Joint Estimation of Deformable Motion and Photometric Parameters in Single View Video

Anna Hilsmann and Peter Eisert
Fraunhofer Heinrich-Hertz Institute
Einsteinufer 37, 10587 Berlin, Germany
{anna.hilsmann, peter.eisert}@hhi.fraunhofer.de

Abstract

In this paper we present a method for joint deformation and illumination parameter estimation from monocular image sequences exploiting direct image information. We are particularly interested in augmented reality applications, where a new texture is rendered onto a moving and deforming surface in the original video in real-time. Realistic retexturing not only requires geometric registration but also photometric parameter retrieval for convincing illusion. The contribution of this paper lies in a method for deformable surface augmentation exploiting a combination of an extended optical flow equation with a mesh-based shape and illumination prior that allows simultaneous deformation and photometric parameter estimation. Taking into account the photometric part by relaxing the brightness constancy assumption of the estimation not only allows realistic augmentation results but also improves spatial tracking.

1. Introduction

The problem of capturing non-rigid motion in monocular image sequences is of wide interest in many fields of application including medical imaging [13] or augmented reality [16, 17], where computer generated content is merged with real video. However, in real environments tracking is often influenced by varying lighting. One approach to make tracking more robust against illumination changes is to make use of illumination invariant feature points [16] or to bandpass filter the images before processing [10]. Another approach is to explicitly model the illumination changes. This not only improves geometric registration but also allows actual retrieval of illumination parameters. We are particularly interested in *retexturing* deforming surfaces in augmented reality applications. Here, photometric parameters have to be recovered to enhance the realism of the synthetic texture.

Current retexturing methods usually treat geometric and photometric parameter estimation separately [16, 17, 19, 3, 9, 22]. Some require markers to establish a shading map by intensity interpolation [19, 3, 9], others restrict the surface to consist of a limited set of colors which can be easily classified [22]. Such assumption of a-priori knowledge, however, is problematic in many applications.

In this paper we propose a method for simultaneous retrieval of deformation and illumination parameters using image-based deformation models with mesh-based deformation and illumination priors. We exploit the optical flow constraint equation instead of working with distinct features. The classical optical flow equation assumes brightness constancy, which is almost never the case in a realistic scene. Gennert and Negahdaripour [6] proposed to extend the optical flow constraint equation by a specific illumination model which allows the brightness of a scene point to vary with time to make motion estimation more robust against illumination changes. Inspired by this work, we formulate a cost function based on an extended optical flow formulation and actually retrieve both deformation and photometric parameters by incorporating a mesh-based shape and illumination model. More specifically, we model the spatial deformations in a regular 2D mesh and incorporate a photometric parameter in a third data dimension at each mesh vertex (see Figure 1).

The following section briefly reviews the existing literature on deformable tracking and augmentation. In Section 3 we propose our method for simultaneous deformation and illumination parameter estimation for deformable surfaces, before Section 4 describes shading map creation and retexturing. Section 5 reports experimental results on several data sets of deforming surfaces.

2. Related Work

In general, the literature distinguishes between feature-based [16, 17], marker-based [19] and image-based [2, 12, 21] deformable surface tracking and augmentation

methods in monocular image sequences.

Scholz and Magnor [19] use color-coded patterns and a-priori knowledge about surface connectivity for tracking of garments in single-view sequences. They determine shading maps by removing the color markers used for tracking and interpolating the deleted regions with a surface reconstruction method for height fields. White and Forsyth [22] presented a similar method for retexturing non-rigid objects from a single viewpoint using color markers. They limited their method to recover irradiance to screen printing techniques with a finite number of colors. However, the assumption of such a-priori knowledge limits the applicability for arbitrary video sequences.

Pilet et al. [17] proposed a feature-based real-time method for deformable object detection and tracking that uses a wide baseline matching algorithm and deformable meshes. They estimate the irradiance of the surface separately by warping the reference image to the current frame and estimating the luminance ratio between both images.

Generally, image-based methods yield more accurate results in non-rigid deformation estimation because they exploit the entire image instead of distinct points. Bartoli and Zisserman [2] presented an optical flow based approach that used radial basis functions to regularize the flow field. They iteratively insert new center-points for the radial basis functions based on examination of the error image after each iteration. The number of centers grows until the algorithm converges. Recently, Gay-Bellille *et al.* [5] proposed a direct method to estimate deformable motion under self-occlusions by establishing occlusion maps and penalizing a variation in the spatial warp derivative along some direction to prevent mesh-foldings and cope with self-occlusions. Hilsmann and Eisert [10] utilized the idea of mesh-shrinking in a simpler registration scheme by exploiting topological relationship of the warp. Unfortunately, robustness against varying illumination as well as illumination recovery and retexturing is not addressed in these image-based papers.

While in marker- or feature-based methods illumination recovery is crucial for augmentation of the images rather than estimating the correct warp –assuming markers or features are illumination invariant– direct methods usually minimize an error function based on the intensity differences between the aligned images. Therefore, these methods are sensitive to illumination variations. Gennert and Negahdaripour [6] were among the first to propose a direct method robust to illumination variations. They assumed that the brightness at time $t + \delta t$ is related to the brightness at time t through a set of parameters that can be estimated from the image sequence. Several other researchers [14, 8, 1, 20] have exploited their ideas to make tracking more robust against lighting changes. Bartoli [1] proposed a dual inverse compositional algorithm to estimate a homography and an affine photometric registration. The registration results are improved by the photometric registration but only global changes are modeled and thus specular reflections or local shadows are not taken into consideration. Silveira and Malis [20] model local illumination changes to make a homography estimation more robust against generic illumination changes. Pizarro and Bartoli [18] proposed to transform images into a 1D shadow invariant space to achieve direct image registration in the presence of even sharp shadows.

Inspired by these works we utilize the idea to extend the optical flow constraint by a specific illumination model. However, in contrast to the aforementioned work we not only use it as a correction factor for spatial registration but also actually retrieve photometric parameters for convincing retexturing purposes. Furthermore, we model both deformable motion and local photometric changes instead of global illumination changes. The contribution of this paper lies in the combination of the extended optical flow constraint of [6] with a combined mesh-based shape and illumination prior to estimate deformable motion and photometric parameters simultaneously. We model the photometric changes and deformation parameters explicitly in one mesh-based model. In order to keep computational cost low we use a hierarchical framework with a Gauss-Newton approximation of the extended optical flow constraint on each level. This lets us establish an error function that can be efficiently solved in closed-form on each resolution level, which makes it suitable for real-time augmented reality applications. Small errors due to linearization are compensated using an analysis-by-synthesis approach.

3. Joint Motion and Illumination Estimation

Generally, our aim is to recover spatial deformations and photometric changes in the image plane, as a 3D reconstruction of a deforming surface is an ill-posed problem, especially for real-time applications. Without a 3D reconstruction of the surface we cannot explicitly model the illumination. However, we can recover its impact on the intensity of a scene point. Hence, we describe the spatial deformation of the surface in the image plane in a dense pixel displacement field $\mathbf{D}(\mathbf{x})$ and the photometric changes in a dense pixel intensity scale field $\tilde{\mathcal{S}}(\mathbf{x})$, where $\mathbf{x} = [x, y]^T$ represents the pixel position. As our intention is to *retexture* the surface and not to render a new synthetic view, the information in the image plane is sufficient. In order to parameterize these fields in a parameter vector $\boldsymbol{\Theta}$, we introduce a mesh-based deformation and illumination model. The formulation of this model is described in Section 3.2 before Section 3.3 describes our method for simultaneous spatial deformation and photometric parameter estimation using direct image

information. First, we describe the hierarchical analysis-by-synthesis approach used for parameter estimation.

3.1. Hierarchical Analysis-by-Synthesis Approach

Intensity-based differential techniques, which estimate the motion only between two successive frames, often suffer from drift because they accumulate errors indefinitely. This limits their effectiveness when dealing with long video sequences. To avoid error accumulation we make use of an analysis-by synthesis approach. In this approach, we use the previously estimated parameter set $\hat{\mathbf{\Theta}}_{n-1}$ to generate a synthetic image of the previous frame $\mathcal{I}_{n-1}$ from a model image $\mathcal{I}_0$. The new parameters $\hat{\mathbf{\Theta}}_n$ are then estimated from this synthetic previous frame $\hat{\mathcal{I}}_{n-1}$ to the current frame $\mathcal{I}_n$. This allows for recovery from small inaccuracies during parameter estimation because it assures that no misalignment of the model and the previous frame occurs. However, if photometric changes at a scene point are not considered in the parameter estimation and brightness is not adapted in the warped model, the intensity difference between the synthetic previous frame and the current frame increases and causes errors in the motion estimation. To avoid these errors it is essential to both warp the model image spatially but also adapt the image brightness locally according to the current photometric parameters and estimate both deformation and photometric parameter changes from this model image to the current frame.
The analysis-by-synthesis approach allows us to estimate the parameter vector $\hat{\mathbf{\Theta}}_n$ in a hierarchical scheme where an initial approximation for the parameters between $\mathcal{I}_n$ and $\hat{\mathcal{I}}_{n-1}$ is computed from low-pass filtered and sub-sampled versions of the current frame $\mathcal{I}_n$ and the synthetic previous frame $\hat{\mathcal{I}}_{n-1}$. The estimated parameters are then used to generate a new synthetic version of the previous frame $\hat{\mathcal{I}}_{n-1}$ on the next resolution level. The procedure is repeated at higher resolutions, each time yielding a more and more accurate parameter set $\hat{\mathbf{\Theta}}_n$.

3.2. Combined Mesh-Based Shape and Illumination Model

We parameterize the fields $\mathbf{D}(\mathbf{x})$ and $\tilde{\mathcal{S}}(\mathbf{x})$ with a deformable model that is presented as a planar triangulated regular 2D mesh with K vertices $\mathbf{v}_k, (k = 1...K)$. The spatial position of each vertex $\mathbf{v}_k$ in the image is given by its image coordinates $\mathbf{x}_k = [x_k, y_k]^T$. We model the brightness scale in a third photometric parameter μ_k at each vertex (see Figure 1). Each vertex $\mathbf{v}_k$ then has three parameters $\theta_k = (\delta\mathbf{v}_k, \delta\mu_k)$, i.e. the vertex displacements in x- and y-direction $\delta\mathbf{v}_k = (\delta v_{kx}, \delta v_{ky})$ and the photometric parameter $\delta\mu_k$ describing the deviation of the brightness scale from identity (see Figure 1). The resulting parameter vector $\mathbf{\Theta}$ modeling the deformable motion and illumination changes

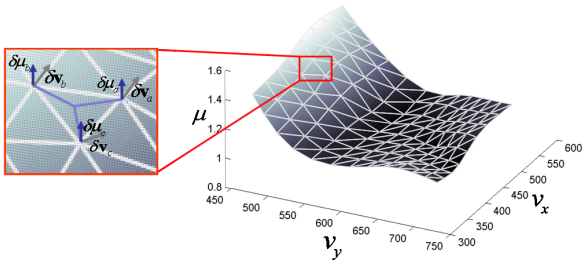

Figure 1. Illustration of the mesh-based deformation and illumination parameters.

is then given by concatenating the parameters of each vertex

$$\mathbf{\Theta} = (\delta v_{x_1}, ..., \delta v_{x_K}, \delta v_{y_1}, ..., \delta v_{y_K}, \delta\mu_1, ..., \delta\mu_K)^T . \quad (1)$$

In the following we present a parameterization of the deformation field $\mathbf{D}(\mathbf{x})$ and the brightness scale field $\tilde{\mathcal{S}}(\mathbf{x})$ by the parameter vector $\mathbf{\Theta}$. For clarity and spatial reasons we describe a parametrization which uses affine interpolation between the vertex positions. Note, that higher order interpolation, like e.g. B-splines or thin-plate splines, are also possible.

If a pixel $\mathbf{x}_i$ is surrounded by the three vertices $\mathbf{v}_a, \mathbf{v}_b, \mathbf{v}_c$, with indices a, b, c and $\beta_a, \beta_b, \beta_c$ are the three corresponding barycentric coordinates ($\beta_a + \beta_b + \beta_c = 1, 0 \leq \beta_{a,b,c} \leq 1$), it can be represented by the weighted sum of its enclosing vertices. Also, the intensity scale field $\tilde{\mathcal{S}}(\mathbf{x_i})$ is represented by the brightness scale parameters of these vertices.

$$\begin{aligned} \mathbf{x}_i &= \sum_{i\in\{a,b,c\}} \beta_i \mathbf{v}_i \\ \tilde{\mathcal{S}}(\mathbf{x}_i) &= \sum_{i\in\{a,b,c\}} \beta_i \mu_i \end{aligned} \quad (2)$$

The pixel displacement field $\mathbf{D}(\mathbf{x})$ and the deviation of brightness scale field from identity scaling $\mathcal{S}(\mathbf{x}) = \tilde{\mathcal{S}}(\mathbf{x}) - 1$ can then be parameterized in the same way:

$$\begin{aligned} \mathbf{D}(\mathbf{x}_i) &= \sum_{i\in\{a,b,c\}} \beta_i \delta\mathbf{v}_i \\ \mathcal{S}(\mathbf{x}_i) &= \sum_{i\in\{a,b,c\}} \beta_i \delta\mu_i \end{aligned} \quad (3)$$

where $\delta\mathbf{v_i}$ are the three vertex displacements of the enclosing triangle. $\delta\mu_i$ are the deviations of the photometric parameters from identity scaling ($\delta\mu_i = \mu_i - 1$) of the three enclosing vertices respectively. In the following we explain how we estimate the parameters $\mathbf{\Theta}$ exploiting direct image information.

3.3. Joint Parameter Estimation using Extended Optical Flow

Generally, estimating the parameter vector $\mathbf{\Theta}$ leads to minimizing a cost function that consists of two terms:

$$\hat{\mathbf{\Theta}} = \text{argmin}_{\mathbf{\Theta}} \left(E_D(\mathbf{\Theta}) + \lambda^2 E_S(\mathbf{\Theta}) \right) \tag{4}$$

where $E_D(\mathbf{\Theta})$ is the *data term* and $E_S(\mathbf{\Theta})$ represents prior knowledge on the shape and illumination model and is often called the *smoothness term*. λ is a regularization parameter which weights the influence of this prior knowledge against fitting to the data term.

Data Term

The data term E_D is derived by exploiting the optical flow equation in its original form which assumes brightness constancy between two successive image frames [11]:

$$\mathcal{I}\left(\mathbf{x} + \mathbf{D}(\mathbf{x}), t + \delta t\right) = \mathcal{I}\left(\mathbf{x}, t\right) \tag{5}$$

where $\mathcal{I}\left(\mathbf{x}, t\right)$ is the image intensity at pixel $\mathbf{x}$ and time t and $\mathbf{D}(\mathbf{x})$ is the displacement vector field at pixel $\mathbf{x} = [x, y]^T$. This formulation assumes that an image pixel representing an object point does not change its brightness value from time t to time $t+\delta t$ and differences between successive frames are due to geometric deformation only. However, this assumption is almost never valid for natural scenes. We relax the optical flow constraint equation allowing for multiplicative deviations from brightness constancy.

$$\mathcal{I}\left(\mathbf{x} + \mathbf{D}(\mathbf{x}), t + \delta t\right) = \tilde{\mathcal{S}}\left(\mathbf{x}\right) \cdot \mathcal{I}\left(\mathbf{x}, t\right) \tag{6}$$

The data term E_D then is given by the Sum of Squared Differences

$$E_D = \sum_{\mathbf{x} \in \mathcal{R}} \left(\mathcal{I}\left(\mathbf{x} + \mathbf{D}(\mathbf{x}), t + \delta t\right) - \tilde{\mathcal{S}}\left(\mathbf{x}\right) \cdot \mathcal{I}\left(\mathbf{x}, t\right) \right)^2. \tag{7}$$

where $\mathcal{R}$ denotes the region of interest in the image. On each pyramid level we approximate the above equation with a Gauss-Newton approximation by applying a first order Taylor expansion [15]:

$$E_D = \sum_{\mathbf{x} \in \mathcal{R}} \left(\nabla \mathcal{I}\left(\mathbf{x}\right) \cdot \mathbf{D}(\mathbf{x}) + \frac{\partial \mathcal{I}\left(\mathbf{x}\right)}{\partial t} - \mathcal{S}\left(\mathbf{x}\right) \cdot \mathcal{I}\left(\mathbf{x}\right) \right)^2. \tag{8}$$

Due to the hierarchical estimation scheme described in Section 3.1 the linearization is valid over a wide range in the image. Without any additional information equation (8) is still an under-determined problem. To regularize the optical flow field $\mathbf{D}(\mathbf{x})$ and the illumination scale field $\mathcal{S}(\mathbf{x})$ we incorporate the mesh-based motion and illumination model represented by the parameter vector $\mathbf{\Theta}$. This leads to a cost function of the following form:

$$E_D(\mathbf{\Theta}) = \left\| \mathbf{J} \cdot \mathbf{\Theta} - \mathbf{r} \right\|^2 \tag{9}$$

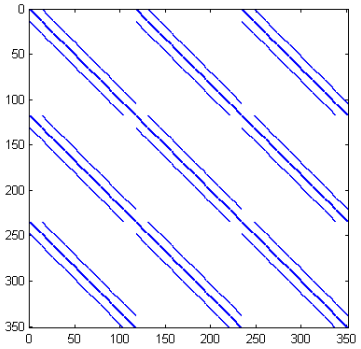

Figure 2. Example of $\mathbf{H} \approx \mathbf{J}^T\mathbf{J}$. The sparse structure of the matrix is exploited to solve the normal equations of the Linear Least Squares.

where $\mathbf{J}$ is the sparse Jacobian matrix of the Gauss-Newton approximation that is composed of three $n \times K$ submatrices:

$$\mathbf{J} = \left(\mathbf{J}_{I_x} \mathbf{J}_{I_y} \mathbf{J}_I \right) \tag{10}$$

n denotes the number of pixels in region $\mathcal{R}$ selected for contribution. We need both high and low image gradients, as for the geometric part of the equation high image gradients are needed whereas for the photometric part also low image gradients are exploited because brightness scale is best evaluated in homogeneous regions. For the parametrization given in equation (3) $\mathbf{J}_{\mathcal{I}_x}$, $\mathbf{J}_{\mathcal{I}_y}$ and $\mathbf{J}_{\mathcal{I}}$ are sparse matrices with the following non-zero entries in the i^{th} row and the a^{th}, b^{th} and c^{th} column, where a, b, c are the three indices of vertices surrounding pixel $\mathbf{x}_i$:

$$\begin{aligned} \mathbf{J}_{\mathcal{I}_x}(i, a) &= \beta_a \cdot \mathcal{I}_x(\mathbf{x}_i) \\ \mathbf{J}_{\mathcal{I}_y}(i, a) &= \beta_a \cdot \mathcal{I}_y(\mathbf{x}_i) \\ \mathbf{J}_{\mathcal{I}}(i, a) &= \beta_a \cdot \mathcal{I}(\mathbf{x}_i), \end{aligned} \tag{11}$$

for b, c respectively. The sparse structure of the Gauss-Newton approximation of the Hessian $\mathbf{H} \approx \mathbf{J}^T\mathbf{J}$, which is used to solve the normal equations of the Linear Equation System, is depicted in Figure 2. Note, that for a higher order interpolation, like B-splines or thin-plate splines $\mathbf{J}$ can be formed in a similar way and the resulting equation system is still linear in the mesh parameters, but the matrix is less sparse. The $K \times 1$ parameter vector $\mathbf{\Theta}$ is given by equation (1) and $\mathbf{r}$ is an $n \times 1$ vector given by:

$$\mathbf{r} = \left(\frac{\partial \mathcal{I}(\mathbf{x}_1)}{\partial t}, ..., \frac{\partial \mathcal{I}(\mathbf{x}_n)}{\partial t} \right)^T \tag{12}$$

Smoothness Term

In order to incorporate prior knowledge on the smoothness of the deformation and illumination fields we penalize the discrete second derivative of the motion and illumination parameters in the mesh in a smoothness term E_S. In matrix notation, this can be expressed as:

$$E_S(\mathbf{\Theta}) = \left\| \tilde{\mathbf{L}} \cdot \mathbf{\Theta} \right\|^2 \tag{13}$$

Figure 3. Cloth Retexturing. Original images and synthetic textures. Note that spatial deformation is purely 2D and 3D impression comes from shading properties (Cloth1 and Cloth2 Sequences).

where $\tilde{\mathbf{L}} = (\lambda_d \mathbf{L}, \lambda_d \mathbf{L}, \lambda_\mu \mathbf{L})$ is composed by concatenating three times the scaled *Laplacian Matrix* $\mathbf{L}$ of the mesh [4], one for each vertex parameter. λ_d and λ_μ are needed to weight the smoothing terms of the deformation against the smoothing terms of the brightness scale[1]. The scaled *Laplacian Matrix* $\mathbf{L}$ is a $K \times K$ matrix with one row and one column for each vertex and $\mathbf{L}(i,j) = w_i$ if vertex i and j are connected and $\mathbf{L}(i,i) = 1$. All other entries are set to zero. If all neighbors shall contribute the same to vertex $\mathbf{v}_i$ the weight w_i is set to $w_i = \frac{-1}{|\mathcal{N}_i|}$, where $|\mathcal{N}_i|$ denotes the numbers of neighbors vertex $\mathbf{v}_i$ is connected to. In order to give closer neighbors a higher influence on vertex $\mathbf{v}_i$ than neighbors with a larger distance, we use a *distance weighted scaling* which additionally weights each neighbor according to its distance to vertex $\mathbf{v_i}$. In this case, $w_i = -\frac{1/d_{i,j}}{\sum_{n \in \mathcal{N}_i} 1/d_{i,n}}$, where $d_{i,j}$ denotes the distance between the two vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ and $\mathcal{N}_i$ denotes the neighborhood of $\mathbf{v}_i$.
$E_S(\mathbf{\Theta})$ penalizes the discrete second derivative of the mesh and regularizes the optical flow field in addition to the mesh-based motion model itself, especially in case of lack of information in the data due to e.g. homogeneous regions with low image gradient in the surrounding triangles. One important property of using direct image information is that less image information, i.e. small image gradients, in a region automatically leads to a higher local weighting of the smoothness constraint in the resulting equation system.

The resulting quadratic cost function can then be minimized directly on each iteration level using Linear Least Squares methods.

$$\hat{\mathbf{\Theta}} = \text{argmin}_{\mathbf{\Theta}} \left\| \begin{pmatrix} \mathbf{J} \\ \lambda \tilde{\mathbf{L}} \end{pmatrix} \mathbf{\Theta} - \begin{pmatrix} \mathbf{r} \\ \mathbf{0} \end{pmatrix} \right\|^2 \qquad (14)$$

A short discussion on the selection of the regularization parameter will be found in Section 5.

[1]This is due to the different scaling of the pixel displacement and the photometric parameter as the former is additive while the latter is multiplicative.

4. Texture Replacement

Under the assumption that the reflectance of the new texture is the same as the original texture we can now use the estimated deformation and shape parameters to retexture a deforming surface in the video sequence with correct deformation and lighting. The vertex displacements $\delta \mathbf{v}_k$ are used to spatially warp a new synthetic texture image $\mathcal{T}(\mathbf{x})$ such that the original and the synthetic texture $\mathcal{T}_{warp}(\mathbf{x})$ are geometrically registered. A shading map $\tilde{\mathcal{S}}(\mathbf{x})$ is established as follows. For each pixel $\mathbf{x}_i$ belonging to a triangle with brightness scale parameters μ_a, μ_b, μ_c at the surrounding vertices the shading map is

$$\tilde{S}(\mathbf{x}_i) = \sum_{i \in \{a,b,c\}} \beta_i \mu_i \qquad (15)$$

This interpolation scheme corresponds to Gouraud shading. Again, higher order interpolation is possible. The intensities of the geometrically registered synthetic texture image $\mathcal{T}_{warp}(\mathbf{x})$ are then multiplied with the shading map to generate a new synthetic texture image $\mathcal{T}_{synth}(\mathbf{x})$ with correct shading and illumination properties (see Figure 3). When dealing with color images, we proceed as described for all color channels. Saturations and highlights in one of the color channel in the original sequence make the estimation of the illumination scale parameter unreliable. We treat these cases by thresholding the resulting color value of the synthetic image to its maximum possible value. The resulting synthetic image is still perceptually correct (see last row of Figure 7). Figure 7 shows examples of shading maps in the right column and retexturing results with only geometric information (second column) and both geometric and photometric information (third column).

5. Experimental Validation

We applied our method to several real video sequences with a resolution of 1024×768 pixels and a frame rate of 25 fps showing deforming surfaces, in particular pieces of paper (see Figure 7) and cloth (see Figure 3 and 7 last row), under varying illumination conditions. We evaluate

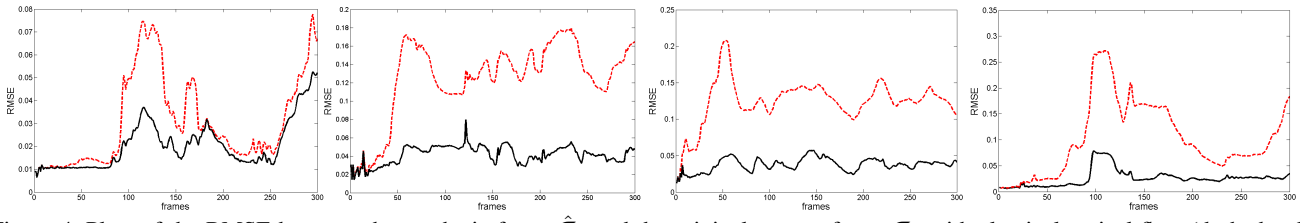

Figure 4. Plots of the RMSE between the synthetic frame $\hat{\mathcal{I}}_n$ and the original current frame $\mathcal{I}_n$ with classical optical flow (dashed red) and our method (solid black) for $\lambda = 0.5$. The RMSE is significantly reduced when illumination is explicitly considered. Left to right: Picasso, Flower1, Flower2, Art Sequence.

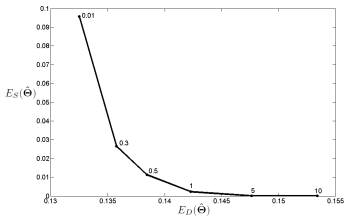

Figure 5. A typical L-curve

the registration results based on the Root Mean Squared Error (RMSE) between the synthetic image $\hat{\mathcal{I}}_n$ and the actual frame $\mathcal{I}_n$ computed over all image pixels in the mesh region and additionally on manually labeled ground truth points. We chose these two methods because the RMSE between two images is affected by both geometric and photometric registration errors whereas the registration error of ground truth points separates the geometric registration error from the photometric part. Additionally, we used the estimated deformation and illumination parameters to establish shading maps and produce augmented versions of several video sequences which are best evaluated by visual inspection.
To this end we use 4 levels of resolution and experiments with synthetic image sequences with this hierarchical scheme showed that it is able to estimate displacements of up to 25 pixels between two frames with a mean error of 0.2 pixels. On a 2.4 GHz Pentium 4 based system the parameter estimation takes about 40 ms, leading to a frame rate of about 25 fps.
In the following we will describe the different experiments with real videos after a short discussion on the selection of the regularization parameter λ.

Selection of Regularization Parameter. The influence of the regularization parameter λ was analyzed using the L-curve [7] which plots $E_D(\hat{\mathbf{\Theta}})$ against $E_S(\hat{\mathbf{\Theta}})$ for different values of λ . The L-curve (see Figure 5) is basically made up of two parts which correspond to the oversmoothed and undersmoothed solutions. The more horizontal part corresponds to the solution where the regularization parameter is too large and the solution is dominated by the regularization errors. The more vertical part corresponds to solutions where the regularization parameter is too small and the solution is dominated by the data error. We use the value of λ at the L-shaped corner of the curve, i.e. the point of maximum curvature, which represents the best balance between both sides. Based on the analysis of the L-curve of several video sequences we used $\lambda = 0.5$ in our experiments.

RMSE Based Registration Evaluation. To evaluate the registration results, we calculate the Root Mean Squared Error (RMSE) between the synthetic image $\hat{\mathcal{I}}_n$, generated from the geometrical and photometrical parameters $\hat{\mathbf{\Theta}}_n$, and the original current frame $\mathcal{I}_n$ over all image pixels in the mesh region $\mathcal{R}$ for several video sequences. Figure 4 depicts RMSE plots of four sequences of deforming surfaces. It compares the results with our approach (black solid plots) to the classical optical flow method (dashed red plots). The peaks in the plots for the classical approach result from registration errors in situations where the illumination changed or shadow were cast onto the object. In case of the Picasso Sequence (first plot in Figure 4) the RMSE between the registered model and the original current frame without illumination consideration increases in these situations but also decreases again when the brightness of the surface changes to the initial brightness. Here, the analysis-by-synthesis approach is able to recover from the inaccuracies due to illumination change and the increased RMSE in the plots of the classical approach arouse from the different illuminations in the model and the current frame rather than from inaccurate tracking. However, in the other examples the changing illumination in the sequence leads to an increased intensity difference between model and original current frame. This results in an error in the spatial warp if the illumination change is not taken into account. Here, considering illumination variations in parameter estimation and adapting the brightness in the warped model improves the spatial tracking results. Table 1 sums up the mean RMSE over the entire sequence for nine sequences. It shows that taking illumination parameters into account significantly reduces the mean RMSE over the entire sequence by up to 74%.

Sequence	RMSE with classical OF	RMSE with our approach	reduction
Picasso	0.0306	0.0204	33.33%
Flower1	0.1364	0.0415	69.57%
Flower2	0.1245	0.0376	69.80%
Art	0.1016	0.0258	74.61%
Flower3	0.1286	0.0385	70.06%
Shirt	0.0630	0.0405	35.71%
Pattern	0.0632	0.0213	66.30%
Cloth1	0.0877	0.0443	49.49%
Cloth2	0.0976	0.0513	47.44%

Table 1. Comparison of the average RMSE over the entire video sequence with our approach and classical optical flow.

Point Based Registration Evaluation. We manually labeled 141 and 84 prominent feature points respectively in every 50th frame of two test sequences which serve as ground truth points. We then warped the ground truth points of the reference frame with the geometric deformation parameters of these frames. The mean difference between the estimated positions and the manually labeled ground truth position describes the geometric registration error. This additional evaluation approach is chosen to evaluate geometric registration accuracy separately from photometric registration. Table 2 shows the mean spatial registration error over the sequence with our approach compared to classical optical flow. It shows that taking illumination parameters into account during tracking significantly improves geometric registration. We can reduce the mean distance between the estimated and the ground truth position by approximately 40%. Figure 6 shows frame 250 of the Art Sequence with the real and estimated positions of the ground truth points overlayed. Here, classical optical flow leads to geometric misregistration in contrast to our method.

Retexturing Results. Figure 7 shows augmentation results under different lighting conditions using the estimated illumination parameters to establish a shading map. The left images show the original frame of the sequence while the middle images depict the augmentation result without and with illumination consideration. The right images show the shading map. Figure 3 shows further retexturing of two different kinds of cloth. The left example shows thick cloth with very smooth deformations while the right example shows cloth that produces small crinkles and creases. These examples demonstrate how crucial illumination recovery is for convincing texture augmentation of deforming surfaces. The addition of realistic lighting increases the perception of spatial relations between the real and virtual objects. Note that spatial deformation is purely 2D and the 3-dimensional impression comes from shading.

Sequence	distance with classical OF	distance with our approach	reduction
Art	1.70	0.98	42.37%
Pattern	3.10	1.82	41.01%

Table 2. Comparison of the average distance in pixels between ground truth and estimated position of feature points with our approach and classical optical flow.

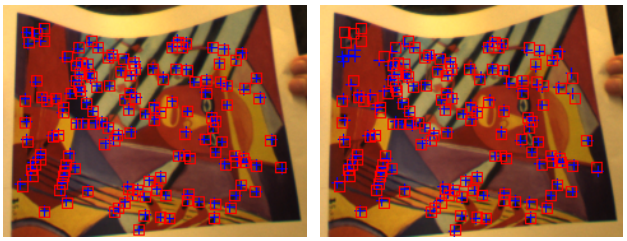

Figure 6. Estimated (blue crosses) and ground truth (red squares) position of the feature points with (left) and without (right) illumination consideration in frame 250 of the Art Sequence for $\lambda = 0.5$

6. Conclusion

We presented a method for deformable surface tracking and augmentation in single view sequences with simultaneous motion and illumination parameter estimation. Our method uses an optical-flow-based approach and incorporates a mesh-based deformation and illumination model. We explicitly model illumination changes in a third photometric parameter at each mesh vertex. We showed that considering illumination parameters in an optical flow-based analysis-by-synthesis approach reduces the RMSE by up to 74% and the differences between estimated and real positions of manually labelled feature points by 40 % compared to the classical optical flow constraint. Furthermore, our approach not only improves geometric registration but also retieves photometric parameters which allow the creation of a shading map for convincing texture augmentation. Our method is currently limited to video sequences with smooth deformations and shading due to the smoothness terms. Future work will e.g. concentrate on modeling discontinuities in both the warp, e.g. in case of folding, and in the shading map, e.g. in case of sharp shadows.

References

[1] A. Bartoli. Groupwise geometric and photometric direct image registration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(12):2098–2108, 2008.

[2] A. Bartoli and A. Zisserman. Direct estimation of non-rigid registrations. In *Proc. British Machine Vision Conf. (BMVC 2004)*, London, UK, 2004.

[3] D. Bradley, G. Roth, and P. Bose. Augmented clothing. In *Proc. Graphics Interface (GI 2005)* [3].

[4] D. M. Cvetkovic, M. Doob, and H. Sachs. *Spectra of Graphs: Theory and Applications*. Wiley, 1998.

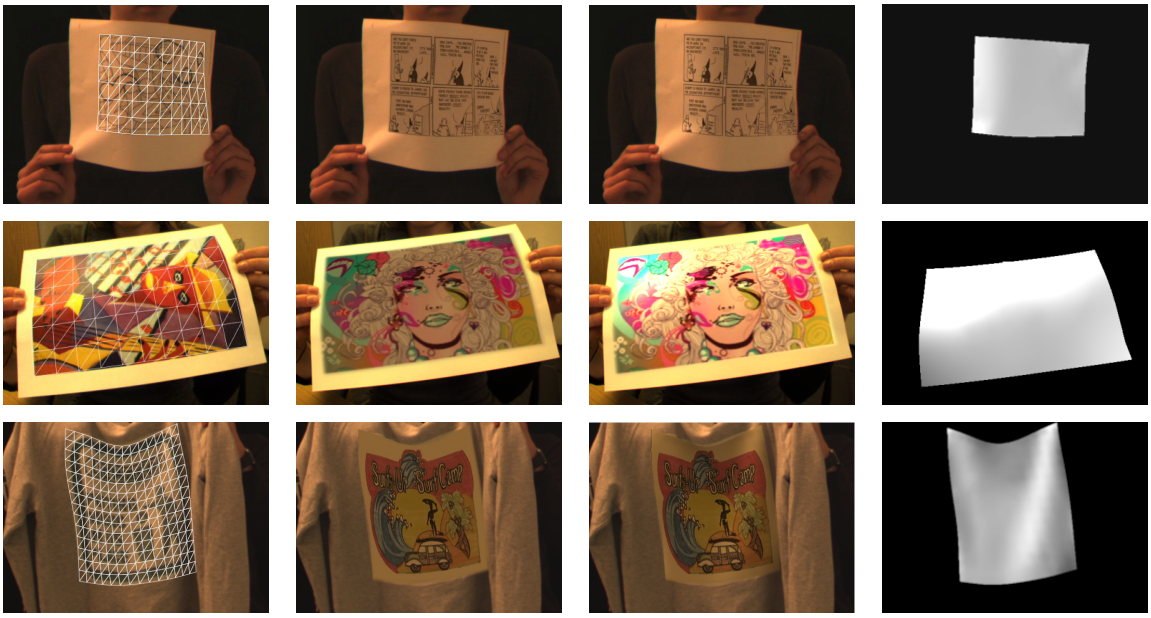

Figure 7. Augmentation results. Left to right: original image, augmentation results without illumination considering, augmentation results with our approach, shading map. The addition of real lighting increases the perception that the paper is truly exhibiting the virtual texture.

[5] V. Gay-Bellile, A. Bartoli, and P. Sayd. Direct estimation of non-rigid registrations with image-based self-occlusion reasoning. pages 1–6, 2007.

[6] M. A. Gennert and S. Negahdaripour. Relaxing the brightness constancy assumption in computing optical flow. Technical report, Cambridge, MA, USA, 1987.

[7] P. Hansen. The use of the l-curve in the regularization of discrete ill-posed problems. *SIAM Journ. Sci. Comp.*, 34:561–580, 1992.

[8] H. W. Haussecker and D. J. Fleet. Computing optical flow with physical models of brightness variation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):661–673, 2001.

[9] A. Hilsmann and P. Eisert. Optical flow based tracking and retexturing of garments. In *Proc. International Conference on Image Processing (ICIP)*, San Diego, USA, Oct. 2008.

[10] A. Hilsmann and P. Eisert. Tracking deformable surfaces with optical flow in the presence of self occlusions in monocular image sequences. In *CVPR Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment*, Anchorage, USA, June 2008.

[11] B. Horn and B. G. Schunck. Determining optical flow. Technical report, Cambridge, MA, USA, 1980.

[12] J. Lim and M.-H. Yang. A direct method for modeling non-rigid motion with thin plate spline. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2005)*, pages 1196–1202, 2005.

[13] D. Metaxas. *Physics-Based Deformable Models: Applications to Computer Vision, Graphics, and Medical Imaging*. Kluwer Academic Publishers, 1996.

[14] C. Nastar, B. Moghaddam, and A. Pentland. Generalized image matching: Statistical learning of physically based deformations. In *Proc. of European Conf. on Computer Vision (ECCV 1996)*, pages 589–598, Cambridge, UK, 1996.

[15] S. Negahdaripour and C.H.Yu. A generalized brightness change model for computing optical flow. In *Proc. Int. Conf. on Computer Vision (ICCV 1993)*, pages 2–11, Berlin, Germany, 1993.

[16] J. Pilet, V. Lepetit, and P. Fua. Augmenting deformable objects in real-time. In *Int. Symposium on Mixed and Augmented Reality*, Vienna, Austria, October 2005.

[17] J. Pilet, V. Lepetit, and P. Fua. Fast non-rigid surface detection, registration and realistic augmentation. *Int. Journal of Computer Vision*, 76(2), Feb. 2008.

[18] D. Pizarro and A. Bertoli. Shadow resistant direct image registration. In *Proc. of the 15th Scandinavian Conference on Image Analysis*, Aalborg, Denmark, June 2007.

[19] V. Scholz and M. Magnor. Texture replacement of garments in monocular video sequences. In *Rendering Techniques 2006: Eurographics Symposium on Rendering*, pages 305–312, June 2006.

[20] G. Silveira and E. Malis. Real-time visual tracking under arbitrary illumination changes. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2007)*, pages 1–6, 2007.

[21] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2001)*, Hawaii, USA, 2001.

[22] R. White and D. A. Forsyth. Retexturing single views using texture and shading. In *Proc. European Conf. on Computer Vision (ECCV 2006)*, volume 3954 of *Lecture Notes in Computer Science*, pages 70–81. Springer, 2006.