SPATIO-TEMPORAL REGISTRATION TECHNIQUES FOR RELIGHTABLE 3D VIDEO

Naveed Ahmed[1], *Christian Theobalt*[1], *Marcus Magnor*[2], *Hans-Peter Seidel*[1]*

[1]MPI Informatik, Saarbruecken, Germany [2]Braunschweig Technical University, Germany

ABSTRACT

By jointly applying a model-based marker-less motion capture approach and multi-view texture generation 3D Videos of human actors can be reconstructed from multi-view video streams. If the input data were recorded under calibrated lighting, the texture information can also be used to measure time-varying surface reflectance. This way, 3D videos can be realistically displayed under novel lighting conditions. Reflectance estimation is only feasible if the multi-view texture-to-surface registration is consistent over time. In this paper, we propose two image-based warping methods that compensate registration errors due to inaccurate model geometry and shifting of apparel over the body.

Index Terms— 3D video, spatio-temporal registration, image processing, machine vision, computer graphics

1. INTRODUCTION

The commitment to an adaptable a priori body model enables us to jointly use a marker-less motion capture approach and a multi-view texture generation method to reconstruct free-viewpoint videos of human actors from only a handful of multi-view video streams [1]. If the input video footage is recorded under calibrated lighting conditions we can exploit the fact that the person moves relatively to the cameras and light sources in order to reconstruct a dynamic surface reflectance description for the model [2]. This description consists of a parametric BRDF for each surface texel and a normal with time-varying direction. The joint description of time-varying scene geometry and appearance enables us to realistically display 3D videos also under novel lighting conditions. For reflectance estimation, it is essential that the multi-view texture-to-surface registration is consistent over time. In this paper, we present two methods based on image warping to correct for the two most important sources of texture registration errors.

The first source of error are inaccuracies in the body model's geometry that can be compensated by warping the input video frames prior to texture generation. In contrast to our image-based approach, related methods from the literature typically deform the geometry of the model to optimize model-to-image registration. For instance, [3, 4] deform model geometry from input images by jointly optimizing multi-view silhouette- and photo-consistency. In a similar line of thinking, [5] jointly employs silhouette and stereo constraints to deform scene geometry from images. The main advantage of our image warping method is that it is part of the preprocessing and thus time-varying geometry variations don't need to be encoded in the 3D video streams.

The second source of misregistrations is the shifting of apparel while the person is moving. This also has to be compensated prior to reflectance estimation. We detect the motion of the apparel by computing optical flow fields in the texture space. Subsequently, the texture coordinates for lookup are warped according to the textile motion. To our knowledge, this is the first method in the literature to attack this shifting problem.

Our results show that it is feasible to use purely image-based approaches to compensate the most prominent multi-view registration errors. Consequently, they can be handled in a preprocessing stage which enables us to stick to a very compact relightable 3D video data format.

The paper continues in Sect. 2 with a brief review of the basics of relightable 3D video reconstruction and important preprocessing steps. Sect. 3 explains the flow-based image warping technique which is used during both the geometry error compensation, Sect. 4, as well as the cloth shift detection, Sect. 5. Results are shown in Sect. 6 and the paper concludes with an outlook to future work in Sect. 7.

2. PRELIMINARIES

Our recording setup comprises of eight megapixel video cameras that are placed in an approximately circular arrangement around the moving subject, as well as two calibrated light sources. During relightable free-viewpoint video acquisition we record two types of multi-view video sequences for each person and each type of apparel [2]. One so-called reflectance estimation sequence (RES) is captured in which the person rotates on the spot while attaining an approximately static posture (achievable by rotating in very small steps). This sequence is used for per-texel BRDF estimation. Also, several dynamic scene sequences (DSS) are recorded to capture arbitrary human motion. From these sequences, the actual relightable free-viewpoint videos are reconstructed, and the

*This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

second component of the reflectance model, the time-varying normal field, is estimated. We apply the marker-less optical motion estimation and shape matching scheme described in [1] to make a kinematic body model with a single-skin surface follow the motion of the actor in each of the input video streams.

Given the moving geometry, all input video frames and all corresponding data required for reflectance estimation (e.g. image samples, normals, visibility information, light vectors) are transformed into sequences of textures. To this end, we parameterize the model's surface over a 2D square. For the BRDF and time-varying normal estimation we need a parameterization with minimal surface distortion. To achieve this, we employ a parameterization (Parameterization A) that leaves the mesh boundary free and results in fairly uniform distribution of samples [6]. For the purpose of cloth shift detection, on the other hand, we prefer a parameterization (Parameterization B) with a fixed square boundary, Fig. 1.

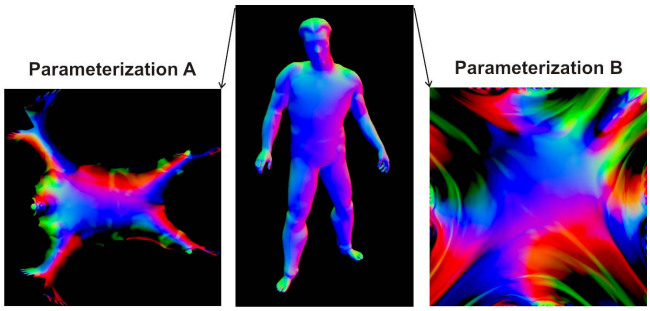

Fig. 1. Human body model and the corresponding texture parameterizations (colors=normals encoded in RGB).

3. IMAGE WARPING

A generic GPU-assisted image-warping method is used in either of the two subsequently described multi-view registration approaches. The method deforms an input image, I_M, in such a way that it optimally overlaps with a reference image I_R. The warping operation works as follows:

A regular 2D triangle mesh T with n vertices $\{v_1, \ldots, v_n\}$ is superimposed over I_M. The optical flow between I_R and I_M is computed by means of an appropriate optical flow method, for instance the hierarchical Lucas-Kanade [7] technique. The so-created flow field describes a displacement for each pixel in I_M that brings it into optimal overlap with its corresponding pixel in I_R, Fig. 2. From the per-pixel displacements we compute a globally consistent warping for I_M that brings it into photo-consistent registration with respect to I_R. In order to do this for each vertex v_i in T a 2D displacement vector $\vec{r}_i$ is estimated by performing a weighted average on all flow vectors in a rectangular pixel neighborhood around the position of v_i. The triangle mesh is then deformed to globally adapt to the per-vertex displacements by means of a Laplace interpolation. The new mesh configuration approximately satisfies the displacement constraints and

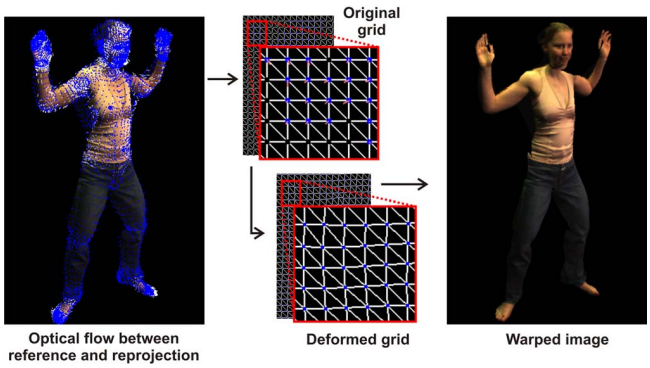

Fig. 2. Illustration of the image-warping procedure.

also preserves a smooth geometry. Formally, the deformation of the mesh is found by solving the Laplace equation

$$\mathbf{L}\mathbf{x} = 0 \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ are the vertex positions and the $n \times n$-Matrix $\mathbf{L}$ is the discrete Laplace operator [8] with

$$\mathbf{L}_{ij} = \begin{cases} 4 & \text{if } i \text{ inner vertex and } i = j, \\ -1 & \text{if } i \text{ inner vertex and } j \text{ in its 4-neighborhood}, \\ 0 & \text{else.} \end{cases} \tag{2}$$

To solve the system, we add suitable boundary conditions to Eq. (1) and reformulate the problem as

$$\min \left(\begin{pmatrix} \mathbf{L} \\ \mathbf{K} \end{pmatrix} \mathbf{x} = \begin{pmatrix} 0 \\ \mathbf{d} \end{pmatrix} \right)^2 \tag{3}$$

The $n \times n$ matrix $\mathbf{K}$ and $\mathbf{d} \in \mathbb{R}^n$ impose the interpolation conditions which will be satisfied in least-squares sense. $\mathbf{K}$ is a diagonal matrix which contains non-zero weights w_i for vertices v_i for which a displacement constraint has been found, as well as for vertices on the image boundary. The vector $\mathbf{d}$ encodes position constraints of the form $x_i = w_i \cdot (u_i + \vec{r}_i)$ for inner vertices with displacement $\vec{r}_i$, and constraints of the form $x_i = u_i \cdot x_i$ for vertices on the image boundary (u_i being undeformed vertex coordinates). It contains 0-entries for all other vertices. We solve Eq. (3) for each coordinate direction individually.

4. GEOMETRY ERROR COMPENSATION

Since the geometry of our body model does not exactly match the geometry of its real-world counterpart, during texture generation color information from spatially distinct surface locations may be mapped onto the same surface point of the model. We prevent these errors by modifying the texture generation process in the following way:

Suppose we want to transform the image $I_C(t)$ seen by camera C at time step t into the texture domain, and thereby

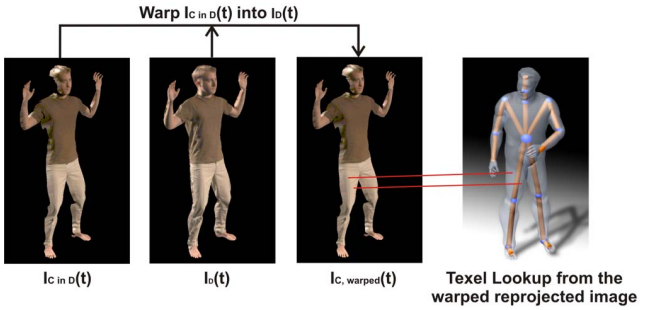

Fig. 3. Geometry error compensation.

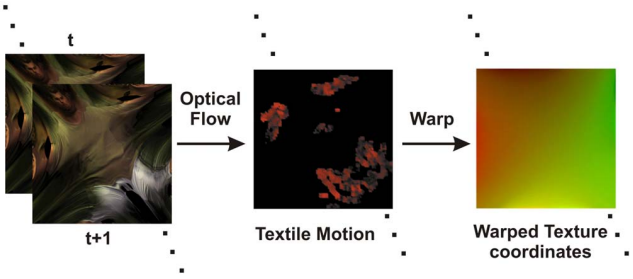

Fig. 4. Cloth shift between two subsequent combined textures t and $t+1$ (in parameterization B) is found via optical flow. In the middle, detected shifted areas are shown in red. Finally, the shift is encoded in the warped texture-coordinates.

produce a texture henceforth referred to as multi-view video (MVV) texture. For each texel K_i in the MVV texture, we first determine the camera that sees it best. We achieve this by searching for the camera that exhibits the minimal angular deviation between its viewing vector and the normal of the surface point that maps to K_i. In case the camera that sees the point best is camera C itself the texel color is taken from $I_C(t)$. In contrast, if it is another camera $D \neq C$, the body model is projectively textured with $I_c(t)$ and the so-textured model is rendered from camera view D to obtain $I_{C\text{ in }D}$, Fig. 3. Using the image warping method (Sect. 3), $I_{C\text{ in }D}$ is deformed such that it is optimally overlaps with $I_D(t)$. The resulting warped reprojected image is called $I_{C,warped}$. Sometimes better results are obtained by recursively applying the warping procedure. Typically, after three iterations a convergence is achieved. The texel color is now taken from $I_{C,warped}(t)$. All possible combinations of reference and reprojected warped images for each time step are precomputed which means 56 warping computations per time step in our eight camera setup.

The comparison shown in Figs. 5a,b proofs that ghosting artifacts in textures that are due to geometry errors can be prevented by our approach without resorting to error-prone geometry deformation. One might argue that optical flow is based on the assumption that all surfaces in the scene are diffuse. For reflectance estimation, though, we deliberately generate specular highlights in the images. Our experiments show that the method nonetheless produces good results since in most input frames the diffuse reflectance is predominant.

5. CLOTH SHIFT DETECTION AND COMPENSATION

Our BRDF estimation procedure assumes that a static set of material parameters can be assigned to each point on the model's surface. In reality, however, this assumption does not hold since the apparel of the person shifts across the body while she is moving. Prior to surface reflectance, we thus estimate the motion of the apparel over time and register all surface textures against a reference texture. Please note that we can still reproduce the true shifting of the apparel during rendering by making the cloth motion information accessible to the renderer. During display, the renderer warps the estimated static BRDF textures back into their true position. We employ the following method to detect the shifting of cloth in the texture domain, Fig. 4:

Our reference time step is the last frame of the RES. MVV textures for this frame and all the frames of the DSS are resampled into a weightedly blended single texture in parameterization B. Cloth shift is detected by computing an optical flow field between subsequent blended textures. This flow field describes for each texel how it shifts across the body surface. This texel motion information is made accessible to the reflectance estimation process as well as the renderer in the form of warped texture coordinates.

Please remember that we use texture parameterization A for sampling, but texture parameterization B for cloth shift computation. Therefore, we project the parameterization A texture coordinates of the reference frame into parameterization B to obtain the texture coordinate image $I_{CoordAB}(0)$. Given the accumulated displacements from the pairwise flow fields we can deform $I_{CoordAB}(0)$ such that it matches the texture at each time of input video using the method from Sect. 3. Note that it is essential to compute the cloth motion relative to the previous frame and accumulate the displacement over time. Only this way, appearance differences due to lighting changes can be robustly handled.

The sequence of deformed texture coordinates enables us to account for cloth shifting during estimation and rendering, although we only estimate a static set of BRDF parameters.

6. RESULTS

Fig. 6a,b shows example screen-shots of a high-quality relit 3D video rendered in real-time. We assess the multi-view warping quality by comparing the image differences between reference views and reprojected model views before and after the warp. The local registration improvements in single image pairs lead to a global improvement in multi-view texture-to-model consistency. In Fig. 5a,b the texture registration improvement due to warping-based geometry correction are visible. With respect to one input stream not used for reconstruc-

Fig. 5. With geometry error compensation ghosting artifacts (a) are significantly reduced (b).

tion we have obtained a peak-signal-to-noise-ration improvement of 0.2 dB. On a Pentium IV 3.0 GHz, warp correction takes around 10 seconds for one pair of reference image and reprojected image.

Cloth shift compensation yields a further PSNR improvement of 0.1-0.2 dB. Although these quantitative improvements may appear small, their influence on the overall visual is quality is well-pronounced. Fig. 6 shows how it corrects the movement of seams of the shirt over the surface. Cloth shift detection takes around 35 s per time step.

Both methods lead to enhanced registration in majority of the situations. Still, being global optimization methods, they can possibly lead to local deterioration. During cloth shift detection the evolution of wrinkles can also cause a problem. In some rare cases, seams at visibility boundaries may occur, however this was never a noticeable problem in our many test scenes. Both methods are optional extensions to the original relightable free-viewpoint video estimation framework, and it is up to the user to decide if they are activated.

Despite the limitations, our results show that image-based warp correction and cloth shift detection are effective registration techniques that enhance spatio-temporal photo-consistency for relightable 3D videos.

7. CONCLUSIONS

We presented two image-based spatio-temporal registration techniques that enable high-quality reconstruction of model-based relightable 3D videos. In conjunction, they enable the faithful reproduction of an actor's appearance despite small inaccuracies in the template model's shape, and despite movement of textiles across the body's surface. Quality improvements in the real-time renderings were shown both quantitatively and visually. In the future, we plan to work on new algorithms to reconstruct moving people wearing very wide apparel.

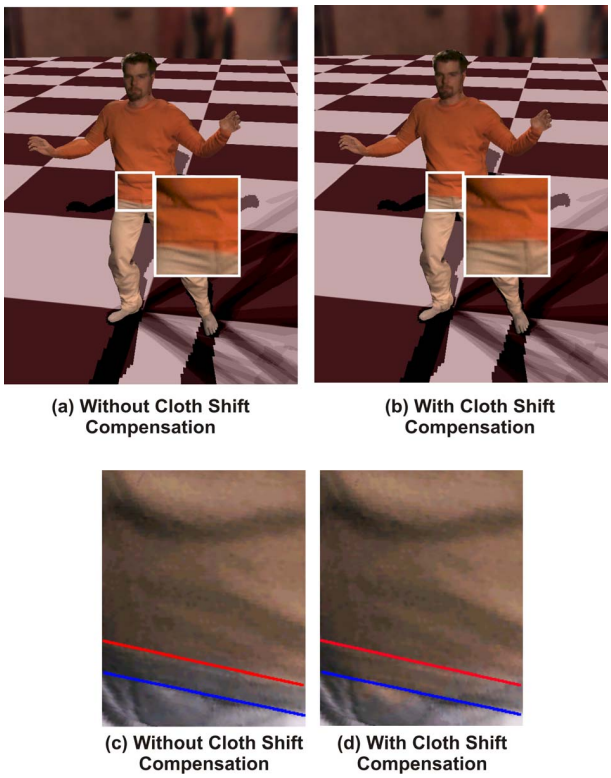

Fig. 6. Top row: screen-shots of relightable 3D videos rendered under captured real-world illumination. (a) Without cloth-shift detection, the seam of the t-shirt is rendered incorrectly. (b) With cloth shift detection, it is reproduced accurately. - Bottom row: Same effect for another sequence. Overlayed lines indicate the correct (blue) and wrong (red) position of the t-shirt's seam. With active shift detection (b) it is correctly reproduced, without it not (a).

8. REFERENCES

[1] J. Carranza, C. Theobalt, M.A. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," in *Proc. of SIGGRAPH'03*, 2003, pp. 569–577.

[2] C. Theobalt, N. Ahmed, H.P.A. Lensch, M.A. Magnor, and H.-P. Seidel, "Seeing people in different light - joint shape, motion and reflectance capture," *IEEE TVCG*, accepted, to appear.

[3] E. de Aguiar, C. Theobalt, M.A. Magnor, and H.-P. Seidel, "Reconstructing human shape and motion from multi-view video," in *2nd European Conference on Visual Media Production (CVMP)*, London, UK, December 2005, pp. 42–49, The IEE.

[4] H. Kück, W. Heidrich, and C. Vogelgsang, "Shape from contours and multiple stereo - a hierarchical, mesh-based approach.," in *CRV*, 2004, pp. 76–83.

[5] C.H. Esteban and F. Schmitt, "Silhouette and stereo fusion for 3d object modeling," *CVIU*, vol. 96, no. 3, pp. 367–392, 2004.

[6] R. Zayer, C. Rössl, and H.-P. Seidel, "Discrete tensorial quasi-harmonic maps," in *Proc. of Shape Modeling International*. 2005, pp. 276–285, IEEE.

[7] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. DARPA IU Workshop*, 1981, pp. 121–130.

[8] M. Meyer, M. Desbrun, P. Schröder, and A. Barr, "discrete differential-geometry operators for triangulated 2-manifolds," in *Proc. of Vis Math*, 2002, pp. 35–37.