Accurate 3D Reconstruction of Dynamic Scenes from Monocular Image Sequences with Severe Occlusions
Supplementary Material

Vladislav Golyanik Torben Fetzer Didier Stricker
Department of Computer Science, University of Kaiserslautern
Department Augmented Vision, DFKI Kaiserslautern
{Vladislav.Golyanik, Torben.Fetzer, Didier.Stricker}@dfki.de

This supplementary material provides insights into SPVA going beyond the scope of the main matter. In Secs. A and B, more details on occlusion tensor estimation and TI criterion are provided. Sec. C contains a more detailed description of the experimental results on the *human face*, *heart surgery*, and *ASL* sequences. We show that on the ASL sequence, the correspondence correction [6] in combination with the base scheme [2] fails, and analyse reasons for failure. Moreover, algorithm parameters and runtimes for every sequence are listed.

A. Obtaining occlusion tensor

Consider dense flow fields computed by an MFSF, or displacements of pixels visible in the reference frame throughout the whole image sequence:

$$\mathbf{u}(x;\mathbf{n}) = \begin{bmatrix} u(x,\mathbf{n}) \\ v(x,\mathbf{n}) \end{bmatrix} : \Omega \times \{1,...,F\} \rightarrow \mathbb{R}^2, \quad (1)$$

where $\Omega \in \mathbb{R}^2$, $\mathbf{n}$ denotes a frame index, $\mathbf{u}(x,\cdot)$ denotes a 2D displacement of a point x through the image sequence ($u(x,\mathbf{n})$ and $v(x,\mathbf{n})$ denote displacements in u and v directions respectively). Let $\mathbf{r}$ be the index of the reference frame and $\mathbf{I}(x,\mathbf{r})$ a reference frame. Occlusion maps $\mathbf{E}(x,\mathbf{n}) : \Omega \times \{2,...,F\} \rightarrow \mathbb{R}$ can be obtained from the dense correspondences and the reference frame according to the algorithm summarised in Alg. 1. Firstly, a backprojection $\mathbf{B}(\mathbf{n},\mathbf{r})$ of every frame $\mathbf{I}(x,\mathbf{n})$ to the reference frame is performed. In this step, reverse point displacements are applied to every frame $\in \{2,\dots,F\}$ with an optional interpolation for missing parts. Secondly, image differences between the warped images and the reference image are computed. Therefore, we take L_2-norms of sums of channel-wise differences (RGB) for every pixel convolved with the Gaussian kernel $G_{k\times k}$ of an odd width k (see Alg. 1, rows 3–7). The result is postprocessed (normalised and discretised) so that the estimated values lie in the interval $[0;255]$.

Algorithm 1 Estimation of an occlusion tensor $\mathbf{E}(x)$

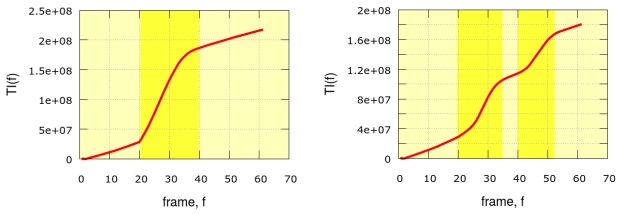
Input: dense flow fields $\mathbf{u}(x;\mathbf{n})$, a reference frame $\mathbf{I}(x,\mathbf{r})$, Gaussian kernel $G_{k\times k}$
Output: occlusion maps $\mathbf{E}(x,\mathbf{n})$
1: **for every frame** $\mathbf{n} \in \{2,\dots,F\}$ **do**
2: $\mathbf{w}(\mathbf{n},\mathbf{r}) = \mathbf{I}(x,\mathbf{n}) - \mathbf{u}(x;\mathbf{n})$ (backprojection to $\mathbf{I}(x,\mathbf{r})$)
3: image difference $\mathbf{B}(x) = \mathbf{w}(\mathbf{n},r) - \mathbf{I}(x,\mathbf{r}) \hat{=}$
4: **for every pixel** x **do**
5: $\mathbf{B}(x) = \left\| (x_r^{\mathbf{w}} - x_r^{\mathbf{I}})^2 + (x_g^{\mathbf{w}} - x_g^{\mathbf{I}})^2 + (x_b^{\mathbf{w}} - x_b^{\mathbf{I}})^2 \right\|_2$
6: **end for**
7: $\mathbf{E}(x,\mathbf{n}) = \mathbf{B}(x) * G$
8: postprocess $\mathbf{E}(x)$
9: **end for**

The resulting image series represents occlusion tensor with per-pixel occlusion probabilities for every frame. If required, occlusion maps can be binarised. The entire algorithm exhibits data parallelism (both on the frame and pixel levels) and is well suitable for implementation on a GPU.

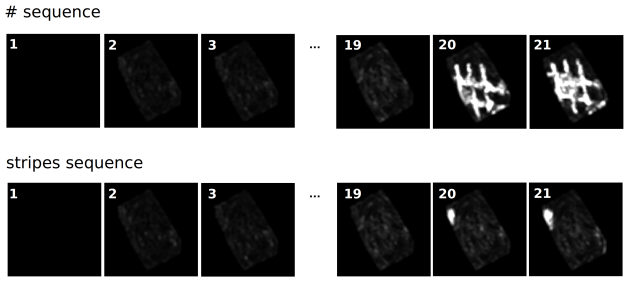
B. Total intensity criterion

In Fig. 1, results of the algorithm operating based on the total intensity (TI) criterion are given. TI accumulates non-zero pixel values for all frames until the frame f [1]. As can be seen in the comparison of the images and corresponding TI plots, sudden increases in TI happen if an occlusion begins. Low occlusion probabilities cause gradual increases of TI. When comparing the frame indexes with the starting occlusions and the responses of the TI indicators, the correlation is clearly seen.
ϵ, i.e., the sensitivity threshold, depends on the size and/or duration of an occlusion. Suppose occlusions do not happen, and the sequence is infinitely long. In this case, the threshold ϵ will still be reached after a finite number of

[1] in this sense, the TI criterion is similar to a cumulative distribution function.



(a) TI plots for #-sequence (left) and stripes sequence (right). Deep yellow colour marks occluded, and light yellow colour marks unoccluded frames. Note the difference in function slopes for occluded and unoccluded regions, as well as correlation of the change in slopes with the beginning of occlusions in (b).



(b) Occlusion maps at the beginning of the sequences and when occlusions start.

Figure 1: Plots of TI function of the number of frames for #- and stripes sequences (a), excerpts from the sequences when occlusions begin (b).

frames. However, this property is still desirable, because the length of the sequence for the shape prior estimation should be narrowed down. TI can be augmented with or in some cases replaced by a differential TI criterion such as:

$$\frac{TI(F_{sp}+1) - TI(F_{sp}-1)}{2} \leq \epsilon', \qquad (2)$$

where ϵ' is a threshold on the derivative value and F_{sp} is the last frame suitable for the shape prior estimation.

C. Experiments on real sequences

In this section, more details on the experiments with real image data are given.

The heart surgery sequence. The heart sequence originates from [5] and shows a patient's heart during bypass surgery naturally non-rigidly deforming. The sequence contains 60 frames; at frame 20, a robotic arm enters the scene and occludes large regions of the scene over multiple frames. The shape prior is estimated on 18 initial frames. We use an average occlusion map for every frame since some of the regions disappear or are occluded in most of the frames. The results of the experiment are shown in Fig. 2. Due to noisy initialisation under rigidity assumption, the combination MFSF+VA[2] produces reconstructions with severe inaccuracies and discontinuities (Fig. 2-(c)). Occlusion-aware MFOF+VA[2] generates visually consistent and smooth reconstructions, but we notice that natural non-rigid deformations of the heart are attenuated (due to oversmoothing). MFSF+SPVA produces visually consistent and accurate reconstructions which better reflect heart contractions.

The human face sequence. The new human face sequence depicting a speaking person was recorded with a monocular RGB camera. It contains 135 frames. Due to large displacements and deformations, the MFSF method [3] gives inaccurate correspondences especially around the frame 120. The direct method [2] relying on data term corrupts the structure and does not preserve realistic point topology. Using the per pixel per frame shape prior, we are able to preserve the point topology in the corrupted regions while relying on the data term where correspondences are accurate. Fig. 3 illustrates several problematic frames, corresponding occlusion maps and reconstructions with combinations MFSF+VA[2] and MFOF+SPVA. The experiment shows that even in the absence of occlusions a method for correspondence computation may perform poorly. In this case, reconstructions may exhibit such artefacts as broken point topology leading to corrupted reconstructions. Especially if it is not possible to recompute correspondences or available methods for obtaining correspondences do not improve the situation, our framework can be advantageous. For non-occluded regions (or regions with accurate correspondences), our approach strongly relies on the data term, whereas otherwise it strongly relies on the regularisation and shape prior terms.

The ASL sequence. The American Sign Language (ASL) sequence F5_10_A_H17 is taken from [1]. It shows a communicating person and contains severe occlusions due to hand gesticulation. Out of 114 frames, 80 frames have occlusions. To compute a shape prior, first 12 frames are used. The processing results are shown in Fig. 4. The combination MFSF[3]+VA[2] suffers from the complexity of the seemingly simple scene (Fig. 4-(c),(e)). In this experiment, the occlusion-aware MFOF could not improve the reconstruction accuracy. The reason is a mixture of a suboptimal reference view, occlusions due to significant head rotations and severe external occlusions. Our approach paired with MFSF is the only one which achieves a meaningful reconstruction on the F5_10_A_H17 sequence, perhaps for the first time in the dense case (Fig. 4-(d), (f)). In Fig. 5, additional results are shown. We tested correspondence correction [6] and found out that it did not produce expected results. In comparison to reconstructions achieved by the combination MFSF[3]+VA[2] (see Fig. 5-(a)), the combination MFOF[6]+VA[2] (see Fig. 5-(b)) deteriorates the results. Large and long occlusions are

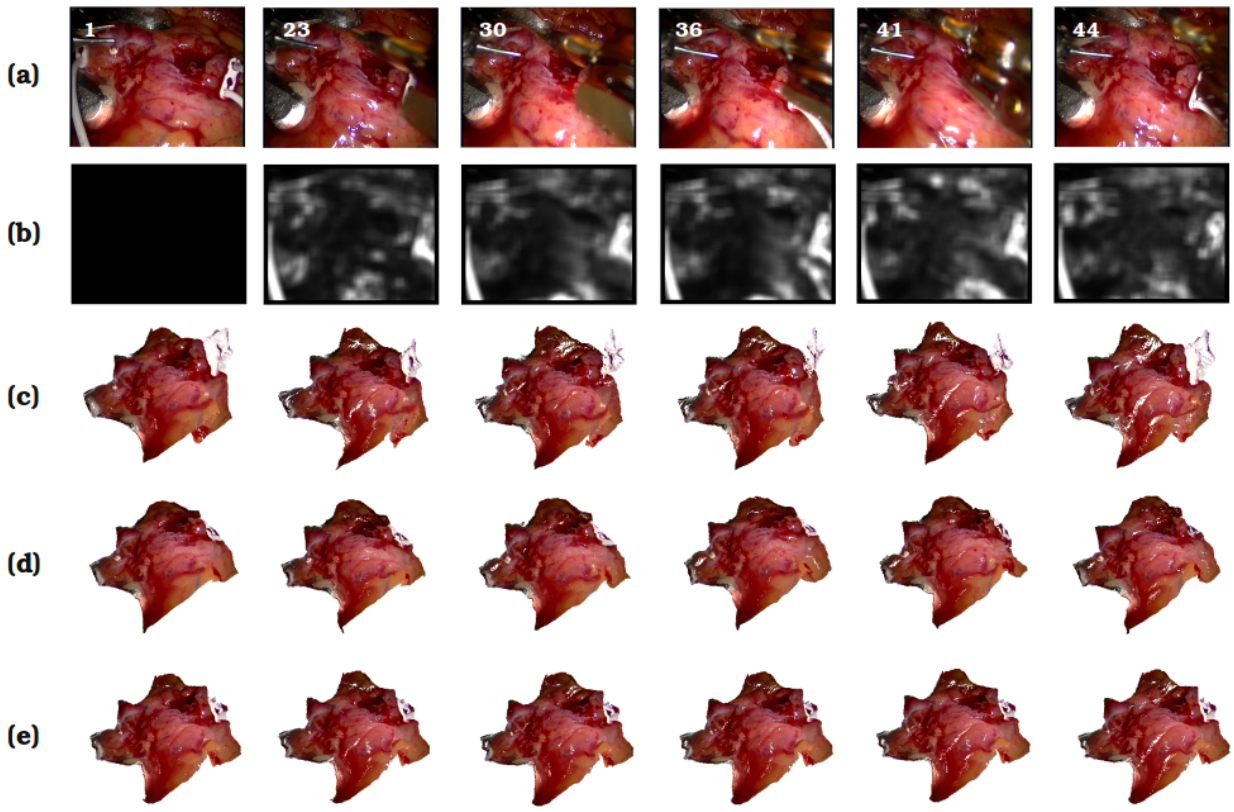


Figure 2: Experimental results on the heart sequence [5]: (a) the reference frame 1 and several occluded frames; (b) per-pixel occlusion maps for the frames shown above; (c) reconstructions with the MFSF[3]+VA[2]; (d): reconstructions with the occlusion-aware MFOF[6]+VA[2]; (e) reconstructions with the MFSF[3]+SPVA (per frame).

configuration	**heart surgery [5]** 360 × 288, 50 fr.	**face (new)** 241 × 285, 136 fr.	**ASL F5_10_A_H17 [1]** 720 × 480, 114 fr.
MFSF [3] + VA [2]	481.0 + 119.3	728.9 + 35.7	3114.0 + 400.0
MFSF [3] + AMP [4]	481.0 + 20.4	728.9 + 26.4	3114.0 + 98.0
occlusion-aware MFOF [6] + VA [2]	1592.8 + 119.2	2693.6 + 35.7	11995.3 + 300.5
MFSF [3] + SPVA	481.0 + 846.2	728.9 + 122.9	3114.0 + 1011.0

Table 1: Runtimes of different algorithm combinations for the sequences involved in the experiments, in seconds.

sequence	λ	θ	τ	γ
heart surgery	10^4	10^{-5}	10^4	$5 \cdot 10^4$
human face	$5 \cdot 10^3$	10^{-5}	$5 \cdot 10^3$	10^3
ASL	$5 \cdot 10^4$	10^{-5}	$4.2 \cdot 10^3$	10^5

Table 2: Parameters of the proposed approach for different sequences.

mainly responsible for that. MFOF [6] can compensate for occlusions of small durations. If occlusions are large and permanent, the built-in occlusion indicator of the method may fail. As a side effect, the measurements are often oversmoothed in these cases. Another reason is a high default sensitivity of the occlusion indicator. Note that we have not tuned parameters of the occlusion indicators, because they are supposed to be universal. In this example, however, large regions of the scene are spuriously detected as occlusions and the correction step relies on erroneous data.

Table 1 contains a summary of the performed experiments including parameters of the sequences, types of the applied shape priors and runtimes for all combinations and sequences. Parameters of the proposed SPVA approach are summarised in Table. 2. Recall that in an energy-based formulation, relative weights of the different terms are important, and an absolute value of the energy does not have a direct interpretation. In all experiments, σ was set to 1.0, and per frame per pixel shape prior was used.

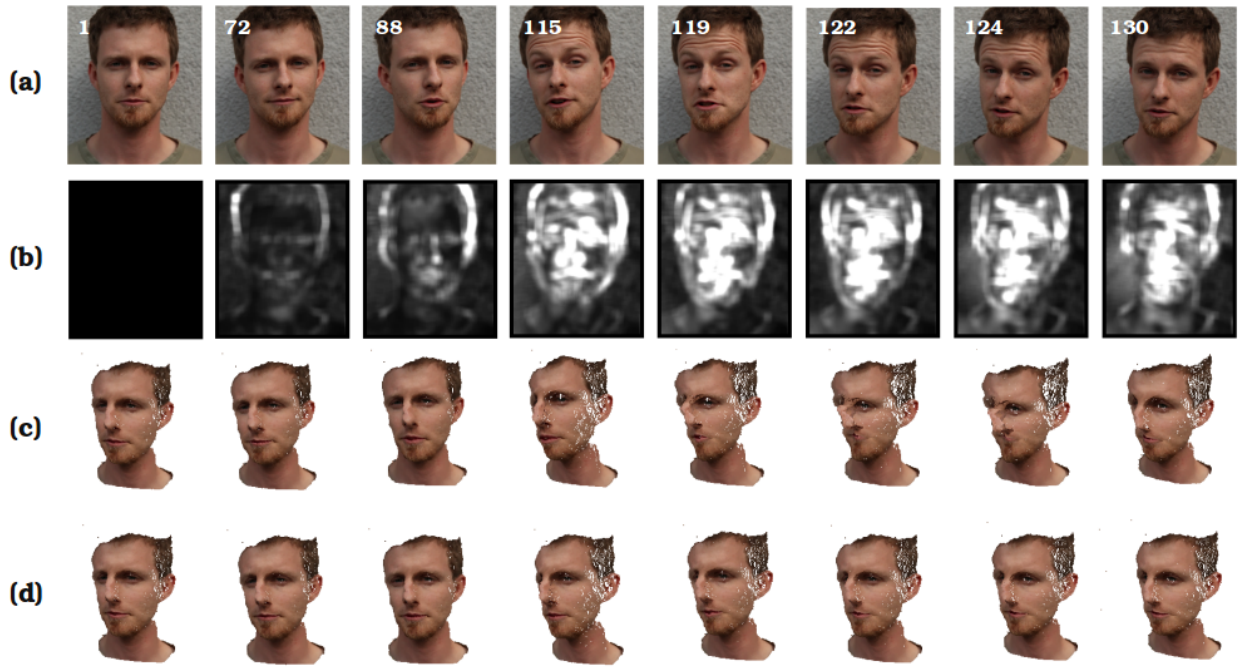


Figure 3: Experimental results on the face sequence: (a) several frames of the sequence; (b) corresponding occlusion maps; (c) reconstructions with the MFSF[3]+VA[2]; (d) reconstructions with the MFSF[3] + SPVA (per frame per pixel).

Figure 4: Experimental results for the ASL sequence [1]: (a) reference frame 1 and several frames with occlusions; (b) corresponding occlusion maps; (c) reconstructions with MFSF[3]+VA[2]; (d) reconstructions with MFSF[3]+SPVA (per frame per pixel); (e) side view of frame 28 for the configuration used in (c); (f) side view of frame 28 for the configuration used in (d).

Discussion. In all experiments, the proposed method performed more accurate on uncorrected correspondences as the base scheme, with acceptable added runtime. In some cases, the proposed approach can even produce more realistic dynamic reconstructions, as in the case of the *heart surgery* and *ASL* sequences. An advantage compared to Taetz *et al*. [6] is that the correction of inaccuracies is not restricted to a predefined procedure based on Bayesian inference. Different methods can be used to generate occlusion tensor, also integrating prior knowledge about a scene. At the same time, the proposed pipeline is faster and more suitable for online operation in scenarios with severe occlusions. Moreover, the proposed scheme can enhance the accuracy of reconstructions when the

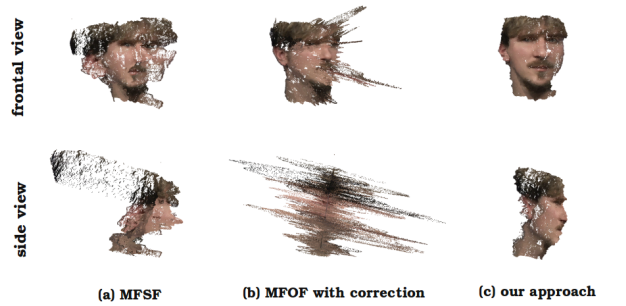


Figure 5: Results on the ASL sequence with correspondence correction; in this example, method by Taetz *et al*. [6] does not improve reconstruction accuracy; (a) reconstruction example of MFSF[3]+VA[2]; (b) reconstruction example of MFOF[6]+VA[2]; (c) reconstruction example of MFSF[3]+SPVA (the proposed approach).

occlusion-aware MFOF [6] fails to correct correspondences or correspondences cannot be computed anew. A limitation of the new approach is the dependency on the shape prior. If it is corrupted, overall accuracy can be deteriorated. Fast correspondence computation is also an open question. Occlusion detection still requires a trajectory-based approach operating on multiple frames, but once correspondences are established, an occlusion tensor can be efficiently computed in parallel.

References

[1] C. F. Benitez-Quiroz, K. Gökgöz, R. B. Wilbur, and A. M. Martinez. Discriminant features and temporal structure of nonmanuals in american sign language. *PLoS ONE*, 9:1–17, 2014. 2, 3, 4

[2] 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. 1, 2, 3, 4, 5

[3] 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. 2, 3, 4, 5

[4] 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. 3

[5] D. Stoyanov. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 479–486, 2012. 2, 3

[6] 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. 1, 2, 3, 4, 5