EUROGRAPHICS 2010 / T. Akenine-Möller and M. Zwicker
(Guest Editors)

Volume 29 (2010), Number 2

Articulated Billboards for Video-based Rendering

Marcel Germann[1], Alexander Hornung[1], Richard Keiser[2], Remo Ziegler[2], Stephan Würmlin[2], Markus Gross[1]

[1]ETH Zurich, Switzerland. [2]LiberoVision AG, Switzerland

Abstract

We present a novel representation and rendering method for free-viewpoint video of human characters based on multiple input video streams. The basic idea is to approximate the articulated 3D shape of the human body using a subdivision into textured billboards along the skeleton structure. Billboards are clustered to fans such that each skeleton bone contains one billboard per source camera. We call this representation articulated billboards.

In the paper we describe a semi-automatic, data-driven algorithm to construct and render this representation, which robustly handles even challenging acquisition scenarios characterized by sparse camera positioning, inaccurate camera calibration, low video resolution, or occlusions in the scene. First, for each input view, a 2D pose estimation based on image silhouettes, motion capture data, and temporal video coherence is used to create a segmentation mask for each body part. Then, from the 2D poses and the segmentation, the actual articulated billboard model is constructed by a 3D joint optimization and compensation for camera calibration errors. The rendering method includes a novel way of blending the textural contributions of each billboard and features an adaptive seam correction to eliminate visible discontinuities between adjacent billboards textures.

Our articulated billboards do not only minimize ghosting artifacts known from conventional billboard rendering, but also alleviate restrictions to the setup and sensitivities to errors of more complex 3D representations and multiview reconstruction techniques. Our results demonstrate the flexibility and the robustness of our approach with high quality free-viewpoint video generated from broadcast footage of challenging, uncontrolled environments.

1. Introduction

Image-based rendering (IBR) has been introduced in the pioneering work of Levoy et al. [LH96] and Gortler et al. [GGSC96]. The basic goal is simple: IBR strives to create a sense of a 3D real-world scene based on captured image data. Many subsequent works have explored the theoretical foundations, e.g., the dependency of geometry and images in respect to a minimal sampling requirement [CCST00], or developed more efficient and less restrictive implementations [BBM*01]. One important general insight from these works is that a sufficiently accurate geometric proxy of the scene reduces the number of required input images considerably.

A small number of input views is an important prerequisite in order to apply IBR in real-world environments and applications. One prominent example is sports broadcasting, where we observe a growing demand for free-viewpoint replay for scene analysis [Lib]. However, for these and most other non-studio applications, IBR should ideally work based on existing infrastructure such as manually operated TV cameras. This poses the fundamental question how we can *robustly* generate a sufficiently accurate geometric proxy, despite the wide-baseline cameras, uncontrolled acquisition conditions, low texture quality and resolution, and inaccurate camera calibration. These problems become even more severe for processing video sequences instead of still images. Under these challenging real-world conditions, classical 3D reconstruction techniques such as visual hulls [MBR*00] or multi-view stereo [Mid09] are generally inapplicable. Due to the involved difficulties, one of the currently most popular approaches in this field is still the use of simple planar billboards [HS06], despite the unavoidable visual artifacts such as ghosting.

In this paper, we present a novel geometrical representation which is specifically suited for reconstruction and high quality video rendering of human subjects under the above conditions. Our key observation is that the 3D pose and shape of a character can be well captured by an articulated subdivision of the body into simple geometric primitives: *articulated billboards*. Instead of relying on accurate silhouette information for computing the visual hull or stereo correspondences, our representation requires an estimate of the 2D pose of a subject in the input views. We will show

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

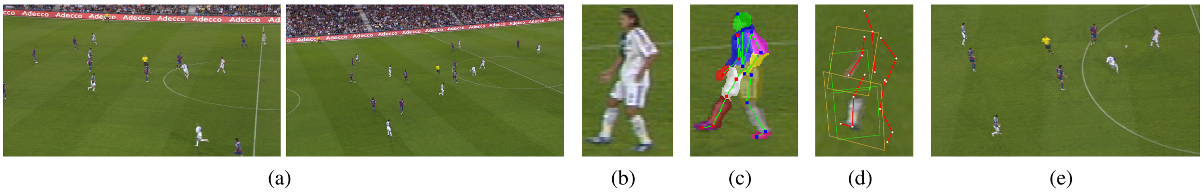


Figure 1: *Overview of our method. (a) Two wide-baseline input video frames of a soccer match. (b) Zoom on one of the players. (c) We first compute the subject's 2D pose in the input views and a segmentation into the different body parts. (d) A multi-view optimization then generates a 3D articulated billboard model. For clarity we show only a subset of the billboards in this example. (e) With the articulated billboard models photo-realistic views from a large range of novel viewpoints can be rendered.*

in this work how this can be achieved in a simple and efficient manner by a semi-automatic, data-driven algorithm. From the pose it is then possible to construct a 3D articulated billboard model, which is a faithful representation of the subjects geometry and which allows for photo-realistic free-viewpoint video. The novel technical contributions of our work are

- Articulated billboards, a novel shape representation for free-viewpoint video of human characters under challenging acquisition conditions.
- Semi-automatic, data-driven 2D pose estimation based on approximate silhouettes.
- Automatic segmentation of body parts by 3D template fitting and learning of color models.
- Generation of the articulated billboard model by 3D pose optimization and seam correction for optimal texture consistency.
- GPU-based, pixel-accurate blending and rendering for realistic and efficient view synthesis.

Applications for articulated billboards are multi-view videos of dynamic scenes with humans captured in uncontrolled environments. We will demonstrate in this paper that even from as few as two conventional TV camera images, a scene can be rendered at a high quality from virtual viewpoints where no source camera was recording.

2. Related Work

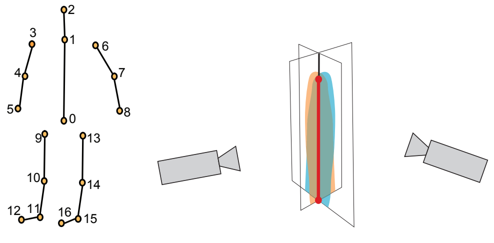
A variety of different 3D representations and rendering methods exists that use images or videos as source. Most of them are tightly connected to particular acquisition setups.

If many cameras with different viewpoints are available, the light field [LH96] of the scene can be computed, which represents the radiance as a function of space. Buehler et al. [BBM*01] generalize this approach to include geometric proxies. The Eye-Vision system used for Super Bowl [Eye09] uses more than 30 controlled cameras for replays of sports events. The method by Reche et al. [RMD04] for trees requires 20-30 images per object. A recent approach by Mahajan et al. [MHM*09] uses gradient-based view interpolation. In contrast to these methods, our method does not require a dense camera placement.

Many methods additionally use range data or depth estimation in their representation. Shade et al. [SGwHS98] use estimated depth information for rendering with layered depth images. Waschbüsch et al. [WWG07] use color and depth to compute 3D video billboard clouds, that allow high quality renderings from arbitrary viewpoints. Pekelny and Gotsman [PG08] use a single depth sensor for reconstruction the geometry of an articulated character. Whereas these methods require either depth data or accurate and dense silhouettes, this is not available in uncontrolled scenes with only a few video cameras and weak calibrations.

Several methods for template-based silhouette matching were proposed for controlled studio setups [CTMS03, VBMP08, dAST*08]. For free-viewpoint rendering, the camera images are blended onto the surface of a matched or deformed template model. However, these methods require accurate source images from studio setups whereas articulated billboards can be used with sparsely placed and inaccurately calibrated cameras. In these situations, the geometry of articulated billboards is much more robust against errors than, e.g., a full template body model where the texture has to be projected accurately onto curved and often thin (e.g. an arm) parts. Moreover, the generally required highly tessellated 3D template models are not efficient for rendering the often small subjects with low texture quality and resolution. Debevec et al. [DTM96] proposed a method that uses stereo correspondence with a simple 3D model. However, it applies to architecture and is not straight-forward extendable to articulated figures without straight lines.

Recently, improved methods for visual hulls, the conservative visual hull and the view dependent visual hull, showed promising results [GTH*07, KSHG07]. However, these methods are based on volume carving that requires selected camera positions to remove non-body parts on all sides of the subject. Our method does not require a special camera setting and can already be used with only two source cameras to show, e.g., a bird's eye perspective from a viewpoint above the positions of all cameras. Recent work by Guillemaut et al. [GKH09] addresses many challenges for free-viewpoint video in sports broadcasting by jointly optimizing scene segmentation and multi-view reconstruction. Their approach is leading to a more accurate geometry than the visual hull, but still requires a fairly big number of quite densely placed cameras (6-12). We compare our method to their reconstruction results in Section 7.



(a) (b)

Figure 2: *(a) Skeleton structure used for our articulated billboard model. (b) Illustration of a single fan of two billboards and the corresponding source cameras.*

A simple method for uncontrolled setups is to blend between billboards [HS06] per subject and camera. However, such standard billboards suffer from ghosting artifacts and do not preserve the 3D body pose of a person due to their planar representation. The idea to subdivide the body into parts represented by billboards is similar in spirit to the billboard clouds representation [DDS03, BCF*05], microfacets [YSK*02, GM03] or subdivision into impostors [ABB*07, ABT99]. However, these methods are not suited for our target application, since they rely on controlled scenes, depth data or even given models. Lee et al. [LBDGG05] proposed a method to extract billboards from optical flow. However, they used generated input images from synthetic models with high quality.

Related to our approach is also the quite large body of work on human pose estimation and body segmentation from images. Here, we can only discuss the most relevant works. Efros et al. [EBMM03] have presented an interesting approach for recognizing human action at a distance with applications to pose estimation. Their method requires an estimate of the optical scene flow which is often difficult to estimate in dynamic and uncontrolled environments. Agarwal and Triggs [AT06], Jaeggli et al. [JKMG07], and Gammeter et al. [GEJ*08] present learning-based methods for 3D human pose estimation and tracking. However, the computed poses are often only approximations, whereas we require accurate estimations of the subject's joint positions. Moreover, we generally have to deal with a much lower image quality and resolution in our setting. We therefore present a semi-automatic, data-driven approach, since a restricted amount of user interaction is acceptable in many application scenarios if it leads to a considerable improvement in quality.

3. Overview

Our aim is to enable virtually unconstrained free-viewpoint rendering of human subjects from a small set of wide-baseline video footage (see Figure 1). This requires a shape and appearance model for rendering, which can be robustly generated from the videos despite of limited resolution and texture quality, inaccuracies in the camera calibration, or the complex occlusions that may occur for articulated bodies.

We propose a representation based on *articulated billboards*. The basis of this model is a 3D human skeleton structure (see Figure 2(a)). Every bone, represented by a 3D vector $\mathbf{b}_i$ and the position of its end-joint $\mathbf{x}_i$, corresponds to a major component of the body, e.g., the torso or the extremities. With each bone we associate a fan of billboards, which contains a billboard for every input image I_j of a subject (see Figure 2(b)). More specifically, for each I_j the corresponding billboard plane is defined by the joint $\mathbf{x}_i$, the bone direction $\mathbf{b}_i$, and the vector $\mathbf{b}_i \times (\mathbf{c}_j - \mathbf{x}_i)$, where $\mathbf{c}_j$ is the camera position of I_j. Hence, the billboards are aligned with the character bones and as orthogonal as possible to their associated input views.

The basic idea of our method is to compute a 3D pose of the articulated billboard model, i.e., a spatial joint configuration of the underlying skeleton structure, which brings its 2D projection into correspondence with the subject's pose in each input frame of the video. After this alignment, a texture map and alpha mask is generated for each billboard from its associated view. However, a fully automatic computation of a single 3D pose, which is perfectly consistent with all input views, is generally not possible in the presence of the above mentioned issues such as imperfect camera calibration or low texture resolution. Hence, we developed a semi-automatic, data-driven approach which operates in three consecutive phases: a 2D pose estimation and template-based image segmentation, the construction of the articulated 3D billboard model, and the actual rendering.

First, for the 2D pose estimation in each individual input view, we utilize a database of silhouettes, temporal motion coherence of subjects in the video, and motion capture data to assist the user in fast and accurate placement of joints. Given these 2D joint positions, a segmentation of the image into the different body parts, i.e., the torso or the limbs, is computed using a human template model in order to map image-pixels to billboards (see Section 4).

The second phase of the algorithm integrates the pose and texture information from all individual views and generates the final articulated billboard model for rendering. This processing step includes an optimization of the 3D joint positions and a compensation for camera calibration errors, which optimizes the texture overlap for each model segment, i.e., for each fan of billboards. A final alpha-mask and texture optimization eliminates visible seams and discontinuities between adjacent billboards (see Section 5).

The last step is the actual real-time rendering of novel views. Section 6 describes an algorithm for a fully GPU-based, view-dependent per-pixel blending scheme, which is optimized for rendering articulated billboard models efficiently while preserving the photorealism of the original input video.

4. Pose Estimation and Template-based Segmentation

In the first phase of the algorithm we compute an initial guess of the subject's joint positions in image space and a segmentation of the pixels into the different body parts.

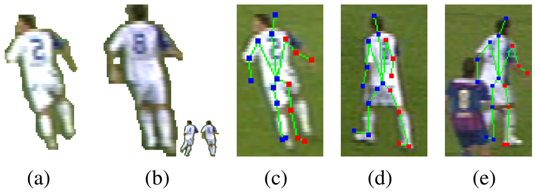


Figure 3: *Silhouette-based pose estimation. (a) Typical silhouette of a subject. (b) The 3 best matching poses from our database. Note that we also consider flipped images. (c) The corresponding 2D skeleton pose estimated from the best matching pose. (d) and (e) Further examples.*

For calibration of the intrinsic and extrinsic camera parameters we currently use the method of Thomas [Tho06]. As mentioned previously a fully automatic pose estimation and segmentation is very challenging due to the relatively low resolution and quality. Accordingly, we propose the following semi-automatic approach which minimizes the required user-interaction to only a few mouse-clicks. Then, given the joint positions, the segmentation of the subject's body parts is computed by fitting a human template model with a known segmentation to the input video frames.

4.1. 2D Pose Estimation

We assume that a coarse segmentation of the subject from the background is available, e.g., using chroma keying or background subtraction. Figure 3(a) shows a typical example of a segmented image in our application scenario. The basic idea to compute an initial guess of a subject's pose, i.e., the 2D positions of the skeleton joints, is to compare it to a database of silhouettes, for which the respective skeleton poses are known (see Figure 3(b)) .

First, for each view I_j, we normalize for differently sized subjects by re-sampling the silhouette on a 32×40 grid and stack the binary silhouette information at each grid point into a vector $\mathbf{v}_j \in [0,1]^n$, with $n = 32 \times 40$. Then, for each $\mathbf{v}_j$, our algorithm finds the best matching k entries in the database, which minimize the error

$$E_S = (1-\lambda)\frac{1}{n}\sum_{i=0}^{n-1}|\mathbf{v}_j(i) - \mathbf{w}(i)| + \lambda\frac{1}{m}\sum_{r=0}^{m-1}|\mathbf{p}_j(r) - \mathbf{q}(r)|, \quad (1)$$

where $\mathbf{w}$ is an entry in the database, $\mathbf{q}$ its corresponding 2D joint positions, and m is the number of skeleton joints. The vector $\mathbf{p}_j$ contains the joint coordinates from the previous video frame. The first term of Eq. (1) ensures a proper match of the silhouettes whereas the second term exploits temporal motion coherence of subject's in the video. This is of particular help to resolve left-right ambiguities in the silhouettes. The influence of the second term can be weighted by the value λ. For the first frame of a sequence we simply set $\lambda = 0$, for all other frames we used a value of $\lambda = 0.5$ for all our examples. The joint positions are processed in normalized coordinates with respect to the subject's bounding box. Using this error E_S, the $k = 3$ best matching silhouettes and their corresponding 2D joint positions for each single view I_j are retrieved from the database.

In order to select the most plausible 2D pose from each of these sets we run a multi-view optimization for each combination of poses: we compute the 3D rays from each camera $\mathbf{c}_j$ center through the retrieved joint positions in I_j. Then, we compute the 3D representative for each joint which is closest to the corresponding rays. Figure 4 shows an example with two cameras. The measure for the quality of a particular combination of poses is the accumulated sum of distances of each 3D joint from its respective rays. In order to make this procedure more robust to the often inaccurate camera calibration, this multi-view optimization also includes a simple correction step. For each silhouette we additionally compute a 2D offset in the image plane, which is optimized using the Levenberg-Marquardt algorithm. This calibration correction proved to be very effective: for some silhouette images the necessary 2D offset for minimizing the error measure can be as high as 8 pixels.

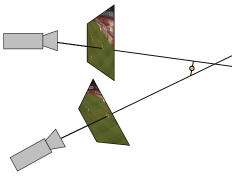


Figure 4: *3D joint estimation from two camera images.*

As demonstrated in Figure 3(c), this silhouette-based pose estimation and joint optimization generally provides a good guess of the subject's 2D joint positions in each view I_j. With a simple interface the user can then manually correct these positions by moving the joints (see Figure 5(a)). We refer to the accompanying video for a demonstration of this manual interaction step. After this joint refinement step the silhouette and joint positions are immediately added to our database. The increase of poses in the database has proven to lead to significantly better matches for new sequences. Note that, in application scenarios where no silhouette information is available at all, the user can resort to placing all joints manually. But even in this case the required interaction time per subject is generally only a few seconds.

4.2. 3D Template Fitting

Even with accurate 2D joints a robust segmentation of the image into the subject's body parts is still a difficult problem. Using a database of segmented silhouettes instead of the above binary silhouette segmentation is not a desirable option, since creating such a database would be extremely complex and time-consuming, and we could still not expect to always find sufficiently accurate matches.

Instead, our idea is to fit a generic, pre-segmented 3D template model to the images. This has the considerable advantage that we get a good starting solution for the segmentation process and that we can easily resolve occlusions. However, fitting a 3D model requires, for each particular input view, the computation of a 3D pose whose projection *perfectly* aligns with the 2D joints. A 3D pose leading to a perfect

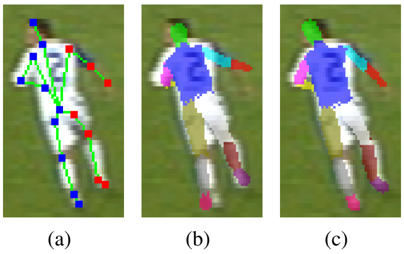


Figure 5: *3D template fitting: (a) Corrected joint positions. (b) Initial fitting of the pre-segmented 3D shape template using the method of Hornung et al. [HDK07]. (c) Our corrected fit which exactly matches the joint positions in (a).*

match in all views can often not be found due to calibration inaccuracies or slight joint misplacements. Therefore, we fit a 3D model per input view. A solution for computing an *approximate* 3D pose for articulated models from a single image has been presented by Hornung et al. [HDK07]. Given the 2D joint positions $\mathbf{x}_i$ for an image I_j, their approach uses a database of 3D motion capture data to find a set of 3D joint positions $\mathbf{X}_i$ whose projection approximately matches the 2D input joints (see Figure 5(b)). We provide a simple but effective modification to their algorithm for computing the required *accurate* fit.

The approximate 3D match can be deformed, such as to align with the 2D joints according to the following algorithm. Through each 3D joint $\mathbf{X}_i$, we create a plane parallel to the image plane of I_j. Then, we cast a ray from the camera center $\mathbf{c}_j$ through the corresponding target joint position $\mathbf{x}_i$ in I_j and compute its intersection with the plane. The 3D pose is then updated by moving each $\mathbf{X}_i$ to the respective intersection point and updating the 3D bone coordinate systems accordingly. The result is the required 3D pose which projects exactly onto the previously estimated 2D joints. The 3D template model can now be fitted to the image by deforming it according to this computed 3D pose using standard techniques for skeleton-based animation [LCF00] (see Figure 5(c)). Note that this algorithm generally does not preserve the limb lengths of the original 3D skeleton and therefore, enables an adaptation of the 3D template mesh to fit the subjects dimensions more accurately.

4.3. Segmentation of Body Parts

The fitted, pre-segmented template model does not perfectly segment the input frame I_j and might not completely cover the entire silhouette. Therefore, a refinement of the segmentation is done in three simple steps. In a first step, a color model is learned per body segment based on automatically selected *confident* pixels of the pre-segmented body parts (see Figure 6(a)). In a second step, the trained color model is used to label the *unconfident* pixels leading to a segmentation adjusted to the subjects body dimensions and silhouette (see Figure 6(b)). In a third step, a morphological closing operation removes outliers as depicted in Figure 6(c).

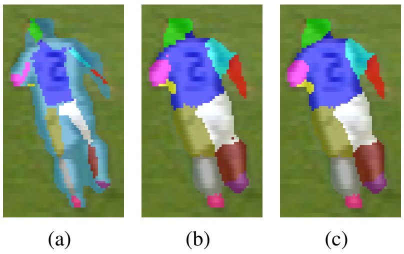


Figure 6: *Body segmentation. (a) Initial segmentation with safe pixels derived from the template model and unconfident boundary pixels. (b) Segmentation after labeling according to the trained color model. (c) Final segmentation after morphological removal of outliers.*

To determine the confident pixels, we project a slightly thinned and thickened version of the template model into the image and label the silhouette pixels accordingly. Pixels which receive the same label in both projections are marked as confident pixels and labeled with the corresponding body segment. *All* remaining pixels within the silhouette are labeled as unconfident as shown in Figure 6(a).

By learning the color model on-the-fly, we provide a robust segmentation algorithm being able to handle segmentation in uncontrolled environments. Changing lighting conditions, subject specific appearance or view dependent appearance can thus be handled reliably.

The pose estimation and segmentation procedure is performed for every view and input frame from which free-viewpoint renderings are to be generated. Note that our segmentation approach using successive 2D pose estimation and 3D template fitting automatically handles occluded body parts, is robust even for low image quality and resolution, and requires only a small amount of simple user interaction during the refinement of joint positions. We refer to the accompanying video for an example of this procedure.

5. Construction of the Articulated 3D Billboard Model

We use the computed 3D joint positions of Section 4.1 as an initial pose for the final articulated billboard representation. If a 3D joint of the articulated billboard model is not optimally positioned, the texture resulting from the rendering of all billboards of a billboard fan will not align (see Figure 7). In this section, we describe how the 3D joint positions can be optimized based on a quantitative measure of the alignment of the billboard textures.

In the following, we first define a scoring function for a position of a joint in one view and for one camera pair. This scoring function is then extended to several views and cameras. Using this scoring function and anthropometric constraints the 3D pose of the articulated billboard model is optimized. Finally, we will describe a seam correction which removes texture discontinuities between adjacent billboards.

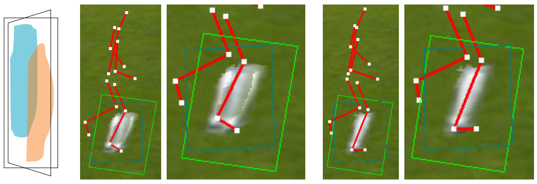

(a) (b) (c)

Figure 7: *(a) Illustration of a misaligned billboard fan. (b) Billboard fan before joint optimization. (c) Result after optimization. Note the improved texture alignment.*

5.1. Position Scoring

To score the quality of a joint position of an output view V, all billboards adjacent to this joint are evaluated. For each fan of billboards, the alignment of its billboards for a pair of input views (I_1, I_2) is scored by a pixel-wise comparison of the projected textures. For every output pixel p of V, the per-pixel score $s_{I_1,I_2}(p)$ is defined as

$$s_{I_1,I_2}(p) = \begin{cases} 1-\varepsilon(V_{I_1}(p),V_{I_2}(p)), & p \text{ active in } I_1 \text{ and } I_2 \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where $V_{I_j}(p)$ is the color contribution of a billboard associated with view I_j to pixel p. $\varepsilon(\cdot)$ is a color distance measure in RGB. The *active pixels* are defined as those pixels in the output view V which receive a valid color contribution from the input views I_1 and I_2. The segmentation generated in Section 4.3 is used to reliably resolve occlusion. The score for a joint in a view V is the normalized sum of all pixels

$$s_{I_1,I_2}(V) = \frac{\sum_{p\in V} s_{I_1,I_2}(p)n(p)}{\sum_{p\in P_v} n(p)}. \tag{3}$$

The normalization factor $n(p)$ is 1, if at least one of the two pixels is active and 0, otherwise. Thus, the scoring function measures the matching of texture values, while $n(p)$ penalizes non-aligned parts as in Figure 7(a). These pixel-wise operations are efficiently implemented on the GPU using fragment shaders.

For more than two input views, we define the score as a weighted average of all camera pairs, where the weight depends on the angle β_{I_1,I_2} between the respective viewing directions, with narrow angles receiving a higher weight:

$$s(V) = \frac{\sum\limits_{(I_1,I_2)\in\mathcal{I}} s_{I_1,I_2}(V)\omega(\beta_{I_1,I_2})}{\sum\limits_{(I_1,I_2)\in\mathcal{I}} \omega(\beta_{I_1,I_2})}, \tag{4}$$

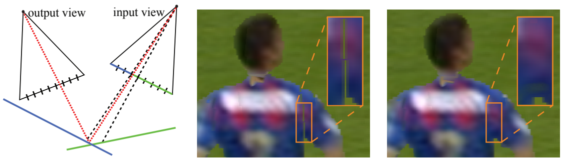
where $\mathcal{I}$ is the set of all pairs of input views and $\omega(\beta)$ is a Gaussian weight:

$$\omega(\beta) = e^{-\frac{\beta^2}{2\sigma^2}}. \tag{5}$$

The value for σ was empirically determined to be 0.32. Finally, the score of the joint position is the normalized sum of the scores in all evaluated views:

$$S_{\mathcal{V}} = \frac{1}{|\mathcal{V}|} \sum_{V\in\mathcal{V}} s(V), \tag{6}$$

where $\mathcal{V}$ is the set of all evaluated views.



(a) (b) (c)

Figure 8: *Seam correction. (a) Sampling errors in the segmentation mask cause cracks, e.g., the look-up of a pixel on the blue billboard ends up on the mask of the green billboard from an adjacent billboard fan. (b) Corresponding rendering artifact. (c) Result after our seam correction.*

5.2. 3D Pose Optimization

Since the scoring of the joint position depends on the evaluated views, we need a suitable set $\mathcal{V}$. In order to cover a reasonable range of viewing positions, we evaluate the scoring function at the camera positions of all input views and the virtual views in the center between each camera pair. For the position optimization of a joint, we evaluate $S_{\mathcal{V}}$ at spatially close candidate positions on a discrete, adaptive 3D grid. The grid is refined in a greedy manner around those candidate positions which achieve a higher score $S_{\mathcal{V}}$, until a given grid resolution is reached (empirically set to 1.2 cm).

To avoid degenerate configurations with billboard fans of zero length, we additionally consider the anthropometric consistency [NAS09] during the evaluation of each pose. A joint position receives a zero score if one of the following constraints does not hold:

- The joint is on or above the ground.
- Lengths of topologically symmetric skeleton bones (e.g., left/right arm) do not differ more than 10%.
- The lengths of adjacent bones are within anthropometric standards.
- Distances to unconnected joints are within anthropometric standards.

For the last two constraints, we use the 5th percentile of female subjects rounded down as minimal lengths and the 95th percentile of male subjects rounded up as maximal lengths.

This grid-search optimization process is iteratively repeated over the skeleton. In our experiments, we found that it typically converges after 4 iterations. See Figure 7 for an articulated billboard model before and after optimization.

5.3. Texture Seam Correction

Due to sampling of the billboards' segmentation masks during rendering with projective texturing (see Figure 8(a)), small discontinuities (visible cracks) between adjacent billboards might appear in the output view as shown in Figure 8(b). To overcome this problem, these *seam pixels* have to be rendered for both adjacent billboards. Therefore, we mark pixels as seam pixels in the input views if they cover billboards on two adjacent skeleton bones (e.g., pixel enclosed by dashed lines in Figure 8(a)).

To detect seam pixels, the segmentation mask is traversed for each input view. A pixel p is marked as seam pixel, if it fulfills both of the following conditions:

- At least one pixel p' in its 4-neighborhood has a different label but comes from the same subject
- $|\text{depth}(p) - \text{depth}(p')| < \varphi$

where $\text{depth}(\cdot)$ is the depth value at this pixel. The threshold φ distinguishes between occluding parts and connected parts. It was empirically set to $\varphi = 3$ cm. An example for the seam corrected segmentation mask and the resulting rendering improvement is shown in Figure 8(c).

6. Rendering

In the following we describe a procedure for photo-realistic rendering of articulated billboards. We designed this algorithm according to the general criteria defined by Buehler et al. [BBM*01]. Due to our challenging setting with calibration errors and very sparse camera positioning, our particular focus is on:

- *Coherent Appearance*: Adjacent billboards should intersect without cracks or disturbing artifacts and blend realistically with the environment.
- *Visual Continuity*: Billboards should not suddenly change or pop up when moving the viewpoint.
- *View Interpolation*: When viewing the scene from an original camera angle and position, the rendered view should reproduce that of the input camera.

Input to the rendering procedure are the articulated billboard model, the segmented input views $\mathcal{I}$ (Section 4.3) and the seams computed in Section 5.3. For each rendered output frame, the articulated billboards are sorted back-to-front for a proper handling of occlusions. In order to meet the above goals, we perform a per-pixel blending procedure. We separate between per-camera weights which are computed once per billboard and the final per-pixel weights.

6.1. Camera Blending Weights

For a smooth blending of the billboards associated with one fan of billboards, we use the same Gaussian weight as in Eq. (5). To achieve an interpolation at an original camera view, we introduce an attenuation function which ensures that all views from an original camera perspective are identical to the corresponding camera source images while still assuming a smooth transition between different views. The attenuation function is defined as $f(I_{\omega_{Max}}) = 1$ for the source view $I_{\omega_{Max}}$ with the highest value of $\omega(\cdot)$ and

$$f(I_{\omega_{Max}}) = 1 - exp\left(-\frac{d(V, I_{\omega_{Max}})^2}{2\sigma^2}\right) \tag{7}$$

for all other cameras I_j. $d(V, I_{\omega_{Max}})$ is the Euclidean distance from the viewers position to the camera position of view $I_{\omega_{Max}}$. The constant σ is empirically determined to be 1 meter, which is lower than the minimal distance between two cameras and thus does not lead to any discontinuities.

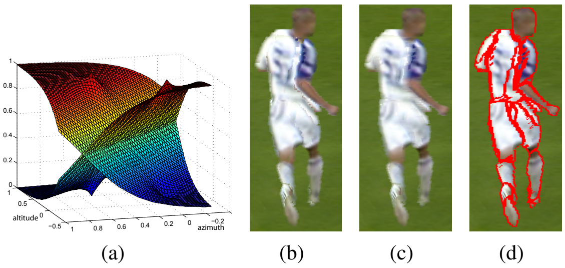


Figure 9: *Blending and smoothing. (a) Blending weight example for two cameras. The angles are the spherical coordinates of the view position. (b) Rendering without smoothing. (c) Adaptive smoothing enabled. (d) Marked discontinuities where smoothing has been applied.*

6.2. Per-Pixel Processing

The billboards of a billboard fan are blended per-pixel. As shown in Figure 8(a), a camera look-up in the corresponding segmentation mask of each billboard is performed. This determines if the current output pixel p is on the body part belonging to this billboard. If so, then the corresponding color contribution $V_{I_j}(p)$ from source view I_j and its alpha value $\alpha_{I_j}(p)$ can be added to the output view V. Otherwise, we set $\alpha_{I_j}(p) = 0$, i.e., transparent. The latter case also occurs when the corresponding body part is occluded in I_j and the color information should be taken from other cameras. The resulting color value $V(p)$ of the screen pixel is then

$$V(p) = \frac{\sum\limits_{I_j \in \mathcal{I}} V_{I_j}(p) w(I_j, p)}{\sum\limits_{I_j \in \mathcal{I}} w(I_j, p)} \tag{8}$$

with the per-pixel weights

$$w(I_j, p) = \alpha_{I_j}(p)\omega(\beta_{I_j}) f(I_{\omega_{Max}}). \tag{9}$$

This is done for all color channels separately. The resulting alpha value is

$$\alpha_V(p) = \begin{cases} \alpha_{I_{\omega_{Max}}}(p), & \text{if } w(I_{\omega_{Max}}, p) \neq 0 \\ \frac{\sum\limits_{I_j \in \mathcal{I}} \alpha_{I_j}(p) w(I_j, p)}{\sum\limits_{I_j \in \mathcal{I}} \alpha_{I_j}(p)\omega(\beta_{I_j})}, & \text{otherwise} \end{cases} \tag{10}$$

where the first case applies, if the closest camera is used for this pixel. Eq. (8) and Eq. (10) make sure that the color val-

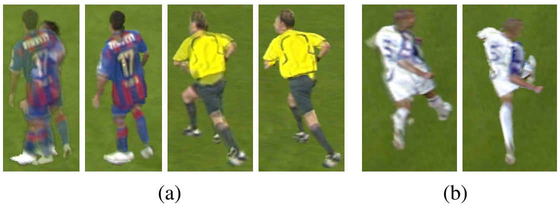


Figure 10: *(a) Direct comparison for a view in the middle of two source cameras. Each player is rendered with a standard billboard technique and with articulated billboards. The standard billboards exhibit considerably ghosting artifacts. (b) An example of a bird's eye perspective. While standard billboards are simply tilted (left), articulated billboards give an impression of the actual 3D pose.*

ues are blended such that the factors sum up to 1. However, the alpha values do not have to sum up to 1, e.g., if continuous alpha mattes are available instead of binary segmentation masks.

In addition to this, billboards seen at an oblique angle or from the backside, i.e., having a normal in an angle close to or more than 90 degrees away from the viewing direction, are simply faded out. For simplification, these factors are not shown in the equations.

An example for blending of intensities (i.e., one color channel) of two cameras is shown in Figure 9(a) where the azimuth and altitude angles are from spherical coordinates of the view position around the fan of billboards. The two peak points at $(0.0, 0.0)$ and $(0.5, 0.5)$ correspond to the positions of the source cameras. As it can be seen in the plot, when approaching these points the corresponding camera's weight increases to 1.0 and all other camera weights decrease to 0.0. Therefore, in this case only the source camera is used which results in the exact reproduction of the source image.

Finally, to prevent non smooth edges at the boundaries of a fan of billboards with respect to the background, other billboard fans, and at locations where other input views receive the highest weight (e.g., due to occlusions on a billboard), an additional Gaussian smoothing step is applied. This is done adaptively as a post-process only at discontinuities detected and stored while rendering the billboards. Figure 9 shows an example.

6.3. Implementation on the GPU

The per-pixel blending can be implemented efficiently on the GPU by using textures for the masks and fragment shaders for blending. We use the three color channels of the mask textures to store the subject ID, the billboard ID and whether it is a seam pixel or not. The blending is done in a fragment shader. Finally, the adaptive smoothing is implemented using multiple render targets such that the locations which require smoothing can be stored in the same render pass. In a second pass, this information is used for the smoothing.

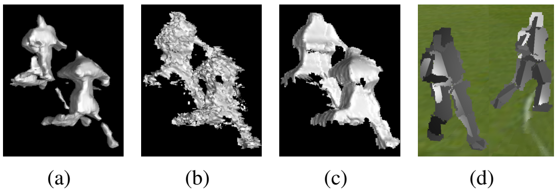


Figure 11: *Qualitative comparison. (a) Visual hulls cannot capture geometric detail and are sensitive to camera calibration errors. (b) Stereo reconstruction is problematic due to low texture resolution and noise. (c) The recent method of Guillemaut et al. [GKH09] improves the silhouette segmentation considerably. However, the shape reconstruction is still rather inaccurate, leading to ghosting artifacts. (d) The articulated billboard reconstruction (for a similar scene) captures the geometry much more faithfully. See the video for a 3D rendering. (a)-(c) Courtesy of [GKH09].*

7. Results

We applied our method to footage of real scenes of a soccer game, captured by TV cameras, which were used to broadcast the game in HD resolution (1920 × 1080, interlaced). The background (pitch, stadium) was simply blended onto large planes.

Figure 10(a) shows a direct comparison of our articulated billboards to standard billboard rendering with one billboard for each player and input camera. Whereas simple billboarding suffers from duplications of arms and legs, our method keeps the 3D perception, e.g., self-occlusions, correct due to the adaptive geometry. Even in the worst case, i.e., the view from a position in the middle of two source cameras, the overall body pose is preserved. In Figure 10(b) a bird's eye view is depicted. It shows how standard billboarding simply tilts the large billboards, whereas articulated billboards provide a realistic rendering of the entire pose.

Figure 12 shows a comparison of our rendering to ground truth data using a leave-one-out test, which shows that our method is able to realistically reproduce the visual appearance of a scene from only two input views even at distant novel viewing positions.

A qualitative comparison of the shape representation with articulated billboards to visual hulls, stereo reconstruction, and a recent method by Guillemaut et al. [GKH09] is shown in Figure 11. Due to its articulated structure with a planar geometric proxy for each limb and input view, our method generally provides a better geometrical approximation of the subject's shape, in particular for challenging and inaccurate input data as in our application setting. The benefit is an improved rendering quality with less ghosting artifacts.

Figure 13 shows virtual views of two different scenes. Due to our articulated billboard structure, the 3D poses of the players remain consistent even for extreme viewpoint changes and the viewer gets a clear impression of the actual positions of different body parts.

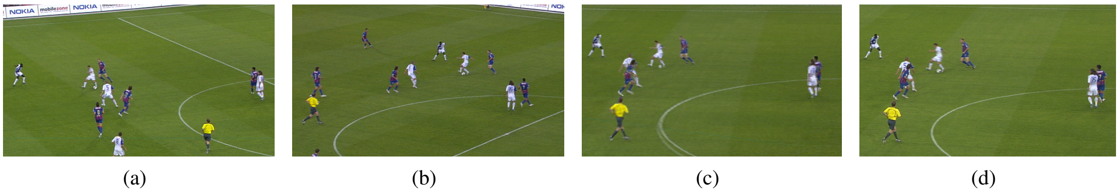

(a) (b) (c) (d)

Figure 12: *Leave-one-out example. (a) and (b) are two wide-baseline input views. (c) From these two input views we computed a virtual view of the scene from a novel viewing position. (d) Ground-truth view of an actual camera at this position. Note that the ghosting on the ground plane in (c) stems from the camera calibration.*

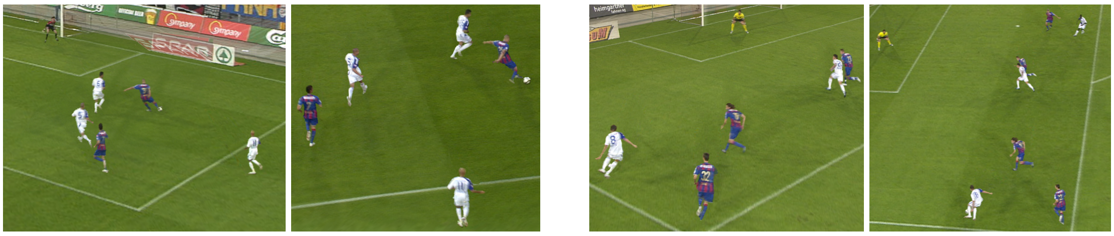


Figure 13: *Novel views of a soccer game. The 3D pose and realism of the players is preserved for strongly varying viewpoints, e.g., the arms and the legs within the rotation. More examples and video sequences are shown in the accompanying video.*

The accompanying video shows 4 examples of free-viewpoint video rendered with our system in real-time at HD resolution (> 40fps). Two of these scenes feature video replay from the virtual viewing positions. To our knowledge this has not been shown before in a similarly challenging application scenario with only two cameras. The interaction time per subject and camera view is only about 10 seconds. Hence, even for large scenes with 14 subjects and 2 camera views, the manual interaction time per video frame is generally only 4 to 5 minutes.

In the video, subjects visible only in one camera have been removed from the scene. Optionally, the automatic segmentation of body parts can be manually corrected in ambiguous regions. We performed slight corrections in the single-frame examples for maximum quality. Note, however, that we did not use such a manual correction for the dynamic scenes with video replay.

The timing for automatic processing of our system is generally a matter of seconds. The most expensive step is the joint position optimization which takes maximally 30 seconds per subject for three cameras. Therefore, the overall processing time including manual and automatic steps is about 6 minutes for a single frame scene with two cameras. Since we currently do not use temporal information extensively for dynamic scenes, this results in about 3 hours for a sequence with 30 frames on a standard workstation with an NVIDIA 8800GTX graphics card. In the future we plan to speed this up by a 3D joint computation exploiting temporal coherence.

The silhouette/skeleton database consists of currently 1966 silhouettes (983 poses and their mirrored versions) from about 20 different players. Since it is set up as a bootstrapping process, the database is refined on the fly using the manually corrected pose estimations. These new poses are continuously integrated into the database to improve the pose estimation for novel scenes.

Limitations and Future Work Designed for low quality input data recorded with large base-lines, articulated billboards are an optimal approximation if the player has a height up to about 200 pixels in the original as well as in the rendered image. If higher resolutions or dense camera setups are used, more complex primitives should be used to prevent ghostings within a billboard fan.

Similar to previous work, the view range is naturally limited to the vicinity of the source cameras. Nevertheless, as shown in Figure 12, our method features a quite large viewing range even from only two input cameras. As shown in Figure 3(e), occlusions reduce the quality of the pose estimation and thus increase the manual interaction in such cases. During rendering, occlusions are only a problem at pixels which are not visible in any of the cameras. For these cases we plan to apply a hole filling algorithm. Additionally, due to our current depth sorting, flickering artifacts can appear. Therefore, we will investigate the computation of a per-pixel depth based on the billboard planes.

Finally, we would like to investigate whether our semi-automatic pose estimation can be further automated, e.g., using techniques discussed in Section 2 such as [EBMM03]. In order to improve temporal coherence, we plan to investigate global optimization over all video frames.

8. Conclusion

We presented a representation and rendering method for human bodies suited for uncontrolled scenes. Our articulated billboards provide an improved geometric shape approximation for challenging acquisition conditions, where methods based on accurate silhouettes, stereo correspondences, or calibration generally fail. Our semi-automatic, data-driven pose estimation and model computation provides a practical solution even in challenging setups, and our pixel-accurate processing results in high quality renderings with a realistic reproduction of the subject's appearance for novel views. We have shown results in a quality comparable to the source images from HD-TV broadcast cameras. With their simple representation, articulated billboards can be rendered highly efficiently and thus will be applicable even for mobile devices.

9. Acknowledgement

The data is courtesy of Teleclub and LiberoVision. We thank Thomas Oskam for helping with the video. This project is supported by a grant of CTI Switzerland.

References

[ABB*07] Andújar C., Boo J., Brunet P., Fairén M., Navazo I., Vázquez P., Vinacua A.: Omni-directional relief impostors. *Computer Graphics Forum 26*, 3 (2007), 553–560.

[ABT99] Aubel A., Boulic R., Thalmann D.: Lowering the cost of virtual human rendering with structured animated impostors. In *WSCG'99* (1999).

[AT06] Agarwal A., Triggs B.: Recovering 3d human pose from monocular images. *PAMI 28*, 1 (2006), 44–58.

[BBM*01] Buehler C., Bosse M., McMillan L., Gortler S., Cohen M.: Unstructured lumigraph rendering. In *SIGGRAPH '01* (2001), pp. 425–432.

[BCF*05] Behrendt S., Colditz C., Franzke O., Kopf J., Deussen O.: Realistic real-time rendering of landscapes using billboard clouds. *Comput. Graph. Forum 24*, 3 (2005), 507–516.

[CCST00] Chai J.-X., Chan S.-C., Shum H.-Y., Tong X.: Plenoptic sampling. In *SIGGRAPH '00* (2000), pp. 307–318.

[CTMS03] Carranza J., Theobalt C., Magnor M. A., Seidel H.-P.: Free-viewpoint video of human actors. In *SIGGRAPH '03* (2003), pp. 569–577.

[dAST*08] de Aguiar E., Stoll C., Theobalt C., Ahmed N., Seidel H.-P., Thrun S.: Performance capture from sparse multi-view video. In *SIGGRAPH '08* (2008), pp. 1–10.

[DDS03] Décoret X., Durand F., Sillion F. X.: Billboard clouds. In *SCG '03* (2003), pp. 376–376.

[DTM96] Debevec P. E., Taylor C. J., Malik J.: Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. *SIGGRAPH'96* (1996), 11–20.

[EBMM03] Efros A. A., Berg A. C., Mori G., Malik J.: Recognizing action at a distance. In *ICCV* (2003), pp. 726–733.

[Eye09] EyeVision:. http://www.ri.cmu.edu/events/sb35/tksuperbowl.html (2009).

[GEJ*08] Gammeter S., Ess A., Jaeggli T., Schindler K., Leibe B., Gool L. J. V.: Articulated multi-body tracking under egomotion. In *ECCV (2)* (2008), pp. 816–830.

[GGSC96] Gortler S. J., Grzeszczuk R., Szeliski R., Cohen M. F.: The lumigraph. In *SIGGRAPH '96* (1996), pp. 43–54.

[GKH09] Guillemaut J.-Y., Kilner J., Hilton A.: Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes. In *ICCV* (2009).

[GM03] Goldluecke B., Magnor M.: Real-time microfacet billboarding for free-viewpoint video rendering. In *ICIP'03* (2003), vol. 3, pp. 713–716.

[GTH*07] Grau O., Thomas G. A., Hilton A., Kilner J., Starck J.: A robust free-viewpoint video system for sport scenes. In *3DTV* (2007).

[HDK07] Hornung A., Dekkers E., Kobbelt L.: Character animation from 2D pictures and 3D motion data. *ACM Trans. Graph. 26*, 1 (2007).

[HS06] Hayashi K., Saito H.: Synthesizing free-viewpoint images from multiple view videos in soccer stadium. In *CGIV* (2006), pp. 220–225.

[JKMG07] Jaeggli T., Koller-Meier E., Gool L. J. V.: Learning generative models for monocular body pose estimation. In *ACCV* (2007), pp. 608–617.

[KSHG07] Kilner J., Starck J., Hilton A., Grau O.: Dual-mode deformable models for free-viewpoint video of sports events. *3DIM* (2007), 177–184.

[LBDGG05] Lee O., Bhushan A., Diaz-Gutierrez P., Gopi M.: Capturing and view-dependent rendering of billboard models. In *ISVC* (2005), pp. 601–606.

[LCF00] Lewis J. P., Cordner M., Fong N.: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH '00* (2000), pp. 165–172.

[LH96] Levoy M., Hanrahan P.: Light field rendering. In *SIGGRAPH '96* (1996), pp. 31–42.

[Lib] LiberoVision:. www.liberovision.com.

[MBR*00] Matusik W., Buehler C., Raskar R., Gortler S. J., McMillan L.: Image-based visual hulls. In *SIGGRAPH '00* (2000), pp. 369–374.

[MHM*09] Mahajan D., Huang F.-C., Matusik W., Ramamoorthi R., Belhumeur P. N.: Moving gradients: a path-based method for plausible image interpolation. *ACM Trans. Graph. 28*, 3 (2009).

[Mid09] Middlebury multi-view stereo evaluation. http://vision.middlebury.edu/mview/, October 2009.

[NAS09] NASA: Anthropometry and biomechanics. http://msis.jsc.nasa.gov/sections/section03.htm (2009).

[PG08] Pekelny Y., Gotsman C.: Articulated object reconstruction and markerless motion capture from depth video. *Comput. Graph. Forum 27*, 2 (2008), 399–408.

[RMD04] Reche A., Martin I., Drettakis G.: Volumetric reconstruction and interactive rendering of trees from photographs. *SIGGRAPH'04 23*, 3 (July 2004).

[SGwHS98] Shade J., Gortler S., wei He L., Szeliski R.: Layered depth images. In *SIGGRAPH'98* (1998), pp. 231–242.

[Tho06] Thomas G.: Real-time camera pose estimation for augmenting sports scenes. *Visual Media Production, CVMP* (2006), 10–19.

[VBMP08] Vlasic D., Baran I., Matusik W., Popović J.: Articulated mesh animation from multi-view silhouettes. In *SIGGRAPH '08* (2008), pp. 1–9.

[WWG07] Waschbüsch M., Würmlin S., Gross M.: 3d video billboard clouds. *Comput. Graph. Forum 26*, 3 (2007), 561–569.

[YSK*02] Yamazaki S., Sagawa R., Kawasaki H., Ikeuchi K., Sakauchi M.: Microfacet billboarding. In *EGRW '02* (2002), pp. 169–180.