Introduction to Coherent Depth Fields for Dense Monocular Surface Recovery

Vladislav Golyanik[1,2]
vladislav.golyanik@dfki.de
Torben Fetzer[1]
torben.fetzer@dfki.de
Didier Stricker[1,2]
didier.stricker@dfki.de

[1] Department of Computer Science
University of Kaiserslautern

[2] Department Augmented Vision
German Research Center for Artificial Intelligence (DFKI)

Abstract

Handling large occlusions in non-rigid structure from motion (NRSfM) currently requires either an expensive correspondence correction or estimation of a shape prior on several non-occluded views. To save computational cost and remove the dependency on additional pre-processing steps, this paper introduces the concept of depth fields. With the proposed depth fields, NRSfM is interpreted as an alternating estimation of vector fields with fixed origins on the one side, and estimation of displacements of the origins along the depth dimension on the other. The core of the new energy-based Coherent Depth Fields (CDF) approach is the spatial smoothness coherency term (CT) applied on the depth fields. Having its origins in the Motion Coherence Theory, CT interprets data as a displacement vector field and penalises irregularities in displacements. Not only for handling occlusions but also for unoccluded scenes CT has multiple advantages compared to previously proposed regularisers such as total variation. We show experimentally that CDF achieves state-of-the-art in dense NRSfM including scenarios with long and large occlusions, inaccurate correspondences as well as inaccurate initialisations, without requiring any additional pre-processing steps.

1 Introduction

Dynamic dense 3D reconstruction, or 4D reconstruction, pursues the objective of capturing the geometry of dynamically evolving scenes. This actively researched problem has been studied for the multi-view [29, 39], RGB-D [10, 23] and monocular [1, 11, 32] based settings. The latter case — also known as monocular surface recovery — is particularly difficult due to the inherent ambiguities in input observations. Additional real-world challenges represent self- and external occlusions naturally arising while observing a scene as well as effects such as specularities or weakly textured areas.

There are several approaches to deal with large occlusions. If a template is available, i.e., an accurate reconstruction for at least a single view, occlusions can be handled efficiently [42]. If no template is available, a shape prior can be estimated on-the-fly from dense point correspondences obtained on several unoccluded views. The shape prior can then be used as a constraint for reconstruction of the occluded areas, both with and without available correspondences for the rest of the sequence [19]. Finally, correspondence based methods or non-rigid structure from motion (NRSfM)[1] can employ correspondence correction in the pre-processing step [36], although this approach works well for rather short-time disturbances.

Contributions. This paper introduces two novel concepts which allow to design a computationally efficient, accurate and easy to implement (practical) approach as well as overcome the dependency on the pre-processing steps in NRSfM while handling large occlusions. The first concept is the notion of a *depth vector field* or, concisely, a *depth field*. A depth field is a 2D parametrisation of a surface embedded into 3D space so that every tracked 2D point is associated with a displacement along the depth dimension. This definition implies that all displacements are parallel to each other, or, in other words, a *depth field is an irrotational vector field*. With the new parametrisation the problem of NRSfM is interpreted as a filtering of a depth field in the first alternating step (at that moment point origins are fixed), and shape refinement in the second alternation step. The updated shape, in turn, alters the depth field which is further filtered, and so on until convergence.

Next, we propose *coherency term* (CT) as a new soft spatial regulariser on the adjacent depth vectors. CT derives its origin from the motion coherence theory (MCT) [43, 44] which studies principles of coherent motion and perception. MCT, in accordance to the human visual system states that neighboring structures tend to move coherently, i.e., with a common velocity and direction. We call the proposed approach *Coherent Depth Fields* (CDF) and formulate it as an energy-minimisation problem with CT. The main reason lies in the expressiveness of energy-based methods — an energy functional explicitly encodes assumptions on the underlying physical processes and relates input data with the sought solution. We elaborate *efficient optimisation techniques* involving direct and inverse fast Fourier transforms (FT) and show experimentally that the proposed form of regularisation has several advantages (e.g., ability to filter depth values flexibly without edge oversmoothing) both when dealing with occluded and unoccluded scenes. CDF achieves state of the art accuracy on the joint evaluation benchmark with large occlusions [19], an actor benchmark [5] and several established data sets for qualitative evaluation (with and without occlusions).

2 Related Work

Given point correspondences throughout multiple views, the purpose of NRSfM is to recover the lost depth component of an observed non-rigidly deforming scene. This initial problem statement has motivated us to introduce the notion of depth fields. The proposed interpretation bears a resemblance to Helmholtz decomposition which allows decomposing an arbitrary 3D vector field into curl-free, divergence-free and lateral components.

The phenomenon of coherent motion was initially studied in visual perception — the Gestalt theory [22, 26]. The seminal works of Yuille and Crzywacz on MCT [43, 44] introduced the notions of *coherency* and a Gaussian *radial basis function* in computer vision. Since then, MCT was applied to many tasks such as camera pose and correspondence estimation [27], tracking [34], motion segmentation [41], visual search [24], and extensively influenced non-rigid point set registration [8, 25, 28, 33]. Thus, Coherent Point Drift [28] and offspring methods [3, 40] adopt CT for topology-preserving transformations.

Since sparse NRSfM was introduced by Bregler *et al.* [6], several dense approaches

[1] this class of methods exploits motion and deformations as reconstruction cues, hence the name.

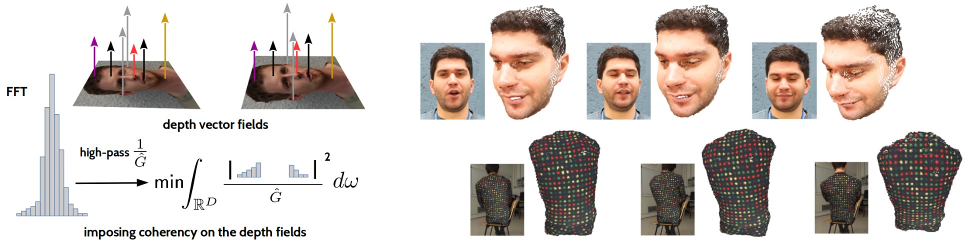
FFT
depth vector fields
high-pass $\frac{1}{\hat{G}}$
$\min \int_{\mathbb{R}^D} \frac{|\cdot|^2}{\hat{G}} d\omega$
imposing coherency on the depth fields

Figure 1: Coherency term penalises high-frequency component of a Fourier-transformed vector field, i.e., it favors field homogeneities. CDF reconstructs depth vector fields and uses coherency term as a regulariser. On the right side, our reconstructions of the *face* [14] and *back* [31] sequences are shown.

emerged [1, 32]. In energy-based formulations, total variation (TV) was shown as an efficient spatial regulariser [14]. TV allows for discontinuities at depth edges while being scale-unaware which is a favourable characteristic in the monocular setting. However, the resulting energy is non-convex and optimisation is performed with a computationally expensive iterative scheme. In contrast, the CDF energy is convex, the method requires fewer operations and is also well parallelisable. The accuracy of the methods [1, 14, 32] degrades considerably when correspondences are obtained on scenes with long and large occlusions. To overcome this limitation, Golyanik *et al.* proposed a hybrid approach with a shape prior obtained on-the-fly on several non-occluded frames [19]. Guided by an occlusions tensor, the shape prior is used as a depth regulariser in the occluded areas. The main limitation of hybrid NRSfM is the dependency on the accurate occlusion tensor and the shape prior. Taetz *et al.* proposed Bayesian inference framework to stabilise occluded point trajectories [36]. The proposed multi-frame optical flow approach works in two passes and allows to compensate for short-time disturbances. In contrast, *CDF does not require any pre-processing steps and can handle large and long occlusions with weaker assumptions and in less time*.

Remark. Occlusions constitute a common reason for missing entries in the measurement matrix and several NRSfM methods can explicitly account for missing data [21, 30, 38]. Since we consider the dense case and track points with dense optical flow techniques [15, 36], measurement matrices are always complete in our case. Nonetheless, due to occlusions, the accuracy of point correspondences degrades. CDF assumes a complete measurement matrix and perhaps inaccurate correspondences.

3 Coherency Term

Suppose $v = v(x)$ is a displacement function, i.e., for each element $x \in \Theta$ it outputs the corresponding displacement. MCT introduced a smoothing term for v which in a reproducing kernel Hilbert space can be written as a norm of a displacement field [16]:

$$\phi[v] = \int_{\mathbb{R}^D} \frac{|\hat{v}(\omega)|^2}{\hat{G}(\omega)} d\omega, \tag{1}$$

where $\hat{G}$ is a Fourier transformed reproducing kernel G, $\hat{v}$ is a Fourier transformed v and ω is a frequency variable. The right side of Eq. (1) — which we refer to as *coherency term* (CT) — applies two operators on $\hat{v}$. First, a high-pass filter $\frac{1}{\hat{G}}$ is applied. Second, L^2-norm[2]

[2] recall, $\|\psi\|_{L^2}^2 = \langle \psi, \psi \rangle_{L^2} = \int |\psi(\omega)|^2 d\omega$.

of the extracted component is taken. As a result, the norm $\phi[v]$ measures the total energy of the function at a high frequency. In other words, the less the function v oscillates, the smaller is the scalar $\phi[v]$ or, likewise, the more *coherent* are the displacements. Note that CT regularises an extracted high-frequency component of the depth field and approaches low-pass filtering in functionality. Fig. 1 visualises CT as applied to a depth field, i.e., a vector field arising in the proposed CDF (see Sec. 4).

4 Coherent Depth Fields

Given coordinates of N points tracked throughout an image stream, CDF aims at recovery of 3D surface geometry of the observed scene $\mathbf{S}(t)$ and camera poses $\mathbf{R}(t)$. Suppose coordinates of the tracked points over F frames are stacked together row-wise in a measurement matrix $\mathbf{W}_{2F\times N} = \left[\begin{bmatrix} u_1^t & v_1^t & \dots \end{bmatrix}^\mathsf{T} \quad \begin{bmatrix} u_2^t & v_2^t & \dots \end{bmatrix}^\mathsf{T} \quad \dots \right]$, with $t \in \{1, \dots, F\}$. Note that further on, the discrete and continuous notations $\mathbf{R} = \mathrm{D}\,\mathbf{R}(t)$, $\mathbf{S} = \mathrm{D}\,\mathbf{S}(t)$ and $\mathbf{W} = \mathrm{D}\,\mathbf{W}(t)$ are used interchangeably; D denotes a discretisation operator. Without loss of generality, we use an orthographic camera model and assume that the translation in the scene is resolved. The observations $\mathbf{W}(t)$ are caused by imaging of the deformable geometry $\mathbf{S}(t) = [\mathbf{S}_1\mathbf{S}_2 \dots \mathbf{S}_F]^\mathsf{T}$ by orthographic camera $\mathbf{R}(t)$:

$$\mathbf{W}(t) = \mathbf{R}(t)\,\mathbf{S}(t) = \begin{pmatrix} \mathbf{R}_1 & & \\ & \mathbf{R}_2 & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \end{pmatrix}. \tag{2}$$

Note that we impose orthonormality on $\mathbf{R}_t$ matrices. NRSfM parameterised by a depth field is given by the energy functional

$$\mathbf{E}(\mathbf{R},\mathbf{S}) = \frac{1}{2}\|\mathbf{W}(t) - \mathbf{R}(t)\mathbf{S}(t)\|_{\mathcal{F}}^2 + \frac{\lambda}{2}\int_{\mathbb{R}^2} \frac{|\hat{\mathbf{S}}(s)|^2}{\hat{G}(s)}ds \tag{3}$$

$$\text{s.t. } \operatorname{rank}(\mathrm{P}(\mathbf{S})) = \tau, \tag{4}$$

where $\mathcal{F}$ stays for Frobenius or Schatten 2-norm, $\hat{\mathbf{S}}(s)$ denotes the Fourier transformed shape, $\hat{G}$ is a Fourier transformed reproducing kernel, s is the frequency domain variable, and operator $\mathrm{P}(\cdot)$ entangles rows of every submatrix $\mathbf{S}_i$ through reordering them in a single row. The right side of Eq. (3) contains the data term defined as elementwise 2-norm of the reprojections $\mathbf{R}(t)\,\mathbf{S}(t)$ and CT as a spatial regulariser. The rank-constraint expresses the assumption about the complexity of the deformations and steadily insures that at most τ shapes are linearly independent. The particular form of CT requires a detailed explanation.

Recall that 1) the number and ordering of the points in every frame are equal and 2) coordinates of all tracked points can be backprojected to the reference frame. In the ideal case, where there are no tracking inaccuracies in the pre-processing step, the objective of NRSfM — to reconstruct complete 3D coordinates of every point — can be simplified to the recovery of missing depths z only. Geometrically, this means that a static set of points induces a time-varying depth vector field $\mathfrak{X} = \mathfrak{X}(u,v,t)$ *with fixed origins*. We call such a depth vector field regularised by CT *coherent depth field*. Accordingly, we name the new algorithm CDF which emphasises the interpretation of data and the smoothness term. The concept of coherent depth field is visualised in Fig. 1. In other words, every $\mathbf{S}(t)$ can be comprehended as a vector field $\mathbf{S}\colon \mathbb{R}^2 \to \mathbb{R}$. Thus, the term $\frac{\lambda}{2}\int_{\mathbb{R}^2}\frac{|\hat{\mathbf{S}}(s)|^2}{\hat{G}(s)}ds$ imposes coherency on the neighboring elements of $\mathfrak{X}$ or *depths* (for the basic interpretation of CT cf. Sec. 3). It it worth nothing that in the real case, the origins of the depth fields may drift due to tracking inaccuracies — the same principles apply.

We minimise the energy in Eq. (3) by alternately fixing $\mathbf{S}$ and $\mathbf{R}$ and minimizing for $\mathbf{R}$ and $\mathbf{S}$ respectively. While $\mathbf{S}$ is fixed, only the data term depends on $\mathbf{R}$. This subproblem can be efficiently solved in a closed form by projecting an affine update of the rotation matrix into the SO(3) group by normal equations, i.e, $\mathbf{R} = \mathbf{W}\mathbf{S}^{\mathsf{T}}(\mathbf{S}\mathbf{S}^{\mathsf{T}})^{-1}$ which is of comparable accuracy and faster than more computationally expensive non-linear optimisation schemes.

While $\mathbf{R}$ is fixed, the problem in Eq. (3) is convex in $\mathbf{S}$[3]. Nevertheless, as different norms are used in the data and smoothness terms, it can not be easily solved in a standard way (e.g., by directly applying Euler-Lagrange differential equation). The problem is remedied by proximal splitting — through introduction of an auxiliary variable $\bar{\mathbf{S}}$ we split the problem in two subproblems. The original problem is thus equivalent to

$$\underset{\mathbf{S}}{\operatorname{argmin}}\ \frac{1}{2\theta}\|\mathbf{S}-\bar{\mathbf{S}}\|_{\mathcal{F}}^2 + \frac{\lambda}{2}\int_{\mathbb{R}^2}\frac{|\hat{\mathbf{S}}(s)|^2}{\hat{G}(s)}ds, \tag{5}$$

$$\underset{\bar{\mathbf{S}}}{\operatorname{argmin}}\ \frac{1}{2\theta}\|\mathbf{S}-\bar{\mathbf{S}}\|_{\mathcal{F}}^2 + \frac{1}{2}\|\mathbf{W}-\mathbf{R}\bar{\mathbf{S}}\|_{\mathcal{F}}^2 \tag{6}$$

$$\text{s.t. } \operatorname{rank}(P(\bar{\mathbf{S}})) = \tau$$

and solved through alternating optimisations of the functionals (5) and (6). (5) updates (filters) the depth field, and the x and y coordinates are fixed in this step. Given the updated depth field, (6) revises the complete shapes. We reformulate the functional (5) as

$$\frac{1}{2\theta}\int_{\Omega}|\mathbf{S}(x)-\bar{\mathbf{S}}(x)|^2dx + \frac{\lambda}{2}\int_{\Omega}\frac{|\hat{\mathbf{S}}(s)|^2}{\hat{G}(s)}ds, \tag{7}$$

where Ω is the set of points considered for reconstruction (fixed depth field origins) and $x \in \Omega$. Next, FT of $\mathbf{S}$ is performed leading to

$$\frac{1}{2\theta}\int_{\Omega}\left|\int_{\mathbb{R}^2}\hat{\mathbf{S}}(s)e^{2\pi i\langle x,s\rangle}ds - \bar{\mathbf{S}}(x)\right|^2dx + \frac{\lambda}{2}\int_{\Omega}\frac{|\hat{\mathbf{S}}(s)|^2}{\hat{G}(s)}ds\ . \tag{8}$$

The energy in Eq. (8) is optimised w.r.t. $\hat{\mathbf{S}}$ whilst $\bar{\mathbf{S}}$ is fixed. To find the minimum, we take the partial derivative of $\mathbf{E}(\hat{\mathbf{S}})$ w.r.t. $\hat{\mathbf{S}}(t)$ and equate it to zero:

$$\frac{\partial E(\hat{\mathbf{S}})}{\partial\hat{\mathbf{S}}(t)} = \frac{1}{\theta}(\mathbf{S}(t)-\bar{\mathbf{S}}(t))\int_{\Omega}\frac{\partial\hat{\mathbf{S}}(s)}{\partial\hat{\mathbf{S}}(t)}e^{2\pi i\langle x,s\rangle}ds + \frac{\lambda}{2}\int_{\Omega}\frac{\partial\hat{\mathbf{S}}(s)^2}{\partial\hat{\mathbf{S}}(t)}\frac{1}{\hat{G}(s)}ds \overset{!}{=} 0 \tag{9}$$

$$\implies \frac{1}{\theta}(\mathbf{S}(t)-\bar{\mathbf{S}}(t))e^{2\pi i\langle x,t\rangle} + \lambda\frac{\hat{\mathbf{S}}(-t)}{\hat{G}(t)} \overset{!}{=} 0. \tag{10}$$

Note that we introduce a new Fourier-space variable t to express a different integration area in contrast to those associated with the variable s. After inverse FT is applied, the multiplication alters to a convolution:

$$\mathbf{S}(t) = \frac{1}{\lambda\theta}G(t)*(\bar{\mathbf{S}}-\mathbf{S})(t). \tag{11}$$

[3] the low-rank constraint in Eq. (4) — which makes the whole minimisation objective non-convex when considered jointly — is imposed in a separate step after minimisation of Eq. (3).

This convolution equation is subsequently solved w.r.t $\mathbf{S}$, and the solution is given by

$$\mathbf{S} = \mathcal{F}^{-1}\left(\mathcal{F}(\bar{\mathbf{S}}) \circ \frac{\mathcal{F}(G)}{\lambda\theta 1_{m\times n} + \mathcal{F}(G)}\right), \tag{12}$$

where $1_{m\times n}$ is an all-ones matrix and $\circ$ is elementwise multiplication. The quotient on the right side of Eq. (12) is the resulting depth field filter (approaching the low-pass).

Next, we minimise the energy in Eq. (6) w.r.t $\bar{\mathbf{S}}$ whilst $\mathbf{S}$ is fixed. Therefore, we find the gradient w.r.t. $\bar{\mathbf{S}}$:

$$\nabla_{\bar{\mathbf{S}}} = (\frac{1}{\theta} + \mathbf{R}^\mathsf{T}\mathbf{R})\bar{\mathbf{S}} - (\mathbf{R}^\mathsf{T}\mathbf{W} + \frac{1}{\theta}\mathbf{S}). \tag{13}$$

The minimiser is obtained by demanding $\nabla_{\bar{\mathbf{S}}} \overset{!}{=} 0$ as

$$\bar{\mathbf{S}}' = (\frac{1}{\theta} + \mathbf{R}^\mathsf{T}\mathbf{R})^{-1}(\mathbf{R}^\mathsf{T}\mathbf{W} + \frac{1}{\theta}\mathbf{S}). \tag{14}$$

To fulfil the rank constraint, the suboptimal $\bar{\mathbf{S}}'$ obtained by Eq. (14) is projected onto the subspace of τ-rank matrices using svd. Suppose

$$U\Sigma V^\mathsf{T} = \mathrm{svd}\left(\mathrm{P}\left((\frac{1}{\theta} + \mathbf{R}^\mathsf{T}\mathbf{R})^{-1}(\mathbf{R}^\mathsf{T}\mathbf{W} + \frac{1}{\theta}\mathbf{S})\right)\right). \tag{15}$$

The solution to the problem in Eq. (6) reads

$$\bar{\mathbf{S}} = \mathrm{P}^{-1}\left(U\Sigma_\tau V^\mathsf{T}\right), \tag{16}$$

where Σ_τ is the truncated diagonal matrix Σ with τ largest elements (singular values) preserved and zeroes otherwise. Once $\bar{\mathbf{S}}$ is recovered, $\mathbf{S}$ can be updated according to Eq. (12).

CDF expects $\mathbf{W}$ and four parameters (λ, θ, τ and σ — the variance of the Gaussian kernel) as an input. The entire algorithm is summarised in Alg. 1. An expensive part is $\mathbf{S}$ computation of $\mathcal{O}(FN\log N)$ complexity in Eq. (12) — it requires an FT, an inverse FT and an element-wise multiplication. Fortunately, it can be accomplished efficiently on a GPU. Otherwise, $\mathbf{R}^\mathsf{T}\mathbf{W}$ is fully parallelisable and svd is performed on 3×3 matrices twice per alternation. We initialise $\mathbf{S}$ under rigidity with the Tomasi-Kanade approach [37]. Convergence criteria for the inner and outer loops are defined as $\left\|\bar{\mathbf{S}} - \mathbf{S}\right\|_\mathcal{F} < \varepsilon$ and $\mathbf{E}(\mathbf{R},\mathbf{S})^i - \mathbf{E}(\mathbf{R},\mathbf{S})^{i+1} < \xi$ respectively; ε and ξ are scalars.

CDF is implemented in C++; external dependencies include `fttw3` library for fast FT and inverse fast FT [12] as well as `eigen3` for operations on matrices [13]. The test platform is composed of 32 GB RAM and Intel i7-6700K CPU running at 4 GHz. The next Sec. describes evaluation of the proposed approach.

Table 1: RMSE of VA [14], AMP [17] and the proposed CDF for the actor data set [5].

method	conf. 1	conf. 2	optimal parameters
VA [14]	0.36762	0.33624	$\lambda = 5\cdot 10^3, \theta = 10^{-4}$
AMP [17]	1.5058	1.509	$K = 7$
CDF (ours)	**0.20188**	**0.19638**	$\sigma = 4.4$, $\lambda = 0.4, \theta = 10^{-2}$, $\tau = 20$

Algorithm 1 Coherent Depth Fields

```
Input: measurements $\mathbf{W}$, parameters $\lambda$, $\theta$, $\tau$, $\sigma$
Output: time varying non-rigid shapes $\mathbf{S}$
Initialisation: $\mathbf{S} = \left[\mathbf{S}_r \mathbf{S}_r \dots \mathbf{S}_r\right]^\top$, where $\mathbf{S}_r$ is factorisation under rigidity assumption [37]
while not converge do
    step 1: fix S, update R
    $\text{svd}(\mathbf{W}\mathbf{S}(\mathbf{S}\mathbf{S}^\top)^{-1}) = U\Sigma V^\top$
    $\mathbf{R} = UCV^\top$, where $C = \text{diag}(1, 1, \dots, 1, \text{sign}(\det(UV^\top)))$
    step 2: fix R, update S; initialise $\bar{\mathbf{S}} = \mathbf{S}$
    while not converge do
        $U\Sigma V^\top = \text{svd}\left(\mathrm{P}\left((\frac{1}{\theta}\mathbf{I} + \mathbf{R}^\top\mathbf{R})^{-1}(\frac{1}{\theta}\mathbf{S} + \mathbf{R}^\top\mathbf{W})\right)\right)$
        $\bar{\mathbf{S}} = \mathrm{P}^{-1}\left(U\Sigma_{trunc}V^\top\right)$
        $\mathbf{S} = \mathcal{F}^{-1}\left(\mathcal{F}(\bar{\mathbf{S}}) \circ \frac{\mathcal{F}(G)}{\lambda\theta 1_{m\times n} + \mathcal{F}(G)}\right)$    // $\circ$ denotes elementwise multiplication
    end while
end while
```

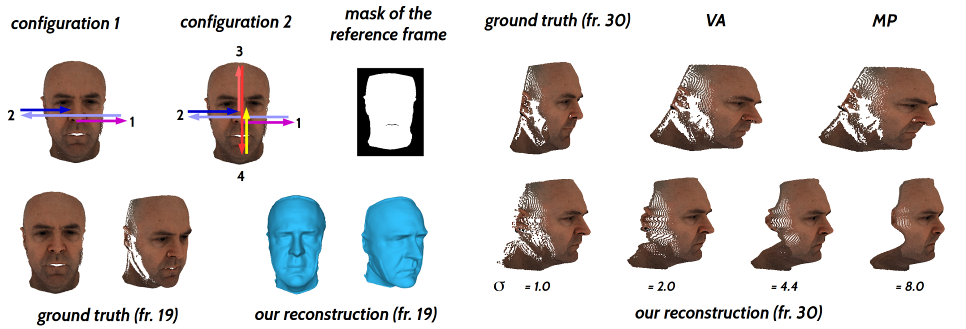

Figure 2: For the 3D actor mocap data set [8], we created ground truth measurements through imaging by a virtual orthographic camera following two trajectories (upper left). Shown are exemplary reconstructions by VA [14], AMP [17] and CDF (our approach) on the new sequence. For CDF, we show surface evolution depending on σ (variance of the Gaussian).

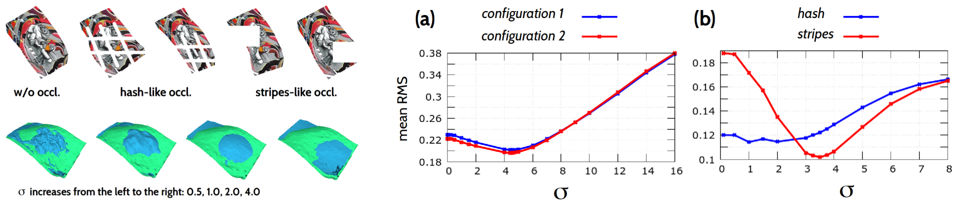

Figure 3: (Left): *Flag mocap* sequences with large external occlusions [19]; evolution of reconstructed occluded regions for different σ (reference is shown in green, reconstructions are shown in cyan). (Right): mean RMSE as a function of σ on the (a): *actor mocap*, for two camera trajectories and (b): *flag mocap* with large occlusions (we evaluate how close reconstructions obtained under occlusions approach reconstructions on the accurate tracks).

5 Experiments

For data sets with ground truth (*flag mocap* [15], *actor mocap* [5], *synthetic face* [14]) we compare per sequence average normalised root-mean square error (RMSE) — defined as $e_{3D} = \frac{1}{F}\sum_{f=1}^{F}\frac{\left\|\mathbf{S}_f^{ref}-\mathbf{S}_f\right\|_{\mathcal{F}}}{\left\|\mathbf{S}_f^{ref}\right\|_{\mathcal{F}}}$ ($\mathbf{S}_f^{ref}$ are ground truth surfaces) — of several approaches supporting dense setting, i.e., VA [14], Metric Projections (MP) [30] or Accelerated Metric Projections (AMP) [17] and the proposed CDF. In doing so, every reconstructed shape is registered to the ground truth with Procrustes analysis.

Joint evaluation. To test how accurate CDF performs on inaccurate correspondences, we jointly evaluate dense point tracking and NRSfM. We use the *flag mocap* data set with ground truth surfaces, correspondences and rendered images with added *hash* and *stripes* large occlusions patterns (see Fig. 3-(left)) [15, 19]. For both cases, several combinations of multi-frame optical flow (MFOF) (either occlusion-aware MFOF or multi-frame subspace flow (MFSF)) and NRSfM methods are tested. The results are summarised in Table 2. Reconstructions by AMP and VA on noisy correspondences (columns two and three) exhibit strong depth variations (see supplementary material). In contrast, CDF compensates for tracking inaccuracies while not jeopardising the unoccluded parts. Remarkably that RMSE of the combination MFSF [14] + CDF — without additional pre-processing steps — is comparable to RMSE of the computationally expensive MFOF with point trajectory correction [36] + VA [14]. MFOF requires twice to triple the runtime of MFSF, as it improves point trajectories in a separate pass. Next, the waving flag is reconstructed with CDF both on the ground truth and inaccurate point tracks. Using the reconstructions on the ground truth correspondences as a reference, we measured the relative RMSE for multiple σ values. This test reveals how close reconstructions on inaccurate correspondences due to occlusions are approaching the structure obtained on unoccluded data. Results are plotted in Fig. 3-(b).

Table 2: Average RMSE on the occluded flag sequences [19].

	MFSF [15] + AMP [30]	MFSF [15] + VA [14]	MFOF [36] + VA [14]	MFSF [15] + CDF
hash	0.297 (0.381)	0.239 (0.252)	**0.181 (0.219)**	**0.188 (0.212)**
stripes	0.460 (0.523)	0.341 (0.355)	**0.195 (0.209)**	**0.211 (0.216)**

Likewise, we evaluate CDF in a scenario with inaccurate initialisations. Therefore, we take the 4D *actor* motion capture data set of Beeler *et al.* [5] and generate measurements by projecting individual 3D shapes by a virtual orthographic camera. Two different camera trajectories are choosen, see Fig. 2-(upper left). In the first setting, the camera observes the face frontally and then rotates to the right and left eventually returning to the initial position; in the second setting, the camera follows the right-left-up-down pattern. The movements are more rapid and the amplitude is smaller compared to the first setting (max. 30°). Both sequences contain 51 frames with $3.7 \cdot 10^4$ points each. Facial expressions of the actor are realistic and moderate (there are no exaggerated expressions as a strong cue) and the data set is particularly challenging for monocular reconstruction. Table. 1 summarises the obtained RMSE and Fig. 2 contains some exemplary reconstructions. AMP achieves the RMSE of 1.506. VA improves the error by the factor of four and reconstructs more fine details. The test shows that both methods can only lightly recover from the inaccuracy in the initial depth estimation. CDF, starting from the same initialisation through rigid factorisation [37] as VA, achieves the lowest RMSE of 0.202, as it is capable to regularise depth fields. In the second camera setting, all algorithms lessen RMSE consistently over all tested parameter configurations so that the overall placement remains the same. The reason is a richer rotation cue in the scene. Overall, we tested multiple σ values and identified that CDF exhibits well-posedness w.r.t. the parameter choice, unless set too high, see Fig. 3-(a). The RMSE percent variance for a suboptimal σ does not exceed 20% in the range $[10^{-3}; 8.0]$ (λ and τ were fixed to 0.4 and 20 respectively in the course of all experiments).

Additionally, we evaluated CDF on four dense synthetic benchmark face sequences [14]. Several methods [4, 11, 14, 20, 30] were compared on this data set before [11]. CDF achieves RMSE of 8.03% (an average RMSE for all four data sets). For the sequences 1 and 2 with 10 frames each, CDF achieves 7.54% and 6.64% respectively and RMSE increases for sequences 3 and 4 (99 frames each) to 8.87% and 9.04% respectively. Qualitatively, our reconstructions exhibit fewer surface fluctuations compared to [4] and [30]. Both VA and CDF rely on accurate initialisations. CDF's robustness against occlusions comes at cost of more flattened depth values in the case without occlusions, compared to VA. Table 3 provides a compact comparison of the four methods with average RMSE and standard deviation s for all synthetic face sequences jointly[4].

Table 3: Average RMSE and s on the synthethic faces [14] over four sequences jointly.

	TB [4]	**MP [30]**	**VA [14]**	**CDF (ours)**
RMSE / s	9.24 / 5.37	8.81 / 6.15	3.22 / 0.55	8.03 / 0.98

Next, we show qualitative results on several challenging real-world image sequences — *face* (120 frames) [14], *back* (150 frames) [31], *heart* bypass surgery scene occluded by a robotic arm (40 frames) [35] and abdominal *laparoscopic* sequence (120 frames) [2]. Fig. 1-(right) shows selected reconstructions of the *face* and *back* sequences. The person's face is reconstructed up to the fine details such as closing and opening of the eyes. The back exhibits stronger large-scale non-rigid deformations and rotations which are likewise plausibly captured. All in all, both reconstructions appear highly realistic.

Laparoscopic sequence depicts palpation of the abdominal cavity of a rabbit. During palpation the soft tissues are deformed by human fingers and deformations are observed by a compact camera inside the body. Fig. 4-(left) shows two selected reconstructions. The moment when almost no pressure on the soft tissues is exerted is shown in Fig. 4-(a). In Fig. 4-(b), deformations of the tissues are the strongest. CDF successfully reconstructs the scene and provides the means for detailed deformation analysis. After meshing and shading of the resultant point clouds, deformations can be visually identified and analysed in a virtual fly-by along the captured dynamic surfaces (see the supplementary video).

Selected results on the *heart* sequence are given in Fig. 4-(right). In the course of the surgery, a robotic arm enters the scene and occludes up to 50% of the region of interest. Due to inaccurate initialisations, AMP outputs unlikely depth tensions of the structure whereas VA fails to reconstruct recognisable surfaces. Our approach obtains plausible structures, although the geometry differs from those obtained on the *heart* sequence without occlusions[5]. It is noteworthy that 1) unoccluded parts are reconstructed more accurately as one would intuitively expect from CDF and 2) strong occlusions do not ruin surface recovery.

Finally, we reconstruct several other real image sequences (*new face* (120 frames) [19], *shaman2* (50 frames) [7], *owl* (202 frames) [9] and *notes* (139 frames) [18]). Exemplary shaded reconstructions are assembled into Fig. 5. The runtime of CDF depends on multiple

[4] the average RMSE for TB [4], MP [30] and VA [14] are taken from [14].

[5] the heart sequence *without occlusions* was reconstructed in multiple previous works [11, 14, 19, 36].

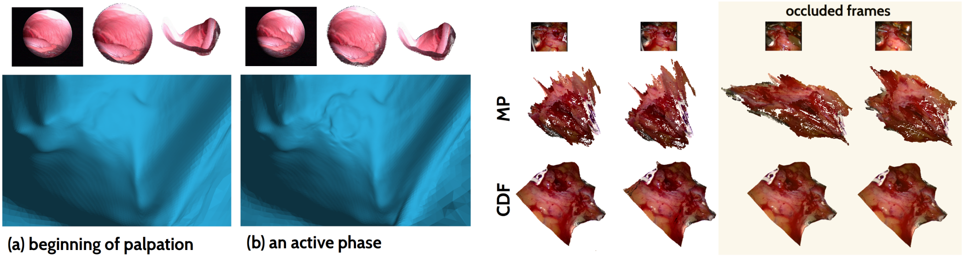

Figure 4: (Left): CDF reconstruction of the challenging laparoscopic sequence [2], (a): the palpation of the abdominal area begins, (b): an active phase of the palpation. In both sides, shown are the input image, a corresponding half-spherical reconstruction (frontal and side views) as well as inspection of shaded and zoomed in ROIs. (Right): results on the heart bypass surgery sequence with a robotic arm entering the scene. Our method can reconstruct the heart, AMP [17] is largely affected by the tracking inaccuracies, and VA [14] fails.

factors (number of frames and number points in a sequence, σ, *etc*.). For the *back* sequence it amounts to 1322 seconds (8 alternations, 10 primal-dual iterations and $\sigma = 3.5$).

6 Conclusion

We introduced the CDF algorithm for dense NRSfM based on two central novel concepts, i.e., a depth field and a new, in the context of NRSfM, spatial smoothness term — the coherency term. CDF proves itself as an accurate and robust to occlusions algorithm which can efficiently utilise available cues in a scene. On the challenging actor motion capture sequences with small deformations, we obtain the lowest RMSE among several approaches. We believe that CDF is the first method which can compensate for severe occlusions (dozens of frames with 50% or more of a scene eclipsed) without an explicit correspondence correction in the preprocessing step, a learned deformation model or a shape prior. Though CDF may oversmooth fine structures if variance of the Gaussian is chosen suboptimally, we determined that the approach is well-posed w.r.t. parameter choice. We showed that CDF can be applied in a variety of real-world scenarios such as heart surgery, minimally invasive diagnostics as well as high-quality facial motion capture. Furthermore, CDF is easy to implement, has a high portion of parallelisable code and is suitable for implementation on embedded hardware. Regarding the coherency term, we believe that the set of tasks which could take advantage of it is not exhausted by NRSfM.

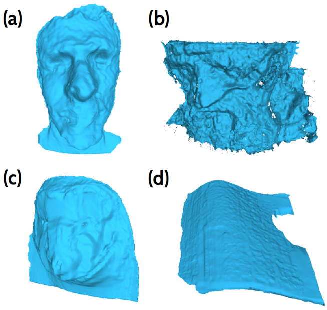

Figure 5: Examples of shaded surfaces reconstructed by CDF: (a) *new face*, (b) *shaman2*, (c) *owl*, (d) *notes*.

Acknowledgments

This work was supported by the project DYNAMICS (01IW15003) of the German Federal Ministry of Education and Research (BMBF). The authors thank Antonio Agudo for providing the laparoscopic sequence.

References

[1] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo. Online dense non-rigid 3d shape and camera motion recovery. In *British Machine Vision Conference (BMVC)*, 2014.

[2] A. Agudo, F. Moreno-Noguer, B. Calvo, and J. M. M. Montiel. Sequential non-rigid structure from motion using physical priors. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[3] A. Aidibe and A. Tahan. Adapting the coherent point drift algorithm to the fixtureless dimensional inspection of compliant parts. *The International Journal of Advanced Manufacturing Technology*, 79(5):831–841, 2015.

[4] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. *Neural Information Processing Systems (NIPS)*, 2008.

[5] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross. High-quality passive facial performance capture using anchor frames. *ACM Transactions on Graphics (TOG)*, 30(4):75, 2011.

[6] 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.

[7] 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.

[8] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2-3):114–141, 2003.

[9] P. Dinning. *Barn Owl at Screech Owl Sanctuary*. https://www.youtube.com/watch?v=xmou8t-DHh0, 2014. [online; accessed on 25.04.2017].

[10] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):114:1–114:13, 2016.

[11] K. Fragkiadaki, M. Salas, P. Arbelaez, and J. Malik. Grouping-based low-rank trajectory completion and 3d reconstruction. In *Neural Information Processing Systems (NIPS)*, 2014.

[12] M. Frigo and S. G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE. Special issue on "Program Generation, Optimization, and Platform Adaptation"*, 93(2):216–231, 2005.

[13] G. Gaël, J. Benoît, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[14] 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.

[15] 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.

[16] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995.

[17] V. Golyanik and D. Stricker. Dense batch non-rigid structure from motion in a second. In *Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[18] V. Golyanik, A. S. Mathur, and D. Stricker. NRSfM-Flow: Recovering non-rigid scene flow from monocular image sequences. In *British Machine Vision Conference (BMVC)*, 2016.

[19] V. Golyanik, T. Fetzer, and D. Stricker. Accurate 3d reconstruction of dynamic scenes from monocular image sequences with severe occlusions. In *Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[20] P. F. U. Gotardo and A. M. Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *TPAMI*, 33(10):2051–2065, 2011.

[21] 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.

[22] G. Humphrey. The psychology of the gestalt. *Journal of Educational Psychology*, 15 (7):401–412, 1924.

[23] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. VolumeDeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 362–379, 2016.

[24] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10–12):1489 – 1506, 2000.

[25] B. Jian and B. C. Vemuri. Robust point set registration using gaussian mixture models. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(8):1633–1645, 2011.

[26] K. Koffka. *Principles of Gestalt psychology*. Harcourt, 1935.

[27] W. Lin, L. Cheong, P. Tan, G. Dong, and S. Liu. Simultaneous camera pose and correspondence estimation with motion coherence. *International Journal of Computer Vision*, 96(2):145–161, 2012.

[28] A. Myronenko and X. Song. Point-set registration: Coherent point drift. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010.

[29] M. R. Oswald and D. Cremers. A convex relaxation approach to space time multi-view 3d reconstruction. In *International Conference on Computer Vision (ICCV) Workshops*, 2013.

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

[31] C. Russell, J. Fayad, and L. Agapito. Energy based multiple model fitting for non-rigid structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3009–3016, 2011.

[32] 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.

[33] G. Sanroma, R. Alquézar, F. Serratosa, and B. Herrera. Smooth point-set registration using neighboring constraints. *Pattern Recognition Letters*, 33(15):2029 – 2037, 2012.

[34] X. Song, A. Myronenko, and D. J. Sahn. Speckle tracking in 3d echocardiography with motion coherence. In *Computer Vision and Pattern Recognition (CVPR)*, 2007.

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

[36] 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.

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

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

[39] T. Tung, S. Nobuhara, and T. Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *International Conference on Computer Vision (ICCV)*, pages 1709–1716, 2009.

[40] P. Wang, P. Wang, Z. Qu, Y. Gao, and Z. Shen. A refined coherent point drift (cpd) algorithm for point set registration. *Science China Information Sciences*, 54(12):2639–2646, 2011.

[41] Y. Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 1997.

[42] R. Yu, C. Russell, N.D.F. Campbell, and L. Agapito. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *International Conference for Computer Vision (ICCV)*, 2015.

[43] A. L. Yuille and N. M. Grzywacz. A mathematical analysis of the motion coherence theory. *International Journal of Computer Vision*, 3(2):155–175, 1989.

[44] A.L. Yuille and N.M. Grzywacz. The motion coherence theory. In *International Conference on Computer Visons (ICCV)*, 1988.