Christian Theobalt, Naveed Ahmed,
Gernot Ziegler, and Hans-Peter Seidel

© BRAND X PICTURES

High-Quality Reconstruction from Multiview Video Streams

Dynamic representation of 3-D human actors

Three-dimensional (3-D) video processing is currently an active area of research that attracts scientists from many disciplines, including computer graphics, computer vision, electrical engineering, and video processing. They join their expertise to attack the very hard problem of reconstructing dynamic representations of real-world scenes from a sparse set of synchronized video streams. To put this idea into practice, a variety of challenging engineering and algorithmic problems have to be efficiently solved, ranging from acquisition, over reconstruction in itself, to realistic rendering.

The complexity of the task originates from the fact that a variety of aspects of the real-world scene have to be faithfully mapped to a digital model. Most importantly, the dynamic shape and motion as well as the dynamic appearance and material properties of a real-world scene have to be captured, reconstructed, and displayed in high quality. Among the most difficult scenes to

Digital Object Identifier 10.1109/MSP.2007.905701

1053-5888/07/$25.00©2007IEEE

reconstruct are scenes involving human actors. The eye of the human observer is unforgiving and will immediately unmask even slightest inaccuracies in a virtual human's appearance or motion, and therefore a faithful estimation of either of these aspects is a necessity.

This article is a tutorial style review of methods from the literature aiming at reconstruction of 3-D humans as well as of a variety of model-based approaches that we developed to reconstruct, render, and encode free-viewpoint videos of human actors. We will show that the commitment to an a priori shape representation of a person in the real world allows us to solve many of the previously described reconstruction problems in an efficient way.

The article continues with a review of important categories of alternative dynamic scene reconstruction methods. We also analyze their advantages and disadvantages and discuss their suitability for reconstructing virtual human actors. Thereafter, we review important related approaches for reconstructing dynamic reflectance properties.

Subsequently, we detail our model-based paradigm to free-viewpoint video of human actors. First, we describe the multicamera system that we employ to capture input video streams. Thereafter, we describe a shape-adaptable human template model serving as our dynamic geometry and motion representation. A core component of our 3-D video approach is a model-based analysis-through-synthesis algorithm enabling us to capture the time-varying shape and motion of an actor from input video streams without the use of optical markings. In a first algorithmic variant, we create free-viewpoint videos by applying a real-time dynamic surface texturing approach to our dynamic scene models, thereby reproducing the actor's appearance from any viewpoint and under fixed lighting conditions. An efficient encoding for the data streams is also illustrated. In a second algorithmic variant, we not only reproduce dynamic surface appearance under fixed lighting positions but, rather, estimate a complete dynamic reflectance model of the recorded individual. By this means, free-viewpoint videos can also be displayed in real time under arbitrary virtual lighting conditions. The model-based framework allows for an efficient compaction of even relightable dynamic scene representations, which lends itself to real-time visualization on consumer-grade hardware. Finally, we present a variety of results obtained with the different algorithms, and we give an outlook to ongoing and future work along with a conclusion.

RELATED WORK

Since the dynamic scene reconstruction methods that we propose to capture virtual actors simultaneously solve a variety of problems, there is an immense body of related work that we can capitalize on, ranging from previous work in markerless motion capture to work on image-based and real-time rendering. However, in this article, we intend to give an overview of the most related methods that also address dynamic scene reconstruction as a whole and not only an algorithmic subaspect. Therefore, we focus on the most important literature in 3-D video reconstruction as well as recent work on dynamic reflectance estimation.

3-D VIDEO

Early research that paved the way for free-viewpoint video was presented in the field of image-based rendering (IBR). Shape-from-silhouette methods reconstruct geometry models of a scene from multiview silhouette images or video streams [17], [18]. Starting from the silhouettes extracted from the camera pictures, a conservative shell enveloping the true geometry of the object is computed by reprojecting the silhouette cones into the 3-D scene and intersecting them. This generated shell is called the visual hull. For two-dimensional (2-D) scenes, the visual hull is equal to the convex hull of the object, and for 3-D scenes, the visual hull is contained in the convex hull, where concavities are not removed but hyperbolic regions are. While the visual hull algorithms are efficient and many systems allow for real-time reconstruction and rendering performance, the geometry models they reconstruct are often not accurate enough for high-quality reconstruction of human actors. As such, when observed by only a few cameras, the scene's visual hull is often much larger than the true scene and disturbing phantom volumes due to undersampling lead to a deterioration of the overall appearance. When rendering new views, one can partially compensate for such geometric inaccuracies by view-dependent texture-mapping.

Strictly, the visual hull is the maximal volume constructed from all possible silhouettes. In almost any practical setting, the visual hull of an object is computed with respect to a finite number of silhouettes, which is called the inferred visual hull. There exist two classes of methods to compute the visual hull: 1) voxel carving methods, which carve away all voxels that are not contained in the silhouettes of the acquisition cameras and 2) image-based methods that exploit epipolar geometry and store so-called occupancy intervals at every pixel. Some examples are image-based [18] or polyhedral visual hull methods [17] as well as approaches performing point-based reconstruction [10]. Despite quality limitations of the measured shape models, visual hull reconstruction methods are still the algorithms of choice when real-time performance is the primary goal.

To overcome some of the principal limitations of visual hull methods, researchers tried to combine visual hull and stereo reconstruction approaches. These hybrid methods typically employ the visual hull surface as a shape prior and use a stereo method to locally refine the shape estimates to accurately recover convex surface areas also [15], [24].

In contrast to the previously mentioned hybrid approaches, purely stereo-based 3-D video reconstruction methods don't require separable image silhouettes for the foreground objects to be reconstructed and can, therefore, directly be applied to estimate the dynamic shape of foreground and background. On the other hand, the latter category of methods often requires a much denser camera arrangement, which leads to restrictions in the range of virtual viewpoints that can be handled.

A stereo-based method to reconstruct and render complete dynamic scenes comprising of dynamic foreground and background is presented in [34] (Figure 1). It combines a novel segmentation-based stereo algorithm with a multilayered representation, which can then be used to generate intermediate viewpoints along a one-dimensional (1-D) rail of recording cameras. A segmentation-based approach to stereo tries to overcome some of the limitations of the pixel-based algorithms. Pixels are inherently hard to match, and by correlating entire segments, the algorithm produces much better depth maps. However, it relies on the assumption that all pixels of a segment belong to the same surface—so there are no discontinuities. Hence, a fine over-segmentation has to be performed during a preprocess step (Figure 1). Although the range of virtual viewpoints is limited, the achieved reconstruction and rendering quality is very high.

To overcome some of the limitations inherent to purely passive stereo-based 3-D video methods, [29] proposes an approach supported by active illumination in the scene. Multiple video projectors simultaneously project random noise patterns into the scene to increase the robustness of the geometry reconstruction. Further geometry enhancements are achieved by employing a space-time stereo method. Pattern projection is performed in synchronization with the camera system in such a way that the illumination patterns add up to white light in the texture cameras. This way, appearance estimation in conjunction with texture estimation is possible.

While the algorithms that were described so far rely on dynamic geometry reconstruction to create novel viewpoint renderings, ray-space techniques explicitly abstain from shape estimation. In contrast, they generate novel views by appropriately combining light rays of a scene captured with multiple cameras. Looking at it from a different perspective, ray-space methods aim at locally reconstructing a simplified version of the full plenoptic function (which describes the light transport at each point and in each direction of space) to create novel virtual views, however at the cost of increased memory consumption. One exemplary method is light field rendering [14], which has, in an extended form, been employed in the 3-D TV system [19] to enable simultaneous scene acquisition and rendering in real-time; [8] also uses light field rendering for novel viewpoint generation in dynamic scenes. Being purely data-driven approaches, ray-space techniques have the big advantage that they are capable of reproducing any local or global lighting and appearance effect visible in the real-world, given that the scene has been sampled densely enough with recording cameras. At the same time, the high required sampling density is also the main drawback of these approaches since the huge amount of captured data makes it difficult to handle larger scenes, large image resolutions, and, in particular, dynamic scenes.

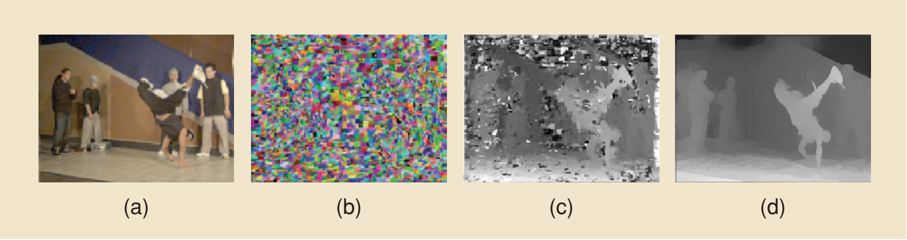
As opposed to the above methods, we take a model-based approach to free-viewpoint video [4], [27] of human actors. By applying a shape and motion prior, we can efficiently circumvent many of the drawbacks of related approaches described previously. Eventually, the commitment to a strong prior makes it possible to realistically reproduce the omnidirectional appearance of captured human actors even though we only use eight recording cameras.

With the increasing availability of reconstruction approaches, the issue of efficient content encoding has become more important. To name just exemplary work, [19] described efficient multiview video encoding, and [34] developed a framework for multiview depth video encoding. An extension to the MPEG framework for encoding multiview video and geometry content has been proposed in [31]. Methods for efficient encoding of model-based free-viewpoint videos are described in [32] and [33] and will be addressed briefly in the following.

All of the algorithms mentioned so far can visualize a recorded scene only under the same illumination conditions that it was captured in. For implanting 3-D video footage, in particular virtual actors, into novel virtual surroundings—a problem that often needs to be solved in movie and game productions—information on dynamic surface reflectance is needed.

REFLECTANCE ESTIMATION IN STATIC AND DYNAMIC SCENES

Until today, the problem of measuring reflectance properties of real objects from image data has been mainly addressed for the case of static scenes. Typically, a single-point light source is used to illuminate an object of known 3-D geometry consisting of only one material. One common approach is to take high-dynamic range (HDR) images of a curved object, yielding a different incident and outgoing directions per pixel and thus capturing a vast number of reflectance samples in parallel. Often, the parameters of an analytic bidirectional reflectance distribution function (BRDF) model are fit to the measured data [13], or a data-driven model is used [20]. Reflectance measurements of scenes with more complex incident illumination can be derived by either a full-blown inverse global illumination approach [2] or by representing the incident light field as an environment map and solving for the direct illumination component only [22]. In a method that we will explain later, we approximate the incident illumination by multiple point light sources and estimate BRDF model parameters taking only direct illumination into account.



[FIG1] Results from the stereo reconstruction approach of Zitnick et. al. [34]. From left: The segmented input image is used to estimate the initial disparity, which is then refined using an iterative scheme.

Rushmeier et al. estimate diffuse albedo and normal map from photographs with varied incident light directions [23]. In [9], reflectance and shape of static scenes are simultaneously refined using a single light source in each photograph.

All the approaches mentioned in the preceding paragraphs were tailored to handle static scene. Only very few methods so far have tried to attack the even more difficult problem of estimating the time-varying reflectance properties of dynamic scenes.

In one line of research, a data-driven approach to dynamic reflectance measurement is taken. Here, instead of explicitly reconstructing a parametric reflectance model, novel views under novel lighting conditions are generated by weightedly blending captured input images. An early approach to data-driven reflectance measurement is proposed in [11], where a special device called lightstage, comprising of light sources and imaging sensors, is used to capture the reflectance field of a human face.

Wenger et al. [30] extend the static light stage device such that it enables capturing of dynamic reflectance fields, in this particular case the time-varying reflectance of an actor's face. This novel light stage comprises a dome of LED light sources that can reproduce basic lighting conditions at several hundred frames per second (fps), as well as a high-speed video camera. By having such a fast illumination and recording apparatus, it becomes feasible to record a full set of images of the face under all basic lighting conditions for each frame of a regular 30 fps video clip. The visual results achieved with this method are impressive, however it is not possible to change the viewpoint in the scene.

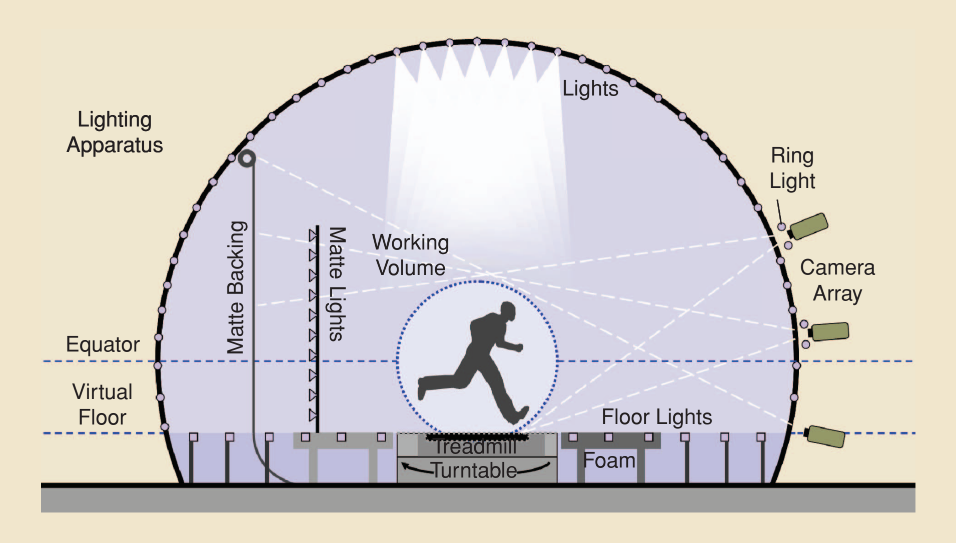
Einarsson et al. [8] extend this approach even further by using a much larger light stage (light stage 6) that enables them to capture a seven-dimensional (7-D) full dynamic reflectance field. The novel light stage features a treadmill in the center, where the person walks. Eventually, this setup enables us to capture a complete set of images for periodically moving humans that spans two dimensions for the images themselves, two dimensions (directions) for the incident lighting, two dimensions for the viewpoints (directions) and one dimension for time (Figure 2). When rendering a novel viewpoint of a particular captured moment under novel lighting conditions, the novel lighting conditions are projected into the employed lighting basis and the images in the 7-D data set are combined appropriately to generate the output view. This way, human performances can be rendered from novel perspectives and relit. Unfortunately, their method can only capture periodic walking and only reproduces low-frequency lighting effects. Also, the required hardware setup makes this method infeasible for many practical applications.

Carceroni and Kutulakos present a surfel-based method for simultaneous motion and reflectance capture for nonrigid objects [3], producing nice results for confined scenes. Their method was the first one aiming at reconstruction of parametric reflectance models for dynamic scenes from multiview video footage. Relighting was not their main application focus. The BRDF estimation was rather part of an involved multistage optimization process producing a reliable time-varying shape model. Although reflectance reconstruction was not their primary goal, the proposed algorithms served as a motivation for our model-based dynamic relighting method.

In our work, we decided to employ a parametric reflectance model to estimate dynamic surface reflectance of human actors. Having a good prior shape model, this enables us to reconstruct high-quality dynamic reflectance properties and normal maps using only a handful of recording cameras. Furthermore, our reflectance description allows for arbitrary viewpoint changes as well as high-frequency relighting [25].

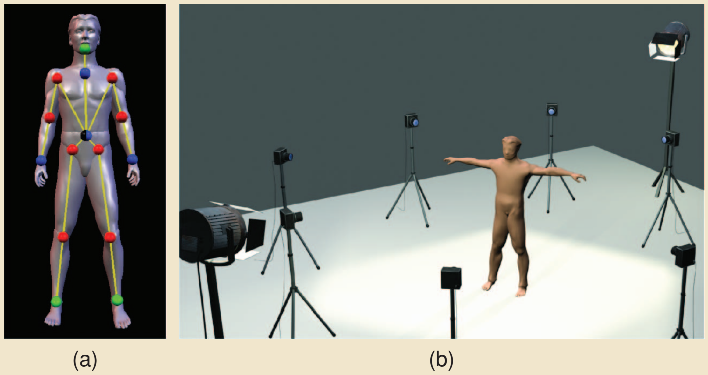
ACQUISITION—A STUDIO FOR MULTIVIEW VIDEO RECORDING

The input to our system are multiple synchronized video streams of a moving person [multiview video texture (MVV) sequences] that we capture in our free-viewpoint video studio. The studio features a multicamera system that enables us to capture a volume of approximately 4 × 4 × 3 m with eight externally synchronized video cameras. The imaging sensors can be placed in arbitrary positions, but typically, we resort to an approximately circular arrangement around the center of the scene. Optionally, one of the cameras is placed in an overhead position [Figure 3(b)]. Each of our eight Imperx MDC 1004 video cameras features a 1004 × 1004 pixel image sensor with 12-b color depth and



[FIG2] A schematic diagram of the acquisition system used by Einarsson et al. [8].

runs at a frame rate of 25 fps. Prior to recording, the cameras are calibrated, and intercamera color consistency is ensured by applying a color-space transformation to each video stream. The lighting conditions in the studio are fully controllable, and the scene background optionally can be draped with black molleton. We have a set of different light setups at our disposal. While for the free-viewpoint video with dynamics textures, we prefer a diffuse illumination, our work on relighting requires spotlight illumination. For capturing reflectance estimation sequences, we employ two spotlights. The spotlights are placed on the either side of the room, ensuring maximal illumination while also minimizing the interference. The light sources are fully calibrated, and their position and photometric properties are determined. The use of this calibrated lighting setup for dynamic reflectance estimation is explained in the following.



[FIG3] (a) Surface model and the underlying skeletal structure. Spheres indicate joints and the different parameterizations used; blue sphere—3 DOF ball joint, green sphere—1 DOF hinge joint, red spheres (two per limb)—4 DOF limb parameterization. (b) Typical camera and light arrangement during recording.

THE ADAPTABLE HUMAN BODY MODEL

We employ a triangle mesh representation because it offers a closed and detailed surface description and it can be rendered very fast on graphics hardware. Since the model must be able to perform the same complex motion as its real-world counterpart, it is composed of multiple rigid-body parts that are linked by a hierarchical kinematic chain. The joints between segments are suitably parameterized to reflect the object's kinematic degrees of freedom (DoF). Besides object pose, the dimensions of the separate body parts also must be kept adaptable as to be able to match the model to the object's individual stature.

A publicly available virtual reality modeling language (VRML) geometry model of a human body is used [Figure 3(a)]. The model consists of 16 rigid body segments, one each for the upper and lower torso, neck, and head; and pairs for the upper arms, lower arms, hands, upper legs, lower legs, and feet. In total, more than 21,000 triangles make up the human body model. A hierarchical kinematic chain connects all body segments, resembling the anatomy of the human skeleton. Seventeen joints with a total of 35 joint parameters define the pose of the virtual character.

In addition to the pose parameters, the model provides 17 anthropomorphic shape parameters per segment to scale and deform each of them. One parameter is a global scaling parameter, the 16 remaining anthropomorphic DoF control the deformation of each segment by means of a simple free-form deformation scheme using Bézier curves for scaling.

SILHOUETTE-BASED ANALYSIS-THROUGH-SYNTHESIS

The challenge in applying model-based analysis for free-viewpoint video reconstruction is to find a way to adapt the geometry model automatically and robustly to the subject's appearance as it was recorded by the video cameras. Since the geometry model is suitably parameterized to alter its shape and pose, the problem reduces to determining the parameter values that achieve the best match between the model and the video images. This task is regarded as an optimization problem. The subject's silhouettes, as seen from the different camera viewpoints, are used to match the model to the video images: The model is rendered from all camera viewpoints, and the rendered images are thresholded to yield binary masks of the model's silhouettes. The rendered model silhouettes are then compared to the corresponding image silhouettes [4], [27]. As a comparison measure, the number of silhouette pixels that do not overlap is determined. Conveniently, the exclusive-or (XOR) operation between the rendered model silhouette and the segmented video-image silhouette yields those pixels that are not overlapping. Fortunately, an energy function based on XOR operation can be evaluated very efficiently in graphics hardware (Figure 4). With eight cameras, a 3.0 GHz Pentium IV with a GeForce 6800 graphics board easily performs more than 250 of such matching function evaluations per second.

The silhouette-based analysis-through-synthesis approach is employed for two purposes: the initialization or shape adaptation of the model's geometry and the computation of the body pose at each time step. For the shape adaptation, the silhouette-based analysis-through-synthesis algorithm is used to optimize the anthropomorphic parameters of the model. During model initialization, the actor stands still for a brief moment in a pre-defined pose to have his silhouettes recorded from all cameras. The generic model is rendered for this known initialization pose, and without user intervention, the model proportions are optimized automatically to match the individual's silhouettes. Shape adaptation commences by roughly aligning the model globally. Thereafter, it iterates between segment scaling and pose parameter computation. Shape customization is finalized by finding an optimal set of Bézier scaling parameters such that the silhouette outlines are reproduced as closely as possible.

Thanks to advanced rendering techniques, an exact match is neither needed for convincing dynamic texturing nor for reflectance estimation (see following sections). The initialization procedure takes only a few seconds. From now on, the anthropomorphic shape parameters remain fixed. Shape adaptation can be extended to reconstruct not only a spatiotemporally consistent shape model but also smaller-scale per-time step deformations [5].

The individualized geometry model automatically tracks the motion of the human actor by optimizing the 35 joint parameters for each time step. The analysis-through-synthesis framework enables us to capture these pose parameters without having the actor wear any specialized apparel. This is a necessary precondition for free-viewpoint video reconstruction, since only if motion is captured completely passively can the video imagery be used for texturing. The model silhouettes are matched to the segmented image silhouettes of the actor so that the model performs the same movements as the human in front of the cameras (Figure 4). At each time step, an optimal set of pose parameters is found by performing a numerical minimization of the silhouette XOR energy functional in the space of pose parameters. The performance of the silhouette-based pose tracker can be further improved by capitalizing on the structural properties of the optimization problem, in particular, the kinematic hierarchy [28]. Tracking can also be augmented by additionally considering texture and 3-D scene flow [27].

FREE-VIEWPOINT VIDEO WITH DYNAMIC TEXTURES

By combining the silhouette-based analysis-through synthesis method with a dynamic texture generation, we can reconstruct and render free-viewpoint videos of human actors that reproduce the omnidirectional appearance of the actor under fixed lighting conditions. A high-quality 3-D geometry model is now available that closely matches the dynamic object in the scene over the entire length of the sequence. To display the object photo-realistically, the recorded video images are used for GPU-based projective texturing of the model's surface. We can also capitalize on spatiotemporal coherence to encode efficiently image and geometry data.

DYNAMIC PROJECTIVE TEXTURING

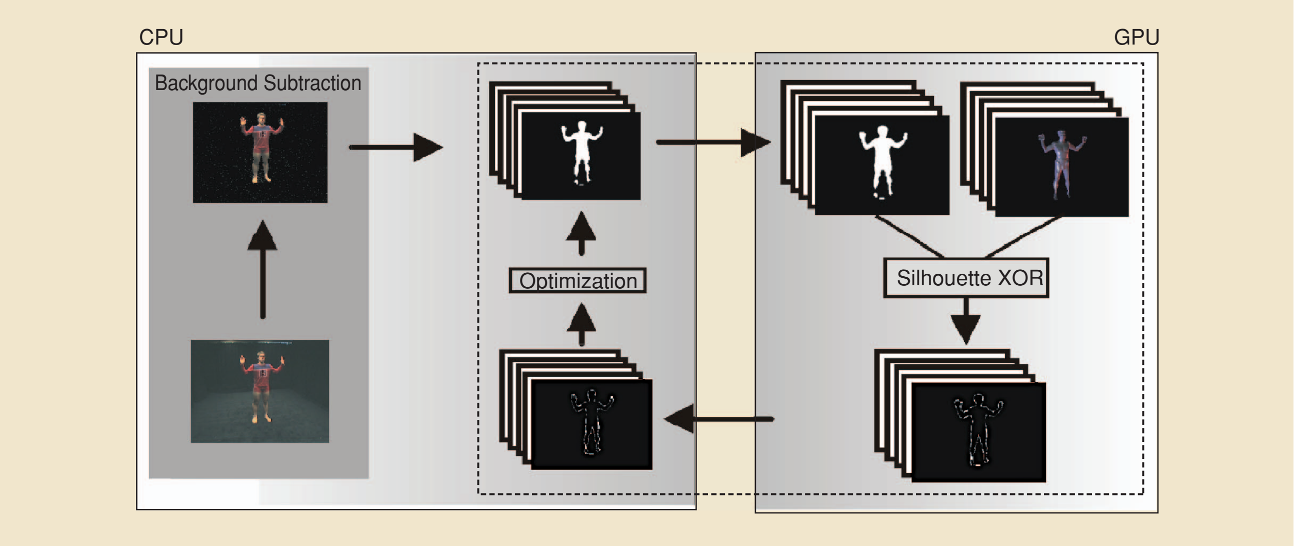
Prior to display, the geometry model as well as the video cameras' calibration data is transferred to the graphics board. During rendering, the user's viewpoint information, the model's updated pose parameter values, the current video images, as well as the visibility and blending coefficients ν_i, ω_i for all vertices and cameras i are continuously transferred to the graphics card.

The color of each rendered pixel $c(j)$ is determined by blending all l video images I_i according to

$$c(j) = \sum_{i=1}^{l} \nu_i(j) * \rho_i(j) * \omega_i(j) * I_i(j), \qquad (1)$$

where $\omega_i(j)$ denotes the blending weight of camera i, $\rho_i(j)$ is the optional view-dependent rescaling factor, and $\nu_i(j) = \{0, 1\}$ is the local visibility. During texture preprocessing, the weight products $\nu_i(j)\rho_i(j)\omega_i(j)$ have been normalized to ensure energy conservation. Technically, (1) is evaluated for each fragment by a fragment program on the graphics board. By this means, time-varying cloth folds and creases, shadows, and facial expressions are faithfully reproduced in texture, lending a very natural, dynamic appearance to the rendered object, Figure 5(a).

Given approximate geometry and Lambertian surface reflectance assumption, high-quality, detailed model texture can



[FIG4] Hardware-based analysis-through-synthesis for free-viewpoint video: To match the geometry model to the multi-video recordings of the actor, the image foreground is segmented and binarized. The boolean XOR operation is executed between the foreground images and the corresponding model renderings. The numerical minimization algorithm runs on the CPU while the energy function evaluation is implemented on the GPU.

be obtained by blending the video images cleverly. A visually convincing weight assignment has been found to be

$$\omega_i = \frac{1}{(1 + \max_j(1/\theta_j) - 1/\theta_i)^\alpha}, \qquad (2)$$

where θ_i denotes the angle between a vertex normal and the optical axis of camera i and the weights ω_i additionally are normalized to sum up to unity. The parameter α determines the influence of vertex orientation with respect to camera viewing direction and the impact of the most head-on camera view per vertex [4].

Due to the use of a parameterized geometry model, the silhouette outlines in the images do not correspond exactly to the outline of the model. When projecting video images onto the model, a texture seam belonging to some frontal body segment may fall onto another body segment farther back. To avoid such artifacts, extended soft shadowing is applied, which makes sure that a triangle is textured by a camera image only if all of its three vertices are completely visible from that camera.

ENCODING

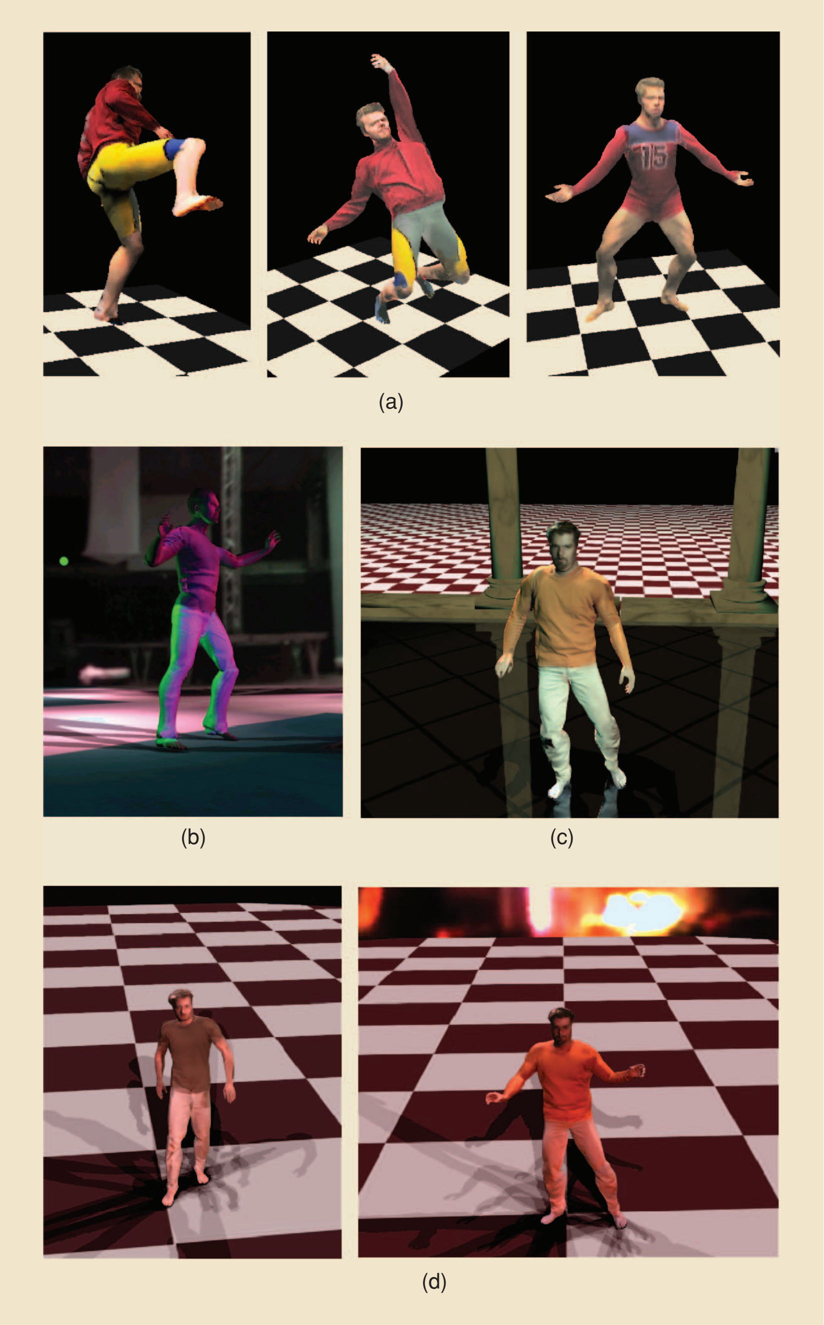
Conveniently, time-varying geometry is represented compactly as a stream of pose parameters for the model. To encode the multiview video stream efficiently, it turns out to be advantageous to convert each input video frame into a texture, thereby creating eight so-called MVV textures for each time step, Figure 6. MVV textures actually comprise correlated four-dimensional (4-D) data volumes, since the texture changes only slightly with the viewing angle over time. We have developed two approaches to exploit this correlation for texture stream encoding.

The first approach employs predictive texture encoding [32]. It generates a two-step average (first over all camera angles, then over all time steps) and subsequently computes differential textures with respect to the average textures. All resulting textures are then compressed using a shape-adaptive wavelet algorithm. The format grants random access over two differential image additions.

The second compaction scheme is based on 4-D-SPIHT [33], and we have a closer look at its individual processing steps in the following.

DATA GROUPING AND FILLING

Compression of a data block commences when all necessary camera images are available as textures. After resampling, we group the texture maps into blocks of spatial and temporal coherency, yielding 4-D data blocks of YUV samples. The block



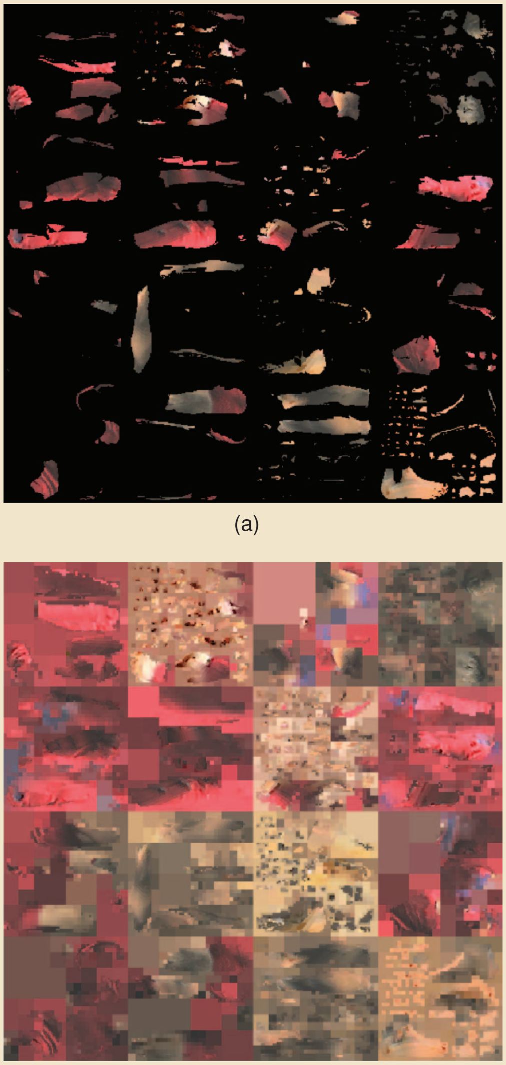
[FIG5] **Conventional video systems cannot offer moving viewpoints of scenes frozen in time. However, with our free-viewpoint video system *freeze-and-rotate* camera shots of body poses are possible. (a) Novel viewpoints of scenes frozen in time for different subjects and different types of motion. (b) Disco-type lighting condition. (c) Environment setting typical for a computer game. (d) 3-D videos rendered under different real-world lighting conditions stored in an HDR environment map of Grace Cathedral (courtesy of Paul Debevec). In either case, the actors appear very lifelike and subtle surface details are faithfully reproduced. Also shadows and mirroring effects can be rendered in real-time.**

division corresponds to the group of picture (GOP) block structure commonly used in MPEG video formats and allows for limited random access as long as the whole 4-D block containing a certain texture is decoded. U and V values can optionally be subsampled, but we currently work with reduced bitrates for these color components (see the SPIHT encoder below).

Unused texels (currently: black pixels) in these 4-D blocks are now filled with averages of the surrounding valid texels see Figure 6 for an example. This ensures best possible data compression under the subsequently applied algorithm, as described in [16].

To serve this purpose, the whole 4-D data block is first downsampled in a Laplacian 4-D pyramid, all the way to the lowest resolution of 1 × 1, taking the different dimension extents into consideration (a division by two remains one if the result would be smaller than one). Afterwards, the pyramid is traversed backward from the lowest to the highest resolution, and each unused (black) texel receives the color of its associated, average parent in the previous level. This way, it is ensured that all unused texels are filled with a color value that corresponds to the average of all valid texels in its support region.

EARLY RESEARCH THAT PAVED THE WAY FOR FREE-VIEWPOINT VIDEO WAS PRESENTED IN THE FIELD OF IMAGE-BASED RENDERING.



[FIG6] (a) Input texture map. (b) Same map after filling operation.

WAVELET ENCODING

The following 4-D wavelet transformation uses Haar wavelets. We take the 4-D data block that was filled in the previous step and sequentially apply a 1-D Haar wavelet transform in all four dimensions until even the texture dimension sizes have been reduced to two. Finally, compression commences. The compression algorithm is based on the widely used SPIHT algorithm, although in a new adaptation, making it suitable for 4-D data. It is based on work done in [16]. The encoder is currently able to handle a 4-D data block with pairs of equal dimensions (e.g., max(s, t, u, v) = {8, 8, 1024, 1024}, i.e., eight timesteps of eight cameras at 1024 × 1024 resolution).

DECODING

Most decoders will probably extract several time steps at once, since time steps are usually read sequentially. The bit mask can only be applied if it has been transmitted to the decoder. If this is not the case, shadow casting must be applied for masking if multiview interpolation is intended (as noted in the next section).

Figure 7 shows example output from the reference decoder. Notice the typical wash-out effect of wavelet compression. The outer contours were transmitted in an additional shape-mask.

RELIGHTABLE FREE-VIEWPOINT VIDEO

In the previous section, we introduced an approach to realistically render human actors for all possible synthetic viewpoints. However, this algorithm can only reproduce the appearance of the actor under the lighting conditions that prevailed at the time of acquisition. To implant a real-world actor into surroundings different from the recording environment, his appearance must be adapted to the new illumination situation. To this end, a description of the actors surface reflectance is required. We have enhanced our original free-viewpoint video pipeline such that we are able to reconstruct such dynamic surface reflectance descriptions. The guiding idea behind this process that we call dynamic reflectometry is that, when letting a person move in front of a calibrated static setup of spot lights and video cameras, the cameras are not only texture sensors but actually reflectance sensors. Due to the motion of the person, each point on the body's surface is seen under many of different incoming light and outgoing viewing directions. Thus, we can fit a dynamic surface reflectance model to each point on the body surface that consists of a per-texel parametric bidirectional reflectance distribution function and a per-texel normal with time-varying direction. We now review the algorithmic steps needed to generate relightable free-viewpoint videos. See [25] and [26] for in depth details.

APPEARANCE ESTIMATION IN CONJUNCTION WITH TEXTURE ESTIMATION IS POSSIBLE.

MODIFICATIONS TO THE RECONSTRUCTION PIPELINE

During recording, we employ two calibrated spotlights, i.e., we know their positions and photometric properties. For each person and each type of apparel, we record one sequence, henceforth termed reflectance sequence (RS), in which the person performs a rather simple in-place rotation while attaining approximately a static posture. The RS will be used to estimate the surface reflectance properties. The actual relightable free-viewpoint videos, as well as the time-varying normal map are reconstructed from the so-called dynamic sequences (DS) in which the actor can move arbitrarily.

Reflectance estimation causes more strict requirements to the quality of the employed body model. To prevent rendering artifacts at body segment boundaries and to facilitate spatio-temporal texture registration, we transform the shape-adapted segmented body model into a single-skin model by means of an interactive procedure.

Prior to reflectance estimation, we transform each input video frame into a 2-D surface texture. Textural representation of surface attributes facilitates rendering of the relightable free-viewpoint videos and also enables us to take measures to enhance spatio-temporal multiview texture registration. Incorrect multiview registration would eventually lead to erroneous reflectance estimates. There are two primary reasons for inaccurate texture registration, first the fact that we use only an approximate model, and second, transversal shift of the apparel while the person is moving. We counter the first problem by warping the multiview input images such that they comply with the geometry model at each time step of video. The motion of textiles is identified and compensated by optical flow computation and texture warping [26].

BRDF ESTIMATION

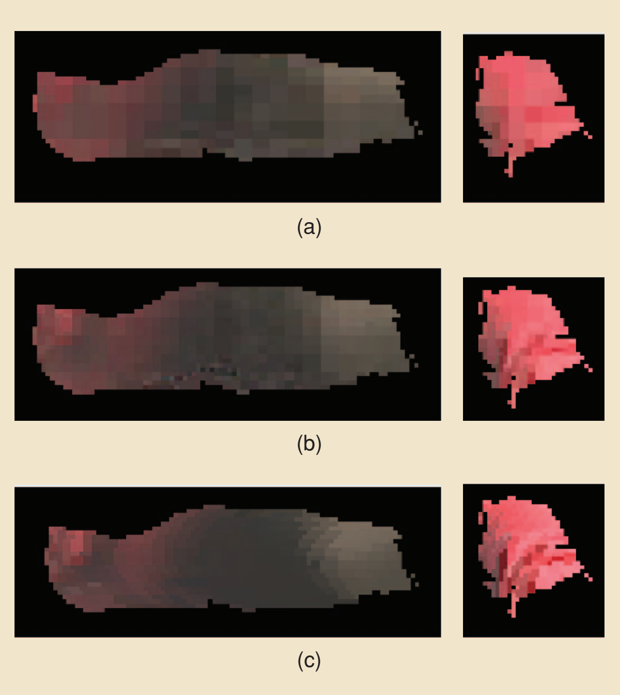
The BRDF part of our reflectance model is estimated for each subject and each type of apparel from the RS, in which the approximately static body is seen under different lighting and viewing directions. We employ a parametric BRDF representation, because it allows us to represent the complex reflectance function in terms of a few coefficients of a predefined functional skeleton. The BRDF thus compactly represents surface reflectance in terms of four direction parameters, the incoming light direction and the outgoing viewing direction, as well as the model parameters. In our approach, we can employ any arbitrary BRDF representation, but we mainly used the Phong [21] and Lafortune [12] models.

In the dynamic sequence, we collect several samples of the BRDF for each surface point or, in GPU terminology, for each texel. The goal is to find optimal model parameters for each texel that reproduce the collected reflectance samples best. We formulate this as the solution to a least-squares problem in the difference between collected samples and predicted appearance according to the current estimate. In general, our estimation of BRDF parameters and later, the estimation of the time-varying normals, is based on minimizing for each surface point $\vec{x}$ the error $E(\vec{x}, \rho(\vec{x}))$ between the current model $\rho(\vec{x})$ and the measurements for this point from all cameras i at all time steps t:

$$E(\vec{x}, \rho(\vec{x})) = \sum_{t}^{N} \sum_{i}^{8} \kappa_i(t) \times \left(S_i(t) - \left[\sum_{j}^{J} \lambda_j(t) (f_r(\hat{l}(t), \hat{v}_i(t), \rho(\vec{x})) \cdot I_j(\hat{n}(t) \cdot \hat{l}(t))) \right] \right)^2 . \tag{3}$$

The term is evaluated separately in the red, green, and blue color channel. $S_i(t)$ denotes the measured color samples at $\vec{x}$ from camera i, and I_j denotes the intensity of light source j. The viewing directions $\hat{v}_i(t)$ and light source directions $\hat{l}_j(t)$ are expressed in the point's local coordinate frame based on the surface normal $\hat{n}(t)$. Visibility of the surface point with respect to each camera is given by $\kappa_i(t)$ and with respect to the light sources by $\lambda_j(t)$, both being either 0 or 1. f_r finally evaluates the BRDF.

By minimizing (3) in the respective parameter sets, we compute optimal estimates for the BRDF parameters, ρ_{best}, as well



[FIG7] The texture map patches. Decoding was performed with equal *Y*, *U*, *V* datarate, and using the fill feature. (a) 0.05 bpp. (b) 0.25 bpp. (c) Encoder input.

as optimal normal estimates for the body in the initialization posture. The procedure alternates between BRDF estimation and normal refinement, and works as follows (Figure 8).

Assuming that there is little variation between per-pixel specular BRDF within a same material, we first cluster the model into different material. A first BRDF estimate is computed by using the template model's surface normals. Note that specular BRDF components are estimated per-material to collect enough samples for this high-frequency signal. Diffuse components are estimated on a per-texel basis to capture saptial variation. After the first reflectance estimate, refined surface normal fields are estimated by means of the procedure described in the following section. The final set of BRDF parameters is estimated with the refined normal field.

We acquire the data for BRDF estimation under a fixed setup of cameras and light sources. This may lead to a biased sampling of surface reflectance since each surface point is only seen under a limited number of half-vector directions. We thus propose a spatio-temporal reflectance sharing method that reduces this bias by taking into account reflectance samples from other surface locations made of similar material. For details on the spatio-temporal reflectance sharing method, see [1].

[WE EMPLOY A PARAMETRIC REFLECTANCE MODEL TO ESTIMATE DYNAMIC SURFACE REFLECTANCE OF HUMAN ACTORS.]

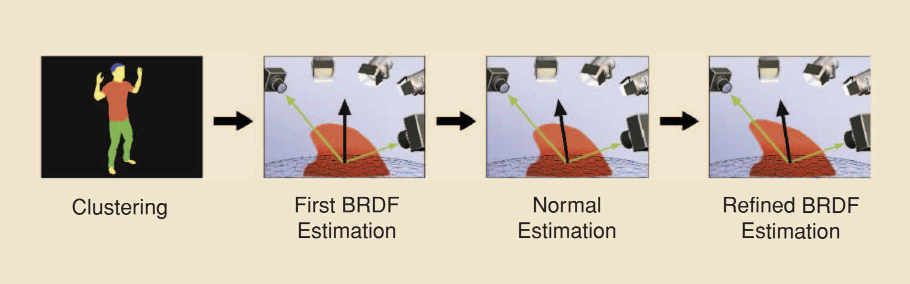
NORMAL ESTIMATION

Knowing the BRDF parameters, one can also refine the surface normal field by looking at the reflectance samples. During normal estimation we minimize the following extended version of the energy functional (3) in the local surface normal direction $\hat{n}(\vec{x})$ of each surface point $\vec{x}$:

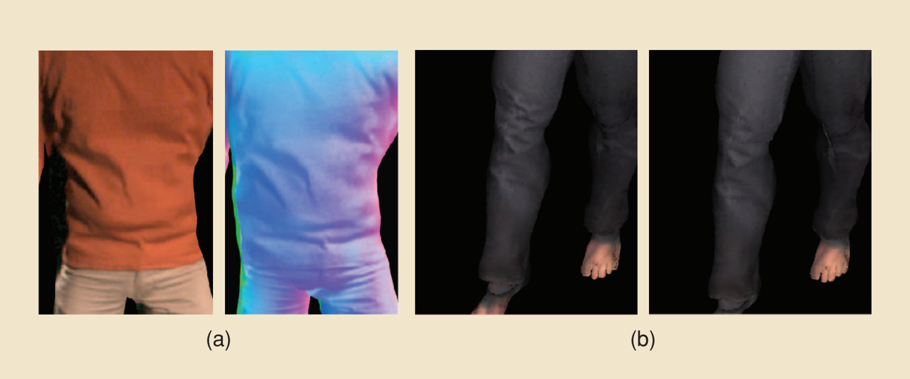
$$E_{\text{normal}}(\vec{x}, \hat{n}(\vec{x})) = \alpha E(\vec{x}, \rho(\vec{x})) + \beta \Delta(\hat{n}(\vec{x}))^{\gamma} . \tag{4}$$

The additional regularization term $\Delta(\hat{n}(\vec{x}))$ penalizes angular deviation from the default normal of the body model. This way, we also enforce convergence to a plausible local minimum of the nonconvex energy functional. The coefficients α and β are weighting factors summing to one, the exponent γ controls the penalty's impact. Appropriate values are found through experiments.

The above procedure forms a part of the actual BRDF estimation, but it is also used to estimate the time-varying normal field for each frame of a DS. In the latter case, we enforce temporal smoothness in the normal field. Figure 9 shows several close-up views of rendered relighatble free-viewpoint videos. Subtle time-varying surface details, such as wrinkles, are encoded in the completely passively captured surface normal fields.



[FIG8] Subsequent steps to estimate per-texel BRDFs.



[FIG9] (a) Comparison between an input frame and the corresponding normal map that we reconstructed. For rendering, the three components of the surface normals were encoded in the three color channels. One can see that subtle surface details have been captured at high accuracy. This level of accuracy also enables us to (b) faithfully reproduce time-varying geometric details, such as the wrinkles in the trousers around the knee.

RENDERING AND ENCODING

At rendering time, the body model is displayed in the sequence of captured body poses and the illumination equation is, in graphics hardware, evaluated for each rendered fragment of the model (Figure 10). We can render and relight free-viewpoint videos in real-time on commodity graphics hardware. For illumination, it is feasible to use both normal point or directional light sources, or even captured real-world illumination from HDR environment maps. Example renderings of actors under novel virtual lighting conditions can be seen in Figure 5(b) and (c). Even subtle surface details are faithfully reproduced in the synthetic lighting environment (Figure 9).

As before, time-varying scene geometry is encoded as a stream of pose parameters, triangle mesh geometry has to be uploaded to the renderer just once. The static BRDF coefficients are stored in parameter textures. The time-varying normal maps are encoded as a stream of textures. To correctly reproduce shifting apparel despite a fixed set of material parameters, cloth shifting information is provided to the renderer as a stream of warped texture coordinates. Figure 10 shows the working principle of our renderer, illustrating the rendering process and the underlying scene representation.

AN EXACT MATCH IS NEITHER NEEDED FOR CONVINCING DYNAMIC TEXTURING NOR FOR REFLECTANCE ESTIMATION.

We make use of the OpenEXR format to encode our 1024 × 1024 pixel texture images. Conveniently, OpenEXR allows us to store images in floating point format, and provides us with efficient methods for lossless compression. Furthermore, we can store the full dynamic range of the captured reflectance data and also display it by capitalizing on the floating point pipelines of state-of-the art graphics processors. Compared to streaming eight images for for each frame of the original free-viewpoint video method, only streaming two texture images per frame makes a big difference. While a sequence with 330 frames requires 1.6 GB of storage space in the original framework, the new encoding reduces it to around 800 MB (including geometry).

RESULTS

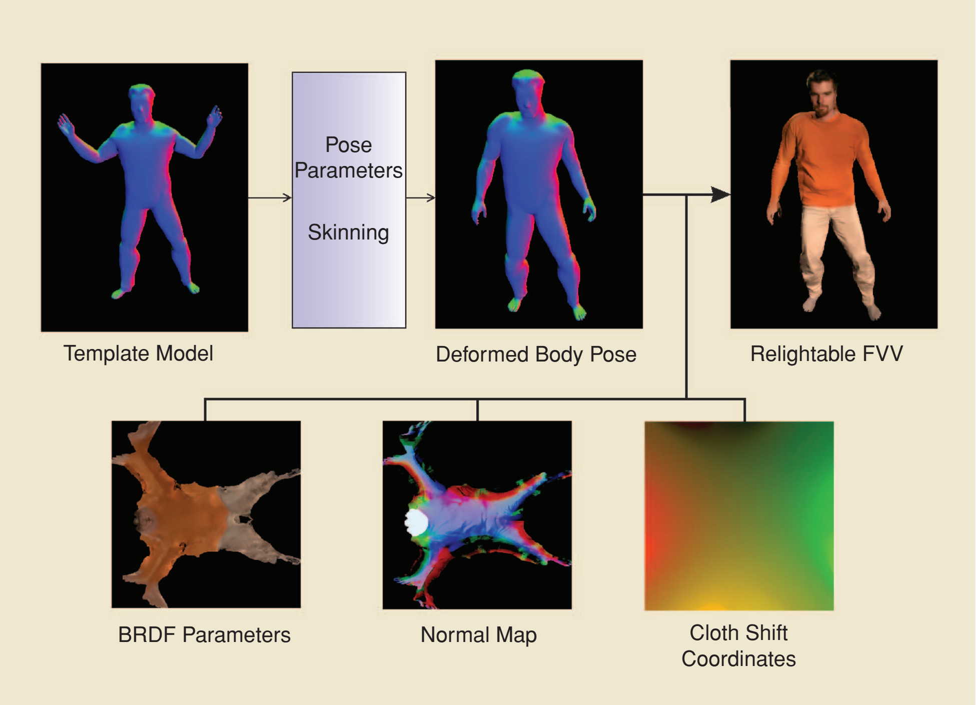
We have tested our free-viewpoint video method with dynamic texture generation on a variety of input sequences showing several motions ranging from simple gestures to ballet dancing. The sequences are typically 100–400 frames long. Our reconstructed videos reproduce both motion and appearance of the actors faithfully [4]. Subtle time-varying surface details are nicely reproduced. Figure 5(a) shows several screen shots of freeze-frame visualizations, i.e., the animation was held, and the user can freely fly around the scene. Even on a comparably old XEON 1.8 GHz GPU featuring a GeForce 3 GPU, a frame rate of 30 fps easily can be achieved. On a state-of-the art machine, motion estimation times of 1 fps are feasible. Fitting times below one second can be reached by employing a parallel implementation [28].

The data for relightable free-viewpoint video were capture with our new camera setup. We recorded a large database of subjects, apparel types, and motion styles. The typical length of a RS sequence is around 300 frames, the length of the employed motion sequences is 300–500 frames. All the results shown in this article were created with the Phong reflectance model.

Our dynamic scene description allows us to render photorealistically human actors under both artificial and real-world illumination that has been captured in high-dynamic range environment maps [see Figure 5(d)]. Even with realistically cast shadows, relightable 3-D videos can be displayed in real-time on state-of-the-art commodity graphics hardware. We can also implant actors into virtual environments as they are commonly used in computer games, such as a little pavilion with mirroring floor, Figure 5(c). Our dynamic reflectometry method faithfully captures time-varying surface details, Figure 9(a). By this means, they can be displayed realistically under varying artificial lighting conditions, Figure 9(b).

Reflectance estimation typically takes one hour on a Pentium IV 3.0 GHz. Normal estimation takes approximately 50 s per time step, and it can be parallelized to bring the computation time down. Optional input frame warping takes around 10 s for one pair of reference image and reprojected image. Cloth shift compensation accounts for an additional 35 s of computation time for one time step of video.

We have validated our dynamic reflectometry method both visually and quantitatively via comparison to ground truth image data and reflectance descriptions obtained with



[FIG10] Rendering pipeline: At each time step, the template model is deformed and brought into the correct pose. The deformed model along with BRDF parameters, normal map and warped texture coordinates is used to render the human actor under novel lighting conditions.

laser-scanned geometry. Material parameters can be recovered with reconstruction errors of less than 2%. For a detailed elaboration on these evaluations, see [26].

All of the presented model-based algorithms are subject to a couple of limitations. General restrictions are that we cannot easily reproduce background scenes and need a separate model for each type of subject to be captured. Furthermore, although we can handle normal every-day apparel, we can not account for loose apparel whose surface can deviate almost arbitrarily from the body model. Sometimes, we observe small rendering artifacts due to undersampling (e.g., on the underneath of the arms). However, for the relighting pipeline, we have verified that the application of an RS sequence showing several rotation motions with different body postures almost completely solves this problem. For a more detailed discussion on individual techniques, see the referenced papers.

Despite these limitations, we have presented an effective combination of algorithmic tools that allows for the creation of realistic dynamic scene descriptions with both nonrelighatble and relightable surface appearance.

FUTURE DIRECTIONS

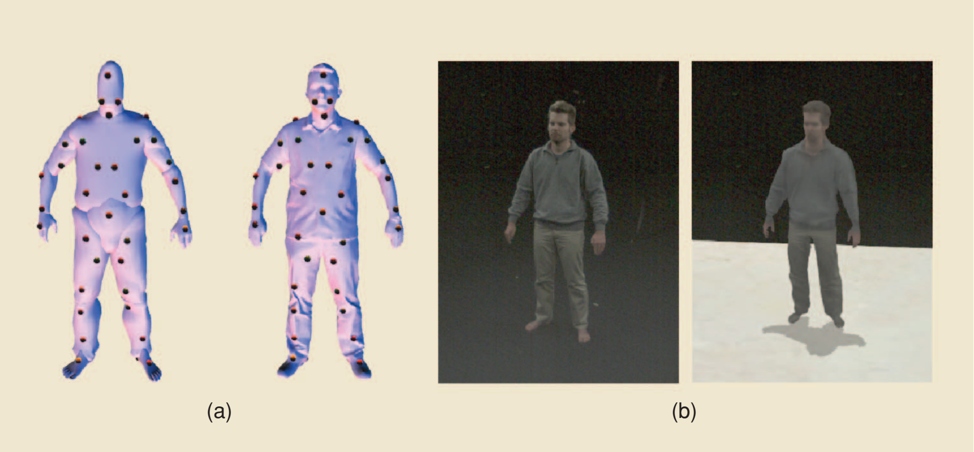
The commitment to a parameterized body model enables us to make the inverse problems of motion estimation and appearance reconstruction tractable. However, a model-based approach also implies a couple of limitations. Firstly, a template model is needed for each type of object that we want to record. Secondly, we currently can not handle people wearing very loose apparel. Furthermore, while a relatively smooth template model enables easy fitting to a wide range of body shapes, more detailed geometry specific to each actor would improve rendering quality even more. For instance, it would be intriguing to have a method at hand that enables us to make a high-quality laser scan follow the motion of the actor in each video frame without having to manually design skeleton models or surface skinning parameters.

To achieve this goal, we have developed a method that enables us to make our moving template model drive the motion of a high-quality laser scan of the same person. The user only needs to mark a handful of correspondences between triangles on the moving template and triangles on the target mesh, Figure 11(a). The transformations of the marked triangles on the source are mapped to their counterparts on the high-quality scan. Deformations for all the other triangles are interpolated on the surface by means of a harmonic field. The surface of the appropriately deformed scan at each time step is computed by solving a Poisson system. Our framework is based on the principle of differential mesh editing and only requires the solution of simple linear systems to map poses of the template to the target mesh. As an additional benefit, our algorithm implicitly solves the motion retargeting problem and automatically generates convincing nonrigid surface deformations. Figure 11(b) shows an example where we mapped the motion of our moving template model onto a high-quality static laser scan. This way, we can easily use detailed dynamic scene geometry as our underlying shape representation. For details on the correspondence specification and the mesh-deformation framework, see [7].

Recently, we extended the approach mentioned above such that we now are able to track the motion of laser scans in a marker-free manner directly from video streams. By this means, we will be able in future to reconstruct model-based 3-D videos even of people wearing wide apparel and of subjects other than humans [6].

CONCLUSIONS

We presented a tutorial-like compendium of approachs to capture, reconstruct, render and encode high-quality 3-D videos of human actors. The commitment to an a priori body model enables us to find efficient solutions to the above problem areas, even if only a handful of input cameras is used. Our free-viewpoint video approach with dynamic textures is among the first methods in the literature capable of such high-quality real-time renderings of virtual humans. Similarly, we have developed the first completely passive approach to capture dynamic and fully relightable representations of such complex real-world scenes. In the future, we plan to further investigate improved geometry and appearance reconstruction approaches from unmodified input video footage. For instance, we plan to investigate new ways to incoporate high-quality static laser-scans into our dynamic framework.



[FIG11] (a) The motion of the template model (l) is mapped onto target (r) by only specifying correspondences between individual triangles. (b) We can now use the moving laser scan instead of the moving template model in our free-viewpoint video pipeline. The image on the left is an input frame, the image on the right the free-viewpoint video with laser-scanned geometry.

ACKNOWLEDGMENTS

Parts of this work were supported by EU 3-DTV NoE project 511568.

AUTHORS

Christian Theobalt (theobalt@cs.stanford.edu) is a visiting assistant professor in the Department of Computer Science at Stanford University. He is also the head of the research group 3-D video and Vision-based Graphics, in the Max-Planck-Center for Visual Computing and Communication (Saarbrücken/ Stanford). He received his M.Sc. degree in artificial intelligence from the University of Edinburgh, Scotland, and his Diplom (M.S.) degree in computer science from Saarland University, Saarbrücken, Germany, in 2000 and 2001, respectively, and his Ph.D. (Dr. -Ing.) from Saarland University in 2005. His research interests include free-viewpoint and 3-D video, marker-less optical motion capture, 3-D computer vision, IBR, computer animation and physically based rendering.

Naveed Ahmed (nahmed@mpi-inf.mpg.de) received the B.S. and M.Sc. degrees in computer science from University of Karachi and Saarland University in 2001 and 2004, respectively. He is currently a Ph.D. candidate in Hans-Peter Seidel's Computer Graphics Group at MPI Informatik, Saarbrücken, Germany. His research interests include video-based rendering and relighting.

Gernot Ziegler (gziegler@mpi-inf.mpg.de) is currently a Ph.D. candidate in Hans-Peter Seidel's Computer Graphics Group at MPI Informatik, Saarbrücken, Germany. His areas of research are 3-D video coding technology and GPU based image processing/rendering (visual computing).

Hans-Peter Seidel (hpseidel@mpi-inf.mpg.de) is the scientific director and chair of the computer graphics group at the Max-Planck-Institut (MPI) Informatik and a professor of computer science at Saarland University. He has published 200 technical papers in the field and has lectured widely. He has received grants from a wide range of organizations, including the German National Science Foundation (DFG), the German Federal Government (BMBF), the European Community (EU), NATO, and the German-Israel Foundation (GIF). In 2003, he was awarded the Leibniz Preis, the most prestigious German research award, from the German Research Foundation (DFG). He is the first computer graphics researcher to receive this award.

REFERENCES

[1] N. Ahmed, C. Theobalt, and H.-P. Seidel, "Spatio-temporal reflectance sharing for relightable 3-D video," in *Proc. Mirage 2007*, pp. 46–58.

[2] S. Boivin and A. Gagalowicz, "Image-based rendering of diffuse, specular and glossy surfaces from a single image," in *Proc. SIGGRAPH'01*, 2001, pp. 107–116.

[3] R.L. Carceroni and K.N. Kutulakos, "Multi-view scene capture by surfel sampling: From video streams to non-rigid 3-D motion shape & reflectance," in *Proc. ICCV*, 2001, pp. 60–67.

[4] J. Carranza, C. Theobalt, M.A. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," *Proc. SIGGRAPH'03*, vol. 22, no. 3, pp. 569–577, July 2003.

[5] E. de Aguiar, C. Teobalt, M.A. Magnor, and H.-P. Seidel, "Reconstructing human shape and motion from multi-view video," in *Proc. CVMP'05*, London, U.K., Dec. 2005, pp. 42–49.

[6] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel, "Marker-less deformable mesh tracking for human shape and motion capture," in *Proc. IEEE CVPR*, Minneapolis, USA, June 2007, pp. 1–8.

[7] E. de Aguiar, R. Zayer, C. Theobalt, M. Magnor, and H.-P. Seidel, "Video-driven animation of human body scans," in *Proc. IEEE 3-DTV Conf.*, Kos Island, Greece, 2007, pp. 1–4.

[8] P. Einarsson, C.-F. Chabert, A. Jones, W.-C. Ma, B. Lamond, T. Hawkins, M.B., S. Sylwan, and P. Debevec, "Relighting human locomotion with flowed reflectance fields," in *Proc. Eurographics Symp. Rendering*, 2006, pp. 183–194.

[9] D. Goldman, B. Curless, A. Hertzmann, and S. Seitz, "Shape and spatially-varying brdfs from photometric stereo," in *Proc. ICCV*, 2004, pp. 341–448.

[10] M.H. Gross, S. Würmlin, M. Näf, E. Lamboray, C.P. Spagno, A.M. Kunz, E. Koller-Meier, T. Svoboda, L.J. Van Gool, S. Lang, K. Strehlke, A.V. Moere, and O.G. Staadt, "Blue-c: A spatially immersive display and 3-D video portal for telepresence," *Proc. SIGGRAPH'03*, vol. 22, no. 3, pp. 819–827, 2003.

[11] T. Hawkins, A. Wenger, C. Tchou, A. Gardner, F. Göransson, and P. Debevec, "Animatable facial reflectance fields," in *Proc. EGSR*, 2004, pp. 309–319.

[12] E. Lafortune, S.-C. Foo, K.E. Torrance, and D.P. Greenberg, "Non-linear approximation of reflectance functions," in *Proc. SIGGRAPH'97*, 1997, pp. 117–126.

[13] H. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel, "Image-based reconstruction of spatial appearance and geometric detail," *ACM Trans. Graphics*, vol. 22, no. 2, p. 27, 2003.

[14] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. SIGGRAPH'96*, 1996, pp. 31–42.

[15] M. Li, H. Schirmacher, M.A. Magnor, and H.-P. Seidel, "Combining stereo and visual hull information for on-line reconstruction and rendering of dynamic scenes," in *Proc. IEEE Multimedia Signal Processing*, 2002, pp. 9–12.

[16] M. Magnor, P. Ramanathan, and B. Girod, "Multi-view coding for image-based rendering using 3-D scene geometry," *IEEE T-CSVT*, vol. 13, no. 11, pp. 1092–1106, Nov. 2003.

[17] T. Matsuyama and T. Takai, "Generation, visualization, and editing of 3-D video," in *Proc. 3-DPVT'02*, 2002, p. 234ff.

[18] W. Matusik, C. Buehler, R. Raskar, S.J. Gortler, and L. McMillan, "Image-based visual hulls," in *Proc. SIGGRAPH'00*, 2000, pp. 369–374.

[19] W. Matusik and H. Pfister, "3-D TV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes," in *Proc. SIGGRAPH'04*, 2004, pp. 814–824.

[20] W. Matusik, H. Pfister, M. Brand, and L. McMillan, "A data-driven reflectance model," *Proc. SIGGRAPH'03*, vol. 22, no. 3, pp. 759–769, 2003.

[21] B.-T. Phong, "Illumnation for computer generated pictures," *Commun. ACM*, vol. 16, no. 6, pp. 311–317, 1975.

[22] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for inverse rendering," in *Proc. SIGGRAPH'01*, 2001, pp. 117–128.

[23] H. Rushmeier, G. Taubin, and A. Guéziec, "Applying shape from lighting variation to bump map capture," in *Proc. EGSR'97*, June 1997, pp. 35–44.

[24] J. Starck and A. Hilton, "Surface capture for performance based animation," *IEEE Comput. Graph. Appl.*, vol. 27, no. 3, pp. 21–31, 2007.

[25] C. Theobalt, N. Ahmed, H. Lensch, M. Magnor, and H.-P. Seidel, "Seing people in different light: Joint shape, motion and reflectance capture," *IEEE TVCG*, vol. 13, no. 4, pp. 663–674.

[26] C. Theobalt, N. Ahmed, H. Lensch, M.A. Magnor, and H.-P. Seidel, "Enhanced dynamic reflectometry for relightable free-viewpoint video," Max-Planck-Institut fuer Informatik, Saarbrücken, Germany, Res. Rep. MPI-I-2006-4-006, 2006.

[27] C. Theobalt, J. Carranza, M. Magnor, and H.-P. Seidel, "Enhancing silhouette-based human motion capture with 3-D motion fields," in *Proc. of Pacific Graphics'03*, Canmore, Canada, Oct. 2003, pp. 185–193.

[28] C. Theobalt, J. Carranza, M. Magnor, and H.-P. Seidel, "A parallel framework for silhouette-based human motion capture," in *Proc. VMV'03*, Munich, Germany, Nov. 2003, pp. 207–214.

[29] M. Waschbüsch, S. Würmlin, D. Cotting, F. Sadlo, and M. Gross, "Scalable 3-D video of dynamic scenes," in *Proc. Pacific Graphics*, 2005, pp. 629–638.

[30] A. Wenger, A. Gardner, C. Tchou, J. Unger, T. Hawkins, and P. Debevec, "Performance relighting and reflectance transformation with time-multiplexed illumination," in *Proc. SIGGRAPH'05*, vol. 24, no. 3, pp. 756–764, 2005.

[31] S. Würmlin, E. Lamboray, M. Waschbüsch, P. Kaufmann, A. Smolic, and M. Gross, "Image-space free-viewpoint video," in *Proc. Vision, Modeling, Visualization (VMV) 2005*, 2005, pp. 453–460.

[32] G. Ziegler, H. Lensch, N. Ahmed, M.A. Magnor, and H.-P. Seidel, "Multi-video compression in texture space," in *Proc. ICIP'04*, 2004, pp. 2467–2470.

[33] G. Ziegler, H. Lensch, M. Magnor, and H.P. Seidel, "Multi-video compression in texture space using 4-D SPIHT," in *Proc. IEEE MMSP*, 2004, pp. 39–42.

[34] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *Proc. SIGGRAPH'04*, vol. 23, no. 3, pp. 600–608, 2004.