Dense image registration and deformable surface reconstruction in presence of occlusions and minimal texture *

Dat Tien Ngo[a]   Sanghyuk Park[b]   Anne Jorstad[a]   Alberto Crivellaro[a]
Chang Yoo[b]   Pascal Fua[a]

[a]Computer Vision Laboratory, EPFL, Switzerland
[b]School of Electrical Engineering, KAIST, Korea

Abstract

Deformable surface tracking from monocular images is well-known to be under-constrained. Occlusions often make the task even more challenging, and can result in failure if the surface is not sufficiently textured. In this work, we explicitly address the problem of 3D reconstruction of poorly textured, occluded surfaces, proposing a framework based on a template-matching approach that scales dense robust features by a relevancy score. Our approach is extensively compared to current methods employing both local feature matching and dense template alignment. We test on standard datasets as well as on a new dataset (that will be made publicly available) of a sparsely textured, occluded surface. Our framework achieves state-of-the-art results for both well and poorly textured, occluded surfaces.

1. Introduction

Being able to recover the 3D shape of deformable surfaces from ordinary images will make it possible to field reconstruction systems that require only a single video camera, such as those that now equip most mobile devices. It will also allow 3D shape recovery in more specialized contexts, such as when performing endoscopic surgery or using a fast camera to capture the deformations of a rapidly moving object. Depth ambiguities make such monocular shape recovery highly under-constrained. Moreover, when the surface is partially occluded or has minimal texture, the problem becomes even more challenging because there is little or no useful information about large parts of it.

Arguably, these ambiguities could be resolved by using a depth-camera, such as the popular Kinect sensor [33]. However, such depth-cameras are more difficult to fit into a cellphone or an endoscope and have limited range. In this work, we focus on 3D shape recovery given a reference image and a single corresponding 3D template shape known *a priori*.

When the surface is well-textured, correspondence-based methods have proved effective at solving this problem, even in the presence of occlusions [3, 5, 6, 7, 24, 26,
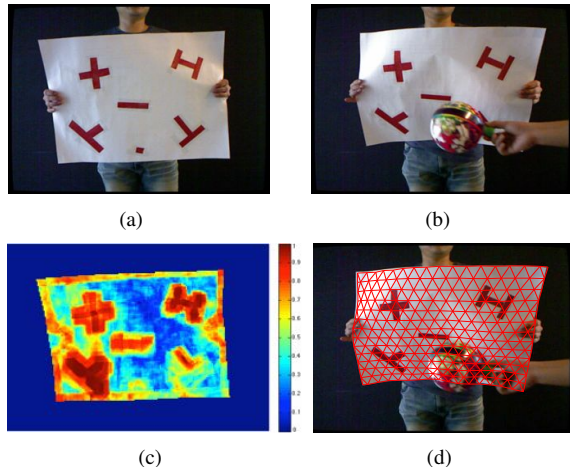
Figure 1. Tracking a sparsely textured surface in the presence of occlusion: (a) template image, (b) input image, (c) relevancy score, (d) surface tracking result with proposed framework. All figures in this paper are best viewed in color.

37]. In contrast, when the surface lacks texture, dense pixel-level template matching should be used instead. Unfortunately, many methods such as [21, 31] either are hampered by a narrow basin of attraction, which means they must be initialized from interest points correspondences, or require supervised learning to enhance robustness. Using Mutual Information has often been claimed [10, 12, 23, 38] to be effective at handling these difficulties but our experiments do not bear this out. Instead, we advocate template matching over robust dense features that relies on a pixel-wise relevancy score pre-computed for each frame, as shown in Fig. 1. Our approach can handle occlusions and lack of texture simultaneously. Moreover, no training step is required as in [31], which we consider to be an advantage because this obligates either collecting training data or having sufficient knowledge of the surface properties, neither of which may be forthcoming.

Our main contribution is therefore a robust framework for image registration and monocular 3D reconstruction of deformable surfaces in the presence of occlusions and minimal texture. A main ingredient is the pixel-wise relevancy score we use to achieve the robustness. We will make the code publicly available, and release the dataset we used

to validate our approach, which contains challenging sequences of sparsely-textured deforming surfaces and the corresponding ground truth.

2. Related Work

The main approaches for deformable surface reconstruction either require 2D tracking throughout a batch of images or a video sequence [1, 13, 27] or they assume a reference template and corresponding 3D shape is known. In this work, we focus on the second approach, which we refer to as *template-based reconstruction*.

The most successful current approaches generally rely on finding feature point correspondences [22, 3, 5, 7, 24, 26, 37], because they are robust to occlusions. Unfortunately, as shown by our experimental results, these methods tend to break down when attempting to reconstruct sparsely or repetitively textured surfaces, since they rely on a fairly high number of correct matches.

Pixel-based techniques are able to overcome some limitations of local feature matching, since they reconstruct surfaces based on a global, dense comparison of images. On the other hand, some precautions must be taken to handle occlusions, lighting changes, and noise. [21] estimates a visibility mask on the reconstructed surface, but unlike us, only textured surfaces and self-occlusions are handled. [14] registers images of deformable surfaces in 2D and shrinks the image warps in self-occluded areas. [3] proves that an analytical solution to the 3D surface shape can be derived from this 2D warp. However, the surface shape in self-occluded areas is undefined. [8] registers local image patches of feature point correspondences to estimate their depths, and geometric constraints are imposed to classify incorrect feature point correspondences. In contrast to these local depth estimations, our method reconstructs surfaces globally in order to be more robust to noise and outliers.

Other recent approaches employ supervised learning for enhancing performance [28, 36]. In [31] strong results are achieved with poorly textured surfaces and occlusion by employing trained local deformation models, a dense template matching framework using Normalized Cross Correlation (NCC) [32] and contour detection. Our proposed framework manages to achieve similar performance without requiring any supervised learning step, while the use of robust, gradient-based dense descriptors recently proposed in [9] avoids the need to explicitly detect contours.

Other techniques employed for dealing with occlusions and noise, such as Mutual Information (MI) [11, 12, 23, 38] and robust M-estimators [2] are studied explicitly in our context, and found to be successful only up to a point.

Our method is similar to that of [25], where a template matching approach is employed and a visibility mask is computed for the pixels lying on the surface, but in this work a very good initialization from a feature point-based

method is required in order for its EM algorithm to converge. In addition to the geometrical degrees of freedom of the surface, local illumination parameters are explicitly estimated in [16, 34]. This requires a reduced deformation model for the surface to keep the size of the problem reasonable.

In the proposed framework, we achieve good performance without the need to explicitly estimate any illumination model, so that an accurate geometric model for the surface can be employed. Furthermore, rather than estimating a simple visibility mask as is often done in many domains such as stereo vision [35], face recognition [40], or pedestrian detection [39], we employ a real-valued pixel-wise relevancy score, penalizing at the same time pixels with unreliable information originating both from occluded and low-textured regions. Our method has a much wider basin of convergence and we can track both well and poorly textured surfaces without requiring initialization by a feature point-based method.

3. Proposed Framework

In this work, we demonstrate that a carefully designed dense template matching framework can lead to state-of-the-art results in monocular reconstruction of deformable surfaces. In this section we describe our framework, based on a recently introduced gradient-based pixel descriptors [9] for robust template matching and the computation of a relevancy score for outlier rejection.

3.1. Template Matching

We assume we are given both a template image T and the rest shape of the corresponding deformable surface, which is a triangular mesh defined by a vector of N_v vertex coordinates in 3D, $\mathbf{V}_T \in \mathbb{R}^{N_v \times 3}$. To recover the shape of the deformed surface in an input image I, the vertex coordinates $\mathbf{V}_T$ of the 3D reference shape must be adjusted so that their projection onto the image plane aligns with I.

We assume the internal parameters of the camera are known and, without loss of generality, that the world reference system coincides with the one of the camera. In order to register each input image, a pixel-wise correspondence is sought between the template and the input image. Each pixel $\mathbf{x} \in \mathbb{R}^2$ on the template corresponds to a point $\mathbf{p} \in \mathbb{R}^3$ on the 3D surface. This 3D point is represented by fixed barycentric coordinates which are computed by backprojecting the image location $\mathbf{x}$ onto the 3D reference shape.

The camera projection defines an image warping function $\mathbf{W} : \mathbb{R}^2 \times \mathbb{R}^{3 \times N_v} \to \mathbb{R}^2$ which sends pixel $\mathbf{x}$ to a new image location based on the current surface mesh $\mathbf{V}$ as illustrated in Fig. 2. The optimal warping function should minimize the difference between $T(\mathbf{x})$ and $I(\mathbf{W}(\mathbf{x}; \mathbf{V}))$, according to some measurement of pixel similarity. Traditionally, image intensity has been used, but more robust
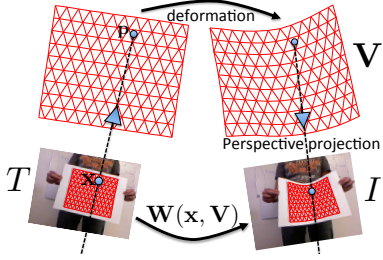
Figure 2. An image warping function maps a pixel from the template image onto the deforming surface in the input image.

pixel feature descriptors $\phi_I(\mathbf{x})$ will lead to more meaningful comparisons, as discussed in Section 3.2.2.

The image energy cost function is a comparison between $\phi_T(\mathbf{x})$ and $\phi_I(\mathbf{W}(\mathbf{x}; \mathbf{V}))$ at every image point $\mathbf{x}$ defining the quality of their alignment

$$E_{\text{image}}(\mathbf{V}) = \sum_{\mathbf{x}} d\big(\phi_T(\mathbf{x}), \phi_I(\mathbf{W}(\mathbf{x}; \mathbf{V}))\big). \quad (1)$$

There are many possible choices for the function d comparing the descriptor vectors, such as Sum of Squared Differences (SSD), NCC, MI, and others. We will discuss more in detail about the choice of d in Section 3.2.3.

Since monocular 3D surface reconstruction is an under-constrained problem and there are multiple 3D shapes having the same reprojection on the image plane, minimizing the image energy in Eq. (1) alone is ill-posed. Additional constraints must be added, such as isometric deformation constraints enforcing that the surface should not stretch or shrink. A change in the length between vertex $\mathbf{v}_i$ and vertex $\mathbf{v}_j$ as compared to the template rest length l_{ij} from $\mathbf{V}_T$ is penalized as

$$E_{\text{length}}(\mathbf{V}) = \sum_{i,j} (\|\mathbf{v}_i - \mathbf{v}_j\| - l_{ij})^2. \quad (2)$$

To encourage physically plausible deformations, the Laplacian mesh smoothing proposed in [22] is used. This rotation-invariant curvature-preserving regularization term based on Laplacian smoothing matrix $\mathbf{A}$ penalizes non-rigid deformations away from the reference shape, based on the preservation of affine combinations of neighboring vertices.

$$E_{\text{smooth}}(\mathbf{V}) = \|\mathbf{A}\mathbf{V}\|^2. \quad (3)$$

To reconstruct the surface, we therefore seek the mesh configuration $\mathbf{V}$ that minimizes the following total energy:

$$\underset{\mathbf{V}}{\arg\min} \; E_{\text{image}}(\mathbf{V}) + \lambda_{\text{L}} E_{\text{length}}(\mathbf{V}) + \lambda_{\text{S}} E_{\text{smooth}}(\mathbf{V}), \quad (4)$$

for relative weighting parameters λ_{L} and λ_{S}.

3.2. Robust Optimization

3.2.1 Optimization Scheme

To make the optimization more robust to noise and wide pose changes, we employ a multi-scale approach, iteratively minimizing $E^\sigma = E^\sigma_{\text{image}} + \lambda_{\text{L}} E_{\text{length}} + \lambda_{\text{S}} E_{\text{smooth}}$ for decreasing values of a scale parameter σ, with:

$$E^\sigma_{\text{image}} = \sum_{\mathbf{x}} d\left(G^\sigma * \phi_T(\mathbf{x}), G^\sigma * \phi_I(\mathbf{W}(\mathbf{x}; \mathbf{V}))\right), \quad (5)$$

where G^σ is a low-pass Gaussian filter of variance σ^2. In our experiments we solve the alignment at three scales, using the final result of each coarser scale to initialize the next set of iterations, and initializing the coarsest scale with the final position found for the previous frame. The first frame of each image sequence is taken as the template, and we employ a standard Gauss-Newton algorithm for minimization.

3.2.2 Feature Selection

The image information compared in Eq. (1) comes from pixel-based image features. Previous approaches [21, 25, 31] employ image intensity as a local descriptor, $\phi_I(\mathbf{x}) = I(\mathbf{x})$. More robust results can be obtained with other features, such as the lighting-insensitive image gradient direction (GD) [15], where $\phi_I(\mathbf{x}) = \tan^{-1}\frac{I_y(\mathbf{x})}{I_x(\mathbf{x})}$ with $\mod 2\pi$ differencing. Based on its strong previous performance we also consider the Gradient Based Descriptor Fields (GBDF) recently proposed in [9]:

$$\phi_I(\mathbf{x}) = \left[[\frac{\partial I}{\partial x}(\mathbf{x})]^+, \ [\frac{\partial I}{\partial x}(\mathbf{x})]^-, \ [\frac{\partial I}{\partial y}(\mathbf{x})]^+, \ [\frac{\partial I}{\partial y}(\mathbf{x})]^-\right]^\top,$$
$$(6)$$

where the $[\cdot]^+$ and $[\cdot]^-$ operations respectively keep the positive and negative values of a real-valued signal. These descriptors are robust under light changes, and remain discriminative after the Gaussian smoothing employed in Eq. (5); however, as originally proposed in [9], they are not rotation invariant. To achieve in-plane rotation invariance, in our final framework we employ a modified version of GBDF. In order to compare pixel descriptors in the same, unrotated coordinate system, the reconstruction of the previous frame is used to establish a local coordinate system for each mesh facet. Each pixel descriptor on the template is then rotated in accordance with its corresponding mesh facet, to be directly comparable to the points in the input image. We show in Section 4 that this modification indeed increases registration accuracy by being able to successfully track a rotating deformable surface.

3.2.3 Similarity Function Selection

Choosing the correct comparison function d for Eq. (1) also significantly affects the robustness of the tracking. Common choices include the SSD of the descriptors, and the NCC of image intensities [20], which is invariant under affine changes in lighting.

Mutual Information: MI [38] is a similarity function that measures the amount of information shared between two variables, and it is known to be robust to outliers such as noise and illumination changes [10]. It has been repeatedly claimed to be robust to occlusion, for example in [10, 12, 23, 38]. Where occlusions occur, the shared information between occluded pixels and the template image is low or none, and its variation does not cause significant change in the image entropy; therefore, the MI obtains an accurate maximum value at the position of the correct alignment, in spite of the occlusion.

However, an MI-based cost function is limited in application. MI generally provides a non-convex energy function with a very strong response at the optimum, but a very narrow basin of convergence, as shown in Fig. 5. This makes it unsuited for direct numerical optimization, while smoothing leads to a significantly degraded energy function. Numerical experiments reported in Section 4 show that, in our context, MI leaves room for improvement.

Robust Statistics: M-estimators are a popular method for handling outliers in a template matching framework. Let $e_i = \phi_I(\mathbf{W}(\mathbf{x}_i; V)) - \phi_T(\mathbf{x}_i)$ be the residual at pixel $\mathbf{x}_i$; then instead of minimizing the sum of squared residuals $\sum_i e_i^2$, a modified loss function ρ of the residuals is considered, instead minimizing $\sum_i \rho(e_i)$, in order to reduce the influence of outliers.

In Section 4, tests are performed using two of the most commonly employed M-estimators, the Huber [18] and the Tukey [17] estimators. In our context, M-estimators show moderate efficacy, likely because part of the useful information is rejected as outliers. This problem becomes particularly significant when dealing with low-textured surfaces, where the amount of information available for alignment is low.

3.3. Handling Occlusions with a Relevancy Score

Our experiments suggest that selecting a robust similarity function is not enough to deal with the occlusions and image variability encountered when attempting to track a deforming surface in real-world imagery.

Inspired by the effectiveness of the occlusion masks developed in works such as [35, 39, 40], we derive a more robust method to handle occlusions by pre-computing a relevancy score for each pixel of the current frame, which is then used to weight the pixels during the alignment. Since we would like to handle occlusions and sparsely textured surfaces together, rather than designing a binary occlusion prediction mask, we develop a continuous-valued score that will raise or lower the importance of pixels depending on their relevancy. This pre-processing step can greatly improve the quality of the image information handed to the cost function in Eq. (4).
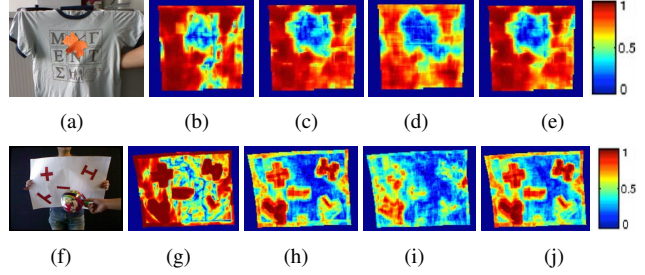

Figure 3. The relevancy score results using various methods on the (a) cloth dataset and (f) sparsely textured paper dataset using (b)(g) Intensity (c)(h) GBDF (d)(i) Gradient direction (e)(j) GBDF+Intensity. GBDF gives a better relevancy map than intensity and gradient direction on the first dataset while intensity is better than GBDF and gradient direction on the second. We therefore combine both intensity and GBDF in our proposed relevancy score.

Given the estimated configuration $\mathbf{V}_{t-1}^*$ of the deformable surface from the previous frame, a thin-plate spline-based warping function [4] is used to un-warp image I_t to closely align with the template T. A relevancy score is then computed between each pixel $\mathbf{x}$ on the synthetically back-warped image $\hat{I}_t$ and the same pixel on the template T, with a sliding-window approach.

It has been verified repeatedly in the literature that NCC is a reliable choice for measuring patch-based image similarity, and so we compute the NCC over the images as an efficient prediction of relevancy. In one such approach [19], local image patches are affinely warped based on the predicted camera pose, and sliding NCC windows are then used to look for correspondences of map points in the input image. Our approach is somewhat different, as we use sliding NCC to measure the relevancy of template pixels on the input image. We average the NCC of both the image intensity and the GBDF features, as it was found that both descriptors provide relevant and often complementary information at this predictor stage, (see Fig. 3 for a qualitative comparison).

The sliding relevancy score is computed as the maximum NCC value over a range of patches near image location $\mathbf{x}$:

$$\omega(\mathbf{x}) = \max_\delta \text{NCC}(\mathcal{P}_T(\mathbf{x}), \mathcal{P}_{\hat{I}}(\mathbf{x} + \delta)), \qquad (7)$$

where $\mathcal{P}_T(\mathbf{x})$ and $\mathcal{P}_{\hat{I}}(\mathbf{x})$ are patches of size 26×26 centered at $\mathbf{x}$, $\delta = [\delta_x, \delta_y]^T$, and δ_x, δ_y vary over $[-30, 30]$ in all our experiments. Allowing the patch to be compared to nearby patches accounts for some of the variability between the surface position $\mathbf{V}_{t-1}^*$ and the desired position $\mathbf{V}_t^*$ to be recovered.

The similarity scores are then normalized to lie in $[0, 1]$, and outlier data is also limited at this stage in a process similar to an M-estimator. The mean μ and standard deviation σ of the NCC scores are found for each frame, and all values further than 3σ from the mean are clamped to the interval
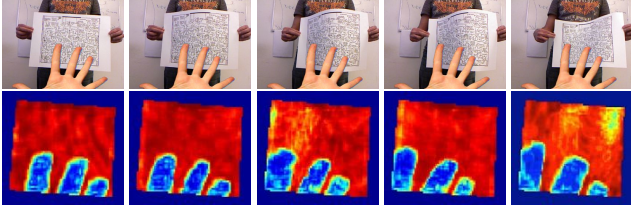
Figure 4. Relevancy scores for the well-textured paper dataset.

$\mu \pm 3\sigma$. These values are then linearly rescaled to lie between 0 and 1, and the normalized weights $\hat{\omega}$ are applied to the data in the image energy term of Equation (4):

$$E_{\text{image}}(V) = \sum_{\mathbf{x}} \hat{\omega}(\mathbf{x}) d\big(\phi_T(\mathbf{x}), \phi_I(\mathbf{W}(\mathbf{x}; V))\big), \quad (8)$$

where the sum here is extended to all the pixels of the template. Relevancy scores for the well-textured paper dataset are shown in Fig. 4.

3.4. Handling Sparsely Textured Surfaces

The relevancy score described in Section 3.3 is also able to handle sparsely textured surfaces. Image regions containing little or no texture have low relevancy scores, so these pixels will not negatively influence the image alignment. For example, see Fig. 1. Using the proposed relevancy score to weight the utility of the image information coming from each pixel in the image allows the optimization to be driven by the most meaningful available information.

4. Experiments and Results

3D surface reconstructions are computed with and without occlusion on both well and poorly textured deforming surfaces. We compare recent methods described in Section 2, which are representative of the current state-of-the-art, against our dense template matching-based reconstruction methods using the various similarity measures and occlusion handling techniques described in Section 3.

In particular, we report detailed results of comparisons with the following methods: "Bartoli12" [3], that reconstructs the surface by analytically solving a system of PDEs starting from an estimated 2D parametric warp between images; "Chhatkuli14" [7], that infers the surface shape exploiting the depth gradient non-holomonic solution of a PDE; "Brunet14" [6], that reconstructs a smooth surface imposing soft differential constraints of isometric deformation; "Ostlund12" [22], that introduces the Laplacian mesh smoothing we employ; and "Salzmann11" [29], that uses pre-learned linear local deformation models.[1]

As for pixel-based template matching techniques, comparing pixel intensity values "Intensity" and gradient direction values "GD" are done using SSD. We also compare standard "NCC" and "MI" over intensity values. The

---

[1]Code provided by the authors of these papers was used for all comparisons.

Table 1. Reconstruction errors over a range of weighting coefficient values using the well-textured paper dataset.

error (mm)		λ_L							
		10	2	1	0.5	0.25	0.1	0.05	0.01
λ_S	10	7.46	6.46	5.73	5.59	5.31	18.45	93.07	N.A
	2	4.60	1.58	2.41	3.68	4.49	5.17	17.23	319.09
	1	1.93	1.39	1.20	1.97	3.44	4.61	5.94	147.96
	0.5	2.02	1.73	1.43	1.08	1.80	3.76	4.77	61.87
	0.25	2.01	1.86	**1.67**	1.45	1.10	2.17	3.76	26.18
	0.1	2.05	1.91	1.75	1.62	1.57	1.20	1.68	240.66
	0.05	2.07	2.03	1.96	1.86	1.82	5.62	14.86	185.47
	0.01	18.23	15.44	6.25	6.45	6.31	10.44	14.11	342.12

"GBDF" features are compared using SSD, and were seen to be the strongest feature descriptor, so it is these values that we test in the M-estimator framework using the "Huber" and "Tukey" loss functions. Our proposed framework from Section 3.3 is labeled "GBDF+Oc" in the figures. We see that it achieves state-of-the-art performances on a standard, well-textured dataset, and it achieves optimal reconstruction performance in all datasets with occlusions and low texture.

Image sequences were acquired using a Kinect camera, and ground truth surfaces were generated from the depth information. The template is constructed from the first frame, and 3D reconstruction is performed for the rest of the sequence using the image information alone. The initial mesh coordinates for each frame are set to the locations of the final reconstruction of the previous frame in the sequence.

We consider two different metrics to define the reconstruction accuracy. Many previous methods compare the average distance of the reconstructed 3D mesh vertices to their closest projections onto the depth images. This metric ignores the correspondences between the mesh points and the point cloud. As a more meaningful metric, we use the Kinect point cloud to build ground truth meshes, and compute the average vertex-to-vertex distance from the reconstructed mesh to the ground truth mesh. This metric is used for the paper itself. Results using the vertex-to-point-cloud distance are provided in the supplementary material.

To ensure a fair comparison, all results are presented using the best parameter values found for each method, tuned separately. To ensure that our results are not overly sensitive to the selection of parameters λ_L and λ_S, we performed the full reconstruction on the well-textured paper dataset over a wide range of values, as presented in Table 1. It can be observed that increasing or decreasing these parameters by a factor of two around $\lambda_L = 1$ and $\lambda_S = 0.25$ results in very little change in the final reconstruction accuracy, implying that the method is sufficiently insensitive to these parameters as long as they are within a reasonable range.

The surface rest shape is modeled by a 10×13 triangle mesh in the well-textured dataset, 14×17 in the sparsely textured dataset, and 15×14 on the T-shirt dataset. The σs used in the hierarchical procedures were $\{15, 7, 3\}$ and $\{5, 3, 2\}$.

Our approach relies on frame-to-frame tracking and thus

requires a sufficiently good initialization. However, because the method has a wide basin of convergence, a rough initialization suffices. Our method can fail when the initialization is too far from the solution, when frame-to-frame deformations are so large that the relevancy scores stop being reliable, or when large changes in surface appearance and severe occlusions cause the image energy term to become uninformative. If the tracking is lost, it must be reinitialized, for example by using a feature point based method. However, this did not prove to be necessary to obtain any of the results shown below.

4.1. Basin of Convergence

To understand the limitations of the various cost functions, we conducted a simple alignment experiment to test their respective sensitivities to initial position and image distractors; results are presented in Fig. 5. The red image window in the input image is translated in x and y about the known best alignment to the green template window, and the cost to compare each window pair is plotted, to allow the basins of convergence to be inspected visually. MI and NCC both reach a maximum value close to 1 at the point of best alignment, but we invert these functions so that a minimum cost of all functions is expected at the point of best alignment.

The GBDF descriptors from Eq. 6 are seen to have a strong minimum at the point of best alignment, with a reasonably wide, smooth basin of convergence, the desired property of a good cost function. However, Intensity, MI, and NCC all have several nearby local minima. Mutual Information is further seen to have a very narrow basin of convergence around the correct point of best alignment, meaning that it is very likely to converge to an incorrect alignment given an imperfect initial position.

This experiment only tests translation sensitivity because this is the variation best understood visually, but the similar results are likely from other types of misalignment.

4.2. Well-Textured Surfaces

We performed experiments using the well-textured paper dataset presented in [36] consisting of 193 consecutive images, for example see Fig. 8. Quantitative results are presented in Fig. 6. For this well-textured dataset, all the feature point-based methods work well and dense matching methods are only slightly better. The biggest errors are due to lighting changes, where intensity features using SSD occasionally fail to track part of the surface, and hence have a higher error.

To evaluate the robustness of each method to occlusion, we add artificial hand image occlusions to the image sequence. The reconstruction results are presented in Fig. 7. Feature-based methods still produce reasonably good 2D reprojection results in this dataset, but the recovered depths
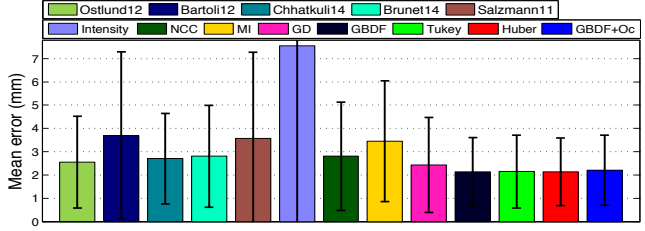

Figure 6. Reconstruction results on the well-textured paper dataset, no occlusions. All feature-points based methods work reasonably well.
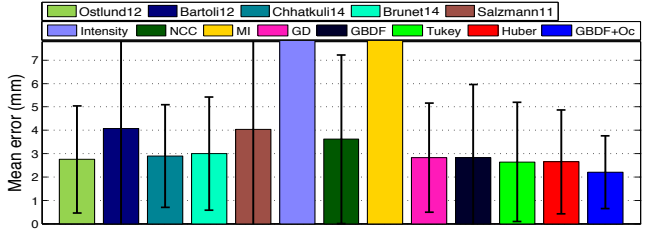

Figure 7. Reconstruction results on the well-textured paper dataset, with occlusions. Feature-based methods are largely robust to occlusion, however the overall depths recovered are not as accurate as the proposed framework that includes occlusion handling.

(a) Ostlund12   (b) GBDF alone   (c) MI

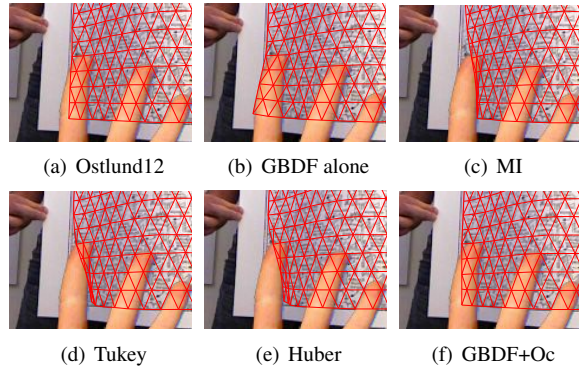(d) Tukey   (e) Huber   (f) GBDF+Oc

Figure 8. Output for a single frame showing relative reconstruction accuracies. Mutual Information and M-estimators fail to correctly handle the occlusion, while the proposed framework is successful.

under the occlusion are not very accurate. Fig. 8 provides the output for a single frame where it can be seen that the reconstruction fails when using the strong GBDF without occlusion handling and also when using M-estimators to attempt to handle occlusion, while the proposed framework is still able to able track the surface accurately. In this situation, Mutual Information and both the Tukey and Huber M-estimators are confused by the edges created by the finger and converge to incorrect locations.

We also demonstrate that the proposed rotation handling technique described in Section 3.2.2 that overcomes the rotation sensitivity of the GBDF descriptors can successfully track a rotating deformable object. Fig. 9 shows that without rotation handling, the original GBDF descriptors can only track up to 50 degrees of rotation, while the proposed rotation handling technique can track the whole sequence including a full 360 degrees of rotation.
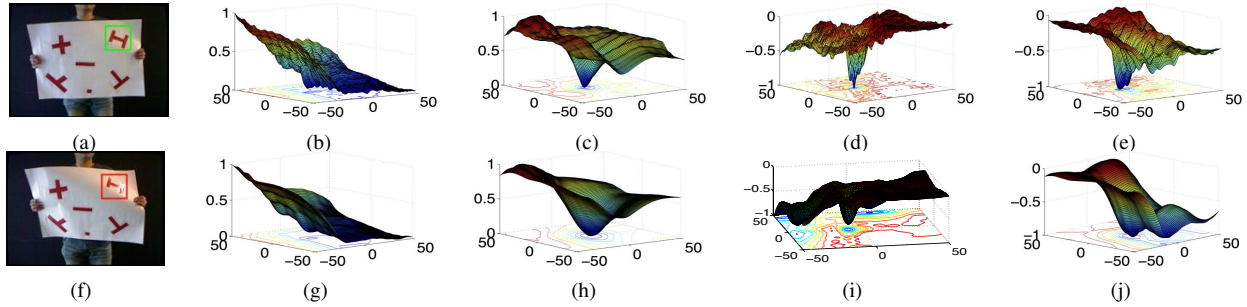
Figure 5. Robustness of alignment functions *w.r.t.* translations between (a) template, and (f) input image, showing the basin of convergence of the alignment costs around the correct position using **Top row:** weak Gaussian smoothing, **Bottom row:** strong Gaussian smoothing, over (b)(g) Intensity, (c)(h) GBDF, (d)(i) MI, (e)(j) NCC.

Figure 9. Tracking a rotating deformable surface. **Top row:** Without rotation handling, tracking with the original GBDF descriptors eventually breaks down. **Bottom row:** Our modified GBDF features can track the whole sequence with up to 360° of rotation.

4.3. Sparsely Textured Surfaces

To understand the performance of the methods in a re-alistic, sparsely textured setting, a different dataset is re-quired. A less textured paper dataset exists, as published in [31], but no ground truth information is available for this dataset in 3D, and so it is not suitable for numerical compar-isons. Nevertheless, for qualitative comparison purposes, we ran the proposed framework on this dataset, and our re-constructions align very well to the image information. Ex-ample frames are provided in Fig. 10, and the entire video is provided as supplementary material. The best known pub-lished results on this dataset are found in [30], which uses an algorithm that requires training data in addition to explicitly delineating the edges of the surface. Our proposed frame-work is seen to perform as well as this previous method, qualitatively, while requiring no learning.

In order to be able to perform more meaningful numer-ical comparisons, we constructed a new dataset along with ground truth in 3D using a Kinect sensor, example images are provided in Fig. 12. This new sparsely textured pa-per dataset contains various deformations and large lighting changes along with occlusions.

Quantitative results using the new dataset are presented in Fig. 11. Feature-based methods that fail to reconstruct plausible surface shapes are indicated by high error bars that exceed y-axis range. Fig. 12 provides a representative reconstruction on a single frame. NCC and MI can track the surface fairly accurately, however they fail to capture fine details at the surface boundaries and hence the recov-ered depths in 3D are not very precise. Without occlusion
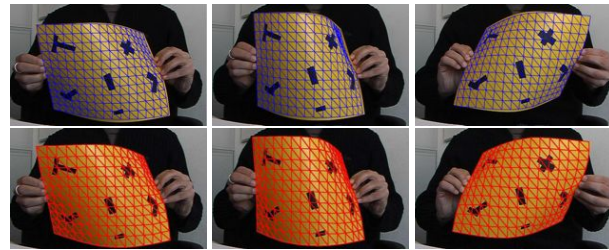

Figure 10. Sample reconstructions from the [31] dataset. While no ground truth is available in 3D, our results (top row) are qual-itatively observed to be very accurate; the best published results on this dataset are [31] (bottom row), which has to extract the im-age edges explicitly, and involves learning, while our method does not. We do not have access to a reference image where the sur-face is in its planar rest shape, as our mesh assumes, causing some misalignment at the surface boundary.
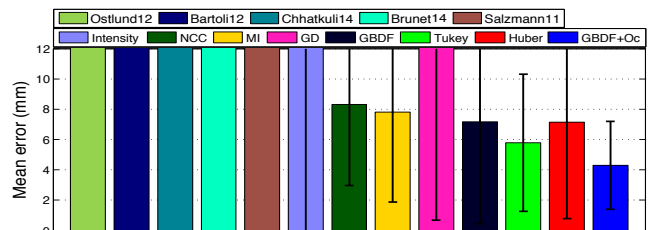

Figure 11. Reconstruction results on the sparsely textured paper dataset. Feature-based methods fail to reconstruct plausible sur-faces, as indicated by the out-of-range error bars on the left.

handling, dense matching with gradient-based descriptors often fails near occlusions. The M-estimators are inconsis-tent near occlusions. However, the proposed framework is seen to be able to accurately track the surface throughout the entire sequence.

(a) GBDF+Oc   (b) NCC   (c) MI
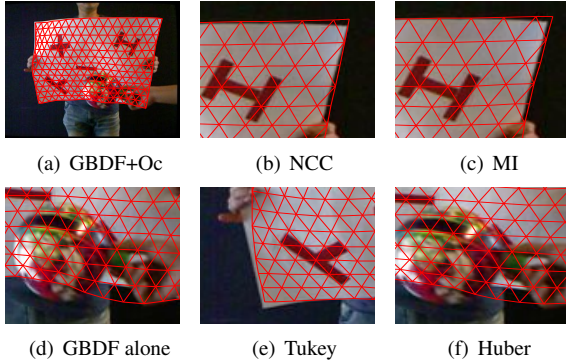
(d) GBDF alone   (e) Tukey   (f) Huber

Figure 12. Reconstruction results on the same single frame. The proposed framework can track the whole sequence accurately, while other methods are seen to have trouble handling occlusions on top of the sparsely textured surface.

It is interesting to note that while the occlusions are cleanly delineated in the relevancy score over a textured surface, they are much less obviously visible in the relevancy score over a sparsely textured surface. This is expected, because well-textured un-occluded regions have consistently high correlation values with the template, and so it is only the occluded regions that are assigned low relevancy scores. However, image regions of little texture have low and noisy correlation with a template, so occluded regions of similarly low correlation are assigned similarly low relevancy scores, and an occluded region is not as obviously distinct from the low textured regions in the relevancy score map. This is one of the strengths of the proposed framework, because only the truly meaningful image regions are allowed to strongly influence the image energy cost.

4.4. Applications

We demonstrate the robustness of our method in a variety of real-world applications. First, we provide results on a cloth surface undergoing a different type of deformation than studied in the paper datasets. We created a new dataset along with ground truth in 3D using a Kinect sensor, as before, to which artificial occlusions were added, example images and our reconstructions are shown in Fig. 13. Quantitative results are presented in Fig. 14.

The strength of our approach is demonstrated on a sparsely-textured sail surface with a few dot markers, shown in Fig. 15. Thanks to the large basin of convergence of our algorithm, we can simply initialize the registration from a very rough initial estimate without having first to establish correspondences. Our algorithm naturally exploits line fea-
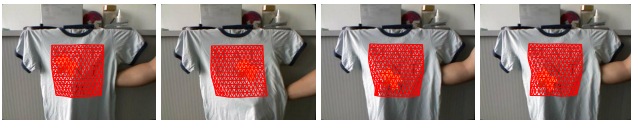

Figure 13. Our representative reconstructions on the T-shirt dataset with artificial occlusions added. Rightmost: a tracking failure case when occlusions appear at areas with large deformations.
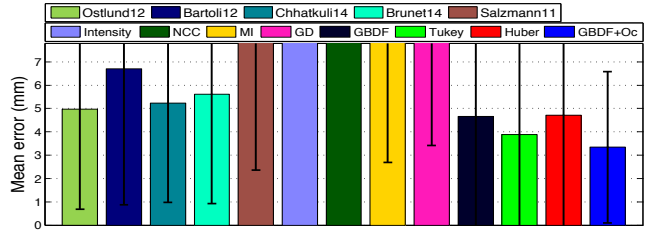

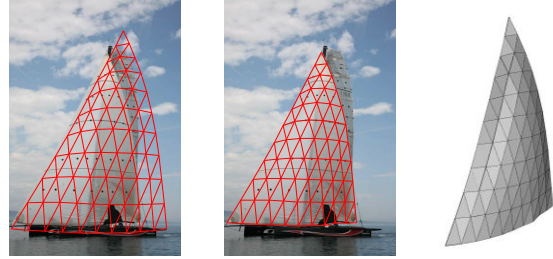Figure 14. Reconstruction results on the T-shirt dataset.

Figure 15. Image registration and surface reconstruction on the sparsely textured sail surface. From left to right: input image and initialization; final registration and reconstruction; the sail shape seen from a different viewpoint.
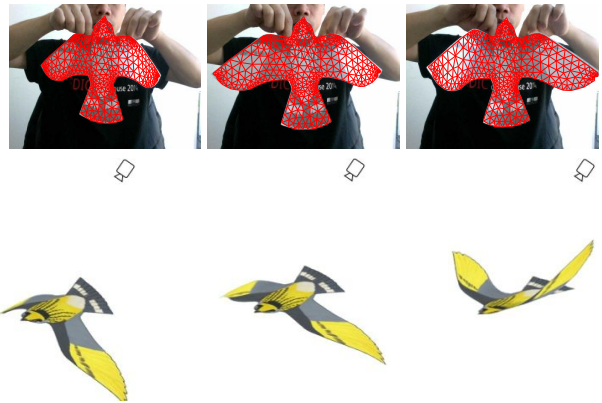

Figure 16. Surface reconstruction of an animation capture from a monocular camera stream.

tures, which feature point-based methods usually do not.

Fig. 16 depicts another application of our method for animation capture from a monocular camera stream. In this setting, we capture the animations of a bird whose animations can be transferred to another character. The video of captured animations is provided in the supplementary material.

5. Conclusion

We have presented a framework for tracking both well textured and sparsely textured deforming surfaces in videos in the presence of occlusions. Our framework computes a relevancy score for each pixel, which is then used to weight the influence of the image information from that pixel in the image energy cost function. The presented method favorably compares to standard cost functions used for handling occlusion, such as Mutual Information and M-estimators.

References

[1] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory Space: A Dual Representation for Nonrigid Structure from Motion. *PAMI*, 33(7):1442–1456, 2011. 2

[2] K. Arya, P. Gupta, P. Kalra, and P. Mitra. Image Registration Using Robust M-Estimators. *PR*, 28(15):1957–1968, 2007. 2

[3] A. Bartoli, Y. Gérard, F. Chadebecq, and T. Collins. On Template-Based Reconstruction from a Single View: Analytical Solutions and Proofs of Well-Posedness for Developable, Isometric and Conformal Surfaces. In *CVPR*, 2012. 1, 2, 5

[4] F. Bookstein. Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *PAMI*, 11(6):567–585, 1989. 4

[5] S. Bronte, M. Paladini, L. Bergasa, L. Agapito, and R. Arroyo. Real-Time Sequential Model-Based Non-Rigid SFM. In *IROS*, pages 1026–1031, 2014. 1, 2

[6] F. Brunet, A. Bartoli, and R. Hartley. Monocular Template-Based 3D Surface Reconstruction: Convex Inextensible and Nonconvex Isometric Methods. *CVIU*, 125:138–154, 2014. 1, 5

[7] A. Chhatkuli, D. Pizarro, and A. Bartoli. Stable Template-Based Isometric 3D Reconstruction in All Imaging Conditions by Linear Least-Squares. In *CVPR*, 2014. 1, 2, 5

[8] T. Collins and A. Bartoli. Using Isometry to Classify Correct/incorrect 3D-2D Correspondences. In *ECCV*, pages 325–340, 2014. 2

[9] A. Crivellaro and V. Lepetit. Robust 3D Tracking with Descriptor Fields. In *CVPR*, 2014. 2, 3

[10] A. Dame and E. Marchand. Second-Order Optimization of Mutual Information for Real-Time Image Registration. *IEEE Transactions on Image Processing*, 21(9):4190–4203, 2012. 1, 4

[11] D. Damen, P. Bunnun, A. Calway, and W. Mayol-cuevas. Real-Time Learning and Detection of 3D Texture-Less Objects: A Scalable Approach. In *BMVC*, 2012. 2

[12] N. Dowson and R. Bowden. A Unifying Framework for Mutual Information Methods for Use in Non-Linear Optimisation. In *ECCV*, pages 365–378, 2006. 1, 2, 4

[13] R. Garg, A. Roussos, and L. Agapito. Dense Variational Reconstruction of Non-Rigid Surfaces from Monocular Video. In *CVPR*, 2013. 2

[14] V. Gay-Bellile, A. Bartoli, and P. Sayd. Direct Estimation of Nonrigid Registrations with Image-Based Self-Occlusion Reasoning. *PAMI*, 32(1):87–104, 2010. 2

[15] R. Gopalan and D. Jacobs. Comparing and Combining Lighting Insensitive Approaches for Face Recognition. *CVIU*, 114(1):135–145, 2010. 3

[16] G. Hager and P. Belhumeur. Efficient Region Tracking with Parametric Models of Geometry and Illumination. *PAMI*, 20(10):1025–1039, 1998. 2

[17] D. Hoaglin, F. Mosteller, and J. Tukey. *Understanding Robust and Exploratory Data Analysis*. Wiley New York, 1983. 4

[18] P. Huber. *Robust Statistics*. Wiley, 1981. 4

[19] G. Klein and D. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *ISMAR*, 2007. 4

[20] J. Lewis. Fast Normalized Cross-Correlation. In *Vision Interface*, pages 120–123, 1995. 3

[21] A. Malti, A. Bartoli, and T. Collins. A Pixel-Based Approach to Template-Based Monocular 3D Reconstruction of Deformable Surfaces. In *ICCV*, 2011. 1, 2, 3

[22] J. Ostlund, A. Varol, D. Ngo, and P. Fua. Laplacian Meshes for Monocular 3D Shape Recovery. In *ECCV*, 2012. 2, 3, 5

[23] G. Panin and A. Knoll. Mutual Information-Based 3D Object Tracking. *IJCV*, 78(1):107–118, 2008. 1, 2, 4

[24] M. Perriollat, R. Hartley, and A. Bartoli. Monocular Template-Based Reconstruction of Inextensible Surfaces. *IJCV*, 95, 2011. 1, 2

[25] J. Pilet, V. Lepetit, and P. Fua. Retexturing in the Presence of Complex Illuminations and Occlusions. In *ISMAR*, 2007. 2, 3

[26] D. Pizarro and A. Bartoli. Feature-Based Deformable Surface Detection with Self-Occlusion Reasoning. *IJCV*, March 2012. 1, 2

[27] C. Russell, J. Fayad, and L. Agapito. Dense Non-Rigid Structure from Motion. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, pages 509–516, 2012. 2

[28] M. Salzmann. Continuous Inference in Graphical Models with Polynomial Energies. In *CVPR*, June 2013. 2

[29] M. Salzmann and P. Fua. Linear Local Models for Monocular Reconstruction of Deformable Surfaces. *PAMI*, 33(5):931–944, 2011. 5

[30] M. Salzmann and R. Urtasun. Beyond Feature Points: Structured Prediction for Monocular Non-Rigid 3D Reconstruction. In *ECCV*, 2012. 7

[31] M. Salzmann, R. Urtasun, and P. Fua. Local Deformation Models for Monocular 3D Shape Recovery. In *CVPR*, June 2008. 1, 2, 3, 7

[32] G. Scandaroli, M. Meilland, and R. Richa. Improving NCC-Based Direct Visual Tracking. In *ECCV*, 2012. 2

[33] J. Shotton, A. Fitzgibbon, M. Cook, and A. Blake. Real-Time Human Pose Recognition in Parts from a Single Depth Image. In *CVPR*, 2011. 1

[34] G. Silveira and E. Malis. Real-Time Visual Tracking Under Arbitrary Illumination Changes. In *CVPR*, 2007. 2

[35] C. Strecha, R. Fransens, and L. Van Gool. Combined Depth and Outlier Estimation in Multi-View Stereo. In *CVPR*, 2006. 2, 4

[36] A. Varol, M. Salzmann, P. Fua, and R. Urtasun. A Constrained Latent Variable Model. In *CVPR*, 2012. 2, 6

[37] S. Vicente and L. Agapito. Soft Inextensibility Constraints for Template-Free Non-Rigid Reconstruction. In *ECCV*, 2012. 1, 2

[38] P. Viola and W. Wells. Alignment by Maximization of Mutual Information. *IJCV*, 24(2):134–154, 1997. 1, 2, 4

[39] X. Wang, T. Han, and S. Yan. An HoG-LBP Human Detector with Partial Occlusion Handling. In *ICCV*, 2009. 2, 4

[40] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma. Face Recognition with Contiguous Occlusion Using Markov Random Fields. In *ICCV*, 2009. 2, 4