigid deformations is limited.

The discussed methods share several common attributes. Firstly, they are formulated as energy minimization problems solved by an Euler-Lagrange differential equation. They extend a classic optical flow formulation[1] by estimating depths/disparities and a 3D motion field instead of image motion and estimate correspondences and geometry simultaneously. Secondly, intrinsic camera parameters are assumed to be known. Thirdly, the reviewed methods operate in a batched manner, i.e. they compute scene flow after the complete image sequence is acquired. Fourthly, support of non-rigidly deforming structures is limited. Processing of non-rigid scenes was demonstrated in the papers by Birkbeck *et al*. [5, 6], but assumptions which need to be satisfied limit their applicability in practice considerably. Multiple common aspects encourage us to classify the reviewed MSF methods into a separate class which we refer to as *direct* MSF methods.

Extension of optical flow to three dimensions is one possible approach to the problem of MSF recovery. Another one is to adopt Non-Rigid Structure from Motion (NRSfM) techniques. NRSfM allows reconstruction of non-rigidly deforming and moving 3D surfaces from monocular image sequences given coordinates of tracked points for every frame (combined in a measurement matrix). Birkbeck *et al*. [5, 6] mention NRSfM as a class of tech-

[1] which has its roots in the seminal work by Horn and Schunck [14]

niques that can be potentially used to recover geometry of deformable scenes — a scenario similar to MSF recovery. However, an explicit step for 3D motion field estimation is not present in NRSfM methods. Another possible limiting factor for adopting NRSfM for scene flow estimation at that time was the lack of dense techniques.

NRSfM methods take advantage of motion and non-rigid deformation as cues to infer geometry. They are based on factorisation of a measurement matrix with coordinates of the tracked points into non-rigid shape and motion for every frame — an inverse problem inherently ill-posed in the sense of Hadamard. Additional constraints are required to obtain a unique and reasonable solution. For an orthographic camera, the factorisation idea was first proposed by Tomasi and Kanade for the rigid case [26] and later extended to the non-rigid case by Bregler *et al*. [7]. In [7], every non-rigid shape is represented by a linear combination of basis shapes, wherein the basis shapes and the weights are unknowns. This idea was further improved in successor methods proposing different types of constraints and optimization methods for higher stability and reconstruction accuracy [12, 13, 19, 20, 28] and for sequential operation [2, 18]. Akhter *et al*. proposed employment of a trajectory basis [3] instead of the metric one. It allows reduction of the number of unknowns in NRSfM, since the trajectory basis is fixed in advance. As a result, NRSfM in the trajectory space can lead to more stable reconstructions. Several papers investigated ways to improve this class of methods and eliminate weaknesses such as ambiguity in trajectory bases [30] or to model point trajectories more realistically [33]. The first dense NRSfM method was shown in [23] followed by [10] which is currently one of the best performing methods qualitatively.

In contrast to direct MSF methods, NRSfM relies on correspondences obtained in a separate step. The demand of sufficient motion implies a certain length of the image sequence. Another assumption commonly made in NRSfM methods is that a scene is centered throughout the whole image sequence (similar to the direct method of [5]), which may limit application of NRSfM methods in real-world scenarios.

Thus, several additional steps are required to adopt current NRSfM methods to the problem of MSF. Accordingly, the following contributions are made in this paper: 1) relation between NRSfM and MSF is established. A novel analytical framework is introduced which allows to analyse and relate both problems in the continuous domain on a high level of abstraction; 2) a solution to MSF recovery based on the extensively studied NRSfM under orthography is proposed — the NRSfM-Flow; 3) two preprocessing steps are proposed — to resolve translation in a scene and to compress the input image sequence by eliminating redundant frames (frames with low variance in scene appearance) — so as to reduce the runtime and enhance the overall accuracy of the approach; 4) NRSfM-Flow is designed and implemented as a framework combining state of the art methods for correspondence computation [11, 27], non-rigid geometry reconstruction [10] and proposed preprocessing steps; 5) results on several real-world image sequences are shown and performance of the proposed approach is evaluated qualitatively. We consider MSF estimation as a standalone field in computer vision. To the best of our knowledge, we are the first to propose an NRSfM-based MSF method, to formulate NRSfM in the continuous domain and to propose explicit translation resolution and redundancy removal for NRSfM. Our approach can handle challenging scenarios with non-rigid motion which cannot be handled by the current direct monocular scene flow methods.

The rest of the paper is organized as follows. In Sec. 2, the formulation of NRSfM in the continuous domain as well as relation of NRSfM and scene flow is derived. In Sec. 3, the NRSfM-Flow framework is described together with preprocessing steps broadening the scope of NRSfM, followed by experiments in Sec. 4 and conclusions in Sec. 5.

2 MSF and NRSfM in the continuous domain

In this section, relation between NRSfM and MSF is established. Therefore, both problems are formulated in the continuous domain. Continuous representation often allows one to analyse problems and reveal their properties on a high level of abstraction. Moreover, multiple variants of discretisation, numerical and optimization methods are possible in combination with it. In other words, the problem statement and its implementation is kept separate.

Assume an orthographic camera observes a 3D non-rigidly deforming scene $\mathbf{S}(\mathbf{p},t)$ consisting of 3D points $\mathbf{p}$ from the continuous point space domain $\Omega \subset \mathbb{R}^{3+1}$. Every point possesses a colour, hence an additional space dimension denoted by $+1$. The observed scene is different at each time $t \in \mathrm{T} \subset \mathbb{R}$:

$$\mathbf{S}(\mathbf{p},t) : \Omega \times \mathrm{T} \to \mathbb{R}^{3+1}. \tag{1}$$

The scene $\mathbf{S}(\mathbf{p},t)$ continuously produces 2D projections (images) on the camera sensor containing 2D points $\mathbf{v}$ from the image domain $\Psi \subset \mathbb{R}^{2+1}$:

$$\mathbf{I}(\mathbf{v},t) : \Psi \times \mathrm{T} \to \mathbb{R}^{2+1}. \tag{2}$$

We assume that the scene is registered to the origin of the coordinate system and the camera translation $T(t)$ is always $\mathbf{0}$. The scene and its image is related as

$$\mathbf{W}(\hat{\mathbf{v}},t) = \mathbf{W}_\tau(\hat{\mathbf{v}},t) + \mathbf{C}(\hat{\mathbf{v}}) = \mathbf{R}(t)\,\mathbf{S}(\mathbf{p},t), \tag{3}$$

where $\mathbf{R}(t) : \mathrm{T} \to SO(3)$ is the camera pose, $\mathbf{W}(\hat{\mathbf{v}},t)$ is a measurement function (image coordinates of the tracked points); the correspondence (2D motion field) function $\mathbf{W}_\tau(\hat{\mathbf{v}},t)$ outputs a 2D displacement field relative to a reference time τ and $\mathbf{C}(\hat{\mathbf{v}})$ is the point displacement function in image coordinates relative to the origin of the coordinate system of the image. Note that $\hat{\mathbf{v}} \in \hat{\Psi} \subset \mathbb{R}^2$ are 2D colourless points and $\hat{\mathbf{v}} \subset \mathbf{v} : \mathbf{v}$ are visible at time τ; in $\mathbf{I}(\mathbf{v},t)$, point displacements are given relative to the changing reference time τ, whereas in the case of $\mathbf{W}_\tau(\hat{\mathbf{v}},t)$ the reference time τ is fixed (see Fig. 2 for geometric interpretations). An infinitesimal change in camera pose and 3D scene structure $\Theta(\mathbf{p},t) : \Omega \times \mathrm{T} \to \mathbb{R}^3$ can be described by a derivative of the right side of Eq. (3):

$$\Theta(\mathbf{p},t) = \frac{\partial \mathbf{R}(t)}{\partial t}\mathbf{S}(\mathbf{p},t) + \mathbf{R}(t)\frac{\partial \mathbf{S}(\mathbf{p},t)}{\partial t}, \tag{4}$$

where $\frac{\partial}{\partial t}$ denotes a partial derivative with respect to time t. Note that unlike $\mathbf{S}(\mathbf{p},t)$, $\Theta(\mathbf{p},t)$ represents a continuous 3D vector field, i.e. the 3D output encodes relative displacements of points $\mathbf{p}$, or *scene flow*. The scene flow is composed of a *rotational component* $\rho(t) = \frac{d\mathbf{R}(t)}{dt} \in SO(3)$ and a *deformational component* $\frac{\partial \mathbf{S}(\mathbf{p},t)}{\partial t}$.

Recall that in NRSfM, there is an inherent rotational ambiguity, i.e. the rotational component can be explained either by a camera or an object movement or a combination of both; without a prior knowledge it is not possible to determine a cause of the observed rotation. This ambiguity means that all combinations are possible and lead to equivalent observations. Assume that the camera is fixed, namely $\forall t : \mathbf{R}(t) = \mathbf{I}$ and the object rotates. In this case $\mathbf{S}(\mathbf{p},t)$ also covers observed rotations in the scene. We can exploit the rotational ambiguity to simplify Eq. (4) — the rotational component is equal to zero and only the right term remains in the expression:

$$\Theta(\mathbf{p},t) = \frac{\partial \mathbf{S}(\mathbf{p},t)}{\partial t}. \tag{5}$$

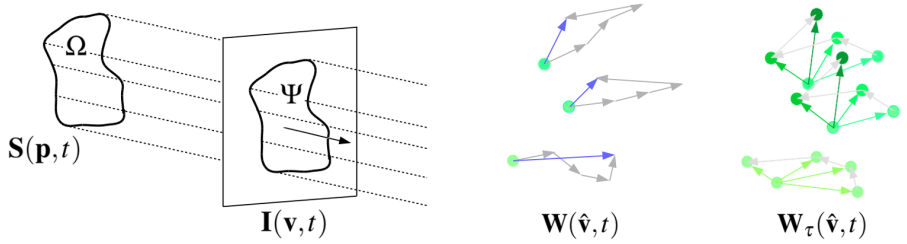

Figure 2: On the left: orthographic projection of a 3D object to a 2D image plane — the projection lines are parallel and intersect at infinity. In the middle: function $\mathbf{W}(\hat{\mathbf{v}},t) = \mathbf{W}_\tau(\hat{\mathbf{v}},t) + \mathbf{C}(\hat{\mathbf{v}})$ as in Eq. (3). $\mathbf{W}(\hat{\mathbf{v}},t)$ outputs *absolute* coordinates of the tracked points. On the right: function $\mathbf{W}_\tau(\hat{\mathbf{v}},t)$ visualized, see Eq. (6) — it outputs 2D displacements of the points visible at the reference time τ for every time t.

domain	meaning	defined notions	equations
$\mathbf{p} \in \Omega \subset \mathbb{R}^{3+1}$	all 3D points of a scene	3D scene $\mathbf{S}(\mathbf{p},t)$, scene flow $\Theta(\mathbf{p},t)$	Eqs. (3)–(5), (9)
$\hat{\mathbf{p}} \in \hat{\Omega} \subset \mathbb{R}^3$	reconstructed 3D points	reconstructed 3D surface $\mathbf{S}(\hat{\mathbf{p}},t)$	Eq. (8)
$\mathbf{v} \in \Psi \subset \mathbb{R}^{2+1}$	all observed 2D points	images $\mathbf{I}(\mathbf{v},t)$, optical flow $\boldsymbol{\Xi}(\mathbf{v},t)$	Eqs. (6)–(8)
$\hat{\mathbf{v}} \in \hat{\Psi} \subset \mathbb{R}^2$	2D points visible at time τ	measurement function $\mathbf{W}_\tau(\hat{\mathbf{v}},t)$	Eqs. (3), (6)

Table 1: Equations of the proposed theoretical framework relating NRSfM and MSF summarized

To insure $\forall t : \mathbf{R}_{2\times3}(t) = \mathbf{I}_{2\times3}$, the recovered rotation must be applied to the observed non-rigidly deforming structure in 3D space. Therefore, we obtain the $\mathbf{R}_{3\times3}(t)$ matrix by extending $\mathbf{R}_{2\times3}(t)$ with a third row. The third row is equal to a cross product of the first two rows, which guarantees orthonormality of the $\mathbf{R}_{3\times3}(t)$ matrix.

Likewise, the deformational component in the image plane can be computed from the image function $\mathbf{I}(\mathbf{v},t)$ as a continuous 2D vector field referred to as *optical flow* $\boldsymbol{\Xi}(\mathbf{v},t)$ (algorithms for computing optical flow from images are e.g. [14, 52]). The relation between the optical flow $\boldsymbol{\Xi}$ and the measurement function $\mathbf{W}_\tau$ reads:

$$\mathbf{W}_\tau(\hat{\mathbf{v}},t) = \int_\tau^t \boldsymbol{\Xi}(\hat{\mathbf{v}},t)\,dt. \tag{6}$$

Change of τ causes change in $\hat{\Psi}$, since the set of visible points is different at each time. From Eqs. (3), (4) and (6) the relation between the infinitesimal optical flow and the scene flow can be established as

$$\int_\tau^t \boldsymbol{\Xi}(\hat{\mathbf{v}},t)\,dt + \mathbf{C}(\hat{\mathbf{v}}) = \mathbf{R}(t)\,\mathbf{S}(\hat{\mathbf{p}},t), \quad \text{or} \tag{7}$$

$$\boldsymbol{\Xi}(\hat{\mathbf{v}},t) = \mathbf{R}_{2\times3}(t)\frac{\partial \mathbf{S}(\hat{\mathbf{p}},t)}{\partial t}, \tag{8}$$

where points $\hat{\mathbf{p}}$ and $\hat{\mathbf{v}}$ are the reconstructed 3D points and their projections into the image plane respectively. A summary of the domains and defined notions is given in Table 1. Similar to Eq. (3) which relates geometry of a non-rigid scene with its projection into an image plane, Eg. (8) relates changes in the geometry with changes in the projection. An accumulated scene flow (or equivalently, 3D point trajectories which can be also expressed as a set of 3D line integrals) in time interval $[t_1;t_2]$ reads as an integral

$$\int_{t_1}^{t_2} \Theta(\mathbf{p},t)\,dt. \tag{9}$$

Using the introduced space-time structures and relations, it is possible now to give the formal definitions of NRSfM and MSF recovery problems.

Definition of NRSfM. *Given the displacements of the projected points $W_\tau(\hat{\mathbf{v}},t)$ relative to time τ, the objective of an NRSfM problem is to recover the underlying non-rigidly deforming scene function $S(\mathbf{p},t)$.*

Definition of MSF. *Given projections of the observed scene $\mathbf{I}(\mathbf{v},t)$, the objective of an MSF problem is to reconstruct the scene flow function $\Theta(\mathbf{p},t)$.*

As follows from the definitions, the inputs and objectives of NRSfM and MSF are different, but related with Eqs. (6) and (4). Thus, it is possible to adopt NRSfM for estimation of scene flow from monocular image sequences using the proposed analytical framework.

3 The NRSfM-Flow framework

In this section the NRSfM-Flow framework for MSF recovery is introduced. It encompasses several steps including correspondence computation, geometry reconstruction as well as preprocessing steps for input image sequences. Though NRSfM-Flow is designed with batch processing in mind, it can be adopted for sequential processing.

To compute the measurement function $\mathbf{W}(\hat{\mathbf{v}},t)$, we adopt the state of the art Multi-Frame Optical Flow (MFOF) method of Garg *et al.* [11]. In the case of severe occlusions presented in a scene, we also use the occlusion-aware MFOF of Taetz *et al.* [27]. To recover non-rigid geometry and camera pose, we choose variational approach [10] combined with the GrabCut algorithm [21] for foreground-background segmentation. Thus, the methods aiming at high quality reconstructions are combined in NRSfM-Flow.

Preprocessing steps. NRSfM methods require sufficient diversity in non-rigid deformations and camera motion as reconstruction cues. We propose to compress an input image sequence so that it fulfils temporal and spatial assumptions of NRSfM in an optimal way. We call this preprocessing step *redundancy removal*. Suppose at time t_a an instantaneous image is considered for further processing. The next instantaneous image will be taken at time t_b for which the inequality holds:

$$\left\| \int_{\hat{\Psi}} \int_{t_a}^{t_b} \boldsymbol{\Xi}(\mathbf{v},t)\, dt\, d\hat{\mathbf{v}} \right\|_2 \geq \varepsilon, \tag{10}$$

where $\|.\|_2$ denotes a 2-norm and ε is a scalar threshold. In other words, if total flow (2-norm of the integrated flow field) in a time interval $[t_a; t_b]$ is above a threshold ε, then a view at time t_b exhibits sufficient diversity relative to the view at time t_a. Otherwise, another time interval $[t_a; t_b = t_b + dt]$ should be evaluated. The optimality criterion proposed in Eq. (10) can detect duplicate frames, small motions as well as oscillatory effects. Moreover, it can also serve as a discretisation criterion, since the regularization parameter ε determines whether the observed motion provides a sufficient reconstruction cue.

Though it is possible to resolve translation in a scene directly by registering the measurement matrix to the mean coordinates of the structure, we notice that resolving it before computing correspondences may increase accuracy of reconstructions. Therefore, we propose an explicit *translation resolution* step. Assuming that an object is entirely visible at the reference

time τ, we segment the scene into foreground and background and track the ROI throughout the image sequence using Kanade-Lucas-Tomasi feature tracker [16]. The output of the translation resolution is a frame size and the corresponding translation function $T(t)$.
After applying redundancy removal and translation resolution, correspondences are computed faster. If required, reverse transformations can be applied in the postprocessing step.
The NRSfM-Flow framework is summarized in an algorithmic fashion in Alg. 1. Note that the framework does not prescribe particular algorithms for the steps 2–5. They can be chosen or tuned dependent on requirements (e.g. perspective or orthographic camera model, etc.).

Algorithm 1 NRSfM-Flow framework for MSF recovery

Input: monocular image sequence $\mathbf{I}(\mathbf{v},t) : \Psi \times \mathrm{T} \rightarrow \mathbb{R}^{2+1}$.
Output: Scene flow $\Theta(\mathbf{p},t) : \Omega \times \mathrm{T} \rightarrow \mathbb{R}^3$
1: **Initialization:** *depends on the underlying algorithms*
2: Resolve translation, find translation function $T(t)$
3: Compress image sequence (eliminate redundant frames) according to Eq. (10)
4: Compute measurement function $\mathbf{W}_\tau(\hat{\mathbf{v}},t)$
5: Factorize $\mathbf{W}_\tau(\hat{\mathbf{v}},t) + \mathbf{C}(\hat{\mathbf{v}})$ into non-rigid shapes $\mathbf{S}(\mathbf{p},t)$ and motion $\mathbf{R}(t)$
6: Apply $\mathbf{R}(t)$ to $\mathbf{S}(\mathbf{p},t)$
7: Compute scene flow according to Eq. (5)
8: Apply reverse transformations ($-T(t)$ and geometry duplication) if required
9: Save the final recovered scene flow in $\Theta(\mathbf{p},t)$

Implementation. Our test platform has 128 GB RAM, an NVIDIA GeForce TITAN Z GPU and an Intel Xeon E5-1650 v3 CPU. We use our own C++/CUDA C implementations of the methods [27] and [10] as well as a publicly available Matlab [15] version [22] of the method [11]. The preprocessing steps were implemented in C++ using the OpenCV 3.0 library [1]. The implementation of [10] supports heterogeneous platforms with a multi-core CPU and a CUDA-capable GPU. Since the framework is formulated in the continuous domain following Sec. 2, several discretisation aspects shall be mentioned here. In the beginning, we choose discretisation points to coincide with image frames, wherein for computing derivatives (step 7) forward finite differences between consecutive frames are used. In this case the accumulated scene flow between two consecutive frames is computed, as defined in Eq. (9). We are also able to estimate geometry between frames by interpolating the structure along the accumulated 3D motion fields.

4 Evaluation

In this section the proposed NRSfM-Flow framework is qualitatively evaluated on several challenging real-world image sequences depicting non-rigid scenes. The recovered 3D flow fields and reconstructions are visualized. Additionally, projections of the scene flow into the image plane and optical flow between consecutive frames are compared. We follow colour schemes for optical and scene flow fields proposed in [4] and [29] respectively. Scenes which can be handled by our method should fulfil the requirements, i.e. provide sufficient reconstruction cues and preferably consist of a single non-rigidly deforming and moving (possibly translating) object. We are not aware of scene flow benchmark datasets fulfilling the aforementioned requirements. Therefore, we opt for several real-world image sequences. All existing works on MSF recovery [5, 6, 17, 31] employ a similar evaluation methodology.

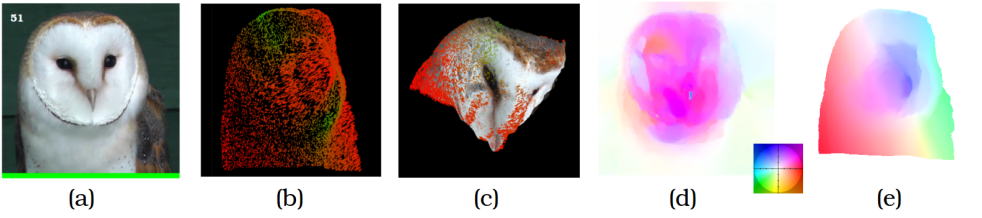

Figure 3: Experimental results on the *barn owl* sequence [9]: (a) input frame number 51; (b) result of the MSF recovery by NRSfM-Flow (between frames 51 and 52), the colour scheme is the same as in [29]; (c) geometry with an overlayed scene flow for the frame 51; (d) result of the optical flow between frames 51 and 52 by the TV-L1 method [32]; (e) projection of the recovered 3D motion field into the image plane. In (d) and (e) the colour scheme is replicated according to [9].

Figure 4: Examples of Poisson reconstructions: (a) shaded geometry from novel viewpoints from the *face* sequence (frame 1); (b) textured and shaded geometry from novel viewpoints from the *barn owl* sequence (frame 1).

The human face sequence. The face sequence was acquired with a Flea FL2-03S2C camera. It depicts a speaking person; arbitrary translations, facial expressions (non-rigid deformations) and self-occlusions are present in the sequence which makes it challenging for MSF recovery. Resolution of the images is 486×366. Translation resolution is applied in the preprocessing step. Exemplary results are shown in Fig. 1. Every recovered dense surface contains $5.6 \cdot 10^4$ points. Results of this experiment are qualitatively similar to the results on the *mouth* sequence shown in [5]. Though, several differences can be noticed. First, our results are less accurate in the areas of the forehead and mouth. The forehead is inherently poorly textured and correspondences as well as reconstructions are less accurate in this area. In the area of mouth, the points are interpolated building a smooth surface so that the opening is less recognisable in the shaded Poisson surface. However, both reconstructions exhibit artefacts associated with correspondences (there are convexities and concavities). Thereby, our reconstructions are more accurate in the cheek and side areas, see Fig. 4-a. Recall that our method does not rely on a known camera motion. The length of the sequence is 80 frames. Reducing the length to 40 frames does not result in decay of reconstruction accuracy. The runtime of NRSfM-Flow for the face sequence amounts to 805 seconds which is split amongst preprocessing, correspondence computation and surface recovery as 2, 771 and 32 seconds respectively. Note that we also tried the NRSfM-Flow pipeline without preprocessing on the face sequence. In that case, face reconstructions were unnaturally lengthened.

The barn owl sequence [9]. The sequence was acquired outdoors. It depicts a barn owl performing movement peculiar to a predator bird — jerky head turns followed by periods of a focused gaze. Due to the movements, observed surfaces deform non-rigidly. The sequence contains 602 frames in resolution 960×540. There is almost no translation, but there are a lot of redundant frames. In the preprocessing step, 400 out of 602 frames are removed using redundancy removal. An exemplary scene flow field for the barn owl sequence can be seen

in Fig. 3. Number of points per surface amounts to $2 \cdot 10^5$ and reconstructions look realistic. Scene geometry is correctly explained by rotational and deformational effects, which are especially well visualised in the supplementary video. See Fig. 4-b for exemplary textured and shaded Poisson surfaces. The runtimes for this sequence amount to 70, 1504, 6300 seconds for preprocessing, correspondence computation and surface recovery respectively.

Sintel Flow Dataset [8]. The Sintel Flow Dataset emerged as response to a growing demand

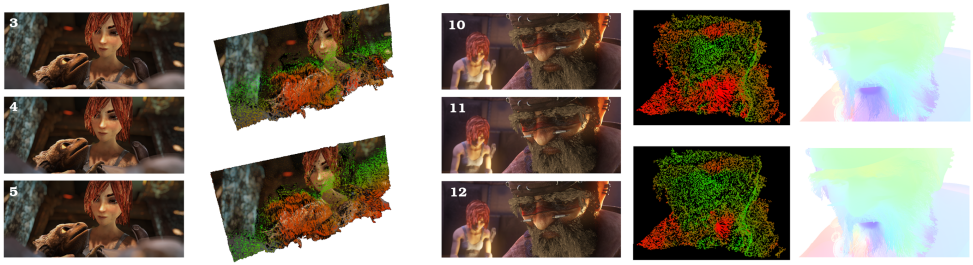

Figure 5: Experimental results on the Sintel Flow Dataset [8]: (a) selected frames from the *bandage2 (final)* sequence; (b) whole scene reconstructions with an overlayed scene flow between the corresponding frames on the left; (c) selected frames from the *shaman2 (final)* sequence; (d) enlarged groundtruth optical flow between the corresponding frames on the left; (e) scene flow between the corresponding frames in (c).

for evaluation of optical flow methods in challenging scenarios. It includes multiple monocular image sequences covering a broad range of realistic scenes varying by type of motion and deformations, environmental conditions and disturbing effects (motion blur, defocus). We tested the proposed framework on several Sintel sequences. Fig. 5 depicts selected results on *bandage2 (final)* and *shaman2 (final)* sequences with 50 non-redundant frames in 1024×436 resolution each. With the first one, we tested performance of NRSfM-flow in a complex scenario with multiple non-rigid objects. The result discloses few limitations of the proposed approach — without segmentation or a shape prior, the variational NRSfM cannot recover relative depths of individual parts correctly, mainly due to the assumed orthographic camera (see Fig. 5-a,-b). The relative depths of Sintel and Scales dragon are recovered correctly, but the background is inserted between them. In the case of additional regions (e.g. when the Sintel's hand enters the scene after the frame 20), more depth ambiguities occur. Another limitation concerns objects' boundaries — due to the variational nature, NRSfM produces smooth transitions from the foreground objects to the background. Those limitations define the open issues in the area of NRSfM. The *shaman2* sequence shows a slowly moving human face in the foreground (Fig. 5-c) and provides optimal conditions for reconstruction with current NRSfM methods. As a result, we were able to obtain accurate scene flow (Fig. 5-d) given a foreground-background mask, matching visually well with the groundtruth optical flow (Fig. 5-e). Both sequences took around 2000 seconds for correspondence computation and 450 seconds for surface recovery.

Supplementary material contains results on the face, barn owl, bandage2, shaman2 and several other sequences (heart [25], music notes and synthetic flag [11]) as videos.

Discussion. Due to MFOF and linear subspace model of the NRSfM, our approach can handle self- and external occlusions (e.g. occurring in the *bandage2* sequence). Using the proposed framework, it is possible to recover scene flow from monocular image sequences in scenarios not tackleable by existing MSF methods. Concerning NRSfM, we observe a favourable side effect. Serendipitiously, MSF allows visualisation of results of a 4D reconstruction better compared to sequentially showing recovered surfaces. MSF also enables one

to differentiate between rotational and deformational components in a convenient manner, analyse properties of NRSfM algorithms effectively, tune parameters and uncover directions for further algorithmic improvements.

Our framework inherits limitations peculiar to the current NRSfM methods. Most of them are able to reconstruct scenes which can be easily segmented in background and foreground. Scenes with multiple segments would preferably need additional preprocessing. Due to the orthographic camera model, the proposed framework does not recover absolute depths. Nevertheless, the most appropriate NRSfM method may be chosen depending on requirements (e.g. real-time operation, handling complex composed scenes or support of perspective views). For instance, the method of Russell *et al*. [24] allows joint segmentation and reconstruction complex real-world scenes. The NRSfM-Flow framework will directly benefit from advances in NRSfM methods. As shown experimentally, NRSfM-Flow is not real-time capable in its current form. However, by adopting sequential processing or iterative schemes, it will be possible to achieve real-time MSF recovery performance.

5 Conclusions

In this paper, a new framework for scene flow recovery from monocular image sequences with two preprocessing steps is proposed — the NRSfM-Flow. We intoduce a novel analytical framework which allows for relating NRSfM and MSF problems in the continuous domain. We believe that it provides additional insights into both problems and thus will facilitate development of next generation algorithms. We would like to draw attention to model based methods for MSF recovery and to emphasize the importance of differential interpretation of NRSfM.

NRSfM-Flow does not prescribe any particular NRSfM algorithm and inherits advantages and disadvantages of the NRSfM methods. The proposed framework may qualitatively outperform existing MSF methods in the ability to capture 3D motion fields of non-rigidly deforming scenes, since less restrictive assumptions about the scene and camera motion are made. For making this conclusion, we consider results of MSF recovery shown in literature so far and experimental results from this paper. One of the central concerns of future work lies in performing comprehensive comparative studies of existing MSF algorithms. As a next step, we plan on using the proposed theoretical apparatus to improve variational NRSfM and to formalize new challenges in the area of NRSfM. NRSfM-Flow will also be used for visualization purposes supporting development of augmented reality and medical applications.

Acknowledgment

The work was funded by the project DYNAMICS (01IW15003) of the German Federal Ministry of Education and Research (BMBF).

References

[1] *OpenCV Web-Page*. http://opencv.org/, 2016. [online; accessed on 11.05.2016].

[2] A. Agudo, L. Agapito, B. Calvo, and J. M. M. Montiel. Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1558–1565, 2014.

[3] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *Pattern Analysis and Machine Intelligence (TPAMI)*, 33(7):1442–1456, 2011.

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

[5] N. Birkbeck, D. Cobzaş, and M. Jägersand. Depth and scene flow from a single moving camera. In *3D Data Processing Visualization and Transmission (3DPVT)*, 2010.

[6] N. Birkbeck, D. Cobzaş, and M. Jägersand. Basis constrained 3d scene flow on a dynamic proxy. In *International Conference on Computer Vision (ICCV)*, pages 1967–1974, 2011.

[7] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Computer Vision and Pattern Recognition (CVPR)*, pages 690–696, 2000.

[8] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, pages 611–625, 2012.

[9] P. Dinning. *Barn Owl at Screech Owl Sanctuary*. https://www.youtube.com/watch?v=xmou8t-DHh0, 2014. [online; accessed on 12.05.2016; usage rights obtained from the author].

[10] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1272–1279, 2013.

[11] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *International Journal of Computer Vision (IJCV)*, 104(3): 286–314, 2013.

[12] P. F. U. Gotardo and A. M. Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3065–3072, 2011.

[13] P. F. U. Gotardo and A. M. Martinez. Kernel non-rigid structure from motion. In *International Conference on Computer Vision (ICCV)*, pages 802–809, 2011.

[14] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artifical Intelligence (An International Journal)*, 17:185–203, 1981.

[15] The MathWorks Inc. *MATLAB version 9.0 (R2016a)*. http://mathworks.com/products/matlab/, 2016.

[16] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, 1981.

[17] A. Mitiche, Y. Mathlouthi, and I. Ben Ayed. Monocular concurrent recovery of structure and motion scene flow. *Frontiers in ICT (FICT)*, 2(16), 2015.

[18] M. Paladini, A. Bartoli, and L. Agapito. Sequential non-rigid structure-from-motion with the 3d-implicit low-rank shape model. In *European Conference on Computer Vision (ECCV)*, pages 15–28, 2010.

[19] M. Paladini, A. Del Bue, J. Xavier, L. Agapito, M. Stošić, and M. Dodig. Optimal metric projections for deformable and articulated structure-from-motion. *International Journal of Computer Vision (IJCV)*, 96(2):252–276, 2011.

[20] V. Rabaud and S. Belongie. Re-thinking non-rigid structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[21] C. Rother, V. Kolmogorov, and A. Blake. Grabcut -interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH)*, 2004.

[22] A. Roussos, R. Garg, and L. Agapito. *Multi-Frame Subspace Flow (MFSF)*. http://www0.cs.ucl.ac.uk/staff/lagapito/subspace_flow/, 2015. [online; accessed on 12.02.2016].

[23] C. Russell, J. Fayad, and L. Agapito. Dense non-rigid structure from motion. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, pages 509–516, 2012.

[24] C. Russell, R. Yu, and L. Agapito. Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *European Conference on Computer Vision (ECCV)*, pages 583–598, 2014.

[25] D. Stoyanov. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 479–486, 2012.

[26] Carlo T. and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision (IJCV)*, 9:137–154, 1992.

[27] B. Taetz, G. Bleser, V. Golyanik, and D. Stricker. Occlusion-aware video registration for highly non-rigid objects. In *Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[28] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *Pattern Analysis and Machine Intelligence (TPAMI)*, 30(5):878–892, 2008.

[29] L. Valgaerts, A. Bruhn, H. Zimmer, J. Weickert, C. Stoll, and C. Theobalt. Joint estimation of motion, structure and geometry from stereo sequences. In *European Conference on Computer Vision (ECCV)*, pages 568–581, 2010.

[30] J. Valmadre and S. Lucey. General trajectory prior for non-rigid reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1394–1401, 2012.

[31] D. Xiao, Q. Yang, B. Yang, and W. Wei. Monocular scene flow estimation via variational method. *Multimedia Tools and Applications (An International Journal)*, pages 1–23, 2015.

[32] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *DAGM Conference on Pattern Recognition*, pages 214–223, 2007.

[33] Y. Zhu and S. Lucey. Convolutional sparse coding for trajectory reconstruction. In *Pattern Analysis and Machine Intelligence (TPAMI)*, 2014.